FASE FRONTIERS OF AGRICULTURAL SCIENCE AND ENGINEERING
Engineering Transactions of CAE
RESEARCH ARTICLE
2026 / Volume 13 / Issue 3 / 25658

Modeling soil pH dynamics with distinct buffering mechanisms: insights from two purple soils

Haiyang HUANG[1,2], Xuanjing CHEN (✉)[1,2], Yuting ZHANG[1,2], Tao GUO[1], Shuai WANG[3], Jia ZHOU[3], Zhiqi LI[4], Yang WANG[4], Yueqiang ZHANG[1,2], Xiaojun SHI (✉)[1,2]

1 College of Resources and Environment. Southwest University, Chongqing 400715, China.
2 Interdisciplinary Research Center for Agriculture Green Development in Yangtze River Basin, Southwest University, Chongqing 400715, China.
3 Agricultural Technology Extension Station, Chongqing 401122, China.
4 District Agricultural Technology Extension Center, Jiangjin, Chongqing 402260, China.

KEYWORDS

Long-term experiments, machine learning, purple soil, soil acidification, VSD+ model

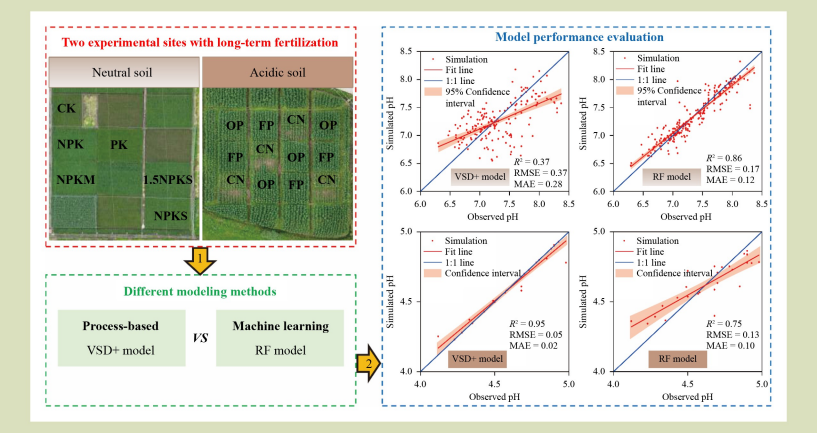
HIGHLIGHTS

- Soil pH dynamics were modeled in two purple soils with distinct buffering mechanisms.
- Soil background pH was the primary factor affecting soil pH changes.
- Machine learning models were superior for neutral soil.
- The Very Simple Dynamic Model Plus (VSD+) performed excellently for acidic soil.
- Random forest modeling gave the best accuracy of four machine learning models tested.

Received May 17, 2025;
Accepted August 22, 2025.

Correspondences: chenxj0517@swu.edu.cn, shixj@swu.edu.cn

GRAPHICAL ABSTRACT



ABSTRACT

Soil acidification models are useful for evaluating measures to mitigate soil acidification under various agronomic practices. However, the appropriate modeling approaches for simulating the soil acidification process have not been adequately studied across soils with distinct buffering mechanisms. This study evaluated the performance differences between a process-based soil acidification model (VSD+) and four machine learning models, including random forest (RF), support vector machine, extreme gradient boosting and decision tree, in simulating pH dynamics of neutral and acidic soils. Two long-term experimental sites were selected with distinct buffering mechanisms on purple soil as an example for the development, calibration and validation of soil acidification models. Results from the RF importance factor analysis indicated that soil background pH was the primary factor influencing the dynamic changes in purple soil pH, followed by meteorological conditions and agronomic practices. pH was then chosen as an essential input variable to developing machine learning models for simulating soil acidification patterns.

Machine learning models achieved higher accuracy in neutral soil than the VSD+ model. The RF model gave the best simulation performance, outperforming other machine learning models at both sites, with the highest R^2 of 0.70 and 0.47 and the lowest MAE of 0.19 and 0.17 for neutral and acidic soils, respectively. In contrast, the VSD+ model exhibited excellent accuracy with acidic soil (R^2 = 0.95, RMSE = 0.05 and MAE = 0.02) compared to the other machine learning models (R^2 = 0.20–0.47, RMSE = 0.15–0.23 and MAE = 0.14–0.20). These findings provide information for selecting the most suitable modeling approach to simulate soil acidification process with distinct buffering mechanisms, supporting informed decision-making for restoring soil health and quality.

1 Introduction

Soil acidification is a widespread problem becoming increasingly severe due to excessive chemical fertilizer application and atmospheric acid deposition[1]. From 1980 to 2000, the topsoil pH of major Chinese croplands has decreased by an average of 0.5 units due to excessive nitrogen fertilizer application and acid rain[2]. Now, about 40% of farmland soils worldwide are exposed to risk acidification[3]. Soil acidification has been a significant cause of soil quality degradation in farmland. Also, soil acidification restricts crop growth by decreasing the availability of soil nutrients and increasing the release of toxic aluminum and manganese[4,5]. A recent study revealed that soil acidification leads to a decrease of 25.4% in root length and 33% in crop yield[6]. Soil acidification is projected to induce a 20% loss increase in crop yield loss in China from 2010 to 2050[7]. Therefore, a better understanding of the temporal variation of soil pH has important implications for safeguarding food security. Simulating the dynamic changes in soil pH and projecting its future trends using acidification models is more convenient and efficient than continuously monitoring soil pH over long periods in the field. Thus, soil acidification models are crucial for managing and restoring soil quality in farmland.

The development of soil acidification models began in the 1980s. Most of soil acidification models were developed to assess long-term acidification in non-agricultural soils (primarily forests), such as SMART[8], MAGIC[9], SAFE[10] and VSD+[11]. Building on equations by the charge balance principle, these models aim to simulate the chemical properties of non-agricultural soils over time and space, and predict trends under various scenarios. The VSD+ model was specifically developed to simulate soil acidification process with relatively few input parameters, making it easier to apply than other acidification models[12]. Zeng et al.[13] first showed that the VSD+ model gave optimal performance in acidic red soils but limited accuracy in neutral purple and calcareous black soils. This indicates that for alkaline or neutral soils containing calcium carbonate, that is, when the soil acid buffering substance is calcium carbonate, the VSD+ model has poor modeling accuracy for such soils due to design limitations. However, in high-pH soils (> 7.5) with significant acid buffering capacity and abundant $CaCO_3$, exogenous H^+ inputs were primarily neutralized through carbonate dissolution rather than base cations exchange[14]. To date, few process-based models can effectively simulate pH or base saturation dynamics in agricultural soils. Therefore, accurately modeling pH dynamics of calcareous or neutral soils containing calcium carbonate is still a significant scientific challenge.

Machine learning, which involves computer science and artificial intelligence, can learn from empirical data to capture intricate nonlinear relationships and build highly accurate regression predictive models[15,16]. Machine learning methods have been successfully applied in the pollutants loading projection and crop yield mapping with excellent performance[17,18]. Compared to most traditional process-based models, machine learning approaches exhibit significant advantages in terms of high prediction accuracy and robust generalization ability, while relying on a reduced set of process parameters[19]. Also, recent studies have successfully used machine learning methods to draw the spatial pattern of soil pH by examining the partial direct relationships between soil

pH and multiple environmental variables and soil management, such as climate, soil properties, terrain characteristics and management practices[3,20]. However, applications of these methods for simulating temporal pH dynamics under long-term field management regimes remain scarce. As a result, it remains largely unclear whether machine learning can compensate for the poor performance of process-based models in alkaline or neutral soils. For this study, we hypothesized that the long-term dynamics of pH in soils where the acid buffering substance is calcium carbonate can be simulated by machine learning models.

Purple soil is one of the widely cultivated soils in south-western China, known for its rich mineral elements and high fertility[21]. In addition, purple soil is considered a mineral soil and that is less prone to soil acidification[22]. However, the pH of purple soil in south-western China decreased by 0.3 units from 1981 to 2012[23]. To address the above knowledge gaps, we take neutral and acidic purple soils for this study. The main objectives of this study were to explore an appropriate modeling approach for two purple soils with distinct buffering mechanisms (calcium carbonate and exchangeable base cations). In this study, we evaluated the performance of VSD+ model alongside four machine learning models in simulating pH dynamics for two types of purple soil. Two long-term experimental sites on purple soil with distinct buffering mechanisms and durations (1991–2023 and 2013–2023) were selected. We also integrated the random forest (RF) importance factor ranking and correlation analyses to identify input variables for machine learning modeling to simulate soil pH dynamics.

