agronomy

MDPI

Article

Farmland Soil pH in Guizhou, China: Spatial Patterns and Associated Factors

Qinyin Pang [1], Changlin Xu [2], Rongguo Sun [1,3,*] and Yu Li [3,4]

[1] Guizhou Key Laboratory for Mountainous Environmental Information and Ecological Protection, Guizhou Normal University, Guiyang 550025, China
[2] School of Chemistry and Materials Sciences, Guizhou Normal University, Guiyang 550001, China
[3] Guizhou Key Laboratory of Cultivated Land Quality, Guiyang 550025, China
[4] Institute of Soil and Fertilizer, Guizhou Academy of Agricultural Sciences, Guiyang 550006, China
* Correspondence: srg@gznu.edu.cn

Abstract

The spatial structure and multivariable associations of farmland soil pH remain poorly resolved in mountainous landscapes. We analyzed 487 topsoil samples (0–30 cm) from Guizhou, China, using spatial autocorrelation, geostatistics, random forest (RF), and structural equation modeling (SEM). Soil pH ranged from 4.37 to 8.45, with 17.9% of samples below the pH 5.5 screening threshold. Because sampling was not area-weighted, this percentage does not estimate province-wide areal prevalence. Coordinate-level means of pH, soil organic matter (SOM), and cation exchange capacity (CEC) exhibited modest positive global spatial autocorrelation. Local pH evidence after multiplicity correction was limited and neighborhood-dependent. The pH semivariogram indicated short-range dependence and substantial unresolved variation. RF showed modest performance under random and spatially blocked nested cross-validation, with higher errors under spatial validation. CEC ranked first in importance, followed by climate- and terrain-related variables. SEM explained approximately one-fifth of pH variance. Within the model, CEC had the largest standardized direct coefficient in the pH equation, SOM showed a positive indirect association with pH through CEC, and an ordered landform-relief score showed a negative total conditional association. Climate-related pathways yielded small, nonsignificant total associations. These findings inform regional screening and field-level diagnosis of soil acidity, nutrient availability, and exchange status.

Keywords: soil pH; spatial autocorrelation; geostatistics; random forest; structural equation model; cation exchange capacity

Check for updates

Academic Editor: Witold Grzebisz

Received: 20 September 2026
Revised: 21 October 2026
Accepted: 21 October 2026
Published: 25 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Soil pH is a central indicator of soil chemical conditions because it influences nutrient availability, metal solubility, microbial processes, and crop performance [1,2]. For example, pH-dependent sorption and precipitation reactions involving Fe, Al, and Ca influence phosphorus availability, with responses that vary among soils [3]. Interpreting pH together with nutrient supply and crop requirements is therefore relevant to sustainable soil fertility management. Long-term acidifying inputs can lower cropland pH and increase exchangeable acidity, although the magnitude and agronomic consequences of these changes depend on soil properties, crop requirements, and management history [4,5]. At regional scales, soil pH is spatially heterogeneous and covaries with pedogenic conditions, climate, terrain, and land use [6–8]. Characterizing this heterogeneity is particularly challenging in mountainous

agricultural regions, where dissected terrain, strong elevation gradients, heterogeneous parent materials, and differences in erosion, runoff, and drainage can modify soil and base-cation redistribution and contribute to variation among nearby fields [6–8]. Such spatial complexity increases uncertainty when measurements from individual fields are generalized across broader agricultural landscapes and creates a need to distinguish regional spatial organization from variation that remains unresolved at shorter distances [9].

Regional farmland pH may be associated with broad-scale environmental conditions, proximal soil properties, and field-management practices [10]. At broad spatial scales, water balance is associated with weathering, leaching, and the redistribution of base cations and may therefore help structure geographical pH patterns [11,12]. Terrain modifies drainage, hydrological redistribution, erosion, deposition, and local climatic conditions and has been associated with spatial variation in soil chemical properties and pH [6,7,13]. Proximal soil properties provide a more direct representation of the soil matrix at individual sampling locations [14]. Cation exchange capacity (CEC) characterizes the capacity of soil to retain exchangeable cations, whereas soil organic matter (SOM) contributes to cation-exchange properties and pH buffering [15,16]. However, exchange capacity alone does not specify the proportions occupied by base and acidic cations and cannot substitute for measurements of exchangeable acidity or base saturation.

Field management can modify soil acid–base balances through several processes. Nitrification of fertilizer-derived ammonium generates protons, while the uptake and removal of base cations by crops can contribute to acidification [4]. Organic amendments can supply base cations and modify exchange and buffering properties, although their net influence on pH depends on material composition and subsequent transformations [4,17]. Irrigation can introduce alkalinity and dissolved ions, with consequences that depend on water chemistry and application amounts [13,18,19]. Cropping history influences the cumulative balance between nutrient inputs and removal in harvested biomass [20,21]. The direction and magnitude of management-related pH changes therefore depend on input rates, material properties, water chemistry, and management duration. Because environmental, soil, and management variables may covary spatially, their information can overlap within multivariable models, influencing predictor rankings and conditional associations [22,23].

Previous regional studies have examined soil pH variation in relation to environmental and land-use changes [6,7], estimated pH from climatic variables [8], and developed detailed cropland pH maps [10]. For heterogeneous mountain farmland, a further need is to compare the predictive information associated with soil properties, environmental gradients, and management categories with the conditional associations represented within a soil-process model. The reliability of pH prediction when spatial blocks are withheld also requires explicit evaluation. Spatial autocorrelation and geostatistical analyses characterize geographical clustering and distance-dependent variation [6,8]. Random forest (RF) regression ranks model-dependent predictive information while accommodating flexible and potentially nonlinear relationships, but permutation importance does not indicate association direction or uniquely attributable effects [22]. Structural equation modeling (SEM) evaluates specified linear direct, indirect, and total conditional associations within a theory-guided path structure [24]. This study addresses these questions within the same regional survey, allowing pH patterns to be interpreted alongside covariate associations and prediction errors under spatial holdout.

Guizhou Province, Southwest China, provides a relevant setting for examining these associations, as farmland is distributed across valley bottoms, basins, and hillslopes with contrasting soil backgrounds and management conditions [25,26]. We evaluated two related hypotheses: (1) regional farmland soil pH variation is spatially structured and associated with measured soil properties and environmental gradients; and (2) pH contrasts among

predominant fertilization and water-supply regimes remain evident after accounting for the included soil and environmental covariates. We analyzed 487 farmland topsoil samples collected from the 0–30 cm layer across the province. The study aimed to characterize the spatial heterogeneity of farmland pH, examine its associations with soil properties, climate, terrain, pedogenic background, and field management, and evaluate the direct and indirect associations represented within the proposed soil-process structure. The resulting regional evidence is intended to guide follow-up sampling and identify soil and management characteristics that warrant more detailed field-level diagnosis. The results are interpreted as complementary spatial, predictive, and conditional associations within a cross-sectional observational framework.

2. Materials and Methods

2.1. Study Area

The study was conducted in Guizhou Province, Southwest China (103°36′–109°35′ E, 24°37′–29°13′ N), which covers approximately 176,000 km^2 [25] (Figure 1). Mountains and hills account for approximately 92.5% of the provincial area [26]. Elevation ranges from approximately 148 to 2901 m above sea level (a.s.l.) and generally decreases from west to east [27]. The province has a humid subtropical monsoon climate, with a mean annual temperature (MAT) of approximately 16.4 °C, mean annual precipitation (MAP) of approximately 1218 mm, and annual sunshine duration (SSD) of approximately 1200 h [28,29]. Regional geology is characterized by widespread carbonate rocks, particularly limestone and dolomite, and noncarbonate lithologies such as sandstone and mudstone [30,31]. Soil parent materials include the weathering products of these contrasting lithologies [30,31]. Major regional soil types include yellow, red, yellow-brown, limestone, purple, and paddy soils [31].

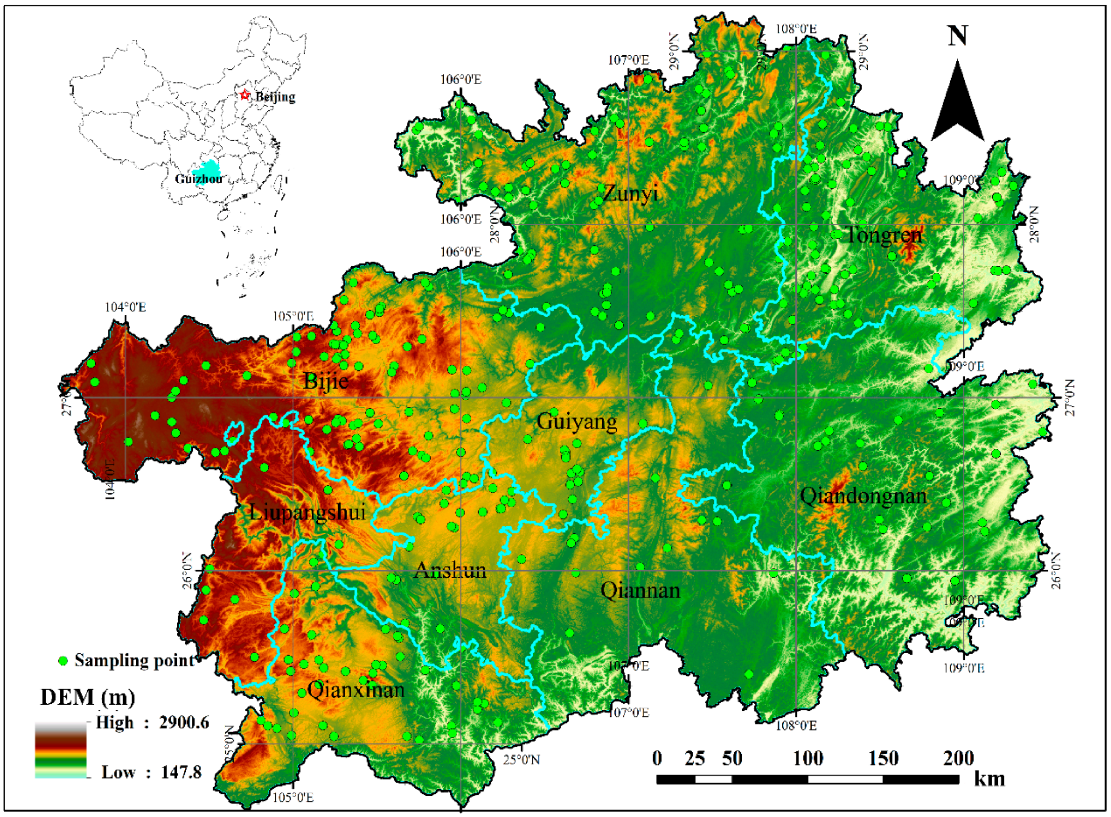

Figure 1. Location of Guizhou Province in China, provincial digital elevation model (DEM), and recorded locations of the 487 farmland topsoil samples.

Farmland is distributed discontinuously across valley bottoms, basins, hillslopes, and mountain areas and supports maize, rice, vegetables, and tobacco production [32]. The close spatial juxtaposition of cultivated fields across contrasting terrain, climatic conditions, soil groups, and management categories makes Guizhou suitable for examining the spatial dependence and multivariable associations of farmland soil pH across a heterogeneous mountainous agricultural landscape.

2.2. Soil Sampling and Laboratory Analyses

A coverage-oriented field-sampling design was used to span major farmland areas, topographic settings, soil backgrounds, and agricultural-use conditions across Guizhou Province. Fields were distributed across all nine prefecture-level divisions of the province, with sampling intensity higher in the main agricultural basins and lower in steep, remote, or poorly accessible terrain; the uneven distribution shown in Figure 1 therefore reflects the areal distribution and accessibility of farmland in this mountainous province rather than a probability-based allocation. The number of fields sampled in each prefecture-level division is summarized in Table S1. Candidate fields were selected to maximize environmental coverage while avoiding field margins, roads, residential areas, and visibly disturbed non-agricultural locations. At the time of field selection, the predominant cropping system, field water-supply regime, and fertilization regime of each candidate field were recorded during the sampling campaign (Table S2), so that every field-level composite sample could be linked directly to its corresponding field-scale management record. Soil sampling was conducted during the main cropping season from July to August 2023. By this sampling window, basal and early top-dressing nitrogen applications for summer maize and transplanted rice had generally been applied approximately 1–2 months earlier; nevertheless, vegetable fields receive repeated split fertilization, and short-term residual effects of recent nitrogen inputs on topsoil pH cannot be fully excluded.

At each sampled field, defined as a contiguous, uniformly managed plot under a single crop, with pairwise distances among sampling locations not exceeding approximately 50 m, approximately 10 sampling locations were established across the field following a plum-blossom sampling pattern [33]. The 0–30 cm plow layer was sampled to capture the rooting zone of the dominant crops, including rice, whose roots extend relatively deep into the topsoil, following the sampling depth used in comparable paddy and upland soil surveys [34]. At each location, five soil subsamples were collected from this layer at the four corners and center of a 1 m $\times$ 1 m square using a stainless-steel soil auger [35]. The 1 m $\times$ 1 m frame was used only to standardize the within-point micro-plot at each of the approximately 10 field locations; it was not the spatial extent of the composite sample itself. All subsamples collected within the same field were pooled and thoroughly mixed to form a single field-level composite sample. This procedure yielded 487 composite farmland topsoil samples. Geographic coordinates associated with each composite sample were recorded in the WGS 84 geographic coordinate system using a Trimble Juno SB handheld GPS receiver.

Samples were placed in clean polyethylene bags and transported to the laboratory. Visible roots, stones, and other coarse debris were removed, after which the soils were air-dried, gently ground, homogenized, and passed through a 2-mm sieve. Soil pH was measured potentiometrically in a 1:2.5 soil-to-deionized-water suspension using a calibrated pH meter (PHS-3C, Shanghai INESA Scientific Instrument Co., Ltd., Shanghai, China), following NY/T 1377-2007 [36]. For SOM determination, an aliquot of each air-dried sample was further ground and passed through a 0.25-mm sieve. SOM was determined using the potassium dichromate oxidation method following NY/T 1121.6-2006 [37]. SOM was

calculated from the organic carbon determined by this method using a conversion factor of 1.724.

