Journal of Environmental Management 377 (2025) 124719

Contents lists available at ScienceDirect

Journal of Environmental Management

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

Research article

Developing a real-time water quality simulation toolbox using machine learning and application programming interface

Gi-Hun Bang [a], Na-Hyeon Gwon [b], Min-Jeong Cho [b], Ji-Ye Park [b], Sang-Soo Baek [b,*]

[a] Department of Integrated Water Management, Yeungnam University, Daehak-ro 280, Gyeongsan-si, Water Campus, Korea Water Cluster, Gukgasandan-daero 40-gil, Guji-myeon, Dalseong-gun, Gyeongsangbuk-do, Daegu, Republic of Korea
[b] Department of Environmental Engineering, Yeongnam University, 280 Daehak-Ro, Gyeonsan-Si, Gyeongbuk, 38541, Republic of Korea

ARTICLE INFO

Handling editor: Jason Michael Evans

Keywords:
Application programming interface
Graphic user interface
Machine learning
Water quality simulation
Sensitivity analysis
Uncertainty analysis

ABSTRACT

Rivers are vital for sustaining human life as they foster social development, provide drinking water, maintain aquatic ecosystems, and offer recreational spaces. However, most rivers are being increasingly contaminated by pollutants from non-point sources, urbanization, and other sources. Consequently, real-time river water quality modeling is essential for managing and protecting rivers from contamination, and its significance is growing across various sectors, including public health, agriculture, and water treatment systems. Therefore, a real-time river water quality simulation toolbox was developed using machine learning (ML) and an application program interface (API). To create the toolbox, models that simulated water quality parameters such as chlorophyll *a* (*Chl-a*), dissolved oxygen (DO), total nitrogen (TN), total organic carbon (TOC), and total phosphorus (TP) at each point in the Nakdong River were constructed. The models were constructed using Artificial neural network (ANN), Random Forest (RF), support vector machines (SVM), and data from API. Subsequently, hyperparameter optimization was conducted to enhance the model's performance. During training, the models' performances were evaluated and compared based on the data sampling method and ML algorithms. Models trained with random sampling data outperformed those trained with time-series data. Among the algorithm models that used random sampling data, the RF exhibited the best performance. The average coefficient of determination (R^2) values for each water quality simulation with randomly sampled data using RF for DO, TN, TP, *Chl-a*, and TOC were 0.79, 0.65, 0.74, 0.45, and 0.48, respectively. For ANN, they were 0.7, 0.51, 0.64, 0.35, and 0.35, respectively, and for SVM, they were 0.73, 0.51, 0.59, 0.21, and 0.3, respectively. The *Chl-a* and TOC models exhibited relatively poor performance, whereas the DO, TN, and TP models demonstrated superior performance. Diversifying the input data variables is necessary to improve the performance of the *Chl-a* and TOC models. Sensitivity and uncertainty analyses were conducted to evaluate and enhance the models' understanding. Furthermore, using a graphic user interface (GUI) toolbox, user convenience was maximized.

1. Introduction

With ever-increasing demand, river water plays a pivotal role in supplying domestic, industrial, and agricultural water (Amador-Castro et al., 2024; Baek et al., 2020). Given the direct implications of river water quality on public health and the environment, the importance of water quality monitoring has steadily increased (Aldhyani et al., 2020; Wang, Y. et al., 2017). However, traditional monitoring methods require specialized facilities with trained manpower, which is neither cost-effective nor time-efficient (Amador-Castro et al., 2024; Jones et al., 2011; Villa et al., 2019). Traditional river water quality modeling was proposed to assess river pollutants and aquatic ecosystems (Chen, Q. et al., 2012; Cui et al., 2019). However, simulating or predicting the quality of surface water has proven to be a highly challenging task, primarily due to the influence of external factors such as weather patterns, human activities, and industrial processes (Noori et al., 2020).

Conventional approaches to river water quality modeling include parameter-based statistical and deterministic models (Khullar and Singh, 2021). Deterministic models represent various chemical and physical processes through statistical terms, using variables acquired from historical data or obtained through empirical means, experience, or examination. In contrast, statistical models establish general rules

* Corresponding author.
E-mail address: ssbaek@yu.ac.kr (S.-S. Baek).

https://doi.org/10.1016/j.jenvman.2025.124719
Received 31 May 2024; Received in revised form 21 February 2025; Accepted 22 February 2025
Available online 28 February 2025

based on field data and extract valuable information from experimental observations (Ahmed et al., 2019). However, these models require extensive data regarding numerous hydrological subprocesses to generate accurate results, and such comprehensive data are not always readily available (Baek et al., 2020). Moreover, these models depend on accurately defined rate constants and coefficients associated with hydrological, chemical, physical, and biological processes, which are highly dependent on specific temporal and spatial conditions (Khullar and Singh, 2021). Many factors influencing water quality exhibit complex nonlinear relationships with input variables, which conventional modeling approaches may not adequately capture.

Artificial intelligence (AI) has emerged as a promising alternative to address the limitations of conventional models. Advancements in computing power and software have propelled the rapid development of AI. Data-driven modeling, also known as artificial intelligence (AI) modeling, has demonstrated remarkable performance and has been applied in various scientific and technological domains, including medicine, image recognition, and water quality simulations (Kourou et al., 2015; Li, 2022; Park, Y. et al., 2015). Unlike traditional modeling approaches, data-driven modeling relies solely on available data to derive results. Before the learning process of data-driven models, users specify the input variables and desired output. By establishing correlations between water quality parameters and their influencing factors, the models generate valid results. Moreover, data-driven modeling incorporates optimization functions, enabling the derivation of more accurate outcomes than conventional approaches. Therefore, data-driven modeling excels at simulating complex phenomena and nonlinear data patterns (Zanoni et al., 2022). Machine learning (ML), a subset of data-driven modeling, has been frequently employed to simulate and predict surface water quality. Park et al. (2015) employed Artificial Neural Networks (ANN) and support vector machines (SVM) to simulate chlorophyll-*a* (*Chl-a*) levels and developed an early warning protocol. Baek et al. (2021a,b) explored the use of deep and ML techniques to estimate micropollutants by replacing internal standards. Wang et al. (2021) conducted spatial heterogeneity modeling of water quality using the Random Forest (RF) algorithm. Gómez et al. (2021) used ML and satellite data to estimate *Chl-a* in the Menor Sea. In this study, RF, SVM, ANN, and deep neural networks (DNN) were adopted for ML. Castrillo and García (2020) estimated high-frequency nutrient concentrations using a multiple linear regression model and RF, with RF showing better performance in most cases. Saboe et al. (2021) estimated water quality parameters (e.g., DO, pH, and turbidity) and algal concentration (e.g., blue-green algae and chlorophyll) using long short-term memory (LSTM) with microbial potentiometric sensor data (MPS). This study emphasizes the advantages of combining MPS technology with an ML algorithm. Zhang et al. (2024b) predicted dissolved oxygen (DO) through feature screening and machine learning. The features were selected through maximum information coefficient, then employed RF to combine five ML algorithms such as K-Nearest Neighbors (KNN), Backpropagation (BP) Neural Network, Support Vector Regression (SVR), Long Short-Term Memory (LSTM), and Kernel Ridge Regression (KRR). As a result, the ensemble-RF model effectively tracked and replicated DO trends throughout the model development phase, successfully capturing both stable periods and episodes of sudden hypoxia. Kim, K. and Ahn (2022) predicted chl-a utilizing ANN, RF, and SVM with feature selection through Pearson's correlation analysis. They employed meteorological and daily water quality data to train each model. Among the algorithm, the RF showed the best performance with determination coefficient (R^2) of 0.747 and root mean squared error (RMSE) of 8.617 mg/m^3. ML has been used across various environmental fields, demonstrating outstanding performance through novel methods. Nevertheless, ML-based water quality modeling still has room for improvement in terms of usability and interpretability.

To enhance the usability of ML-based water quality modeling, integrating application programming interfaces (APIs) and graphic user interfaces (GUI) holds significant promise. APIs facilitate seamless communication among different software systems, which is particularly advantageous in computational environments where most ML-based modeling is performed. By designing an algorithm that connects the API process, data pre-processing, and data input process, the data acquisition and input processes can be automated, eliminating the need for manual labor in data management. Consequently, this automated approach ensures the sustainability of the modeling system beyond the study duration and facilitates its commercialization. Additionally, employing a GUI alongside automated data management further enhances the prospects for commercialization. A GUI offers a user-friendly visual interface operated via windows, increasing accessibility for a wide range of users, irrespective of their familiarity with computer languages. Another significant improvement in ML-based modeling is the incorporation of sensitivity analysis (SA) and uncertainty analysis (UA). Despite the outstanding performance of ML-based models, their lack of interpretability often renders them "black boxes" that contain uncertainties from observational data. The SA plays a crucial role in addressing this challenge by providing interpretability, and it explains specific model results by deriving the importance weights of the input data. Using SA, ML-based models transition from being "black boxes" to "grey boxes," thereby enhancing their reliability. Additionally, UA increases the reliability of the model by providing results within a probability-based range.

This study aimed to develop a real-time river water quality simulation and prediction toolbox using ML techniques to effectively preserve rivers for the well-being of human populations. To enhance the usability of ML-based modeling, this study incorporated the utilization of API, GUI, SA, and UA. In addition, hyperparameter optimization was conducted to achieve more accurate simulations and predictions. Collectively, the specific objectives of this study were to: 1) leverage a national open-source API for automated data updates to ensure data currency; 2) employ ML techniques to simulate and predict water quality parameters; 3) compare results based on data type and algorithm; 4) use SA and UA to interpret the models and provide explanatory insights; and 5) integrate all these procedures into a user-friendly GUI toolbox that facilitates accessibility for a broad range of users.

2. Materials and method

2.1. Study site

The study area was located on the Nakdong River, one of South Korea's four major rivers, situated in the southeastern part of the country (Fig. 1). The Nakdong River has the second-largest drainage area in South Korea, covering 23,647 km^2 and the longest stream length of 510.4 km (Jung et al., 2016). Approximately 13 million people reside in this region, accounting for over one-quarter of the Korean population (Park, S. S. and Lee, 2002). With a monsoon climate, the river basin experiences an average annual temperature of 13.4 °C and rainfall of 1325 mm (KMA, 2021; Seo et al., 2021). Approximately 7 million people reside in this area, and more than 10 million rely on this river for water consumption (Park, S. S. and Lee, 2002). This river is pivotal for public health and drinking water management (Hur et al., 2013). Daily flow rate and meteorological data were acquired from stations located in Gumi, Sungju, Haman, and Milyang, managed by the National Institute of Environmental Research (NIER) in the Republic of Korea. Additionally, weather data were collected from Chupungryeong, Gumi, North-Changwon, and Kimhae weather stations. Water quality data, including DO, total nitrogen (TN), total phosphorous (TP), *Chl-a*, and total organic carbon (TOC), were obtained from NIER. The primary focus of this study was on four specific stations: Dogae (36° 16′ 24″ N, 128° 20′ 45″ E), Dasan (35° 51′ 4″ N, 128° 24′ 4″ E), Chungam (35° 23′ 48.86″ N, 128° 31′ 17.78″ E), and Sangdong (35° 21′ 49.05″ N, 128° 54′ 16.98″ E). These stations span the main channel of the Nakdong River, extending from downstream to upstream.

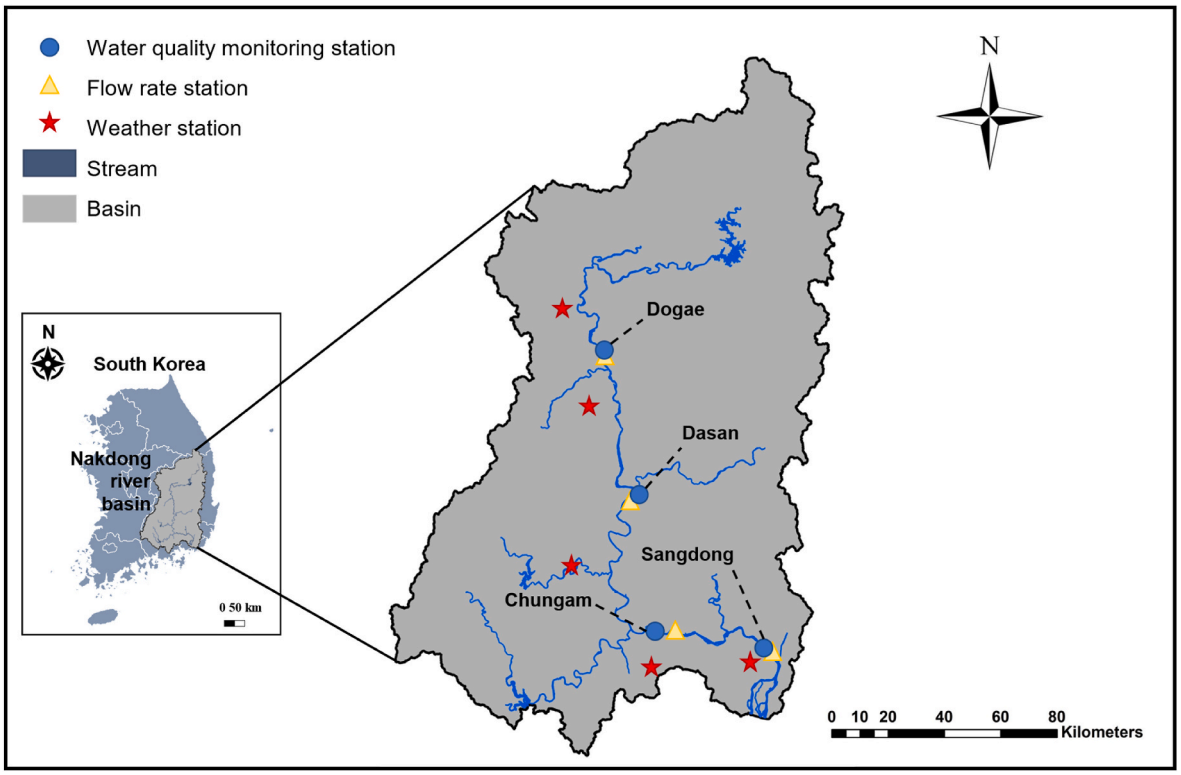

Fig. 1. Study site. Map of Nakdong River Basin in Korea. Blue circles indicate water quality monitoring stations, yellow triangles indicate flow rate stations and red stars indicate weather stations.

2.2. Data preparation

2.2.1. Data acquisition with application programming interface (API)

An API was used to automatically acquire water quality, and weather data. Acting as a mediator, the API enables communication among different software components (Ofoeda et al., 2019). API ensures access to up-to-date data and facilitates the management of large databases. The API used in this study is a web API accessible over a network, enabling data exchange between us and national institutes (Raatikainen et al., 2021). An API provided by national institutes was utilized to obtain data for training the models and operating the toolbox. In the API setup, the personal computer functioned as the client requesting data, while the national institutes acted as servers to provide the data. To operate the API effectively, the data type, observation stations, and data collection period were specified. Through the API, weather data such as average temperature (°C), rainfall (mm), wind speed (m/s), day length (h), solar radiation (MJ/m^2), and small evaporation (mm) were obtained from the Korea Meteorological Administration (KMA). Additionally, water quality data, including Chl-a (μg/L), DO (mg/L), TN (mg/L), TOC (mg/L), and TP (mg/L), were retrieved from the National Institute of Environmental Research (NIER). The dataset was collected between January 2015 and May 2023. The API implementation was performed using MATLAB (2022a). The summary statistics of input and output variables are shown in Fig. 2 and Table 1.

2.2.2. Influence of weather and flow rate on water quality

Weather and flow rate are fundamental factors driving the dynamics of water quality through their influence on physical, chemical, and biological processes in aquatic ecosystems (Whitehead et al., 2009). Precipitation plays a critical role by contributing to surface runoff, which transports nutrients, sediments, and pollutants into water bodies (Müller et al., 2020). This process significantly affects water quality parameters such as TN, TP, and suspended solids, often resulting in increased turbidity and heightened risks of eutrophication (Atique and An, 2019). During heavy rainfall events, agricultural runoff containing fertilizers and pesticides further exacerbates water quality degradation, leading to elevated pollutant loads (Khan et al., 2015; Zahoor and Mushtaq, 2023).

Flow rate also has a significant impact on water quality by regulating the mixing, transport, and dilution of pollutants, nutrients, and organic matter within aquatic systems. High flow rates during extreme weather events intensify turbulence and sediment transport, leading to elevated concentrations of suspended solids and organic matter in the water column (Whitehead et al., 2009). Conversely, low flow conditions reduce the dilution capacity of water bodies, resulting in pollutant accumulation and conditions such as hypoxia and eutrophication, particularly in stagnant areas (Nilsson and Renöfält, 2008).

Weather variables such as air temperature and solar radiation are equally critical in influencing water quality. Higher air temperatures reduce the solubility of oxygen in water, thereby directly impacting DO levels, which are essential for aquatic organisms (Guo et al., 2021). Additionally, insolation and daytime duration drive algal growth, contributing to variability in chl-a concentrations, a key indicator of primary productivity in aquatic ecosystems (Adams et al., 2022). These interactions highlight the interconnected nature of weather and hydrological factors with water quality.