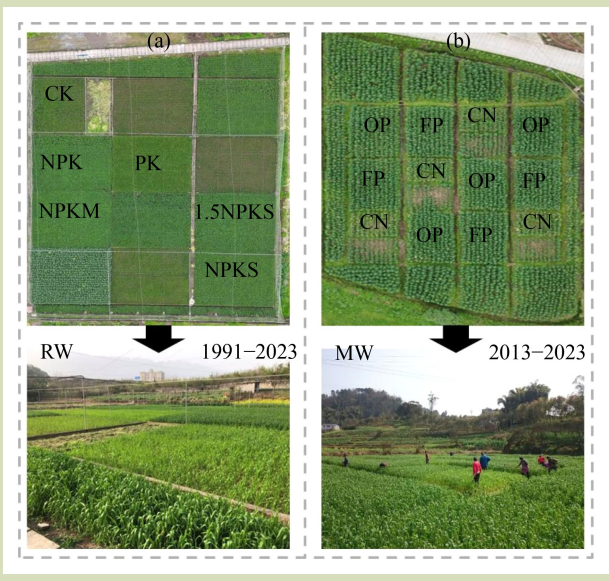


Fig. 1 Photographs of the two long-term fertilizer experiments. Photographs of the two long-term fertilizer experiments used in this study. (a) Beibei had a rice–wheat (RW) rotation with six treatments: CK, a control with no fertilizer; PK, mineral P and K, NPK, mineral N, P and K fertilizers; NPKS, mineral N, P and K plus straw; 1.5NPKS, 1.5 times mineral N, P and K plus straw; and NPKM, mineral N, P and K plus manure. (b) Jiangjin had a maize-wheat (MW) rotation with three treatments: CN, a control with non-N application but with optimized P and K; OP, optimal N, P and K application; and FP, farmer practice with excessive N, P and K application.

2 Materials and methods

2.1 Site information

We applied the VSD+ model for two long-term field experiments with fertilizer application treatments to compare the difference in simulation accuracy of soil pH value. These two sites provided data from different monitoring years, which were used for data set construction and model calibration. The sites were in Beibei (106°26′ E, 30°26′ N) and Jiangjin (106°11′ E, 29°03′ N) districts, Chongqing, China (Fig. 1). Both sites have purple soils classified as the Regosols according to the FAO soil classification system[24] and developed from the purple rocks of Shaximiao formation[25]. Both locations have a similar subtropical humid climate, with an average annual temperature of 18.7 °C in Beibei and 19.3 °C in Jiangjin. Yearly precipitation is 1162 mm in Beibei and 1069 mm in Jiangjin site.

The Beibei site has been managed under a standard rice–wheat rotation since 1991. The initial physical and chemical properties of topsoil (0–20 cm) were: clay content, 26.8%; pH, 7.7; calcium carbonate content, 5.6 g·kg^{-1}; bulk density, 1.38 g·cm^{-3}; soil organic carbon (SOC), 1.10%; total nitrogen (TN), 1.25 g·kg^{-1}; and cation exchange capacity (CEC), 240 mmol·kg^{-1}. Six experimental treatments of the Beibei experiment were included in this study: (1) a control with no fertilizer (CK), (2) mineral P and K (PK), (3) mineral N, P and K (NPK), (4) mineral N, P and K plus straw (NPKS), (5) 1.5 times mineral N, P and K plus straw, and (6) mineral N, P and K plus manure (NPKM). Each plot was 120 m^2 (10 m × 12 m),

with no replication and separated by 60 cm concrete baffles to prevent nutrient movement by water between plots. At this site rice seedlings are transplanted in mid or late May and harvested in mid or late August each year. Wheat is sown in early November and harvested in early May of next year. The planting for both rice and wheat is 24 cm between rows and 16.7 cm between planting positions, about 250,000 positions ha^{-1}. The seeding rate of wheat is 1.4–1.5 kg per plot, about 10 seeds planting per position. Rice is transplanted with 2–3 tillers per plant[26].

The Jiangjin experiment began in 2012 under a maize-wheat rotation system. The initial physical and chemical properties of the topsoil (0–20 cm) were: clay content, 21.7%; pH, 4.9; bulk density, 1.15 $g{\cdot}cm^{-3}$; SOC, 1.15%; TN, 1.99 $g{\cdot}kg^{-1}$; and CEC, 295 $mmol{\cdot}kg^{-1}$.The three fertilizer treatments at Jiangjin were: (1) a control with non-N application but with optimized P and K (CN), (2) optimal N, P and K application (OP), and (3) farmer practice with excessive N, P and K application (FP). This experiment used a randomized complete block design with three treatments and four replicates. The plots were 40 m^2 (5 m × 8 m) except for the control treatment, which was half a plot (5 m × 4 m). These plots separated by 1-m-wide buffer zones to prevent nutrient movement and runoff losses between plots. The Jiangjin site has been managed with maize sown in mid-May and harvested in low August. Then wheat is sown in early November and harvested at the end of May the following year. Maize is transplanted at the two leaf with two plants in each planting position to give a planting density of 50,000 plants ha^{-1}, hole spacing and row spacing of 40 and 100 cm, respectively. Wheat is sown in holes, sowing 100 $kg{\cdot}ha^{-1}$, planting position and row spacing of 20 and 27 cm, respectively, with about 10 seeds planted in each position[25].

The amount of chemical and organic fertilizer application in the Beibei and Jiangjin experiments for each treatment is given in Table 1. Mineral N, P and K fertilizers used were urea, calcium superphosphate and potassium sulfate, respectively. Soil samples (0–20 cm) were collected annually from each plot after rice (maize) or wheat harvest. Five soil samples were randomly collected using the five-point sampling method and mixed thoroughly to form a composite sample. After removing fine roots and debris, the soil samples were passed through a 2-mm sieve, air-dried and physical and chemical properties determined. In addition, nutrients such as N, P, K, Ca, Mg and S were also measured in fresh plant samples including straw, leaves and grains. All data for this study were obtained from the historical monitoring databases for the two sites (e.g.,

Table 1 Mineral and organic fertilizers application rates ($kg{\cdot}ha^{-1}$) in treatments used from the Beibei (1991–2023) and Jiangjin (2013–2023) experiments

Treatment	Wheat				Rice–maize		
	N	P	K	Manure-straw	N	P	K
Beibei							
CK	0	0	0	–	–	–	–
PK	0	33 (26)	62 (50)	–	0	33 (26)	62 (50)
NPK	150 (135)	33 (26)	62 (50)	–	150	33 (26)	62 (50)
NPKS	150 (135)	33 (26)	62 (0)	7500	150	33 (26)	62 (50)
1.5NPKS	225 (202)	49 (39)	93 (75)	22500 (7500)	225	49 (39)	93 (75)
NPKM	150 (135)	33 (26)	62 (50)	22500	150	33 (26)	62 (50)
Jiangjin							
CN	0	26	37	–	0	26	62
OP	96	26	37	–	150	26	62
FP	180	52	37	–	220	52	62

Note: Beibei: CK, a control with no fertilizer; PK, mineral P and K, NPK, mineral N, P and K fertilizers; NPKS, mineral N, P and K plus straw; 1.5NPKS, 1.5 times mineral N, P and K plus straw (manure was applied from 1991 to 1996 and thereafter to rice straw retained); and NPKM, mineral N, P and K plus manure. Values in brackets are the elemental values from 1997 to now. Jiangjin: CN, a control with non-N application but with optimized P and K; OP, optimal N, P and K application; and FP, farmer practice with excessive N, P and K application.

annual soil pH, crop yield and nutrient element concentrations). The data resources for Beibei (1991–2023) and Jiangjin (2013–2023) were used for model training, calibration and validation.

2.2 Model description

The VSD+ model is an extension of the VSD model[11,27], simulating soil acidification processes in a single-soil layer. It consists of a set of mass balance and equilibrium equations to simulate base saturation and pH changes in soil solution chemistry. This process is determined by elemental input from inorganic and organic fertilizers, N fixation and deposition, crop nutrient uptake, net mineralization/immobilization, nitrification and denitrification as well as soil buffering processes. It also considers annual fluxes for wet and dry deposition, weathering, and ion uptake rates. The elements leaching fluxes are derived by multiplying water fluxes with dissolved element concentrations and annual water flux is calculated using the MetHyd model[11]. In the VSD+ mode, crop nutrient uptake is calculated as the elemental content of the crop yield multiplied by the harvested biomass. N uptake is further differentiated into NH_4^+ and NO_3^-. The VSD+ model assumes that NH_4^+ is taken up preferentially, followed by NO_3^-. The net mineralization or immobilization of N in soil is calculated from the turnover of carbon pools based on the RothC model and the fixed C:N ratios for five C pools[28]. Nitrification and denitrification are modeled as first-order rate processes, which were affected by temperature and soil moisture. Soil buffering processes include weathering, cation exchange (Gaines-Thomas or Gapon equations), and dissolution of Al hydroxides according to a gibbsite equilibrium. This study used the Gapon equation to model cation exchange, with K, Ca and Mg aggregated as base cations, disregarding the interactions between Na and the adsorption complex. The soil pH is calculated in the model by the charge balance principle where H^+ is determined by the sum of all considered anions minus cations (Table S1). In addition, the VSD+ model also requires multiple parameters of soil property as inputs, such as CEC, base saturation, partial CO_2 pressure and the amount of soil C and N. A detailed description of the model equations is also provided in Table S1.

2.3 Data sources

The input data of the VSD+ model include nutrient inputs to fields, nutrient removal by crops, soil properties, and meteorology. We obtained the VSD+ model input parameters from field observation, literature and other preprocessor model outputs. Below, we introduced data sources briefly. Table S2 provides details on the input parameters of the VSD+ model for each site.