CEC was determined using the hexamminecobalt(III) chloride method following the Chinese environmental protection standard HJ 889-2017 [38]. Briefly, a given mass of air-dried soil passing through the 2-mm sieve was mixed with 25 mL of $[Co(NH_3)_6]Cl_3$ solution in a 50-mL centrifuge tube, shaken for 1 h, and centrifuged at 4000 rpm for 10 min. The residual concentration of hexamminecobalt(III) ions in the supernatant was quantified by UV–Vis spectrophotometry (UV-5500, Shanghai Metash, Shanghai, China), and CEC was calculated from the difference between the initial and residual solution concentrations and expressed as cmol(+) kg^{-1}. Detailed instrument calibration and analytical quality-control procedures, including triplicate analytical determinations and certified-reference-soil checks, are provided in Text S1. Field sampling, sample preparation, laboratory measurements, and the analytical use of the sampling-location data are summarized in Table S3.

2.3. Environmental Covariates and Agricultural Management Information

Topographic covariates were derived from ASTER Global Digital Elevation Model version 3 (ASTER GDEM v3; NASA/METI, 2019), obtained through the Geospatial Data Cloud (https://www.gscloud.cn, accessed on 19 September 2026). The source dataset has a nominal resolution of 1 arc-second, approximately 30 m. Before terrain derivatives were calculated, the DEM was projected to a CGCS2000-based Albers equal-area conic coordinate system, with a central meridian of 105° E and standard parallels of 25° N and 47° N, using a 30-m grid. Slope and aspect were derived from the projected DEM in ArcMap 10.8.1, and aspect was grouped into north-, east-, south-, and west-facing classes. Local relief amplitude (R) was calculated as the difference between the maximum and minimum elevations within a circular moving window with a radius of 1000 m. The topographic position index (TPI) was calculated as the elevation of each cell minus the mean elevation within the same neighborhood, expressed as an unstandardized elevation difference in metres [39,40]. A simplified relief-based landform classification, informed by Zhou et al. (2009) [41], used the following study-specific operational criteria: pooled plain/basin, $R < 75$ m and TPI ≤ 0; hill, $75 \leq R < 200$ m or $R < 75$ m and TPI > 0; low-relief mountain, $200 \leq R \leq 500$ m; moderate-relief mountain, $500 < R \leq 1000$ m; and high-relief mountain, $R > 1000$ m. These GIS-derived categories were subsequently encoded separately for RF and SEM, as described in Section 2.4. Mapped soil type was obtained from the 1-km China Soil Type Spatial Distribution Dataset, a rasterization of the 1:1,000,000 Soil Map of China compiled from the second national soil survey (Table S4).

Long-term MAT, MAP, and potential evapotranspiration (PET) were derived from 1-km monthly products for 2000–2023, whereas SSD was derived from a 4-km daily product for 2000–2020 [42]. Monthly temperature was averaged within each year, whereas monthly precipitation and PET and daily SSD were summed annually. Annual values were then averaged over the corresponding periods. These variables describe long-term climatic conditions. Sources, native spatial resolutions, temporal coverage, derivation procedures, and preprocessing of all raster-based environmental covariates, including the landform-classification framework, are summarized in Table S4.

All covariate rasters were projected to the same CGCS2000-based Albers coordinate system, clipped to Guizhou Province, and aligned to a common 250-m grid before values were extracted at the recorded sample coordinates. Terrain derivatives and landform classification were calculated on the projected 30-m DEM before this alignment. Bilinear interpolation was used for continuous layers, including elevation, slope, relief amplitude, TPI, and climatic variables, whereas nearest-neighbor resampling was used for categorical

layers, including landform, aspect classes, and source soil-map classes. Raster processing was conducted in ArcMap 10.8.1.

Agricultural-management information was compiled from field survey records collected during the 2023 sampling campaign (Table S5). The predominant cropping system was classified as vegetable-dominated cropping, rice, maize-vegetable rotation, maize, or tobacco. Predominant field water-supply regime was classified as rainfed, surface-water-supplied, or tap-water-supplied (i.e., fields irrigated from centralized treated domestic-water pipelines; this category occurred mainly in peri-urban or village settings and served as the SEM reference because it represented the most seasonally reliable water supply), and predominant fertilization regime as mineral-fertilizer-only management, organic-manure-only management, or combined mineral–organic fertilization. These variables represented predominant field-level management categories.

2.4. Data Preparation and Analytical Variable Representation

The sample-level dataset contained 487 records with no missing observations. No imputation was performed. Continuous variables entered the RF and SEM analyses on their original measurement scales without transformation, centering, or standardization. Fully standardized SEM coefficients were obtained after fitting. Variable definitions, codes, reference categories, and analytical roles are provided in Text S2 and Table S6.

The RF analysis included north–south and east–west aspect contrasts and two crop-related predictors. The rice indicator distinguished rice cultivation from the other cropping systems. The second predictor was an a priori crop-management ordering score: vegetable-dominated cropping = 0.20, rice = 0.35, maize–vegetable rotation = 0.55, maize = 0.75, and tobacco = 1.00. The score encoded an assumed ordering and did not measure proton production or an acidification rate. The ordering reflected a priori assumptions about nitrogen inputs and cultivation practices associated with the five cropping systems. Quantitative field-level fertilizer inputs and management histories were unavailable to validate this ordering. It was specified before RF fitting and was not calibrated against the observed pH values. Its intervals and ratios had no chemical interpretation. The two crop predictors partly overlapped and their importance values were not summed. Neither was included in the candidate SEMs.

Source soil-map classes were consolidated into six analytical soil groups. Five binary indicators represented limestone, paddy, pooled Skeletol-Litho, mountain meadow, and purple soil groups. Yellow, red, and yellow-brown soils formed the pooled all-zero category. This was the SEM reference group and one predictor configuration in RF. These mapped analytical groups describe pedogenic background; they are not direct measurements of parent material, carbonate content, or clay mineralogy.

Water supply was represented by rainfed and surface-water-supply indicators, with tap-water-supplied fields forming the SEM reference category. RF used two non-mutually exclusive indicators for the presence of mineral fertilizer and organic manure. SEM instead used mutually exclusive contrasts for mineral-fertilizer-only and organic-manure-only management relative to combined mineral–organic fertilization. The five GIS-derived landform categories were represented in RF by an ordered relief-intensity score: pooled plain/basin = 0, hill = 1, low-relief mountain = 2, moderate-relief mountain = 3, and high-relief mountain = 4. For SEM, the accumulation-landform indicator was coded as 1 for the pooled plain/basin category and 0 for hills and all mountain categories. The primary erosional-relief score was coded as 0 for the pooled plain/basin and hill categories, 1 for low-relief mountains, and 2 for moderate- and high-relief mountains. The highest SEM level combined 97 moderate-relief observations with four high-relief observations, yielding 101 observations. Observation counts and the alternative 0–3 SEM coding used for

sensitivity analysis are reported in Table S7. These scores represented predefined terrain trends rather than measured erosion rates or equally spaced physical increments in relief. Spatial distributions of selected environmental covariates and the source soil-map classes are shown in Figure S1.

2.5. Descriptive Statistics and Categorical Group Comparisons

All descriptive and categorical comparison analyses used the 487 original sample-level observations; no coordinate-level averaging was applied to these analyses. Continuous variables were summarized using the mean, standard deviation, median, minimum, maximum, skewness, and excess kurtosis (Table S8). Source categorical variables and the final analytical soil groups were summarized using frequencies and percentages (Table S9), and frequencies of the derived binary indicators and ordered landform variables are reported in Tables S10 and S7, respectively.

Exploratory pH differences among analytical soil groups and the fertilization, water-supply, and cropping categories were assessed using Welch's one-way ANOVA because group sizes were unequal and equal variances were not assumed [43]. Omnibus tests meeting $\alpha = 0.05$ were followed by Games–Howell pairwise comparisons [44]. Tobacco observations ($n = 3$) were retained in descriptive summaries and figures but excluded from the inferential cropping comparison, which included 484 observations. Analyses were conducted in R 4.5.3; results are reported in Table S11. These comparisons did not adjust for covariates or spatial dependence. No multiplicity adjustment was applied across the four omnibus tests; Games–Howell probabilities accounted for pairwise comparisons within each tested factor.

2.6. Spatial Autocorrelation and Geostatistical Analysis

Spatial autocorrelation analyses used coordinate-level mean pH, SOM, and CEC, whereas semivariogram modeling and ordinary kriging were restricted to soil pH. Where multiple samples shared an identical recorded coordinate, their pH, SOM, and CEC values were averaged separately, yielding 365 coordinate-level analytical units, so that each recorded coordinate contributed once to the location-based spatial analyses. This preprocessing was restricted to spatial analyses; all 487 original sample-level observations were retained for descriptive analyses, categorical comparisons, RF analysis, and SEM. Recorded coordinates were assigned to WGS 84 and transformed to a local azimuthal-equidistant projected coordinate system centered at 26.8° N and 106.5° E before spatial distances were calculated.

Neighbor-distance and graph-connectivity diagnostics were completed before Moran statistics were calculated. Fixed-distance weighting was not retained because the minimum distance required to eliminate all isolates was 67 km, resulting in an overly broad neighborhood with a median of 32 neighbors per coordinate-level analytical unit. The primary analysis therefore used a symmetrized six-nearest-neighbor graph with row-standardized weights. Four- and eight-nearest-neighbor graphs were retained as lower- and higher-neighborhood sensitivity specifications. These specifications were selected from connectivity and link-distance diagnostics before the Moran statistics were examined. Because the neighbor graph was symmetrized, some coordinate-level analytical units had more than the nominal k neighbors.

Global Moran's I was used to evaluate overall spatial autocorrelation in soil pH, SOM, and CEC under all three neighborhood specifications [45,46]. Analytical two-sided tests were based on the randomization expectation and variance, and permutation-based inference used 9999 random permutations. The reported finite-permutation two-sided probability was calculated by doubling the smaller plus-one-corrected tail probability,

with the result capped at 1. The minimum attainable value under this procedure was $2/(9999 + 1) = 0.0002$. Results attaining this minimum were reported as $p \leq 0.0002$ in the text, figures, and tables. Numerical permutation probabilities were retained for the Benjamini–Hochberg adjustment. For SOM and CEC, the two permutation probabilities were adjusted using the Benjamini–Hochberg procedure separately within each neighborhood specification.

Local Moran's I for soil pH was evaluated under the primary $k = 6$ specification using 9999 conditional permutations [47]. High–high, low–low, high–low, and low–high categories were assigned from the signs of standardized pH and its spatial lag. Rank-based two-sided permutation probabilities were adjusted across all coordinate-level tests using the Benjamini–Hochberg procedure. Classifications based on unadjusted $p < 0.05$ were treated as exploratory, whereas BH-adjusted $q < 0.05$ was used to evaluate multiplicity-controlled local evidence. Local classifications were additionally compared under $k = 4$ and $k = 8$.

For soil pH, an omnidirectional empirical semivariogram was calculated using a 2.5-km lag interval and a primary cutoff distance of 70 km. Spherical, exponential, and Gaussian models were fitted by weighted least squares and evaluated jointly using parameter admissibility, sensitivity to cutoff distances of 30, 50, and 70 km, and leave-one-coordinate-out ordinary-kriging cross-validation. Cross-validation and interpolation used a maximum of 30 neighboring coordinate-level observations for each prediction. The selected model was used for ordinary kriging to visualize broad regional variation in soil pH.

Global and Local Moran analyses were conducted in R using the sf and spdep packages. Geostatistical analyses were conducted using the gstat package (version 2.1-5). Coordinate aggregation and spatial-weights diagnostics are reported in Text S3 and Table S12. Global Moran results for soil pH are presented in Figure 2 and Table S13, whereas exploratory Local Moran classifications and neighborhood-sensitivity results for soil pH are presented in Figure S2 and Table S14. Global Moran results for SOM and CEC are presented in Figure S3 and Table S15. Semivariogram modeling, cutoff sensitivity, and kriging cross-validation are described in Text S3, Figure S4, and Table S16.

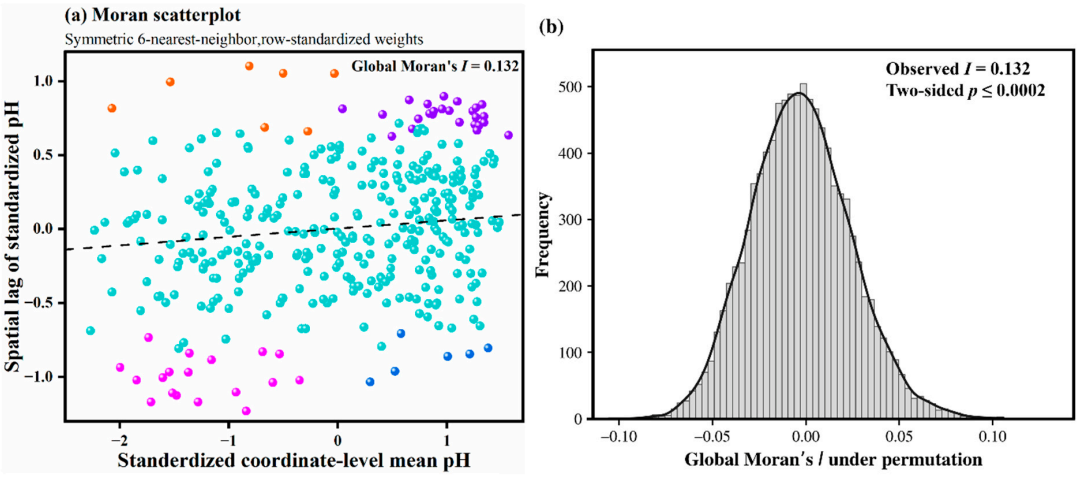

Figure 2. Global spatial autocorrelation of coordinate-level mean farmland soil pH. (**a**) Moran scatterplot under the primary symmetrized six-nearest-neighbor graph with row-standardized weights. The slope of the dashed regression line corresponds to Global Moran's I. (**b**) Reference distribution of Global Moran's I generated from 9999 random permutations. The vertical dashed line indicates the observed value ($I = 0.132$). The two-sided permutation probability was reported as $p \leq 0.0002$, indicating that the calculated value attained the minimum available under the 9999-permutation procedure.