2.2.3. Data pre-processing

A data pre-processing procedure was conducted on both weather and water quality datasets to ensure data integrity and improve model reliability. Outlier detection was performed using the Interquartile Range (IQR) method, which identified some values near peaks as outliers. The IQR method classifies values outside 1.5 times the interquartile range (middle 50% of the data) as potential outliers (M. Mahajan et al., 2020). Given the naturally high variability of environmental data (Weatherhead et al., 1998), these values were retained to preserve dataset variability. Excluding or modifying these peak values could have weakened input-output correlations and reduced the model's

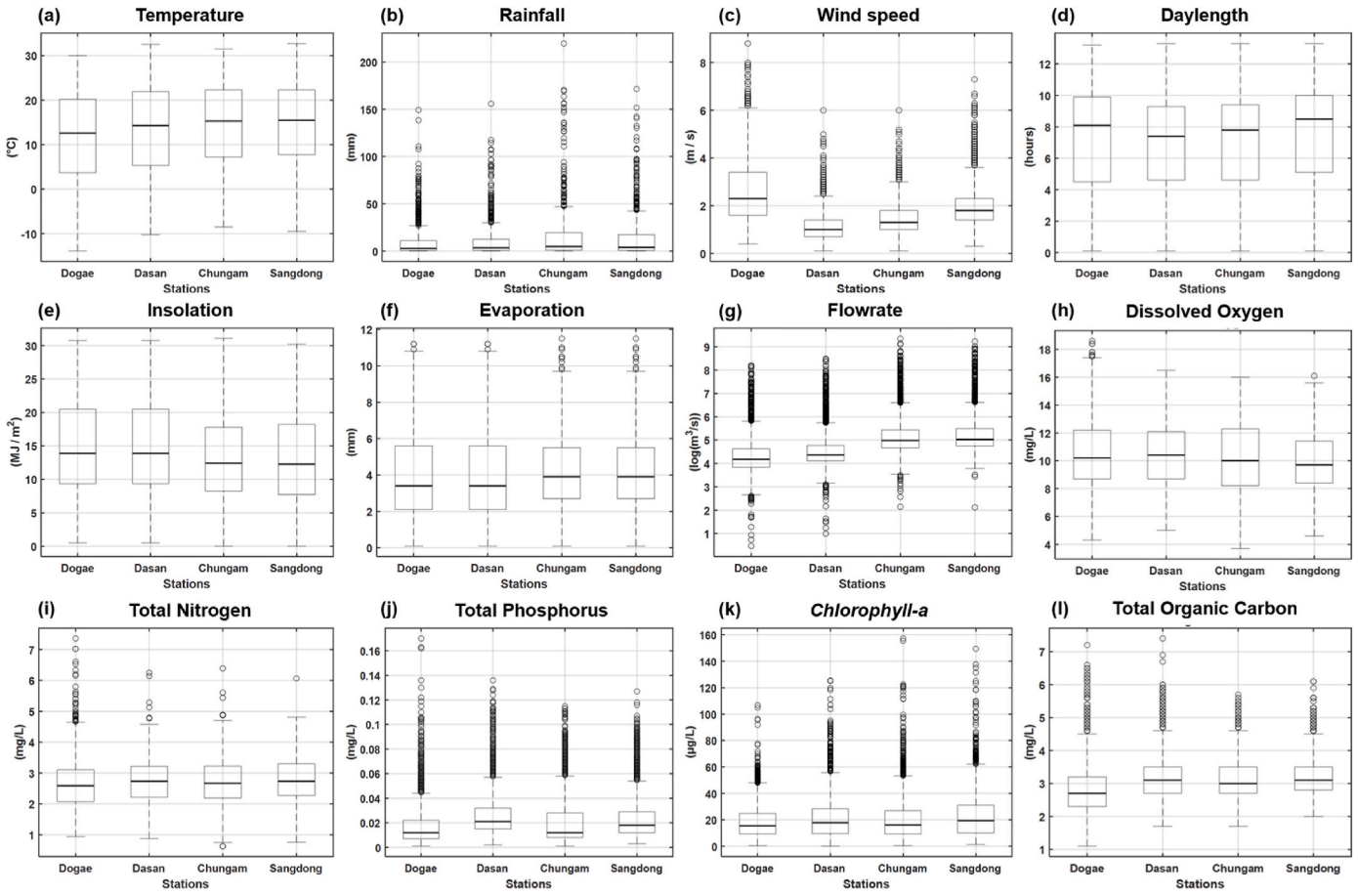

Fig. 2. Box plot results that shows interquartile ranges (IQRs) and variability for water quality model outputs across stations. (a) represents Temperature, (b) represents Rainfall, (c) represents Wind speed, (d) represents Daylength, (e) represents Insolation, (f) represents Evaporation, (g) represents Flowrate, (h) represents Dissolved Oxygen, (i) represents Total Nitrogen, (j) represents Total Phosphorus, (k) represents Chlorophyll-a, and (l) represents Total Organic Carbon in the form of box plot results.

ability to capture inherent variability and interactions (O'Leary et al., 2016). For structural errors, water quality data contained a specific issue where missing values were recorded as −999 instead of NaN. These −999 values were reclassified as missing to maintain consistency in the dataset. Missing values were handled differently for each dataset. In meteorological data, missing values were filled using the mean of adjacent values to minimize gaps (Yang, J. et al., 2017). However, if four or more consecutive values were missing, they were replaced with NaN and excluded from the input data. These pre-processing steps ensured a robust and reliable dataset for model training and validation.

The datasets were normalized using logarithmic transformation, min-max scaling, and z-score methods to improve the accuracy of the simulations and reduce computational costs (Nawi et al., 2013). The logarithmic transformation method addresses challenges related to data scarcity and exponential characteristics, thereby enhancing the performance of the employed models (Behnood and Daneshvar, 2020; Wong et al., 2013). Min-max and z-score normalization scale the data within a narrow range, reduce outliers, and improve data quality and simulation accuracy (Cabello-Solorzano et al., 2023). For the output data, a logarithmic transformation was applied to simulate the flow rate and water quality data. Normalization using either the min-max or z-score was conducted based on a trial-and-error methodology. Among the input data, the flow rate exhibited a wider range, and z-score normalization was used to mitigate its excessive influence on the results.

The datasets used for model training and testing were divided into two distinct types: time-series data and randomly sampled data. The time-series data were segregated into training and testing subsets at proportions of 70% and 30%, respectively. In contrast, randomly sampled data were created using the 'dividerand' function in MATLAB. This function facilitated the establishment of a uniform data range via random sampling and partitioned the data into training and testing subsets in an 80%–20% ratio.

2.3. Development of flow rate and water quality simulation toolbox

The procedure for the real-time flow rate and water quality simulation toolbox is illustrated in Fig. 3. The toolbox adopts API and ML techniques to simulate real-time flow rate and water quality. Additionally, it can forecast water quality for up to three days. The toolbox development process comprises four stages: (1) dataset preparation (Fig. 3 (1)), (2) ML model training and optimization (Fig. 3 (2)), (3) model evaluation (Fig. 3 (3)), and (4) integration of the models into the GUI toolbox (Fig. 3 (4)). The API automatically acquires real-time environmental data, including weather and water quality data. ML was then used to predict flow rate and water quality using this environmental data. APIs facilitate communication between different software applications, allowing for effective data exchange (Ofoeda et al., 2019; Zhang, B. et al., 2011). Water quality and real-time weather data were acquired via the API provided by the KMA and NIER (Fig. 3 (2A)). ML was then used to simulate water quality and quantity by identifying complex relationships between input and output variables during the learning phase (Masson et al., 2023; Pang et al., 2021; Zhang, W. et al., 2019). SA and UA were conducted to assess the impact of the water parameters and introduce uncertainty into the trained model (Kline,

Table 1
Summary statistics of input and output variables for each site (2015–2023). Std refers standard deviation and N refers the number of data.

Stations			Mean	Min	Max	Std	N
Dogae	Input variables	Temperature (°C)	11.97	−13.90	30.00	9.64	3070
		Rainfall (mm)	3.00	0.00	149.30	10.31	3073
		Wind speed (m/s)	2.64	0.40	8.80	1.31	3073
		Day length (h)	6.49	0.00	13.20	3.99	3068
		Solar radiation (MJ/m^2)	14.69	0.52	30.76	7.29	3070
		Evaporation (mm)	3.96	0.10	11.20	2.31	3063
		Flowrate (m^3/s)	107.43	5.67	3498	212.11	3073
	Output variables	DO (mg/L)	10.58	4.30	18.60	2.46	2968
		TN (mg/L)	2.64	0.94	7.36	0.80	2777
		TP (mg/L)	0.02	0.001	0.46	0.02	2687
		TOC (mg/L)	2.81	1.10	7.20	0.80	2505
		chl-a (μg/L)	18.76	0.50	106.80	13.00	2701
Dasan	Input variables	Temperature (°C)	13.76	−10.20	32.50	9.52	3071
		Rainfall (mm)	2.88	0.00	155.90	10.67	3073
		Wind speed (m/s)	1.10	0.00	6.00	0.62	3073
		Day length (h)	6.24	0.00	13.30	3.75	3051
		Solar radiation (MJ/m^2)	14.69	0.52	30.76	7.29	3070
		Evaporation (mm)	3.96	0.10	11.20	2.31	3063
		Flowrate (m^3/s)	135.39	4.92	4921	281.41	3073
	Output variables	DO (mg/L)	10.51	5.00	16.50	2.26	2822
		TN (mg/L)	2.70	0.88	6.25	0.69	2782
		TP (mg/L)	0.03	0.002	0.14	0.02	2526
		TOC (mg/L)	3.19	1.70	7.40	0.67	2712
		chl-a (μg/L)	21.29	0.10	125.40	15.84	2348
Chungam	Input variables	Temperature (°C)	14.76	−8.50	31.50	8.85	3065
		Rainfall (mm)	3.99	0.00	219.50	14.98	3073
		Wind speed (m/s)	1.45	0.10	6.00	0.68	3071
		Day length (h)	6.27	0.00	13.30	3.83	3066
		Solar radiation (MJ/m^2)	13.21	0.04	31.05	6.75	3066
		Evaporation (mm)	4.14	0.00	11.50	1.94	2992
		Flowrate (m^3/s)	262.91	13.36	8757	482.22	3073
	Output variables	DO (mg/L)	10.18	3.70	16.00	2.48	2727
		TN (mg/L)	2.72	0.63	6.40	0.72	2661
		TP (mg/L)	0.02	0.001	0.12	0.02	2428
		TOC (mg/L)	3.12	1.70	5.70	0.62	2449
		chl-a (μg/L)	20.67	0.50	157.20	16.88	2473
Sangdong	Input variables	Temperature (°C)	14.88	−9.50	32.70	8.73	3069
		Rainfall (mm)	3.61	0.00	171.50	13.10	3073
		Wind speed (m/s)	1.98	0.30	7.30	0.82	3073
		Day length (h)	6.71	0.00	13.30	4.00	3056
		Solar radiation (MJ/m^2)	12.98	0.00	30.21	6.97	3068
		Evaporation (mm)	4.14	0.00	11.50	1.94	2992
		Flowrate (m^3/s)	336.90	9.79	10106	716.70	3073
	Output variables	DO (mg/L)	9.88	4.60	16.10	2.02	2858
		TN (mg/L)	2.77	0.76	6.07	0.71	2600
		TP (mg/L)	0.02	0.003	0.13	0.02	2469
		TOC (mg/L)	3.20	2.00	6.10	0.58	2423
		chl-a (μg/L)	23.17	1.40	149.50	17.44	2540

1985; Morris, 1991) (Fig. 3 (3B and 3C)). Integrated within the GUI, the models and evaluation tools were designed to optimize the user experience and facilitate interaction through a visual display that includes windows, menus, icons, and mouse interactions (Bonsiepe, 1990; Jansen, 1998).

Flow rate simulation models were developed to generate real-time flow rate data using RF and weather data from the API. (Fig. 3 (1B)). As real-time flow rates could not be obtained through an API because of the absence of such a service, models were designed to simulate flow rate data. These simulated flow rate data were then integrated into the weather dataset. Although the API did not provide flow rate data, the water quality models used both weather and flow rate data (m^3/s) as inputs, with water quality data as the output.

Data preprocessing techniques such as logarithmic transformation, z-scoring, and min-max scaling were applied to improve the simulation accuracy (Fig. 3 (2A)) (Behnood and Daneshvar, 2020; Cabello-Solorzano et al., 2023; Wong et al., 2013). Subsequently, the dataset was generated into time series and randomly sampled datasets to evaluate the performance of the models using both types of data (Fig. 3 (2B)). Water quality simulation models employing the ANN, RF, and SVM were developed through iterative learning using the aforementioned datasets (Fig. 3 (2C)). The models were structured hierarchically, with the output from the flow rate model providing the input for the water quality simulation and forecasting models.

To forecast water quality values, a dataset that incorporated a temporal discrepancy between the input and output datasets was used (Fig. S1). This discrepancy was generated by offsetting the output data to create a one-day lag between the input and output data. During the training of the simulation models, the input and output data pairs were aligned by date across all columns. However, in the forecasting models, the input and output dates were deliberately misaligned by one day to highlight the temporal differences. This procedure was replicated three times to enable the prediction of water quality for up to three days.

Additionally, a lookback function was incorporated into the input parameters for both flow rate and water quality models. This function allows models to use historical data by setting a specific lookback size, enabling the consideration of data from a specified number of days (Baek et al., 2022; Qin et al., 2019). For example, a lookback size of three enabled the models to use data from the three days preceding the simulation. In the training phase of the flow rate models, a consistent look-back size ranging from 3 to 60 days was applied across all stations. Conversely, the water quality models underwent an optimization

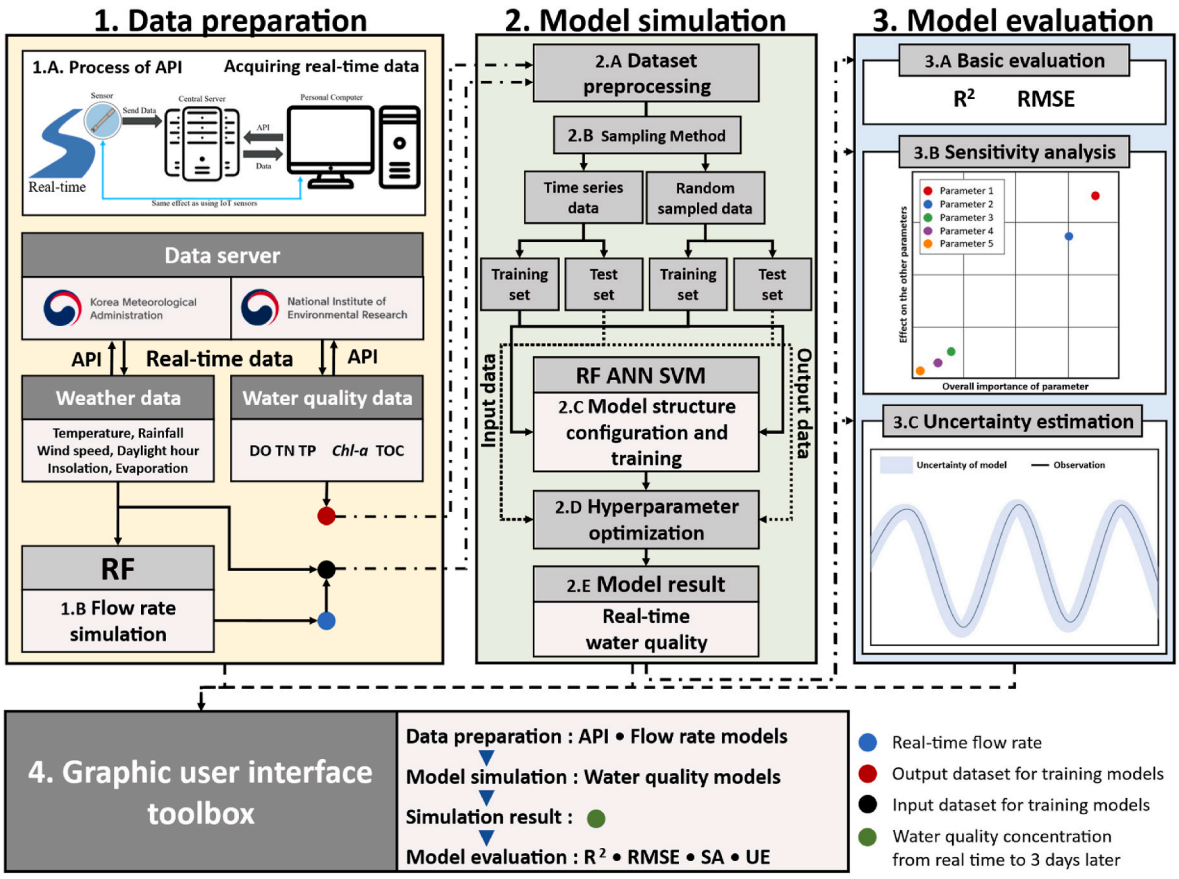

Fig. 3. Flowchart for developing the water quantity and quality simulation toolbox using machine learning (ML) algorithms comprises four steps: (1) preparation of the dataset, (2) training and optimization of the ML models, (3) model evaluation, and (4) integration of the models into the graphic user interface toolbox.

process to ascertain the most effective lookback size, which varied between 3 and 180 days. Hyperparameter optimization was then performed to enhance the performance of the water quality model (Fig. 3 (2D)).

