Journal of Environmental Management 392 (2025) 126729

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research article

An automated machine learning-based framework for predicting groundwater quality with sensor data

Jaeuk Youn [a,1], Do Hwan Jeong [b,1], MoonSu Kim [b], Kyong Min Woo [c], Tae Kwon Lee [a,*], Hyun-koo Kim [b,**]

[a] Department of Environmental & Energy Engineering, Yonsei University, Wonju, 26493, Republic of Korea
[b] Soil and Groundwater Division, National Institute of Environmental Research, Incheon, 22689, Republic of Korea
[c] Clean Environment Technology, Incheon, 21315, Republic of Korea

ARTICLE INFO

Handling editor: Lixiao Zhang

Keywords:
Groundwater
Sensor data calibration
Real-time prediction
Machine learning
AutoML

ABSTRACT

Groundwater quality monitoring stands as a critical aspect of groundwater management, necessitating real-time and accurate measurement technologies. In this study, we introduce an automated framework for predicting NH_3–N in groundwater using multiparameter sensor data and machine learning. Data collected from a carcass burial site in Anseong, South Korea underwent rigorous quality control, including outlier detection and calibration against laboratory measurements. We then applied automated machine learning (AutoML) to optimize NH_3–N prediction models using a core set of features including (NH_3–N, electrical conductivity, temperature, and Cl) achieving significant accuracy gains compared to raw sensor outputs. Specifically, R^2 improved from 0.76 to 0.90, while the root mean square error (RMSE) and mean absolute error (MAE) declined from 0.84 to 0.38 and 0.57 to 0.23, respectively. External validation using datasets from two hydrogeologically distinct regions demonstrated that the proposed framework achieved consistently high predictive performance (R^2 = 0.89–0.98; RMSE = 0.008–0.02), underscoring its robustness across diverse contamination scenarios. These findings highlight the effectiveness of combining calibrated sensor data with automated model selection for robust, continuous surveillance Our results underscore the potential for scalable, early detection strategies in sensitive environments, emphasizing how advanced analytics and automated calibration can enhance contamination alerts and support proactive groundwater management.

1. Introduction

Groundwater quality monitoring continues to be a focal point in groundwater management, demanding real-time and accurate measurement technologies. Previous studies have predominantly focused on either laboratory analyses or sensor-based monitoring, each with its distinct limitations. Although precise, laboratory methods often prove impractical for continuous monitoring owing to high operational costs and a lack of temporal continuity (Tiyasha, 2020). Sensor-based systems offer real-time data collection but are susceptible to long-term drift and inaccuracy resulting from environmental parameters (Bakker and Schaars, 2019). Consequently, there is a growing need for robust, integrated frameworks that combine the high precision of laboratory methods with the continuous feedback offered by sensors, particularly in environmentally vulnerable areas such as animal carcass burial sites.

Recent approaches have sought to address these issues by refining sensor data through outlier detection, anomaly correction, and calibration against reference measurements by applying machine learning (ML) algorithms for predictive modeling (Cortés-Ibáñez et al., 2020; Nelson et al., 2021). However, many existing solutions are fragmented, focusing on isolated techniques rather than comprehensive pipelines. The adoption of ML and deep learning techniques in environmental science has considerably impacted the predictive accuracy and timeliness of water quality assessments (Knoll et al., 2019; Zhu et al., 2020), yet traditional ML often require significant expertise in feature selection, model tuning, and validation, making them resource-intensive, and limiting their widespread application (Adombi et al., 2022). Automated ML (AutoML) simplifies these tasks by automating core steps in model

* Corresponding author.
** Corresponding author.
E-mail addresses: tklee@yonsei.ac.kr (T.K. Lee), khk228@korea.kr (H.-k. Kim).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jenvman.2025.126729
Received 13 August 2024; Received in revised form 22 June 2025; Accepted 23 July 2025
Available online 25 July 2025

development (Weng, 2019), and it has already shown promise in applications such as groundwater potential mapping (Bai et al., 2022). Thus, AutoML has served as a transformative tool for real-time monitoring and proactive environmental management (He et al., 2021).

We propose an end-to-end monitoring framework that integrates data preprocessing, rigorous sensor calibration, and AutoML to predict groundwater contaminants in near-real time. We focus on NH_3–N because it serves as a critical indicator of microbial decomposition, posing rapid and localized threats to groundwater quality near carcass burial sites. Monitoring NH_3–N thus provides early detection capabilities for contamination pathways, reinforcing its central role in our modeling framework. Using 12 key parameters from sensors at a carcass burial site, our study demonstrates how combining sensor technology, comprehensive data processing, and AutoML-based optimization yields enhanced accuracy and reliability for groundwater quality assessments. This novel approach addresses the challenges of sensor drift and limited onsite expertise, offering a practical solution for proactive environmental management and contamination alerts.

2. Materials and methods

2.1. Site characterization

This study was conducted in Sanbuk-ri, Iljuk-myeon, Anseong-si, Republic of Korea, which hosts one of the largest carcass burial sites, approximately 20,000 m^2 in size, and was used to dispose of approximately 290,000 chickens during the 2017 avian influenza outbreak. This site experiences an average annual precipitation of 1200 mm and a mean temperature of 12 °C. The specific coordinates of the research site were 37.0856785°–37.0866785° N and 127.5116601°–127.5126601° E, and the five groundwater monitoring wells were positioned within 30 m of each other (Fig. 1).

2.2. Data acquisition

Groundwater quality data were collected from each well using an Aqua TROLL500 multiparameter sonde (In-Situ Inc., USA). The 12 parameters measured included chloride (Cl^-), depth, dissolved oxygen (DO), electrical conductivity (EC), ammonia nitrogen (NH_3-N), nitrate nitrogen (NO_3-N), oxidation-reduction potential (ORP), pH, salinity, total dissolved solids (TDS), temperature, and turbidity. Data were collected from December 2018 to February 2023 at approximately 10 min intervals. Laboratory analyses were performed every 15 days from October 2018 to December 2022 and adhered to the water quality standard methods of the Korea Ministry of Environment (2009). The major anions (Cl^- and NO_3-N) were analyzed using ion chromatography (Dionex ICS-5000, Thermo Fisher, Waltham, MA, United States). NH_3-N determination was performed using a UV spectrometer (UV-1600; Shimadzu, Japan). Turbidity was analyzed using a turbidimeter (HACH 2100N; Hach Company, Loveland, CO, USA). The depth was determined using WL-50 (Ota Shoji, Japan). Finally, DO, EC, ORP, pH, salinity, TDS, and temperature were measured using a YSI ProQuatro Multiparameter meter (YSI Inc., Yellow Springs, OH, USA).

2.3. Framework architecture

The framework integrates manual machine learning and AutoML for data preprocessing, sensor data calibration, and water quality parameter prediction. The framework used in this study is illustrated in Fig. 2. Data preprocessing involved null-value elimination, interval standardization, outlier removal, and missing-value replacement. Calibration with machine learning algorithms was selected for each parameter, and data from all wells were merged and augmented for model development. Prediction models using both manual machine learning and AutoML were developed and evaluated using R (version 4.2.2) and Azure (Microsoft, Seattle, USA).

2.4. Data quality control

2.4.1. Basic quality control and set data interval

The basic quality control discards values beyond the TROLL 500 specifications (Table S1) and zero values. Subsequently, the irregularly collected data were averaged to daily values for consistency. Considering the observed range in groundwater flow rates, which varies between 0.2 and 15.2 m/day, it was considered suitable to evaluate the quality of groundwater at a daily interval.