2.7. RF Analysis

RF regression was applied to all 487 sample-level observations to rank the model-dependent predictive information associated with the 22 predictors listed in Table S6, using measured soil pH as the continuous response. Analyses were conducted in R 4.5.3 using the caret, randomForest, and rfPermute packages, with sf and blockCV used for the additional spatial validation. Full implementation details are provided in Text S4.

The number of predictors randomly sampled at each split (mtry) was evaluated from 1 to 22 using a fixed 10-fold random cross-validation partition, with the number of trees fixed at ntree = 1000. The value producing the lowest mean validation root mean square error (mtry = 5) was used to fit the final RF model to all observations for predictor-importance assessment. The original cross-validated out-of-fold and full-model out-of-bag (OOB) metrics were retained as internal resampling diagnostics (Tables S17 and S18, Figure S5). The original cross-validation was non-nested because the same folds were used for tuning and performance summarization. Neither this random partition nor OOB evaluation imposed spatial separation.

Additional random and spatial nested cross-validation analyses were conducted using five outer folds and four inner folds, with the same 487 observations, 22 predictors, mtry candidates, and number of trees. For spatial validation, recorded coordinates were transformed to the local azimuthal-equidistant projection. Hexagonal blocks were generated using a nominal block-size parameter of 20 km and randomly assigned to folds using sample-count balance, without selecting assignments according to predictive performance. All observations within a block, including those sharing an identical recorded coordinate, remained in the same fold. Coordinates were used only to construct the partitions; observations were not averaged, and coordinates were not added as predictors.

For each outer spatial training set, a new four-fold inner spatial partition was constructed using the same block-size parameter. The random nested comparator used sample-level random partitions at both levels. Within each outer training set, mtry was selected by minimizing RMSE calculated from pooled inner out-of-fold predictions. The selected specification was then refitted to the complete outer training set and used to predict its held-out outer fold. Outer-validation responses were not used for tuning, and each observation received one outer out-of-fold prediction under each scheme. These analyses evaluated the RF tuning and fitting procedure rather than the single full-data forest used for importance ranking.

Nested-validation performance was summarized using fold-wise and pooled RMSE, MAE, squared Pearson correlation between observed and predicted values (R^2_{corr}), and the prediction-error-based coefficient ($R^2_{\text{pred}} = 1 - \sum_{i=1}^{n}\sum(y_i - \hat{y}_i)^2 / \sum_{i=1}^{n}\sum(y_i - \overline{y})^2$), where y_i and $\hat{y}_i$ are the observed and predicted pH values, respectively, and $\overline{y}$ is the mean observed pH of the evaluation set. These quantities were distinguished from the OOB pseudo-R^2. No spatial buffer was imposed; nearest-training distances and predictor coverage were examined to characterize the tested partitions.

Predictor importance was quantified as the unscaled increase in OOB mean squared error after permutation of each predictor, averaged across trees and extracted using scale = FALSE. The resulting scores were expressed on the squared-pH-error scale. Correlations among predictors and their overlapping information could affect importance scores and rankings (the bivariate correlation matrix of the continuous predictors is reported in Table S19). Importance was interpreted as model-dependent predictive information rather than a unique variance contribution, association direction, or causal effect.

Predictor-specific evidence was evaluated by comparing observed importance scores with reference distributions generated from 1000 response permutations under the final RF specification. In each permutation, soil pH was shuffled relative to the unchanged predictor

matrix, and an RF model was refitted with mtry = 5 and ntree = 1000. Empirical upper-tail probabilities were adjusted across the 22 predictors using the Benjamini–Hochberg procedure [48]. Predictors with BH-adjusted $q < 0.05$ were described as meeting the prespecified nonspatial response-permutation criterion (Table S20). Unrestricted response permutations did not preserve spatial autocorrelation, and BH adjustment addressed multiple testing but did not correct for spatial dependence. The additional nested validation assessed predictive performance under the tested holdout designs; it did not spatially correct the importance probabilities or establish geographical transferability.

2.8. SEM

An observed-variable SEM was used to evaluate direct and model-defined indirect conditional associations of climatic, topographic, soil-property, pedogenic-background, and agricultural-management variables with farmland soil pH [49]. The SEM complemented RF by representing a theory-guided soil-process structure and was not derived mechanically from the RF importance ranking. All 487 sample-level observations were included, with SOM, CEC, and soil pH specified as continuous endogenous variables and soil pH as the final response. Models were fitted in R 4.5.3 using lavaan 0.6–21 and maximum likelihood estimation with robust standard errors and a scaled test statistic (MLR) [50]. Continuous variables entered on their original measurement scales, and fully standardized coefficients were obtained after model fitting. Complete specifications, model comparisons, diagnostics, and sensitivity analyses are provided in Text S5 and Tables S21–S25.

Model development began with a theory-guided three-equation structure and involved limited data-informed refinement. Modification indices and expected parameter changes were used as diagnostic information, and no automated or unrestricted path-search procedure was applied [51]. The MAP → CEC and erosional relief → CEC paths were the only regression paths added following these diagnostics; their inclusion also required a defensible soil-process basis and evaluation through nested-model comparison. Structural simplification and parameter constraints were evaluated using soil-process plausibility, model parsimony, and Satorra–Bentler 2001 scaled chi-square difference tests implemented using lavaan::lavTestLRT() for nested MLR models [52]. Each comparison used the same observations and observed-variable set. Joint block tests were interpreted as evidence concerning the corresponding set of parameters rather than every individual path. No regression path or residual covariance lacking a defensible soil-process basis was added solely to improve model fit. Model refinement and estimation used the same dataset, and the final specification was not validated in an independent sample.

The final structural model was specified as:

SOM ∼ MAT + PET + MAP + Accumulation landform + Mineral fertilizer only + Organic manure only.

CEC ∼ SOM + MAP + Erosional relief + Limestone soil + Paddy soil + Pooled Skeletol–Litho soil + Mountain meadow soil + Purple soil.

pH ∼ CEC + MAT + PET + Erosional relief + Rainfed + Surface water supply + Mineral fertilizer only + Organic manure only.

Following nested-model comparison, the direct SOM → pH path was omitted, and the modeled SOM-pH association was represented through CEC (Table S21). The pooled yellow–red–yellow-brown soil group, tap-water-supplied fields, and combined mineral–organic fertilization served as the reference categories for the soil-group, field water-supply, and fertilization contrasts, respectively (Table S6). In the pH equation, the unstandardized coefficients for rainfed and surface-water-supplied fields were constrained to equality relative to tap-water-supplied fields. This equality constraint applied only to the unstandardized coefficients and did not imply identical standardized associations or common underly-

ing processes. Inference for the common coefficient was conditional on the retained equality constraint and did not constitute two independent tests of water-supply associations.

Specific indirect associations with soil pH were calculated as products of the corresponding unstandardized path coefficients [53]. Where multiple indirect pathways were specified, their algebraic sum defined the net indirect association, and the total conditional association was calculated as the direct plus net indirect association. Robust standard errors, Wald tests, and 95% confidence intervals for defined indirect and total associations were obtained using the Delta method [54]. Fully standardized coefficients were used to compare association magnitudes, whereas statistical inference was based on the corresponding unstandardized parameters [55]. *p*-values for individual paths and model-defined associations were not adjusted for multiple testing. Inference was conditional on the selected specification, and near-threshold results were interpreted cautiously as exploratory evidence.

Model adequacy was evaluated jointly from robust global-fit indices, explained variance, local diagnostics, and sensitivity analyses rather than from a single-index criterion [56]. Sensitivity analyses assessed the original 0–3 erosional-relief coding, omission of the soil-group block, and individual deletion of candidate influential observations (Table S25). Because the study was cross-sectional and observational, the estimated paths were interpreted as conditional associations within the specified model. MLR reduced sensitivity to non-normality but did not account for dependence among observations at identical or nearby recorded coordinates. Residual spatial covariance was not explicitly modeled and could affect the uncertainty of the estimated associations. Model-defined indirect pathways did not independently establish temporal mediation or causal effects [24].

3. Results

3.1. Distribution and Spatial Dependence of Farmland Soil pH

Farmland soil pH across the 487 samples ranged from 4.37 to 8.45, with a mean of 6.78 $\pm$ 1.12 and a median of 7.04 (Table S8). The distribution was slightly negatively skewed (skewness = -0.37), reflecting a modest lower-pH tail while remaining centered near neutrality (Figure S6). Using pH < 5.5 as an initial screening threshold for potentially crop-limiting acidity, 17.9% of the samples fell below this value. Acidic conditions can increase the risk of Al toxicity and nutrient limitations in susceptible crops, including maize [57], although the occurrence and severity of these constraints depend on soil properties and crop requirements [4,5]. Because the sampling design was not area-weighted, this percentage describes the sampled observations and cannot be extrapolated to the proportion of acidic farmland area across Guizhou. Exchangeable Al, base saturation, and crop responses were not measured, so further field-level assessment is needed to establish the agronomic implications of these lower-pH conditions.

For the location-based analyses, coordinate-level mean soil pH exhibited modest but significant positive global spatial autocorrelation under the primary symmetrized six-nearest-neighbor graph with row-standardized weights (Moran's I = 0.132, analytical z = 5.016, analytical $p < 0.001$, and 9999-permutation two-sided $p \leq 0.0002$, Figure 2, Table S13). The direction and statistical evidence were retained under the four-nearest-neighbor (I = 0.150, permutation p = 0.0004) and eight-nearest-neighbor (I = 0.106, permutation $p \leq 0.0002$) sensitivity specifications, indicating that the evidence for positive global spatial autocorrelation was robust to the tested neighborhood sizes.

Under the primary k = 6 specification, rank-based conditional-permutation Local Moran analysis classified 56 coordinate-level analytical units at the unadjusted $p < 0.05$ threshold, comprising 25 high–high (HH), 18 low–low (LL), 6 high–low (HL), and 7 low–high (LH) classifications. None remained significant at BH-adjusted $q < 0.05$

across the local tests under this primary specification. Two units met this criterion under $k = 4$, whereas none did under $k = 8$, demonstrating the sensitivity of the multiplicity-controlled local findings to neighborhood specification. Of the 56 nominally classified units, 50 received the same HH, LL, HL, or LH classification under at least one of the $k = 4$ or $k = 8$ sensitivity specifications, whereas 22 received the same classification under all three neighborhood specifications. The unadjusted classifications were therefore treated as exploratory spatial diagnostics; no multiplicity-controlled local finding persisted across all three tested neighborhood specifications (Figure S2, Table S14).

SOM and CEC also exhibited positive global spatial autocorrelation of modest magnitude under the same primary specification ($I = 0.115$ and 0.121, respectively; two-sided permutation $p \leq 0.0002$ and BH-adjusted $q = 0.0002$ for both variables). Positive autocorrelation was retained under $k = 4$ and $k = 8$, with BH-adjusted $q \leq 0.0006$ after adjustment across SOM and CEC within each specification (Figure S3, Table S15). These separate univariate tests characterize the spatial structure of each property; they do not establish spatial coupling or causal direction among SOM, CEC, and pH.

The empirical semivariogram characterized a complementary, distance-based component of spatial dependence [58]. The spherical model was retained because its parameter estimates were stable across the evaluated cutoff distances and its cross-validated RMSE and MAE were marginally lower than those of the exponential and Gaussian models, although differences in predictive performance were small (Table S16, Figure S4). Under the primary 70-km cutoff, the selected model had a nugget accounting for 66.5% of the fitted total sill and a fitted range of 7.08 km, indicating that the modeled spatially structured component was concentrated primarily within approximately 7 km. The substantial nugget represented variation unresolved by the fitted model, potentially combining heterogeneity below the sampling resolution, coordinate-recording or analytical uncertainty, and other local variation. Their individual contributions could not be distinguished using the available data. Because the fitted semivariogram range left many coordinate-level analytical units without neighbors when evaluated as a fixed Moran distance threshold, it was not used to define the Moran spatial weights (Text S3, Table S12). The two approaches therefore summarized spatial structure under different distance supports: the primary Moran graph had a median link distance of 19.24 km, whereas 7.08 km was the distance at which the fitted spherical semivariogram reached its sill (Table S12) [59].

Leave-one-coordinate-out ordinary-kriging cross-validation yielded a mean error (ME) of -0.013, a root mean square error (RMSE) of 1.043, and a mean absolute error (MAE) of 0.901 on the pH scale (Table S16, Figure S4). The near-zero ME indicated little overall prediction bias, whereas the magnitudes of RMSE and MAE indicated only modest precision for individual locations within the sampled spatial domain. Cross-validated predictions were concentrated around intermediate pH values and tended to overpredict lower-pH observations and underpredict higher-pH observations (Figure S4c). An RMSSE of 0.971 and 97.5% coverage of the nominal 95% prediction intervals suggested broadly calibrated prediction uncertainty, with slightly conservative interval coverage within this internal validation (Table S16).

The ordinary-kriging surface suggested broadly lower predicted pH in southeastern Guizhou, with additional lower-pH patches in western areas, and relatively higher predicted values in parts of central and southwestern Guizhou (Figure 3). Given the cross-validation errors and the internal-validation design, the interpolated surface was interpreted as a spatially smoothed representation of broad regional variation among the sampled farmland locations rather than as a validated field-scale prediction product. The spatial analysis results indicate spatially organized pH variation across the study region, with modeled dependence concentrated at short distances and substantial variation re-

maining unresolved by the fitted semivariogram. This combination of regional clustering, short-range modeled dependence, and substantial unresolved variation is consistent with the spatial complexity reported for mountainous agricultural soils [6–8].

