PLOS ONE

RESEARCH ARTICLE

Machine learning-based estimation of riverine nutrient concentrations and associated uncertainties caused by sampling frequencies

Shengyue Chen[1], **Zhenyu Zhang**[1], **Juanjuan Lin**[2], **Jinliang Huang**[1]*

1 Fujian Key Laboratory of Coastal Pollution Prevention and Control, Xiamen University, Xiamen, China,
2 Xiamen Environmental Publicity and Education Center, Xiamen, China

* jlhuang@xmu.edu.cn

Citation: Chen S, Zhang Z, Lin J, Huang J (2022) Machine learning-based estimation of riverine nutrient concentrations and associated uncertainties caused by sampling frequencies. PLoS ONE 17(7): e0271458. https://doi.org/10.1371/journal.pone.0271458

Editor: Zaher Mundher Yaseen, TDTU: Ton Duc Thang University, VIET NAM

Received: April 6, 2022

Accepted: June 30, 2022

Published: July 13, 2022

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by National Natural Science Foundation of China (Grant No. 41971231). The funders had no role in study design, data collection and analysis, decsion to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Accurate and sufficient water quality data is essential for watershed management and sustainability. Machine learning models have shown great potentials for estimating water quality with the development of online sensors. However, accurate estimation is challenging because of uncertainties related to models used and data input. In this study, random forest (RF), support vector machine (SVM), and back-propagation neural network (BPNN) models are developed with three sampling frequency datasets (i.e., 4-hourly, daily, and weekly) and five conventional indicators (i.e., water temperature (WT), hydrogen ion concentration (pH), electrical conductivity (EC), dissolved oxygen (DO), and turbidity (TUR)) as surrogates to individually estimate riverine total phosphorus (TP), total nitrogen (TN), and ammonia nitrogen (NH_4^+-N) in a small-scale coastal watershed. The results show that the RF model outperforms the SVM and BPNN machine learning models in terms of estimative performance, which explains much of the variation in TP (79 ± 1.3%), TN (84 ± 0.9%), and NH_4^+-N (75 ± 1.3%), when using the 4-hourly sampling frequency dataset. The higher sampling frequency would help the RF obtain a significantly better performance for the three nutrient estimation measures (4-hourly > daily > weekly) for R^2 and NSE values. WT, EC, and TUR were the three key input indicators for nutrient estimations in RF. Our study highlights the importance of high-frequency data as input to machine learning model development. The RF model is shown to be viable for riverine nutrient estimation in small-scale watersheds of important local water security.

1 Introduction

Waterbodies must maintain a good chemical and ecological status to protect human health and safeguard natural ecosystems. Nutrients are important indicators that affect water quality, watershed health, and biological processes [1,2]. As key constituents of riverine nutrients, high concentrations of nitrogen (N) and phosphorus (P) may lead to eutrophication and anoxia in coastal waters [3], thereby not only affecting the living environment of human beings but also the biodiversity [4]. Therefore, it is crucial to master accurate water quality data and elucidate riverine N and P dynamics for effective watershed water management, particularly for small watersheds with limited water quality monitoring but significant local water-security.

Conventional field sampling is usually conducted to examine the dynamics of N and P in fresh water [5]. However, the sampling is typically too infrequent (i.e., weekly or monthly) to fully characterize lotic nutrient conditions and to accurately estimate nutrient loading [6,7]. Additionally, the field-sampling method involves laboratory analysis to determine the concentrations of water-quality parameters, which is labor- and cost-intensive, time-consuming, and limited in terms of spatial coverage [8].

Over the past few years, with the development of online water-quality monitoring technology, the use of sensors that directly measure water quality has changed the approach to watershed research [9]. Compared to lower-frequency field sampling, higher-frequency (e.g., hourly, minutely) water quality monitoring can well capture short-term water quality dynamics and extremes. Conventional water-quality indicators, such as water temperature (WT), hydrogen ion concentration (pH), electrical conductivity (EC), dissolved oxygen (DO), and turbidity (TUR), can be monitored using probes continuously and frequently. Research methods have gradually migrated from conventional field sampling with lab analyses to online monitoring with advanced *in situ* sensors [10]. However, for many key nutrient indicators (i.e., permanganate index, Chlorophyll a, or the components of N and P), it is still difficult and/or uneconomically monitored *in situ* with high-frequency [11,12]. Moreover, there are hidden dangers and problems, such as abnormal indications caused by probe damage and sensor failure, and high maintenance costs [13,14]. The low frequency of field sampling makes it difficult to capture the instantaneous variability of water quality, and the high price of sensors prevents them from being densely deployed, thus the spatial variability of watershed water quality is difficult to capture. Insufficient water quality data caused by these problems is usually not conducive to riverine health assessment and water management.

Machine learning models have shown great potentials for estimating water quality parameters. They can solve highly nonlinear problems [15,16] and supplement mechanism models [17]. Machine learning algorithms do not consider physical processes [18], and a large number of data are often required to operate them [19]. Many studies have adopted surrogate regression to enhance the rapid generation of data input based on *in situ* measurements and to simplify resource-intensive laboratory experimentation. According to this method, the concentration of riverine nutrients can easily be estimated using alternative indicators. Researchers have used a variety of machine learning algorithms, such as neural networks (NNs; [20–22], support vector machines (SVM; [23–25], and random forest (RF; [26–29], to estimate water environment related indicators. It was found that machine learning algorithms, especially RF, have great potential and are more frequently applied for this purpose [30]. For example, different machine learning algorithms were used to compare the estimation accuracy of nutrient concentrations, and the results showed that RF was significantly more accurate than other conventional algorithms when estimating all six levels of water quality (I, II, III, IV, V, and worse than V [WV]), which are based on the National Environmental Quality Standards for surface water of China (GB3838-2002) [31]. The RF, gradient boost regression, and AdaBoost regression have been used to simulate the daily suspended sediment load in the Mississippi River, and the result show that RF is slightly ahead in prediction performance [32].

It is well known that uncertainty is inherent in model development [33]. Many studies were devoted to exploring the causes of uncertainty in machine learning models to improve estimation accuracy [34,35]. Sharafati et al. [35] used a Monte Carlo simulation model to quantify estimation uncertainties. The results showed that the model structures were more influential than the input indicators for estimating effluent quality parameters. Noori et al. [36] used the percentage of observed data bracketed by 95% predicted uncertainties (95PPU) and the bandwidth of 95% confidence intervals (*d*-factor) to analyze the uncertainties brought by SVM hyperparameters. They found that the model was more sensitive to the capacity parameter (C) than to kernel parameters (Gamma) and error tolerance (Epsilon). Not just hyperparameter and model structure, data input associated with different sampling frequencies might also induce uncertainties and influence estimation accuracy [37]. Derot et al. [2] demonstrated that the different sampling frequency datasets directly impact the forecast performance of an RF model. According to their findings, the accuracy of phytoplankton bloom forecasts for a 20-min time step was higher than that of the 1-day time step. It appears from these studies that there are many kinds of factors that affect the estimation accuracy and associated uncertainty. Among those factors, the model uncertainty caused by the frequency of data input might be more worthy of discussion with the increasing popularity of automatic monitoring sensors.

The estimation accuracy of nutrient concentration depends not only on the model structure but also on the amount and type of data input [31]. Many researchers used multiple types of indicator inputs for estimation [38] or indicators having high correlation with the substances to be tested as inputs. Some even used one nutrient to estimate another type of nutrient. Although desired estimation results can be achieved, these methods are difficult to implement in reality because some of the input indicators (chemical oxygen demand, nitrate, and nitrite, etc.) are not readily available in a high temporal resolution [39]. Therefore, it is crucial to develop a convenient as well as accurately model of nutrient concentration estimation that the input indicators are easier available.

Despite that many studies have been focused on machine learning in different fields, few researches have combined machine learning methods with high-frequency monitoring data and evaluate model uncertainty caused by frequency of data input. To develop a model that can estimate riverine nutrient (total phosphorus [TP], total nitrogen [TN], and ammonia nitrogen [NH_4^+-N]) concentrations easily and accurately, as well as evaluate the uncertainty caused by the sampling frequency, thus helpful to water management in a small-scale watershed, we developed an RF model using datasets of only five monitoring water-quality indicators (i.e., WT, pH, EC, DO, and TUR) from the unique online multi-parameter water-quality sensor located in the outlet of the watershed (sensor type can be seen in S1 Text, Supporting information). Concurrently, we constructed an SVM and a back-propagation neural network (BPNN) for performance comparison. All these three machine learning models are widely used, and with well estimation accuracy. Specifically, the main objectives of this study are (1) to compare the estimative performance of different machine learning models for riverine nutrient concentrations, and (2) to evaluate the accuracies and uncertainties of the models with datasets of different sampling frequencies (i.e., 4-hourly, daily, and weekly). The findings of this study would be helpful to easily estimating riverine nutrient concentrations in small-scale watersheds and evaluating the contributions of high-frequency data to estimation accuracy. The proposed model strategy can be used in other small-scale watersheds with scarce data on nutrients but easily available and high frequency chemical/physical indicators to improve the efficiency of machine learning models used for water-quality estimation.

2 Data and methodology

Herein, a data-driven methodology based on machine learning is proposed to measure uncertainties due to three different sampling frequencies while estimating the riverine nutrient concentrations. As shown in Fig 1, this technique route comprises three components: (1) data preparation, (2) model development, and (3) accuracy and uncertainty analyses. The methods and formulations involved are described exhaustively in the following sections.