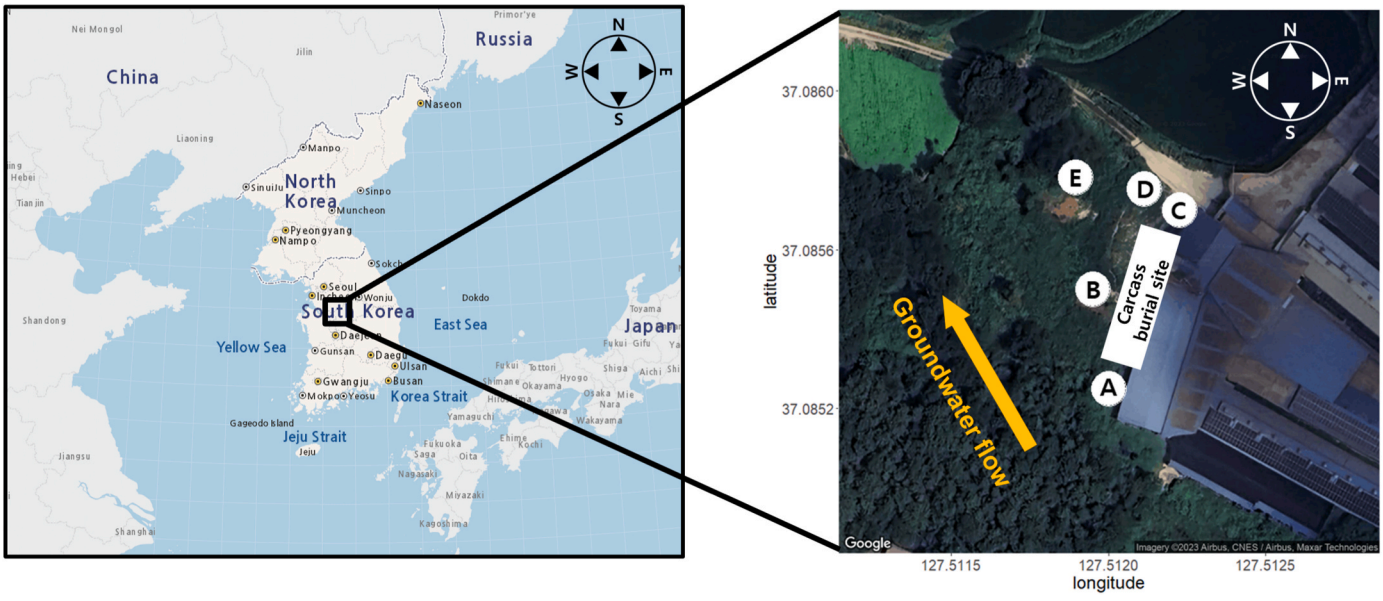

Fig. 1. Location map of groundwater sampling sites in Anseong, South Korea. Map highlighting South Korea with an inset showing the study area's geographic position within the country (left). Detailed aerial view of the carcass burial site in Anseong, indicating the five groundwater wells (labeled A to E) from which samples were collected (right). The direction of groundwater flow is represented by the yellow arrow, providing context for the potential movement of contaminants across the carcass burial site.

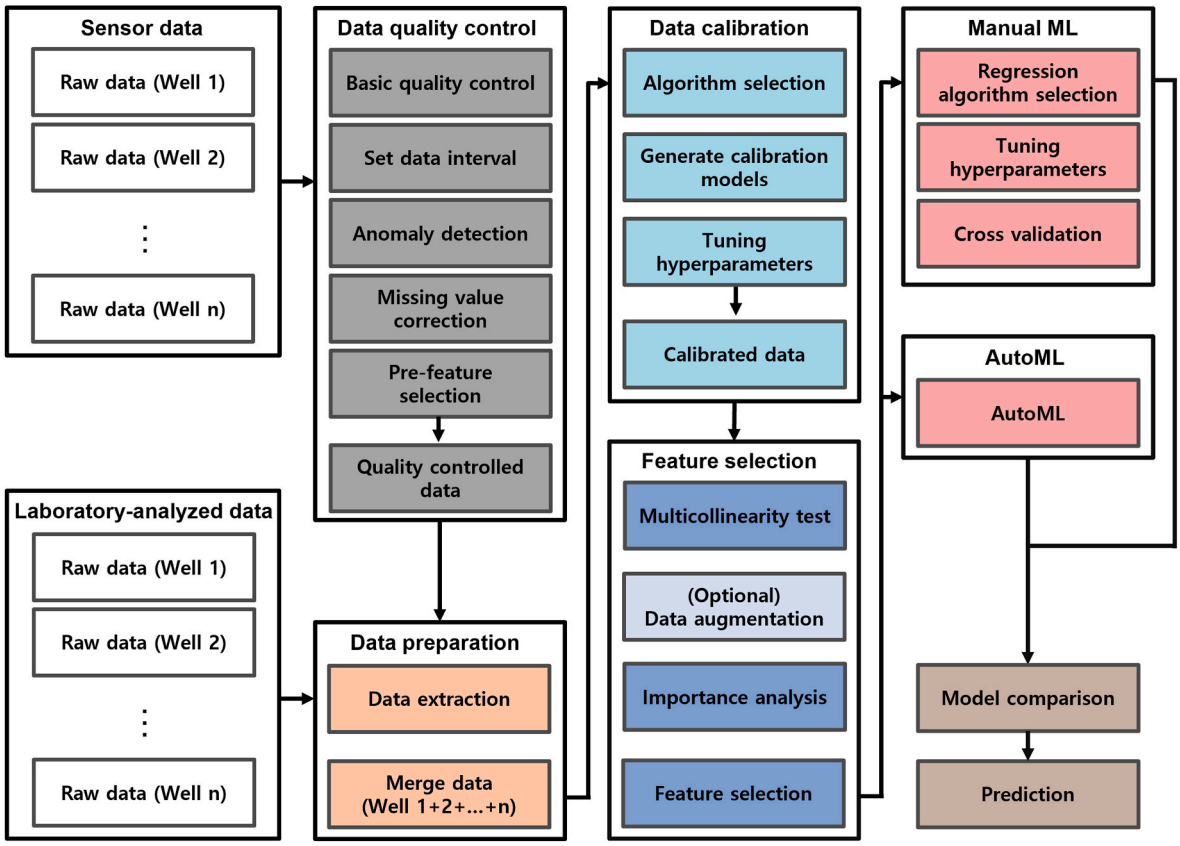

Fig. 2. Schematic overview of AutoML-based framework for groundwater quality prediction. The flowchart illustrates the systematic process from data collection to prediction, starting with sensor and laboratory-analyzed data from multiple wells (Well 1 to Well n). The first column delineates initial raw data collection, which then undergoes a rigorous data quality control process, including anomaly detection and missing value correction, resulting in quality-controlled data. Subsequent data calibration involves algorithm selection and hyperparameter tuning to generate calibrated data. In the third column, manual ML and AutoML approaches are deployed for further refinement and prediction, with feature selection and model comparison being critical steps. The scheme culminates in the prediction phase, highlighting the comprehensive approach integrating manual and automated ML techniques for enhanced predictive accuracy in groundwater quality assessment.

2.4.2. Anomaly detection, missing value correction, and pre-feature selection

Anomaly detection was performed using the tsoutliers package (version 0.6–8) in R, employing autoregressive integrated moving average (ARIMA) models for outlier identification and correction (López-de-Lacalle, 2019). The tsclean function sequentially addresses outliers and missing values using methods such as ARIMA for non-seasonal data and robust seasonal-trend decomposition based on Loess (STL) decomposition for seasonal data, thus refining the time-series data. First, an ARIMA model was fitted to the initial time-series data for each water quality parameter. Then, a *t*-test was conducted on the standardized residuals to detect potential outliers—observations whose residuals exceeded statistically defined thresholds. Identified outliers were replaced with ARIMA-fitted values to smooth sudden spikes or drops that may arise from transient sensor errors or external disturbances. o address missing values, we utilized a two-step approach. For parameters exhibiting distinct seasonal patterns, we applied robust STL after outlier correction. This allowed us to interpolate missing values in a manner consistent with the underlying seasonal structure. For non-seasonal parameters or shorter gaps, linear or spline interpolation methods were employed to estimate missing data points. This approach ensured continuity in the time series, mitigating gaps that could lead to biased model training or hinder the model's ability to capture temporal dynamics. After outlier correction and missing value handling, data standardization was performed. Because the measured groundwater parameters varied greatly in magnitude and units, we applied z-score normalization to each variable. By subtracting the mean and dividing by the standard deviation of each parameter, we placed all variables on a comparable scale. This standardization step prevented any single parameter with a large inherent range from dominating the model training process, thus facilitating more stable hyperparameter tuning and improving overall model convergence. Finally, values from consecutive days showing minimal variance (up to six decimal places) were considered indicative of sensor malfunction and excluded. Additionally, any water quality parameter that exhibited a drift pattern exceeding 50 % over time was removed as an independent variable.

2.5. Data preparation

Data alignment involves matching the sensor and laboratory data by date for the machine learning model input. Merging the data from all the wells addressed the potential biases and enhanced the general applicability of the model. To create a predictive model with general applicability across various wells in the same area despite minor biases and variabilities, instead of focusing on a model with the optimal prediction for a single well, the data from all wells were combined and used for the learning process of the model.