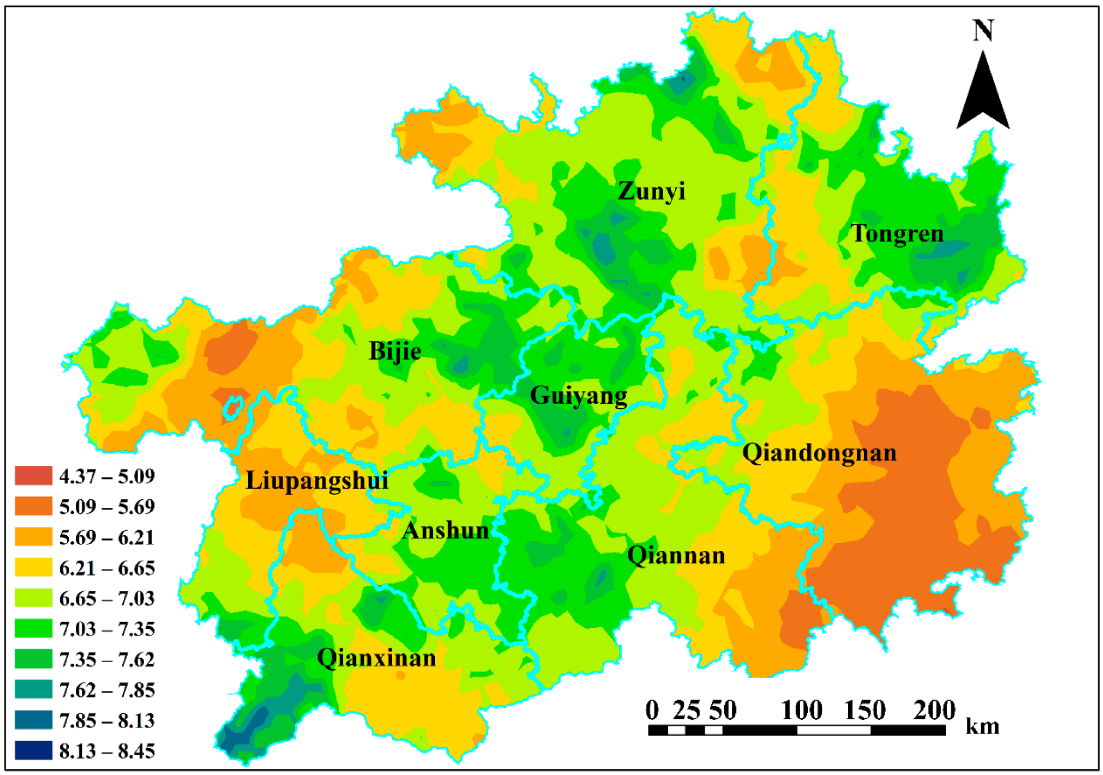

Figure 3. Ordinary-kriging surface of farmland soil pH across Guizhou Province based on the selected spherical semivariogram model. Colors represent interpolated values across the provincial extent and do not establish soil pH conditions for all farmland or non-farmland within that boundary. The interpolated surface is intended to represent broad regional variation among the sampled farmland locations rather than field-scale predictions for individual unsampled locations.

3.2. Unadjusted pH Contrasts Among Soil Groups and Management Regimes

Exploratory sample-level comparisons showed overall heterogeneity in farmland soil pH among the six final analytical soil groups (Welch's F (5, 48.52) = 4.934, $p < 0.001$, Table S11). Mean pH was descriptively highest in the pooled Skeletol–Litho soil group (7.44) and lowest in the purple soil group (6.12). Games–Howell comparisons showed that the pooled Skeletol–Litho group had a higher mean pH than the pooled yellow–red–yellow-brown group and the purple soil group by approximately 0.72 and 1.32 pH units, respectively. The limestone soil group also had a higher mean pH than the purple soil group by approximately 0.90 pH units. No other pairwise soil-group contrast met the $\alpha = 0.05$ criterion. The pooled Skeletol–Litho and pooled yellow–red–yellow-brown groups are analytical aggregations rather than formal taxonomic units, and estimates for the relatively small pooled Skeletol–Litho ($n = 19$) and mountain meadow ($n = 8$) groups were correspondingly less precise.

Mean pH also differed among fields classified by predominant fertilization regime (Welch's F (2, 259.75) = 5.237, $p = 0.006$, Table S11). In Games–Howell comparisons, fields under organic-manure-only management had a higher mean pH (7.04) than those under mineral-fertilizer-only management (6.61) and combined mineral–organic fertilization (6.76), corresponding to differences of approximately 0.43 and 0.28 pH units, respectively. Mineral-fertilizer-only and combined mineral–organic fields did not differ significantly.

Predominant field water-supply regime was also associated with unadjusted pH differences (Welch's F (2, 196.21) = 5.494, p = 0.005, Table S11). Tap-water-supplied fields had a higher mean pH (7.12) than rainfed fields (6.72) and surface-water-supplied fields (6.66), corresponding to differences of approximately 0.40 and 0.46 pH units, respectively. No significant difference was detected between rainfed and surface-water-supplied fields.

Across the four non-tobacco cropping-system categories (n = 484), the overall test was significant (Welch's F (3, 114.89) = 4.923, p = 0.003), but only vegetable-dominated cropping and maize–vegetable rotation differed significantly, with respective mean pH values of 6.92 and 6.37 (a difference of approximately 0.55 pH units). Neither rice nor maize differed significantly from either of these categories. The three tobacco observations were retained for descriptive reporting but were not included in the cropping-system inference.

These factor-specific comparisons were conducted separately and did not jointly adjust for measured soil properties, climate, topography, the other management factors, or spatial dependence. Games–Howell probabilities accounted for pairwise multiplicity within each factor, but no multiplicity adjustment was applied across the four omnibus tests. Nonsignificant pairwise contrasts were not interpreted as evidence of equivalence. Moreover, the management variables represented predominant field categories rather than quantified fertilizer inputs, irrigation amounts, water chemistry, or complete management histories. The observed contrasts should therefore be interpreted as exploratory, unadjusted, and non-spatially corrected sample-level associations rather than independent soil-group or management effects or causal responses. They do not establish changes in pH over time or quantify acidification rates. These comparisons provide descriptive context for the multivariable RF and SEM analyses below.

3.3. Random-Forest Performance and Predictor Importance

The mtry tuning profile showed only shallow differences among the candidate specifications. The numerical minimum in mean validation RMSE occurred at mtry = 5 (0.993927), although performance was nearly identical at mtry = 3 (RMSE = 0.993981) and remained broadly similar across mtry values of 3–7 (Table S17, Figure S5a). The full-data RF model used for predictor-importance assessment therefore retained mtry = 5 and ntree = 1000.

Under the original fixed, non-nested 10-fold sample-level random-resampling scheme, the fold-averaged metrics, reported as mean ± between-fold SD, were an RMSE of 0.9939 ± 0.0696 pH units, and a fold-wise squared Pearson correlation between observed and predicted pH, R^2_{corr}, of 0.2144 ± 0.0985. When the out-of-fold predictions for all 487 observations were pooled, the RMSE and MAE were 0.9959 and 0.8341 pH units, respectively, and the pooled R^2_{corr} was 0.2020. Full-model out-of-bag (OOB) evaluation yielded an RMSE of 0.9887 and an OOB pseudo-R^2 of 0.2131 (Table S18, Figure S5b,c). The OOB MSE curve approximately stabilized within the 1000-tree ensemble (Figure S5b). Because R^2_{corr} denotes the squared Pearson correlation between observed and predicted values rather than a conventional prediction-error-based coefficient of determination, it was reported separately from the OOB pseudo-R^2, calculated as $1 - \mathrm{MSE}_{OOB}/\mathrm{Var}(y)$. The two metrics were not treated as interchangeable estimates of model performance [60,61]. These original results indicate modest internal random-resampling performance [62,63]. The original 10-fold pooled out-of-fold predictions occupied a narrower range than the observations and tended to overpredict lower-pH observations and underpredict higher-pH observations, indicating limited discrimination toward the observed extremes (Figure S5c).

The additional nested analyses separated hyperparameter selection from performance evaluation using five outer folds and four inner folds. Random nested validation yielded a pooled outer out-of-fold RMSE of 1.0096 pH units, an MAE of 0.8500 pH units, R^2_{corr} = 0.1799, and a prediction-error-based R^2_{pred} = 0.1795. Under the nominal 20-km

hexagonal-block spatial specification, the corresponding values were 1.0475 pH units, 0.9012 pH units, 0.1174, and 0.1168, respectively (Table S18). Spatial nested validation therefore yielded slightly higher prediction errors and lower values of both performance indices than random nested validation. The mtry values selected within individual outer training sets ranged from 3 to 10 under random nesting and from 3 to 18 under spatial nesting (Text S4).

No identical recorded coordinates were shared between training and validation sets at either level of the spatial nested analysis. Across outer folds, median distances from validation observations to their nearest training observation were 12.15–16.55 km under spatial partitioning, compared with 1.81–4.38 km under random partitioning. Nevertheless, the spatial partition retained some close cross-fold neighbors, with a minimum distance of approximately 78 m, because no distance buffer was imposed (Text S4). The performance comparison describes the two tested validation designs and does not isolate the contribution of spatial dependence, since training-set composition and predictor coverage also differed. Predictive performance remained modest under the tested spatial holdout specification, and these results do not establish geographical transferability.

Ten of the 22 predictors met the prespecified nonspatial response-permutation criterion after BH adjustment ($q < 0.05$; Table S20). CEC ranked first, producing the largest unscaled permutation-induced increase in out-of-bag mean squared error (OOB MSE) (0.2509, BH-adjusted $q = 0.0110$, Figure 4, Table S20). Its importance was approximately three times that of the second-ranked predictor, SSD (0.0797, $q = 0.0220$). A relatively compact second tier comprised SSD, PET, elevation, MAT, MAP, the relief-intensity index, and slope, with importance values ranging from 0.0624 to 0.0797. All seven variables met the prespecified nonspatial response-permutation criterion after BH adjustment ($q < 0.05$). This pattern highlights CEC as the leading measured soil-property predictor within the fitted forest, while indicating that pH predictions also incorporated information from several climatic and topographic variables with relatively similar importance scores. The full bivariate correlation matrix of the continuous predictors is reported in Table S19. Because the climate–terrain predictors were moderately to strongly intercorrelated, non-conditional permutation importance may redistribute predictive information within correlated blocks, and closely spaced scores in this tier should not be read as uniquely attributable causal rankings [22].

Among the management-related representations, the a priori crop-management acidification proxy ranked tenth (importance = 0.0278, $q = 0.0418$), and the mineral-fertilizer component indicator ranked eleventh (importance = 0.0198, $q = 0.0220$). Both met the prespecified BH-adjusted nonspatial response-permutation criterion, although their importance scores were lower than those of CEC and the leading climatic and topographic predictors. These results indicate that the fitted forest drew on information associated with crop management and mineral-fertilizer use alongside the measured soil and environmental covariates. The crop-proxy result does not validate the predefined ordering or quantify acidification rates.

SOM ranked ninth by importance magnitude (0.0346) but did not meet the prespecified nonspatial response-permutation criterion after BH adjustment ($q = 0.0864$). The remaining water-supply, soil-group, aspect, rice-associated flooding, and organic-manure indicators likewise did not meet the prespecified nonspatial response-permutation criterion after BH adjustment (all $q \geq 0.05$). The contrasting criterion status of SOM and the two lower-ranked management predictors illustrates that numerical importance rank and BH-adjusted response-permutation evidence should be interpreted separately. For binary indicators, permutation importance reflects changes in average OOB MSE and can depend on the relative frequencies of the represented categories. Their scores were therefore interpreted in the context of the category counts reported in Tables S9 and S10.

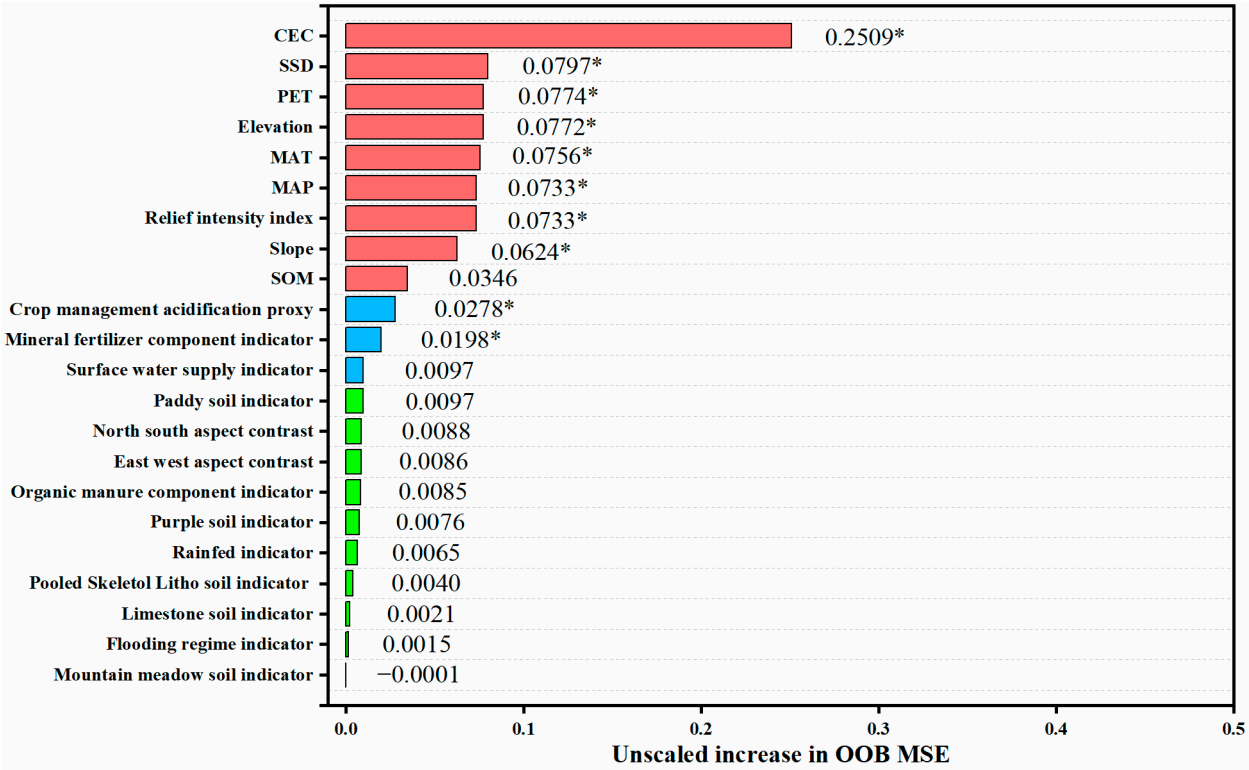

Figure 4. Unscaled permutation importance of the 22 predictors in the full-data RF model for farmland soil pH. Importance was quantified as the increase in out-of-bag mean squared error after predictor permutation under the selected RF specification (mtry = 5, ntree = 1000), averaged across trees and expressed on the squared-pH-error scale. Bars are ordered by decreasing importance. A single asterisk denotes BH-adjusted $q < 0.05$ based on 1000 response permutations, with adjustment across all 22 predictors; unmarked bars indicate $q \geq 0.05$. These probabilities were obtained under the selected nonspatial RF specification with mtry held fixed and should not be interpreted as spatially corrected inferential tests.

