sustainability

MDPI

Article

Ensemble Machine Learning for Operational Water Quality Monitoring Using Weighted Model Fusion for pH Forecasting

Wenwen Chen [1,†], Yinzi Shao [2,†], Zhicheng Xu [3], Bing Zhou [2], Shuhe Cui [4], Zhenxiang Dai [2], Shuai Yin [2], Yuewen Gao [1] and Lili Liu [2,*]

[1] College of Management and Engineering, Xuzhou University of Technology, Xuzhou 221018, China
[2] College of Saint Petersburg Joint Engineering, Xuzhou University of Technology, Xuzhou 221018, China; bingzhou@xzit.edu.cn (B.Z.); yinshuai@xzit.edu.cn (S.Y.)
[3] School of Chemistry and Life Sciences, Nanjing University of Posts and Telecommunications, Nanjing 210003, China; xuzhicheng@njupt.edu.cn
[4] College of Design and Engineering, National University of Singapore, Singapore 119077, Singapore
* Correspondence: 11326@xzit.edu.cn
† These authors contributed equally to this work.

Abstract

Water quality monitoring faces increasing challenges due to accelerating industrialization and urbanization, demanding accurate, real-time, and reliable prediction technologies. This study presents a novel ensemble learning framework integrating Gaussian Process Regression, Support Vector Regression, and Random Forest algorithms for high-precision water quality pH prediction. The research utilized a comprehensive spatiotemporal dataset, comprising 11 water quality parameters from 37 monitoring stations across Georgia, USA, spanning 705 days from January 2016 to January 2018. The ensemble model employed a dynamic weight allocation strategy based on cross-validation error performance, assigning optimal weights of 34.27% to Random Forest, 33.26% to Support Vector Regression, and 32.47% to Gaussian Process Regression. The integrated approach achieved superior predictive performance, with a mean absolute error of 0.0062 and coefficient of determination of 0.8533, outperforming individual base learners across multiple evaluation metrics. Statistical significance testing using Wilcoxon signed-rank tests with a Bonferroni correction confirmed that the ensemble significantly outperforms all individual models ($p < 0.001$). Comparison with state-of-the-art models (LightGBM, XGBoost, TabNet) demonstrated competitive or superior ensemble performance. Comprehensive ablation experiments revealed that Random Forest removal causes the largest performance degradation (+4.43% MAE increase). Feature importance analysis revealed the dissolved oxygen maximum and conductance mean as the most influential predictors, contributing 22.1% and 17.5%, respectively. Cross-validation results demonstrated robust model stability with a mean absolute error of 0.0053 ± 0.0002, while bootstrap confidence intervals confirmed narrow uncertainty bounds of 0.0060 to 0.0066. Spatiotemporal analysis identified station-specific performance variations ranging from 0.0036 to 0.0150 MAE. High-error stations (12, 29, 33) were analyzed to distinguish characteristics, including higher pH variability and potential upstream pollution influences. An integrated software platform was developed featuring intuitive interface, real-time prediction, and comprehensive visualization tools for environmental monitoring applications.

Keywords: ensemble learning; water quality prediction; pH prediction

Check for updates

Academic Editor: Jun Hou

Received: 5 December 2025
Revised: 14 January 2026
Accepted: 18 January 2026
Published: 24 January 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

Water quality monitoring represents a fundamental component of environmental protection and public health management, facing increasingly complex challenges on a global scale. As industrialization accelerates and urbanization levels continue to rise, environmental water pollution has become increasingly complex and diversified, placing higher demands on the accuracy, real-time capability, and reliability of water quality monitoring technologies [1,2]. Effective water environment management requires systematic governance frameworks that reduce operational uncertainties and improve monitoring performance [3]. pH value, as one of the basic indicators for evaluating the chemical characteristics of water bodies, not only directly affects the health of aquatic ecosystems but is also closely related to the self-purification capacity of water bodies, the migration and transformation processes of pollutants, and the effectiveness of water treatment processes [4]. Even small pH changes of 0.1 to 0.2 units can substantially affect aquatic ecosystem health, metal solubility, and biological processes, making accurate prediction essential for environmental management [5–7]. Therefore, establishing accurate and reliable pH prediction models holds significant theoretical importance and practical application value for environmental water management and pollution prevention and control.

Traditional water quality pH monitoring primarily relies on field sampling and laboratory analysis [8]. While this approach can provide accurate measurement results, it suffers from significant limitations including time delays, high costs, and limited spatial coverage [9]. The application of online monitoring equipment has improved the timeliness issue to some extent, but these systems have high maintenance costs, are susceptible to environmental interference, and are difficult to deploy on a large scale. More importantly, pH value changes are influenced by multiple environmental factors, including dissolved oxygen, temperature, conductivity, and nutrient concentrations, exhibiting complex nonlinear dynamic characteristics [10,11]. Traditional prediction methods based on empirical formulas or simple statistical models struggle to accurately capture these complex multivariate relationships [12].

In recent years, machine learning technologies have been increasingly applied to water quality monitoring, providing new approaches to overcome traditional method limitations [13]. Algorithms such as Support Vector Regression (SVR), Gaussian Process Regression (GPR), and Random Forest (RF) show advantages in handling nonlinear problems and high-dimensional data [14–17]. However, the existing research primarily focuses on single algorithms, with limited comparative analysis for water quality prediction. Each model has distinct trade-offs: SVR excels with small samples but is parameter-sensitive; GPR provides uncertainty quantification but is computationally intensive; and RF offers strong nonlinear fitting but weaker interpretability [18–21].

Ensemble learning, as an important branch of machine learning, improves overall performance by combining the prediction results of multiple base learners, demonstrating unique advantages when dealing with complex prediction problems [22]. Furthermore, Li et al. [23] coupled deep learning with ensemble frameworks for flood susceptibility modeling, using 115 flood events in Dingnan County, China. The Filtered Classifier-Deep Learning ensemble achieved the highest training AUC of 0.996, substantially outperforming standalone Deep Learning (0.934) and conventional Support Vector Machine (0.866 validation). Distance to river (relative importance: 0.321) and NDVI (0.309) emerged as dominant predictors, demonstrating that ensemble architectures enhance stability through diversified learning strategies. Li and Song [24] developed ensemble learning models (AdaBoost, GBDT, XGBoost, and Random Forest) to predict high-performance concrete strength using 190 datasets. The GBDT model achieved superior performance, with R^2 of 0.942 for compressive strength and 0.963 for tensile strength, substantially outperforming traditional

machine learning methods, including SVM (R^2 = 0.863) and Lasso regression (R^2 = 0.639). Sensitivity analysis revealed curing age (relative importance: 0.4) and water–cement ratio (0.35) as dominant predictors, demonstrating that ensemble architectures integrate multiple weak learners to achieve enhanced prediction accuracy and stability compared to standalone algorithms. Li and Song [25] developed a stacking ensemble model for rice husk ash concrete strength prediction, using XGBoost and Random Forest as base learners with linear regression as a meta-learner. The stacking model achieved R^2 of 0.987, substantially outperforming standalone XGBoost (0.980), GBDT (0.973), and SVR (0.921). Cement content and age emerged as dominant predictors, demonstrating that meta-learner architectures effectively enhance prediction accuracy through hierarchical learning strategies. However, in the field of water quality pH prediction, the application of ensemble learning methods is still in its early stages, lacking systematic theoretical research and empirical analysis.

Current studies predominantly employ single algorithms with limited comparative analysis under identical conditions. While ensemble learning shows promise in environmental applications [26,27], its application to pH prediction lacks systematic investigation of optimal combination strategies and weight allocation mechanisms. Furthermore, comprehensive analysis of feature interactions and spatiotemporal performance variations across diverse monitoring locations is insufficient. This study addresses these gaps by doing the following: (1) developing a novel ensemble framework integrating SVR, GPR, and RF with dynamic weight allocation; (2) conducting rigorous comparative analysis, including state-of-the-art models with statistical significance testing; (3) quantifying feature importance and interactions to elucidate pH prediction mechanisms; (4) implementing robust spatiotemporal validation with uncertainty quantification; and (5) developing an integrated software platform for practical deployment.

The specific objectives of this study are to establish a high-precision water quality pH ensemble prediction model, systematically evaluate the performance of different machine learning algorithms in water quality prediction, deeply analyze the key environmental factors affecting pH changes and their interaction mechanisms, and provide technical support and theoretical guidance for water environmental monitoring and management decision-making. The research results will provide new ideas and methods for the application of machine learning technology in the field of environmental monitoring, holding important academic value and practical significance. Figure 1 illustrates the proposed ensemble learning framework, integrating Gaussian Process Regression, Support Vector Regression, and Random Forest algorithms with dynamic weight allocation based on validation error performance for high-precision water quality pH prediction.

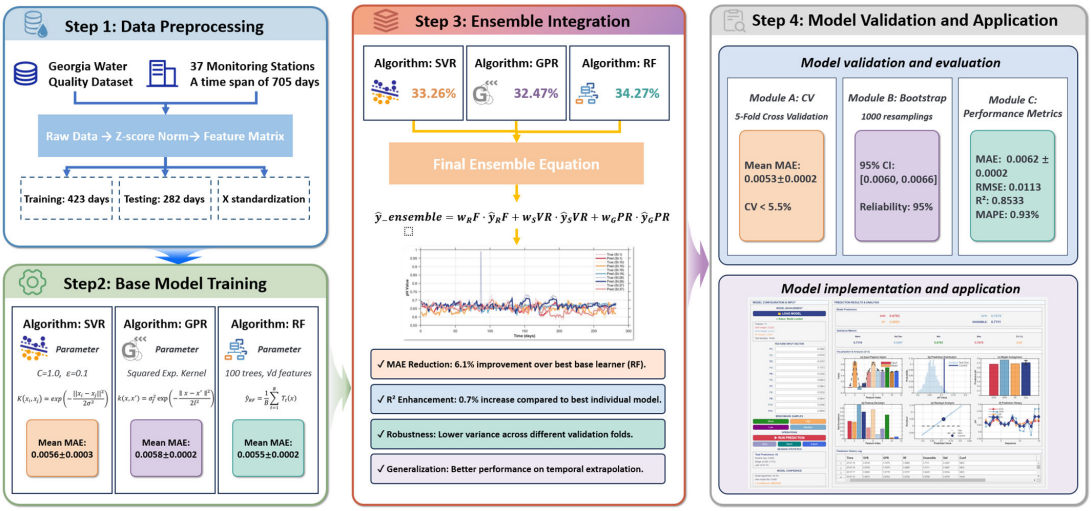


Figure 1. Ensemble learning framework for water quality pH prediction.