2.6. Data calibration

2.6.1. Regression algorithm selection

The sensor data were calibrated using laboratory-analyzed data as references. The data for each of the 10 water quality parameters (excluding temperature and DO) were divided into training (80 %) and testing (20 %) sets. For temperature, the accuracy of the sensor was

more reliable than that measured by humans and was excluded from the calibration process. DO was excluded because it exhibited significant long-term drift, making it unsuitable for reliable calibration. Eleven machine learning algorithms were implemented using the caret package (version 6.0–93) in R: linear regression, random forest, generalized linear model, Bayesian generalized linear model, boosted generalized linear model, boosted linear model, Gaussian process with linear and radial basis function kernels, support vector machine with linear and radial basis function kernels, and projection pursuit regression. We aimed to capture a diverse range of modeling approaches that have shown effectiveness in environmental time-series prediction. This set includes classical linear regression models, tree-based methods, and kernel-based algorithms. The performance of each algorithm was evaluated based on the root mean square error (RMSE) across 20 iterations. The algorithms were ranked from 1 (lowest RMSE) to 11 (highest RMSE) for each parameter, with the lowest cumulative rank across all parameters determining the optimal calibration algorithm. This approach reduced the complexity of the model and avoided the need to create specialized models for each well. In pursuit of a model capable of universal application across multiple wells within the same region, accepting minor biases and variabilities, the model's learning was based on aggregated data from all wells rather than developing a model optimized for the predictive performance of a singular well.

2.6.2. Optimization of calibration models

Calibration models were optimized by fine-tuning the hyperparameters using a grid search and cross-validation approach. For Support Vector Machine (SVM) models, we fine-tuned the cost (C) and gamma (γ) parameters over a predefined grid, employing 10-fold cross-validation to identify combinations that minimized the RMSE. The cost parameter controlled the trade-off between achieving a low training error and maintaining a smooth decision boundary, while gamma influenced the degree of curvature in the decision boundary. This process was repeated for each water quality parameter to obtain optimal hyperparameters for each calibration model. The final selection of the SVM radial model for calibration, for example, was not only based on its final RMSE values or coefficient of determination (R^2) but also on its stability across multiple validation iterations and its capacity to accurately reproduce laboratory-referenced measurements from sensor inputs. Any systematic biases or deviations were corrected by adjusting the predicted values to align them more closely with the laboratory-analyzed data.

2.7. Feature selection

2.7.1. Multicollinearity test

Multicollinearity analysis was conducted using Spearman's correlation coefficients to evaluate the relationships among the predictor variables. We adopted Spearman's correlation coefficients because our data were partly nonparametric and occasionally exhibited outliers, requiring a rank-based measure less sensitive to extreme values. This approach provided a more robust assessment of monotonic relationships than standard linear correlation measures. Correlation coefficients exceeding 0.8 were considered indicative of significant multicollinearity, warranting adjustments to the model for improved accuracy and interpretability.

2.7.2. Data augmentation

Considering the relatively small size of the dataset (13–35 data points per well), data augmentation was performed using the AugmenterR package (version 0.1.0) in R (da Silva et al., 2021). This process enhanced the quantity and diversity of the dataset and balanced the representation of the data from each well. To avoid creating a model that is only appropriate for wells with large amounts of data, owing to the different amounts of training data for each well, the data points in all wells were adjusted to 50 by augmentation. The GenerateMultipleCandidates function was used to generate additional samples.

2.8. Prediction of NH_3-N in groundwater

2.8.1. Manual ML-based modeling

NH_3-N prediction models were developed using the calibrated water quality parameters. The random forest algorithm was initially used to determine the order of importance of the 11 parameters. We selected this approach because it captures nonlinear interactions among predictors and incorporates out-of-bag error estimates, providing a robust measure of variable importance. The model performance for various combinations of parameters (from the top three to 11 in importance) was compared using algorithms such as SVM radial, Gaussian radial, and random forest, which are suitable for learning nonlinear data implemented using the caret package in R. This iterative process (repeated 20 times) identified the combination with the lowest median RMSE as the most effective combination for NH_3-N prediction. The hyperparameters of the selected model were then optimized using a grid search method to achieve minimal RMSE. The data were divided into training (80 %) and test (20 %) sets. Model performance was evaluated using R^2 and the mean absolute error (MAE).

2.8.2. AutoML-based modeling

Azure AutoML, a cloud-based machine learning service, automates key processes, including algorithm selection and hyperparameter tuning for NH_3-N prediction. The model was developed in parallel with the manual ML process, with optimal model selection based on the RMSE, MAE, and R^2 metrics. Six algorithms (Decision Tree, Extreme Random Trees, Random Forest, Elasticnet, Stack Ensemble, and Voting Ensemble) and four scalers (MaxABsScaler, StandardScalerWrapper, MinMaxScaler, and RobustScaler) were used to construct the model. The model was created by combining one of the algorithms, the scaler. The learning time was set to 120 min, and the RMSE threshold was set to 0.08. A Monte Carlo cross-validation test was used to validate the final generated model.

2.8.3. External validation using independent groundwater datasets

To evaluate the broader applicability of the developed AutoML-based NH_3–N prediction model, external validation was conducted using groundwater datasets from two independent regions: Hoengseong in Gangwon and Uijeongbu in Gyeonggi, both located in military base area. These regions exhibit distinct hydrogeochemical characteristics and contamination profiles. Hoengseong is primarily affected by nitrate (NO_3^-–N, up to 19.2 mg/L) and total petroleum hydrocarbons (TPH, up to 11.4 mg/L), while Uijeongbu is mainly contaminated by TPH (up to 32.8 mg/L). Sensor data from each site were subjected to the same preprocessing, calibration, and feature selection procedures applied to the Anseong dataset. The calibrated and augmented data were then used to generate NH_3–N predictions using the optimized AutoML model developed in the Anseong case. Model performance was assessed using RMSE, MAE, and R^2 to confirm robustness across heterogeneous environmental settings.

3. Results

3.1. Sensor data anomalies

In the analysis of the sensor data from five groundwater wells (Fig. S1), significant variations and anomalies were observed in the 11 monitored water quality parameters, emphasizing the need for the developed framework. The data exhibited notable temporal variations, which were indicative of environmental influences on groundwater dynamics and mechanical errors. The presence of noise, data gaps, and potential sensor drift in measurements of DO, turbidity, pH, salinity, and NH_3-N were observed in all wells. These data irregularities, along with

the complexity of multiparametric interactions, highlight the imperative need for a sophisticated machine learning-based framework.

3.2. Data volume reduction and integrity enhancement through rigorous quality control

The initial acquisition of groundwater sensor data across five different wells provided a comprehensive dataset with a substantial number of data points, ranging from 75,329 to 277,123 per well (Table S2). The collection period covered various start and end dates, beginning as early as November 27, 2018, and extending up to March 16, 2023. After the application of data quality control measures, which included basic quality checks and the integration of data into a consistent daily format, a significant reduction in data volume was observed. For example, Well A originally had 75,329 data points, which were reduced to 545 data points covering the same initial period after the application of data quality control. This represents data retention of approximately 0.72 % of the original volume. Similarly, for Well B, of the 277,123 initial data points, only 1330 were retained after quality control, corresponding to a retention rate of approximately 0.48 %. Wells C, D, and E followed the same trend with data retention rates of approximately 0.69 %, 0.67 %, and 0.71 %, respectively. The dramatic reduction in data points indicates a rigorous quality control process that effectively filters out erroneous, inconsistent, or irrelevant data that could potentially skew the predictive modeling process. The retained data represent the most reliable and accurate measurements necessary to create a robust predictive model. This transition from raw to quality-controlled data underscores the importance of stringent data validation practices in environmental monitoring applications, particularly when employing machine-learning techniques for predictive analysis.

