

Identification of driving factors of algal growth in the South-to-North Water Diversion Project by Transformer-based deep learning

Jing Qian^a, Nan Pu^b, Li Qian^c, Xiaobai Xue^d, Yonghong Bi^{e,*}, Stefan Norra^f

^a Institute of Applied Geosciences, Karlsruhe Institute of Technology, Karlsruhe, 76131, Germany

^b Institute of Advanced Computer Science, Leiden University, Leiden, 2333, CA, Netherlands

^c Institute of Informatics, Ludwig Maximilian University of Munich, Munich, 80538, Germany

^d MioTech Research, Yingtou Information Technology (Shanghai) Limited, Shanghai, 200120, China

^e State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, 430072, China

^f Institute of Environmental Sciences and Geography, Soil Sciences and Geoecology, Potsdam University Potsdam-Golm, 14476, Germany

ARTICLE INFO

Keywords:

Algal growth
Deep learning
Driving factor determination
Model interpretability
Transformer

ABSTRACT

Accurate and credible identification of the drivers of algal growth is essential for sustainable utilization and scientific management of freshwater. In this study, we developed a deep learning-based Transformer model, named Bloomformer-1, for end-to-end identification of the drivers of algal growth without the needing extensive a priori knowledge or prior experiments. The Middle Route of the South-to-North Water Diversion Project (MRP) was used as the study site to demonstrate that Bloomformer-1 exhibited more robust performance (with the highest R^2 , 0.80 to 0.94, and the lowest RMSE, 0.22–0.43 $\mu\text{g/L}$) compared to four widely used traditional machine learning models, namely extra trees regression (ETR), gradient boosting regression tree (GBRT), support vector regression (SVR), and multiple linear regression (MLR). In addition, Bloomformer-1 had higher interpretability (including higher transferability and understandability) than the four traditional machine learning models, which meant that it was trustworthy and the results could be directly applied to real scenarios. Finally, it was determined that total phosphorus (TP) was the most important driver for the MRP, especially in Henan section of the canal, although total nitrogen (TN) had the highest effect on algal growth in the Hebei section. Based on these results, phosphorus loading controlling in the whole MRP was proposed as an algal control strategy.

1. Introduction

Algae, as a major footstone in the aquatic food chain, have a two-way and complex relationship with water quality. On the one hand, algae can affect water quality, since overgrowth and eventual death of algae cells can adversely influence water quality by producing toxic secondary metabolites and stench thereby affecting the survival of other aquatic organisms (Xia et al., 2019). On the other hand, algae can respond immediately to changes in physico-chemical properties of water, such as variations of temperature and nutrients, which can lead to changes in the species' qualitative and quantitative composition. Consequently, algae can often be used as reliable indicators for water quality assessment (Gökçe et al., 2016). However, increased knowledge and understanding of this relationship is necessary.

Modeling the interactions of algal biomass, expressed as chlorophyll-a (Chl-a) content, with multiple environmental factors based on a

mathematical representation of the ecosystem is an effective approach to analyzing the relationship between water quality and algal growth, including process-based models and data-driven models (Su et al., 2022). Process-based models, such as the Lotka-Volterra model in ecology, are mathematical models that explicitly represent the processes occurring in the target system with equations. In the identification of the driving factors of algal growth, the process-based model is represented as an ecodynamic model that attempts to simulate process-based relationships by combining hydrodynamic processes with ecological processes and takes into account the interactions between multiple subsystems. Although ecodynamic models are capable of systematically representing relationships between a single output and multiple inputs, they usually require significant computational resource (Ralston and Moore, 2020). In addition, equations for process-based models are often derived from theory, but they are not necessarily credible (Knüsel and Baumberger, 2020), which leads to questionable correlations being obtained from the

* Corresponding author.

E-mail address: biyh@ihb.ac.cn (Y. Bi).