Using the optimal model, we analyzed the performance of water quality forecasting with increasing forecast lead times (days) of up to three days (Fig. 3 (2E)). The performances of these models were assessed using the R^2 and RMSE metrics, and their effectiveness was compared based on the dataset and algorithm employed. Additionally, a comprehensive model evaluation was conducted using SA and UA to provide a deeper understanding of the behavior of the models (Fig. 3 (3)). In the final stage, the API, trained models, and evaluation methods were integrated into a GUI toolbox to enhance user convenience (Fig. 3 (4)). The toolbox for each model development was executed on a system equipped with an Intel i5-10505 CPU @ 3.20 GHz and 16 GB of RAM, using MATLAB as the implementation platform, which is renowned for its extensive use in scientific computing.

2.4. Machine learning

2.4.1. Random forest

RF, introduced by Breiman (2001), is a machine learning algorithm that falls under the ensemble learning category, and it combines the concepts of bagging and random feature selection to enhance predictive accuracy (Fig. S77) (Calhoun et al., 2014). RF employs multiple decision trees to generate reliable predictions (Rasaei and Bogaert, 2019). By aggregating subsamples from the original data, RF constructs an individual decision tree (Alnahit et al., 2022). This process is repeated to create an ensemble of decision trees. Given input data, RF generates prediction values from each decision tree in the forest. In the regression tasks, the final prediction value is determined by averaging the predictions of all the trees. For classification tasks, the RF assigns the class with the highest occurrence among the decision trees as the predicted class. The RF method specializes in processing high-dimensional, noisy, and imbalanced data to overcome overfitting issues (Khalilia et al., 2011; Yuan et al., 2017).

2.4.2. Artificial neural network

An ANN is a computational model inspired by the biological processes of the human brain with the aim of simulating information processing, and it comprises numerous interconnected units known as artificial neurons or nodes (Fig. S78) (Agatonovic-Kustrin and Beresford, 2000). Each node contains weight and activation functions (Maier and Dandy, 2000). Through its activation functions, the ANN obtains nonlinearity (Chen, Y. et al., 2020). The nodes are organized into layers and linked by coefficients, commonly referred to as weights, to form neural structures (Park, Y. et al., 2015). An ANN comprises the input, hidden, and output layers. The behavior of a neural network is dictated by the activation functions of its neurons, learning rules, and overall network architecture. When the input layer accepts input data, the weight of each node is multiplied and transferred to the hidden layer. This process is repeated in the hidden layers, yielding a prediction result from the output layer. Throughout the training process, the interconnections between the units are iteratively optimized until the prediction error is minimized, enabling the network to achieve the desired level of accuracy (Chen, Y. et al., 2020).

2.4.3. Support vector mahcine

The SVM, introduced by Cortes and Vapnik (1995), was developed for classification and uses hyperplanes that divide a dimensional space

into defined decision boundaries (Fig. S79) (Singh, K. P. et al., 2011). The SVM has been extended to solve regression problems and time-series data by introducing an insensitive loss function (Vapnik et al., 1996; Wang, W. et al., 2003). In SVM regression, the SVM aims to find the optimal hyperplane in the highest dimension that minimizes the distance to all data points (Pan et al., 2008), and it constructs data-driven nonlinear process models using kernel functions, including the radial basis function (RBF), linear function, and polynomial function (Wu et al., 2010). In this study, the RBF was adopted as the kernel function. Furthermore, the SVM adopts the structure risk minimization principle, which minimizes the upper boundary of the generalization error in the Vapnik-Chernoverkis dimension (Wang, W. et al., 2003). Subsequently, SVM possesses the following advantages: (1) it prevents overfitting problems, (2) it is capable of dealing with high-dimensional inputs, and (3) it can model nonlinear relationships (Balabin and Lomakina, 2011; Gunn, 1998).

2.4.4. Hybridizing

The trained models from each algorithm were hybridized and compared to the performance of individual algorithms to assess potential improvements. For hybridization, an ensemble of three models was created, combining the outputs of each algorithm and averaging them, following a method similar to Random Forest, where the final result is derived by averaging outputs from multiple decision trees.

2.5. Model evaluation and optimization

The R^2, RMSE, and Nash–Sutcliffe efficiency (NSE) were adopted as the simulation and prediction performance (Baek et al., 2022; Chang et al., 2015; Jeung et al., 2024; Sokolova et al., 2022). R^2, the coefficient of determination, indicates the proportion of variance in the observed data explained by the model, with higher values between 0 and 1 signifying better model accuracy. RMSE measures the average magnitude of prediction errors, providing insight into the model's overall performance, with lower values indicating fewer errors. NSE evaluates the predictive power of hydrological and environmental models by comparing observed and predicted values, ranging from negative infinity to 1, with values closer to 1 indicating stronger model agreement and values less than 0 reflecting poor model performance (Jeung et al., 2024; Nelson et al., 2020). These evaluation metrics were determined as follows:

$$R^2 = \left(\frac{\sum_{i=1}^{N} \left(Y_i^{sim} - Y_i^{mean\_sim}\right)\left(Y_i^{obs} - Y_i^{mean\_obs}\right)}{\sqrt{\left(\sum_{i=1}^{N} \left(Y_i^{sim} - Y_i^{mean\_sim}\right)^2\right)\left(\sum_{i=1}^{N} \left(Y_i^{obs} - Y_i^{mean\_obs}\right)^2\right)}} \right)^2 \quad (1)$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \left(Y_i^{obs} - Y_i^{sim}\right)^2 \Big/ n} \quad (2)$$

$$NSE = 1 - \frac{\sum_{i=1}^{n} \left(Y_i^{obs} - Y_i^{sim}\right)^2}{\sum_{i=1}^{N} \left(Y_i^{obs} - Y_i^{mean\_obs}\right)^2} \quad (3)$$

Where Y_i^{obs} is the i_{th} observed flow rate [m^3/s] or water quality concentration [μg/L or mg/L], Y_i^{sim} is the i_{th} simulated flow rate [m^3/s] or water quality concentration [μg/L or mg/L], $Y_i^{mean\_obs}$ is the average of the observed flow rate [m^3/s] or water quality concentration [μg/L or mg/L], $Y_i^{mean\_sim}$ is the average of the simulated flow rate [m^3/s] or water quality concentration [μg/L or mg/L], and n is the total number of observed values.

The constructed models were optimized to enhance their accuracy and precision, aligning with best practices in predictive modeling (Liu, S. et al., 2013; Shahriari et al., 2015; Singh et al., 2024b,c). MATLAB's 'hyperparameteroptimization' function, based on Bayesian optimization, was used to automatically determine the optimal values for each algorithm's hyperparameters over 30 iterations (Shahriari et al., 2015). In each iteration, the posterior distribution refines as new parameter sets are evaluated, resulting in an increasingly optimized model configuration (Victoria and Maragatham, 2021). This iterative approach is especially advantageous for data-driven applications, where precise parameter optimization is crucial to capturing complex environmental patterns. By efficiently identifying the best hyperparameters, this method significantly enhances model performance, as data-driven models are highly sensitive to hyperparameter selection during the training phase (Hutter et al., 2015).

For the RF models, five key hyperparameters were optimized: "Method," "NumLearningCycles," "LearnRate," "MinLeafSize," and "NumBins." The "Method" specifies the ensemble aggregation method, with options for "LSBoost" and "Bag." The "NumLearningCycles" determines the number of ensembles' learning cycles, while the "LearnRate" defines the shrinkage learning rate. The "MinLeafSize" sets the minimum number of observations per branch node, and "NumBins" determines the number of bins used for numeric predictors. For ANN, four hyperparameters were optimized: "Activation," "Standardize," "Lambda," and "LayerSizes." The "Activation" function controls the type of activation for fully connected layers, with choices including "relu," "tanh," "sigmoid," and "none." The "Standardize" option determines whether predictor data should be standardized, while "Lambda" adjusts the regularization strength. "LayerSizes" defines the size of each neural network layer. In the case of SVM, the hyperparameter "KernelScale" was fine-tuned to control the dimension of the kernel in the Gaussian kernel function to find the best prediction. For specific details on the optimized hyperparameters for each algorithm, refer to Table 2.

2.6. Sensitivity analysis and uncertainty analysis

2.6.1. Sensitivity analysis

SA is employed to assess the impacts of input variables on natural phenomena by identifying their significance or contributions (Lamboni et al., 2011). This approach enhances understanding and reliability by elucidating the importance of inputs through SA (Razavi and Gupta, 2015) and has been acknowledged as an effective method for identifying influential parameters (Makler-Pick et al., 2011; Nossent et al., 2011; Saltelli et al., 2004). SA identifies key variables and their influence on other variables, thus aiding in comprehending the interrelationships between variables and models and providing insights into their reflections on natural phenomena (Kim, S. et al., 2021; Yi et al., 2016). In this study, global SA was conducted using the Elementary Effect Test (EET) and Latin Hypercube-One-Factor-At-a-Time (LH-OAT) methods to identify the influential parameters among the input variables. Global SA examines all variables simultaneously, thus enhancing the comprehensiveness of the analysis. The LH-OAT integrates Latin Hypercube sampling with the one-factor-at-a-time design (Van Griensven and Meixner, 2003), substituting the sampling method of the latter with the former to evaluate the entire range of parameters with fewer input sets than the Morris OAT design (Kim, S. et al., 2021; van Griensven et al., 2006; Xu et al., 2016). EET, widely recognized for its effectiveness and simplicity in global SA, minimizes the number of iterations required (Campolongo et al., 2007). The influence of input variables is quantified through Elementary Effects (EE), with the mean of the EEs determining the overall importance of an input variable to the model and the standard deviation of the EEs assessing the influence of specific variables on others (Campolongo et al., 2007; Morris, 1991). SA is also widely recognized as a screening method for identifying influential variables, further highlighting its utility in analyzing input-output relationships and enhancing model interpretability (Campolongo et al., 2011).

2.6.2. Uncertainty analysis

Uncertainty analysis is critical in water quality modeling as it identifies and quantifies potential errors in model predictions, providing

Table 2
Range of hyperparameter setting.

RF		ANN		SVM	
Parameters	Range	Parameters	Range	Parameters	Range
Method	LSBoost or Bag	Activation	Relu, tanh, or sigmoid	KernelScale	36.54–589.37
NumLearningCycles	38–500	Standardize	0 or 1		
LearnRate	0.05–0.16	Lambda	0.00–16.63		
NumBins	10–100	LayerSizes	1-3, 1-299		

insights into their potential range of reliability (Salinas et al., 2020). It addresses uncertainties arising from input parameter, modeling processes, and data preparation steps, which can significantly impact predictive accuracy (Singh et al., 2024a,b,c). Methods such as LH-OAT, that is originated from Monte Carlo, are effective in capturing variability and improving model predictions, particularly in dynamic environmental systems (Huang et al., 2015; Singh et al., 2024a,b,c). Additionally, uncertainty analysis enhances model transparency and robustness, supporting more informed and reliable decision-making in water resource management (Uddin et al., 2023).

UA was employed to examine and quantify the uncertainty stemming from the input data, enhancing reliability by providing a probabilistic area. The UA represents the model output as a range influenced by the uncertainty of the model inputs, considering the dynamics of natural conditions and the inherent limitations of measurement techniques (Beck, 1987). Zhang et al. (2024a,b) stated that model simulation values are validated as reliable when the simulation uncertainty is considered. In this study, we used Generalized Likelihood Uncertainty Estimation (GLUE) in conjunction with the LH-OAT method to estimate the uncertainty in model outputs based on the error rate of the input variables. The GLUE method is extensively applied for several reasons: 1) it estimates uncertainty based on the entirety of the input data rather than individual variables (Beck, 1987); 2) it accommodates variable interactions and nonlinearities through likelihood measures (Vázquez et al., 2009); and 3) it is recognized for its simplicity and ease of implementation (Shen et al., 2012).

To apply GLUE based on the error rates of the input variables, the error rates of the measuring instruments for the input variables were investigated, and the likelihood for each was defined. LH-OAT was used for random data sampling, enhancing the efficiency of the GLUE approach. In defining the thresholds, the 97.5% and 2.5% percentiles were established as limits. Consequently, the uncertainty of the model

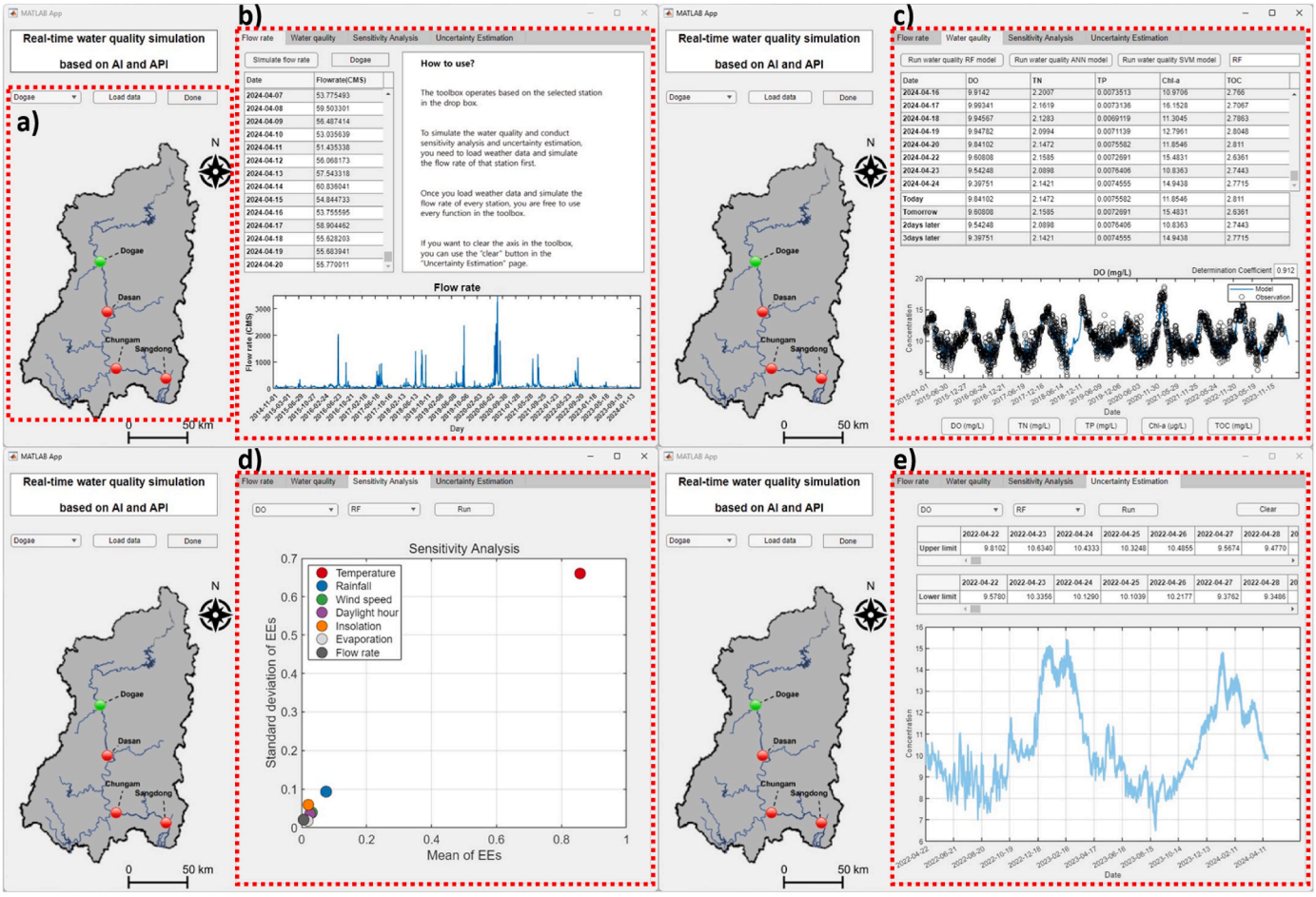

Fig. 4. View of the Graphic User Interface system window. a) loads the real-time data based on the drop-down, b) simulates real-time flow rate; c) simulates and forecasts the water qualities; d) analyzes the sensitivity; and e) estimates the uncertainty of the model.

was illustrated as the area between the upper and lower limits, with a 95% probability that the model output lies within this range, reflecting the uncertainty of the model input.

3. Results

3.1. Real-time flow rate/water quality simulation toolbox

A real-time toolbox for simulating flow rate and water quality was developed using ML and an API (Fig. 4). The toolbox consists of four modules: flow simulation, water quality prediction, sensitivity analysis, and uncertainty analysis. In the main window (Fig. 4(a)), a map displays the selected monitoring station as chosen from a drop-down menu. By clicking the "load data" button at the top of Fig. 4(a), real-time weather data and validation water quality data are retrieved from the data server via the API.

1) The first module (Fig. 4(b)) is the flow simulation module, which uses a trained flow model. It simulates real-time flow by inputting weather data acquired through the API. The simulation results are displayed in a table at the top left and a graph at the bottom.
2) The second module (Fig. 4(c)) is the water quality prediction module. It predicts water quality by utilizing the trained water quality model with real-time weather data from the API and flow data generated by the flow model. The simulation algorithm can be selected using the buttons at the top. The prediction results are displayed in the upper table and the graph below it. Users can view the results of each water quality parameter by selecting the respective parameter button at the bottom.
3) The third and fourth modules (Fig. 4(d) and (e)) perform sensitivity and uncertainty analyses, respectively. These analyses use the previously executed water quality prediction model and results to provide updated sensitivity and uncertainty analyses tailored to the latest data. Both modules allow users to select the algorithm and water quality parameter from drop-down menus to conduct analyses suited to different conditions. For the sensitivity analysis module (Fig. 4(d)), users can assess the impact of input variables by examining the rank of variables in the legend and the results graph. The uncertainty analysis module (Fig. 4(e)) displays the prediction results as a range graph, with upper and lower bounds shown in the table at the top.