2.3.1 Nutrient inputs by fertilizer and deposition

The nutrient inputs from chemical and organic fertilizer to the field were documented based on historical application rate and concentration. Also, the total deposition rates for NH_4^+, NO_3^- and SO_4^{2-} during 1991–2008 were derived from Zeng et al.[13] having been based on a deposition monitoring network at the provincial level[29]. However, after 2008, the N and S deposition used the average due to data limitations. For base cation deposition (including K, Ca and Mg) during 1991–2023, we used a value of 0.3 eq $m^{-2}{\cdot}yr^{-1}$[30]. In this study, we assumed that the atmospheric deposition data in the two long-term fertilizer experiments were the same due to the close distance. As required by the model inputs format, the units of calculation for nutrients (ions) have been converted from $kg{\cdot}ha^{-1}{\cdot}yr^{-1}$ to keq $ha^{-1}{\cdot}yr^{-1}$, referring to the supplementary material of Zhu et al.[31]. Nutrient inputs by fertilizer and deposition are given in Table S3.

2.3.2 Nutrient removal by crops

The removal of nutrients was calculated by multiplying the dry matter yields of grain and straw by their respective nutrient concentrations. Crop yield and the concentrations of nutrients (N, P and K) were measured after harvest each year. However, the nutrient concentrations of Ca, Mg and S were not consistently available. These elements were only measured in the Beibei experiment from 2001 to 2013 and in the Jiangjin experiment in 2016. Thus, we use the average value of Ca, Mg and S concentration from 2001 to 2013 as the missing data for the Beibei experimental site.

2.3.3 Soil properties

The initial soil pH, bulk density, soil texture, TN, SOC and CEC of purple soil in the Beibei and Jiangjin experiments are detailed in site information given above. Due to the close distance of the two sites, and the same soil type and parent materials, we assumed that the initial soil weathering rates of the two sites were the same. The weathering rates of base cations were set at 5.3, 1.73, 0.4 and 0.2 $keq{\cdot}ha^{-1}{\cdot}yr^{-1}$ for Ca^{2+}, K^+, Mg^{2+} and Na^+, respectively[13]. The CO_2 pressure in the

soil solution, determining the concentration of HCO_3^-, was set to a value of 10, following the recommendations of Koehler et al.[32] and Nan et al.[33]. Parameters in VSD+ model describing the amounts of S and P adsorbed (in meq kg^{-1}) are the maximum adsorption capacity of the soil (Xadmax in meq kg^{-1} with X being SO_4^{2-} or $H_2PO_4^-$), the dissolved concentration (meq L^{-1}) and the half-saturation constant (Xhalf in meq L^{-1} with X being SO_4^{2-} or $H_2PO_4^-$). The H_2PO_4 admax and H_2PO_4 half were set at 20.8 and 0.41 in the supporting information[34]. The SO_4^{2-} admax was 25 times lower than that for P and SO_4 half was assumed to be equal to H_2PO_4 half following de Vries et al.[35] and Schoumans[36]. Therefore, the SO_4^{2-} admax was set at 0.83 and SO_4^{2-} half at 0.41. Details of soil properties used in the VSD+ model input data for the Beibei and Jiangjin sites are given in Table S2.

2.3.4 Meteorology

The VSD+ model requires the parameters of soil water content and precipitation surpluses to calculate or output the concentration of ionic nutrients (such as H^+, NH_4^+, SO_4^{2-} and NO_3^-) in soil percolation water. Nutrient concentration was multiplied by the precipitation surplus yields to give the amount of nutrient leaching from the soil. These two parameters could be output by running the hydrological model MetHyd, a preprocessor for the VSD+ model[11]. We calculated soil water contents and precipitation surpluses using the MetHyd model. Required input information of the MetHyd are site coordinate, soil data (bulk density, soil texture and SOC content) and daily (monthly) meteorological data (precipitation, temperature and sunshine duration). Those data were taken from the meteorological stations located in the two long-term experimental stations. The reduction functions for N mineralization, nitrification and denitrification, influenced by temperature and/or soil moisture, were computed with the MetHyd model and given in Table S2.

2.4 Model calibration

In the VSD+ model, the primary factors influencing changes in soil pH are twofold: first, the constants related to the maximum rates of nitrification and denitrification, which significantly impact the generation of acidity through N transformations, and secondly, the selectivity constants for the exchange processes involving H and Al to base cation ratios, and the dissolution of aluminum hydroxides, which primarily dictate the rate of acid buffering. The initial value of these constants was sourced from existing literature[37]. To calibrate the VSD+ model, we used a Bayesian calibration procedure[38] with available observed soil pH. Through model calibration, we subsequently fine-tuned the exchange constants, the dissolution constant for aluminum hydroxide, and the denitrification parameter within a reasonable range. This calibration was achieved by minimizing the discrepancy between measured and simulated soil pH values accounting for observed variations in soil acidity. The feasible range of exchange rates used in the calibration was derived from the work of de Vries et al.[37]. The equilibrium constants for aluminum were assumed to fluctuate within a range of 7 to 9, consistent with the findings of Reinds & de Vries[39].

2.5 Machine learning models

To compare the simulation performance differences between process-based and machine learning models, we implemented four commonly used machine learning methods to simulate the long-term dynamics of purple soil pH. They included RF, support vector machine (SVM), extreme gradient boosting (XGB) and decision tree (DT). The RF algorithm, an ensemble machine learning technique, can be used for classification and regression tasks to generate statistical prediction, where each tree is grown on a bootstrap sample of the training data set[40,41]. The SVM is a supervised machine learning method applied to regression and classification processes. All input data for SVM are mapped into a new hyperspace using Kernel functions and then separated. The main goal of SVM modeling is to find the optimal hyperspace for an N-dimensional data set[42,43]. The XGB is an advanced and efficient algorithm based on gradient boosting decision trees[44]. This integration enhanced model performance and reduced overfitting tendencies. The XGB is also designed to optimize parallel computation, efficiently using multi-core CPUs for training. The DT is a supervised learning model that can be used for both classification and regression tasks. It selects an outcome based on a tree of potential decisions[45,46]. The tree structure resembles a flowchart and evaluates outcomes by considering numerous features and attributes.

We provided the same input variables to train four machine learning models for simulating purple soil pH dynamics. The input variables encompassed climate, nutrient management, and soil background information. Climate variables included temperature and precipitation. Nutrient management variables included nutrient input and removal (N, P, K, Ca, Mg and S).

Soil background information included soil organic matter (SOM) and soil background pH, and we assumed actually observed soil pH from the previous year to represent the soil background pH value. Soil acidification is not only influenced by agricultural-climatic variables but also depends on soil background pH value. The data for the model input variables were derived from historical monitoring data collected at the Beibei and Jiangjin sites. A detailed list of the selected variables for machine learning modeling is given in Table S4.

According to the differences in soil background pH values, purple soils can be sub-classified based on pH into acidic (pH ⩽ 6.5), neutral (6.5 < pH ⩽ 7.5) and calcareous (pH > 7.5)[47,48]. However, the neutral properties of purple soil at Beibei Experimental Station were mainly determined by the initial parent material type, rather than background pH value. According to the Chinese Soil Taxonomic Classification, this soil was a neutral typical Purpli-Udic Cambosols, which developed from purple rock series of Jurassic system[49]. Therefore, we refer her as neutral purple soil. There are significant differences in the acid buffering mechanisms of purple soils with different background pH values. When soil pH is greater than 7, the soil acid buffering substance is calcium carbonate. In the soil pH range from 4 to 7, the acid buffering substances are mainly soil exchangeable cations and silicate minerals. Compared to the other pH ranges, the pH value at this stage is most variable in the presence of an equal input of exogenous active H^+. In the pH range of 3 to 4, the main buffering substances are silicon oxides and aluminum hydroxides. When the pH is less than 3, the buffering substances are iron oxides[8,14].

To ensure the accuracy and objectivity of the simulation results, the input data set for machine learning modeling was randomly divided into 70% for training set and 30% for testing set. The testing set was used to evaluate the predicted performance of the model. Also, we estimated the importance of each input variable for soil pH dynamics by calculating the percentage increase in mean square error in the RF importance ranking[50]. Statistical analyses of the machine learning models were conducted using the open-source software R v4.3.3. The RF analysis, executed using the R packages "rfPermute", "randomForest" and "caret", was used to construct regression models and determine the relative importance of input variables on soil pH. SVM, XGB and DT models were built using the R packages "e1071", "xgboost" and "rpart", respectively.

2.6 Model evaluation

The performance of the VSD+ model and the machine learning approach was assessed by comparing the simulated values with the available observations. To assess the model performance, three quantitative statistical metrics were computed: coefficient of determination (R^2, 0 to 1), the mean absolute error (MAE, 0 to +∞) and the root mean square error (RMSE, 0 to +∞)[51].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}\left(O_i - \overline{O}_i\right)^2} \quad (1)$$

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

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|O_i - P_i| \quad (3)$$

where, n is the sample size, O_i and P_i are the observed and simulated pH values, respectively, and $\overline{O}_i$ is the average of the observed pH values. The R^2 value reflects how well the model simulated pH fits relative to the observed pH, with a higher R^2 (nearer to 1) indicating a better simulation of soil pH dynamics. The RMSE and MAE measure model prediction error, with smaller values indicating better prediction accuracy. Compared to RMSE, MAE is less sensitive to outliers and is more representative of the overall prediction performance of the model.