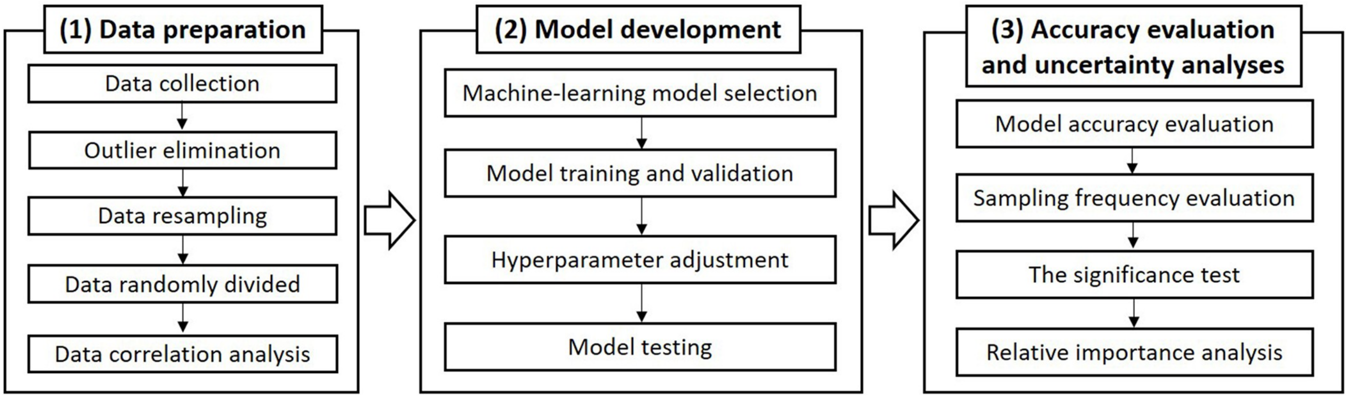
(1) Data preparation
Data collection
Outlier elimination
Data resampling
Data randomly divided
Data correlation analysis
(2) Model development
Machine-learning model selection
Model training and validation
Hyperparameter adjustment
Model testing
(3) Accuracy evaluation and uncertainty analyses
Model accuracy evaluation
Sampling frequency evaluation
The significance test
Relative importance analysis

Fig 1. Flowchart of the proposed methodology.

https://doi.org/10.1371/journal.pone.0271458.g001

2.1 Data preparation

The Aitoutan (ATT) watershed is located in Tong'an District, Xiamen, China. Since China launched environmental regulations (e.g., "River Chief") in 2016, water quality in the ATT watershed has been significantly improved. In recent years, the main pollutant faced by the watershed is TP, and the sensor-monitoring data at the outlet of the watershed shows that the concentration of TP frequently exceeds the level III based on National Environmental Quality Standards for surface water of China (higher than 0.2 mg/L) (Fig 2). Thus, water quality is still a concern for local governments.

The data of the monitoring site in the study area was acquired by sensors in the surface water, and the other monitoring indicators except nutrients are used as the input indicators of the machine learning models. The dataset in this study comprises five physical/chemical indicators used as inputs of machine learning models, namely WT, pH, EC, DO, and TUR, and three nutrients being estimated, namely TP, TN, and NH_4^+-N, which covers the period from January 1, 2019, to March 31, 2021, and was provided by the Xiamen Environmental Publicity and Education Center (specific information can be seen in S6 Text, Supporting information). The outliers (each water quality indicator value lower than/equal to 0 and the null value) were eliminated from this dataset. This dataset has a temporal resolution of four hours, which denotes that the water-quality indicators were automatically monitored by an interval of four hours from midnight daily. We resampled this 4-hourly frequency monitoring dataset to mimic both daily and weekly monitoring schemes. The water-quality indicators at 8 a.m. each day were extracted as a daily dataset, and the indicators at 8 a.m. each Monday were extracted as a weekly dataset. The three datasets of sampling frequency scenarios have the same temporal span. The 4-hourly dataset includes 4,209 samples of water quality indicators (five physical/chemical indicators and three nutrients as described above), whereas the daily dataset includes 803 samples; the weekly dataset has 115 samples. The samples in each dataset are at the same time step, that is, there is no time lag in the input samples in this study.

As summarized in Table 1, the descriptive statistics of these five input indicators and three nutrients with the 4-hourly frequency showed that the indicators having the highest coefficients of variation (CV) were TUR and NH_4^+-N, and the most stable indicator was pH. The CVs of WT and DO as well as TUR and NH_4^+-N were similar in pairs. The standard deviation (SD) was used to measure the data deviation from the mean value. CV is the mean normalized SD, and it represents the statistical dispersion of data. Before model development, the input indicators and nutrients of training set of the 4-hourly dataset will undergo Spearman's test of

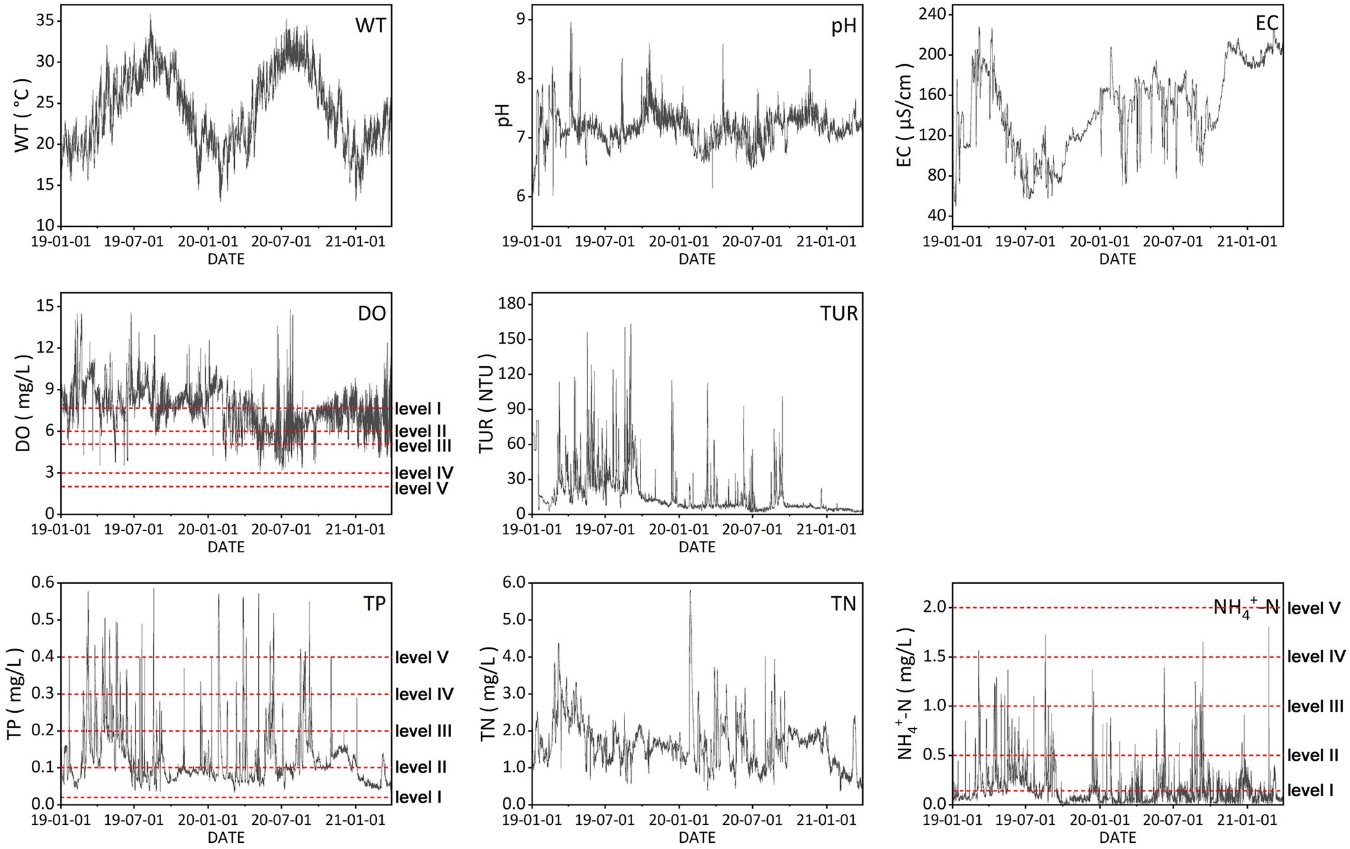

Fig 2. The 4-hourly variation of sensor readings from January, 2019, to March, 2021, for the water quality indicators in the outlet of the Aitoutan (ATT) watershed. The red dotted lines represent the boundary of environmental quality standards for surface water in China. The water quality levels gradually deteriorate from level I to level V and the value of indicators exceeding the level V is defined as "worse than V". For DO, the higher value represents the better water quality level, and for TP and NH_4^+-N, the higher value represents the worse water quality level.

https://doi.org/10.1371/journal.pone.0271458.g002

rank correlation to determine whether the correlation between the five input indicators and nutrients are too high.

$$\mathrm{SD} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(O_i - \bar{O}\right)^2} \tag{1}$$

Table 1. Descriptive statistics of input indicators and output nutrients from the monitoring site located in the outlet of Aitoutan (ATT) watershed.

Parameter	Max	Min	Mean	SD	CV (%)
WT (°C)	35.83	13.00	24.27	4.58	18.88
pH	8.96	6.03	7.16	0.28	3.92
EC (μS/cm)	227.95	49.80	146.44	42.24	28.84
DO (mg/L)	14.79	3.17	7.66	1.56	20.40
TUR (NTU)	162.97	1.90	16.70	17.94	107.41
TP (mg/L)	0.59	0.03	0.13	0.08	64.82
TN (mg/L)	5.81	0.38	1.66	0.69	41.28
NH_4^+-N (mg/L)	1.92	0.01	0.16	0.19	113.55

Notes: CV = "coefficient of variation"; SD = "standard deviation".

https://doi.org/10.1371/journal.pone.0271458.t001

$$\mathrm{CV}\% = \frac{\mathrm{SD}}{\bar{O}} \times 100 \tag{2}$$

where n is the number of input samples, O_i is the observations, and $\bar{O}$ represents the mean values of the observations.