3.2. Flow rate simulations

The simulated flow rates at each station are depicted in Fig. S2. During the training phase, the simulated flow rates closely matched the observed flow rates, yielding R^2 and RMSE values ranging from 0.96 to 0.99 and 16.18–169.96 m^3/s, respectively (Table 3). However, during the validation phase, the model's performance declined, yielding R^2 and RMSE values ranging from 0.64 to 0.81 and 88.56–202.89 m^3/s, respectively. The training phase R^2 was larger than 0.85, falling within the "very good" criterion proposed by Moriasi et al. (2015). The validation phase R^2 values for Dasan, Chungam, and Sangdong stations were greater than 0.7, meeting the "good" criterion, while Dogae station fell within the "satisfactory" criterion with an R^2 value of 0.64. The discrepancy in RMSE varied depending on the station, which was in turn influenced by the magnitude of the flow rate. During both the training and validation phases, the simulations captured the peak flow rates, demonstrating the model's capability to represent hydrological phenomena. Sangdong station showed the best performance, with an R^2 value of 0.81. However, Sangdong station had missing values in the observations from December 2020 to December 2021 (Fig. S2). In addition, Chungam station showed the best performance, yielding an R^2 of 0.74 without any missing values.

Table 3
Determination coefficient (R^2) and root mean squared error (RMSE) of flow rate simulation.

Stations	Training R^2	Validation R^2	Training RMSE	Validation RMSE
Dogae	0.97	0.64	53.2	88.56
Dasan	0.99	0.71	42.1	101.48
Chungam	0.96	0.74	169.96	202.89
Sangdong	0.99	0.81	16.18	167.87

3.3. Water quality simulations

3.3.1. Model performance depending on the data sampling method

The results of the water quality simulation are summarized in Figs. 5 and 6, Figs. S3–S42, Tables 4–7, and Tables S1–6. The simulations using randomly sampled data exhibited superior performance compared to those using time-series data (Figs. 5 and 6(a, b) and Figs. S15–30). In the RF simulations at Dasan, randomly sampled data produced validation R^2 values of 0.50–0.84 for DO, TN, TP, *Chl-a,* and TOC, with corresponding RMSE values of 0.014–11.52 (μg/L or mg/L). Conversely, the simulations using time-series data yielded validation R^2 values of 0.11–0.73 for DO, TN, TP, *Chl-a,* and TOC and corresponding RMSE values of 0.013–16.6 (μg/L or mg/L). The *Chl-a* and TOC simulation performances were improved when randomly sampled data were used. Moreover, instances of overestimated and underestimated water quality were more prevalent in the model using time-series data (Fig. 6(i) and (ii)). Fig. S32 demonstrates that simulations using randomly sampled data yielded improved performance with reduced overestimation and underestimation, as evidenced in the TN simulation using RF and TP simulations using ANN (Fig. S32).

3.3.2. Model performance depending on ML algorithms

The water quality simulated using RF, ANN, SVM, and hybridizing is shown in Fig. 6 and S3–S30. The RF simulations using randomly sampled data had R^2 values of 0.45–0.79 for DO, TN, TP, *Chl-a*, and TOC, respectively. For ANN, the corresponding R^2 values were 0.35–0.7, while for SVM, they were 0.21–0.73. The RF simulations using randomly sampled data had RMSE values of 0.01–12.41 (μg/L or mg/L) for DO, TN, TP, *Chl-a*, and TOC, respectively. For ANN, the corresponding RMSE values were 0.01–12.47 (μg/L or mg/L), while for SVM, they were 0.01–13.98. Among the ML algorithms employed, RF demonstrated the best performance with an average R^2 and RMSE of 0.62 and 2.9, respectively. In contrast, the ANN and SVM showed average R^2 values of 0.51 and 0.47 and average RMSE values of 2.97 and 3.26, respectively. The simulated results by the ANN, characterized by sporadic increases and decreases on consecutive days (Fig. 6(b)–S7, and S22), showed a higher variance compared to those of the ML models. While the ANN simulations occasionally showed better agreement with the peak than the RF simulations, overall, the performance of the ANN was lower than that of the RF.

The results of hybridization using ensemble techniques showed performance that was similar or lower than that of individual algorithms. For TN, the hybrid model demonstrated the best performance, achieving a coefficient of determination of 0.76 and an RMSE of 0.33, slightly outperforming the best individual algorithm, ANN. However, for DO, TP, TOC, and Chl-a, the hybrid model showed coefficients of determination of 0.84, 0.75, 0.49, and 0.62, respectively, and RMSE values of 1.04, 0.01, 0.43, and 10.55. These results were either similar to or slightly lower in terms of R^2 and marginally higher in RMSE compared to the best-performing individual algorithms for each parameter.

The DO simulations showed the best performance across both time series and randomly sampled data. Using RF with time-series data, the validation R^2 of DO was nearly 0.7, except for Sangdong station, where it exhibited an "unsatisfactory" result of an R^2 of 0.43. Using randomly sampled data, the validation R^2 of DO was close to 0.8, notably improving its performance at Sangdong station to an R^2 of 0.69. The TN

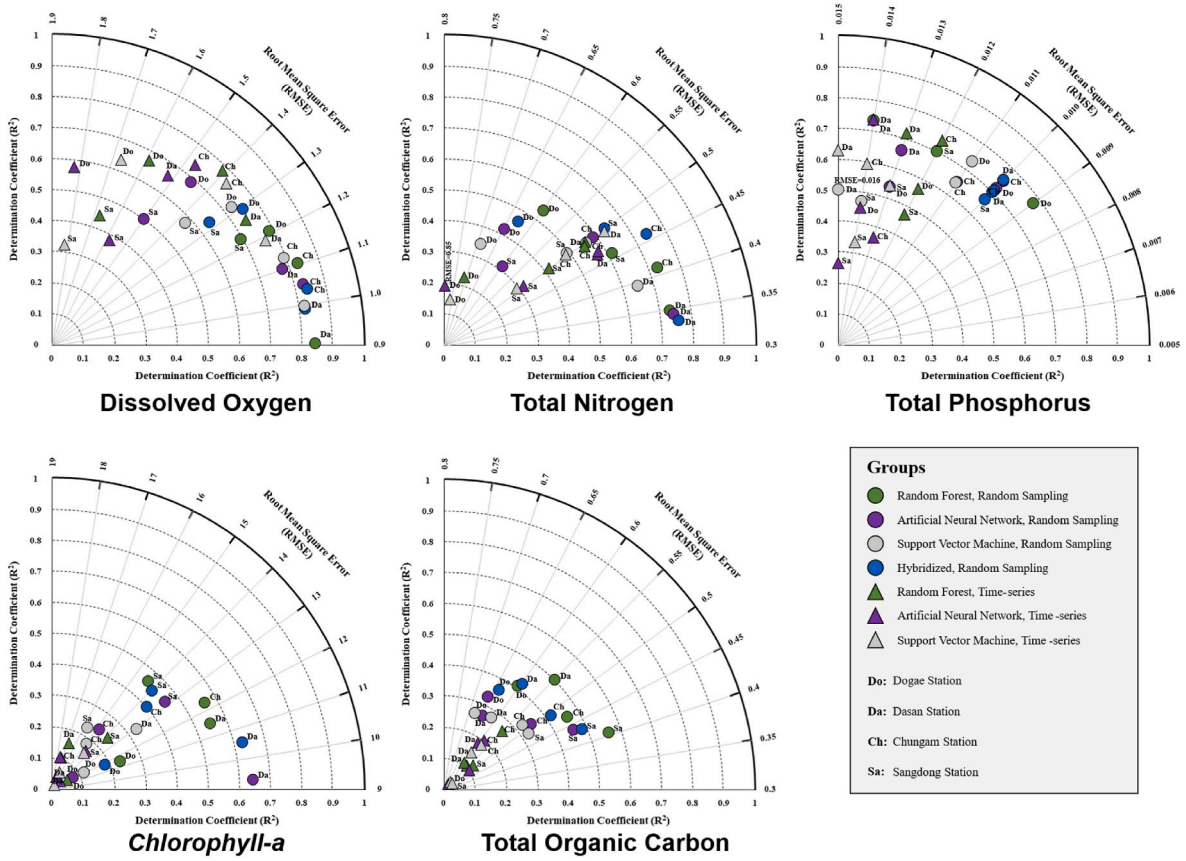

Fig. 5. Diagram for displaying the performance of each water quality model according the to utilized algorithm and data sampling method. The radial axis represents the coefficient of determination (R^2), while the azimuthal axis represents the root mean square error (RMSE).

and TP simulations achieved validation R^2 values mostly above 0.5, except for the TN simulation at Dogae and the TN and TP simulations at Sangdong with values of 0.22, 0.41, and 0.47, respectively. *Chl-a* and TOC simulations using time-series data yielded poor validation R^2 scores of below 0.1 (Moriasi et al., 2015), whereas those using randomly sampled data achieved validation R^2 scores of approximately 0.5, except for the Dogae station.

3.4. Water quality forecasting

The performances of the models varied based on the forecast lead time (Tables 4–7 and S1–S6). With RF and SVM, R^2 generally decreased as the forecast lead time increased. For instance, in the TP simulation using the RF with randomly sampled data at Dogae, the R^2 values were 0.78 for nowcasting, 0.74 for 1 day ahead, 0.73 for 2 days ahead, and 0.71 for 3 days ahead. The R^2 of the ANN seemed unaffected by the forecast lead time. For the TP simulation using the ANN with randomly sampled data at the Dogae station, the R^2 values were 0.71 for nowcasting, 0.51 for 1 day ahead, 0.68 for 2 days ahead, and 0.57 for 3 days ahead.

3.5. Sensitivity and uncertainty analysis

The SA results are shown in Fig. 7 and S43–S52. The x-axis represents the mean of the EEs, indicating feature importance, whereas the y-axis shows the standard deviation of the EEs, highlighting the impact of specific variables on others (Campolongo et al., 2007; Morris, 1991). The legends were sorted according to the mean of the EEs. For simulations using RF at the Dasan station with randomly sampled data, the most influential features for simulating DO were temperature, evaporation, and rainfall. For TN, the most influential features included temperature, flow rate, and evaporation, while for TP and *Chl-a*, they were temperature, flow rate, and wind speed. For TOC, the most influential features included temperature, insolation, and flow rate. In the ANN simulations with randomly sampled data at Dasan station, the most influential features for DO were temperature, flow rate, and evaporation; for TN were wind speed, flow rate, and rainfall; for TP were rainfall, flow rate, and wind speed; for *Chl-a* were wind speed, rainfall, and temperature; and for TOC were flow rate, rainfall, and temperature. The SVM simulations identified rainfall, temperature, insolation, and flow rate as the most influential features of DO, TN, TP, *Chl-a*, and TOC alike. In the SA of the RF and ANN, different rankings of feature importance were observed depending on the water quality parameter. SA in the RF tended to rank temperature as the most influential feature, regardless of water quality. The SVM simulations showed similar rankings across all the simulations.

The UA results are shown in Fig. 8 and S53–S76 and in Tables S7–9. The uncertainty band, between the 2.5th and 97.5th percentiles, indicates a confidence level of 95%. The mean, standard deviation, and maximum uncertainty were calculated to assess the uncertainty of each model. The UA results showed no significant relationship with the simulation performance. As shown in Fig. 8(a), the uncertainty of DO was greater than that of TN, despite the DO simulations using the ANN with time-series data demonstrating better performance (Table S8). Among the algorithms, SVM exhibited the lowest uncertainty, whereas RF showed the highest uncertainty, with no discernible improvements in uncertainty observed between simulations with time-series data and randomly sampled data. In the TN and TP simulations using the ANN at Dasan, an increase in uncertainty was observed when using randomly sampled data compared with time-series data (Fig. 8 and Table S8). Among the water quality simulations, the *Chl-a* and TP simulations predominantly exhibited higher uncertainties (Tables S7–9).

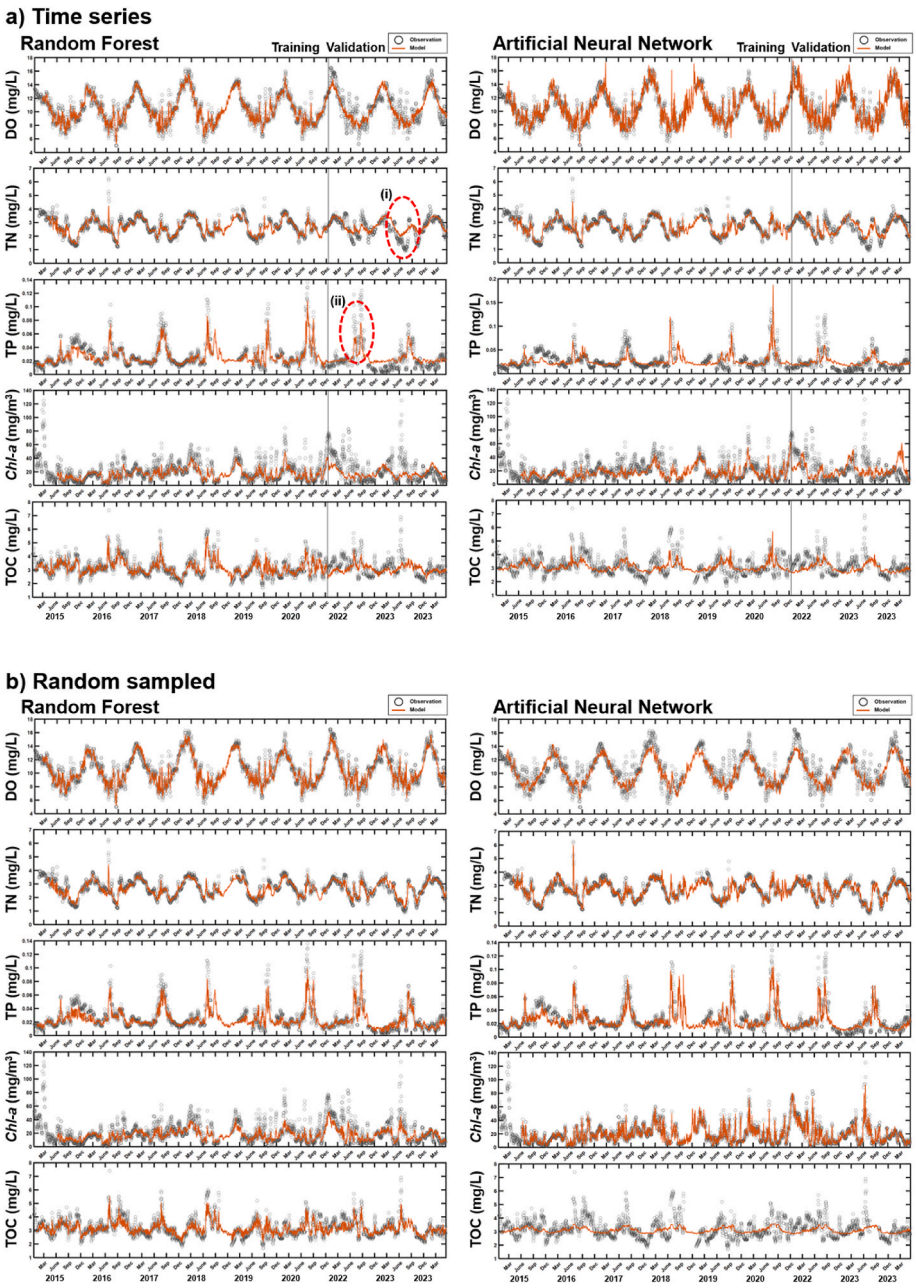

Fig. 6. Results of water quality simulation at the Dasan station utilizing random forest (RF) and artificial neural network (ANN) models. The left and right figures show the results of the RF and ANN models, respectively. a) represents simulations utilizing time-series data and b) represents simulations utilizing randomly sampled data.

4. Discussion

4.1. Prediction performance based on data sampling methods and ML algorithms

The difference in model performance between the time series and randomly sampled data stems from the varying distributions of the training data. Simulations using time-series data exhibited larger discrepancies between the training and validation R^2 values compared to those using randomly sampled data. Differences in the distributions of the predicted variables between the training and testing datasets may decrease the reliability of the ML model. Bjerre et al. (2022) demonstrated that randomly sampled data with similar distributions exhibit higher performance. While using randomly sampled data instead of time-series data, a performance improvement was observed in simulations of all water quality parameters by ensuring a similar distribution between the training and validation sets. Previous studies have adopted randomly sampled data to ensure better model performances (Barzegar et al., 2018; Djarum et al., 2023; Kim, K. and Ahn, 2022; Snieder and Khan, 2023).