3 Results

3.1 Relative importance of variables to soil pH dynamics

The results revealed that the background soil pH was the primary factor influencing soil pH dynamics. The RF importance analysis showed that soil background pH value accounted for most of the variability in the soil pH dynamics, explaining 26.8% of the variation in soil pH dynamics, followed by Ca input (15.2%), Ca uptake (8.92%), K input (8.74%), P input (8.08%) and temperature (7.84%) (Fig. 2(a)). The Pearson correlation analysis showed that soil pH was positively significant correlated with soil background pH, precipitation, Ca input, Mg input and S uptake. In contrast, it was negatively correlated with temperature, SOM, N input, P input, S input, K uptake and Ca uptake ($P < 0.05$, Fig. 2(b)). The soil background pH demonstrated the strongest positive correlation with soil pH, achieving a maximum correlation coefficient of 0.97. These

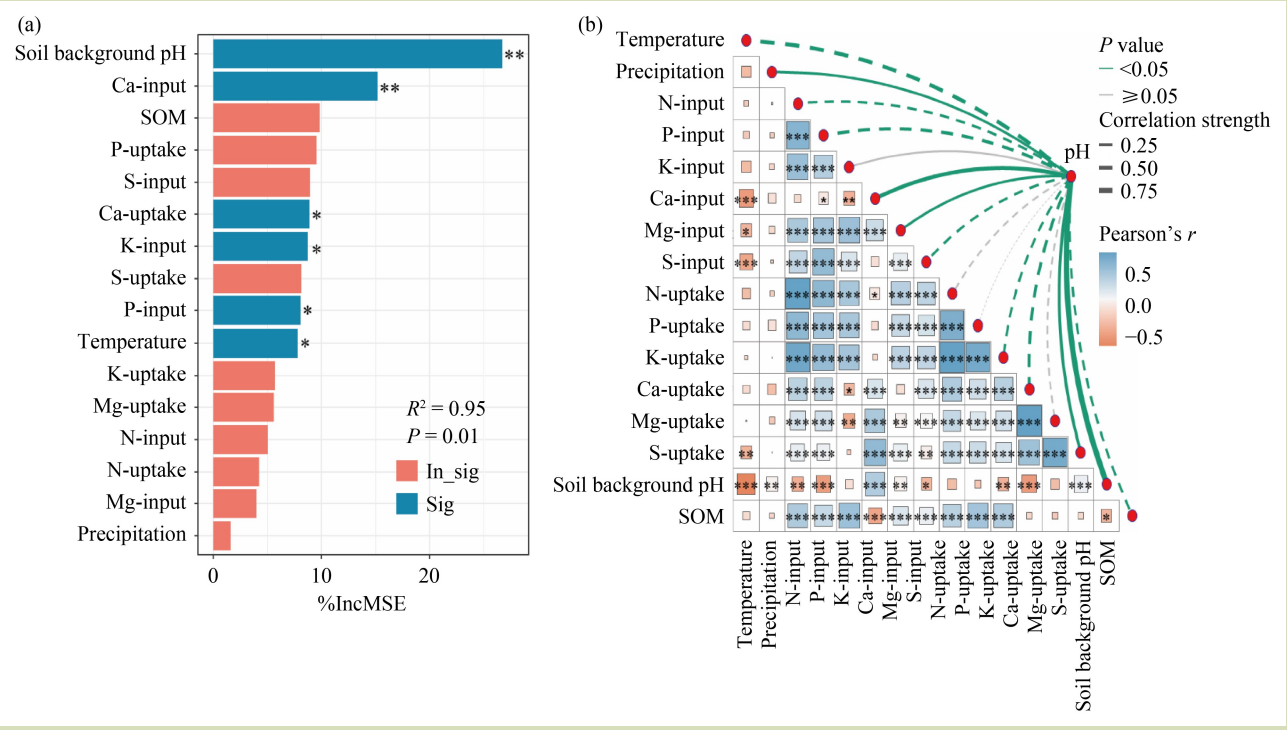


Fig. 2 (a) Random forest importance ranking of variables to soil pH dynamics. The higher percentage increase in mean square error indicates that variables are more important. (b) Pearson correlation of soil pH dynamics with soil background information, climate and nutrient management in purple soil. *, $0.01 < P \leqslant 0.05$; **, $0.001 < P \leqslant 0.01$; and ***, $P \leqslant 0.001$.

results offer an important insight for identifying inputs variable to train machine learning models in simulating soil pH trends.

3.2 Model comparison in neutral purple soil

The machine learning model had a higher performance than the VSD+ model for neutral purple soil. This is evident from the performance of the testing set (Table 2), with higher R^2 (0.64–0.70), smaller RMSE (0.25–0.28) and smaller MAE (0.19–0.22). In contrast, the accuracy of the VSD+ model for simulating soil pH dynamics for neutral purple soil was relatively poor, with R^2, RMSE and MAE of 0.37, 0.37 and 0.28, respectively. The RF model performed best of the machine learning algorithms tested (e.g., SVM, XGB and DT) for simulating pH dynamics for neutral purple soil, achieving the highest accuracy (R^2 = 0.70) with relatively low error rates

Table 2 Performance of simulated models for neutral purple soil pH using multiple methods (n = 198)

Method	Model	R^2	RMSE	MAE
Processed-model	VSD+	0.37	0.37	0.28
Machine learning (30% testing set)	RF	0.70	0.25	0.19
	SVM	0.67	0.26	0.20
	XGB	0.65	0.28	0.21
	DT	0.64	0.28	0.22

Note: VSD+, a process-based soil acidification model[11]; RF, random forest; SVM, support vector machine; XGB, extreme gradient boosting; and DT, decision tree.

(RMSE = 0.25 and MAE = 0.19). For validation the between simulated results and observation, the SVM and XGB models gave comparable performance (SVM: R^2 = 0.67, RMSE = 0.26 and MAE = 0.20 versus XGB: R^2 = 0.65, RMSE = 0.28 and MAE = 0.21). In contrast, the DT model had a slightly lower accuracy (R^2 = 0.64, RMSE = 0.28 and MAE = 0.22).

Compared to the simulation results of the VSD+ model for neutral purple soil, the difference between the simulated pH values from machine learning models and measurements were relatively small (Fig. 3). For example, in the control treatment, observed pH showed no significant soil acidification whereas the simulated pH of the VSD+ model had a gradual decline trend. Specifically, the VSD+ model overestimated the pH by 0.1–0.6 units between 1991 and 2004, and underestimated it by 0.1–0.8 units from 2006 to 2022. In contrast, the RF model gave smaller deviations, with slight overestimations (0.1–0.4 units) during 1991–2005 and 2021–2023 and minor underestimations (0.1–0.3 units) in the 2006–2020 period. In the NPKM treatment, the VSD+ model simulations were higher than the observed pH by 0.1–0.6 units during 1996–2006, but underestimated the observed pH by 0.1–0.4 units (1991–1994) and 0.1–0.7 units (2007–2023). Conversely, the discrepancy between RF model simulations and observed pH values was constrained to within ±0.3 units from 1991 to 2023. In the NPKS treatment, the simulated pH of the VSD+ model was