<https://doi.org/10.1016/j.watbs.2023.100184>

Received 12 December 2022; Received in revised form 21 February 2023; Accepted 19 April 2023

Available online xxx

2772-7351/© 2023 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

resulting models. In contrast, because this information is hidden in previous data, data-driven models can escape the limitations of theory and reveal patterns of interaction between algal growth and environmental factors from limited data and explain these patterns by correlation theory. Earlier data-driven approaches were empirical regression algorithms that used simple correlation and regression analyses to empirically model the relationship between a single water quality parameter (e.g., TP) and Chl-a (Xia et al., 2019). Since these models were generally unable to represent multi-factor interactions, multivariate analysis methods, such as cluster analysis (CA) and principal component analysis (PCA), were applied to explore algal growth (Bierman et al., 2011; Du et al., 2017; Qian et al., 2021). However, the relationship between environmental factors and algal biomass is, in many cases, non-linear (Nelson et al., 2018). As linear functions are the basis for most correlation coefficients and multivariate methods, they cannot be applied to nonlinear relationships (Su et al., 2022). In this context, machine learning has recently been widely used to understand aquatic ecological processes and to determine the strength of the association between environmental variables and algal growth (Yu et al., 2021; Ly et al., 2021; Deng et al., 2021).

Many studies have confirmed that traditional machine learning tools, such as support vector machine (SVM), logistic regression, extra trees regression (ETR), and multi-linear regression, are effective for the simulation of algal growth, (Su et al., 2022; Park et al., 2015; Liping and Binghui, 2013). As environmental research begins to migrate from small data to big data, the shortcomings of traditional machine learning is becoming more apparent, and deep learning, with its powerful big data processing capabilities, is receiving increased attention (Qian et al., 2022). Deep learning has been employed in previous studies to make predictions regarding Chl-a time series, but has rarely been applied to identify the critical factors associated with algal growth. This is because deep learning operates with less transparency than traditional machine learning and is implicitly expressive about the contributions of each factor. To solve this problem, deep learning models of algal growth are needed.

The Middle Route of the South-to-North Water Diversion Project (MRP) is a national large-scale project in China, which aims to transfer abundant water resources from the south to the north through artificial channels in order to balance the overall water distribution of the country (Zhu et al., 2022). The total length of the channel is 1432 km, including 155 km in Tianjin, serving a population of about 69 million people (Wang et al., 2021). As a long-distance and long-term drinking water supply corridor, water quality safety of the MRP is particularly important. Previous studies have shown that algal growth accelerated in parts of the MRP after 2016, with large clusters of filamentous algae causing problems such as blockage of the basin grate and rapid siltation in front of the outlet sluice (Zhu et al., 2019). Moreover, foul-smelling compounds and algal toxins produced by the siltation of decomposing algal debris also affected water quality levels and threatened water supply safety (Zhu et al., 2022). Consequently, during the 5–10 years since MRP operation, algal biomass has been a major factor affecting water quality. Furthermore, the instability of the overall system has made it difficult to identify the mechanisms and factors that determine algal growth in the MRP. It is noteworthy that most of the world's large water diversion projects are built for irrigation and power generation and that only a small percentage have provision of a drinking water supply as the main purpose (Long et al., 2022). The low attention to water quality changes in these large water diversion projects has resulted in a lack of case studies that can be applied to the management of water quality safety in MRP. Therefore, at this stage, the accurate identification of mechanisms that influence water quality and algae in MRP is lacking. Nevertheless, the effective prediction and management of algal growth are important for success of long distance and long-term drinking water delivery projects such as MRP.

This study aims to accurately and quantitatively identify the driving factors of algal biomass in the MRP with the core of big data mining. Our method involves developing a Transformer-based deep learning model, named Bloomformer-1, which runs on a big data platform derived from long-term manual monitoring data, in order to reveal the driving mechanisms of algal growth in the MRP accurately, transparently, and directly. The findings will be useful for the efficient management and sustainable utilization of the MRP.

2. Materials and methods

2.1. Study area and data collection

A total of nine water quality monitoring stations were evenly spaced along the MRP, labeled P1 to P9, extending from south to north, with P1, P2, P3 and P4 located in the Henan section, P5, P6 and P7 located in the Hebei section, P8 located in the Tianjing section, P9 located in the Beijing Section (Fig. 1). The database used in this study consists of 49 months (August 1, 2018, to August 30, 2022) of water quality monitoring data from each station. Water samples were collected at a depth of 0.5 m, stored at 4 °C, and transported to the laboratory to determine water quality parameters.