The differences in performance among machine learning models for water quality prediction are primarily due to variations in each algorithm's structure and learning methodology, which become especially pronounced in water quality modeling. Ensemble models like RF combine multiple decision trees to reflect data variability and reduce the impact of noise, resulting in higher predictive accuracy and stability (Azrour et al., 2022). This structure is advantageous for handling complex physical systems with interacting environmental variables, providing robust generalization while mitigating overfitting (Breiman, 2001). In contrast, ANN serve as powerful tools for capturing complex

Table 4
Determination coefficient (R^2), root mean squared error (RMSE), and Nash–Sutcliffe efficiency (NSE) for the validation of water quality simulations and predictions at the Dogae station.

Sampling methods	Water quality	Lead time	RF			ANN			SVM		
			R^2	RMSE	NSE	R^2	RMSE	NSE	R^2	RMSE	NSE
Time series	DO	Today	0.66	1.59	0.64	0.57	1.86	0.50	0.63	1.66	0.61
		1 day	0.65	1.61	0.63	0.52	2.01	0.42	0.64	1.65	0.61
		2 days	0.65	1.61	0.62	0.62	1.7	0.58	0.64	1.65	0.61
		3 days	0.65	1.59	0.63	0.54	1.93	0.46	0.63	1.66	0.60
	TN	Today	0.22	0.71	−0.19	0.19	0.85	−0.68	0.15	0.78	−0.41
		1 day	0.19	0.72	−0.21	0.18	0.78	−0.44	0.15	0.78	−0.42
		2 days	0.17	0.73	−0.23	0.17	0.88	−0.81	0.14	0.78	−0.43
		3 days	0.15	0.73	−0.25	0.17	0.69	−0.12	0.14	0.79	−0.45
	TP	Today	0.56	0.012	0.55	0.45	0.014	0.37	0.53	0.013	0.53
		1 day	0.56	0.012	0.56	0.53	0.013	0.51	0.53	0.013	0.52
		2 days	0.56	0.013	0.53	0.53	0.013	0.51	0.53	0.013	0.52
		3 days	0.57	0.012	0.57	0.54	0.013	0.52	0.53	0.013	0.52
	chl-a	Today	0.05	11.88	0.00	0.02	13.17	−0.23	0.002	13.41	−0.28
		1 day	0.03	12.08	−0.05	0.02	13.34	−0.28	0.002	13.33	−0.28
		2 days	0.03	12.10	−0.06	0.02	13.18	−0.25	0.001	13.45	−0.31
		3 days	0.03	11.95	−0.04	0.01	13.36	−0.30	0.001	13.48	−0.32
	TOC	Today	0.10	0.55	−0.20	0.10	0.54	−0.18	0.04	0.6	−0.45
		1 day	0.07	0.57	−0.27	0.07	0.57	−0.31	0.04	0.6	−0.43
		2 days	0.07	0.57	−0.30	0.08	0.53	−0.11	0.05	0.6	−0.43
		3 days	0.07	0.57	−0.27	0.06	0.56	−0.24	0.04	0.61	−0.47
Random sampled	DO	Today	0.78	1.2	0.77	0.68	1.45	0.67	0.73	1.31	0.73
		1 day	0.76	1.24	0.76	0.68	1.43	0.68	0.71	1.35	0.71
		2 days	0.75	1.23	0.75	0.71	1.33	0.71	0.72	1.29	0.72
		3 days	0.77	1.26	0.77	0.72	1.37	0.72	0.73	1.37	0.72
	TN	Today	0.53	0.6	0.47	0.41	0.65	0.39	0.34	0.69	0.31
		1 day	0.49	0.63	0.45	0.32	0.7	0.31	0.33	0.7	0.32
		2 days	0.50	0.61	0.45	0.38	0.66	0.36	0.32	0.69	0.30
		3 days	0.46	0.6	0.45	0.34	0.66	0.33	0.34	0.66	0.33
	TP	Today	0.78	0.009	0.75	0.71	0.01	0.69	0.73	0.01	0.72
		1 day	0.74	0.009	0.71	0.51	0.016	0.13	0.70	0.01	0.68
		2 days	0.73	0.01	0.70	0.68	0.011	0.65	0.69	0.011	0.66
		3 days	0.71	0.011	0.67	0.57	0.012	0.56	0.62	0.012	0.60
	chl-a	Today	0.23	11.49	0.15	0.07	12.5	0.00	0.12	12.08	0.06
		1 day	0.21	11.08	0.16	0.09	11.89	0.04	0.16	11.35	0.12
		2 days	0.19	12.21	0.13	0.05	13.23	−0.02	0.17	12.4	0.10
		3 days	0.20	10.57	0.17	0.06	11.48	0.02	0.18	10.75	0.14
	TOC	Today	0.41	0.61	0.40	0.31	0.66	0.31	0.27	0.68	0.26
		1 day	0.36	0.68	0.36	0.33	0.7	0.31	0.31	0.71	0.29
		2 days	0.38	0.62	0.38	0.33	0.66	0.31	0.34	0.65	0.32
		3 days	0.34	0.66	0.32	0.26	0.71	0.23	0.26	0.7	0.24

nonlinear relationships but are more sensitive to training data, often with a higher risk of overfitting; ANN models require substantial data for effective learning, and their performance can degrade significantly if data are insufficient or contain noise (Chanklan et al., 2018). SVM operates by setting regression boundaries to produce optimal predictions, performing effectively when data exhibit clear patterns and achieving reliable predictions through these boundaries (Isazadeh et al., 2017). However, SVM's sensitivity to data scaling and limitations in managing high-dimensional interactions make it less suited for complex, nonlinear water quality data. Thus, preprocessing and optimal parameter tuning are essential for SVM to reach stable performance; without these, its predictive accuracy can be inconsistent. Additionally, the inherent characteristics of water quality data—often nonlinear and prone to noise—significantly impact model performance. RF's structural robustness allows it to maintain high predictive performance even with noisy data, as it effectively manages inter-variable interactions (Azrour et al., 2022). Conversely, ANN and SVM models are more vulnerable to data irregularities; ANN, in particular, is prone to overfitting noisy data, reducing its generalization capacity, while SVM, being highly sensitive to data scale and parameter configuration, may struggle to deliver optimal performance unless data are carefully processed and calibrated (Isazadeh et al., 2017).

4.2. Prediction performance based on water quality

The prediction performance of DO, TN, and TP is comparatively higher due to their strong correlations with measurable physical and chemical processes. DO predictions are notably accurate when variables such as water temperature, BOD, and atmospheric diffusion are included, reflecting well-characterized seasonal and diurnal dynamics (Haider et al., 2013; Rajwa-Kuligiewicz et al., 2015). Similarly, TN and TP concentrations are strongly linked to nutrient inputs, such as agricultural runoff, and hydrological patterns, including rainfall and soil erosion. Incorporating variables such as precipitation, land-use patterns, and nutrient application rates significantly enhances model performance for TN and TP predictions (De Jager and Houser, 2012; Ha et al., 2020). The relatively linear dynamics of these parameters simplify their modeling compared to more complex indicators like TOC and Chl-a.

TOC and Chl-a are challenging water quality parameters to predict due to their nonlinear behavior and dependence on multifaceted interactions. TOC levels are shaped by both natural processes, such as microbial activity, organic matter decomposition, soil runoff, and anthropogenic factors, including wastewater discharge and agricultural drainage (Akhtar et al., 2021; Dar et al., 2022). Similarly, Chl-a is governed by complex ecological, physical, and chemical processes. Factors such as photosynthesis rates, nutrient availability, water temperature, and light penetration interact dynamically to influence Chl-a levels (Chen et al., 2015; Geider et al., 1998). Both parameters also

Table 5
Determination coefficient (R^2), root mean squared error (RMSE), and Nash–Sutcliffe efficiency (NSE) for the validation of water quality simulations and predictions at the Dasan station.

Sampling methods	Water quality	Lead time	RF			ANN			SVM		
			R^2	RMSE	NSE	R^2	RMSE	NSE	R^2	RMSE	NSE
Time series	DO	Today	0.73	1.26	0.73	0.66	1.52	0.60	0.76	1.19	0.76
		1 day	0.74	1.23	0.74	0.71	1.63	0.54	0.76	1.19	0.76
		2 days	0.75	1.21	0.75	0.64	1.49	0.62	0.76	1.18	0.76
		3 days	0.75	1.22	0.74	0.73	1.29	0.71	0.75	1.19	0.75
	TN	Today	0.56	0.5	0.51	0.57	0.47	0.57	0.52	0.49	0.51
		1 day	0.52	0.52	0.47	0.50	0.53	0.45	0.52	0.49	0.51
		2 days	0.56	0.49	0.51	0.55	0.48	0.54	0.52	0.49	0.51
		3 days	0.54	0.51	0.48	0.53	0.51	0.49	0.52	0.49	0.52
	TP	Today	0.71	0.013	0.58	0.73	0.014	0.53	0.62	0.015	0.44
		1 day	0.62	0.013	0.56	0.71	0.014	0.51	0.61	0.015	0.45
		2 days	0.70	0.013	0.57	0.74	0.011	0.69	0.61	0.015	0.45
		3 days	0.70	0.013	0.57	0.71	0.014	0.50	0.61	0.015	0.46
	chl-a	Today	0.15	16.6	0.09	0.03	18.71	−0.15	0.04	17.68	−0.03
		1 day	0.12	16.75	0.07	0.04	17.88	−0.06	0.04	17.66	−0.04
		2 days	0.14	16.48	0.09	0.03	18.4	−0.14	0.04	17.56	−0.04
		3 days	0.12	16.49	0.08	0.03	18.68	−0.18	0.04	17.52	−0.04
	TOC	Today	0.11	0.61	0.04	0.18	0.61	0.05	0.14	0.61	0.06
		1 day	0.10	0.62	0.01	0.20	0.61	0.04	0.14	0.61	0.06
		2 days	0.10	0.62	0.02	0.15	0.61	0.04	0.13	0.61	0.06
		3 days	0.11	0.62	0.03	0.12	0.61	0.05	0.12	0.61	0.05
Random sampled	DO	Today	0.84	0.91	0.84	0.78	1.1	0.77	0.81	0.99	0.81
		1 day	0.82	0.97	0.82	0.76	1.1	0.76	0.79	1.05	0.79
		2 days	0.80	0.99	0.80	0.74	1.14	0.73	0.79	1	0.79
		3 days	0.78	1.04	0.78	0.77	1.07	0.77	0.77	1.09	0.76
	TN	Today	0.73	0.35	0.72	0.74	0.34	0.74	0.65	0.39	0.65
		1 day	0.73	0.35	0.71	0.67	0.39	0.65	0.62	0.41	0.62
		2 days	0.72	0.36	0.71	0.74	0.35	0.73	0.62	0.41	0.62
		3 days	0.71	0.37	0.70	0.78	0.32	0.78	0.64	0.41	0.63
	TP	Today	0.73	0.014	0.57	0.66	0.013	0.63	0.50	0.016	0.44
		1 day	0.71	0.013	0.60	0.65	0.013	0.61	0.54	0.015	0.45
		2 days	0.70	0.013	0.59	0.47	0.015	0.44	0.55	0.015	0.48
		3 days	0.67	0.011	0.61	0.42	0.014	0.39	0.50	0.013	0.48
	chl-a	Today	0.55	11.52	0.42	0.64	9.33	0.62	0.32	13.01	0.26
		1 day	0.51	12.39	0.35	0.63	9.49	0.62	0.28	13.65	0.21
		2 days	0.55	11	0.42	0.58	9.52	0.56	0.32	12.2	0.28
		3 days	0.51	11.82	0.36	0.22	14.61	0.02	0.28	13.11	0.21
	TOC	Today	0.50	0.55	0.42	0.27	0.65	0.18	0.28	0.62	0.24
		1 day	0.52	0.47	0.45	0.26	0.56	0.24	0.32	0.54	0.30
		2 days	0.49	0.52	0.41	0.22	0.6	0.22	0.25	0.59	0.24
		3 days	0.48	0.53	0.40	0.23	0.61	0.21	0.26	0.6	0.23

exhibit significant spatial and temporal variability, particularly during episodic events like heavy rainfall, which amplify the influence of topographical and hydrodynamic conditions (HaRa et al., 2020; Zhang and Blomquist, 2018). Accurate modeling of TOC requires the inclusion of multidimensional variables like pH, microbial respiration rates, and DO, while Chl-a prediction benefits from detailed ecological data, including nutrient cycling and algal growth dynamics (Dimberg and Bryhn, 2015; Han et al., 2024; Jiang et al., 2022). The intricate dependencies and intertwined behaviors of TOC and Chl-a underscore the need for advanced interdisciplinary approaches to capture their variability and achieve reliable predictions.

Predicting the levels of Chl-a is particularly challenging due to the intricate dynamics of various algae types, including green algae, diatoms, and cyanobacteria. These organisms exhibit distinct responses to environmental factors such as temperature, light, and nutrient availability from human activities (Liu, X. et al., 2019; Sterner and Grover, 1998). Li et al. (2024) and Nelson et al. (2020) demonstrate that predicting Chl-a with process-based models is a challenging task. To improve Chl-a prediction accuracy using machine learning, it is essential to incorporate variables such as temperature, solar intensity, TN, TP, and TOC, as well as detailed ecological interactions like predator-prey dynamics (Baird et al., 2001; Kim, K. and Ahn, 2022; Park, Y. et al., 2021). However, limitations in real-time data acquisition often restrict the availability of these variables, which can diminish the accuracy of Chl-a predictions. While other parameters such as DO, TN, and TP display predictable patterns with environmental variables, Chl-a's biological complexity necessitates a broader set of inputs to achieve reliable predictions.

4.3. Flow rate and water quality simulation using API and ML

The simulated flow rate demonstrated reasonable performance, falling within the "good" and "satisfactory" (e.g., Doage station) criteria (Moriasi et al., 2015). However, the model performance for the flow rate was slightly lower than that reported in previous studies. For instance, Elbeltagi et al. (2022) achieved an R^2 value of 0.887 in flow rate simulations using M5 pruned, random subspace, RF, and bagging algorithms. Similarly, Dehghani and Poudeh (2020) used an SVM for river flow estimation and achieved a correlation coefficient of 0.970. The most significant differences between our study and the previous studies were the composition of the input data and the algorithm adopted during the simulations. Previous studies included flow rate data in their models. This inclusion likely contributed to the higher performance observed in those previous studies because Elbeltagi et al. (2022) and Dehghani and Poudeh (2020) used similar ML algorithms.

Moon et al. (2022) simulated the DO concentration in the Hwanggugicheon River using RF, Gradient Boosting (GB), and AdaBoost algorithms, obtaining R^2 values of 0.89, 0.79, and 0.90, respectively. Among these algorithms, AdaBoost showed the best performance, with an R^2 value of 0.912 after optimization. Both this study and Moon et al. (2022)

Table 6
Determination coefficient (R^2), root mean squared error (RMSE), and Nash–Sutcliffe efficiency (NSE) for the validation of water quality simulations and predictions at the Chungam station.

Sampling methods	Water quality	Lead time	RF			ANN			SVM		
			R^2	RMSE	NSE	R^2	RMSE	NSE	R^2	RMSE	NSE
Time series	DO	Today	0.78	1.41	0.71	0.73	1.47	0.69	0.76	1.36	0.73
		1 day	0.74	1.44	0.70	0.72	1.49	0.68	0.76	1.36	0.73
		2 days	0.73	1.5	0.67	0.66	1.59	0.63	0.76	1.37	0.73
		3 days	0.76	1.43	0.70	0.75	1.44	0.70	0.76	1.37	0.73
	TN	Today	0.55	0.49	0.45	0.58	0.48	0.46	0.49	0.5	0.42
		1 day	0.60	0.46	0.50	0.55	0.48	0.47	0.49	0.5	0.42
		2 days	0.61	0.46	0.50	0.57	0.45	0.52	0.50	0.5	0.42
		3 days	0.61	0.46	0.51	0.57	0.45	0.53	0.50	0.5	0.42
	TP	Today	0.73	0.012	0.68	0.36	0.013	0.59	0.59	0.014	0.58
		1 day	0.72	0.012	0.68	0.61	0.014	0.57	0.60	0.014	0.59
		2 days	0.71	0.012	0.69	0.53	0.015	0.51	0.61	0.013	0.59
		3 days	0.71	0.012	0.69	0.50	0.015	0.51	0.61	0.013	0.59
	chl-a	Today	0.10	17.08	0.08	0.10	17.16	0.08	0.02	18.21	−0.04
		1 day	0.07	17.36	0.05	0.05	17.89	0.00	0.02	18.27	−0.05
		2 days	0.07	17.41	0.05	0.06	17.99	−0.02	0.02	18.28	−0.05
		3 days	0.07	17.47	0.04	0.13	17.73	0.01	0.02	18.27	−0.05
	TOC	Today	0.26	0.55	0.08	0.20	0.63	−0.21	0.18	0.59	−0.06
		1 day	0.25	0.56	0.07	0.12	0.66	−0.31	0.18	0.6	−0.09
		2 days	0.25	0.56	0.07	0.16	0.69	−0.42	0.17	0.62	−0.14
		3 days	0.25	0.56	0.07	0.27	0.58	−0.01	0.17	0.61	−0.13
Random sampled	DO	Today	0.83	1.1	0.81	0.82	1.06	0.82	0.79	1.15	0.79
		1 day	0.82	1.01	0.82	0.71	1.3	0.70	0.75	1.18	0.75
		2 days	0.82	1.04	0.81	0.73	1.3	0.71	0.75	1.23	0.74
		3 days	0.82	1.1	0.81	0.73	1.31	0.73	0.73	1.33	0.72
	TN	Today	0.72	0.41	0.70	0.59	0.5	0.55	0.56	0.5	0.55
		1 day	0.70	0.41	0.69	0.65	0.44	0.64	0.55	0.49	0.55
		2 days	0.70	0.42	0.68	0.55	0.51	0.54	0.56	0.5	0.55
		3 days	0.72	0.39	0.71	0.59	0.46	0.59	0.59	0.47	0.58
	TP	Today	0.75	0.01	0.73	0.65	0.011	0.64	0.65	0.011	0.64
		1 day	0.76	0.012	0.70	0.66	0.014	0.60	0.63	0.014	0.61
		2 days	0.76	0.011	0.71	0.67	0.012	0.66	0.62	0.013	0.60
		3 days	0.77	0.011	0.73	0.37	0.019	0.14	0.58	0.014	0.56
	chl-a	Today	0.56	12.26	0.44	0.24	14.81	0.18	0.18	15	0.15
		1 day	0.58	12.7	0.43	0.17	16.32	0.06	0.24	15.44	0.16
		2 days	0.58	12.79	0.43	0.24	15.29	0.18	0.14	15.84	0.12
		3 days	0.55	11.58	0.46	0.28	13.55	0.26	0.22	14.23	0.18
	TOC	Today	0.46	0.47	0.45	0.35	0.51	0.35	0.32	0.52	0.32
		1 day	0.44	0.51	0.42	0.33	0.57	0.28	0.33	0.55	0.31
		2 days	0.47	0.49	0.42	0.30	0.55	0.27	0.29	0.56	0.26
		3 days	0.45	0.48	0.44	0.33	0.53	0.33	0.30	0.54	0.30

demonstrated accurate performance with RF simulations within the "very good" and "good" criteria; the previous study showed slightly higher performance (Moriasi et al., 2015). The key difference between our study and the aforementioned one lies in the composition of the input data. Moon et al. (2022) used water quality data such as pH, SS, water temperature, TN, DTP, NH_3-N, COD, DTN, and NO_3-N as inputs, whereas our study used weather data. The fact that "good" performance was achieved in our study with limited input data demonstrates the potential of simulating DO using weather data alone. This is particularly noteworthy because temperature, which follows a certain pattern, is a major factor influencing DO concentration in river water (Haider et al., 2013; Moriasi et al., 2015; Rajwa-Kuligiewicz et al., 2015).