2.2 Model development

MATLAB 2019b was used in this study to develop the RF, BPNN, and SVM model. To prevent overfitting of the models and ensure the generalization ability of the model, 80% of the dataset was randomly selected as the training set first, and the remaining 20% was selected as the testing set. The training set was then divided into a training-validation set based on a 10-fold cross-validation [40,41]. In this study, the training set was used for model fitting, the validation set was used to pick the optimal hyperparameter combination, both training set and validation set here were in 10-fold cross-validation phase, and we determined the optimal hyperparameters by the average of the statistical metrics of the validation set under 10-fold cross-validation. Then we iterated the optimal hyperparameter combination to three machine learning models, fit the models with the initially divided training set, and test the generalization ability of the models in the testing set. We selected the optimal model from three machine learning models (Section 3.2) and evaluate the estimation accuracy and uncertainty of the selected model with three sampling frequency scenarios (Section 3.3).

2.3 Accuracy evaluation and uncertainty analysis

The three machine learning models were evaluated for the estimation accuracy of cross-validation step under the 4-hourly frequency scenario, and the model with the best performance of validation set would be selected for the next phase (accuracy and uncertainty analysis due to different sampling frequencies). Several statistical metrics were selected to evaluate the estimation accuracy and uncertainty of the models proposed in this study. The coefficient of determination (R^2), Nash-Sutcliffe efficiency (NSE), root mean squared error (RMSE), and mean absolute error (MAE) were used to assess the goodness of fit between the observed nutrient concentrations and those estimated by three models.

$$R^2 = \frac{\left[\sum_{i=1}^{n}(O_i - \bar{O})(P_i - \bar{P})\right]^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2 \sum_{i=1}^{n}(P_i - \bar{P})^2} \tag{3}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{1}^{n}(O_i - P_i)^2} \tag{5}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{1}^{n}|(O_i - P_i)| \tag{6}$$

where n is the samples of training/validation/test sets in 4-hourly/daily/weekly frequency scenario; O_i and P_i are respectively the observations and model estimations for each set; $\bar{O}$ and $\bar{P}$ respectively represent the mean values of the observations and model estimations for each set.

Usually, R^2 and NSE values closer to 1 while RMSE and MAE values closer to 0 denote higher accuracy.

In this phase, to evaluate the estimation accuracy and uncertainty caused by sampling frequencies, we first selected the model with the highest estimation accuracy from the three machine learning models. We resampled the 4-hourly dataset to extract daily and weekly sets according to the pattern in Section 2.1 and nine scenarios (i.e., three nutrients × three sampling frequencies) were designed. The testing set of the 4-hourly scenario has 842 samples (20% previously split from the 4-hourly dataset). The datasets of daily and weekly sampling frequency scenarios were all used as training-validation sets for their respective models based on the k-fold cross-validation. In order to equally evaluate and compare the impact of three sampling frequency scenarios on the estimation accuracy of RF, we chose the testing set of 4-hourly scenario, and from it we randomly selected 20% of the total samples of daily/weekly scenario as the testing sets for the daily/weekly scenarios. Therefore, the training set of the daily scenario has 803 samples and the testing set has 161 samples; the training set of the weekly scenario has 115 samples and the testing set has 23 samples. We performed 30 replicate estimations under this dataset division, and evaluated the model accuracies and uncertainties in testing sets under three sampling frequency scenarios. The statistical metrics for estimation accuracies of testing sets were used for the one-way analysis of variance (ANOVA) test to evaluate whether there is a significant difference in the estimation accuracy between the three sampling frequencies.

One of the main advantages of RF is that it can assess the importance of the input indicators used in the modeling processes [42]. It is vital to identify some key water indicators when model developing. To further optimize the machine learning model and improve the comprehensive management of watersheds, the RF model was selected to analyze the relative importance of the input indicators. For each nutrient, the weights and relative importance of the input indicators were ranked and analyzed. The calculation method of the importance of each indicator in RF is as follows: (1) For each decision tree in the RF model, the out-of-bag (OOB) data are used to calculate OOB error, denoted as OOBE1. (2) Redistribute all the original N samples of each indicator through permutation, the OOB error is calculated again and recorded as OOBE2. (3) Assuming that there are N trees in the RF model, the relative importance for each indicator can be shown in Eq (7):

$$RI_i = \frac{\sum_1^n \left[\frac{\sum_1^N (\mathrm{OOBE2}_i - \mathrm{OOBE1}_i)}{N} \right]}{n} \tag{7}$$

where RI_i refers to the relative importance of each indicator, N denotes the amounts of tree of RF model, and n is the number of indicators.

3 Results

3.1 Correlation analysis of water quality indicators

Based on Spearman's test of rank correlation, there was a large number of high statistically-significance (i.e., $p < 0.01$) among the nutrients and input indicators (Fig 3). As shown in this figure, TUR is strongly positively correlated with all three nutrients, DO is positively correlated with all three nutrients, EC is weakly positively correlated with TP and TN and weakly negatively correlated with TN, pH is positively correlated with TP and TN and negatively correlated with NH_4^+-N, and WT is weakly positively correlated with TP and NH_4^+-N and negatively correlated with TN. The correlation analysis of the nutrient concentrations showed that TP

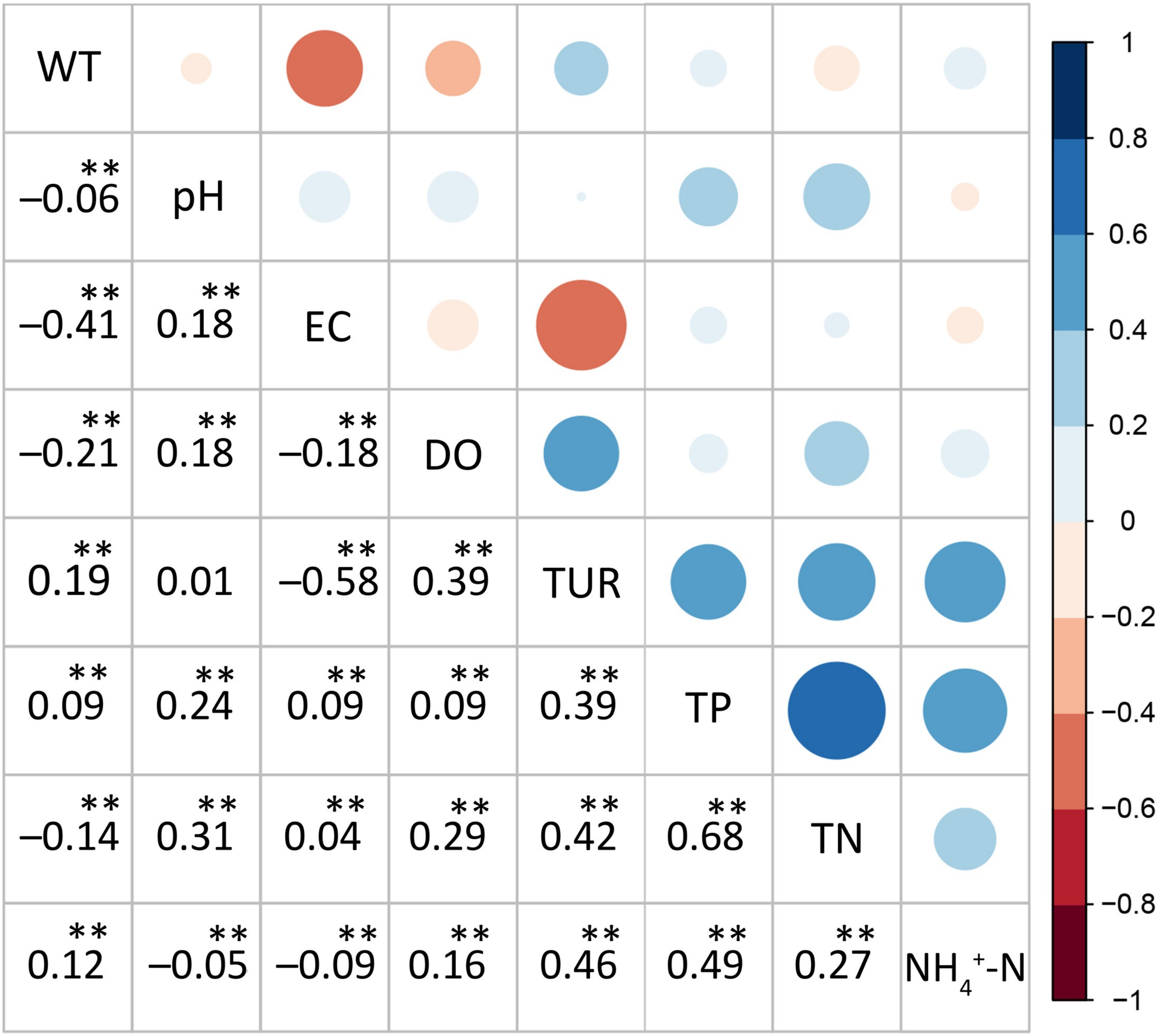

Fig 3. Correlation analysis for the input and output indicators. The statistical significance of rank correlations is denoted by asterisks for $p < 0.05$ (*) and $p < 0.01$ (**) (lower left). The different sizes and colors of circles represent the strength of the correlation between the indicators (upper right).

https://doi.org/10.1371/journal.pone.0271458.g003

was strongly positively correlated with TN and NH_4^+-N, while the correlation between TN and NH_4^+-N was relatively low.

3.2 Evaluation of estimation accuracy among three machine learning models

The sampling frequency of data we used in this phase was the 4-hourly scenario, and the three models used the same division rules for the dataset. Different machine learning models using the same dataset for estimation may have different performances. The hyperparameter

Table 2. Comparison of the average estimation accuracy of the three machine-learning models (4-hourly frequency, testing step, n = 842).

Model	Statistical metric	Nutrient		
		TP	TN	NH_4^+-N
RF	R^2	0.801	0.859	0.759
	NSE	0.785	0.853	0.748
	RMSE	0.039	0.284	0.087
	MAE	0.024	0.189	0.057
SVM	R^2	0.737	0.811	0.720
	NSE	0.734	0.810	0.717
	RMSE	0.044	0.316	0.095
	MAE	0.025	0.219	0.054
BPNN	R^2	0.668	0.757	0.616
	NSE	0.666	0.754	0.602
	RMSE	0.049	0.361	0.113
	MAE	0.031	0.268	0.072

https://doi.org/10.1371/journal.pone.0271458.t002

selections of three machine learning models can be found in S2, S3, and S4 Text in Supporting information. The performances of testing set can be seen in Table 2. For each nutrient, the R^2 and NSE obtained by RF are higher than SVM and BPNN, whereas the RMSE and MAE of RF are the lowest among three models.