3.4. Sem-Derived Direct, Indirect, and Total Associations with Soil pH

The final SEM converged normally. The global and residual-based indices were consistent with an adequate approximation to the observed covariance structure: robust χ^2 (24) = 33.200, $p = 0.100$, robust comparative fit index (CFI) = 0.967, robust Tucker–Lewis index (TLI) = 0.939, robust root mean square error of approximation (RMSEA) = 0.028 (90% CI: 0.000–0.049), and standardized root mean square residual (SRMR) = 0.019 (Figure 5, Table S22). The model accounted for 18.1% of the observed variance in SOM, 14.4% in CEC, and 20.1% in soil pH. Thus, the specified structure captured a modest proportion of the sample-level variation, with substantial variance in all three endogenous variables remaining unexplained.

CEC had the largest absolute fully standardized direct coefficient in the pH equation ($\beta = 0.369$, $B = 0.062$ pH unit per cmol(+) kg^{-1}, $p < 0.001$, Table S23), consistent with its highest importance in the RF analysis. Within the specified model, samples with greater CEC therefore tended to have higher pH after conditioning on the remaining predictors. CEC, however, measures the capacity of the exchange complex rather than its occupancy by base or acidic cations. Exchangeable bases, exchangeable acidity, and base saturation were not measured [64]. Because CEC can vary with pH through pH-dependent charge development, the CEC → pH arrow represents the assumed structural ordering; its direction was not established by the cross-sectional observations [24,65]. Among the other direct paths to pH, MAT had a small positive coefficient ($\beta = 0.100$, $p = 0.022$), whereas PET ($\beta = -0.099$, $p = 0.018$) and the erosional-relief score ($\beta = -0.168$, $p < 0.001$) were negative.

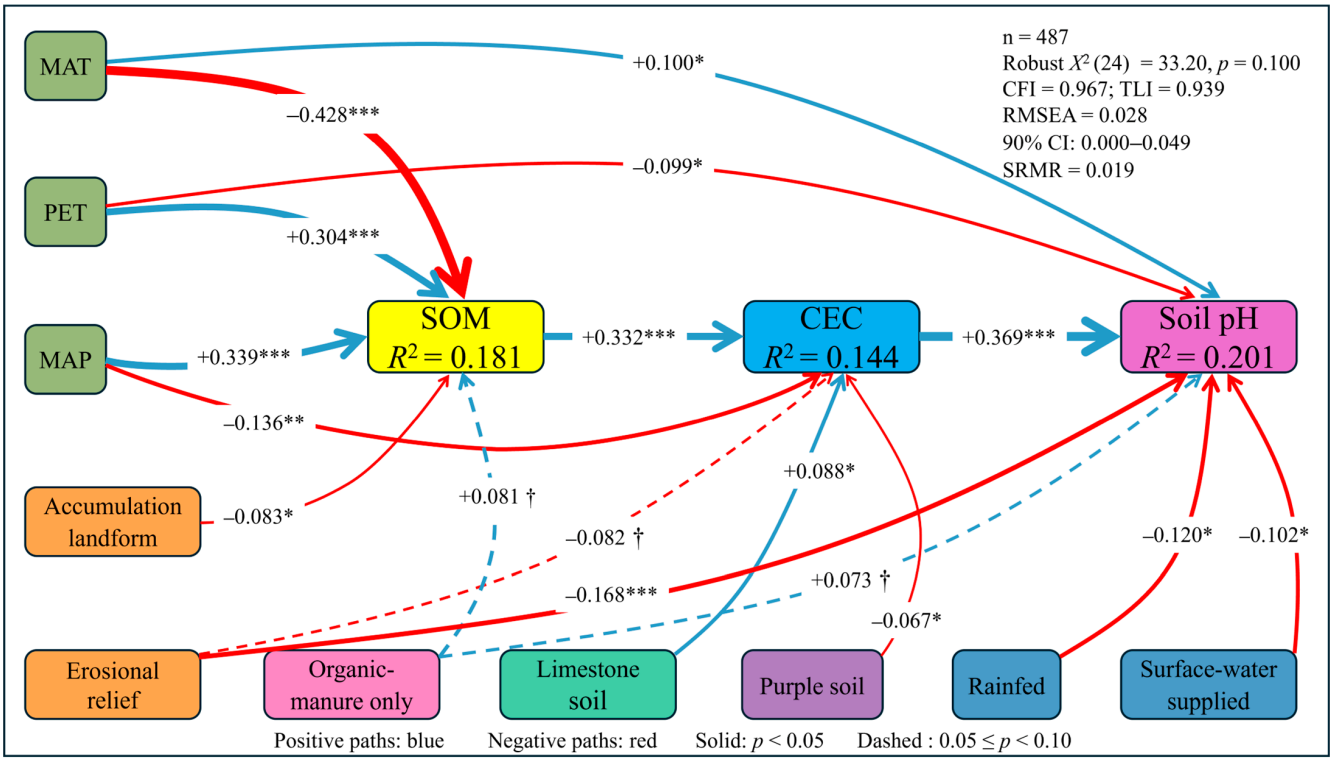

Figure 5. Simplified path diagram of the final observed-variable SEM for farmland soil pH. Arrow labels are fully standardized direct coefficients (β, Std.all), and R^2 values indicate the variance explained in each endogenous variable. Blue and red arrows denote positive and negative coefficients, respectively, with arrow widths proportional to $|\beta|$. Solid arrows indicate nominal robust Wald $p < 0.05$, and dashed arrows indicate $0.05 \leq p < 0.10$. Symbols denote †, $0.05 \leq p < 0.10$; *, $0.01 \leq p < 0.05$; **, $0.001 \leq p < 0.01$; and ***, $p < 0.001$. No multiplicity adjustment was applied. Five retained paths with $p \geq 0.10$ are omitted. All retained direct paths and model-defined indirect and total associations are reported in Tables S23 and S24. Reference categories are pooled yellow–red–yellow-brown soils, tap-water-supplied fields, and combined mineral–organic fertilization. The two water-supply contrasts share one unstandardized coefficient and Wald p-value; their standardized coefficients differ with predictor variances. Exogenous covariances and residual variances are omitted. Paths represent conditional associations without spatial correction.

SOM was positively associated with CEC ($\beta = 0.332$, $p < 0.001$). During model refinement, fixing the additional SOM → pH coefficient to zero did not significantly worsen fit in the corresponding nested-model comparison ($p = 0.662$; Table S21). The final model therefore represented the SOM–pH association through the specified SOM → CEC → pH pathway, yielding a positive standardized indirect association of 0.122 ($p < 0.001$; Table S24). The total association was identical to the indirect association because no direct SOM → pH path was retained. This result describes the retained covariance structure rather than establishing temporal mediation or excluding other unmeasured SOM-related processes.

Oppositely signed components partly offset one another in the model-defined climatic associations. MAT had a positive direct association with pH but a negative indirect association through SOM and CEC ($\beta = -0.052$), producing a small total association that did not meet the nominal 0.05 criterion ($\beta = 0.048$, $p = 0.289$). PET showed the converse pattern, with a negative direct association partly offset by a positive SOM–CEC pathway ($\beta = 0.037$); its total association also did not meet this criterion ($\beta = -0.062$, $p = 0.148$). MAP was positively associated with SOM ($\beta = 0.339$, $p < 0.001$) but negatively associated with CEC after conditioning on SOM and the other CEC predictors ($\beta = -0.136$, $p = 0.003$). The resulting positive MAP → SOM → CEC → pH component ($\beta = 0.042$) and negative MAP

→ CEC → pH component ($\beta = -0.050$) nearly cancelled, yielding a negligible net and total association ($\beta = -0.009$, $p = 0.650$).

Among the climatic and topographic predictors, the erosional-relief score had the largest-magnitude negative standardized total association ($\beta = -0.198$, $B = -0.322$ pH unit per one-level increase, $p < 0.001$). This total was dominated by the direct component; the smaller CEC-linked indirect component was directionally consistent ($\beta = -0.030$) but did not meet the conventional 0.05 criterion ($p = 0.062$). Because erosional relief was a predefined ordered trend score, its coefficient represents an average conditional contrast across successive landform levels rather than the response to an equal physical increment in measured relief or erosion rate. Accumulation landform (basin or plain relative to hill or mountain) had a small nominally negative indirect and total association through SOM and CEC ($\beta = -0.010$, $B = -0.033$, $p = 0.029$).

The field water-supply contrasts were evaluated separately through nested constraint comparisons. Jointly fixing the two freely estimated coefficients to zero gave $p = 0.061$, whereas constraining their unstandardized coefficients to equality did not significantly worsen fit ($p = 0.634$; Table S21). Under the retained equality constraint, rainfed and surface-water-supplied fields shared a negative unstandardized contrast with tap-water-supplied fields ($B = -0.272$, nominal Wald $p = 0.024$). Fixing this common coefficient to zero worsened fit in the corresponding scaled difference test ($p = 0.026$). The fully standardized coefficients were −0.120 and −0.102, respectively, reflecting the different variances of the binary indicators. These values represent two standardized expressions of one shared unstandardized contrast. This result concerns reported field water-supply categories and does not identify effects of irrigation volume or water chemistry.

During model development, jointly fixing the four fertilization coefficients to zero worsened fit ($p = 0.039$, Table S21), supporting retention of the fertilization block within the evaluated structure. Neither the two-path SOM sub-block ($p = 0.092$) nor the two-path direct-pH sub-block ($p = 0.070$) met the nominal 0.05 criterion when tested separately. The joint result does not establish support for each individual path. Mineral-fertilizer-only management had a direct contrast of $\beta = -0.039$ ($p = 0.370$) and a total contrast of $\beta = -0.041$ ($B = -0.104$, $p = 0.348$) with combined fertilization. Organic-manure-only management had a positive direct contrast ($\beta = 0.073$, $p = 0.083$) and a small positive SOM–CEC indirect component ($\beta = 0.010$, $p = 0.099$). Their sum yielded a nominally positive total contrast ($\beta = 0.083$, $B = 0.222$, 95% CI: 0.001–0.443, $p = 0.049$). The confidence interval for this total contrast narrowly excluded zero. This near-threshold result was interpreted as exploratory evidence of a conditional association because parameter uncertainty was not corrected for spatial dependence, no multiple-testing adjustment was applied, and fertilizer quantities and compositions were unavailable.

The five soil-group indicators were retained as a prespecified pedogenic-background block in the CEC equation. Jointly fixing its five coefficients to zero gave $p = 0.058$ and reduced R^2_{CEC} from 0.144 to 0.129. Thus, the joint comparison did not meet the nominal 0.05 criterion, although removing the block reduced the explained variance in CEC. Within this block, the limestone soil group had a small nominally positive CEC-linked total contrast with the pooled yellow–red–yellow-brown reference group ($\beta = 0.032$, $B = 0.094$, 95% CI: 0.001–0.186, $p = 0.047$), whereas the purple soil CEC-linked total contrast was negative and did not meet the nominal 0.05 criterion ($\beta = -0.025$, $p = 0.057$). The remaining soil-group contrasts likewise did not meet this criterion. These estimates were interpreted as secondary conditional associations within the retained pedogenic-background block.

Local fit diagnostics indicated no large discrepancies between the observed and model-implied correlations, and the core coefficients changed little under the tested coding and one-case-deletion perturbations. All absolute correlation residuals were below 0.10, no

process-plausible omitted regression path had a modification index greater than 10, and no residual covariance was added solely to improve fit (Table S25). The directions of the core pathways were retained under the alternative 0–3 erosional-relief coding. The 20 observations with the largest approximate generalized Cook's distances were screened using an ML fit and then deleted individually, with each model refitted using MLR. These individual deletions yielded narrow ranges for the standardized SOM → CEC (0.314–0.339), CEC → pH (0.365–0.377), erosional relief → pH (−0.174 to −0.162), and MAP → CEC (−0.145 to −0.117) coefficients. All 487 observations were retained in the primary model. These checks assessed sensitivity to the tested coding and individual case deletions; residual spatial dependence was not evaluated. The reported standard errors and Wald *p*-values were obtained without spatial correction and may be affected by unmodeled spatial dependence among observations.

4. Discussion

Regional pH organization coexisted with substantial unresolved variation and modest predictive performance, emphasizing the need to distinguish broad spatial patterns from the precision achieved at individual held-out locations. Positive global spatial autocorrelation was detected for pH, SOM, and CEC across the tested neighborhood specifications (Tables S13 and S15). These separate univariate analyses document spatial patterning within each property but do not establish spatial coupling among the three properties. Local pH classifications showed limited consistency after multiplicity adjustment: two analytical units met the BH-adjusted criterion under $k = 4$, whereas none did so under $k = 6$ or $k = 8$ (Table S14). The local classifications were therefore retained as exploratory, neighborhood-dependent spatial diagnostics. Neither the Moran statistics nor the fitted semivariogram identifies the processes generating the observed spatial structure. Covariation with climate, terrain, pedogenic background, management, and unresolved local soil conditions provides plausible explanations whose respective contributions remain to be evaluated.

The most consistent cross-method result concerned CEC. CEC was clearly first-ranked in RF and had the largest absolute fully standardized direct coefficient in the SEM pH equation, highlighting its prominence among the measured soil-property covariates in this dataset (Tables S20 and S23). Because both models used the same observations, their agreement provides complementary evidence within this survey rather than independent mechanistic validation. This agreement supports prioritizing exchange-related soil characterization in subsequent diagnosis, but it does not demonstrate that CEC alone governs soil acid–base status or that increasing CEC would necessarily increase pH. Shared variation in SOM, texture, mineralogy, carbonates, and parent material may contribute to the CEC–pH association. A recent method comparison found higher CEC values with buffered ammonium-acetate extraction than with unbuffered hexamminecobalt extraction in acidic soils with pH-dependent exchange sites [65]. This finding supports considering pH-dependent charge when interpreting the positive CEC–pH association observed here. Measurements of exchangeable cations, acidity, and mineral composition are therefore needed to distinguish the processes underlying this association.