Examination of the raw data revealed a broad range of values for each parameter, with notable quantities of missing values (NA's) for several indicators such as Cl^- and NH_3-N. The quality control process, which presumably included the removal of outliers and correction of erroneous readings (zero, missing values, irregular data intervals, and observations outside the sensor limits), resulted in a more constrained range of values for each parameter, with a complete absence of missing values, indicating a refined dataset (Table 1). In the context of Cl^- concentration, the quality control procedure notably reduced the maximum value from 3.80×10^5 mg/L to 1.00×10^3 mg/L, thereby significantly narrowing the range and aligning the mean values closer to the median, suggesting a more symmetric data distribution. For parameters such as DO and ORP, the raw data exhibited excessively high maximum values, which were rectified, leading to more plausible ranges and median values closer to the mean, implying a reduced skew in the data distribution. The temperature parameter showed minimal changes in its range and central tendency measures, suggesting that the raw temperature data were relatively accurate or less affected by erroneous readings. DO was excluded from the NH_3-N predictive independent variable because it showed long-term drift (up to six decimal places) for more than 50 % of the total time. This process evidently improved data integrity, as observed by the reduction in extreme values and the alignment of median and mean values, thereby enhancing the precision and utility of the dataset for further ecological and environmental analyses.

3.3. A comparative analysis of ML algorithms for sensor data calibration

Machine learning algorithms were employed to develop a calibratable regression model for multiple groundwater quality parameters. This study assessed the performance of 11 distinct algorithms, including linear models such as linear models (Lm) and generalized linear models (Glm) and more complex approaches such as boosted models, random forest (Rf), and support vector machines (SVMs). The algorithms were ranked based on their RMSE values for each water-quality parameter (Table 2). The key findings indicate that the SVM radial basis function (SVM radial) algorithm exhibited the most consistent performance across different water quality indicators, as evidenced by its lowest cumulative rank score of 46. This consistent performance suggests that the SVM radial model has substantial potential for calibrating water-quality sensor data across various measurements. Furthermore, the results of this study indicate that it may be feasible to standardize the calibration process across different water quality indicators using a single machine-learning algorithm. This approach can streamline the calibration framework and potentially reduce computational complexity and resource utilization.

Data quality improvement for the sensor data was assessed by comparing the RMSE values following data quality control and additional calibration with those achieved by data quality control alone (Table 3). After examination of the RMSE values for quality-controlled data, the parameters exhibited a range of variability, with EC (307.3 μS/cm) and NH_3-N (487.3 mg/L), demonstrating the highest RMSE values, indicating less precision in the quality-controlled sensor data for these parameters. In contrast, parameters such as pH and salinity showed considerably lower RMSE values (0.8 and 0.10, respectively), suggesting a higher precision in the sensor measurements before calibration. After calibration, the RMSE values for all parameters showed notable improvements. The most notable enhancement was observed for NH_3-N, which demonstrated a substantial decrease in the RMSE from 487.3 to 0.9 mg/L, indicating a significant improvement of 486.4 in the effectiveness of the calibration process. Similarly, EC showed a notable reduction in RMSE from 307.3 to 116.2 μS/cm, corresponding to an improvement of 191.1. These improvements suggest that the calibration process notably increased the accuracy of the sensor data for these

Table 1
Statistical summary of water quality parameters before and after data quality control. A comparison of the statistical metrics for various water quality parameters measured by sensors, detailing the changes from raw data to data post-quality control. For each parameter, the minimum, maximum, median, and mean values are listed alongside the count of missing values (NA's).

Parameter	Unit	Raw data					Quality controlled				
		Minimum	Maximum	Median	Mean	NA's	Minimum	Maximum	Median	Mean	NA's
Cl^-	mg/L	−4.00E+00	3.80E+05	6.03E+01	5.26E+02	691	3.75E+00	1.00E+03	5.97E+01	1.36E+02	–
Depth	m	0.00E+00	8.28E+01	6.10E+00	7.00E+00	–	4.08E+00	1.70E+01	6.29E+00	7.17E+00	–
DO	mg/L	−9.61E+21	4.13E+31	0.00E+00	4.77E+25	803	–	–	–	–	–
EC	μS/cm	0.00E+00	1.18E+07	6.29E+02	1.26E+04	95	1.91E+02	3.48E+02	2.97E+02	2.88E+02	–
NH_3-N	mg/L	−4.90E+00	1.51E+04	6.37E-01	4.37E+02	2289	3.90E-03	2.46E+03	8.11E-01	1.32E+02	–
NO_3-N	mg/L	−2.67E+01	5.00E+04	9.93E+00	8.84E+01	1986	2.35E-01	3.17E+02	1.35E+01	3.45E+01	–
ORP	mV	−7.17E+02	9.55E+02	5.95E+01	6.49E+01	566	−3.62E+02	5.60E+02	1.61E+02	1.32E+02	–
pH	–	0.00E+00	3.14E+01	6.48E+00	6.85E+00	241	5.47E+00	1.01E+01	6.37E+00	6.56E+00	–
Salinity	–	−8.00E-03	1.23E+03	3.80E-01	4.83E-01	293	1.00E-03	6.73E-01	3.08E-01	3.37E-01	–
TDS	ppt	−8.20E+01	5.42E+06	1.00E+00	9.19E+03	495	1.00E-03	8.68E-01	3.96E-01	4.42E-01	–
Temperature	°C	−1.79E+01	1.46E+02	1.40E+01	1.45E+01	7	1.28E+01	1.78E+01	1.40E+01	1.43E+01	–
Turbidity	NTU	−1.09E+02	2.76E+04	1.08E+01	1.65E+02	597	7.91E-02	2.92E+02	1.34E+01	3.50E+01	–

Table 2
Ranking of machine learning algorithms for water quality parameter calibration. An algorithm ranking analysis based on the median of the root mean square error (RMSE) for calibration of various water quality parameters. Each algorithm is assessed across multiple parameters and ranked from lowest (best) to highest RMSE. The sum of ranks for each algorithm across all parameters is computed, resulting in a total rank that guides the selection of the most effective calibration algorithm for the dataset. Algorithms listed include Linear Model (Lm), Random Forest (Rf), Generalized Linear Model (Glm), Bayesian Glm, Boosted Glm, Boosted Linear Model (Lm), Gaussian Linear, Gaussian Radial, Support Vector Machine (SVM) Linear, SVM Radial, and Projection Pursuit Regression (PPR).

Algorithm	Rank (Median of RMSE)										Sum	Total Rank
	Cl^-	Depth	EC	NH_3-N	NO_3-N	ORP	pH	Salinity	TDS	Turbidity		
Lm	4 (16.12)	8 (2.01)	4 (438.51)	7 (1.24)	6 (16.57)	1 (83.07)	7 (0.43)	2 (0.08)	2 (337.72)	7 (87.03)	48	2
Rf	11 (18.09)	3 (1.56)	11 (527.47)	1 (1.13)	3 (14.79)	11 (95.68)	11 (0.49)	10 (0.09)	11 (424.23)	11 (96.84)	83	11
Glm	3 (16.04)	9 (2.02)	6 (456.08)	6 (1.23)	10 (17.24)	5 (87.41)	8 (0.43)	4 (0.08)	6 (347.24)	6 (86.84)	63	8
Bayesian glm	2 (16.01)	7 (2.01)	9 (474.17)	5 (1.21)	9 (17.02)	7 (88.33)	3 (0.42)	6 (0.08)	5 (344.95)	5 (86.64)	58	6
Boosted glm	5 (16.18)	10 (2.03)	7 (463.85)	3 (1.17)	7 (16.65)	6 (87.64)	6 (0.43)	1 (0.08)	7 (364.04)	2 (83.93)	54	3
Boosted lm	8 (16.92)	5 (2.00)	2 (407.42)	9 (1.25)	5 (15.90)	10 (92.98)	4 (0.43)	11 (0.11)	9 (380.90)	9 (89.12)	72	10
Gaussian Linear	1 (15.87)	6 (2.00)	3 (424.68)	10 (1.25)	8 (16.77)	3 (84.96)	9 (0.43)	5 (0.08)	4 (343.20)	8 (87.07)	57	5
Gaussian Radial	10 (17.71)	2 (1.53)	10 (490.32)	4 (1.19)	1 (14.02)	8 (89.91)	2 (0.42)	9 (0.09)	8 (380.79)	4 (85.82)	58	6
SVM Linear	6 (16.25)	11 (2.43)	5 (447.11)	8 (1.24)	11 (19.20)	2 (84.82)	5 (0.43)	3 (0.08)	1 (299.11)	3 (84.55)	55	4
SVM Radial	9 (16.94)	4 (1.73)	1 (391.48)	11 (1.26)	4 (14.87)	4 (86.68)	1 (0.41)	8 (0.09)	3 (338.46)	1 (79.25)	46	1
PPR	7 (16.32)	1 (1.51)	8 (467.43)	2 (1.15)	2 (14.59)	9 (89.93)	10 (0.44)	7 (0.08)	10 (399.67)	10 (89.35)	66	9

Table 3
Improvements in RMSE values through ML-based calibration of water quality parameters. The RMSE values for various water quality parameters both before and after the application of machine learning-based calibration. For each parameter, the improvement in RMSE following calibration is quantified. The reduction in RMSE illustrates the effectiveness of the calibration process in refining sensor data accuracy.