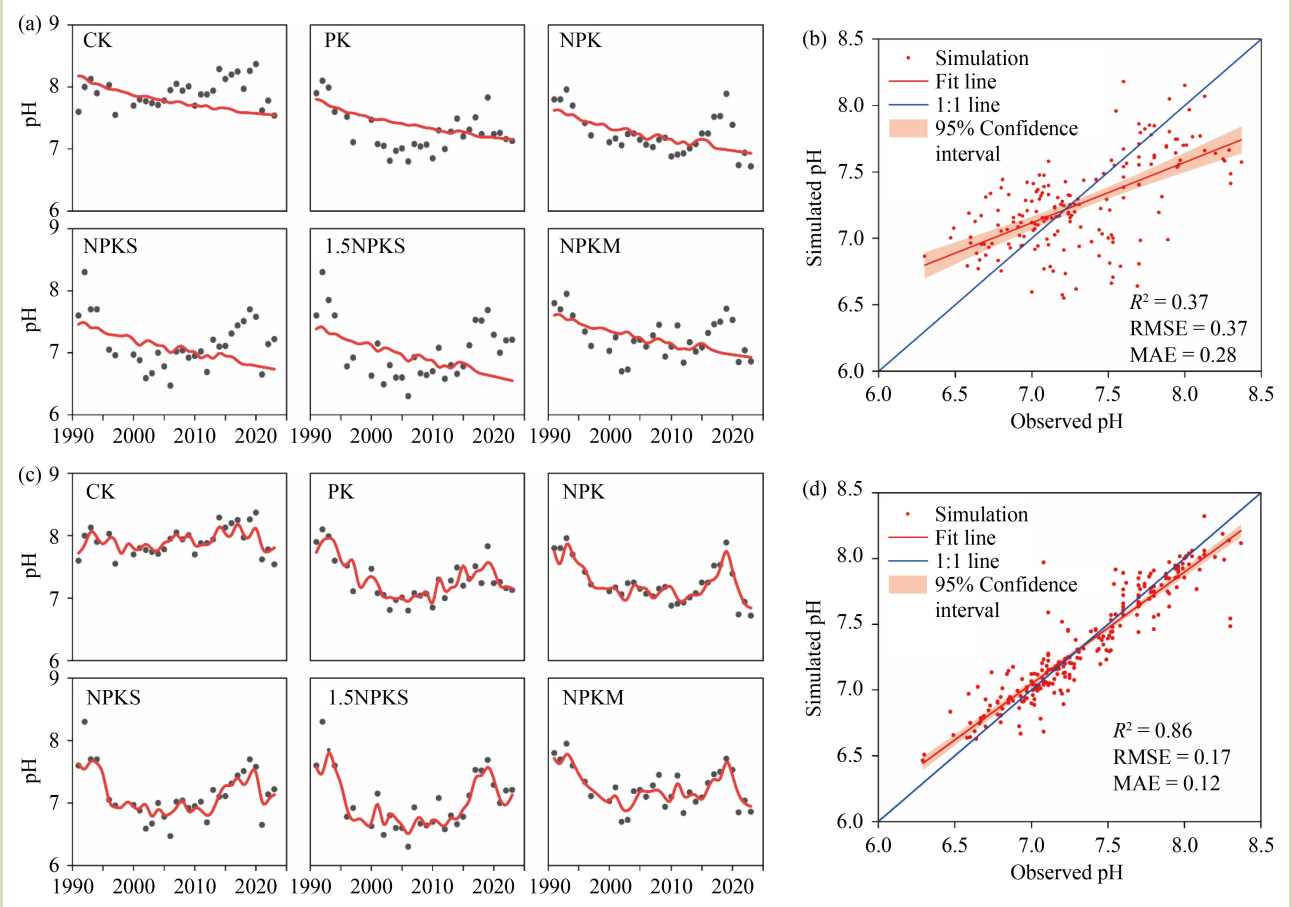


Fig. 3 Comparison of the simulation and fitting performance of the VSD+ and RF models on neutral purple soil. (a) Observed (dots) and simulated (lines) changes in pH and (b) the scatter plot under six treatments by the VSD+ model; (c) observed and simulated changes in pH; and (d) the scatter plot under six treatments by the RF model, respectively. CK, a control with no fertilizer; PK, mineral P and K, NPK, mineral N, P and K fertilizers; NPKS, mineral N, P and K plus straw; 1.5NPKS, 1.5 times mineral N, P and K plus straw; and NPKM, mineral N, P and K plus manure.

slightly higher than the observed pH by 0.1–0.6 units from 1996 to 2010 while lower than the observed pH by 0.1–0.8 units (1991–1994) and 0.1–0.9 units (2011–2023). The RF model gave consistent performance in projecting pH for the NPKS treatment, maintaining deviations within ±0.4 pH units throughout the 1991–2023 period. The results of other machine learning models are given in Figs. S1–S3.

3.3 Model comparison in acidic purple soil

Compared to neutral purple soil, the VSD+ model exhibited better accuracy than the machine learning approaches at acidic purple soil. Specifically, this was evident from the excellent simulation capability of the VSD+ model with higher R^2 (0.95), smaller RMSE (0.05), and smaller MAE (0.02) (Table 3). In contrast, the performance of machine learning models had limitations with R^2 = 0.20–0.47, RMSE = 0.15–0.23 and MAE = 0.14–0.20.

The RF model still outperformed other machine learning models in simulating the pH dynamics for acidic purple soil (Table 3). The result of testing data set, which reflects the effects of model actual generalization, showed that the RF algorithm exhibited the best performance of the machine learning models tested (R^2 = 0.47, RMSE = 0.19, and MAE = 0.17). By comparison, the XGB and SVM models gave moderately lower performance (XGB: R^2 = 0.45, RMSE = 0.15, MAE = 0.14; SVM: R^2 = 0.40, RMSE = 0.20 and MAE = 0.19). The DT model gave the weakest performance of the algorithms tested (R^2 = 0.20, RMSE = 0.23, MAE = 0.20).

The VSD+ model gave superior performance in simulating acidic soil pH dynamics, having closer agreement with observed values compared to the RF model (Fig. 4). For example, in the optimal NPK fertilizer treatment, the RF model gave overestimations of 0.2–0.4 pH units in 2019 and 2022 and underestimates by about 0.2 units in 2020. In contrast, the VSD+ model accurately captured the historical changes in soil pH dynamics from 2013 to 2023. Similar patterns emerged in the farmer practice treatment. The simulations from the RF model were overestimated pH by 0.1–0.2 units during 2017–2023, except for underestimations of 0.3 units in 2019 and 0.1 units in 2022. However, the VSD+ model simulations had minimal deviation from observed pH values, with errors constrained to within ±0.1 units. Both VSD+ and RF models gave good performance in simulating pH for acidic soil in the in the control treatment, with deviations generally within ±0.2 units. RF consistently overestimated by about 0.1 units (except in 2021) whereas VSD+ matched trends well but had about 0.2-unit underestimation in 2021. The performance of other machine learning models is shown in Figs. S4–S6.

4 Discussion

4.1 Strengths and limitations of different modeling methods

The VSD+ model is widely used for simulating future pH trends and examining different scenarios to alleviate soil acidification, but is mainly used for acidic red soil with low acid buffering capacity[52,53]. For example, this model has been applied to the assessment of the long-term impacts of various fertilizer application scenarios on red soil in four typical cropping systems[54]. Also, this model has also been used to determine optimal liming amounts and intervals, as well as the best nutrient management strategies to mitigate soil acidification[52,55]. However, the VSD+ model has inherent limitations for simulating soil acidification processes for both alkaline or $CaCO_3$-rich soils and neutral soils with low $CaCO_3$

Table 3 Performance of simulated models for acidic purple soil pH using multiple methods (n = 33)

Method	Model	R^2	RMSE	MAE
Processed-model	VSD+	0.95	0.05	0.02
Machine learning (30% testing set)	RF	0.47	0.19	0.17
	SVM	0.40	0.20	0.19
	XGB	0.45	0.15	0.14
	DT	0.20	0.23	0.20

Note: A process-based soil acidification model (VSD+)[11]; RF, random forest; SVM, support vector machine; XGB, extreme gradient boosting; and DT, decision tree.

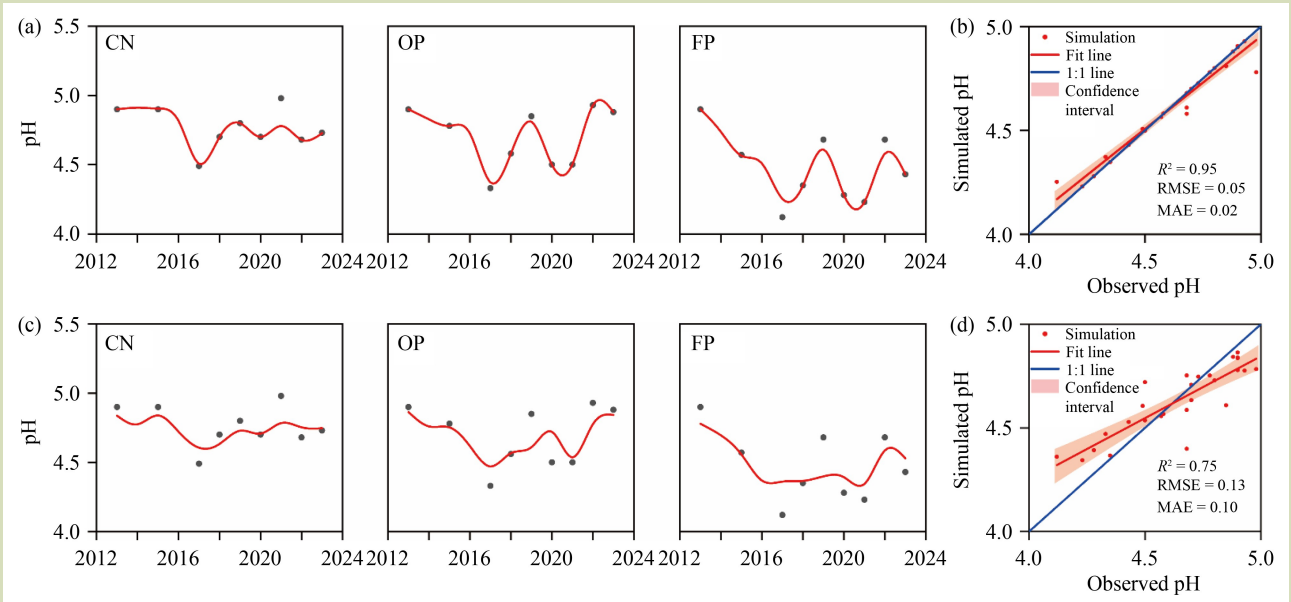


Fig. 4 Comparison of the simulation and fitting performance of the VSD + and RF models on acidic purple soil. (a) Observed (dots) and simulated (lines) changes in pH and (b) the scatter plot under three treatments by the VSD + model; (c) observed and simulated changes in pH; and (d) the scatter plot under three treatments by the RF model, respectively. CN, a control with non-N application but with optimized P and K; OP, optimal N, P and K application; and FP, farmer practice with excessive N, P and K application.

content (≤ 0.4%)[56]. These limitations primarily stem from two of its assumptions: (1) homogeneous distribution of $CaCO_3$ within the soil compartment, and (2) constant $CaCO_3$ content over short time scales[13]. For calcareous soils, this leads to an overestimation of acid buffering capacity, as the model predicts no decrease in pH until complete depletion of $CaCO_3$ reserves, thereby introducing systematic deviations between simulated and observed pH values. In neutral soils with limited calcium carbonate content (≤ 0.4%), spatial heterogeneity of $CaCO_3$ distribution creates uncertainty in surface soil buffering capacity, further contributing to discrepancies between model outputs and field measurements. Also, the implementation of VSD+ model generally relies on several years of observational data for calibration and validation, often from long-term fertilizer experiments[28,57]. It is challenging to meet the requirements for model calibration and validation in most fields due to the limited availability of long-term interannual data, which also somewhat restricts the applicability of VSD+ model.