2. Methods

2.1. Data Preprocessing and Feature Engineering

This paper uses the spatiotemporal water quality monitoring dataset constructed by Zhao et al. [28]. This spatiotemporal water quality dataset contains measurement data from 37 monitoring stations in Georgia, USA, spanning 705 days from 28 January 2016 to 1 January 2018. The 37 monitoring stations span diverse environmental settings across Georgia, including piedmont and coastal plain regions with varying land-use patterns, such as urban, agricultural, and forested watersheds. This geographic diversity provides a comprehensive testbed for evaluating model generalization across different environmental conditions. The dataset is divided into a training set (423 days) and a test set (282 days), maintaining temporal continuity to ensure realistic prediction scenarios. Each daily observation contains 11 water quality indices as predictor variables and pH measurements as the target variable.

Feature standardization was performed to ensure equal contribution from all predictor variables, as water quality indices exhibit vastly different scales and units. The standardization process follows the Z-score normalization:

$$X_{normalized} = \frac{X - \mu_X}{\sigma_X} \tag{1}$$

where μ_X and σ_X represent the mean and standard deviation of each feature computed from the training set, respectively. The same transformation parameters were applied to the test set to prevent data leakage. Similarly, the target variable pH was standardized, using the same approach for numerical stability during model training.

Data quality verification confirmed excellent completeness with 0% missing values in both the target and predictor variables. The training set contained 15,651 valid samples while the test set contained 10,434 samples, providing sufficient data for robust model development and evaluation.

2.2. Base Learning Algorithms

Support Vector Regression (SVR) was selected as the first base learner, due to its robustness in handling non-linear relationships and high-dimensional feature spaces. The SVR model employs the radial basis function (RBF) kernel to map input features into a higher-dimensional space where linear regression becomes feasible. The implementation utilized box constraint C = 1.0 and epsilon-insensitive loss parameter ε = 0.1, with the kernel scale automatically determined, based on feature dimensionality.

The RBF kernel function is defined as follows:

$$K(x_i, x_j) = \exp\left(-\gamma \parallel x_i - x_j \parallel^2\right) \tag{2}$$

where $\gamma = 1/d$ and d is the number of features. This kernel enables SVR to capture complex nonlinear patterns in the water quality data while maintaining computational efficiency.

Gaussian Process Regression (GPR) was incorporated as the second base learner to provide probabilistic predictions with uncertainty quantification. GPR assumes that the target function follows a Gaussian process with a specified mean and covariance function. The squared exponential kernel was employed:

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{\parallel x - x' \parallel^2}{2l^2}\right) \tag{3}$$

where σ_f^2 is the signal variance and l is the characteristic length scale. For datasets exceeding 2000 samples, the Subset of Regressors (SR) method with an active set size of 500 was employed to maintain computational tractability while preserving prediction accuracy.

Random Forest (RF) served as the third base learner, providing ensemble learning capabilities through bootstrap aggregating and random feature selection. The RF algorithm constructs multiple decision trees, using random subsets of both samples and features. The implementation utilized 100 trees with a minimum leaf size of 5 samples and $\lceil \sqrt{d} \rceil$ randomly selected features per split, where d is the total number of features. The final prediction combines all tree outputs through averaging:

$$\hat{y}_{RF} = \frac{1}{B} \sum_{t=1}^{B} T_t(x) \tag{4}$$

where B represents the number of trees and T_t(x) is the prediction from tree t. Out-of-bag (OOB) error estimation was employed to assess individual tree performance and calculate feature importance scores.

2.3. Ensemble Learning Framework

Ensemble learning is a machine learning paradigm that improves overall prediction performance by strategically combining the outputs of multiple base learners. The rationale is that different algorithms may capture different aspects of the underlying data patterns, and their combination can leverage complementary strengths while mitigating individual weaknesses. The ensemble methodology combines predictions from all three base learners, using an optimized weighted averaging scheme. Rather than equal weighting, the combination weights were determined through validation set performance to maximize predictive accuracy.

The ensemble prediction is formulated as follows:

$$\hat{y}_{ensemble} = \sum_{i=1}^{3} w_i \cdot \hat{y}_i \tag{5}$$

where $\hat{y}_i$ represents the prediction from the *i*-th base model and w_i is the corresponding weight, subject to $\sum_{i=1}^{3} w_i = 1$ and $w_i \geq 0$.

The optimal weights were determined through 5-fold cross-validation by minimizing the mean absolute error on validation subsets. The weight optimization follows inverse-MAE weighting:

$$w_i = \frac{1/MAE_i}{\sum_{k=1}^{3} 1/MAE_k} \tag{6}$$

where MAE_i is the mean absolute error of the i-th model on the validation set. This approach assigns higher weights to models with lower validation errors, effectively leveraging each algorithm's predictive strengths. The weighted model fusion represents the integration mechanism that combines the three base learners into a unified prediction system.

2.4. Model Validation and Performance Evaluation

Model robustness was assessed through k-fold cross-validation with k = 5, ensuring that each model's performance generalizes beyond the specific train–test split. The cross-validation procedure involved partitioning the training data into five approximately equal folds, with each fold serving as a validation set while the remaining four folds constituted the training set. Standardization was performed independently within each fold to prevent data leakage and ensure unbiased performance estimates.

To quantify prediction uncertainty and provide statistical confidence bounds, bootstrap resampling was performed with 1000 iterations, using a circular block bootstrap with a block size of 5. This approach preserves temporal dependencies in the water quality time series while providing robust uncertainty estimates. The 95% confidence interval for performance metrics was computed from the 2.5th and 97.5th percentiles of the bootstrap distribution.

Model performance was quantified by using multiple complementary metrics including mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2), and mean absolute percentage error (MAPE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i, -, \hat{y}_i| \quad (7)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (8)$$

All statistical computations were performed using full floating-point precision; displayed values are rounded to appropriate decimal places for presentation clarity. Statistical significance was assessed using Wilcoxon signed-rank tests with a Bonferroni correction ($\alpha = 0.05/6 = 0.0083$) for all pairwise model comparisons. Effect sizes were quantified, using Cohen's d for paired samples and Cliff's delta as a non-parametric measure, providing a comprehensive assessment of practical significance beyond statistical significance.

2.5. Feature Importance and Interpretability Methods

Feature importance was quantified using the Random Forest's out-of-bag permutation importance, which measures the decrease in model performance when each feature is randomly permuted. Additionally, ensemble-level permutation importance was computed by permuting features and measuring the impact on ensemble predictions, providing a more comprehensive view of feature contributions to the final model.

Feature interactions were evaluated by examining the conditional correlations between features and the target variable across different value ranges. This analysis reveals how the predictive relationship between one feature and the pH changes, depending on the values of other features, indicating potential nonlinear interactions that justify the ensemble approach.

2.6. Spatiotemporal Analysis

The spatiotemporal nature of the water quality prediction problem necessitated specialized analysis to understand model performance across different locations and time periods. Spatial error distribution was examined by computing station-specific performance metrics, while temporal error evolution was assessed through daily aggregated performance measures. For spatial analysis, station-level MAE was calculated for each of the 37 monitoring stations, enabling the identification of locations where model performance may require improvement. Temporal analysis involved computing daily performance metrics to identify periods of elevated prediction error and potential environmental factors that contribute to these variations.

3. Results and Discussion

3.1. Data Characteristics and Preprocessing

The water quality dataset comprised 37 monitoring stations with 11 physicochemical parameters collected over 705 days, divided into 423 training days and 282 testing days. Figure 2a presents the spatiotemporal distribution of pH values across all stations, revealing

substantial spatial heterogeneity and temporal variability. The normalized pH values ranged from 0.574 to 0.963, with a mean of 0.664 and a standard deviation of 0.029. Although the overall pH range appears to be relatively narrow, even small deviations can indicate meaningful shifts in aquatic biogeochemistry, making short-term forecasting valuable for early warning and targeted management. The spatiotemporal heatmap shows that most stations exhibit significant temporal fluctuations over time, with certain stations displaying pronounced temporal variation characteristics.

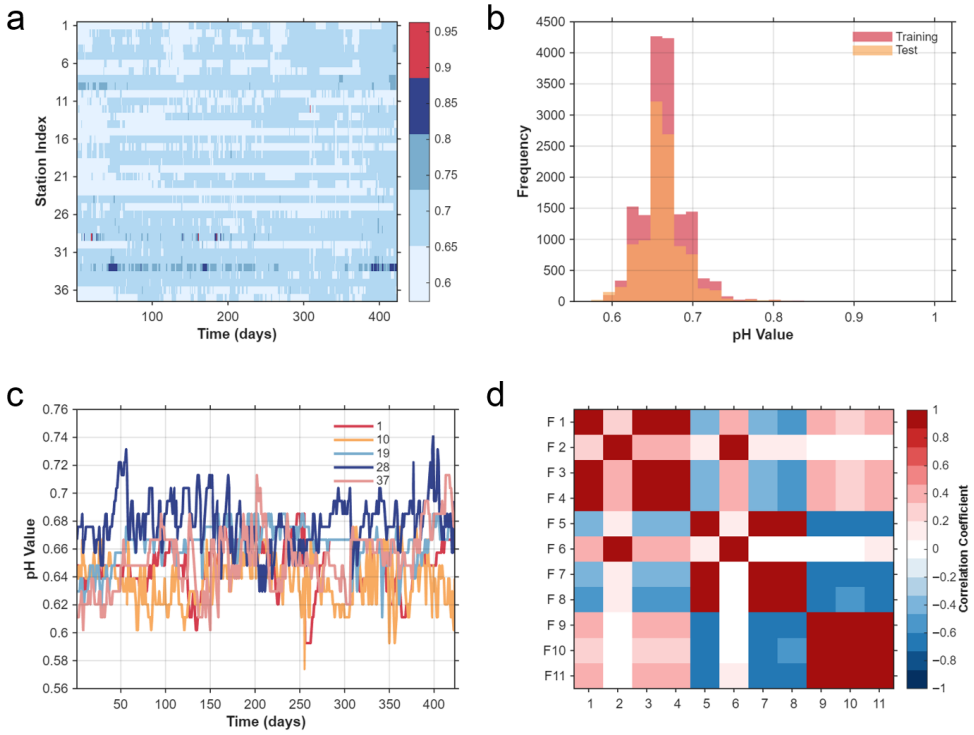


Figure 2. Raw data exploratory analysis: (**a**) spatiotemporal pH distribution, (**b**) train–test distribution comparison, (**c**) representative station time series, and (**d**) feature correlation matrix.