Parameter	Unit	RMSE		Improvement
		Quality controlled data	Calibrated data	
Cl^-	mg/L	105.0	13.3	91.7
Depth	m	2.7	0.7	2.0
EC	µS/cm	307.3	116.2	191.1
NH_3-N	mg/L	487.3	0.9	486.4
NO_3-N	mg/L	16.7	3.6	13.1
ORP	mV	215.2	43.5	171.7
pH	–	0.8	0.3	0.5
Salinity	–	0.10	0.06	0.04
TDS	ppt	393.8	73.1	320.7
Turbidity	NTU	89.6	45.6	44.1

parameters. Salinity showed the least improvement, with a minimal decrease in RMSE from 0.10 to 0.06, an improvement of 0.04. Despite the smaller magnitude of improvement, this change remains indicative of the enhanced sensor accuracy post-calibration. Turbidity showed a significant decrease in RMSE from 89.6 NTU to 45.6 NTU, demonstrating an improvement of 44.1. Overall, the calibration process resulted in a considerable enhancement in the accuracy of the sensor measurements for all groundwater quality parameters.

Analysis of the sensor data revealed significant differences in the groundwater quality values across the three processing stages: raw data, quality-controlled data, and calibrated data (Fig. S2). The raw data exhibited a notable presence of anomalies characterized by erratic fluctuations and a lack of consistent trends. In contrast, the data that underwent quality control processing demonstrated a substantial reduction in these anomalies, presenting a more stable and coherent trend. Pivotal transformation was observed when the data were subjected to additional calibration. The data quality with the calibration process significantly enhanced the data fidelity, aligning it closely with the trends exhibited by the laboratory-analyzed data. This alignment suggests that the calibration process plays a critical role in aligning sensor-derived data with standard laboratory measurements. For example, the salinity levels were noticeably more erratic in the raw data but became more uniform and consistent post-calibration, closely mirroring the laboratory data. Similarly, nitrate values, which were highly variable in the raw and quality control-processed data, exhibited significant convergence towards the laboratory data trends following calibration. These findings underscore the importance of data quality control and calibration to ensure the reliability and accuracy of sensor-derived environmental data.

3.4. Data augmentation and multicollinearity analysis

After data augmentation, the distribution remained similar to that of the original data; however, the quantity increased. We adopted a strategy that equalized the number of data points per well to 50. This approach prevented uniform augmentation across all concentration ranges, effectively mitigating the bias in the original data. The augmentation process increased the total number of data points as well as varied the augmentation rate according to the concentration range of the water quality variables (Fig. S3). This enabled the development of a more versatile model that is not biased towards specific concentrations and plays a crucial role.

A thorough evaluation of multicollinearity was conducted to select features and algorithms for the NH_3-N prediction model. This step was crucial for determining the most suitable water quality parameters for feature selection. Our results identified significant correlations among several independent variables. Notably, a correlation coefficient greater than 0.8 was observed between NO_3-N and depth, as well as between salinity and EC. Considering these high correlations, which can potentially skew the model performance, NO_3-N, which has a higher correlation with NH_3-N for NO_3-N (0.71) and depth (0.66), and salinity, which has a higher correlation with NH_3-N for salinity (0.77) and EC (0.69), were excluded from the final model. This decision ensured the robustness and accuracy of the model by mitigating the effects of multicollinearity on NH_3-N prediction.

3.5. Comparative analysis of NH_3-N concentration prediction

The Random forest analysis elucidated the hierarchy of variable importance for predicting NH_3-N concentrations (Fig. 3). The algorithm was run iteratively once, including NH_3-N as an independent variable and after excluding it to understand its influence on the importance of other parameters and the prediction performance. The variables were ranked based on their computed importance scores, which signified their contributions to the accuracy of the model predictions. When NH_3-N was included as a predictor, NH_3-N demonstrated the highest importance score, followed closely by EC and temperature. Cl^-, turbidity, ORP, and TDS also showed substantial importance. Conversely, excluding NH_3-N from the model significantly altered the

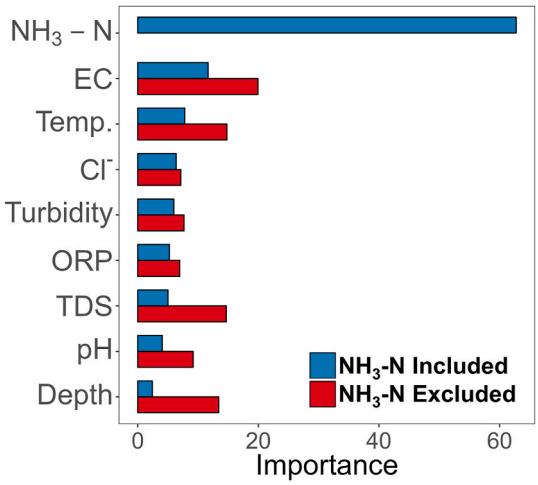

Fig. 3. Comparative importance of water quality parameters in NH_3-N prediction using the Random Forest algorithm. Variable importance scores are derived from a Random Forest model, comparing scenarios with and without NH_3-N as an independent variable. Parameters are ranked on the vertical axis with their corresponding importance scores on the horizontal axis, where the blue bars represent the model with NH_3-N and red bars indicate the model without NH_3-N.

importance of the landscape. EC remained a dominant predictor with an increased score, underscoring its critical role irrespective of the presence of NH_3-N in the model. Notably, the exclusion of NH_3-N increased the importance of parameters such as temperature, TDS, and depth, suggesting that these variables may interact in a complex manner with NH_3-N dynamics. ORP and Cl^- were less influential predictors in the absence of NH_3-N.

The efficacies of the Gaussian radial, random forest, and SVM radial algorithms in predicting the NH_3-N concentrations were evaluated across models with different numbers of parameters (Fig. 4). When NH_3-N was included as a predictor, the RMSE values fluctuated with the number of parameters included in the model. The Gaussian radial approach exhibited a less clear trend without any trends in the RMSE with an increase in the parameter count, suggesting improved performance with a more complex model. The random forest algorithm exhibited a less clear trend, with certain parameter counts (specifically, four parameters) yielding lower RMSE values, indicating more accurate predictions. The SVM radial models showed an increase in RMSE variability with a higher number of parameters, suggesting a potential overfitting issue when too many parameters were used. When NH_3-N was not included in the predictive parameters, the RMSE values generally increased across all the algorithms, highlighting the importance of NH_3-N for model accuracy. The Gaussian radial algorithm exhibited a moderate decrease in RMSE as the number of parameters increased. In contrast, the random forest algorithm demonstrated a slight enhancement in prediction accuracy with the inclusion of more parameters, achieving the lowest RMSE with seven parameters. The SVM radial algorithm demonstrated the highest RMSE variability, indicating that it was the most sensitive to the absence of NH_3-N. The optimal model performance, represented by the dashed red line in both scenarios, was achieved using a different number of parameters for each algorithm. Models employing the random forest algorithm exhibited the minimum RMSE with four parameters when NH_3-N was included as a predictor and with seven parameters when NH_3-N was excluded.

3.6. AutoML-driven advances in groundwater NH_3-N predictive accuracy

The predictive accuracies of various ML models with data processing approaches for NH_3-N concentrations in groundwater were evaluated. The comparative analysis focused on the influence of data processing methods, ML algorithms, and feature selection on model performance, as assessed by the RMSE, MAE, and coefficient of determination (R^2) (Table 4). The dataset utilized for model training encompassed two primary data types: laboratory-analyzed and sensor-calibrated. In some instances, these were further processed with augmentation to enhance the robustness of model training. Two categories of ML approaches were employed: manual ML and AutoML, with specific algorithms, such as SVM radial, Extra Trees, and Random Forest, which were selected as optimal within their respective categories. The results indicated that the models developed using calibrated and augmented data significantly outperformed those trained solely on laboratory-analyzed data. Notably, AutoML with Random Forest, when applied to calibrated and augmented data, demonstrated superior predictive capacity, achieving an RMSE of 0.38, MAE of 0.23, and R^2 of 0.90. This model also utilized a broader set of features, indicating that the inclusion of a more extensive range of water quality indicators can be beneficial for predictive accuracy. In contrast, the Manual ML (SVM radial) model trained on laboratory-analyzed data, using a narrower feature set, exhibited the poorest performance, with an RMSE of 1.23, MAE of 0.77, and R^2 of 0.58. This suggests that the choice of the AutoML algorithm, as well as the richness of the dataset regarding feature diversity and data quality (as enhanced by calibration and augmentation), are critical factors in developing robust predictive models.