The model performances for TN and TP were slightly higher than those reported by Adedeji et al. (2022) simulated TN and TP based on four scenarios using SVM, RF, Extreme Gradient Boosting (XGB), combined RF-XGB, and ANN. Their scenarios differentiated between types of input data, achieving good model performance with R^2 for TN ranging from 0.65 to 0.74 and for TP from 0.65 to 0.79. A major difference between our study and the aforementioned one is the diversity of the input data. Adedeji et al. (2022) used meteorological data and data on land use/land cover, pedology, animal statistics, antecedent conditions, and stream water quality. The ability to achieve similar or better performance using only weather data in this study underscores the potential for improved data efficiency in simulating TN and TP.

The model performance for *Chl-a* was lower compared to a previous study by Park et al. (2015) simulated *Chl-a* concentrations in the Yeongsan and Juam rivers using ANN and SVM. They achieved R^2 values of 0.74 and 0.75 with ANN and SVM, respectively, on the Juam River and R^2 values of 0.43 and 0.45 on the Yeongsan River, respectively. The primary difference between our study and the aforementioned one was the inclusion of nitrogen and phosphorus information in the input data. Algal growth, as indicated by *Chl-a* levels, is influenced by temperature, solar intensity, and nutrient concentrations (Baird et al., 2001), with nutrient concentrations being particularly crucial (Tilman et al., 1982). Therefore, incorporating nutritional information into the input data appears to enhance the model's ability to predict *Chl-a* levels.

The model performance for TOC was lower compared to a previous study by Baek et al. (2020) simulated TOC using a combined convolutional neural network-long short-term memory (CNN-LSTM) deep learning approach. They obtained R^2 values of 0.86 and 0.79 for training and validation, respectively. The most significant difference between our study and the aforementioned one was the use of machine or deep learning techniques and the types of input data used. While our study only used meteorological data, Baek et al. (2020) incorporated a broader range of input data, such as precipitation radar images, water levels, operation information of estuary barrages, meteorological data, and water quality data. Considering the complexity of the relationship between TOC concentration and environmental factors, this demonstrates that input data must be diverse because various factors such as pH, nutrient levels, and runoff can influence TOC concentration (Jiang et al.,

Table 7
Determination coefficient (R^2), root mean squared error (RMSE), and Nash–Sutcliffe efficiency (NSE) for the validation of water quality simulations and predictions at the Sangdong station.

Sampling methods	Water quality	Lead time	RF			ANN			SVM		
			R^2	RMSE	NSE	R^2	RMSE	NSE	R^2	RMSE	NSE
Time series	DO	Today	0.43	1.66	0.28	0.38	1.58	0.35	0.32	1.81	0.14
		1 day	0.39	1.7	0.24	0.31	1.75	0.20	0.32	1.83	0.13
		2 days	0.38	1.76	0.19	0.41	1.53	0.39	0.31	1.86	0.10
		3 days	0.39	1.72	0.23	0.36	1.63	0.31	0.31	1.87	0.09
	TN	Today	0.41	0.5	0.26	0.31	0.5	0.25	0.29	0.52	0.20
		1 day	0.38	0.51	0.22	0.32	0.51	0.22	0.28	0.53	0.17
		2 days	0.39	0.51	0.23	0.30	0.51	0.21	0.26	0.54	0.14
		3 days	0.37	0.51	0.22	0.20	0.53	0.17	0.26	0.54	0.12
	TP	Today	0.47	0.012	0.44	0.27	0.015	0.19	0.33	0.014	0.30
		1 day	0.45	0.013	0.42	0.29	0.015	0.21	0.33	0.014	0.31
		2 days	0.44	0.013	0.41	0.28	0.015	0.19	0.32	0.014	0.31
		3 days	0.45	0.013	0.41	0.29	0.014	0.24	0.33	0.014	0.32
	chl-a	Today	0.23	13.68	0.23	0.16	14.33	0.15	0.13	14.57	0.12
		1 day	0.21	13.88	0.20	0.18	14.13	0.17	0.13	14.59	0.12
		2 days	0.19	14.04	0.18	0.08	15.05	0.06	0.13	14.6	0.12
		3 days	0.16	14.24	0.16	0.21	14.01	0.19	0.13	14.6	0.12
	TOC	Today	0.12	0.51	−0.05	0.03	0.58	−0.32	0.04	0.52	−0.10
		1 day	0.10	0.52	−0.07	0.05	0.56	−0.24	0.04	0.52	−0.09
		2 days	0.09	0.53	−0.12	0.11	0.49	0.06	0.04	0.52	−0.08
		3 days	0.11	0.51	−0.04	0.01	0.58	−0.35	0.05	0.52	−0.08
Random sampled	DO	Today	0.69	1.23	0.66	0.50	1.5	0.50	0.59	1.37	0.58
		1 day	0.64	1.24	0.63	0.54	1.38	0.54	0.60	1.28	0.60
		2 days	0.64	1.29	0.63	0.54	1.45	0.53	0.56	1.4	0.56
		3 days	0.64	1.22	0.64	0.55	1.36	0.55	0.58	1.31	0.58
	TN	Today	0.61	0.46	0.59	0.31	0.6	0.30	0.49	0.51	0.49
		1 day	0.55	0.47	0.55	0.52	0.49	0.52	0.49	0.5	0.49
		2 days	0.55	0.49	0.52	0.32	0.59	0.31	0.43	0.53	0.43
		3 days	0.52	0.53	0.47	0.48	0.53	0.47	0.51	0.51	0.50
	TP	Today	0.70	0.012	0.59	0.53	0.013	0.50	0.47	0.014	0.41
		1 day	0.70	0.013	0.60	0.40	0.016	0.36	0.45	0.015	0.41
		2 days	0.71	0.014	0.57	0.54	0.014	0.54	0.49	0.016	0.40
		3 days	0.69	0.013	0.57	0.46	0.015	0.41	0.47	0.015	0.39
	chl-a	Today	0.46	14.36	0.33	0.46	13.23	0.43	0.23	15.81	0.18
		1 day	0.50	15.25	0.33	0.44	14.49	0.39	0.23	17.03	0.16
		2 days	0.43	14.17	0.29	0.44	13.24	0.38	0.29	14.84	0.23
		3 days	0.49	14.87	0.35	0.47	14.33	0.40	0.28	16.19	0.23
	TOC	Today	0.56	0.41	0.52	0.46	0.44	0.45	0.32	0.49	0.32
		1 day	0.51	0.42	0.49	0.34	0.49	0.32	0.36	0.48	0.35
		2 days	0.52	0.41	0.50	0.33	0.49	0.31	0.41	0.46	0.39
		3 days	0.47	0.45	0.45	0.46	0.45	0.45	0.36	0.49	0.33

2022; Lee, J. et al., 2016).

The evaluation results according to the forecast lead times are shown in Figs. S37–S42. In the RF model and simulations, R^2 decreased and RMSE increased in the validation section as the forecast lead time increased. However, in the ANN simulations, higher R^2 and lower RMSE were observed for longer forecast lead times, which may be attributed to overtraining. This trend indicates that prediction accuracy decreases as forecast lead time increases. This observation is supported by the findings of Miao et al. (2019) and Baek et al. (2021a,b), who demonstrated that prediction accuracy decreases as forecast lead time increases.

4.4. Relationship between SA and UA

Among the limited input variables, temperature is the main factor controlling DO concentration in river water (Haider et al., 2013; Rajwa-Kuligiewicz et al., 2015). Furthermore, flow rate and temperature are the main factors controlling TN and TP concentrations (Xia et al., 2020; Yang, Y. and Jin, 2023), *Chl-a* concentration (Lee, S. et al., 2012; Li, F. et al., 2013), and TOC concentration (Köhler et al., 2009). In our study, the SA of RF showed reasonable results that reflected natural phenomena, except for the TN simulations using both time series and randomly sampled data (Figs. S46 and S51). However, the SA of the ANN and SVM showed conflicting results. The SA of the ANN models tended to select the wrong feature as an important feature, as evidenced by the DO and TP simulations using the ANN with time series at the Dogae and Dasan stations (Figs. S43–44). Conversely, the SA of the SVM simulation tended to produce the same importance rankings across all simulations, with rainfall, temperature, and insolation being consistently selected (Figs. S47 and S52). These results indicate that the RF models can effectively reflect natural phenomena, while the SA results vary depending on the algorithm used. Park et al. (2021) demonstrated that different algorithms may produce different perspectives in SA.

A strong relationship was observed between the SA and UA, whereas a weaker relationship was observed between uncertainty and simulation performance. According to the error rates of the input variables (Table S10), when a variable with a high error rate is selected as an important feature in the SA, its uncertainty increases. For instance, in the TN simulations with time-series data at the Dogae station using RF and ANN (Figs. S43, S53, and S57), the flow rate was identified as the most sensitive variable, resulting in significant uncertainty in the simulation. However, the overall uncertainty in the SVM simulations generally displayed a narrow range of uncertainty.

4.5. Future study

One of the main limitations of this study lies in the relatively lower predictive accuracy for chl-a and the potential for improvement across other water quality parameters. To address these limitations, future research will focus on stabilizing data and increasing the number of input variables to enhance model performance. Data stabilization will

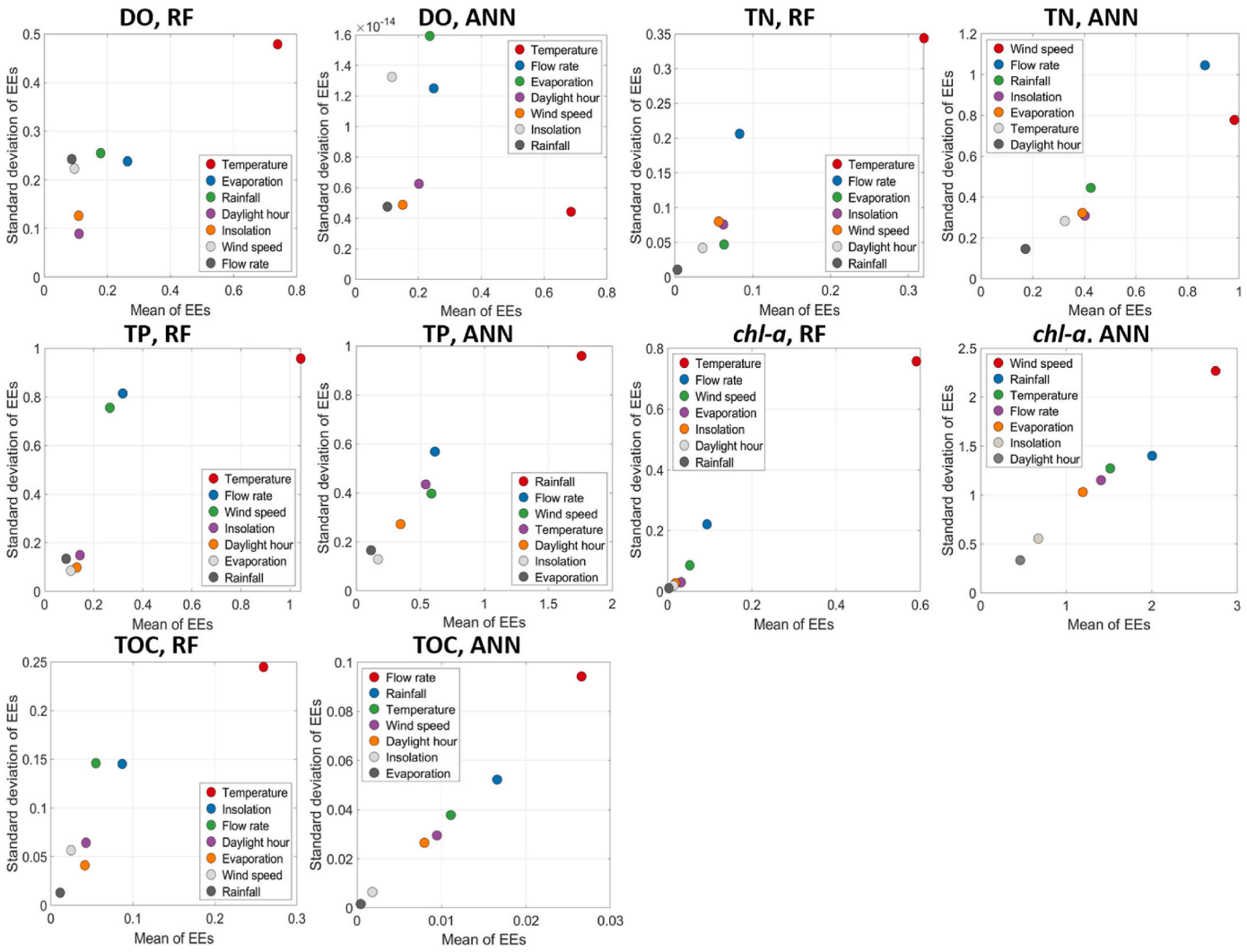

Fig. 7. Sensitivity analysis (SA) test results from the Dasan station utilizing random forest (RF) and artificial neural network (ANN) with randomly sampled data.

involve outlier analysis of sensor data, aiming to mitigate the uncertainty associated with sensor measurements. By reducing sensor data irregularities, model reliability, and consistency are expected to be improved. To diversify input data, future studies will either incorporate additional input variables by integrating numerical models, which will expand the range of available data, or develop artificial intelligence models to predict essential input variables, thereby further supporting model accuracy. Furthermore, to enhance predictive performance, especially for time-series learning, advanced algorithms such as LSTM and deep learning techniques specialized for complex patterns will be explored. This approach is expected to overcome current limitations and improve model adaptability and accuracy for water quality prediction.

5. Conclusion

In this study, the ANN, RF, and SVM algorithms were applied to develop a real-time river water quality simulation toolbox. In addition, to enhance the toolbox's usability and interpretability, API, GUI, SA, and UA were applied. For data acquisition, the API was used to acquire water quality and real-time weather data. For water quality prediction, stepwise modeling was carried out. Initially, flow-rate modeling was performed using RF, followed by water quality modeling performed using ANN, RF, and SVM, considering both time series and random sampling data. To assess model performance, input variable importance, and model uncertainty, SA and UA evaluations were conducted with R^2 and RMSE evaluation. Subsequently, the results were compared based on the data type and algorithm. The research efforts culminated in the successful development and function validation of the GUI toolbox, well beyond the research observation and simulation periods. Integrating the API with the GUI toolbox enhanced its utility and usability for real-time river water quality simulation tasks. From this research, the following conclusions were obtained: a) the integration of API ensured the stable acquisition of the latest data and enabled the effective utilization of large databases; b) in the environmental field, randomly sampled data demonstrated more effective performance compared to time-series data; c) among the ML algorithms evaluated, RF showed the best performance; d) while most of the water quality exhibited reliable performance, simulating *Chl-a* and TOC based solely on weather data had its limitations; e) the use of SA and UA is anticipated to contribute to a broad evaluation and enhanced understanding of the models; and f) building a GUI is expected to enhance user convenience.