The results of VSD+ model for the two types of purple soil studied suggest that the model performs better in acidic soil (R^2 = 0.95). However, it performs poorly in neutral soil, even when abundant data are available for calibration (R^2 = 0.37). The key reason for this discrepancy lies in the specific characteristics of neutral purple soil. At the Beibei site, the initial acid buffering capacity of neutral purple soil was relatively strong, as reflected in the higher initial soil pH, which even exhibited slightly weak calcareous properties[26]. However, long-term fertilizer application, atmospheric acid deposition and continuous weathering of the parent rock (resulting in changes to clay minerals) all contribute to ongoing changes in the acid buffering capacity, ultimately leading to instability in the acid buffering capacity of neutral purple soil. Therefore, the VSD+ model has limited applicability for neutral purple soil, which is consistent with findings of previous research[13].

The present study is the first attempt to simulate the long-term dynamics of soil pH using machine learning models. For neutral purple soil with high acid buffering capacity, machine learning gave both superior performance and greater practicality compared to process-based models. Specifically, machine learning models require fewer input variables than the VSD+ model, particularly avoiding the need for soil process parameters that are difficult to measure (e.g., weathering rate of

base cations). Machine learning methods additionally offer two key advantages: (1) the ability to identify and prioritize important variable features while eliminating redundant parameters[58], and (2) the capacity to capture nonlinear relationships between dependent and independent variables, significantly improving the accuracy of prediction[59]. Although machine learning models provide superior performance for neutral purple soil, their application needs to be supported by a large amount of monitoring data, especially field in situ experiments. In addition, machine learning modeling requires further division of the original data sets, that is training and testing data sets, and it is difficult to be completely objective in the selection of input parameters for the training data set. These algorithms are more susceptible to the influence of noise and outliers during the training process, which may lead to overfitting and poor learning outcomes[60,61]. In contrast, the VSD+ model only requires model calibration of local observations before outputting results. Under data-limited conditions, at the Jiangjin site with shorter monitoring periods, machine learning proved particularly ineffective.

4.2 Main drivers in soil pH dynamics

Many studies have revealed the main factors contributing to soil pH changes. However, most research has primarily focused on variables related to human activities (e.g., agronomic management) and atmospheric factors (e.g., acid deposition), with relatively less attention given to the contribution of soil background information (e.g., soil background pH, soil weathering rates and mineral composition) to long-term soil pH dynamics. For example, Chen et al.[62] demonstrated that changes in soil pH were significantly influenced by atmospheric N fertilizer and N deposition in the Mollisols region of China. Additionally, N fertilizer application, base cations by crop uptake, and precipitation were identified as the three main drivers of soil acidification[23]. Jin et al.[63] quantified the contributions of various driving factors to soil pH dynamics using a structural equation model and identified the initial soil pH, mean annual temperature, mean annual precipitation, bulk density, and SOM as key variables influencing soil pH dynamics. Our results also highlight the importance of the initial soil pH value affecting soil pH dynamics. However, existing applications of machine learning models for soil background pH prediction have rarely incorporated soil background pH as an input variable in model training. The replenishment of base cations from mineral weathering in the parent rock during soil development also determines the acid buffering capacity. Xie et al.[64] showed that with the increase of active H^+ content in neutral purple soil, it leads to the disruption of the mineral lattice of soil clay minerals such as illite, which accelerates its mineral weathering to form vermiculite or montmorillonite and then kaolinite. This process releases large amounts of base cations, which counteracts the decrease in soil pH due to exogenous H^+ inputs, such as long-term fertilizer application, and even enhance the acid buffering capacity of neutral purple soil. However, compared to the weathering rate of base cation, soil background pH (i.e., the soil background pH for each year is equivalent to the measured soil pH value of the preceding year) is not only easier to collect and monitor continuously over long timescale but also reflects, to some extent, the acid buffering capacity of the soil.

Soil background pH is an important indicator of the strength of acid buffering capacity, which is a fundamental attribute that determines the rate and extent of pH change under the influence of external factors such as nutrient inputs and climatic conditions. In addition, soil background pH also reflects the level of acid buffering substances (e.g., calcium carbonate and silicate) to a certain extent. Simulated pH represents the results of changes in soil background pH after cultivation and natural activities. Therefore, the incorporation of background soil pH into machine learning models remains critical for accurately simulate pH dynamics. Given these practical benefits and data accessibility, we took soil background pH as a key input variable for our machine learning framework to simulate soil pH dynamics.

4.3 Model uncertainties

Our results on simulated purple soil pH by the VSD+ model may have inherent uncertainties arising from inputs data and empirical parameters. For example, a sensitivity analysis of the VSD+ model by Mol-Dijkstra & Reinds[65] also found that the simulated pH is to a large extent determined by the soil weathering rate. However, the weathering rate of soil base cations (K^+, Na^+, Ca^{2+} and Mg^{2+}) were obtained from the results of Duan[66] for purple soil, rather than actual measured values for each year. The interannual variation of weathering parameters may generate uncertainties in the model simulation. For example, Zhang et al.[26] showed that the weathering rate of neutral purple soil significantly increased after 2002 at the Beibei site. However, the weathering rate of

base cation inputs in this model could only use default or fixed values due to the lack of historical weathering records. This may explain the relatively poor simulation performance of the VSD+ model in neutral purple soil. In contrast, the soil-forming parent material at the Jiangjin site underwent a longer development process, resulting in a relatively constant soil weathering rate[25]. As a result, the observed soil pH during 2013–2023 revealed relatively minor variation. Additionally, due to the limitations of the data (lack of longer monitoring years for in situ experiments on acidic purple soil), we were unable to use data equivalent in duration (33 years) to that of neutral purple soil to train the machine learning model for acidic purple soil. Therefore, our study may not guarantee that the performance of the VSD+ acidification model on acidic purple soil would still outperform machine learning models under the support of equivalent or more historical data. However, under the current data-limited conditions, using the VSD+ model, a process-based model, to simulate the dynamic changes of soil pH in acidic purple soil is a preferable choice.

4.4 Implications and outlooks

This study compared process-based with machine learning modeling methods for simulating soil pH at two purple soil long-term situ sites. We evaluated the performance of each modeling approach with different quantities of observation data for calibration or training. Our results indicated that four machine learning models outperform the VSD+ model in simulating soil pH dynamics in neutral soil, while the VSD+ model provides better performance in acidic soil. This finding contributes to the knowledge gap in modeling approach selection for two purple soils with distinct buffering mechanisms, calcium carbonate and exchangeable base cations, respectively. We have demonstrated for the first time that machine learning modeling is capable of simulating pH dynamic changes in neutral soils well, thereby bridging the current gap in methodologies for pH simulation in the soils with high acid buffering capacity. Also, we found that the VSD+ model is suitable for simulating pH dynamics of acidic purple soils, in addition to its applicability to acidic red soils. The experience could be extended to other distinct types of soil, such as the black soil acidification modeling study in north-eastern China. Our study offers decision-making recommendations for the selection of methods for employing soil acidification models based on the availability of data at different sites.

We highlight the critical need to establish a richer and more diverse data set for training that incorporate greater diversity of soil types, including broader pH ranges, even calcareous soil. In that case, enhanced data sets would enable the development of more robust models capable of accurate transfer and application across varied soil environments. The RF model we trained only accounted for purple soils within a specific pH range (4.12–4.98 and 6.3–8.37), which limits the generalizability of the trained model for simulating pH dynamics of a broader spectrum of purple soils. The growing availability of high-resolution data sets in meteorology, nutrient management and soil information, offers significant opportunities for advancing machine learning applications in this field. Notably, Liu et al.[67] generated 90-m resolution national gridded maps of key soil properties (e.g., pH and SOM) across multiple depths in China. Also, comprehensive multiyear data sets for crop-specific N, P and K fertilizer application rates at high resolution have become increasingly available[68,69]. Climatic variables, second only to soil background pH in driving soil pH variation, are also accessible in high-resolution data sets, even extending to daily-scale records[70,71]. The ongoing Third National Soil Census in China will further enhance data quality and completeness by establishing a more robust soil information system. These publicly available databases enable us to train more sophisticated machine learning models, capable of simulating the pH dynamics of various soil types with distinct buffering mechanisms under diverse agronomic management practices and meteorological conditions in the future.