A comparative analysis of the ML models for predicting NH_3-N concentrations in groundwater revealed distinct variations in prediction performance, contingent upon whether the NH_3-N sensor data were included as a predictive feature. Models utilizing NH_3-N sensor data as variables, specifically those processed through calibration and augmentation, demonstrated significantly enhanced predictive performance. This is evidenced by AutoML, with Random Forest achieving the highest R^2 value and the lowest RMSE and MAE. Conversely, models that excluded NH_3-N sensor data as features exhibited reduced performance metrics. For example, when the dataset was laboratory-analyzed and did not incorporate sensor-based NH_3-N data, the resulting models, Manual ML (SVM radial) and AutoML (Random Forest), yielded higher RMSE values (1.23 and 0.77, respectively) and lower R^2 values (0.58 and 0.72, respectively), signifying lower predictive accuracy and reliability. The inclusion of NH_3-N sensor data notably improved the predictive performance of the machine learning models for the NH_3-N concentration. Therefore, the optimal predictive framework integrates calibrated and augmented data with a comprehensive feature set, including NH_3-N sensor readings, and utilizes AutoML techniques, particularly the random forest algorithm, to achieve the highest accuracy in predicting groundwater NH_3-N levels.

To assess the broader applicability and robustness of our AutoML-driven predictive framework, additional validation was performed using datasets from two independent regions: Hoengseong, characterized by nitrate and TPH contamination, and Uijeongbu, predominantly contaminated by TPH. Despite the significantly differing hydrogeological and contamination characteristics in these areas, the predictive results remained consistently robust. Specifically, the AutoML model employing Random Forest achieved exceptionally low RMSE (0.008–0.02) and MAE (0.004–0.01) values, along with high R^2 (0.89–0.98), reinforcing the broader applicability of the framework and effectiveness across varying groundwater contamination scenarios.

4. Discussion

We successfully demonstrated the integration of manual ML and AutoML to enhance the accuracy and reliability of groundwater quality predictions using sensor data, particularly in environmentally sensitive areas such as carcass burial sites. This dual approach addresses challenges in traditional monitoring systems, especially when dealing with complex, high-volume data (Singha et al., 2021). Our framework align with those of earlier studies, emphasizing the critical need for advanced sensor-based environmental monitoring techniques (Sagan et al., 2020).

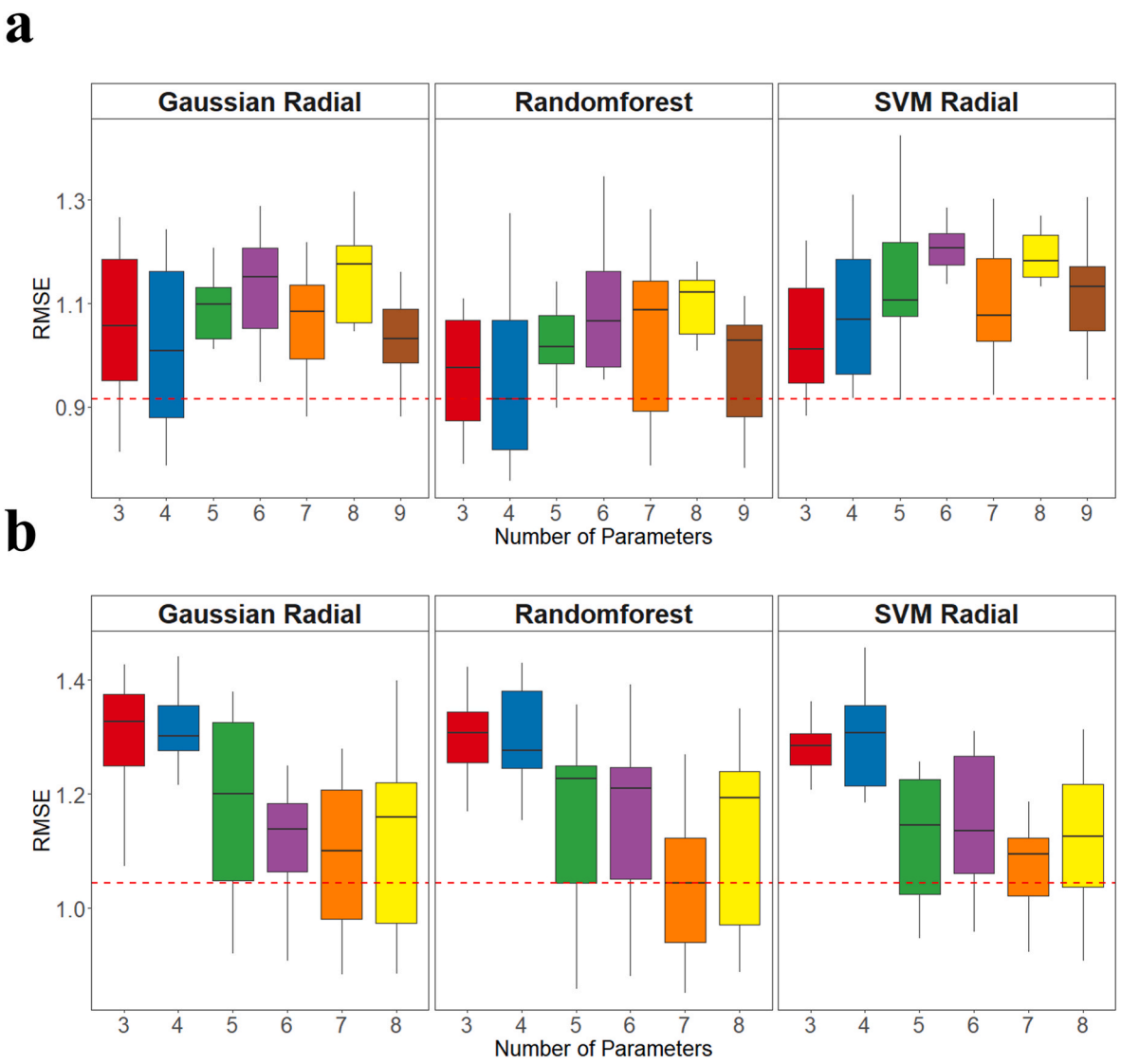

Fig. 4. Optimal parameter combination selection based on RMSE values across three algorithms. Boxplots comparing RMSE values for models built with varying numbers of water quality parameters using three different algorithms: Gaussian radial, Random Forest, and SVM radial. Panel (a) represents combinations including NH_3-N as a parameter, while panel (b) shows combinations excluding NH_3-N. The dashed red line across the plots indicates a benchmark RMSE value for the lowest value of RMSE in the results for both scenarios.

By combining expert-driven algorithm selection with the automation capabilities of AutoML, the model delivers both precision and operational efficiency in groundwater analysis. This is particularly beneficial in carcass burial sites, where conventional monitoring often fails to ensure timely detection of contamination. This real-time monitoring capability represents a substantial advancement over traditional methods, which typically involve delayed analysis owing to the need for sample collection and laboratory testing (Sale et al., 2021). AutoML further streamlines model development, enabling timely and scalable applications in dynamic field conditions (Karmaker et al., 2021). This adaptability, essential for long-term monitoring in dynamic environments such as carcass burial sites, ensures the accuracy and efficiency of our system and enhances real-time groundwater quality assessment with advanced ML techniques and sensor data integration. Beyond technical validation, the development of user-friendly interfaces and cloud-based tools coupled with our framework is crucial for widespread adoption by practitioners and environmental agencies. Simple dashboards for real-time data visualization and automated calibration scripts could further streamline operational workflows, facilitating immediate integration into existing monitoring programs.