The chemical water quality parameters, which comprised total phosphorus (TP), phosphorous-phosphate ($\text{PO}_4 - \text{P}$), total nitrogen (TN), nitrogen-nitrate ($\text{NO}_3 - \text{N}$), nitrogen-ammonia ($\text{NH}_3 - \text{N}$), potassium permanganate index (COD_{Mn}), and total organic carbon (TOC), were determined according to APHA (Zhu et al., 2022). The concentration of Chl-a was used as a response variable in the data-driven methods since it is considered to be an indicator of phytoplankton biomass and was determined according to ASTM D3731-87 (ASTM, 1993).

2.2. Bloomformer-1 model

Transformer is the state-of-the-art solution for natural language processing (NLP) tasks (Wolf et al., 2020). This method takes advantage of the Multi-Head Attention mechanism, which compares each token along the input sequence to other tokens in order to collect and learn dynamic contextual information. Attention is an important part of human cognitive function (Lindsay, 2020), and when faced with large amounts of information, humans can readily adjust the level of focus on the information they received to analyze it more accurately and efficiently. The essence of the attention mechanism was to provide weights. An attention function could be interpreted as mapping a Q (query) and a string of K(key)-V(value) to an output, where Q, K, V, and output were vectors (Vaswani et al., 2017). The attention could be represented as:

$$\text{Output}_{\text{Attention}} = \text{Attention}(Q, K, V)$$

Multi-Head Attention was the projection of Q, K, and V by h different linear transformations. The different attention results were then stitched together, which could be represented as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

where

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

$$W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v} \text{ and } W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$$

In the appealed Attention mechanism, the weights were the direct weight correspondence between the input and output vectors, implying that the weight calculation required the participation of the output vectors. In contrast, the weight of Self-Attention was a weight relation-ship between the input vectors internally, which did not require the