5 Conclusions

This study compared the performance of a process-based model and machine learning models for simulating pH dynamics of purple soils with distinct buffering mechanisms, calcium carbonate and exchangeable base cations, respectively. The results indicated that soil background pH was the primary factor influencing dynamic changes in soil pH, which should be prioritized as input variable when training machine learning models for soil acidification simulation. A strong positive correlation and significant relationship were observed between soil background pH and its dynamic changes. In neutral purple soil, machine learning models outperformed the VSD+ method in simulating soil pH dynamics. Among machine learning models, the RF model achieved the best simulation capability at both sites. In acidic purple soil, the VSD+ model provided

exceptional accuracy compared to the other machine learning models. Our research offers information that should be of value for choosing an appropriate modeling method to simulate acidification process in purple soils with distinct acid buffering mechanisms, thus supporting informed decision-making in agricultural management policies.

Supplementary materials

The online version of this article at https://doi.org/10.15302/J-FASE-2025658 contains supplementary materials (Figs. S1–S6; Tables S1–S4).

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2022YFD1901404), the Regional Innovation Cooperation Foundation of Sichuan, China (2023YFQ0025) and the Foundation of Graduate Research and Innovation in Chongqing, China (CYB23122).

Compliance with ethics guidelines

Haiyang Huang, Xuanjing Chen, Yuting Zhang, Tao Guo, Shuai Wang, Jia Zhou, Zhiqi Li, Yang Wang, Yueqiang Zhang, and Xiaojun Shi declare that they have no conflicts of interest or financial conflicts to disclose. This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Duan L, Chen X, Ma X X, Zhao B, Larssen T, Wang S X, Ye Z X. Atmospheric S and N deposition relates to increasing riverine transport of S and N in southwest China: implications for soil acidification. *Environmental Pollution*, 2016, **218**: 1191–1199
2. Guo J H, Liu X J, Zhang Y, Shen J L, Han W X, Zhang W F, Christie P, Goulding K W T, Vitousek P M, Zhang F S. Significant acidification in major Chinese croplands. *Science*, 2010, **327**(5968): 1008–1010
3. Wu Z F, Sun X M, Sun Y Q, Yan J Y, Zhao Y F, Chen J. Soil acidification and factors controlling topsoil pH shift of cropland in central China from 2008 to 2018. *Geoderma*, 2022, **408**: 115586
4. Liu C Y, Jiang M T, Yuan M M, Wang E T, Bai Y, Crowther T W, Zhou J Z, Ma Z Y, Zhang L, Wang Y, Ding J X, Liu W X, Sun B, Shen R F, Zhang J B, Liang Y T. Root microbiota confers rice resistance to aluminium toxicity and phosphorus deficiency in acidic soils. *Nature Food*, 2023, **4**(10): 912–924
5. Xu D H, Zhang J Z, Li Y J, Li S, Ren S Y, Feng Y, Zhu Q C, Zhang F S. Large-scale farming benefits soil acidification alleviation through improved field management in banana plantations. *Frontiers of Agricultural Science and Engineering*, 2023, **10**(1): 48–60,154
6. Du L X, Zhang Z Y, Chen Y Q, Wang Y, Zhou C X, Yang H Y, Zhang W. Heterogeneous impact of soil acidification on crop yield reduction and its regulatory variables: a global meta-analysis. *Field Crops Research*, 2024, **319**: 109643
7. Zhu Q C, Liu X J, Hao T X, Zeng M F, Shen J B, Zhang F S, de Vries W. Cropland acidification increases risk of yield losses and food insecurity in China. *Environmental Pollution*, 2020, **256**: 113145
8. de Vries W, Posch M, Kämäri J. Simulation of the long-term soil response to acid deposition in various buffer ranges. *Water, Air, and Soil Pollution*, 1989, **48**(3–4): 349–390
9. Cosby B J, Hornberger G M, Galloway J N, Wright R F. Modeling the effects of acid deposition: assessment of a lumped parameter model of soil water and streamwater chemistry. *Water Resources Research*, 1985, **21**(1): 51–63
10. Warfvinge P, Falkengren-Grerup U, Sverdrup H, Andersen B. Modelling long-term cation supply in acidified forest stands. *Environmental Pollution*, 1993, **80**(3): 209–221
11. Bonten L T C, Reinds G J, Posch M. A model to calculate effects of atmospheric deposition on soil acidification, eutrophication and carbon sequestration. *Environmental Modelling & Software*, 2016, **79**: 75–84
12. Bonten L T C, Reinds G J, Groenenberg J E, de Vries W, Posch M, Evans C D, Belyazid S, Braun S, Moldan F, Sverdrup H U, Kurz D. Dynamic Geochemical Models to Assess Deposition Impacts and Target Loads of Acidity for Soils and Surface Waters. In: Critical Loads and Dynamic Risk Assessments.

Dordrecht: *Springer Netherlands*, 2015, 225–251

13. Zeng M F, de Vries W, Bonten L T C, Zhu Q C, Hao T X, Liu X J, Xu M G, Shi X J, Zhang F S, Shen J B. Model-based analysis of the long-term effects of fertilization management on cropland soil acidification. *Environmental Science & Technology*, 2017, **51**(7): 3843–3851
14. Wang P S, Xu D H, Lakshmanan P, Deng Y, Zhu Q C, Zhang F S. Mitigation strategies for soil acidification based on optimal nitrogen management. *Frontiers of Agricultural Science and Engineering*, 2024, **11**(2): 229–242
15. Zhang C, Zhu Z D, Liu F, Yang Y, Wan Y, Huo W W, Yang L. Efficient machine learning method for evaluating compressive strength of cement stabilized soft soil. *Construction & Building Materials*, 2023, **392**: 131887
16. Zhao Y, Fan D, Li Y L, Yang F. Application of machine learning in predicting the adsorption capacity of organic compounds onto biochar and resin. *Environmental Research*, 2022, **208**: 112694
17. von Bloh M, Lobell D, Asseng S. Knowledge informed hybrid machine learning in agricultural yield prediction. *Computers and Electronics in Agriculture*, 2024, **227**(Part2): 109606
18. Chergui N. Durum wheat yield forecasting using machine learning. *Artificial Intelligence in Agriculture*, 2022, **6**: 156–166
19. Wang Y C, Xu C, Lin Q R, Xiao W P, Huang B Q, Lu W F, Chen N W, Chen J X. Modeling of algal blooms: advances, applications and prospects. *Ocean and Coastal Management*, 2024, **255**: 107250
20. Hu B F, Xie M D, Shi Z, Li H Y, Chen S C, Wang Z G, Zhou Y, Ni H J, Geng Y B, Zhu Q, Zhang X L. Fine-resolution mapping of cropland topsoil pH of southern China and its environmental application. *Geoderma*, 2024, **442**: 116798
21. Gong C, Quan L C, Chen W B, Tian G L, Zhang W, Xiao F, Zhang Z X. Ecological risk and spatial distribution, sources of heavy metals in typical purple soils, southwest China. *Scientific Reports*, 2024, **14**(1): 11342
22. Wei C F, Shao J A, Ni J P, Gao M, Xie D T, Pan G X, Hasegawa S. Soil aggregation and its relationship with organic carbon of purple soils in the Sichuan basin, China. *Agricultural Sciences in China*, 2008, **7**(8): 987–998
23. Li Q Q, Li S, Xiao Y, Zhao B, Wang C Q, Li B, Gao X S, Li Y D, Bai G C, Wang Y D, Yuan D G. Soil acidification and its influencing factors in the purple hilly area of southwest China from 1981 to 2012. *Catena*, 2019, **175**: 278–285
24. Wrb I, Schád P, van Huyssteen C, Micheli E. World Reference Base for Soil Resources 2014, update 2015: International Soil Classification System for Naming Soils and Creating Legends for Soil Maps. In: World Soil Resources Reports No 106. Rome: *FAO*, 2015, 172–173
25. Hao T X, Liu X J, Zhu Q C, Zeng M F, Chen X J, Yang L S, Shen J B, Shi X J, Zhang F S, de Vries W. Quantifying drivers of soil acidification in three Chinese cropping systems. *Soil & Tillage Research*, 2022, **215**: 105230
26. Zhang Y T, de Vries W, Thomas B W, Hao X Y, Shi X J. Impacts of long-term nitrogen fertilization on acid buffering rates and mechanisms of a slightly calcareous clay soil. *Geoderma*, 2017, **305**: 92–99
27. Posch M, Reinds G J. A very simple dynamic soil acidification model for scenario analyses and target load calculations. *Environmental Modelling & Software*, 2009, **24**(3): 329–340
28. Coleman K, Jenkinson D S. RothC-26.3—A Model for the turnover of carbon in soil. In: Powlson D S, Smith P, Smith J U, eds. Evaluation of Soil Organic Matter Models. Berlin, Heidelberg: *Springe*r, 1996: 237–246
29. Liu X J, Zhang Y, Han W X, Tang A H, Shen J B, Cui Z L, Vitousek P, Erisman J W, Goulding K, Christie P, Fangmeier A, Zhang F S. Enhanced nitrogen deposition over China. *Nature*, 2013, **494**(7438): 459–462
30. Duan L, Lin Y, Zhu X Y, Tang G G, Gao D F, Hao J M. Modeling atmospheric transport and deposition of calcium in China. *Journal of Tsinghua University (Science and Technology)*, 2007, **47**(9): 1462–1465 (in Chinese)
31. Zhu Q C, de Vries W, Liu X J, Hao T X, Zeng M F, Shen J B, Zhang F S. Enhanced acidification in Chinese croplands as derived from element budgets in the period 1980–2010. *Science of the Total Environment*, 2018, **618**: 1497–1505
32. Koehler B, Zehe E, Corre M D, Veldkamp E. An inverse analysis reveals limitations of the soil-CO_2 profile method to calculate CO_2 production and efflux for well-structured soils. *Biogeosciences*, 2010, **7**(8): 2311–2325
33. Nan W G, Yue S C, Li S Q, Huang H Z, Shen Y F. The factors related to carbon dioxide effluxes and production in the soil profiles of rain-fed maize fields. *Agriculture, Ecosystems & Environment*, 2016, **216**: 177–187
34. Jia X Y, Li J M. Study on soil phosphorus availability and its relation to the soil properties in 14 soils from different sites in China. *Soil and Fertilizer Sciences in China*, 2011, (6): 76–82 (in Chinese)
35. de Vries W, Kros J, van Der Salm C. Long-term impacts of various emission deposition scenarios on Dutch forest soils. *Water, Air, and Soil Pollution*, 1994, **75**(1–2): 1–35
36. Schoumans O F. Description of the phosphorus sorption and desorption processes in lowland peaty clay soils. *Soil Science*, 2013, **178**(6): 291–300
37. de Vries W, Maximilian P. Derivation of cation exchange constants for sand, loess, clay and peat soils on the basis of field measurements in the Netherlands. Wageningen: *Alterra*, 2003
38. Reinds G J, van Oijen M, Heuvelink G B M, Kros H. Bayesian calibration of the VSD soil acidification model using European