The effectiveness of the SVM radial algorithm in standardizing the calibration process for various water quality parameters constitutes a pivotal aspect of our study. The capacity of the SVM radial algorithm to manage nonlinear relationships and its resilience against overfitting are crucial for standardizing the calibration of various water quality parameters within intricate groundwater datasets, accommodating complex environmental interactions, and extensive datasets (Mohan et al., 2020). Additionally, its ability to manage high-dimensional data and adaptability to diverse groundwater environments underscores its suitability for conducting comprehensive groundwater quality analyses (El Bilali et al., 2021). This finding holds significance as it suggests a potential reduction in the complexity and computational resources required for groundwater quality analysis (Hanoon et al., 2021). Our study further demonstrated how a single algorithm, when appropriately selected and applied, can efficiently standardize the calibration process across multiple water quality parameters. This transformative finding revolutionizes the water quality monitoring process, leading to reduction in complexity and computational demands (Yusri et al., 2022). The ability of the SVM radial algorithm to consistently perform across various water quality indicators underlines its robustness and

Table 4

Comparative performance of NH_3-N prediction models across different scenarios. A comprehensive comparison of 10 prediction model scenarios for NH_3-N concentration, using various combinations of data types, algorithms, and features. Each model's performance is evaluated using the number of features and key metrics, including RMSE, Mean Absolute Error (MAE), and the coefficient of determination (R^2). The models range from those trained on laboratory-analyzed data to those utilizing calibrated and augmented data, with algorithms including machine learning (ML) with Support Vector Machine (SVM) Radial and Automated Machine Learning (AutoML) with Random Forest and ExtraTrees. Additional evaluations were conducted using our developed AI-framework applied to Hoengseong and Uijeongbu regions, demonstrating significantly improved predictive accuracy.

Target	Region	Model		Features	Number of features	RMSE	MAE	R^2
		Data type	Algorithm					
NH_3-N	Anseong	Laboratory-analyzed	ML (SVM Radial)	EC, NO_3-N, TDS, depth	4	1.23	0.77	0.58
		Laboratory-analyzed	AutoML (Randomforest)	EC, NO_3-N, TDS, depth	4	0.77	0.52	0.72
		Calibrated	ML (Randomforest)	NH_3-N, EC, temperature, Cl^-	4	0.84	0.57	0.76
		Calibrated	AutoML (ExtraTrees)	NH_3-N, EC, temperature, Cl^-	4	0.65	0.4	0.77
		Calibrated + Augmented	ML (Randomforest)	NH_3-N, EC, temperature, Cl^-	4	0.55	0.32	0.87
		Calibrated + Augmented	AutoML (Randomforest)	NH_3-N, EC, temperature, Cl^-	4	0.38	0.23	0.9
		Calibrated	ML (Randomforest)	EC, temperature, TDS, depth, pH, turbidity, Cl^-	7	1.11	0.83	0.76
		Calibrated	AutoML (ExtraTrees)	EC, temperature, TDS, depth, pH, turbidity, Cl^-	7	0.75	0.53	0.72
		Calibrated + Augmented	ML (Randomforest)	EC, temperature, TDS, depth, pH, turbidity, Cl^-	7	0.6	0.37	0.84
		Calibrated + Augmented	AutoML (Randomforest)	EC, temperature, TDS, depth, pH, turbidity, Cl^-	7	0.4	0.25	0.89
	Hoengseong	Calibrated + Augmented	AutoML (Randomforest)	EC, NH_3-N, temperature, TDS	4	0.008	0.004	0.89
	Uijeongbu	Calibrated + Augmented	AutoML (Randomforest)	NO_3-N, DO, pH, Cl^-, ORP	5	0.02	0.007	0.98

reliability—critical factors in environmental monitoring, where data accuracy is imperative for making informed decisions (Sathish et al., 2023). This study not only contributes to the theoretical understanding of ML in environmental science but also offers practical solutions for designing and implementing more efficient and effective environmental monitoring systems.

We also observed a notable improvement in model prediction performance when NH_3-N was included as an independent variable despite its sensor data showing some inaccuracies and instability (Capella et al., 2020). This inclusion, aimed at evaluating the impact of an imprecise parameter on the predictive capabilities of the model, enhanced its accuracy. The decision to analyze models with and without NH_3-N was driven by the need to understand how a potentially unreliable parameter could affect overall predictions. Our findings revealed that NH_3-N, despite its low correlation with other variables, significantly contributed to predictive accuracy. This is consistent with an earlier study that noted independent variables with low correlations can provide valuable and unique information to a model, thereby improving predictions (Kim et al., 2019). Moreover, the model achieved excellent predictive performance using just four parameters (NH_3-N, EC, temperature, and Cl^-) when NH_3-N was incorporated. This simplifies well sensor management for long-term monitoring by reducing the number of parameters required for effective prediction. The roles of EC, temperature, and Cl^- are crucial in this context. For example, Cl^- levels can indicate pollution events and affect the nitrogen cycle (Koh et al., 2017). EC is a known indicator of water pollution, including nutrient pollution (Mahanta et al., 2022), and temperature, which affects chemical reaction rates, is a valuable predictor of NH_3-N (Hue et al., 2014). The ripple effect of this approach has several significant implications, enhancing the efficiency and accuracy of NH_3-N prediction and offering insights into the optimization of groundwater monitoring. By demonstrating that reliable results can be achieved with fewer sensors, this study suggests potential cost savings, easier sensor maintenance, and more sustainable long-term monitoring practices, thereby contributing to advancements in groundwater quality monitoring.

We employed Random Forest and Extra Tree regression algorithms to predict the NH_3-N concentrations in groundwater. These algorithms are particularly suited for environmental studies due to their ability to handle complex and nonlinear interrelationships among variables, characteristics frequently observed in environmental datasets (Meray et al., 2022). Their adaptability in processing various data types, such as numerical and categorical data, renders them highly versatile in diverse environmental scenarios (Tyralis et al., 2019). A noteworthy finding of our study was the slightly superior performance of AutoML compared to manual ML in terms of prediction accuracy. This enhanced performance is largely attributed to the refined capabilities of AutoML in hyperparameter tuning, allowing for more precise model optimization (He et al., 2021). In addition to improved prediction accuracy, AutoML offers additional benefits, including resource conservation, time efficiency, and enhanced accessibility for users with varying levels of expertise (He et al., 2021). The effectiveness of AutoML in environmental studies is further evidenced by its successful application in other studies. A study in China's Hubei region used AutoML for comprehensive groundwater potential mapping, and the AutoML model demonstrated the best predictive ability and accuracy (Bai et al., 2022). Similarly, a study in India employing AutoML for groundwater level prediction highlighted its proficiency in handling spatially varied data and maintaining stability even with minor uncertainties in input parameters (Singh et al., 2024). The impact of AutoML extends beyond mere prediction accuracy; it also democratizes AI in environmental science. AutoML is making advanced data analysis more accessible and widespread by enabling users with varied expertise levels, often referred to as "citizen data scientists," to effectively explore, analyze, and implement data models.

The geographical focus on one site (Sanbuk-ri) provides in-depth insights into local environmental dynamics but may limit the broader applicability of our findings. This specificity, while valuable for targeted research, poses challenges in generalizing our results to other regions or environmental contexts (Ntona et al., 2022). However, our methodological framework, encompassing data preprocessing, sensor calibration against laboratory references, and AutoML algorithm and parameter selection, has been rigorously tested in two additional and distinct groundwater environments. Remarkably, despite substantial differences in contamination type and hydrogeological characteristics, the

framework consistently demonstrated excellent predictive accuracy across these diverse conditions. This robustness arises from the comprehensive data preprocessing approach, which effectively mitigates sensor drift and missing-value issues common to diverse hydrogeological settings, and from the flexible AutoML optimization process that dynamically selects the most suitable algorithms and parameters tailored to each regional specific groundwater characteristics. Therefore, while our baseline model originates from a single site, the underlying framework has been validated for broader applicability and robustness, confirming its extensibility to diverse groundwater environments. In terms of future research, there is a clear path towards integrating low-cost sensors. This approach has immense potential to enhance environmental data collection and analysis (Mao et al., 2019). The incorporation of such technologies can further refine predictive models, making them more robust and versatile. This is particularly relevant in the context of environmental monitoring, where the spatial and temporal variability of data can be challenging to manage. Additionally, future research could include expanding sensor capabilities to detect a broader range of contaminants, such as arsenic and mercury, which may originate from different burial practices or contamination sources. These technologies have immense potential to provide more comprehensive datasets covering larger geographical areas and longer time periods (Sun and Scanlon, 2019). Such advancements could revolutionize the field of environmental monitoring and provide scientists and policymakers with accurate and detailed information to inform their decisions.