This study uses Taylor diagrams to make visual comparisons of results obtained by the three models (Fig 4). Model performance is represented by a point, where the most accurate model has the closest distance to the point of observation, which is shown by the dark-grey point in the diagrams. Based on the principle of the Taylor diagram (i.e., correlation, standard deviation, and RMSE), the RF model has higher correlations with observed nutrient concentrations and a lower RMSE compared with the two other models. Fig 4 confirms that the RF model provides the highest accuracy when estimating TP, TN, and NH_4^+-N concentrations. Moreover, the BPNN model has the weakest performance compared with the other models.

3.3 Evaluation of model accuracy with different sampling frequency scenarios

We chose the RF model that had the highest R^2 and NSE and the lowest RMSE and MAE values in testing step under 4-hourly scenario (Table 2) for subsequent use. The hyperparameter

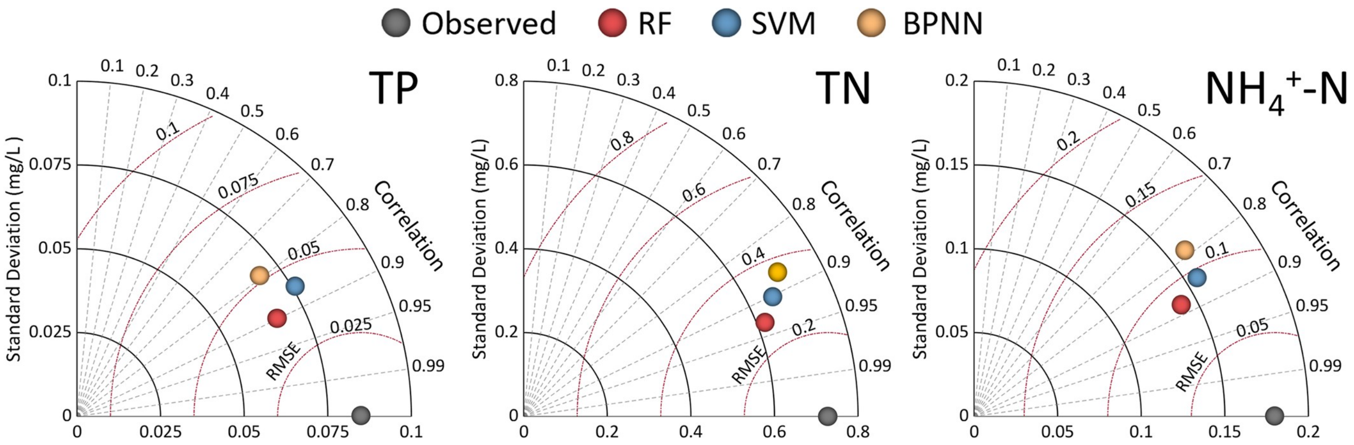

Fig 4. Comparison of the models' performances by Taylor diagrams. RF = "random forest"; SVM = "support vector machine"; BPNN = "back-propagation neural network"; TP = "total phosphorous"; TN = "total nitrogen"; and NH_4^+-N = "ammonia-nitrogen".

https://doi.org/10.1371/journal.pone.0271458.g004

Table 3. Comparison of the average estimation accuracy of the RF model with three sampling frequencies (testing step).

Sampling frequency	Statistical metric	Nutrient		
		TP	TN	NH_4^+-N
Four-hourly (n = 842)	R^2	0.785	0.840	0.749
	NSE	0.781	0.837	0.747
	RMSE	0.038	0.285	0.087
	MAE	0.023	0.187	0.057
Daily (n = 161)	R^2	0.692	0.761	0.676
	NSE	0.667	0.735	0.668
	RMSE	0.041	0.301	0.093
	MAE	0.026	0.196	0.065
Weekly (n = 23)	R^2	0.602	0.658	0.598
	NSE	0.574	0.639	0.559
	RMSE	0.044	0.318	0.101
	MAE	0.029	0.201	0.068

https://doi.org/10.1371/journal.pone.0271458.t003

selections of RF were consistent with Section 3.2. We performed 30 replicate estimations for each of nine scenarios (i.e., three nutrients × three sampling frequencies) as described in Section 2.3. The mean results of testing phase are presented in Table 3. Among them, the rank of R^2 and NSE values of the RF model under three sampling frequency scenarios is 4-hourly > daily > weekly. When the sampling frequency was increased from weekly to 4-hourly, the R^2 and NSE obtained by the RF model is greatly improved (TP 30%, TN 30%, and NH_4^+-N 25% for R^2; TP 36%, TN 31%, and NH_4^+-N 34% for NSE). Regarding RMSE and MAE, there is no such pattern. Among the average estimation results of the RF model with three sampling frequency scenarios, the values of RMSE and MAE do not change much compared with R^2 and NSE.

The scatterplots can characterize the relationship between observed values (i.e., three nutrients with three sampling frequencies) and the average estimation results of the RF model in the testing phase (Fig 5). Results show that as the sampling frequency increases, the slope of the fitted line between the estimated value and the observed value constantly approach 45° (slope = 1), which also results in the increase of model estimation accuracy. For different nutrients, the slope of the fitted line can also prove the rank of model estimation accuracies (TN > TP > NH_4^+-N). When the actual values (i.e., observed nutrient concentrations) are lower than half of their maximum values, overestimation and underestimation by RF exist simultaneously; however, when the actual values are higher than half of their maximum value, the RF tends to underestimate, which is more obvious at the peak of observations. The error between observations and estimations at the peak (especially underestimation) may be the main reason to affect the slope.

The 30 replicate estimation results under various scenarios are also displayed in a violin plot (Fig 6). This representation not only shows the quantile, but it also provides the kernel density curve of the data. In view of the results in which the variation of RMSE and MAE are minimal compared with R^2 and NSE, we only chose R^2 and NSE to evaluate the performance of different sampling frequencies. As shown in Fig 6, for all nutrients, the mean values of R^2 and NSE after 30 RF estimations under the 4-hourly frequency are higher than those of the daily frequency. The weekly one has the lowest R^2 and NSE. It can be observed from the inside boxes that R^2 and NSE values obtained by RF with via 4-hourly sampling frequency scenario have the smallest changes under each scenario. Thus, they maintain a high level. For comparison, the estimation accuracy of RF under the weekly scenario fluctuates greatly, and the high

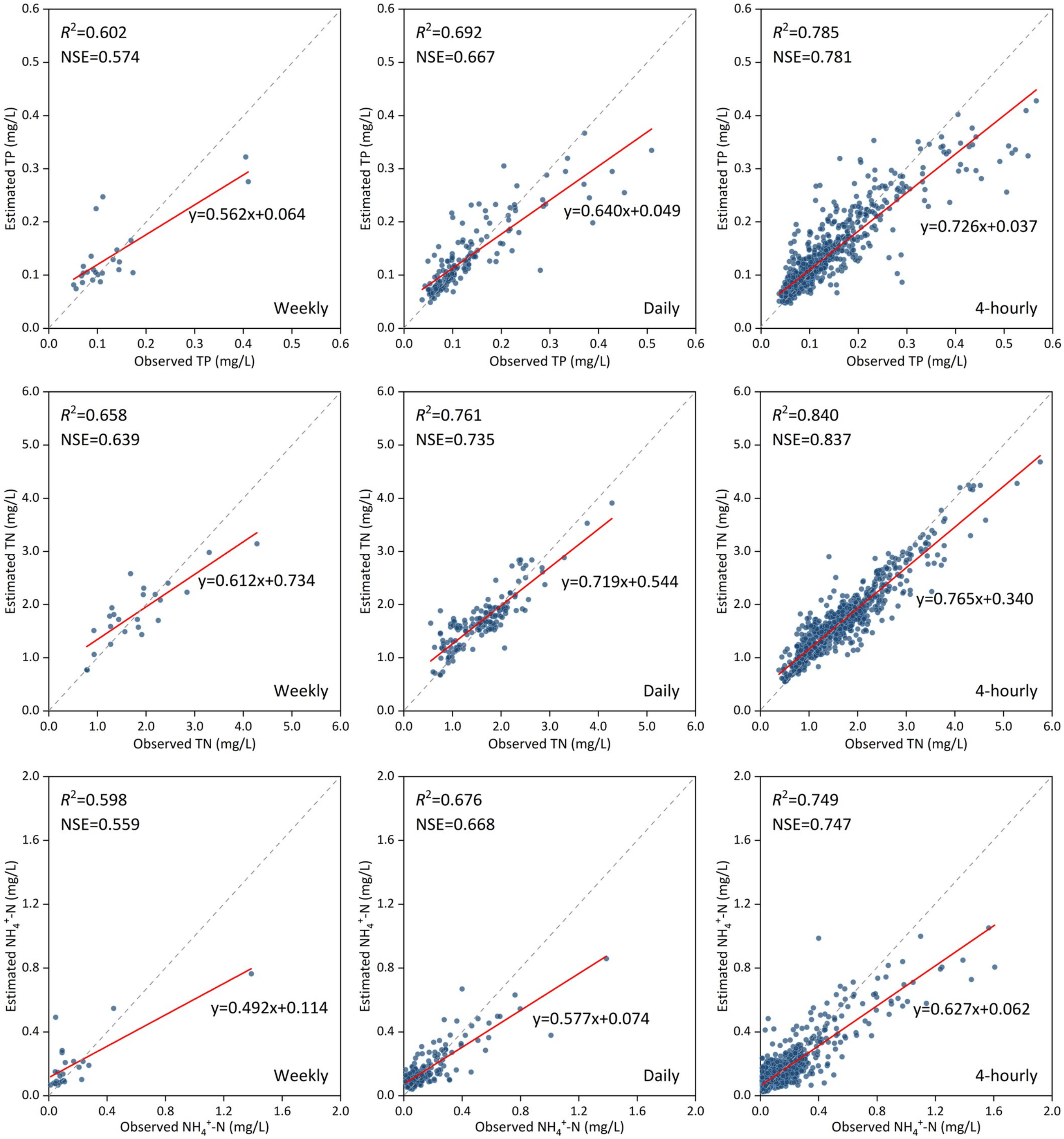

Fig 5. Scatterplots of the observations and average estimations with three sampling frequency scenarios. The x-axis represents the observations while the y-axis represents the estimations. The grey dashed line represents the 1:1 fitted line of observations and estimations under ideal conditions. The red line represents the fitted line of observations and estimations in actual situation.