Terrain also provided relatively consistent information: elevation, slope, and relief intensity met the prespecified nonspatial response-permutation criterion after BH adjustment in RF, whereas the SEM erosional-relief ordering showed a clear negative total conditional association with pH (Tables S20 and S24). The latter variable, however, was an ordered landform proxy rather than a measurement of erosion or material flux. These findings highlight terrain-related pH variation while leaving the contributions of drainage, soil redistribution, and parent-material differences unresolved. Climate contributed substantial

RF predictive information, but its SEM interpretation was more complex. Oppositely signed MAT-, PET-, and MAP-related path components yielded small, statistically nonsignificant total associations in the fitted SEM (Table S24). These offsets concern the specified conditional paths and do not demonstrate compensation among physical climatic processes. Accordingly, RF importance and individual SEM paths provide insufficient grounds for assigning a single directional climate response.

Differences between RF and SEM are expected because the two methods estimated different quantities from differently represented predictor sets [22,24]. RF could accommodate nonlinearities and interactions, although permutation importance alone did not identify their form or contribution to the fitted predictions. SEM quantified specified linear direct and indirect conditional associations within a more restricted structural model [24]. For example, SOM did not meet the RF response-permutation criterion after BH adjustment, whereas SEM estimated a positive SOM → CEC → pH association. Shared predictive information with CEC and climate offers a plausible explanation for this difference, but the amount of information shared among these predictors was not quantified [22]. Slope met the RF criterion, but no additional linear slope path to SOM or pH was retained in the final SEM. The crop-management acidification proxy was included only in RF. Fertilization variables also represented different contrasts: RF indicators recorded the presence of mineral or organic fertilizer components, whereas SEM compared mineral-fertilizer-only and organic-manure-only regimes with combined fertilization. Their importance scores and path coefficients therefore require interpretation according to their respective coding and model roles. Conversely, water-supply and soil-group variables had weak RF importance but were retained in constrained or prespecified SEM blocks with generally modest or marginal evidence. The findings were consistent with the hypothesized soil-property and environmental associations, while providing more limited evidence that management contrasts persist within the specified multivariable structure. The added spatial nested validation yielded higher pooled prediction errors than the matched random nested analysis, with RMSE values of 1.0475 and 1.0096 pH units and MAE values of 0.9012 and 0.8500 pH units, respectively (Table S18). The sensitivity of predictive performance to the validation design is consistent with recent soil-mapping evidence that related observations should be kept together during validation to reduce data leakage [66]. The modest RF performance under both validation designs and modest SEM R^2 values indicate that the analyses captured only part of the observed pH variation.

Several design and measurement limitations constrain broader interpretation. The coverage-oriented sampling design encompassed diverse farmland environments but was not an area-weighted probability survey. Accordingly, the reported summary statistics describe the sampled fields, and the proportion of samples with pH < 5.5 (17.9%) cannot be extrapolated to the proportion of farmland area below this threshold across Guizhou Province. The study also combined a one-time soil-state measurement with long-term climatic averages and predominant management categories. A recent national assessment reported contrasting pH trajectories in paddy and dryland soils, including pH recovery in paddy fields [67]. The present survey therefore cannot distinguish temporal trends, legacy effects, and lagged responses from persistent spatial differences. Soil pH was measured in deionized water only; additional measurements in 1 M KCl or 0.01 M $CaCl_2$ could provide complementary information on soil acidity under defined electrolyte conditions. Quantification of exchangeable acidity requires a separate determination, while lime requirement should be assessed using locally calibrated methods that account for soil buffering and crop-specific target pH values. The 0–30 cm field composites integrated soil across sampling positions and depth, so within-field heterogeneity and depth-specific acidity could not be resolved. Quantitative fertilizer rates, nutrient forms, amendment composition, irrigation

amounts, water chemistry, and complete management histories were unavailable. Key soil properties required to interpret the CEC–pH association were also not measured, including exchangeable acidity and Al, individual exchangeable base cations such as Ca and Mg, base saturation, carbonate content, particle-size distribution, clay mineralogy, and parent-material geochemistry. The omission of these variables limits attribution of the associations represented by CEC, mapped soil group, climate, terrain, and management category.

The environmental covariates also differed substantially in native spatial support. Alignment to a nominal 250-m grid facilitated overlay and extraction but did not convert the 1-km soil and climate products or the 4-km sunshine-duration product into field-scale measurements. Adjacent fields could consequently share identical or highly similar raster-derived values while differing in local soil properties and management. Spatial analyses used coordinate-level means of pH, SOM, and CEC, whereas RF and SEM retained all 487 sample-level observations. Repeated coordinates and shared raster-derived covariates may introduce dependence or redundancy that the nominal sample size does not capture; their implications cannot be determined from coordinate counts alone. Ordinary kriging used no environmental drift and achieved only modest precision for individual held-out locations, so the interpolated surface should be regarded as a broad screening layer rather than a field-level prediction or amendment map.

The original RF analysis used the same random folds for tuning and performance summarization, whereas the additional random and spatial nested analyses separated parameter selection from outer-fold evaluation. Identical recorded coordinates were kept together at both levels of the spatial design. Spatial evaluation nevertheless covered only one nominal 20-km hexagonal-block setting without a distance buffer. Median nearest-training distances across the outer spatial folds ranged from 12.15 to 16.55 km, but the minimum distance was approximately 78 m (Text S4). Spatial separation therefore varied among held-out observations. The performance comparison also involved changes in training-set composition and predictor coverage, precluding attribution of the entire difference to spatial dependence. RF importance probabilities remained based on nonspatial response permutations; the additional spatial validation did not change their inferential scope. SEM underwent limited refinement in the same dataset, did not explicitly model residual spatial covariance, and applied no multiplicity adjustment to individual or defined indirect and total associations. These conditions constrain external validity and the interpretation of near-threshold results. Further evaluation should assess RF prediction stability across additional spatial partitions and block scales and geographical transferability using independent data. SEM residual spatial diagnostics would separately help guide appropriate treatment of dependence in parameter uncertainty [63].

The observed pH range highlights the need to assess nutrient availability alongside acidity and exchange status. Since phosphorus availability depends on both pH-sensitive reactions and soil composition, nutrient constraints may warrant investigation in lower- as well as higher-pH fields [3]. Regional sampling priorities should combine measured low-pH observations, interpolation uncertainty, sampling density, and relevant terrain information. The pH < 5.5 criterion used here is an initial screening threshold, not a universal trigger for amendment. Candidate fields should be resampled to establish the persistence and spatial extent of low pH; separate sampling positions and depth intervals should be retained where within-field or vertical heterogeneity is being assessed. Subsequent diagnosis should measure exchangeable acidity and Al, exchangeable Ca and Mg and other base cations, base saturation, CEC, SOM, texture, carbonate status, and plant-available *p* and other nutrients relevant to the crop. Recent research has estimated lime requirement from soil exchangeable Ca, CEC measured at pH 7.0, and limestone composition [68]. This approach reinforces the need for exchangeable-base measurements and local calibration before an amendment rate

is selected for the fields identified here. Fertilizer histories and irrigation-water amounts, alkalinity, and major-ion composition should be recorded where management-related changes in acid–base status are suspected. These measurements would help distinguish low base status, exchangeable-Al accumulation, buffering differences, management inputs, and hydrological redistribution.

Management interventions should be selected through field diagnosis and evaluated against crop-specific requirements. A recent meta-analysis found that liming generally increased soil pH and crop yield, with responses varying among crops and application practices [69]. Where repeated measurements confirm agronomically limiting acidity, locally calibrated trials can assess acidity amendments, nitrogen-source and rate adjustments, and characterized organic inputs [4]. Erosion-control measures can also be evaluated where field observations independently establish erosion or soil redistribution. Trials should include baseline measurements, concurrent controls, and replication under the relevant soil, terrain, and cropping conditions. Multiyear monitoring should assess pH, exchangeable acidity and Al, base status, nutrient availability, crop response, and nutrient losses to evaluate both agronomic outcomes and environmental trade-offs. The practical contribution of this study is a regional evidence base for directing additional sampling, prioritizing exchange-related soil diagnosis, and formulating management interventions for field testing.

5. Conclusions

Across the 487 farmland topsoil samples, pH ranged from 4.37 to 8.45, with a mean of 6.78 $\pm$ 1.12, and 17.9% of the samples were below the pH 5.5 screening threshold adopted in this study. Because the sampling design was not area-weighted, this percentage describes the sampled fields and cannot be extrapolated to the proportion of farmland area below this threshold across Guizhou. Coordinate-level means of pH, SOM, and CEC exhibited modest positive global spatial autocorrelation across the tested neighborhood specifications. Local pH evidence after multiplicity correction was limited and depended on neighborhood specification. The fitted pH semivariogram indicated a short-range spatially structured component and substantial unresolved variation.

The RF model showed modest predictive performance under random and spatially blocked nested cross-validation, with higher pooled prediction errors under the spatial design. The forest fitted to all samples ranked CEC as the most informative measured predictor, followed by climate- and terrain-related variables. In the SEM, CEC had the largest absolute fully standardized direct coefficient in the pH equation. The model represented a positive SOM–pH indirect association through CEC and a negative total conditional association for the ordered erosional-relief score. Climatic path components partially offset one another within the fitted model, yielding small total conditional associations. Management and soil-group contrasts were generally modest and require cautious interpretation.

These findings characterize observational associations and predictive performance under the tested validation designs; they do not establish causal pathways or geographical transferability. The results support a staged approach to soil acidity assessment: regional screening to guide further sampling, followed by field-specific measurement of exchangeable acidity, Al, base status, and buffering capacity. Locally calibrated field trials guided by crop requirements are needed to evaluate amendment and fertilizer-management options in heterogeneous mountain farmland.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/agronomy16212099/s1, Table S1: Number of field-level composite topsoil samples collected in each of the nine prefecture-level divisions of Guizhou Province, Southwest China; Table S2: Field-survey agricultural-management variables recorded at each sampled field and their linkage to the corresponding field-level composite soil sample; Table S3: Field sam-

pling, laboratory measurements, and sampling-location data; Table S4: Sources, temporal coverage, native spatial resolutions, and preprocessing of raster-derived environmental covariates; Table S5: Field-survey agricultural-management variables and data preparation; Table S6: Analytical variables, coding schemes, reference categories, and roles in the RF and SEM; Table S7: Frequency distributions of landform classes and derived relief scores used in the RF and SEM analyses; Table S8: Descriptive statistics of continuous variables used in the RF and SEM analyses; Table S9; Frequency distributions of categorical source variables and final analytical soil groups in the farmland dataset; Table S10: Frequency distributions of derived binary variables used in the RF and SEM analyses; Table S11: Exploratory unadjusted differences in farmland soil pH among final analytical soil groups and agricultural-management categories; Table S12: Coordinate aggregation, nearest-neighbor distances, fixed-distance diagnostics, and spatial-weights characteristics for the Moran analyses; Table S13: Global Moran's I under the primary and sensitivity spatial-neighborhood specifications; Table S14: Local Moran classification counts and neighborhood-sensitivity summaries; Table S15: Global Moran's I for coordinate-level mean SOM and CEC under the primary and sensitivity neighborhood specifications; Table S16: Candidate semivariogram models and leave-one-coordinate-out ordinary-kriging cross-validation performance for farmland soil pH; Table S17: RF mtry tuning results based on fixed 10-fold random cross-validation; Table S18: RF predictive performance under original non-nested random cross-validation, random and spatial nested cross-validation, and out-of-bag evaluation; Table S19: Pearson correlation matrix among the continuous variables used in the random-forest and structural-equation analyses (n = 487); Table S20: Unscaled permutation importance and BH-adjusted response-permutation results for the 22 predictors in the final RF model; Table S21: Satorra–Bentler scaled chi-square difference tests used in SEM refinement and constraint assessment; Table S22: Global fit indices and explained variance of the final SEM; Table S23: Unstandardized and fully standardized direct-path coefficients in the final SEM; Table S24: Standardized decomposition and unstandardized inference for model-defined direct, indirect, and total conditional associations with farmland soil pH in the final SEM; Table S25: Local-fit diagnostics and robustness analyses for the final SEM; Figure S1: Spatial distributions of selected raster-derived environmental covariates and source soil-map classes across the study area: (a) GIS-derived landform categories, (b) source soil-map classes, (c) MAT, (d) MAP, (e) PET, and (f) SSD; Figure S2: Exploratory Local Moran classifications of coordinate-level mean farmland soil pH. (a) High–high, low–low, high–low, and low–high classifications under the primary symmetrized six-nearest-neighbor graph with row-standardized weights. Classifications were based on two-sided rank-based conditional-permutation probabilities from 9999 permutations using the unadjusted $p < 0.05$ threshold. A total of 56 coordinate-level analytical units met this exploratory threshold, whereas none remained significant after Benjamini–Hochberg false-discovery-rate correction at $q < 0.05$. (b) Coordinate-level analytical units assigned to the same HH, LL, HL, or LH category under all three tested neighborhood specifications (k = 4, 6, and 8). Twenty-two units retained the same classification across all three specifications. Point symbols denote recorded coordinate-level observations and do not represent the areal extent of the corresponding spatial associations. The displayed classifications are exploratory spatial diagnostics rather than confirmatory local clusters or spatial outliers; Figure S3: Global spatial autocorrelation of coordinate-level mean SOM and CEC in farmland soils. (a,c) Moran scatterplots for SOM and CEC, respectively, based on 365 unique recorded coordinate pairs under the primary symmetrized six-nearest-neighbor graph with row-standardized weights. The horizontal axes show standardized soil-property values, and the vertical axes show their spatial lags. The slopes of the solid blue regression lines correspond to Global Moran's I. (b,d) Reference distributions of Global Moran's I generated from 9999 random permutations for SOM and CEC, respectively. The vertical red dashed lines indicate the observed values (I = 0.115 for SOM and 0.121 for CEC). Both variables had two-sided permutation $p \leq 0.0002$ and BH-adjusted q = 0.0002, with adjustment applied across the two variables within the primary neighborhood specification. Results for the four- and eight-nearest-neighbor sensitivity specifications are reported in Table S15; Figure S4: Semivariogram fitting, cutoff sensitivity, and ordinary-kriging cross-validation for farmland soil pH. (a) Omnidirectional empirical semivariogram calculated using a 2.5-km lag interval and a 70-km cutoff, together with fitted spherical, exponential, and Gaussian candidate models.