CRediT authorship contribution statement

Gi-Hun Bang: Writing – original draft, Validation, Software, Methodology, Formal analysis. **Na-Hyeon Gwon:** Software, Formal analysis. **Min-Jeong Cho:** Validation, Formal analysis. **Ji-Ye Park:** Visualization, Formal analysis. **Sang-Soo Baek:** Writing – review & editing, Supervision, Methodology.

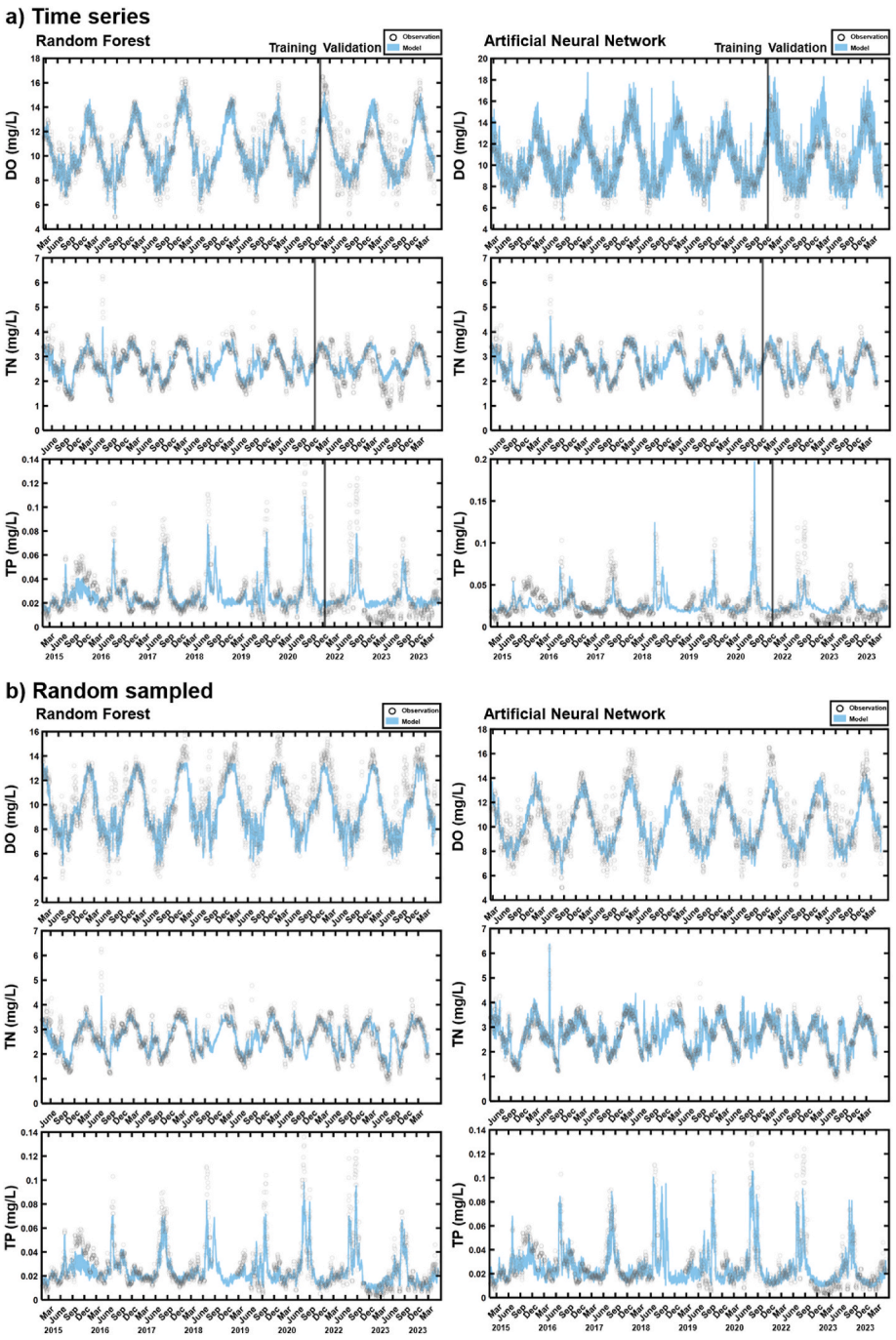

Fig. 8. Results of Generalized likelihood uncertainty estimation (GLUE) at Dasan station that utilized random forest (RF) and artificial neural network (ANN). The left and right figures show the results of the RF and ANN models, respectively. a) represents GLUE utilizing time-series data, and b) represents GLUE utilizing randomly sampled data.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Sang-Soo Baek reports financial support was provided by Yeungnam University. Sang-Soo Baek reports financial support was provided by National Research Foundation of Korea. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the 2022 Yeungnam University Research Grant and MSIT through Sejong Science Fellowship, funded by National Research Foundation of Korea (NRF) (No. RS-2022-NR072229).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvman.2025.124719.

Data availability

The authors do not have permission to share data.

References

Adams, H., Ye, J., Persaud, B.D., Slowinski, S., Kheyrollah Pour, H., Van Cappellen, P., 2022. Rates and timing of chlorophyll-a increases and related environmental variables in global temperate and cold-temperate lakes. Earth Syst. Sci. Data 14, 5139–5156.

Adedeji, I.C., Ahmadisharaf, E., Sun, Y., 2022. Predicting in-stream water quality constituents at the watershed scale using machine learning. J. Contam. Hydrol. 251, 104078.

Agatonovic-Kustrin, S., Beresford, R., 2000. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. J. Pharm. Biomed. Anal. 22, 717–727.

Ahmed, A.N., Othman, F.B., Afan, H.A., Ibrahim, R.K., Fai, C.M., Hossain, M.S., Ehteram, M., Elshafie, A., 2019. Machine learning methods for better water quality prediction. J. Hydrol. 578, 124084.

Akhtar, N., Syakir Ishak, M.I., Bhawani, S.A., Umar, K., 2021. Various natural and anthropogenic factors responsible for water quality degradation: a review. Water 13, 2660.

Aldhyani, T.H., Al-Yaari, M., Alkahtani, H., Maashi, M., 2020. Water quality prediction using artificial intelligence algorithms. Applied Bionics and Biomechanics 2020.

Alnahit, A.O., Mishra, A.K., Khan, A.A., 2022. Stream water quality prediction using boosted regression tree and random forest models. Stoch. Environ. Res. Risk Assess. 36, 2661–2680.

Amador-Castro, F., González-López, M.E., Lopez-Gonzalez, G., Garcia-Gonzalez, A., Díaz-Torres, O., Carbajal-Espinosa, O., Gradilla-Hernández, M.S., 2024. Internet of Things and citizen science as alternative water quality monitoring approaches and the importance of effective water quality communication. J. Environ. Manag. 352, 119959. https://doi.org/10.1016/j.jenvman.2023.119959.

Atique, U., An, K., 2019. Reservoir water quality assessment based on chemical parameters and the chlorophyll dynamics in relation to nutrient regime. Pol. J. Environ. Stud. 28.

Azrour, M., Mabrouki, J., Fattah, G., Guezzaz, A., Aziz, F., 2022. Machine learning algorithms for efficient water quality prediction. Model. Earth Syst. Environ. 8, 2793–2801.

Baek, S., Choi, Y., Jeon, J., Pyo, J., Park, J., Cho, K.H., 2021a. Replacing the internal standard to estimate micropollutants using deep and machine learning. Water Res. 188, 116535. https://doi.org/10.1016/j.watres.2020.116535.

Baek, S., Jung, E., Pyo, J., Pachepsky, Y., Son, H., Cho, K.H., 2022. Hierarchical deep learning model to simulate phytoplankton at phylum/class and genus levels and zooplankton at the genus level. Water Res. 218, 118494.

Baek, S., Pyo, J., Chun, J.A., 2020. Prediction of water level and water quality using a CNN-LSTM combined deep learning approach. Water 12, 3399.

Baek, S., Pyo, J., Kwon, Y.S., Chun, S., Baek, S.H., Ahn, C., Oh, H., Kim, Y.O., Cho, K.H., 2021b. Deep learning for simulating harmful algal blooms using ocean numerical model. Front. Mar. Sci. 8, 729954.

Baird, M.E., Emsley, S.M., Mcglade, J.M., 2001. Modelling the interacting effects of nutrient uptake, light capture and temperature on phytoplankton growth. J. Plankton Res. 23, 829–840.

Balabin, R.M., Lomakina, E.I., 2011. Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data. Analyst 136, 1703–1712.

Barzegar, R., Moghaddam, A.A., Deo, R., Fijani, E., Tziritis, E., 2018. Mapping groundwater contamination risk of multiple aquifers using multi-model ensemble of machine learning algorithms. Sci. Total Environ. 621, 697–712.

Beck, M.B., 1987. Water quality modeling: a review of the analysis of uncertainty. Water Resour. Res. 23, 1393–1442.

Behnood, A., Daneshvar, D., 2020. A machine learning study of the dynamic modulus of asphalt concretes: an application of M5P model tree algorithm. Constr. Build. Mater. 262, 120544.

Bjerre, E., Fienen, M.N., Schneider, R., Koch, J., Højberg, A.L., 2022. Assessing spatial transferability of a random forest metamodel for predicting drainage fraction. J. Hydrol. 612, 128177.

Bonsiepe, G., 1990. Interface design-language-graphics: interpretations of human user interface. Visible Lang. 24, 262.

Breiman, L., 2001. Random forests. Mach. Learning 45, 5–32.

Cabello-Solorzano, K., Ortigosa de Araujo, I., Peña, M., Correia, L., J Tallón-Ballesteros, A., 2023. The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis, pp. 344–353.

Calhoun, P., Su, X., Spoon, K.M., Levine, R.A., Fan, J., 2014. Random Forest. Wiley StatsRef: Statistics Reference Online, pp. 1–20.

Campolongo, F., Cariboni, J., Saltelli, A., 2007. An effective screening design for sensitivity analysis of large models. Environ. Model. Software 22, 1509–1518. https://doi.org/10.1016/j.envsoft.2006.10.004.

Campolongo, F., Saltelli, A., Cariboni, J., 2011. From screening to quantitative sensitivity analysis. A unified approach. Comput. Phys. Commun. 182, 978–988. https://doi.org/10.1016/j.cpc.2010.12.039.

Castrillo, M., García, Á.L., 2020. Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods. Water Res. 172, 115490.

Chang, F., Tsai, Y., Chen, P., Coynel, A., Vachaud, G., 2015. Modeling water quality in an urban river using hydrological factors–Data driven approaches. J. Environ. Manag. 151, 87–96.

Chanklan, R., Kaoungku, N., Suksut, K., Kerdprasop, K., Kerdprasop, N., 2018. Runoff prediction with a combined artificial neural network and support vector regression. Int. J. Mach. Learn. Comput. 8, 39–43.

Chen, M., Fan, M., Liu, R., Wang, X., Yuan, X., Zhu, H., 2015. The dynamics of temperature and light on the growth of phytoplankton. J. Theor. Biol. 385, 8–19.

Chen, Q., Wu, W., Blanckaert, K., Ma, J., Huang, G., 2012. Optimization of water quality monitoring network in a large river by combining measurements, a numerical model and matter-element analyses. J. Environ. Manag. 110, 116–124. https://doi.org/10.1016/j.jenvman.2012.05.024.

Chen, Y., Song, L., Liu, Y., Yang, L., Li, D., 2020. A review of the artificial neural network models for water quality prediction. Appl. Sci. 10, 5776.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

Cui, F., Park, C., Kim, M., 2019. Application of curve-fitting techniques to develop numerical calibration procedures for a river water quality model. J. Environ. Manag. 249, 109375. https://doi.org/10.1016/j.jenvman.2019.109375.

Dar, S.A., Bhat, S.U., Rashid, I., Kumar, P., Sharma, R., Aneaus, S., 2022. Deciphering the source contribution of organic matter accumulation in an urban wetland ecosystem. Land Degrad. Dev. 33, 2390–2404.

De Jager, N.R., Houser, J.N., 2012. Variation in water-mediated connectivity influences patch distributions of total N, total P, and TN: TP ratios in the Upper Mississippi River, USA. Freshw. Sci. 31, 1254–1272.

Dehghani, R., Poudeh, H.T., 2020. Application of support vector machine for river flow estimation. Int. J. Eng. Bus. Manag. 4.

Dimberg, P.H., Bryhn, A.C., 2015. Predicting total nitrogen, total phosphorus, total organic carbon, dissolved oxygen and iron in deep waters of Swedish lakes. Environ. Model. Assess. 20, 411–423.

Djarum, D.H., Ahmad, Z., Zhang, J., 2023. Reduced Bayesian Optimized Stacked Regressor (RBOSR): a highly efficient stacked approach for improved air pollution prediction. Appl. Soft Comput. 144, 110466.

Elbeltagi, A., Di Nunno, F., Kushwaha, N.L., de Marinis, G., Granata, F., 2022. River flow rate prediction in the Des Moines watershed (Iowa, USA): a machine learning approach. Stoch. Environ. Res. Risk Assess. 36, 3835–3855.

Geider, R.J., Maclntyre, H.L., Kana, T.M., 1998. A dynamic regulatory model of phytoplanktonic acclimation to light, nutrients, and temperature. Limnol. Oceanogr. 43, 679–694.

Gómez, D., Salvador, P., Sanz, J., Casanova, J.L., 2021. A new approach to monitor water quality in the Menor sea (Spain) using satellite data and machine learning methods. Environ. Pollut. 286, 117489.

Gunn, S.R., 1998. Support vector machines for classification and regression. In: ISIS technical report, 14, pp. 5–16.

Guo, H., Huang, J.J., Zhu, X., Wang, B., Tian, S., Xu, W., Mai, Y., 2021. A generalized machine learning approach for dissolved oxygen estimation at multiple spatiotemporal scales using remote sensing. Environ. Pollut. 288, 117734. https://doi.org/10.1016/j.envpol.2021.117734.

Ha, N., Nguyen, H.Q., Truong, N.C.Q., Le, T.L., Thai, V.N., Pham, T.L., 2020. Estimation of nitrogen and phosphorus concentrations from water quality surrogates using machine learning in the Tri an Reservoir, Vietnam. Environ. Monit. Assess. 192, 1–20.

Haider, H., Ali, W., Haydar, S., 2013. Evaluation of various relationships of reaeration rate coefficient for modeling dissolved oxygen in a river with extreme flow variations in Pakistan. Hydrol. Process. 27, 3949–3963.

Han, J.W., Kim, T., Lee, S., Kang, T., Im, J.K., 2024. Machine learning and explainable AI for chlorophyll-a prediction in Namhan River Watershed, South Korea. Ecol. Indic. 166, 112361. https://doi.org/10.1016/j.ecolind.2024.112361.

HaRa, J., Atique, U., An, K., 2020. Multiyear links between water chemistry, algal chlorophyll, drought-flood regime, and nutrient enrichment in a morphologically complex reservoir. Int. J. Environ. Res. Publ. Health 17, 3139.

Huang, J., Xu, J., Xia, Z., Liu, L., Zhang, Y., Li, J., Lan, G., Qi, Y., Kamon, M., Sun, X., Li, Y., 2015. Identification of influential parameters through sensitivity analysis of the TOUGH + Hydrate model using LH-OAT sampling. Mar. Petrol. Geol. 65, 141–156. https://doi.org/10.1016/j.marpetgeo.2015.04.009.

Hur, M., Lee, I., Tak, B., Lee, H.J., Yu, J.J., Cheon, S.U., Kim, B., 2013. Temporal shifts in cyanobacterial communities at different sites on the Nakdong River in Korea. Water Res. 47, 6973–6982.

Hutter, F., Lücke, J., Schmidt-Thieme, L., 2015. Beyond manual tuning of hyperparameters. KI-Künstliche Intell. 29, 329–337.

Isazadeh, M., Biazar, S.M., Ashrafzadeh, A., 2017. Support vector machines and feed-forward neural networks for spatial modeling of groundwater qualitative parameters. Environ. Earth Sci. 76, 1–14.

Jansen, B.J., 1998. The graphical user interface. ACM SIGCHI Bull. 30, 22–26.

Jeung, M., Jang, M., Shin, K., Jung, S.W., Baek, S., 2024. Graph neural networks and transfer entropy enhance forecasting of mesozooplankton community dynamics. Environ. Sci. Ecotechnol., 100514 https://doi.org/10.1016/j.ese.2024.100514.

Jiang, Y., He, K., Li, Y., Qin, M., Cui, Z., Zhang, Y., Yao, Y., Chen, X., Deng, M., Gray, A., 2022. Driving factors of total organic carbon in Danjiangkou reservoir using generalized additive model. Water 14, 891.

Jones, A.S., Stevens, D.K., Horsburgh, J.S., Mesner, N.O., 2011. Surrogate measures for providing high frequency estimates of total suspended solids and total phosphorus concentrations 1. JAWRA J. Am. Water Resour. Assoc. 47, 239–253.

Jung, K.Y., Lee, K., Im, T.H., Lee, I.J., Kim, S., Han, K., Ahn, J.M., 2016. Evaluation of water quality for the Nakdong River watershed using multivariate analysis. Environ. Technol. Innov. 5, 67–82. https://doi.org/10.1016/j.eti.2015.12.001.