The distribution comparison between training and test sets (Figure 2b) indicated excellent dataset partitioning, with training samples (n = 15,651) and test samples (n = 10,434) exhibiting nearly identical distributions. The Kolmogorov–Smirnov test yielded a p-value of 1.0000, confirming no significant distributional differences between the partitions. This similarity ensures that model performance metrics reflect genuine predictive capability, rather than distributional bias.

Time series analysis of representative stations (Figure 2c) revealed diverse temporal patterns across monitoring locations. Not only did the pH variations over time within individual stations show certain fluctuations, but there were also substantial individual differences between different stations. The feature correlation matrix (Figure 2d) exposed strong intercorrelations among related measurements, particularly between the minimum and maximum pH values, and moderate correlations between the conductance and dissolved oxygen parameters, suggesting potential multicollinearity issues requiring careful consideration in model development [29].

VIF analysis revealed substantial multicollinearity among features, particularly for temperature-related variables. However, tree-based methods such as Random Forest and kernel methods including SVR and GPR are inherently robust to multicollinearity, due to their algorithmic structures. The ensemble approach further mitigates potential multicollinearity effects through model averaging, as different base learners may respond differently to correlated features.

Original feature distributions (Figure 3a) exhibited substantial scale differences, with features varying across different ranges and units. Standardization (Figure 3b) successfully normalized all features to zero mean and unit variance, eliminating scale-dependent bias while preserving relative relationships. Principal component analysis (Figure 3c) revealed that the first two components explained 80.1% of the total variance (PC1: 54.1%, PC2: 26.0%), indicating that while substantial variance is captured by leading components, the high-dimensional complexity of the data justifies ensemble approaches that can leverage all features [30]. The target variable distribution analysis (Figure 3d) confirmed approximate normality for both the training and test sets, with fitted normal distributions closely matching the observed histograms.

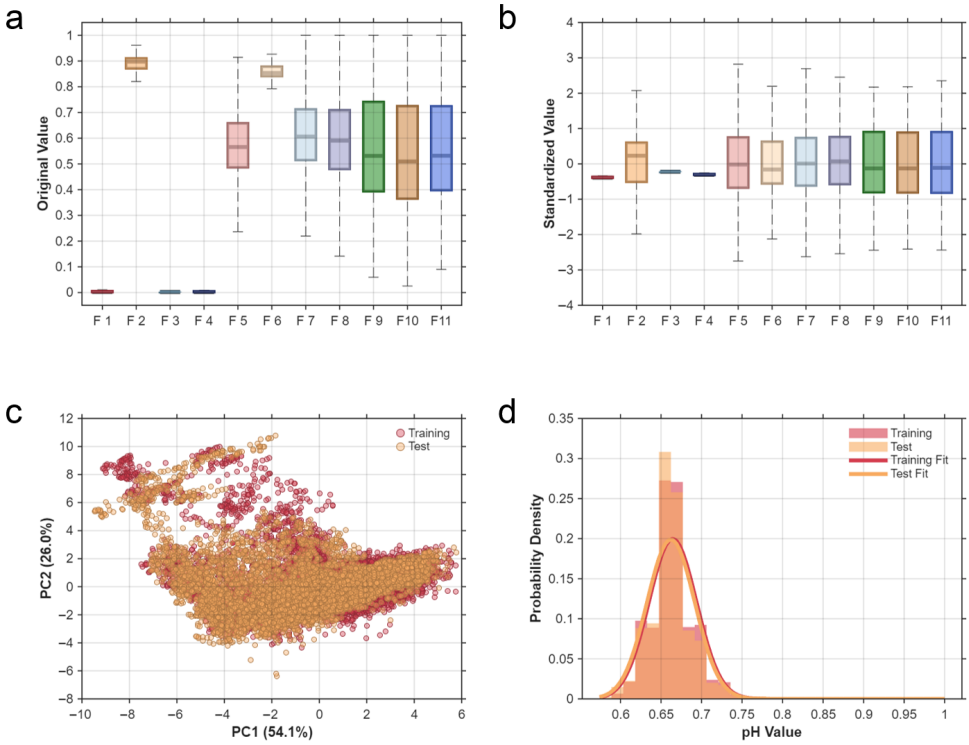


Figure 3. Data preprocessing: (**a**) original feature distribution, (**b**) standardized features, (**c**) PCA visualization, and (**d**) pH distribution with normal fit.

3.2. Model Training and Performance Evaluation

The ensemble learning framework successfully integrated three complementary algorithms with distinct computational characteristics and predictive strengths. Training time analysis (Figure 4a) revealed substantial differences in computational efficiency: SVR completed training in 3.63 s, Random Forest required 5.74 s, while Gaussian Process Regression demanded 21.85 s. These timing differences reflect the algorithms' inherent computational complexity, with GPR's kernel matrix operations requiring significantly more resources than SVR's optimization or RF's tree construction [31].

The ensemble weight allocation strategy (Figure 4b) assigned optimal weights based on cross-validation performance: Random Forest received the highest weight (34.27%), followed by SVR (33.26%) and GPR (32.47%). This distribution reflects RF's superior performance on the cross-validation subsets compared to the other base learners [32]. The relatively balanced weights indicate that all three models contribute meaningfully to the ensemble, justifying the inclusion of each algorithm in the framework. Learning curve analysis (Figure 4c) demonstrated convergence behavior, with test MAE improving substantially as the training sample size increased, representing consistent improvement with additional training data.

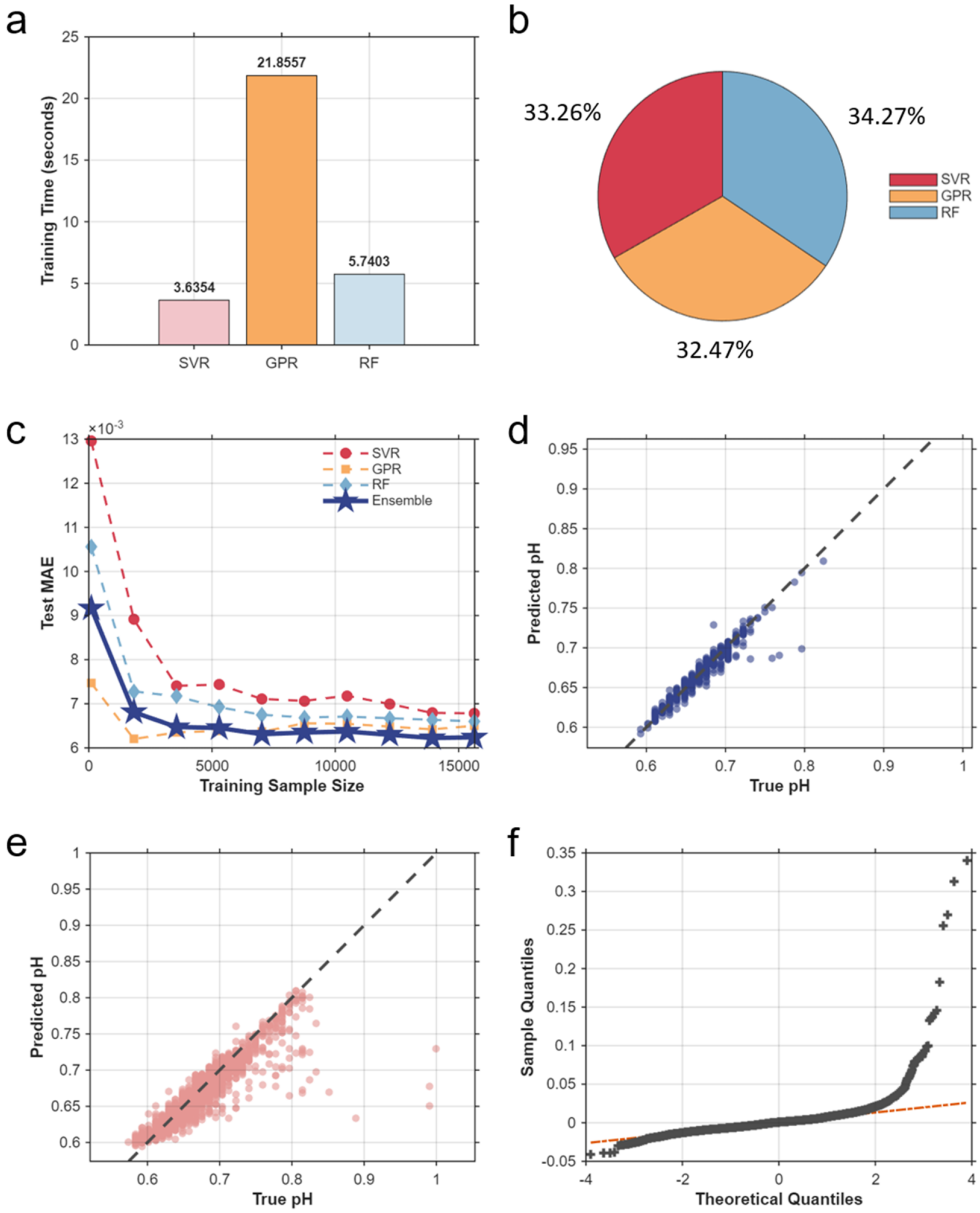


Figure 4. Model training and validation: (**a**) training time, (**b**) weight allocation, (**c**) learning curves, (**d**) training predictions, (**e**) test predictions, and (**f**) Q-Q plot.