(b) Sensitivity of the fitted spherical semivariogram to cutoff distances of 30, 50, and 70 km; curves are shown over the first 20 km to facilitate comparison of the short-range structure. (c) Observed versus leave-one-coordinate-out ordinary-kriging cross-validation predictions obtained using the selected spherical model and a maximum of 30 neighboring locations. The dashed diagonal line denotes the 1:1 relationship. Soil pH values associated with identical recorded coordinates were averaged before spatial modeling. Numerical model parameters and cross-validation statistics are reported in Table S16; Figure S5: RF tuning and internal predictive-performance diagnostics for farmland soil pH. (a) Mean root mean square error (RMSE) across the fixed 10-fold random cross-validation partition for candidate values of mtry from 1 to 22. Error bars represent the standard deviation across the 10 validation folds, and the dashed vertical line identifies the selected value (mtry = 5). (b) Out-of-bag mean squared error (OOB MSE) as a function of the cumulative number of trees in the final RF model fitted with mtry = 5 and ntree = 1000. (c) Observed soil pH versus pooled out-of-fold predictions from the fixed 10-fold cross-validation analysis. The dashed diagonal line denotes the 1:1 relationship. Performance statistics were calculated after pooling the 487 retained out-of-fold predictions; Figure S6: Frequency distribution of farmland topsoil pH across the study area. The histogram is based on all 487 samples. The vertical lines indicate the sample mean, median, and the pH 5.5 screening threshold; Text S1: Analytical quality control; Text S2: Predictor representation and coding for the RF and SEM analyses; Text S3: Spatial autocorrelation of farmland soil pH, SOM, and CEC and geostatistical analysis of soil pH; Text S4: RF model construction, hyperparameter tuning, random and spatial nested validation, and permutation-based importance assessment; Text S5: SEM specification, constrained refinement, association decomposition, diagnostics, and robustness assessment. The Supplementary contains following references [70–97].

Author Contributions: Conceptualization, Q.P., C.X. and R.S.; Methodology, Q.P. and C.X.; software, Q.P. and C.X.; writing—original draft preparation, Q.P., C.X. and R.S.; writing—review and editing, C.X., R.S. and Y.L.; project administration, R.S. and Y.L.; funding acquisition, R.S. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was financially supported by the Science and Technology Department of Guizhou Province (No. Qian Ke He Platform ZSYS [2025]035).

Data Availability Statement: The original contributions presented in this study are included in the article/Supplementary Materials. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gondal, A.H.; Hussain, I.; Ijaz, A.B.; Zafar, A.; Ch, B.I.; Zafar, H.; Sohail, M.D.; Khan, A.A.; Niazi, H.; Yousaf, H.; et al. Influence of Soil pH and Microbes on Mineral Solubility and Plant Nutrition: A Review. *Int. J. Agric. Biol. Sci.* **2021**, *5*, 71–81. [CrossRef]
2. Malik, A.A.; Puissant, J.; Buckeridge, K.M.; Goodall, T.; Jehmlich, N.; Chowdhury, S.; Gweon, H.S.; Peyton, J.M.; Mason, K.E.; van Agtmaal, M.; et al. Land use driven change in soil pH affects microbial carbon cycling processes. *Nat. Commun.* **2018**, *9*, 3591. [CrossRef] [PubMed]
3. Penn, C.J.; Camberato, J.J. A Critical Review on Soil Chemical Processes that Control How Soil pH Affects Phosphorus Availability to Plants. *Agriculture* **2019**, *9*, 120. [CrossRef]
4. Cai, Z.; Wang, B.; Xu, M.; Zhang, H.; He, X.; Zhang, L.; Gao, S. Intensified soil acidification from chemical N fertilization and prevention by manure in an 18-year field experiment in the red soil of southern China. *J. Soils Sediments* **2015**, *15*, 260–270. [CrossRef]
5. Wu, Z.; Sun, X.; Sun, Y.; Yan, J.; Zhao, Y.; Chen, J. Soil acidification and factors controlling topsoil pH shift of cropland in central China from 2008 to 2018. *Geoderma* **2022**, *408*, 115586. [CrossRef]
6. Ou, Y.; Rousseau, A.N.; Wang, L.; Yan, B. Spatio-temporal patterns of soil organic carbon and pH in relation to environmental factors—A case study of the black soil region of northeastern China. *Agric. Ecosyst. Environ.* **2017**, *245*, 22–31. [CrossRef]
7. Han, Y.; Yi, D.; Ye, Y.; Guo, X.; Liu, S. Response of spatiotemporal variability in soil pH and associated influencing factors to land use change in a red soil hilly region in southern China. *Catena* **2022**, *212*, 106074. [CrossRef]
8. Xie, X.; Qiu, J.; Feng, X.; Hou, Y.; Wang, S.; Jia, S.; Liu, S.; Hou, X.; Dou, S. Spatial distribution and estimation model of soil pH in coastal eastern China. *Int. J. Environ. Res. Public Health* **2022**, *19*, 16855. [CrossRef] [PubMed]

9. Zhang, Y.; Hartemink, A.E. Quantifying short-range variation of soil texture and total carbon of a 330-ha farm. *Catena* **2021**, *199*, 105200. [CrossRef]
10. Hu, B.; Xie, M.; Shi, Z.; Li, H.; Chen, S.; Wang, Z.; Zhou, Y.; Ni, H.; Geng, Y.; Zhu, Q.; et al. Fine-resolution mapping of cropland topsoil pH of Southern China and its environmental application. *Geoderma* **2024**, *442*, 116798. [CrossRef]
11. Slessarev, E.W.; Lin, Y.; Bingham, N.L.; Johnson, J.E.; Dai, Y.; Schimel, J.P.; Chadwick, O.A. Water balance creates a threshold in soil pH at the global scale. *Nature* **2016**, *540*, 567–569. [CrossRef] [PubMed]
12. Zhang, X.; Qin, H.; Zhang, Y.; Niu, J.; Wang, Y.; Shi, L. Driving factors of community-level leaf stoichiometry patterns in a typical temperate mountain meadow ecosystem of northern China. *Front. Plant Sci.* **2023**, *14*, 1141765. [CrossRef] [PubMed]
13. Häring, V.; Manka'abusi, D.; Kyei Akoto-Danso, E.; Werner, S.; Atiah, K.; Steiner, C.; Jean-Pascal Lompo, D.; Adiku, S.G.K.; Buerkert, A.; Marschner, B. Effects of biochar, waste water irrigation and fertilization on soil properties in West African urban agriculture. *Sci. Rep.* **2017**, *7*, 10738. [CrossRef] [PubMed]
14. Saifuzzaman, M.; Adamchuk, V.; Biswas, A.; Rabe, N. High-density proximal soil sensing data and topographic derivatives to characterise field variability. *Biosyst. Eng.* **2021**, *210*, 108018. [CrossRef]
15. Paramisparam, P.; Ahmed, O.H.; Omar, L.; Ch'ng, H.Y.; Maru, A.; Johan, P.D. Amending Potassic Fertilizer with Charcoal and Sago (Metroxylon sagu) Bark Ash to Improve Potassium Availability in a Tropical Acid Soil. *Agronomy* **2021**, *11*, 2222. [CrossRef]
16. Guo, Z.; Li, X.; Yang, R.; Qu, F.; Zhang, W.; Bi, X.; Zhou, S. Divergent Proton-Buffering Processes and Acidification Risks in Permanent and Variable-Charge Soils. *Agronomy* **2026**, *16*, 171638. [CrossRef]
17. Lishan, T.; Alemu, F. Elucidating sole application of farmyard manure and blended NPSB fertilizer effects on soil properties at Bench Shako and West Omo zone, South West Ethiopia. *Heliyon* **2024**, *10*, e22908. [CrossRef] [PubMed]
18. Lv, H.; Zhao, Y.; Wang, Y.; Wan, L.; Wang, J.; Butterbach-Bahl, K.; Lin, S. Conventional Flooding Irrigation and Over Fertilization Drives Soil pH Decrease Not Only in the Top- but Also in Subsoil Layers in Solar Greenhouse Vegetable Production Systems. *Geoderma* **2020**, *363*, 114156. [CrossRef]
19. Chu, L.L.; Kang, Y.H.; Wan, S.Q. Effect of different water application intensity and irrigation amount treatments of microirrigation on soil-leaching coastal saline soils of North China. *J. Integr. Agric.* **2016**, *15*, 2022–2031. [CrossRef]
20. Mcengwa, S.; Nyambo, S.; Mbanjwa, V.E.; Elephant, D.E. Cropping History and Cropland Abandonment Period Influenced Levels of Soil Organic Carbon, Distribution of Aggregates and Aggregate Associated Carbon. *Land Degrad. Dev.* **2025**, *36*, 2148–2161. [CrossRef]
21. Sinclair, W.A. Development of Ectomycorrhizae in a Douglas-fir Nursery: II. Influence of Soil Fumigation, Fertilization, and Cropping History. *For. Sci.* **2025**, *20*, 57–63. [CrossRef]
22. Wadoux, A.M.J.-C.; Molnar, C. Beyond prediction: Methods for interpreting complex models of soil variation. *Geoderma* **2022**, *422*, 115953. [CrossRef]
23. Zolfaghari, Z.; Mosaddeghi, M.R.; Ayoubi, S. Relationships of soil shrinkage parameters and indices with intrinsic soil properties and environmental variables in calcareous soils. *Geoderma* **2016**, *277*, 23–34. [CrossRef]
24. Wade, J.; Potter, T.S.; Margenot, A.J. Commentary: Structural equation models and causal claims in soil science and biogeochemistry—An equation-free "how to". *Geoderma* **2026**, *466*, 117702. [CrossRef]
25. Guizhou Provincial Bureau of Statistics; Guizhou Survey Team of the National Bureau of Statistics. *Guizhou Statistical Yearbook 2021*, Electronic ed.; China Statistics Press: Beijing, China, 2021.
26. Yu, L.; Li, Y.; Luo, G.; Ge, G.; Zhang, H.; Tang, F.; Yu, M. Spatiotemporal evolution and driving mechanism of slope cultivated land in karst mountainous areas of Southwest China—A case study of Puding County, Guizhou Province. *Land Degrad. Dev.* **2024**, *35*, 568–585. [CrossRef]
27. Mao, Q.J.; Xie, L.; Fan, L. Spatial characteristics and determinants of traditional village distribution in Guizhou Province. *PLoS ONE* **2025**, *20*, e0324275. [CrossRef] [PubMed]
28. Cao, Y.; Xiong, K. Soil quality assessment under different vegetation restoration strategies in the karst rocky deserted area of southwestern China. *J. Soil Sci. Plant Nutr.* **2024**, *24*, 3249–3264. [CrossRef]
29. Tan, H.; He, Z.; Yu, H.; Yang, S.; Wang, M.; Gu, X.; Xu, M. Characterization of extreme rainfall changes and response to temperature changes in Guizhou Province, China. *Sci. Rep.* **2024**, *14*, 20495. [CrossRef] [PubMed]
30. Xia, X.; Ji, J.; Yang, Z.; Lu, X.; Huang, C.; Wei, Y.; Xu, C.; Liang, Z. Parent rock type control on cadmium background in soil and sediment: An example from Guizhou Province. *Earth Sci. Front.* **2022**, *29*, 438–447. [CrossRef]
31. Huang, Y.; Liu, J.; Zhang, C.; Jie, W.; Kang, Y. Geochemical process and ecogeological effects of bedrock weathering in the rocky desertification area of Guizhou Province. *Geol. Surv. China* **2024**, *11*, 62–71. [CrossRef]
32. Gao, W.; Cai, K.; Li, D.; Lin, Y.; Chen, Y.; Lin, Y.; Li, J.; Zhao, Y.; Ju, B.; Pan, W. Soil taxonomy and suitability assessment on typical tobacco-planting farmlands in Guizhou, Southwest China. *SN Appl. Sci.* **2019**, *1*, 877. [CrossRef]
33. Li, K.; Li, M.; Wang, W.; Su, M.; Liang, P.; Wang, Y.; He, T.; Yuan, A. Altitude, agricultural disturbance intensity, soil biotic and abiotic factors jointly affect soil multifunctionality in mountain farmlands. *Agric. Ecosyst. Environ.* **2026**, *401*, 110280. [CrossRef]