Khalilia, M., Chakraborty, S., Popescu, M., 2011. Predicting disease risks from highly imbalanced data using random forest. BMC Med. Inf. Decis. Making 11, 1–13.
Khan, S.J., Deere, D., Leusch, F.D.L., Humpage, A., Jenkins, M., Cunliffe, D., 2015. Extreme weather events: should drinking water quality management systems adapt to changing risk profiles? Water Res. 85, 124–136. https://doi.org/10.1016/j.watres.2015.08.018.
Khullar, S., Singh, N., 2021. Machine learning techniques in river water quality modelling: a research travelogue. Water Supply 21, 1–13.
Kim, K., Ahn, J., 2022. Machine learning predictions of chlorophyll-a in the Han river basin, Korea. J. Environ. Manag. 318, 115636. https://doi.org/10.1016/j.jenvman.2022.115636.
Kim, S., Kwon, Y.S., Pyo, J., Ligaray, M., Min, J., Ahn, J.M., Baek, S., Cho, K.H., 2021. Developing a cloud-based toolbox for sensitivity analysis of a water quality model. Environ. Model. Software 141, 105068. https://doi.org/10.1016/j.envsoft.2021.105068.
Kline, S.J., 1985. The Purposes of Uncertainty Analysis.
KMA, 2021. Climatological Normals of Korea.
Köhler, S.J., Buffam, I., Seibert, J., Bishop, K.H., Laudon, H., 2009. Dynamics of stream water TOC concentrations in a boreal headwater catchment: controlling factors and implications for climate scenarios. J. Hydrol. 373, 44–56.
Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I., 2015. Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. 13, 8–17.
Lamboni, M., Monod, H., Makowski, D., 2011. Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. Reliab. Eng. Syst. Saf. 96, 450–459.
Lee, J., Shin, K., Park, C., Lee, S., Kim, Y., Yu, S., 2016. Spatial and seasonal variations of organic carbon level in four major rivers in Korea. Environ. Eng. Res. 21, 84–90.
Lee, S., Lee, S., Kim, S.H., Park, H., Park, S., Yum, K., 2012. Examination of critical factors related to summer chlorophyll a concentration in the Sueo Dam reservoir, Republic of Korea. Environ. Eng. Sci. 29, 502–510.
Li, F., Zhang, H., Zhu, Y., Xiao, Y., Chen, L., 2013. Effect of flow velocity on phytoplankton biomass and composition in a freshwater lake. Sci. Total Environ. 447, 64–71.
Li, Y., 2022. Research and Application of Deep Learning in Image Recognition. In: 2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA), pp. 994–999. https://doi.org/10.1109/ICPECA53709.2022.9718847.
Li, H., Li, X., Song, D., Nie, J., Liang, S., 2024. Prediction on daily spatial distribution of chlorophyll-a in coastal seas using a synthetic method of remote sensing, machine learning and numerical modeling. Sci. Total Environ. 910, 168642.
Liu, S., Tai, H., Ding, Q., Li, D., Xu, L., Wei, Y., 2013. A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. Math. Comput. Model. 58, 458–465. https://doi.org/10.1016/j.mcm.2011.11.021.
Liu, X., Feng, J., Wang, Y., 2019. Chlorophyll a predictability and relative importance of factors governing lake phytoplankton at different timescales. Sci. Total Environ. 648, 472–480. https://doi.org/10.1016/j.scitotenv.2018.08.146.
Mahajan, M., Kumar, S., Pant, B., Tiwari, U.K., 2020. Incremental Outlier Detection in Air Quality Data Using Statistical Methods, 2020 International Conference on Data Analytics for Business and Industry: Way towards a Sustainable Economy (ICDABI), pp. 1–5. https://doi.org/10.1109/ICDABI51230.2020.9325683.
Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ. Model. Software 15, 101–124.
Makler-Pick, V., Gal, G., Gorfine, M., Hipsey, M.R., Carmel, Y., 2011. Sensitivity analysis for complex ecological models–a new approach. Environ. Model. Software 26, 124–134.
Masson, J., Biggins, J.S., Ringe, E., 2023. Machine learning for nanoplasmonics. Nat. Nanotechnol. 18, 111–123.
Miao, Q., Pan, B., Wang, H., Hsu, K., Sorooshian, S., 2019. Improving monsoon precipitation prediction using combined convolutional and long short term memory neural network. Water 11, 977.
Moon, J., Lee, J., Lee, S., Yun, H., 2022. Urban river dissolved oxygen prediction model using machine learning. Water 14, 1899.
Moriasi, D.N., Gitau, M.W., Pai, N., Daggupati, P., 2015. Hydrologic and water quality models: performance measures and evaluation criteria. Trans. ASABE 58, 1763–1785.
Morris, M.D., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics 33, 161–174.
Müller, A., Österlund, H., Marsalek, J., Viklander, M., 2020. The pollution conveyed by urban runoff: a review of sources. Sci. Total Environ. 709, 136125. https://doi.org/10.1016/j.scitotenv.2019.136125.
Nawi, N.M., Atomi, W.H., Rehman, M.Z., 2013. The effect of data pre-processing on optimized training of artificial neural networks. Procedia Technol. 11, 32–39.
Nelson, N.G., Munoz-Carpena, R., Phlips, E., 2020. Parameter uncertainty drives important incongruities between simulated chlorophyll-a and phytoplankton functional group dynamics in a mechanistic management model. Environ. Model. Software 129, 104708.
Nilsson, C., Renöfält, B.M., 2008. Linking flow regime and water quality in rivers: a challenge to adaptive catchment management. Ecol. Soc. 13.
Noori, N., Kalin, L., Isik, S., 2020. Water quality prediction using SWAT-ANN coupled approach. J. Hydrol. 590, 125220.
Nossent, J., Elsen, P., Bauwens, W., 2011. Sobol'sensitivity analysis of a complex environmental model. Environ. Model. Software 26, 1515–1525.
Ofoeda, J., Boateng, R., Effah, J., 2019. Application programming interface (API) research: a review of the past to inform the future. Int. J. Enterprise Inf. Syst. 15, 76–95.
O'Leary, B., Reiners, J.J., Xu, X., Lemke, L.D., 2016. Identification and influence of spatio-temporal outliers in urban air quality measurements. Sci. Total Environ. 573, 55–65. https://doi.org/10.1016/j.scitotenv.2016.08.031.
Pan, Y., Jiang, J., Wang, R., Cao, H., 2008. Advantages of support vector machine in QSPR studies for predicting auto-ignition temperatures of organic compounds. Chemometr. Intell. Lab. Syst. 92, 169–178.
Pang, C., Yu, J., Liu, Y., 2021. Correlation analysis of factors affecting wind power based on machine learning and Shapley value. IET Energy Syst. Integr. 3, 227–237.
Park, S.S., Lee, Y.S., 2002. A water quality modeling study of the Nakdong River, Korea. Ecol. Model. 152, 65–75.
Park, Y., Cho, K.H., Park, J., Cha, S.M., Kim, J.H., 2015. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. Sci. Total Environ. 502, 31–41.
Park, Y., Lee, H.K., Shin, J., Chon, K., Kim, S., Cho, K.H., Kim, J.H., Baek, S., 2021. A machine learning approach for early warning of cyanobacterial bloom outbreaks in a freshwater reservoir. J. Environ. Manag. 288, 112415.
Qin, Y., Du, J., Zhang, Y., Lu, H., 2019. Look Back and Predict Forward in Image Captioning, pp. 8367–8375.
Raatikainen, M., Kettunen, E., Salonen, A., Komssi, M., Mikkonen, T., Lehtonen, T., 2021. State of the practice in application programming interfaces (APIs): a case study, pp. 191–206.
Rajwa-Kuligiewicz, A., Bialik, R.J., Rowiński, P.M., 2015. Dissolved oxygen and water temperature dynamics in lowland rivers over various timescales. J. Hydrol. Hydromechanics 63, 353–363.
Rasaei, Z., Bogaert, P., 2019. Spatial filtering and Bayesian data fusion for mapping soil properties: a case study combining legacy and remotely sensed data in Iran. Geoderma 344, 50–62.
Razavi, S., Gupta, H.V., 2015. What do we mean by sensitivity analysis? The need for comprehensive characterization of "global" sensitivity in Earth and Environmental systems models. Water Resour. Res. 51, 3070–3092.
Saboe, D., Ghasemi, H., Gao, M.M., Samardzic, M., Hristovski, K.D., Boscovic, D., Burge, S.R., Burge, R.G., Hoffman, D.A., 2021. Real-time monitoring and prediction of water quality parameters and algae concentrations using microbial potentiometric sensor signals and machine learning tools. Sci. Total Environ. 764, 142876.
Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T., 2020. DeepAR: probabilistic forecasting with autoregressive recurrent networks. Int. J. Forecast. 36, 1181–1191.
Saltelli, A., Tarantola, S., Campolongo, F., Ratto, M., 2004. Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models. Wiley Online Library.
Seo, M., Heo, J., Kim, Y., 2021. Present and potential future critical source areas of nonpoint source pollution: a case of the Nakdong River watershed, South Korea. Environ. Sci. Pollut. Control Ser. 28, 45676–45692.
Shahriari, B., Swersky, K., Wang, Z., Adams, R.P., De Freitas, N., 2015. Taking the human out of the loop: a review of Bayesian optimization. Proc. IEEE 104, 148–175.
Shen, Z.Y., Chen, L., Chen, T., 2012. Analysis of parameter uncertainty in hydrological and sediment modeling using GLUE method: a case study of SWAT model applied to Three Gorges Reservoir Region, China. Hydrol. Earth Syst. Sci. 16, 121–132.
Singh, K.P., Basant, N., Gupta, S., 2011. Support vector machines in water quality management. Anal. Chim. Acta 703, 152–162.
Singh, R.B., Olbert, A.I., Samantra, A., Uddin, M.G., 2024a. AI-driven modelling approaches for predicting oxygen levels in aquatic environments. J. Water Process Eng. 66, 105940. https://doi.org/10.1016/j.jwpe.2024.105940.
Singh, R.B., Patra, K.C., Pradhan, B., Samantra, A., 2024b. HDTO-DeepAR: a novel hybrid approach to forecast surface water quality indicators. J. Environ. Manag. 352, 120091. https://doi.org/10.1016/j.jenvman.2024.120091.
Singh, R.B., Patra, K.C., Samantra, A., 2024c. GHPSO-ATLSTM: a novel attention-based genetic LSTM to predict water quality indicators. Stoch. Environ. Res. Risk Assess. 1–16.
Snieder, E., Khan, U.T., 2023. A novel ensemble algorithm based on hydrological event diversity for urban rainfall–runoff model calibration and validation. J. Hydrol. 619, 129193.
Sokolova, E., Ivarsson, O., Lillieström, A., Speicher, N.K., Rydberg, H., Bondelind, M., 2022. Data-driven models for predicting microbial water quality in the drinking water source using E. coli monitoring and hydrometeorological data. Sci. Total Environ. 802, 149798.
Sterner, R.W., Grover, J.P., 1998. Algal growth in warm temperate reservoirs: kinetic examination of nitrogen, temperature, light, and other nutrients. Water Res. 32, 3539–3548. https://doi.org/10.1016/S0043-1354(98)00165-1.
Tilman, D., Kilham, S.S., Kilham, P., 1982. Phytoplankton community ecology: the role of limiting nutrients. Annu. Rev. Ecol. Systemat. 13, 349–372.
Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2023. Assessing optimization techniques for improving water quality model. J. Clean. Prod. 385, 135671. https://doi.org/10.1016/j.jclepro.2022.135671.
van Griensven, A., Meixner, T., Grunwald, S., Bishop, T., Diluzio, M., Srinivasan, R., 2006. A global sensitivity analysis tool for the parameters of multi-variable catchment models. J. Hydrol. 324, 10–23. https://doi.org/10.1016/j.jhydrol.2005.09.008.
Van Griensven, A., Meixner, T., 2003. Sensitivity, Optimisation and Uncertainty Analysis for the Model Parameters of SWAT, pp. 162–167.
Vapnik, V., Golowich, S., Smola, A., 1996. Support vector method for function approximation, regression estimation and signal processing. Adv. Neural Inf. Process. Syst. 9.
Vázquez, R.F., Beven, K., Feyen, J., 2009. GLUE based assessment on the overall predictions of a MIKE SHE application. Water Resour. Manag. 23, 1325–1349.

Victoria, A.H., Maragatham, G., 2021. Automatic tuning of hyperparameters using Bayesian optimization. Evol. Syst. 12, 217–223.
Villa, A., Fölster, J., Kyllmar, K., 2019. Determining suspended solids and total phosphorus from turbidity: comparison of high-frequency sampling with conventional monitoring methods. Environ. Monit. Assess. 191, 1–16.
Wang, F., Wang, Y., Zhang, K., Hu, M., Weng, Q., Zhang, H., 2021. Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. Environ. Res. 202, 111660.
Wang, W., Xu, Z., Lu, W., Zhang, X., 2003. Determination of the spread parameter in the Gaussian kernel for classification and regression. Neurocomputing 55, 643–663.
Wang, Y., Zhou, J., Chen, K., Wang, Y., Liu, L., 2017. Water Quality Prediction Method Based on LSTM Neural Network, pp. 1–5.
Weatherhead, E.C., Reinsel, G.C., Tiao, G.C., Meng, X., Choi, D., Cheang, W., Keller, T., DeLuisi, J., Wuebbles, D.J., Kerr, J.B., 1998. Factors affecting the detection of trends: statistical considerations and applications to environmental data. J. Geophys. Res. Atmos. 103, 17149–17161.
Whitehead, P.G., Wilby, R.L., Battarbee, R.W., Kernan, M., Wade, A.J., 2009. A review of the potential impacts of climate change on surface water quality. Hydrol. Sci. J. 54, 101–123.
Wong, K.I., Wong, P.K., Cheung, C.S., Vong, C.M., 2013. Modeling and optimization of biodiesel engine performance using advanced machine learning methods. Energy 55, 519–528. https://doi.org/10.1016/j.energy.2013.03.057.
Wu, C., Lv, X., Cao, X., Mo, Y., Chen, C., 2010. Application of support vector regression to predict metallogenic favourability degree. Int. J. Phys. Sci. 5, 2523–2527.
Xia, Y., Zhang, M., Tsang, D.C., Geng, N., Lu, D., Zhu, L., Igalavithana, A.D., Dissanayake, P.D., Rinklebe, J., Yang, X., 2020. Recent advances in control technologies for non-point source pollution with nitrogen and phosphorous from agricultural runoff: current practices and future prospects. Appl. Biol. Chem. 63, 1–13.
Xu, X., Sun, C., Huang, G., Mohanty, B.P., 2016. Global sensitivity analysis and calibration of parameters for a physically-based agro-hydrological model. Environ. Model. Software 83, 88–102. https://doi.org/10.1016/j.envsoft.2016.05.013.
Yang, J., Cheng, C., Chan, C., 2017. A time-series water level forecasting model based on imputation and variable selection method. Comput. Intell. Neurosci. 2017, 8734214.
Yang, Y., Jin, S., 2023. Long-time water quality variations in the Yangtze river from Landsat-8 and Sentinel-2 images based on neural networks. Water 15, 3802.
Yi, X., Zou, R., Guo, H., 2016. Global sensitivity analysis of a three-dimensional nutrients-algae dynamic model for a large shallow lake. Ecol. Model. 327, 74–84. https://doi.org/10.1016/j.ecolmodel.2016.01.005.
Yuan, H., Yang, G., Li, C., Wang, Y., Liu, J., Yu, H., Feng, H., Xu, B., Zhao, X., Yang, X., 2017. Retrieving soybean leaf area index from unmanned aerial vehicle hyperspectral remote sensing: analysis of RF, ANN, and SVM regression models. Remote Sens. 9, 309.
Zahoor, I., Mushtaq, A., 2023. Water pollution from agricultural activities: a critical global review. Int. J. Chem. Biochem. Sci. 23, 164–176.
Zanoni, M.G., Majone, B., Bellin, A., 2022. A catchment-scale model of river water quality by Machine Learning. Sci. Total Environ. 838, 156377.
Zhang, B., Qin, Y., Huang, M., Sun, Q., Li, S., Wang, L., Yu, C., 2011. SD–GIS-based temporal–spatial simulation of water quality in sudden water pollution accidents. Comput. Geosci. 37, 874–882. https://doi.org/10.1016/j.cageo.2011.03.013.
Zhang, C., Nong, X., Behzadian, K., Campos, L.C., Chen, L., Shao, D., 2024a. A new framework for water quality forecasting coupling causal inference, time-frequency analysis and uncertainty quantification. J. Environ. Manag. 350, 119613.
Zhang, P., Liu, X., Dai, H., Shi, C., Xie, R., Song, G., Tang, L., 2024b. A multi-model ensemble approach for reservoir dissolved oxygen forecasting based on feature screening and machine learning. Ecol. Indic. 166, 112413. https://doi.org/10.1016/j.ecolind.2024.112413.
Zhang, Q., Blomquist, J.D., 2018. Watershed export of fine sediment, organic carbon, and chlorophyll-a to Chesapeake Bay: spatial and temporal patterns in 1984–2016. Sci. Total Environ. 619–620, 1066–1078. https://doi.org/10.1016/j.scitotenv.2017.10.279.
Zhang, W., Yang, D., Wang, H., 2019. Data-driven methods for predictive maintenance of industrial equipment: a survey. IEEE Syst. J. 13, 2213–2227.