5. Conclusion

Our results demonstrate that combining automated calibration, core parameter selection, and AutoML yields robust, cost-effective NH_3–N predictions in sensor-based groundwater monitoring. By effectively addressing sensor drift and leveraging laboratory-referenced data, this integrated framework ensures stable and highly accurate performance across dynamic field conditions, even with fewer monitored variables. Validations conducted in hydrogeologically distinct regions (Hoengseong and Uijeongbu) further reinforce the adaptability and reliability of this framework, showcasing its consistent predictive accuracy irrespective of differing contamination patterns or regional characteristics. Consequently, our approach provides a practical and broadly applicable solution for sites where contamination risks necessitate continuous surveillance, offering timely alerts that support environmental protection and resource management. Notably, the modular workflow design and reliance on widely available sensor inputs enhance its versatility, enabling straightforward adaptation to various hydrological settings and contaminants beyond carcass burial scenarios. Our findings emphasize the scalability of data-driven groundwater monitoring, underscoring how a carefully selected feature set can effectively capture critical contamination signals. Future extensions will further refine cost-effectiveness assessments, incorporate additional parameter sets as needed, and investigate long-term data collection strategies to capture seasonal or decadal variability. Ultimately, this study presents a reliable, validated, and adaptable pipeline for proactive groundwater quality management.

CRediT authorship contribution statement

Jaeuk Youn: Writing – original draft, Validation, Methodology, Formal analysis, Data curation. **Do Hwan Jeong:** Writing – original draft, Validation, Project administration, Conceptualization. **MoonSu Kim:** Writing – review & editing, Validation, Conceptualization. **Kyong Min Woo:** Resources, Methodology, Conceptualization. **Tae Kwon Lee:** Writing – review & editing, Validation, Supervision. **Hyun-koo Kim:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We gratefully acknowledge the funding received for the projects of the National Institute of Environmental Research (NIER) in South Korea (NIER-2023-04-02-076; NIER-2023-01-01-096) from the Korean Ministry of Environment.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvman.2025.126729.

Data availability

Data will be made available on request.

References

Adombi, A.V.D.P., et al., 2022. Comparing numerical modelling, traditional machine learning and theory-guided machine learning in inverse modeling of groundwater dynamics: a first study case application. J. Hydrol. 615, 128600. https://doi.org/10.1016/j.jhydrol.2022.128600.

Bai, Z., et al., 2022. Groundwater potential mapping in Hubei Region of China using machine learning, ensemble learning, deep learning and AutoML methods. Nat. Resour. Res. 31, 2549–2569. https://doi.org/10.1007/s11053-022-10100-4.

Bakker, M., Schaars, F., 2019. Solving groundwater flow problems with time series analysis: you may not even need another model. Groundwater 57, 826–833. https://doi.org/10.1111/gwat.12927.

Capella, J.V., et al., 2020. A new ammonium smart sensor with interference rejection. Sensors 20, 7102. https://doi.org/10.3390/s20247102.

Cortés-Ibáñez, J.A., et al., 2020. Preprocessing methodology for time series: an industrial world application case study. Inf. Sci. 514, 385–401. https://doi.org/10.1016/j.ins.2019.11.027.

da Silva, H.M.F., et al., 2021. SAGAD: Synthetic Data Generator for Tabular Datasets. Anais do XXXVI Simpósio Brasileiro de Bancos de Dados. SBC, pp. 1–12.

El Bilali, A., et al., 2021. Groundwater quality forecasting using machine learning algorithms for irrigation purposes. Agric. Water Manag. 245, 106625. https://doi.org/10.1016/j.agwat.2020.106625.

Hanoon, M.S., et al., 2021. Application of artificial intelligence models for modeling water quality in groundwater: comprehensive review, evaluation and future trends. Water, Air, Soil Pollut. 232, 1–41. https://doi.org/10.1007/s11270-021-05311-z.

He, X., et al., 2021. AutoML: a survey of the state-of-the-art. Knowl. Base Syst. 212, 106622. https://doi.org/10.48550/arXiv.1908.00709.

Hue, C., et al., 2014. Near infrared spectroscopy as a new tool to determine cocoa fermentation levels through ammonia nitrogen quantification. Food Chem. 148, 240–245. https://doi.org/10.1016/j.foodchem.2013.10.005.

Karmaker, S.K., et al., 2021. Automl to date and beyond: challenges and opportunities. ACM Comput. Surv. 54, 1–36. https://doi.org/10.48550/arXiv.2010.10777.

Kim, G.G., et al., 2019. Prediction model for PV performance with correlation analysis of environmental variables. IEEE J. Photovoltaics 9, 832–841. https://doi.org/10.1109/Jphotov.2019.2898521.

Knoll, L., et al., 2019. Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. Sci. Total Environ. 668, 1317–1327. https://doi.org/10.1016/j.scitotenv.2019.03.045.

Koh, E.-H., et al., 2017. Impacts of land use change and groundwater management on long-term nitrate-nitrogen and chloride trends in groundwater of Jeju Island, Korea. Environ. Earth Sci. 76, 1–16. https://doi.org/10.1007/s12665-017-6466-3.

Korea Ministry of Environment, 2009. Drinking Water Quality Standard Methods. Announcement of Ministry of Environment, Korea. Announcement No. 2009-2332010.

López-de-Lacalle, J., 2019. Tsoutliers: Detection of Outliers in Time Series. 10.32614/CRAN.package.tsoutliers.

Mahanta, A.R., et al., 2022. Evaluation of long-term nitrate and electrical conductivity in groundwater system of Peninsula, India. Appl. Water Sci. 12, 17. https://doi.org/10.1007/s13201-021-01568-1.

Mao, F., et al., 2019. Low-cost environmental sensor networks: recent advances and future directions. Front. Earth Sci. 7. https://doi.org/10.3389/feart.2019.00221.

Meray, A.O., et al., 2022. Pylenm: a machine learning framework for long-term groundwater contamination monitoring strategies. Environ. Sci. Technol. 56, 5973–5983. https://doi.org/10.1021/acs.est.1c07440.

Mohan, L., et al., 2020. Support vector machine accuracy improvement with classification. 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN). IEEE, pp. 477–481.

Nelson, D.B., et al., 2021. Precipitation isotope time series predictions from machine learning applied in Europe. Proc. Natl. Acad. Sci. 118, e2024107118. https://doi.org/10.1073/pnas.2024107118.

Ntona, M.M., et al., 2022. Modeling groundwater and surface water interaction: an overview of current status and future challenges. Sci. Total Environ. 846. https://doi.org/10.1016/j.scitotenv.2022.157355.

Sagan, V., et al., 2020. Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. Earth Sci. Rev. 205, 103187. https://doi.org/10.1016/j.earscirev.2020.103187.

Sale, T., et al., 2021. Real-time soil and groundwater monitoring via spatial and temporal resolution of biogeochemical potentials. J. Hazard Mater. 408, 124403. https://doi.org/10.1016/j.jhazmat.2020.124403.

Sathish, P., et al., 2023. Design of water quality monitoring system using SVM algorithm. 2023 4th International Conference on Electronics and Sustainable Communication Systems. ICESC), pp. 1196–1201.

Singh, A., et al., 2024. AutoML-GWL: automated machine learning model for the prediction of groundwater level. Eng. Appl. Artif. Intell. 127, 107405. https://doi.org/10.1016/j.engappai.2023.107405.

Singha, S., et al., 2021. Prediction of groundwater quality using efficient machine learning technique. Chemosphere 276. https://doi.org/10.1016/j.chemosphere.2021.130265.

Sun, A.Y., Scanlon, B.R., 2019. How can big data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. Environ. Res. Lett. 14. https://doi.org/10.1088/1748-9326/ab1b7d.

Tiyasha, 2020. A survey on river water quality modelling using artificial intelligence models: 2000-2020. J. Hydrol. 585. https://doi.org/10.1016/j.jhydrol.2020.124670.

Tyralis, H., et al., 2019. A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water 11, 910. https://doi.org/10.3390/w11050910.

Weng, Z., 2019. From conventional machine learning to AutoML. Journal of Physics: Conference Series. IOP Publishing, 012015.

Yusri, H.I.H., et al., 2022. Water quality classification using SVM and XGBoost method. 2022 IEEE 13th Control and System Graduate Research Colloquium (ICSGRC). IEEE, pp. 231–236.

Zhu, S., et al., 2020. Forecasting of water level in multiple temperate lakes using machine learning models. J. Hydrol. 585, 124819. https://doi.org/10.1016/j.jhydrol.2020.124819.