34. Tarrillo, E.; Arce-Inga, M.; Torres-Herrera, P.A.; Tineo, D.; Taboada-Mitma, V.H.; Cruz-Luis, J.; Rojas-Briceño, N.B.; Atalaya-Marin, N.; Gómez-Fernández, D.; Goñas, M. Geospatial distribution of heavy metals in rice soils of northwestern Peru. *Sci. Rep.* **2025**, *15*, 16638. [CrossRef] [PubMed]
35. Jensen, K.H.; Faehndrich, C.S.; Colzani, E.; McClure, M.L.; Covey, K. Rapid soil harvesting using a novel soil auger system for farm-scale soil carbon estimates. *Soil Sci. Soc. Am. J.* **2023**, *88*, 192–202. [CrossRef]
36. *NY/T 1377-2007*; Determination of pH in Soil. Ministry of Agriculture; Chinese Academy of Agricultural Quality Standards and Testing Technology Institute: Beijing, China, 2007. (In Chinese)
37. *NY/T 1121.6-2006*; Method for Determination of Soil Organic Matter. Ministry of Agriculture; National Agricultural Technology Extension and Service Center: Beijing, China, 2006. (In Chinese)
38. *HJ 889-2017*; Soil—Determination of Cation Exchange Capacity-Hexamminecobalt Trichloride Solution-Spectrophotometric Method. Ministry of Environmental Protection of the People's Republic of China: Beijing, China, 2017. (In Chinese)
39. Weiss, A. Topographic position and landforms analysis. In Proceedings of the ESRI User Conference, San Diego, CA, USA, 9–13 July 2001.
40. Jenness, J. *Topographic Position Index (tpi_jen.avx) Extension for ArcView 3.x, v.1.3a*; Jenness Enterprises: Flagstaff, AZ, USA, 2006.
41. Zhou, C.; Cheng, W.M.; Qian, J.K.; Li, B.; Zhang, B. Research on the Classification System of Digital Land Geomorphology of 1:1000000 in China. *J. Geo-Inf. Sci.* **2009**, *11*, 707–724. [CrossRef]
42. Zhang, J.; Liu, B.; Ren, S.; Han, W.; Ding, Y.; Peng, S. A 4 km daily gridded meteorological dataset for China from 2000 to 2020. *Sci. Data* **2024**, *11*, 1230. [CrossRef] [PubMed]
43. Celik, N. Welch's ANOVA: Heteroskedastic skew-t error terms. *Commun. Stat. Theory Methods* **2020**, *51*, 3065–3076. [CrossRef]
44. Kassu, A. An in situ real-time monitoring of the hydration and dehydration of building gypsum: Conventional and standoff Raman spectroscopy coupled with chemometrics analysis. *Constr. Build. Mater.* **2023**, *400*, 132682. [CrossRef]
45. Duan, Y.; Xie, S.; Liu, F.; Tang, J.; Gong, L.; Liu, X.; Wen, T.; Wang, H. GIS-based assessment of spatial and temporal disparities of urban health index in Shenzhen, China. *Front. Public Health* **2024**, *12*, 1429143. [CrossRef] [PubMed]
46. Thompson, E.S.; Saveyn, P.; Declercq, M.; Meert, J.; Guida, V.; Eads, C.D.; Robles, E.S.J.; Britton, M.M. Characterisation of heterogeneity and spatial autocorrelation in phase separating mixtures using Moran's I. *J. Colloid Interface Sci.* **2018**, *513*, 180–187. [CrossRef] [PubMed]
47. Dai, W.; Li, Y.; Fu, W.; Jiang, P.; Zhao, K.; Li, Y.; Penttinen, P. Spatial variability of soil nutrients in forest areas: A case study from subtropical China. *J. Plant Nutr. Soil Sci.* **2018**, *181*, 728–737. [CrossRef]
48. Palejev, D.; Savov, M. On the Convergence of the Benjamini–Hochberg Procedure. *Mathematics* **2021**, *9*, 2154. [CrossRef]
49. Grace, J.B. *Structural Equation Modeling and Natural Systems*; Cambridge University Press: Cambridge, UK, 2006.
50. Rosseel, Y. lavaan: An R Package for Structural Equation Modeling. *J. Stat. Softw.* **2012**, *48*, 1–36. [CrossRef]
51. Garnier Villarreal, M.; Jorgensen, T.D. Evaluating local model misspecification with modification indices in Bayesian structural equation modeling. *Struct. Equ. Model.* **2024**, *32*, 304–318. [CrossRef]
52. Satorra, A.; Bentler, P.M. A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika* **2001**, *66*, 507–514. [CrossRef]
53. Cheng, C.; Spiegelman, D.; Li, F. Estimating the natural indirect effect and the mediation proportion via the product method. *BMC Med. Res. Methodol.* **2021**, *21*, 191. [CrossRef] [PubMed]
54. Sobel, M.E. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociol. Methodol.* **1982**, *13*, 290–312. [CrossRef]
55. Gu, S.; Wu, S.; Zeng, W.; Deng, Y.; Luo, G.; Li, P.; Yang, Y.; Wang, Z.; Hu, Q.; Tan, L. High-elevation-induced decrease in soil pH weakens ecosystem multifunctionality by influencing soil microbiomes. *Environ. Res.* **2024**, *257*, 119330. [CrossRef] [PubMed]
56. Ortner, M.; Seidel, M.; Semella, S.; Udelhoven, T.; Vohland, M.; Thiele-Bruhn, S. Content of soil organic carbon and labile fractions depend on local combinations of mineral-phase characteristics. *Soil* **2022**, *8*, 113–129. [CrossRef]
57. Opala, P.A. Influence of Lime and Phosphorus Application Rates on Growth of Maize in an Acid Soil. *Adv. Agric.* **2017**, *2017*, 7083206. [CrossRef]
58. Markhvida, M.; Ceferino, L.; Baker, J.W. Modeling spatially correlated spectral accelerations at multiple periods using principal component analysis and geostatistics. *Earthq. Eng. Struct. Dyn.* **2018**, *47*, 1173–1191. [CrossRef]
59. Wang, X.S.; Zheng, F.L.; Liang, R.; Yang, X.Y.; Mo, S.H.; An, X.B. Effects of spatial distribution of slope erosion-deposition on soil quality degradation in the Mollisol region of Northeast China. *Soil Tillage Res.* **2025**, *255*, 106966. [CrossRef]
60. Kong, D.; Luo, C.; Liu, H. Integrative remote sensing and machine learning approaches for SOC and TN spatial distribution: Unveiling C:N ratio in Black Soil region. *Soil Tillage Res.* **2025**, *254*, 106809. [CrossRef]
61. van Leeuwen, C.C.E.; Mulder, V.L.; Batjes, N.H.; Heuvelink, G.B.M. Effect of measurement error in wet chemistry soil data on the calibration and model performance of pedotransfer functions. *Geoderma* **2024**, *442*, 116762. [CrossRef]
62. Lei, S.; Shao, M.; Jia, X.; Zhu, Z.; Zhao, C. The stacking method enhances machine learning models for monitoring and understanding regional soil available nitrogen variations in croplands. *Soil Tillage Res.* **2026**, *256*, 106880. [CrossRef]

63. Wei, L.; Wang, J.X.; Wei, Y.J.; Sun, Z.; Ma, L.X.; Hong, Y.S.; Chen, Y.Y. An innovative mapping framework for soil erodibility integrating spatial association dimensions and machine learning. *Soil Tillage Res.* **2026**, *257*, 107135. [CrossRef]
64. Yuan, J.; E, S.; Zhao, X. Improved determination efficiency of the cation exchange capacity and exchangeable base cations in soil using stirring and extraction modifications. *Soil Tillage Res.* **2025**, *251*, 106429. [CrossRef]
65. Nel, T.; Bruneel, Y.; Smolders, E. Comparison of five methods to determine the cation exchange capacity of soil. *J. Plant Nutr. Soil Sci.* **2023**, *186*, 311–320. [CrossRef]
66. John, K.; Saurette, D.D.; Heung, B. The problematic case of data leakage: A case for leave-profile-out cross-validation in 3-dimensional digital soil mapping. *Geoderma* **2025**, *455*, 117223. [CrossRef]
67. Zhang, W.; Wei, C.; Li, J.; Zhou, Q.; Ma, J.; Li, Y.; He, Y.; Zou, Y.; Luo, Y.; Yao, Y. Stabilization of acidification in China's cropland soils. *Nat. Geosci.* **2025**, *18*, 1125–1132. [CrossRef]
68. Moreira, S.G.; Gaudencio, J.R.F.; de Moraes, F.A.; de Morais, E.G.; Peixoto, D.S.; de Resende, H.C.; Silva, J.C.R.; Campos, O.L.V. A practical method for estimating liming requirements based on soil chemical attributes and limestone composition. *Soil Tillage Res.* **2026**, *255*, 106816. [CrossRef]
69. Enesi, R.O.; Dyck, M.; Chang, S.; Thilakarathna, M.S.; Fan, X.; Strelkov, S.; Gorim, L.Y. Liming remediates soil acidity and improves crop yield and profitability—A meta-analysis. *Front. Agron.* **2023**, *5*, 1194896. [CrossRef]
70. Tan, Q.; Geng, J.; Yu, Y.; Pei, J.; Fang, H. Improving soil organic carbon mapping in intensively cultivated Mollisols: A geographically adjusted Gaussian mixture model incorporating tillage practices and spectral indicators. *Soil Tillage Res.* **2025**, *255*, 106911. [CrossRef]
71. Luo, C.; Zhang, X.; Wang, Y.; Men, Z.; Liu, H. Regional soil organic matter mapping models based on the optimal time window, feature selection algorithm and Google Earth Engine. *Soil Tillage Res.* **2022**, *221*, 105325. [CrossRef]
72. Li, R.; Xu, M.; Chen, Z.; Gao, B.; Cai, J.; Shen, F.; He, X.; Zhuang, Y.; Chen, D. Phenology-based classification of crop species and rotation types using fused MODIS and Landsat data: The comparison of a random-forest-based model and a decision-rule-based model. *Soil Tillage Res.* **2020**, *204*, 104838. [CrossRef]
73. Ruan, W.; Liu, H.; Sui, Y.; Wang, C.; Luo, C.; Meng, X.; Dong, C. Quantitative prediction of black soil horizon thickness at the watershed scale in Northeast China's black soil region. *Soil Tillage Res.* **2025**, *254*, 106886. [CrossRef]
74. Zhang, P.; Wang, Y.; Xu, L.; Li, R.; Sun, H.; Zhou, J. Factors controlling spatial variation in soil aggregate stability in a semi-humid watershed. *Soil Tillage Res.* **2021**, *213*, 105187. [CrossRef]
75. Shao, T.Y.; Qian, F.K.; Wang, S.; Jiang, Z.D.; Liu, H.B.; Lal, R.; Han, W. Spatial prediction and dynamic change of soil organic carbon using remote sensing variables as auxiliary information in wavy plain, Northeast China. *Soil Tillage Res.* **2025**, *253*, 106759. [CrossRef]
76. Yu, Y.; Geng, J.; Li, G.; Fang, H.; Cheng, S. Multi-scale analysis of soil property variability in Northeast China's black soils using advanced geospatial models. *Soil Tillage Res.* **2025**, *254*, 106762. [CrossRef]
77. Carrera-Hernández, J.J. Not all DEMs are equal: An evaluation of six globally available 30 m resolution DEMs with geodetic benchmarks and LiDAR in Mexico. *Remote Sens. Environ.* **2021**, *260*, 112474. [CrossRef]
78. Jiru, T.L.; Moomen, A.-W.; Domfeh, M.K.; Appiah, O. Spatial prediction of soil organic carbon stocks using regression kriging to support land degradation neutrality (SDG 15.3.1) in Negele Arsi District, Ethiopia. *Sustain. Futures* **2026**, *11*, 101796. [CrossRef]
79. Mushagalusa, C.A.; Fandohan, A.B.; Kakaï, R.G. Random forest and spatial cross-validation performance in predicting species abundance distributions. *Environ. Syst. Res.* **2024**, *13*, 52. [CrossRef]
80. Zhang, J.; Wang, T. On cross-validated estimation of skew normal model. *J. Stat. Plan. Inference* **2025**, *227*, 106271. [CrossRef]
81. Yu, X.; Liu, Y.; Zhang, Z.; Xiao, R. Influences of buildings on urban heat island based on 3D landscape metrics: An investigation of China's 30 megacities at micro grid-cell scale and macro city scale. *Landsc. Ecol.* **2021**, *36*, 2047–2064. [CrossRef]
82. Liu, R.; Xu, F.; Yu, W.; Shi, J.; Zhang, P.; Shen, Z. Analysis of field-scale spatial correlations and variations of soil nutrients using geostatistics. *Environ. Monit. Assess.* **2016**, *188*, 9. [CrossRef] [PubMed]
83. Angelini, M.E.; Heuvelink, G.B.M.; Kempen, B.; Morrás, H.J.M. Mapping the soils of an Argentine Pampas region using structural equation modelling. *Geoderma* **2016**, *281*, 102–115. [CrossRef]
84. Nembrini, S. On the behaviour of permutation-based variable importance measures in random forest clustering. *J. Chemom.* **2019**, *33*, e3135. [CrossRef]
85. Probst, P.; Wright, M.N.; Boulesteix, A.-L. Hyperparameters and tuning strategies for random forest. *WIREs Data Min. Knowl. Discov.* **2019**, *9*, e1301. [CrossRef]
86. Savalei, V.; Rosseel, Y. Computational options for standard errors and test statistics with incomplete normal and nonnormal data in SEM. *Struct. Equ. Model.* **2021**, *28*, 887–905. [CrossRef]
87. Yang, W.; Cheung, S.F. Forming bootstrap confidence intervals and examining bootstrap distributions of standardized coefficients in structural equation modelling: A simplified workflow using the R package semboottools. *Behav. Res. Methods* **2026**, *58*, 38. [CrossRef] [PubMed]

88. Hair, J.F.; Risher, J.J.; Sarstedt, M.; Ringle, C.M. When to use and how to report the results of PLS-SEM. *Eur. Bus. Rev.* **2019**, *31*, 2–24. [CrossRef]
89. Zur, R.M.; Aballea, S.; Sherman, S. A validation of piecewise structural equation modelling (SEM) compared to traditional structural equation modelling: A simulation study. *Value Health* **2018**, *21*, 967–975. [CrossRef]
90. Han, T.; Huang, J.; Liu, K.; Fan, H.; Shi, X.; Chen, J.; Jiang, X.; Liu, G.; Liu, S.; Zhang, L.; et al. Soil potassium regulation by changes in potassium balance and iron and aluminum oxides in paddy soils subjected to long-term fertilization regimes. *Soil Tillage Res.* **2021**, *214*, 105168. [CrossRef]
91. Cheung, S.F.; Cheung, S.-H. manymome: An R package for computing the indirect effects, conditional effects, and conditional indirect effects, standardized or unstandardized, and their bootstrap confidence intervals, in many (though not all) models. *Behav. Res. Methods* **2023**, *55*, 3837–3862. [CrossRef] [PubMed]
92. Imbens, G.W.; Kolesár, M. Robust standard errors in small samples: Some practical advice. *Rev. Econ. Stat.* **2016**, *98*, 701–712. [CrossRef]
93. Xia, F.; Chan, K.C.G. Decomposition, identification and multiply robust estimation of natural mediation effects with multiple mediators. *Biometrika* **2022**, *109*, 929–946. [CrossRef]
94. Jia, F.; Jorgensen, T.D.; Wu, W. Obtaining robust practical fit indices with multiply imputed nonnormal data in structural equation modeling. *Behav. Res. Methods* **2026**, *58*, 113. [CrossRef] [PubMed]
95. Ximénez, C.; Maydeu-Olivares, A.; Shi, D.; Revuelta, J. Assessing cutoff values of SEM fit indices: Advantages of the unbiased SRMR index and its cutoff criterion based on communality. *Struct. Equ. Model.* **2022**, *29*, 541–556. [CrossRef]
96. Huang, P.-H. Asymptotics of AIC, BIC, and RMSEA for model selection in structural equation modeling. *Psychometrika* **2017**, *82*, 406–429. [CrossRef] [PubMed]
97. Martynova, E.; West, S.G.; Liu, Y. Review of principles and practice of structural equation modeling. *Struct. Equ. Model.* **2017**, *25*, 480–489. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.