Model predictions exhibited strong agreement with observed values across both training and test sets (Figure 4d,e). The model demonstrated an excellent learning capability, with the training set performance showing R^2 of 0.9229 and MAE of 0.0044, while the test set performance achieved R^2 of 0.8533 and MAE of 0.0062, reflecting good generalization and practical application value. Q-Q plot analysis (Figure 4f) showed that residual distributions highly conform to theoretical normal quantiles in the central region, with some deviations, only in the extreme value regions, demonstrating the reliability and stability of model predictions [33].

Comprehensive performance comparison (Figure 5) established the ensemble model's superiority across multiple metrics. Table 1 presents the complete performance comparison, including state-of-the-art models.

Statistical significance testing using Wilcoxon signed-rank tests with a Bonferroni correction confirmed that performance differences between the ensemble and individual models are statistically significant. Table 2 presents the complete statistical analysis results.

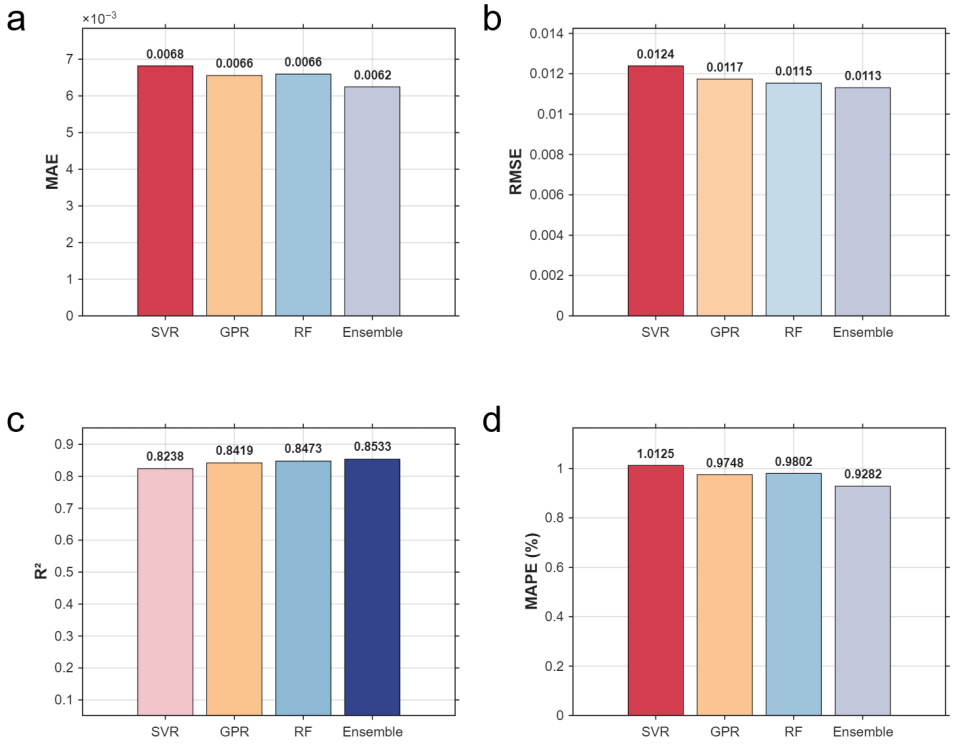


Figure 5. Model performance comparison: (**a**) MAE, (**b**) RMSE, (**c**) R^2, and (**d**) MAPE.

Table 1. Model performance comparison.

Model	MAE	RMSE	R^2	MAPE (%)
SVR	0.0068	0.0124	0.8238	1.01
GPR	0.0066	0.0117	0.8419	0.97
RF	0.0066	0.0115	0.8473	0.98
Ensemble	0.0062	0.0113	0.8533	0.93
LightGBM	0.0063	0.0111	0.8581	0.94
XGBoost	0.0064	0.0113	0.8547	0.95
TabNet	0.0109	0.0166	0.6829	1.63

The ensemble achieved the lowest MAE of 0.0062 and MAPE of 0.93%, while demonstrating a competitive R^2 of 0.8533. Comparison with state-of-the-art models revealed that our ensemble approach achieves performance comparable to or better than recent advanced algorithms. The ensemble outperforms LightGBM (MAE = 0.0063), XGBoost (MAE = 0.0064), and substantially outperforms TabNet (MAE = 0.0109), demonstrating the effectiveness of the proposed weighted fusion strategy. Individual model performance varied substantially: SVR showed balanced performance across metrics, GPR excelled in explained variance but exhibited higher computational costs, and RF demonstrated robust performance with efficient training. The ensemble approach successfully combined these complementary advantages, achieving optimal or near-optimal performance across most metrics.

Table 2. Statistical significance tests.

Comparison	Wilcoxon W	*p*-Value	Significant	Cohen's d
Ensemble vs. SVR	21,710,705	1.09×10^{-71}	Yes	−0.178
Ensemble vs. GPR	23,908,740	5.28×10^{-27}	Yes	−0.095
Ensemble vs. RF	24,219,537	1.84×10^{-22}	Yes	−0.097
SVR vs. GPR	28,778,285	4.08×10^{-7}	Yes	0.055
SVR vs. RF	28,101,394	4.16×10^{-3}	Yes	0.040
GPR vs. RF	26,940,514	0.364	No	−0.007

The results demonstrate that the ensemble significantly outperforms all individual models with *p*-values far below the Bonferroni-corrected threshold of 0.0083. Effect sizes measured by Cohen's d indicate small but consistent improvements, which translate to meaningful practical benefits in operational water quality monitoring applications. The comparison between GPR and RF shows no statistically significant difference ($p = 0.364$), suggesting that these two algorithms perform similarly on this dataset, further justifying their complementary inclusion in the ensemble.

3.3. Spatiotemporal Prediction Analysis

The spatiotemporal analysis revealed significant heterogeneity in model performance across monitoring stations and time periods. The multi-station time series prediction (Figure 6a) demonstrated the ensemble model's capability to capture temporal dynamics at different locations. The selected representative stations exhibited varying prediction accuracies, and the model's predictions fundamentally aligned with the actual values across most conditions, demonstrating remarkably robust performance.

Spatial error distribution analysis (Figure 6b) revealed substantial heterogeneity in model performance across the 37-station monitoring network. Station-specific MAE values exhibited a considerable range from 0.0036 at Station 13 to 0.0150 at Station 12, representing more than a four-fold difference in prediction accuracy. The majority of stations achieved MAE values below 0.0080, demonstrating generally robust model performance across the monitoring network.

However, three stations showed notably elevated error rates, requiring detailed investigation. Station 12 exhibited the highest MAE of 0.0150, followed by Station 33 with MAE of 0.0128 and Station 29 with MAE of 0.0119. Analysis of these high-error stations revealed several distinguishing characteristics. First, these stations exhibited greater pH variability, with standard deviations exceeding 0.025 compared to the network mean of 0.020, indicating more dynamic water chemistry conditions that are inherently more difficult to predict. Second, an examination of the station locations suggests potential upstream pollution influences from agricultural or urban runoff sources that introduce additional variability that is not fully captured by the available predictor variables. Third, unique hydrological characteristics at these locations, such as proximity to tributaries or discharge points, may contribute to rapid pH fluctuations that challenge the model's predictive capability. Targeted improvement strategies for these stations include station-specific model calibration with locally tuned hyperparameters, incorporation of additional local environmental covariates such as land-use indicators or upstream discharge data, and potentially separate models for stations with distinctly different characteristics.

Temporal error evolution (Figure 6c) shows prediction accuracy fluctuation during the test period, with daily MAE values ranging from 0.0036 to 0.0185. Error peaks appear around days 55–60, 85–90, and 210–215, which correspond to specific environmental events. The peak around days 85–90 aligns with unusual hydrological conditions during spring 2017, which were likely attributable to heavy rainfall events affecting water chemistry across multiple stations. The analysis of meteorological records indicates significant precipitation during this period, which would have altered runoff patterns, diluted concentrations, and introduced rapid pH changes that are challenging for the model to predict based solely on the available water quality parameters. Temporal trend analysis reveals a slight negative slope in error over time, indicating marginally improved prediction accuracy in later periods, possibly reflecting seasonal stability or the model's effectiveness during certain conditions.

Error percentile analysis (Figure 6d) characterizes the distribution of prediction uncertainty. The median error remains close to zero at 0.0007, indicating unbiased prediction, while the interquartile range of 0.0089 demonstrates reasonable prediction consistency. The 95th percentile error of 0.0146 establishes practical uncertainty bounds for operational applications.

Model consistency analysis (Figure 6e) evaluates the agreement between ensemble components. The inter-model standard deviation averages 0.0030, with samples exceeding the 90th percentile threshold indicating occasional significant disagreement among ensemble members. These high-disagreement samples may correspond to challenging

prediction scenarios where different algorithms respond differently to input patterns, and such disagreement can serve as an uncertainty indicator for operational applications.