forest monitoring data. *Geoderma*, 2008, **146**(3–4): 475–488
39. Reinds G J, de Vries W. Uncertainties in critical loads and target loads of sulphur and nitrogen for European forests: analysis and quantification. *Science of the Total Environment*, 2010, **408**(8): 1960–1970
40. Prasad A M, Iverson L R, Liaw A. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 2006, **9**(2): 181–199
41. Breiman L. Random forests. *Machine Learning*, 2001, **45**(1): 5–32
42. Smola A J, Schölkopf B. A tutorial on support vector regression. *Statistics and Computing*, 2004, **14**(3): 199–222
43. Guo Q H, Kelly M, Graham C H. Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecological Modelling*, 2005, **182**(1): 75–90
44. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California, USA: *ACM*, 2016, 785–794
45. Perez-Alonso D, Peña-Tejedor S, Navarro M, Rad C, Arnaiz-González Á, Díez-Pastor J F. Decision Trees for the prediction of environmental and agronomic effects of the use of Compost of Sewage Slugde (CSS). *Sustainable Production and Consumption*, 2017, **12**: 119–133
46. Rokach L, Maimon O. Data Mining with Decision Trees. Israel: *World scientific*, 2013, 1–7
47. Zhao X L, Jiang T, Du B. Effect of organic matter and calcium carbonate on behaviors of cadmium adsorption–desorption on/from purple paddy soils. *Chemosphere*, 2014, **99**: 41–48
48. Zhou Z F, Shi X J, Zheng Y, Qin Z X, Xie D T, Li Z L, Guo T. Abundance and community structure of ammonia-oxidizing bacteria and Archaea in purple soil under long-term fertilization. *European Journal of Soil Biology*, 2014, **60**: 24–33
49. Zhao Y N, Zhang Y Q, Liu X Q, He X H, Shi X J. Carbon sequestration dynamic, trend and efficiency as affected by 22-year fertilization under a rice–wheat cropping system. *Journal of Plant Nutrition and Soil Science*, 2016, **179**(5): 652–660
50. Janitza S, Celik E, Boulesteix A L. A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, 2018, **12**(4): 885–915
51. Janssen P H M, Heuberger P S C. Calibration of process-oriented models. *Ecological Modelling*, 1995, **83**(1–2): 55–66
52. Xu D H, Zhu Q C, Ros G H, Xu M G, Wen S L, Zhang F S, de Vries W. Model-based optimal management strategies to mitigate soil acidification and minimize nutrient losses for croplands. *Field Crops Research*, 2023, **292**: 108827
53. Xu D H, Ros G H, Zhu Q C, Zhang F S, de Vries W. Spatial optimization of manure and fertilizer application strategies to minimize nutrient surpluses and acidification rates in croplands of a typical Chinese county. *Journal of Cleaner Production*, 2025, **503**: 145401
54. Zhu Q C, Liu X J, Hao T X, Zeng M F, Shen J B, Zhang F S, de Vries W. Modeling soil acidification in typical Chinese cropping systems. *Science of the Total Environment*, 2018, **613–614**: 1339–1348
55. Xu D H, Carswell A, Zhu Q C, Zhang F S, de Vries W. Modelling long-term impacts of fertilization and liming on soil acidification at Rothamsted experimental station. *Science of the Total Environment*, 2020, **713**: 136249
56. de Vries W, Klijn J A, Kros J. Simulation of the long-term impact of atmospheric deposition on dune ecosystems in the Netherlands. *Journal of Applied Ecology*, 1994, **31**(1): 59–73
57. Cosby B J, Ferrier R C, Jenkins A, Wright R F. Modelling the effects of acid deposition: refinements, adjustments and inclusion of nitrogen dynamics in the MAGIC model. *Hydrology and Earth System Sciences*, 2001, **5**(3): 499–518
58. Sathya P, Gnanasekaran P. Ensemble feature selection framework for paddy yield prediction in Cauvery Basin using machine learning classifiers. *Cogent Engineering*, 2023, **10**(2): 2250061
59. Liu P, Lou S Y, Shen H P, Wang M X. Machine learning for prediction of energy consumption and broken force in the chopping process of maize straw. *Agronomy*, 2023, **13**(12): 3030
60. Umutoni L, Samadi V. Application of machine learning approaches in supporting irrigation decision making: a review. *Agricultural Water Management*, 2024, **294**: 108710
61. Bhakta I, Phadikar S, Majumder K, Mukherjee H, Sau A. A novel plant disease prediction model based on thermal images using modified deep convolutional neural network. *Precision Agriculture*, 2023, **24**(1): 23–39
62. Chen J, Xie E Z, Peng Y X, Yan G J, Jiang J, Hu W Y, Zhao Y G, Khan K S, Zhao Y C. Four-dimensional modelling reveals decline in cropland soil pH during last four decades in China's Mollisols region. *Geoderma*, 2025, **453**: 117135
63. Jin J, Huang X Z, Wu J S, Zhao W M, Fu W J. A 10-year field experiment proves the neutralization of soil pH in Chinese hickory plantation of southeastern China. *Journal of Soils and Sediments*, 2022, **22**(12): 2995–3005
64. Xie J, Wang D, Chen Y X, Li Z Q, Dai W C, Huang R, Wang Z F, Gao M. Neutral purple soil acidification and mineralogical property changes due to long-term urea application in southwest China. *Soil & Tillage Research*, 2024, **244**: 106227
65. Mol-Dijkstra J P, Reinds G J. Technical Documentation of the Soil Model VSD+: Status A. Wageningen: *WOT Natuur & Milieu*, 2017
66. Duan L. Study on Mapping Critical Loads of Acid Deposition

in China. Dissertation for the Doctoral Degree. Beijing: *Tsinghua University*, 2000 (in Chinese)

67. Liu F, Wu H Y, Zhao Y G, Li D C, Yang J L, Song X D, Shi Z, Zhu A X, Zhang G L. Mapping high resolution national soil information grids of China. *Science Bulletin*, 2022, **67**(3): 328–340
68. Yu Z, Liu J, Kattel G. Historical nitrogen fertilizer use in China from 1952 to 2018. *Earth System Science Data*, 2022, **14**(11): 5179–5194
69. Nguyen T H, Tang F H M, Conchedda G, Casse L, Obli-Laryea G, Tubiello F N, Maggi F. NPKGRIDS: a global georeferenced dataset of N, P_2O_5, and K_2O fertilizer application rates for 173 crops. *Scientific Data*, 2024, **11**(1): 1179
70. Hu X F, Shi S L, Zhou B R, Ni J A. 1 km monthly dataset of historical and future climate changes over China. *Scientific Data*, 2025, **12**(1): 436
71. Zhang J L, Liu B, Ren S Q, Han W Q, Ding Y X, Peng S Z A. 4 km daily gridded meteorological dataset for China from 2000 to 2020. *Scientific Data*, 2024, **11**(1): 1230