https://doi.org/10.1371/journal.pone.0271458.g005

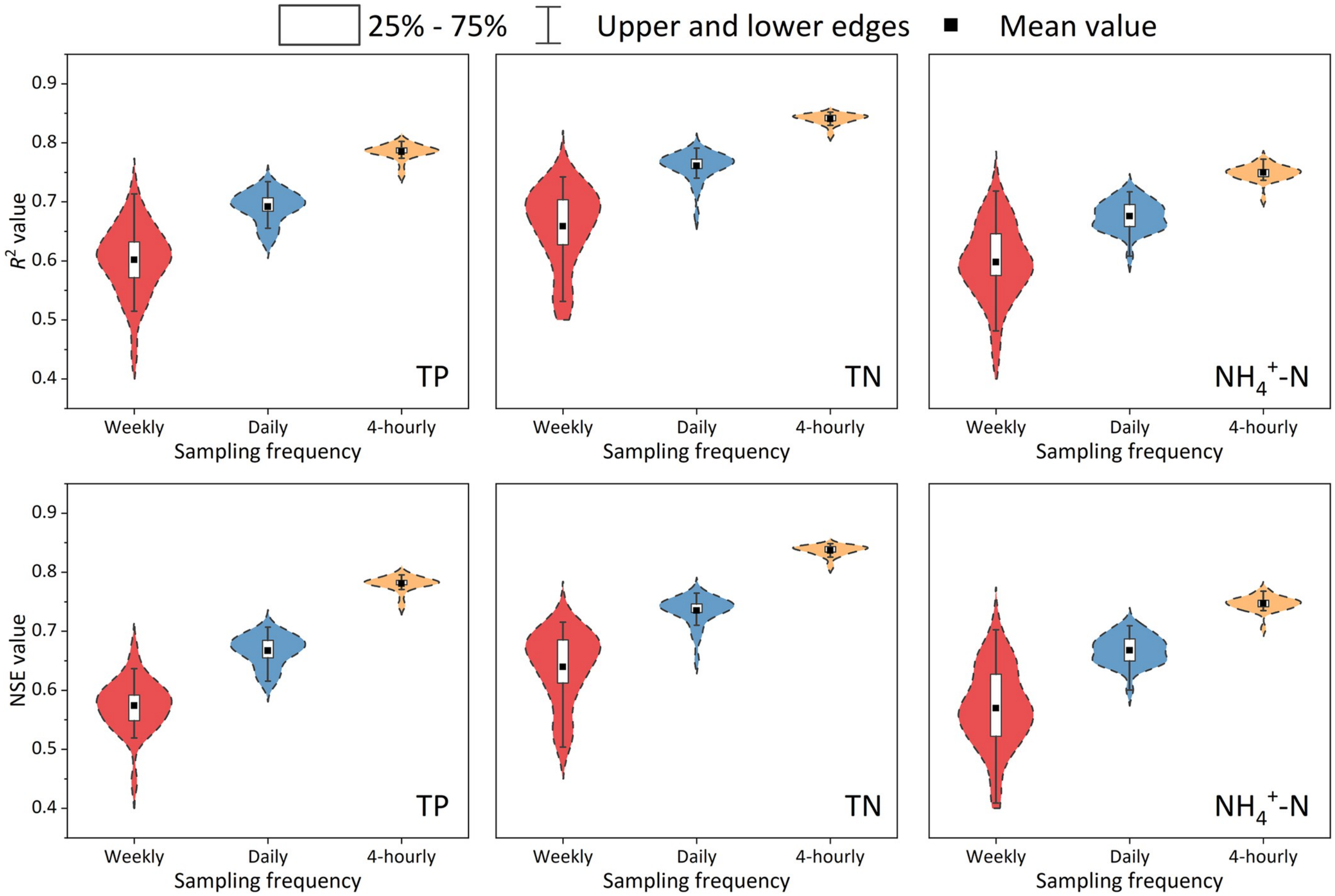

Fig 6. Estimated R^2 and Nash-Sutcliffe efficiency (NSE) values for the random-forest (RF) model under different sampling frequency scenarios. The width of the violin shape indicates the frequency at which R^2 and NSE appear at this value.

https://doi.org/10.1371/journal.pone.0271458.g006

(e.g., R^2 and NSE about 0.7) and the low (R^2 and NSE about 0.4) accuracies appear at the same time. Hence, the mean values are the lowest in the end. Regarding the comparison of estimation accuracies among the different nutrients, driven by the same sampling frequency data input, TN always obtains the highest R^2 and NSE values, whereas NH_4^+-N is always the lowest.

An ANOVA test was performed to confirm whether the uses of dataset with different sampling frequencies cause significant differences in the estimation accuracy of the RF model. The results are presented in Table 4. For each group (one nutrient × one statistical metrics), the differences of three sampling frequencies are significant. The estimation accuracy of the RF model under the 4-hourly frequency is significantly better than that of the daily frequency, and the daily frequency is also significantly better than the weekly one. On the other hand, the higher frequency of data input reduces the fluctuation of RF estimation accuracy (i.e., the smallest SD with 4-hourly and biggest SD with the weekly frequency). In summary, for one nutrient, a higher sampling frequency typically causes the RF to yield a higher estimation accuracy.

3.4 Relative importance of input indicators

To clarify the relative importance of the five alternative inputs and find the key indicators in the nutrient concentration estimations, the RF with the 4-hourly sampling frequency scenario

Table 4. Results of the analysis-of-variance (ANOVA) test.

Nutrient	Statistical metric	Sampling frequency (n = 30)			F
		Weekly (Mean ± SD)	Daily (Mean ± SD)	Four-hourly (Mean ± SD)	
TP	R^2	0.602 ± 0.057 c	0.692 ± 0.026 b	0.785 ± 0.013 a	182.08**
	NSE	0.574 ± 0.042c	0.667 ± 0.026 b	0.781 ± 0.013 a	354.99**
TN	R^2	0.658 ± 0.065 c	0.761 ± 0.023 b	0.840 ± 0.009 a	150.21**
	NSE	0.639 ± 0.059 c	0.735 ± 0.024 b	0.837 ± 0.009 a	205.18**
NH_4^+-N	R^2	0.598 ± 0.062 c	0.676 ± 0.025 b	0.749 ± 0.013 a	107.29**
	NSE	0.559 ± 0.069 c	0.668 ± 0.025 b	0.747 ± 0.012 a	131.95**

Note: Statistical significance in the ANOVA test is denoted by asterisks for both $p < 0.05$ (*) and $p < 0.01$ (**). The F value denotes the ratio of the mean square between groups to the mean square within groups. The larger F value represents the larger difference between the groups. The different letters (a-c) after the numbers (Mean ± SD) indicated the significant differences between three sampling frequencies, while the same letters indicated that there are not significant differences.

https://doi.org/10.1371/journal.pone.0271458.t004

was used. As shown in Fig 7, EC, TUR, and WT are the three most important indicators. TUR shows the highest relative importance in controlling the estimation accuracy of TP and NH_4^+-N, while EC is the most important indicator for the estimation of TN. Comparatively, pH and DO are the two least important indicators of nutrient estimation. Based on the relative importance analysis, we found the three key indicators that affect the nutrient concentration dynamics among the five conventional water quality indicators.

4 Discussion

4.1 Uncertainty of model estimation

Machine learning models have large uncertainties associated with their unique structures, hyperparameter adjustment requirements, and data input [36,43]. The division rules of training and testing sets and the addition or deletion of input indicators can also cause fluctuations of estimation accuracy [44]. The same machine learning algorithm mentioned in different studies will perform differently due to the above-mentioned factors. Different machine learning algorithms will also perform differently even if be in the same study area and using the same dataset (specific information can be seen in the Table in S5 Text, Supporting information). There is no single algorithm that works best under all conditions. [45]. Firstly, we compared the estimation accuracy of three widely used machine learning models in our study area. In addition to the differences of the model, we controlled other variables to maintain consistency. The results of the testing step showed that the estimation accuracy of the RF model was the highest among the three models. The RF had the highest R^2 and NSE values (R^2 = 0.801, 0.859, and 0.759 for TP, TN, and NH_4^+-N; NSE = 0.785, 0.853, and 0.748 for these three nutrients) and the lowest RMSE and MAE values (RMSE = 0.039, 0.284, and 0.087 for TP, TN, and NH_4^+-N; MAE = 0.024, 0.189, and 0.057 for these three nutrients) (Table 2). The Taylor diagrams (Fig 4) also supported this finding. In these diagrams, the RF model was always the closest to the point represented by the observation, whereas the BPNN was the farthest from observation.

Many studies compared the performance of different models under the same conditions. Some of them reached the same conclusion as ours, that the RF model may be a more viable tool than other models for estimating water quality [31,32,46]. We also found that the estimation accuracy of the SVM was higher than BPNN, which is also found in other studies [47,48].