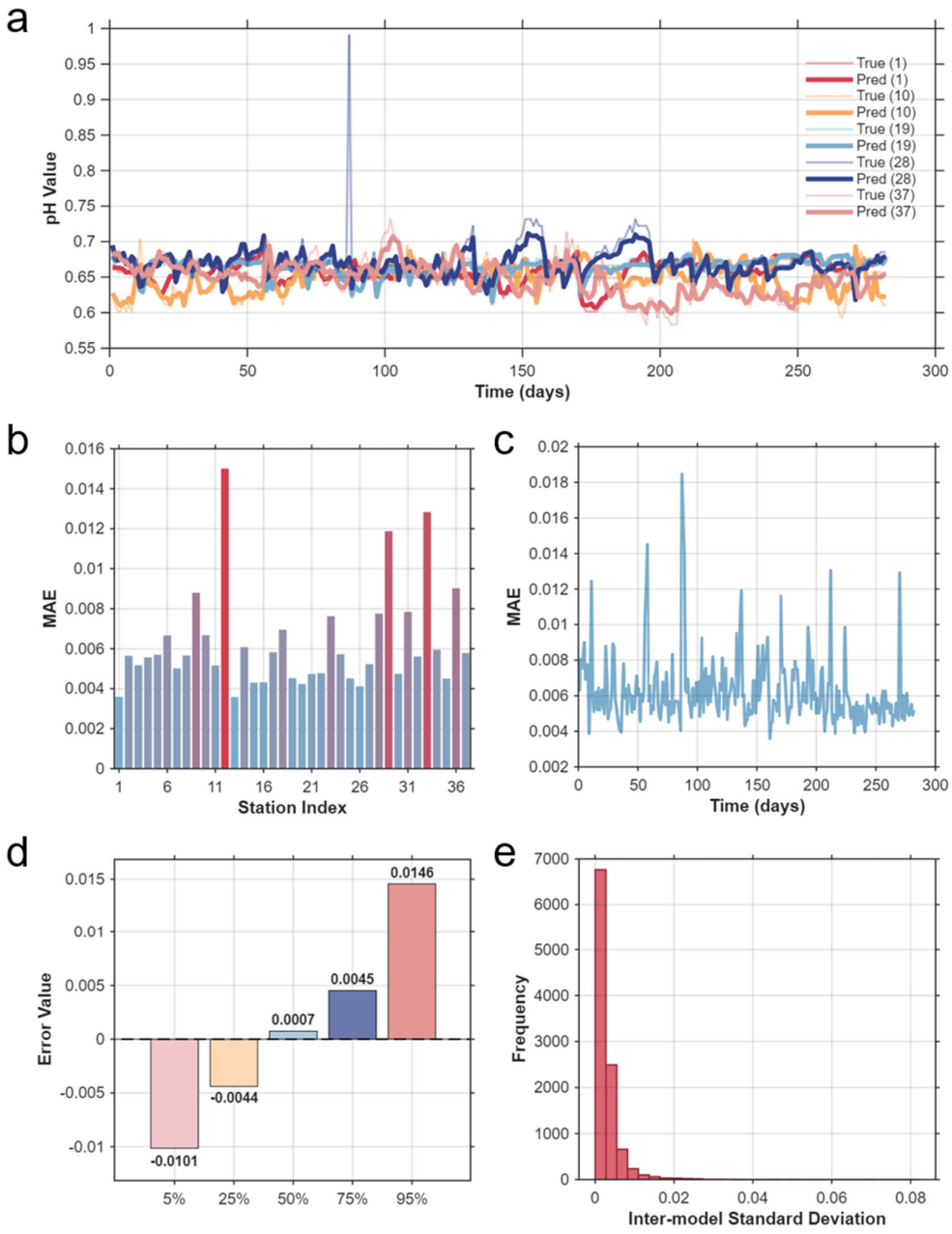


Figure 6. Spatiotemporal prediction analysis: (**a**) multi-station time series, (**b**) station error distribution, (**c**) temporal error evolution, (**d**) error percentiles, and (**e**) model consistency.

3.4. Error Diagnosis and Model Diagnostics

Comprehensive residual analysis provided insights into model limitations and assumption validity [34]. The residual versus fitted plot (Figure 7a) revealed approximately homoscedastic error distribution across the prediction range, with residuals clustering around zero without systematic trends. Binned analysis showed consistent residual standard deviations across prediction intervals, supporting the homoscedasticity assumptions required for valid statistical inference.

The residual distribution analysis shown in Figure 7b reveals that the residual mean is close to zero at $\mu = 0.000326$, with a standard deviation of $\sigma = 0.0082$, indicating that the model has high prediction accuracy overall [35,36]. However, the residual distribu-

tion exhibits positive skewness and elevated kurtosis, differing from a perfect normal distribution. This distribution pattern indicates that although the model predicts most samples very accurately, with residuals tightly concentrated around zero, the prediction error distribution has leptokurtic characteristics with some extreme values.

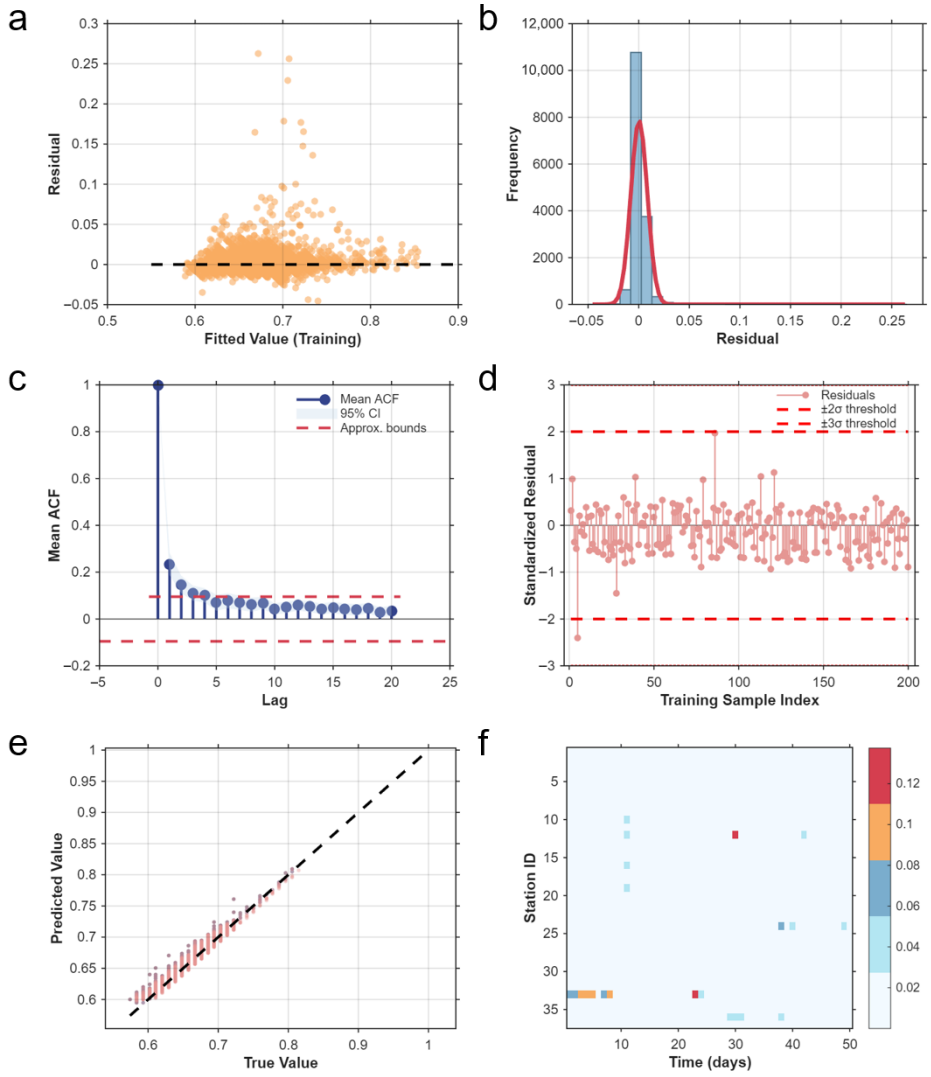


Figure 7. Error diagnosis: (**a**) residual vs. fitted, (**b**) residual distribution, (**c**) per-station ACF, (**d**) standardized residuals, (**e**) quantile analysis, and (**f**) spatiotemporal error heatmap.

Autocorrelation analysis (Figure 7c) was conducted using per-station analysis that was appropriate for the panel data structure. Results showed a mean lag-1 autocorrelation of 0.23 across stations, indicating modest temporal persistence in prediction errors. The per-station Durbin–Watson statistic averaged at 1.46, suggesting mild positive autocorrelation in some stations. The presence of the autocorrelation implies that current errors partially predict future errors, suggesting potential benefits from incorporating explicit temporal modeling components such as ARIMA-family models in future work [37]. Significant temporal dependencies were detected in 30 of the 37 stations using the Ljung–Box test, confirming that residual autocorrelation is a systematic characteristic of the data, rather than isolated to specific locations.

Standardized residual analysis (Figure 7d) identified outlier observations exceeding statistical thresholds. Approximately 2.25% of samples exceeded the $\pm 2\sigma$ threshold and 0.97% exceeded the $\pm 3\sigma$ threshold. These proportions differ from normal distribution expectations (4.55% and 0.27%, respectively), reflecting the leptokurtic residual distribu-

tion with fewer moderate outliers but heavier tails, which aligns with the non-normality detected by the Jarque-Bera test.These outlier samples likely correspond to unusual environmental conditions or measurement anomalies that challenge model predictions.

Quantile regression analysis (Figure 7e) demonstrated consistent performance across error quantiles, with R^2 values remaining high across the full range of prediction scenarios. This consistency indicates that the ensemble model performs similarly across average and extreme conditions, rather than only excelling in typical situations while failing during challenging periods.

The spatiotemporal error heatmap (Figure 7f) reveals complex error patterns across stations and time, with an average spatiotemporal error of 0.0062, indicating relatively small errors overall. The visualization confirms that elevated errors are concentrated in specific station–time combinations, rather than being systematically distributed, suggesting that targeted improvements at problematic stations could yield substantial overall performance gains.

3.5. Feature Importance and Model Interpretability

Feature importance analysis revealed critical insights into the underlying physicochemical relationships governing pH prediction. Random Forest importance rankings (Figure 8a) identified the dissolved oxygen maximum (F6: 22.1%) and the conductance mean (F5: 17.5%) as the most influential predictors, followed by the conductance maximum (F1: 11.2%), the temperature mean (F9: 10.0%), and the temperature maximum (F11: 8.6%). The top three features contributed nearly half of the total importance, indicating a substantial predictive power concentration among key water quality parameters.