On the other hand, the number of input indicators affects the estimation accuracy of the machine learning model [49]. Attention should be paid to the overfitting caused by excessive types of input indicators [38,50]. Simultaneously, the difficulty of data acquisition must be

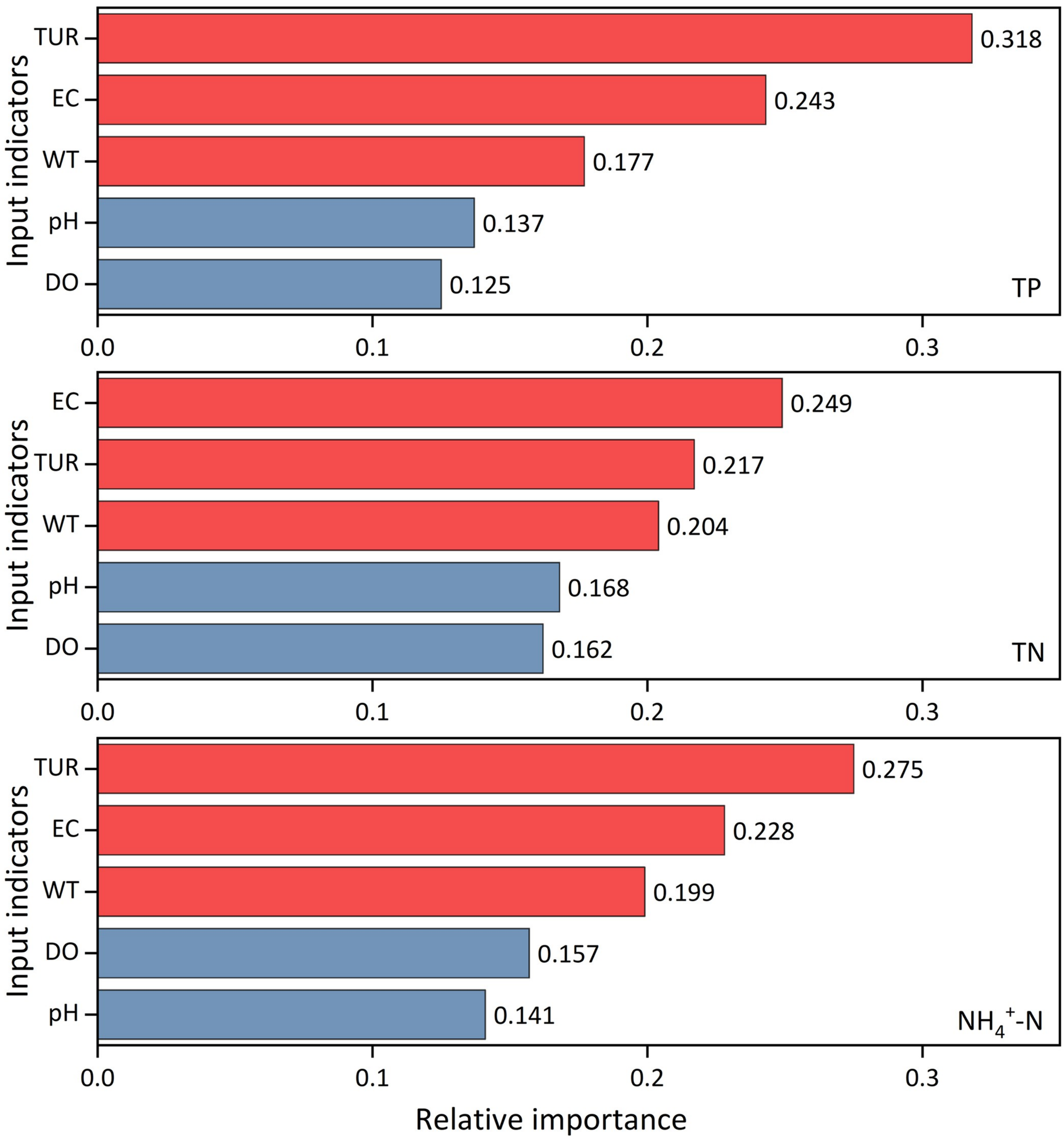

Fig 7. Relative importance analysis results of five input indicators in the random-forest (RF) model.

https://doi.org/10.1371/journal.pone.0271458.g007

considered [39,51]. For the simplicity and feasibility of the model, the input indicators must be at a sufficiently small scale to make estimations [52]. For the convenience of data acquisition, we only selected five water-quality parameters that can be measured easily *in situ*. Manual sampling and experiments or automatic sensor monitoring can be the method to obtain model input data, and the obtained data can be used as input indicators for subsequent nutrient concentration estimations according to the proposed methodology.

Different sampling frequencies influence estimation accuracy when using machine learning methods [31,53]. Generally speaking, the higher sampling frequency means that a larger amount of data can be obtained in the same time period, which will cause the machine learning model to use more data to improve its learning ability and obtain better estimation performance. Thomas et al. [54] found that the R^2 for phytoplankton estimation decreased from 0.89 at a resolution of 4-hourly to 0.74 at a 1-month resolution. Our study also showed that a higher sampling frequency led to higher accuracy (Figs 5 and 6 and Tables 3 and 4). Moreover, high-frequency data input also plays an important role in improving the estimation performance of the mechanism model. Jiang et al. [55] used two frequencies data input and catchment hydrology model named HYPE to estimate nitrate and evaluate uncertainty. They found that HYPE model better captured nitrate dynamics when using daily data than fortnightly data, and daily data produced smaller predictive uncertainty. However, Liu and Lu [56] compared the estimation accuracies of TP and TN concentration by the SVM and artificial neural network (ANN) models under monthly, bimonthly, and trimonthly sampling frequencies from January, 2005, to December, 2010. And they drew a different conclusion: a higher sampling frequency sometimes does not lead to improvements of estimation accuracy, which may even cause accuracy degradation (for example, using SVM and ANN to estimate the concentration of TP and TN under different sampling frequencies, the order of accuracy was that bimonthly > trimonthly > monthly). Their conclusions indicated that increasing the sampling frequency does not necessarily increase the estimation accuracy though the sampling frequency they selected was not the "high frequency".

To evaluate the model performance due to sampling frequency, we used the high-frequency dataset to construct different sampling frequency scenarios, and we analyzed the changes in estimation accuracy. The ANOVA test showed that the mean accuracy of 30 replicate estimations with the 4-hourly sampling frequency data input (R^2 = 0.785, NSE = 0.781 for TP; R^2 = 0.840, NSE = 0.837 for TN; R^2 = 0.749, NSE = 0.747 for NH_4^+-N) was significantly higher than that of the daily (R^2 = 0.692, NSE = 0.667 for TP; R^2 = 0.761, NSE = 0.735 for TN; R^2 = 0.676, NSE = 0.668 for NH_4^+-N) and weekly (R^2 = 0.602, NSE = 0.574 for TP; R^2 = 0.658, NSE = 0.639 for TN; R^2 = 0.598, NSE = 0.559 for NH_4^+-N) data (Table 4). One reason for this may be that more data inputs can lead to a better understanding of hidden patterns [57]. Alternatively, the 4-hourly frequency may better represent the actual situation (e.g., concentration mutations) than the daily and weekly frequencies. This indicates that when other conditions are consistent, the larger number of data input could help the model better reflect the patterns of change in the values estimated, leading to higher performance [58,59]. With the development of technology, high-frequency water-quality monitoring equipment are deployed to rivers worldwide, which helps society better grasp the water-quality change information needed to complete model simulations more accurately [6,60]. This ideal situation cannot be easily realized with low-frequency sampling methods and laboratory experiment. Therefore, we strongly recommend using high-frequency data to develop the RF model to grasp the dynamic changes of riverine nutrient concentration.

4.2 Different estimation accuracies among three nutrient concentrations

In this study, the RF model showed the highest estimation accuracy for TN and the lowest estimation accuracy for NH_4^+-N. During the period from January 2019 to March 2021, the CV of TN was the lowest, whereas that of NH_4^+-N was the highest (Table 1), which is consistent with the ranked estimation accuracy of the three nutrients. Owing to its active chemical properties,

NH_4^+-N can be easily converted to nitrites and nitrates [61]. The data used in this study were collected using an automatic monitoring sensor located at the outlet of the watershed. Point-source emissions might lead to a sudden increase of nutrient concentrations in a short time, owing to rapid urbanization [62]. These factors make the variation in riverine nutrient concentrations larger and more difficult to estimate [60], especially for NH_4^+-N. We identified three key indicators (WT, EC and TUR) through the relative importance analysis in Section 3.4. They have always been the top three important in the estimation of TP, TN and NH_4^+-N concentrations. Interestingly, except TUR, there are only weak correlations between WT as well as EC and nutrients. These indicated that WT, EC and TUR have a great impact on the modeling of nutrient concentration dynamics, and the importance could not be fully reflected in the results of correlation analysis. In future research, we may verify our findings above by using different combinations of input indicators. Also, we may evaluate the changes of model estimation accuracy by leaving out relatively less important indicator (such as pH or DO) to develop a more simplified model with minimal impact on model accuracy.

The RF model underestimated higher concentrations. This underestimation occurs frequently when using a machine learning algorithm to estimate numerous variables [4,19,57,60,63]. There are several reasons leading to the model underestimation of the peak nutrient concentration: the occasionally unusual observations or the fact that the five inputs selected for this study did not fully include the indicators affecting nutrient concentrations. Or some peaks were mistakenly removed as outliers when performing the outlier elimination operation.

4.3 Limitations and future agenda

Notwithstanding the success of machine learning in water-quality estimations, some limitations continue to hamper its wider use and impact. One limitation is the model interpretability [64]. Although machine learning models can fit observations well, it is difficult to trace their mechanism of temporal and spatial changes. The main purpose of this study was to develop a regression model that could accurately estimate nutrient concentrations; hence, the physical mechanism of nutrient changes was omitted. We instead explored the uncertainty induced by the sampling frequencies. Therefore, the uncertainties caused by different models were briefly evaluated and without cross-validation. Furthermore, there was only one automatic monitor at the outlet of the watershed studied. Thus, we used the so far water quality indicators only from one location for modelling and analysis. This may not sufficiently reflect all hydrological processes in the watershed.