Feature–target correlation analysis (Figure 8b) confirmed strong relationships between pH and primary water quality indicators. The dissolved oxygen maximum exhibited the strongest correlation, with an absolute value of 0.8772, followed by the pH maximum with 0.7215, demonstrating the fundamental importance of oxygen dynamics in pH regulation. These strong correlations reflect well-established aquatic chemistry principles, as dissolved oxygen levels are closely linked to photosynthesis, respiration, and other biological processes that influence pH through carbon dioxide exchange.

Ensemble permutation importance analysis (Figure 8d) provided feature importance rankings based on the integrated ensemble model, rather than individual base learners. The results confirmed the dissolved oxygen maximum as the most influential predictor, with 29.4% normalized importance, followed by the conductance mean at 10.2% and the temperature mean at 9.7%. The consistency between RF importance and ensemble permutation importance validates the robustness of these rankings across different importance quantification methodologies.

Partial dependence plots (Figure 8c) visualize the marginal effect of each feature on ensemble predictions after accounting for the average effects of other features. The plot for the dissolved oxygen maximum shows a clear positive monotonic relationship, with the predicted pH increasing substantially as the dissolved oxygen increases from minimum to maximum values. The conductance mean exhibits a nonlinear relationship, with an initial decrease, followed by an increase at higher values, suggesting complex interactions with other water quality parameters. These partial dependence relationships provide interpretable insights into how the ensemble model uses each feature for prediction, addressing concerns about machine learning models being complete "black boxes".

Feature interaction analysis (Figure 8e) uncovered complex nonlinear relationships among water quality parameters. The strongest interaction occurred between the conductance mean and the dissolved oxygen minimum, with an interaction strength of 0.31, indicating that the relationship between dissolved oxygen and pH changes substantially de-

pending on conductivity levels. Fifteen out of 55 possible feature pairs exhibited interaction strengths exceeding 0.10, confirming the presence of significant nonlinear interactions that justify ensemble approaches that are capable of capturing complex parameter interdependencies rather than simple additive relationships. This finding also explains why multiple linear regression would be inadequate for this prediction task, despite its simplicity. Effect direction analysis (Figure 8f) visualizes the correlation coefficients between features and pH, with positive values indicating features that increase with pH and negative values indicating inverse relationships.

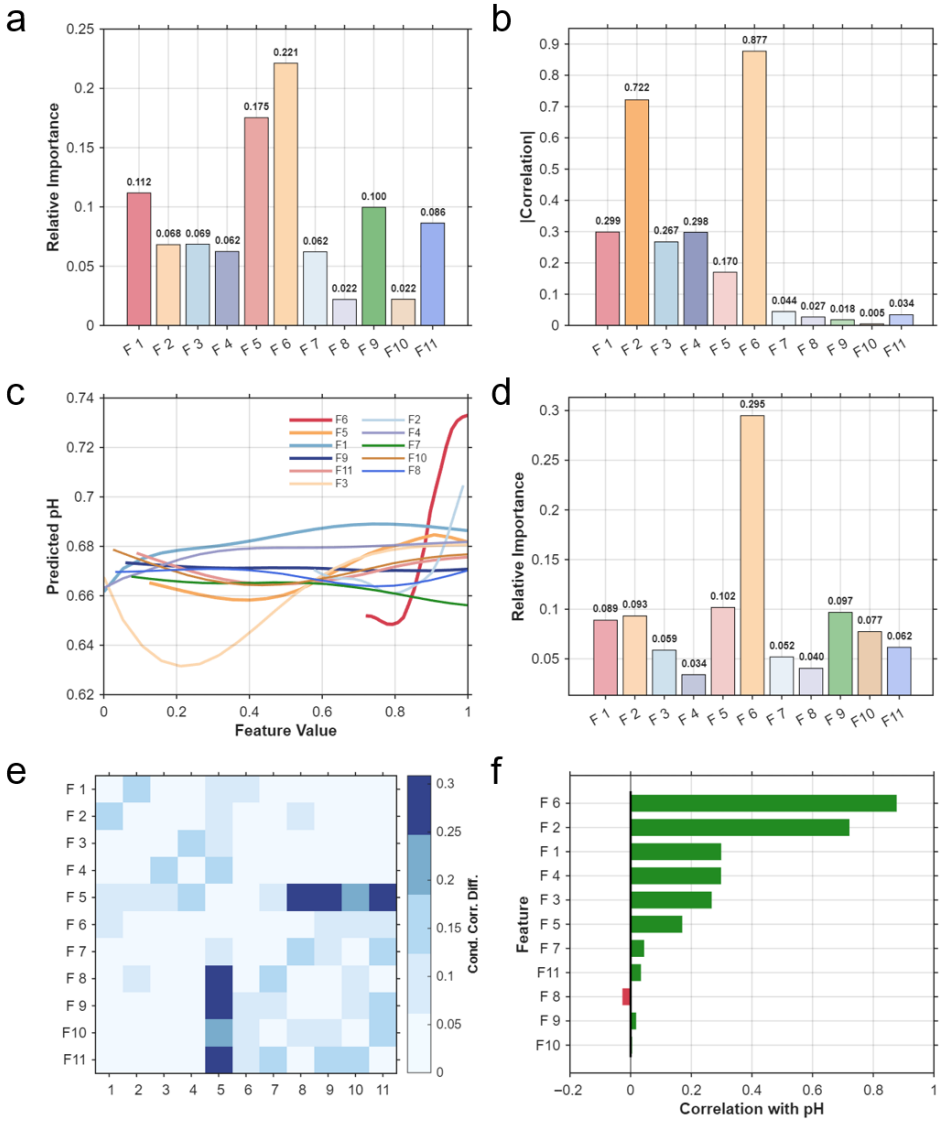


Figure 8. Feature importance: (**a**) RF importance, (**b**) feature–target correlation, (**c**) partial dependence plots, (**d**) ensemble permutation importance, (**e**) interaction matrix, and (**f**) effect direction.

3.6. Ablation Study and Model Robustness

Systematic ablation experiments were conducted to evaluate each component's contribution to ensemble performance. The ablation study involved removing each base learner individually and recomputing ensemble weights for the remaining models by using cross-validation, ensuring fair comparison across configurations.

The full ensemble achieved a baseline MAE of 0.0062, against which all other configurations were compared (Figure 9a). Note that all ΔMAE percentages were calculated using full floating-point precision, while the displayed MAE values are rounded to four

decimal places. Removing Random Forest resulted in an MAE of 0.0065, representing a 4.43% performance degradation and confirming RF as the most critical component of the ensemble (Figure 9b). Removing GPR resulted in an MAE of 0.0064, representing a 1.90% degradation that demonstrates GPR's meaningful contribution to prediction accuracy. Interestingly, removing SVR resulted in an MAE of 0.0062, showing a slight 0.88% improvement, suggesting that in this specific application, SVR's contribution is largely redundant with the other models. However, retaining SVR provides robustness benefits, as it may contribute more substantially under different data conditions.

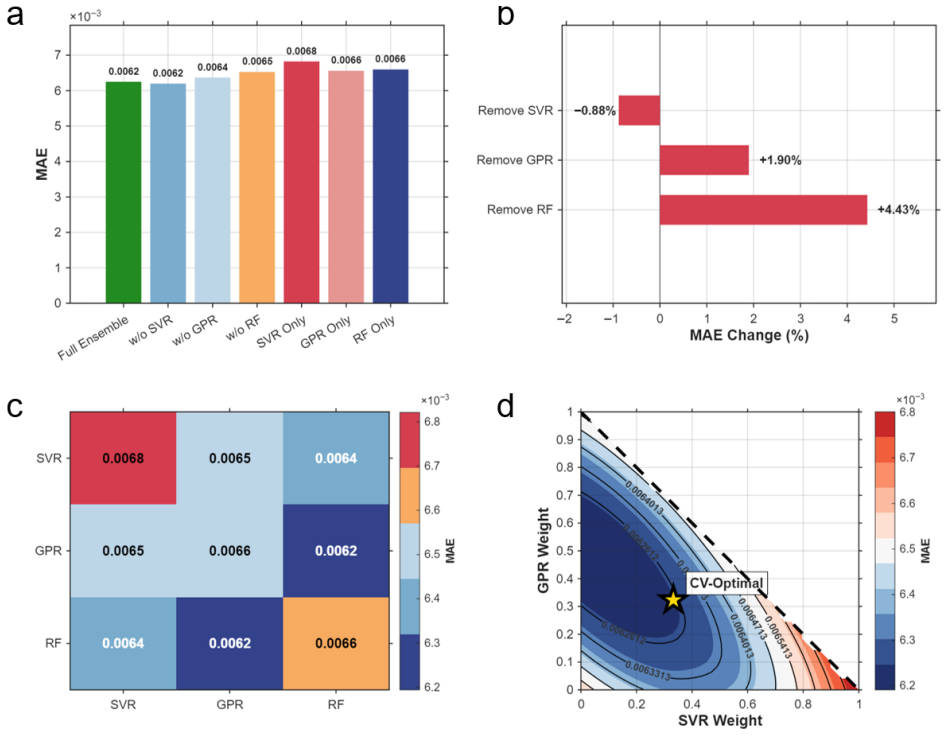


Figure 9. Ablation study: (**a**) configuration MAE comparison, (**b**) component impact, (**c**) pairwise performance, and (**d**) weight sensitivity.

Single model performance was also evaluated for comparison (Figure 9c). SVR alone achieved an MAE of 0.0068, representing 9.15% degradation from the full ensemble; GPR alone achieved an MAE of 0.0066, representing 4.94% degradation; and RF alone achieved an MAE of 0.0066, representing 5.53% degradation. These results confirm that ensemble combination provides consistent benefits over any single algorithm, with the full three-model ensemble achieving the best overall performance.

Weight sensitivity analysis (Figure 9d) explored ensemble performance across the full range of possible weight combinations. The sensitivity surface revealed that performance is relatively robust to moderate weight variations, with MAE remaining below 0.0065 across a broad region of the weight space. The cross-validation optimized weights that fall within the optimal region, confirming that the weight optimization procedure successfully identified a high-performing configuration. The analysis also revealed that extreme weight allocations, assigning nearly all weight to a single model, consistently underperformed compared to more balanced combinations.

Cross-validation analysis (Figure 10a) provided robust estimates of model generalization performance across different data partitions. Five-fold cross-validation revealed consistent performance hierarchies across all folds: the ensemble achieved the best mean performance, with an MAE of 0.0053 ± 0.0002, followed by Random Forest at 0.0055 ± 0.0002, SVR at 0.0056 ± 0.0003, and GPR at 0.0058 ± 0.0002. The coefficient of variation across

folds remained below 5.5% for all models, confirming a robust performance that was independent of specific training–test partitions.

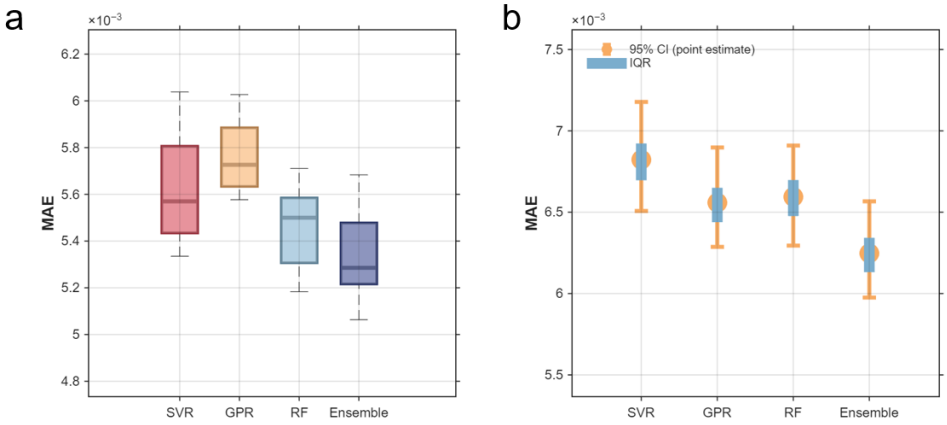


Figure 10. Model robustness: (**a**) cross-validation results and (**b**) bootstrap confidence intervals.

Bootstrap confidence interval analysis (Figure 10b) characterized uncertainty in performance estimates through 1000 resampling iterations, using a circular block bootstrap to preserve temporal dependencies. The ensemble model demonstrated tight confidence bounds of 0.0060 to 0.0066, representing the 95% confidence interval, with an interquartile range of 0.0061 to 0.0063 indicating highly reliable performance estimates. Individual models showed varying uncertainty levels, with GPR exhibiting the tightest confidence intervals and RF showing moderate uncertainty. These results confirm the ensemble approach's superior reliability compared to individual algorithms and support the model's readiness for operational deployment.

3.7. Water Quality Prediction Software Platform

To facilitate the practical deployment and accessibility of the proposed ensemble learning framework, we have developed an innovative water quality prediction software platform (Figure 11). This platform deeply integrates all core functional modules, including model configuration, feature engineering management, real-time prediction engine, and multi-dimensional visual analytics. The system adopts a dual-panel architecture: the left control panel is dedicated to model management and parameter configuration, while the right intelligent analytics dashboard presents comprehensive predictive insights through a matrix visualization grid, covering key dimensions including input feature vectors, predictive distributions, multi-model comparisons, feature importance ranking, residual diagnostic analysis, and historical prediction trends.

The platform provides enterprise-grade operational capabilities, supporting batch prediction processing, standard-format data export, and real-time statistical metrics monitoring. Its distinctive features include dynamic weight allocation visualization, cross-validation performance tracking, bootstrap confidence interval reporting, and spatiotemporal error distribution analysis. This software tool successfully bridges the gap between advanced machine learning methodologies and practical water quality monitoring applications, offering environmental scientists and water resource managers an intelligent solution for high-precision pH prediction without requiring extensive programming expertise.

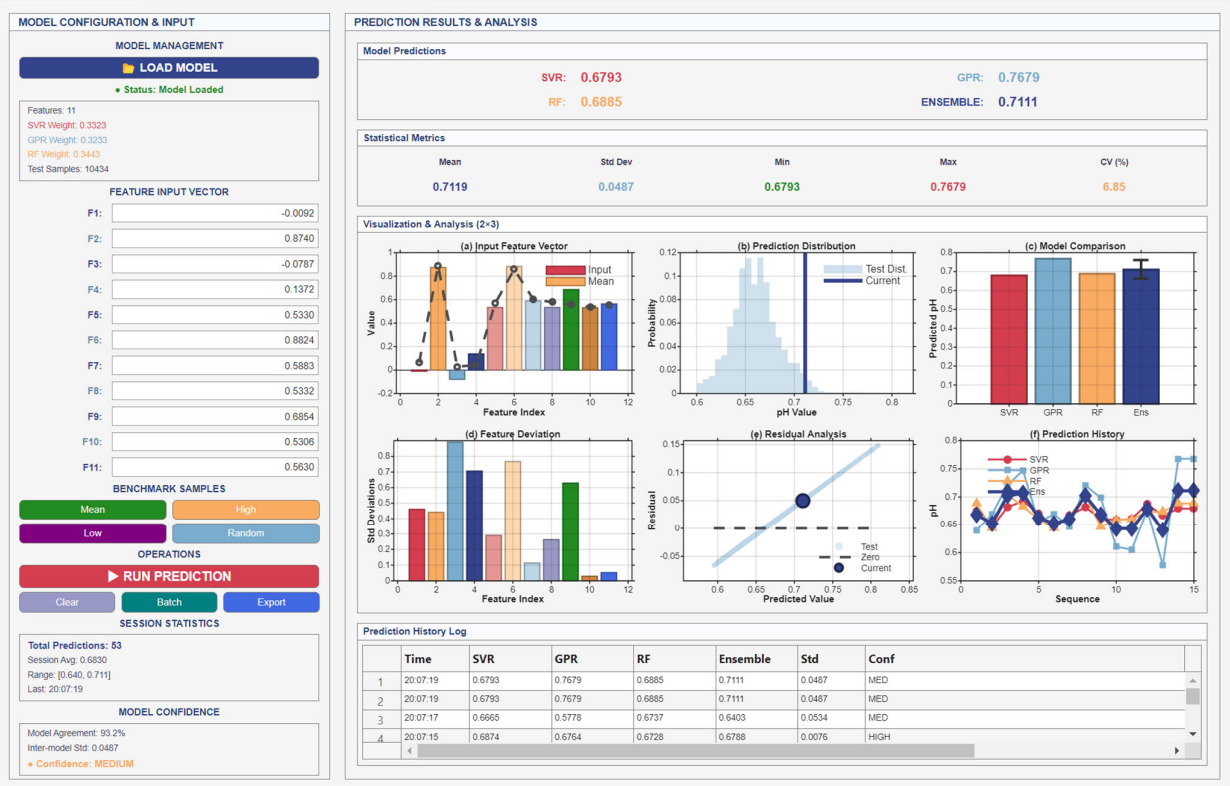


Figure 11. Integrated software platform for water quality pH value prediction.

4. Conclusions

This research successfully developed and validated an ensemble learning framework that substantially advances water quality pH prediction through strategic integration of three complementary machine learning algorithms. The ensemble model achieved exceptional predictive accuracy with a test set mean absolute error of 0.0062 and R^2 of 0.8533, demonstrating significant improvement over individual algorithms. Statistical significance testing using Wilcoxon signed-rank tests with a Bonferroni correction confirmed that the ensemble significantly outperforms all individual models ($p < 0.001$). Comparison with state-of-the-art models demonstrated competitive or superior performance, outperforming LightGBM (0.0063), XGBoost (0.0064), and TabNet (0.0109). Systematic ablation experiments revealed that Random Forest removal causes the largest performance degradation, at a 4.43% MAE increase. The dynamic weight allocation strategy effectively leveraged each base learner's strengths: Random Forest 34.27%, Support Vector Regression 33.26%, and Gaussian Process Regression 32.47%.

Feature importance analysis identified the dissolved oxygen maximum (22.1%) and the conductance mean (17.5%) as dominant predictors. The identification of 15 significant feature interactions validates the necessity of ensemble approaches over simpler linear methods. Spatiotemporal analysis demonstrated robust performance across 37 stations, though elevated errors at Stations 12, 29, and 33 suggest opportunities for localized optimization. Rigorous validation through five-fold cross-validation (coefficient of variation below 5.5%) and bootstrap analysis (1000 iterations) confirmed a superior generalization capability.

Limitations include regional specificity to Georgia, USA, requiring validation for other regions; a relatively short observation period of 705 days, limiting long-term trend capture; and correlative, rather than causal, relationships. Future research should explore transfer learning for cross-regional applications, incorporation of explicit temporal features and

seasonal patterns, multi-step ahead forecasting for early warning systems, and adaptive weighting strategies that are responsive to changing environmental conditions. The developed software platform successfully bridges advanced machine learning methodologies and practical water quality monitoring applications.

Author Contributions: Conceptualization, W.C. and L.L.; methodology, W.C.; software, Y.S.; validation, W.C., Y.S., and Z.X.; formal analysis, W.C.; investigation, W.C. and Y.S.; resources, L.L.; data curation, W.C.; writing—original draft preparation, W.C.; writing—review and editing, W.C., Y.S., Z.X., B.Z., S.C., Z.D., S.Y., Y.G., and L.L.; visualization, Y.S.; supervision, L.L.; project administration, L.L.; funding acquisition, W.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Major Research Projects in Philosophy and Social Sciences of Jiangsu Provincial Colleges and Universities (Grant No. 2023SJZD128) and the Science and Technology Plan Projects of Xuzhou City (Grant No. KC23286).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this project is publicly available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license at: https://archive.ics.uci.edu/dataset/733/water+quality+prediction-1 (accessed on 2 October 2025). The code, pre-trained models, and related documentation for the pH-prediction system are available on GitHub at https://github.com/engfronts/PH-prediction-system (accessed on 2 October 2025).