Considering the continuous implementation of the follow-up work in our study area, this study only used five easily available indicators as data input, which eliminated the need for laboratory experiments. The input indicators can be obtained by sampling and measuring using a portable water-quality monitor along rivers and creeks, or by the sensor located in the outlet of the watershed. However, the convenience of the proposed methodology means that some important physical and chemical parameters (i.e., precipitation, flow, point source discharge, non-point source pollution, some water quality parameters, etc.) that affect the changes of nutrient concentrations were discarded. This is an inevitable problem due to the scarcity of data and the inconsistent time resolution of data from different sources. In subsequent work, we may consider adding more parameters related to the process mechanism as the input data to enhance the interpretability of the machine learning models. In addition, the good estimation results of this study were realized by the excellent fitting ability of machine learning algorithms and high-frequency data. In the future, the model should be continuously optimized or coupled with data-denoising algorithms, such as wavelet transforms, for performance improvement.

5 Conclusions

We developed the RF model to estimate the concentrations of TP, TN, and NH_4^+-N using only five easily obtainable water-quality indicators (i.e., WT, pH, EC, DO, and TUR) as surrogates. We built SVM and BPNN models for comparison to RF, and the results showed that RF performed best. We evaluated the estimation uncertainties related to the sampling frequencies (i.e., 4-hourly, daily, and weekly). There was a significant improvement of model accuracy when the frequency of data input was increased. When using the 4-hourly sampling frequency dataset, RF explained the dynamic variation in TP (79 ± 1.3%), TN (84 ± 0.9%), and NH_4^+-N (75 ± 1.3%). We attribute the accurate estimation of nutrient concentrations to the availability of high-frequency monitoring data, which has shown great potential in water-quality indicator estimations that cannot otherwise be easily realized by daily/weekly sampling routines. Furthermore, EC, TUR, and WT were identified as the key indicators to the estimation of TP, TN, and NH_4^+-N. The RF model is an effective alternative for estimating riverine nutrient concentrations when using high sampling frequency data, which is essential for sustainable water management in watersheds producing scarce water-quality data.

Supporting information

S1 Text. Monitoring sensors for different water quality parameters.
(DOCX)

S2 Text. Selection of hyperparameters for random forest.
(DOCX)

S3 Text. Selection of hyperparameters for Back propagation neural network.
(DOCX)

S4 Text. Selection of hyperparameters for Support vector machine.
(DOCX)

S5 Text. Comparing the performance of our study with other water quality estimation works using different machine learning models.
(DOCX)

S6 Text. Water quality datasets with different sampling frequencies.
(XLSX)

Acknowledgments

The authors would like to extend their gratitude to the Xiamen Environmental Publicity and Education Center under the Xiamen Municipal Bureau of Ecological Environment for providing data that was relevant to this study.

Author Contributions

Conceptualization: Shengyue Chen, Zhenyu Zhang, Jinliang Huang.

Data curation: Juanjuan Lin.

Formal analysis: Shengyue Chen.

Funding acquisition: Jinliang Huang.

Investigation: Shengyue Chen, Juanjuan Lin.

Methodology: Shengyue Chen, Jinliang Huang.

Project administration: Jinliang Huang.

Supervision: Juanjuan Lin.

Validation: Shengyue Chen, Zhenyu Zhang, Jinliang Huang.

Visualization: Shengyue Chen.

Writing – original draft: Shengyue Chen.

Writing – review & editing: Shengyue Chen, Zhenyu Zhang, Jinliang Huang.

References

1. Lei C, Wagner PD, Fohrer N. Effects of land cover, topography, and soil on stream water quality at multiple spatial and seasonal scales in a German lowland catchment. Ecological Indicators. 2021;120. https://doi.org/10.1016/j.ecolind.2020.106940
2. Derot J, Yajima H, Schmitt FG. Benefits of machine learning and sampling frequency on phytoplankton bloom forecasts in coastal areas. Ecological Informatics. 2020;60. https://doi.org/10.1016/j.ecoinf.2020.101174
3. Huang Y, Huang J, Ervinia A, Duan S, Kaushal SS. Land use and climate variability amplifies watershed nitrogen exports in coastal China. Ocean & Coastal Management. 2021; 207:104428. https://doi.org/10.1016/j.ocecoaman.2018.02.024.
4. Shen LQ, Amatulli G, Sethi T, Raymond P, Domisch S. Estimating nitrogen and phosphorus concentrations in streams and rivers, within a machine learning framework. Sci Data. 2020; 7(1):161. Epub 2020/05/30. https://doi.org/10.1038/s41597-020-0478-7 PMID: 32467642; PubMed Central PMCID: PMC7256043.
5. Cassidy R, Jordan P. Limitations of instantaneous water quality sampling in surface-water catchments: Comparison with near-continuous phosphorus time-series data. Journal of Hydrology. 2011; 405(1–2):182–93. https://doi.org/10.1016/j.jhydrol.2011.05.020
6. Harrison JW, Lucius MA, Farrell JL, Eichler LW, Relyea RA. Prediction of stream nitrogen and phosphorus concentrations from high-frequency sensors using Random Forests Regression. Sci Total Environ. 2021; 763:143005. Epub 2020/11/08. https://doi.org/10.1016/j.scitotenv.2020.143005 PMID: 33158521.
7. Bowes MJ, Jarvie HP, Halliday SJ, Skeffington RA, Wade AJ, Loewenthal M, et al. Characterising phosphorus and nitrate inputs to a rural river using high-frequency concentration-flow relationships. Sci Total Environ. 2015; 511:608–20. Epub 2015/01/18. https://doi.org/10.1016/j.scitotenv.2014.12.086 PMID: 25596349.
8. Koparan C, Koc AB, Privette CV, Sawyer CB. In Situ Water Quality Measurements Using an Unmanned Aerial Vehicle (UAV) System. Water. 2018; 10(3):264. https://doi.org/10.3390/w10030264
9. Rode M, Wade AJ, Cohen MJ, Hensley RT, Bowes MJ, Kirchner JW, et al. Sensors in the Stream: The High-Frequency Wave of the Present. Environmental Science & Technology. 2016; 50(19):10297–307. https://doi.org/10.1021/acs.est.6b02155 PMID: 27570873
10. Jiang J, Tang S, Han D, Fu G, Solomatine D, Zheng Y. A comprehensive review on the design and optimization of surface water quality monitoring networks. Environmental Modelling & Software. 2020;132. https://doi.org/10.1016/j.envsoft.2020.104792
11. Lessels JS, Bishop TFA. A post-event stratified random sampling scheme for monitoring event-based water quality using an automatic sampler. Journal of Hydrology. 2020;580. https://doi.org/10.1016/j.jhydrol.2018.12.063
12. Castrillo M, Garcia AL. Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods. Water Res. 2020; 172:115490. Epub 2020/01/24. https://doi.org/10.1016/j.watres.2020.115490 PMID: 31972414.
13. Pellerin BA, Stauffer BA, Young DA, Sullivan DJ, Bricker SB, Walbridge MR, et al. Emerging Tools for Continuous Nutrient Monitoring Networks: Sensors Advancing Science and Water Resources Protection. JAWRA Journal of the American Water Resources Association. 2016; 52(4):993–1008. https://doi.org/10.1111/1752-1688.12386
14. Leigh C, Kandanaarachchi S, McGree JM, Hyndman RJ, Alsibai O, Mengersen K, et al. Predicting sediment and nutrient concentrations from high-frequency water-quality data. PLoS One. 2019; 14(8): e0215503. Epub 2019/08/31. https://doi.org/10.1371/journal.pone.0215503 PMID: 31469846; PubMed Central PMCID: PMC6716630.

15. Adnan RM, Liang Z, Heddam S, Zounemat-Kermani M, Kisi O, Li B. Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs. Journal of Hydrology. 2020;586. https://doi.org/10.1016/j.jhydrol.2019.124371

16. Hunter JM, Maier HR, Gibbs MS, Foale ER, Grosvenor NA, Harders NP, et al. Framework for developing hybrid process-driven, artificial neural network and regression models for salinity prediction in river systems. Hydrology and Earth System Sciences. 2018; 22(5):2987–3006. https://doi.org/10.5194/hess-22-2987-2018

17. Yang S, Yang D, Chen J, Santisirisomboon J, Lu W, Zhao B. A physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data. Journal of Hydrology. 2020;590. https://doi.org/10.1016/j.jhydrol.2020.125206

18. Kasiviswanathan KS, He J, Sudheer KP, Tay J-H. Potential application of wavelet neural network ensemble to forecast streamflow for flood management. Journal of Hydrology. 2016; 536:161–73. https://doi.org/10.1016/j.jhydrol.2016.02.044

19. Noori N, Kalin L, Isik S. Water quality prediction using SWAT-ANN coupled approach. Journal of Hydrology. 2020;590. https://doi.org/10.1016/j.jhydrol.2020.125220

20. Cao X, Liu Y, Wang J, Liu C, Duan Q. Prediction of dissolved oxygen in pond culture water based on K-means clustering and gated recurrent unit neural network. Aquacultural Engineering. 2020;91. https://doi.org/10.1016/j.aquaeng.2020.102122

21. Csábrági A, Molnár S, Tanos P, Kovács J. Application of artificial neural networks to the forecasting of dissolved oxygen content in the Hungarian section of the river Danube. Ecological Engineering. 2017; 100:63–72. https://doi.org/10.1016/j.ecoleng.2016.12.027

22. Ta X, Wei Y. Research on a dissolved oxygen prediction method for recirculating aquaculture systems based on a convolution neural network. Computers and Electronics in Agriculture. 2018; 145:302–10. https://doi.org/10.1016/j.compag.2017.12.037

23. Leong WC, Bahadori A, Zhang J, Ahmad Z. Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM). International Journal of River Basin Management. 2019:1–8. https://doi.org/10.1080/15715124.2019.1628030

24. Yoon H, Hyun Y, Ha K, Lee K-K, Kim G-B. A method to improve the stability and accuracy of ANN- and SVM-based time series models for long-term groundwater level predictions. Computers & Geosciences. 2016; 90:144–55. https://doi.org/10.1016/j.cageo.2016.03.002

25. Heddam S, Kisi O. Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. Journal of Hydrology. 2018; 559:499–509. https://doi.org/10.1016/j.jhydrol.2018.02.061

26. Wang R, Kim JH, Li MH. Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach. Sci Total Environ. 2021; 761:144057. Epub 2020/12/30. https://doi.org/10.1016/j.scitotenv.2020.144057 PMID: 33373848.