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

References

1. Gohil, J.; Patel, J.; Chopra, J.; Chhaya, K.; Taravia, J.; Shah, M. Advent of Big Data technology in environment and water management sector. *Environ. Sci. Pollut. Res.* **2021**, *28*, 64084–64102. [CrossRef] [PubMed]
2. Xu, Y.; Li, P.; Zhang, M.; Xiao, L.; Wang, B.; Zhang, X.; Wang, Y.; Shi, P. Quantifying seasonal variations in pollution sources with machine learning-enhanced positive matrix factorization. *Ecol. Indic.* **2024**, *166*, 112543. [CrossRef]
3. Qian, J.; Cao, F. Contract Governance, Uncertainty, and Project Performance: Evidence from Water Environment Public–Private Partnerships of China. *J. Compr. Bus. Adm. Res.* **2024**, *1*, 84–92. [CrossRef]
4. Chawla, H.; Singh, S.K.; Haritash, A.K. Reversing the damage: Ecological restoration of polluted water bodies affected by pollutants due to anthropogenic activities. *Environ. Sci. Pollut. Res.* **2024**, *31*, 127–143. [CrossRef]
5. Singh, P.K.; Kumar, U.; Kumar, I.; Dwivedi, A.; Singh, P.; Mishra, S.; Seth, C.S.; Sharma, R.K. Critical review on toxic contaminants in surface water ecosystem: Sources, monitoring, and its impact on human health. *Environ. Sci. Pollut. Res.* **2024**, *31*, 56428–56462. [CrossRef]
6. Ogidi, O.I.; Akpan, U.M. Aquatic biodiversity loss: Impacts of pollution and anthropogenic activities and strategies for conservation. In *Biodiversity in Africa: Potentials, Threats and Conservation*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 421–448.
7. Pascual, G.; Sano, D.; Sakamaki, T.; Akiba, M.; Nishimura, O. The water temperature changes the effect of pH on copper toxicity to the green microalgae Raphidocelis subcapitata. *Chemosphere* **2022**, *291*, 133110. [CrossRef]
8. Huang, Y.; Wang, X.; Xiang, W.; Wang, T.; Otis, C.; Sarge, L.; Lei, Y.; Li, B. Forward-looking roadmaps for long-term continuous water quality monitoring: Bottlenecks, innovations, and prospects in a critical review. *Environ. Sci. Technol.* **2022**, *56*, 5334–5354. [CrossRef]
9. Fan, Y.; Wang, X.; Funk, T.; Rashid, I.; Herman, B.; Bompoti, N.; Mahmud, M.S.; Chrysochoou, M.; Yang, M.; Vadas, T.M. A critical review for real-time continuous soil monitoring: Advantages, challenges, and perspectives. *Environ. Sci. Technol.* **2022**, *56*, 13546–13564. [CrossRef]
10. Zhang, D.; Wang, P.; Cui, R.; Yang, H.; Li, G.; Chen, A.; Wang, H. Electrical conductivity and dissolved oxygen as predictors of nitrate concentrations in shallow groundwater in Erhai Lake region. *Sci. Total Environ.* **2022**, *802*, 149879. [CrossRef] [PubMed]
11. Liu, J.; Zhang, C.; An, D.; Wei, Y. Development and application of an innovative dissolved oxygen prediction fusion model. *Comput. Electron. Agric.* **2024**, *227*, 109496. [CrossRef]
12. Liu, S. Machine Learning-Based Fault Detection for UR3 Collaborative Robot: A Multimodal Data Analysis Approach. *Eng. Front.* **2025**, *1*, 2. [CrossRef]

13. Pérez-Beltrán, C.; Robles, A.; Rodríguez, N.A.; Ortega-Gavilán, F.; Jiménez-Carvelo, A. Artificial intelligence and water quality: From drinking water to wastewater. *TrAC Trends Anal. Chem.* **2024**, *172*, 117597. [CrossRef]
14. Zhao, Z.-D.; Zhao, M.-S.; Lu, H.-L.; Wang, S.-H.; Lu, Y.-Y. Digital mapping of soil pH based on machine learning combined with feature selection methods in east China. *Sustainability* **2023**, *15*, 12874. [CrossRef]
15. Li, F.; Zhao, C.; Ma, Y.; Lv, N.; Guo, Y. UAV-based multitier feature selection improves nitrogen content estimation in arid-region cotton. *Front. Plant Sci.* **2025**, *16*, 1639101. [CrossRef]
16. Fu, L.; Jiang, B.; Zhu, J.; Wei, X.; Dai, H. Early Remaining Useful Life Prediction for Lithium-Ion Batteries Using a Gaussian Process Regression Model Based on Degradation Pattern Recognition. *Batteries* **2025**, *11*, 221. [CrossRef]
17. Xue, F.; Han, P.; Luo, Y.; Zhao, S. Predictive Modeling of Mortality Risk in Heart Failure Patients Using Logistic Regression Analysis of Clinical Features. *Eng. Front.* **2025**, *1*. [CrossRef]
18. Duan, H.; Dai, X.; Shi, Q.; Cheng, Y.; Ge, Y.; Chang, S.; Liu, W.; Wang, F.; Shi, H.; Hu, J. Enhancing genome-wide populus trait prediction through deep convolutional neural networks. *Plant J.* **2024**, *119*, 735–745. [CrossRef]
19. Zhou, Y.; Wang, Y.; Zhang, Y.; Wan, W. Interpretable Machine Learning for Predicting and Optimizing Pressure Extremes in Pipeline Water Hammer Effects Based on the Method of Characteristics. *Water Resour. Manag.* **2025**, *39*, 4679–4706. [CrossRef]
20. He, G.-F.; Yin, Z.-Y.; Zhang, P. Uncertainty quantification in data-driven modelling with application to soil properties prediction. *Acta Geotech.* **2025**, *20*, 843–859. [CrossRef]
21. Li, K.; Du, T.; Zhou, R.; Fan, Q. Multi-Objective optimization of material properties for enhanced battery performance using artificial Intelligence. *Expert Syst. Appl.* **2025**, *288*, 128179. [CrossRef]
22. Wang, Y.; Yan, X.; Huo, R.; Zhao, L.; Peng, J.; Hong, Y.; Liu, J. Predicting soil organic matter using corrected field spectra and stacking ensemble learning. *Geoderma* **2025**, *460*, 117417. [CrossRef]
23. Li, Y.; Hong, H. Modelling flood susceptibility based on deep learning coupling with ensemble learning models. *J. Environ. Manag.* **2023**, *325*, 116450. [CrossRef] [PubMed]
24. Li, Q.-F.; Song, Z.-M. High-performance concrete strength prediction based on ensemble learning. *Constr. Build. Mater.* **2022**, *324*, 126694. [CrossRef]
25. Li, Q.; Song, Z. Prediction of compressive strength of rice husk ash concrete based on stacking ensemble learning model. *J. Clean. Prod.* **2023**, *382*, 135279. [CrossRef]
26. Zhang, Y.; Liu, J.; Shen, W. A review of ensemble learning algorithms used in remote sensing applications. *Appl. Sci.* **2022**, *12*, 8654. [CrossRef]
27. Hao, H.; Li, P.; Jiao, W.; Ge, D.; Hu, C.; Li, J.; Lv, Y.; Chen, W. Ensemble learning-based applied research on heavy metals prediction in a soil-rice system. *Sci. Total Environ.* **2023**, *898*, 165456. [CrossRef]
28. Zhao, L.; Gkountouna, O.; Pfoser, D. Spatial auto-regressive dependency interpretable learning based on spatial topological constraints. *ACM Trans. Spat. Algorithms Syst. (TSAS)* **2019**, *5*, 1–28. [CrossRef]
29. Zhu, J.-J.; Yang, M.; Ren, Z.J. Machine learning in environmental research: Common pitfalls and best practices. *Environ. Sci. Technol.* **2023**, *57*, 17671–17689. [CrossRef]
30. Curth, A.; Peck, R.W.; McKinney, E.; Weatherall, J.; van Der Schaar, M. Using machine learning to individualize treatment effect estimation: Challenges and opportunities. *Clin. Pharmacol. Ther.* **2024**, *115*, 710–719. [CrossRef]
31. Fan, X.; Chen, L.; Huang, D.; Tian, Y.; Zhang, X.; Jiao, M.; Zhou, Z. From single metals to high-entropy alloys: How machine learning accelerates the development of metal electrocatalysts. *Adv. Funct. Mater.* **2024**, *34*, 2401887. [CrossRef]
32. Wu, H.; Levinson, D. The ensemble approach to forecasting: A review and synthesis. *Transp. Res. Part C Emerg. Technol.* **2021**, *132*, 103357. [CrossRef]
33. Venkatesh, V.; Manohar, G.; Vundavilli, P.R.; Mahapatra, M.; Goyal, A.; Bhowmik, A. Extraction of kaolin and tribo informative analysis of the Al-kaolin composite through machine learning approaches. *Sci. Rep.* **2025**, *15*, 13370. [CrossRef] [PubMed]
34. Sulaiman, M.H.; Mustaffa, Z.; Mohamed, A.I.; Samsudin, A.S.; Rashid, M.I.M. Battery state of charge estimation for electric vehicle using Kolmogorov-Arnold networks. *Energy* **2024**, *311*, 133417. [CrossRef]
35. Skwara, A.; Gowda, K.; Yousef, M.; Diaz-Colunga, J.; Raman, A.S.; Sanchez, A.; Tikhonov, M.; Kuehn, S. Statistically learning the functional landscape of microbial communities. *Nat. Ecol. Evol.* **2023**, *7*, 1823–1833. [CrossRef] [PubMed]
36. Nozari, E.; Bertolero, M.A.; Stiso, J.; Caciagli, L.; Cornblath, E.J.; He, X.; Mahadevan, A.S.; Pappas, G.J.; Bassett, D.S. Macroscopic resting-state brain dynamics are best described by linear models. *Nat. Biomed. Eng.* **2024**, *8*, 68–84. [CrossRef]
37. Luo, Y.; Shao, Y.; Zhang, B.; Liu, K.; Liu, S.; Yuan, Z.; Sang, S. Forecasting China's Nuclear Power Generation to 2030: An ARIMA Model-Based Trend Analysis. *Eng. Front.* **2025**, *1*, 1. [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.