27. Lu H, Ma X. Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere. 2020; 249:126169. Epub 2020/02/23. https://doi.org/10.1016/j.chemosphere.2020.126169 PMID: 32078849.

28. Heddam S, Ptak M, Zhu S. Modelling of daily lake surface water temperature from air temperature: Extremely randomized trees (ERT) versus Air2Water, MARS, M5Tree, RF and MLPNN. Journal of Hydrology. 2020;588. https://doi.org/10.1016/j.jhydrol.2020.125130

29. Searcy RT, Boehm AB. A Day at the Beach: Enabling Coastal Water Quality Prediction with High-Frequency Sampling and Data-Driven Models. Environ Sci Technol. 2021; 55(3):1908–18. Epub 2021/01/21. https://doi.org/10.1021/acs.est.0c06742 PMID: 33471505.

30. Belgiu M, Drăguţ L. Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing. 2016; 114:24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

31. Chen K, Chen H, Zhou C, Huang Y, Qi X, Shen R, et al. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water Res. 2020; 171:115454. Epub 2020/01/10. https://doi.org/10.1016/j.watres.2019.115454 PMID: 31918388.

32. Sharafati A, Haji Seyed Asadollah SB, Motta D, Yaseen ZM. Application of newly developed ensemble machine learning models for daily suspended sediment load prediction and related uncertainty analysis. Hydrological Sciences Journal. 2020; 65(12):2022–42. https://doi.org/10.1080/02626667.2020.1786571

33. Solomatine DP, Shrestha DL. A novel method to estimate model uncertainty using machine learning techniques. Water Resources Research. 2009; 45(12). https://doi.org/10.1029/2008wr006839

34. Asadollah SBHS, Sharafati A, Motta D, Yaseen ZM. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. Journal of Environmental Chemical Engineering. 2021; 9(1). https://doi.org/10.1016/j.jece.2020.104599

35. Sharafati A, Asadollah SBHS, Hosseinzadeh M. The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty. Process Safety and Environmental Protection. 2020; 140:68–78. https://doi.org/10.1016/j.psep.2020.04.045

36. Noori R, Yeh H-D, Abbasi M, Kachoosangi FT, Moazami S. Uncertainty analysis of support vector machine for online prediction of five-day biochemical oxygen demand. Journal of Hydrology. 2015; 527:833–43. https://doi.org/10.1016/j.jhydrol.2015.05.046

37. Singh P, Kaur PD. Review on Data Mining Techniques for Prediction of Water Quality. International Journal of Advanced Research in Computer Science. 2017; 8(5):396–401. https://doi.org/10.26483/ijarcs.v8i5.3312

38. Muttil N, Chau K-W. Machine-learning paradigms for selecting ecologically significant input variables. Engineering Applications of Artificial Intelligence. 2007; 20(6):735–44. https://doi.org/10.1016/j.engappai.2006.11.016

39. Shi B, Wang P, Jiang J, Liu R. Applying high-frequency surrogate measurements and a wavelet-ANN model to provide early warnings of rapid surface water quality anomalies. Sci Total Environ. 2018; 610–611:1390–9. Epub 2017/09/01. https://doi.org/10.1016/j.scitotenv.2017.08.232 PMID: 28854482.

40. Berrar D. Cross-Validation. Encyclopedia of Bioinformatics and Computational Biology 2019. p. 542–5.

41. Reitermanova Z, editor Data splitting. WDS; 2010.

42. Rahmati O, Choubin B, Fathabadi A, Coulon F, Soltani E, Shahabi H, et al. Predicting uncertainty of machine learning models for modelling nitrate pollution of groundwater using quantile regression and UNEEC methods. Sci Total Environ. 2019; 688:855–66. Epub 2019/07/01. https://doi.org/10.1016/j.scitotenv.2019.06.320 PMID: 31255823.

43. Yin J, Medellin-Azuara J, Escriva-Bou A, Liu Z. Bayesian machine learning ensemble approach to quantify model uncertainty in predicting groundwater storage change. Sci Total Environ. 2021; 769:144715. Epub 2021/03/20. https://doi.org/10.1016/j.scitotenv.2020.144715 PMID: 33736244.

44. Boulesteix A-L, Janitza S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012; 2(6):493–507. https://doi.org/10.1002/widm.1072

45. Moosavi A, Rao V, Sandu A. Machine learning based algorithms for uncertainty quantification in numerical weather prediction models. Journal of Computational Science. 2021;50. https://doi.org/10.1016/j.jocs.2020.101295

46. Zeng Q, Liu Y, Zhao H, Sun M, Li X. Comparison of models for predicting the changes in phytoplankton community composition in the receiving water system of an inter-basin water transfer project. Environ Pollut. 2017; 223:676–84. Epub 2017/02/16. https://doi.org/10.1016/j.envpol.2017.02.001 PMID: 28196722.

47. Liu S, Tai H, Ding Q, Li D, Xu L, Wei Y. A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. Mathematical and Computer Modelling. 2013; 58(3–4):458–65. https://doi.org/10.1016/j.mcm.2011.11.021

48. Li X, Sha J, Wang Z-l. A comparative study of multiple linear regression, artificial neural network and support vector machine for the prediction of dissolved oxygen. Hydrology Research. 2016; 48(5):1214–25. https://doi.org/10.2166/nh.2016.149%JHydrologyResearch

49. Najah Ahmed A, Binti Othman F, Abdulmohsin Afan H, Khaleel Ibrahim R, Ming Fai C, Shabbir Hossain M, et al. Machine learning methods for better water quality prediction. Journal of Hydrology. 2019;578. https://doi.org/10.1016/j.jhydrol.2019.124084

50. Guyon I, Elisseeff AJJomlr. An introduction to variable and feature selection. 2003; 3(Mar):1157–82.

51. Tripathi M, Singal SK. Use of Principal Component Analysis for parameter selection for development of a novel Water Quality Index: A case study of river Ganga India. Ecological Indicators. 2019; 96:430–6. https://doi.org/10.1016/j.ecolind.2018.09.025

52. Wherry SA, Tesoriero AJ, Terziotti S. Factors Affecting Nitrate Concentrations in Stream Base Flow. Environ Sci Technol. 2021; 55(2):902–11. Epub 2020/12/29. https://doi.org/10.1021/acs.est.0c02495 PMID: 33356185.

53. Kong X, Zhan Q, Boehrer B, Rinke K. High frequency data provide new insights into evaluating and modeling nitrogen retention in reservoirs. Water Res. 2019; 166:115017. Epub 2019/09/07. https://doi.org/10.1016/j.watres.2019.115017 PMID: 31491621.

54. Thomas MK, Fontana S, Reyes M, Kehoe M, Pomati F. The predictability of a lake phytoplankton community, over time-scales of hours to years. Ecology letters. 2018; 21(5):619–28. https://doi.org/10.1111/ele.12927 PMID: 29527797

55. Jiang SY, Zhang Q, Werner AD, Wellen C, Jomaa S, Zhu QD, et al. Effects of stream nitrate data frequency on watershed model performance and prediction uncertainty. Journal of Hydrology. 2019; 569:22–36. https://doi.org/10.1016/j.jhydrol.2018.11.049

56. Liu M, Lu J. Support vector machine-an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river? Environ Sci Pollut Res Int. 2014; 21(18):11036–53. Epub 2014/06/05. https://doi.org/10.1007/s11356-014-3046-x PMID: 24894753.

57. Ali I, Cawkwell F, Dwyer E, Green S. Modeling Managed Grassland Biomass Estimation by Using Multitemporal Remote Sensing Data—A Machine Learning Approach. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2017; 10(7):3254–64. https://doi.org/10.1109/jstars.2016.2561618

58. Chen J, Li K, Tang Z, Bilal K, Yu S, Weng C, et al. A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment. IEEE Transactions on Parallel and Distributed Systems. 2017; 28(4):919–33. https://doi.org/10.1109/TPDS.2016.2603511

59. Zhu X, Vondrick C, Ramanan D, Fowlkes CC, editors. Do We Need More Training Data or Better Models for Object Detection? BMVC; 2012: Citeseer.

60. Lannergard EE, Ledesma JLJ, Folster J, Futter MN. An evaluation of high frequency turbidity as a proxy for riverine total phosphorus concentrations. Sci Total Environ. 2019; 651(Pt 1):103–13. Epub 2018/09/19. https://doi.org/10.1016/j.scitotenv.2018.09.127 PMID: 30227280.

61. Damashek J, Smith JM, Mosier AC, Francis CA. Benthic ammonia oxidizers differ in community structure and biogeochemical potential across a riverine delta. Front Microbiol. 2014; 5:743. Epub 2015/01/27. https://doi.org/10.3389/fmicb.2014.00743 PMID: 25620958; PubMed Central PMCID: PMC4287051.

62. Zhang W, Swaney DP, Hong B, Howarth RW, Li X. Influence of rapid rural-urban population migration on riverine nitrogen pollution: perspective from ammonia-nitrogen. Environ Sci Pollut Res Int. 2017; 24(35):27201–14. Epub 2017/10/02. https://doi.org/10.1007/s11356-017-0322-6 PMID: 28965271.

63. Rajaee T. Wavelet and ANN combination model for prediction of daily suspended sediment load in rivers. Sci Total Environ. 2011; 409(15):2917–28. Epub 2011/05/07. https://doi.org/10.1016/j.scitotenv.2010.11.028 PMID: 21546062.

64. Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, et al. Deep learning and process understanding for data-driven Earth system science. Nature. 2019; 566(7743):195–204. Epub 2019/02/15. https://doi.org/10.1038/s41586-019-0912-1 PMID: 30760912.