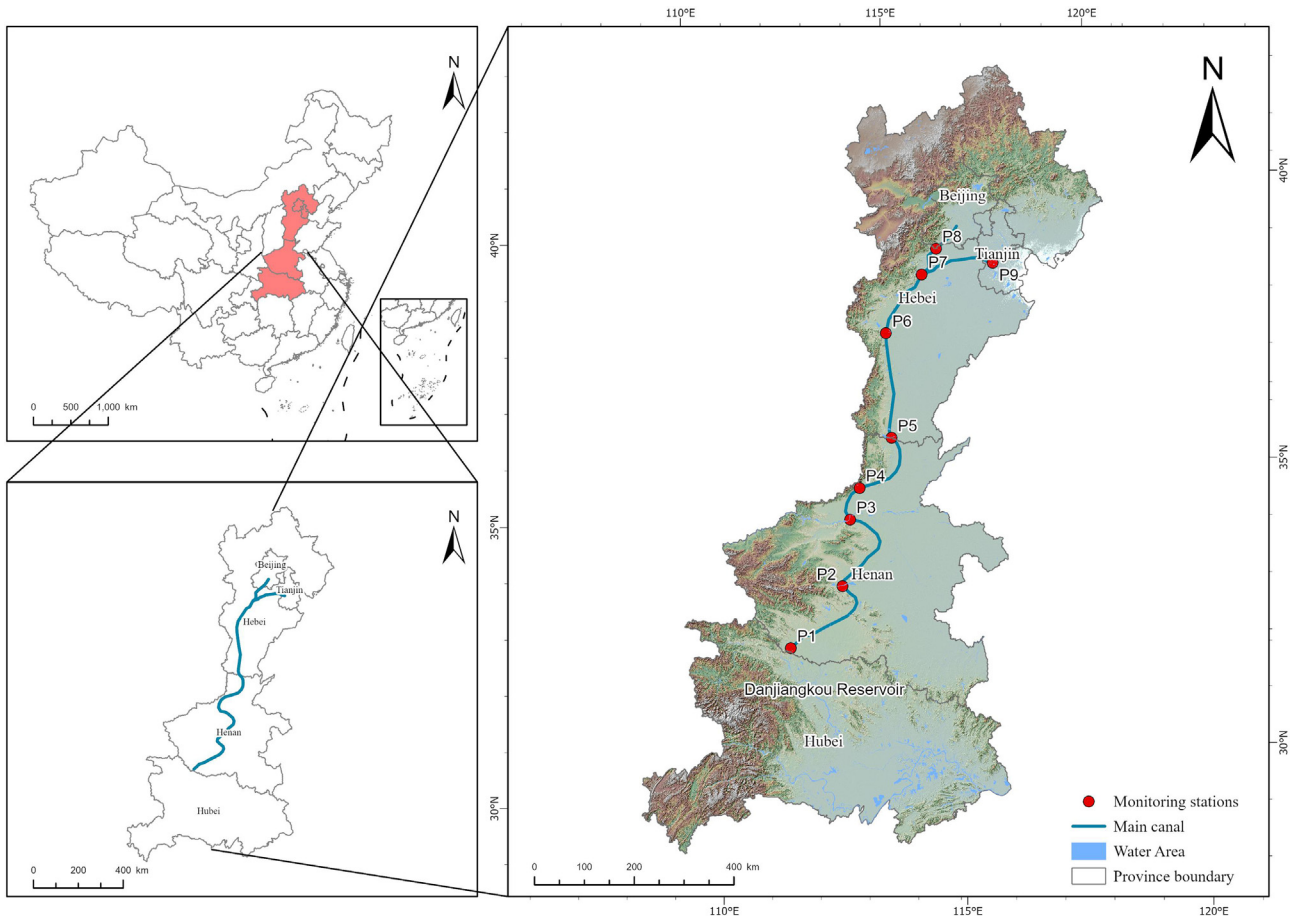


Fig. 1. Sketch map of sampling stations distribution in the middle section of the South-North Water Diversion Project.

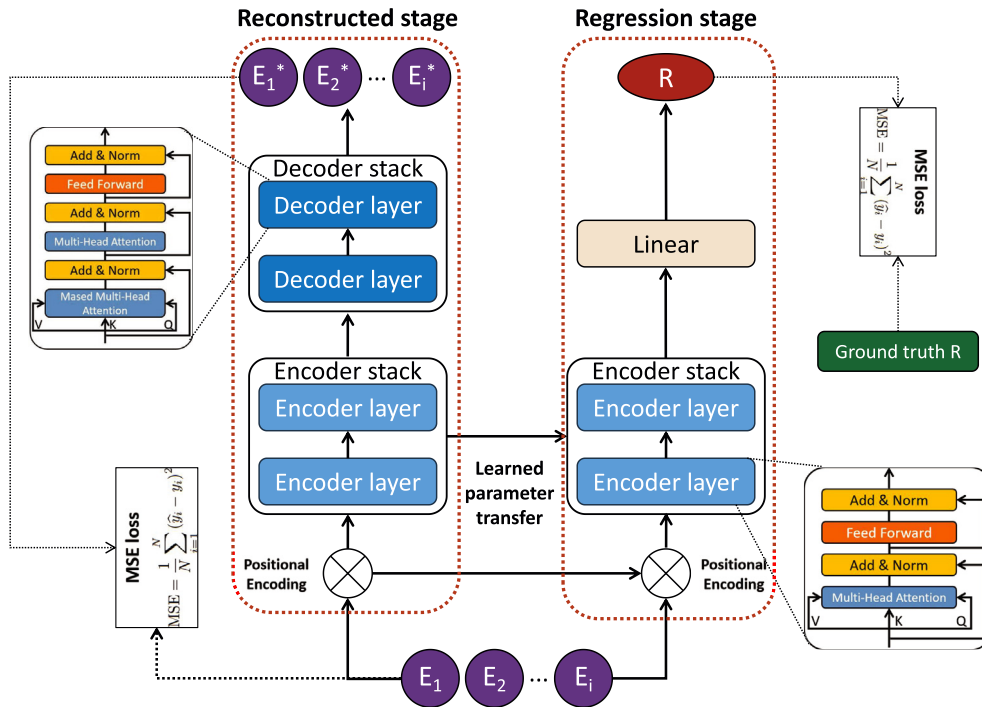


Fig. 2. The framework and architecture of Bloomformer-1.

participation of the output vectors. Therefore, the multi-head-self-attention meant Q , K , and V were the same.

In this study, we used the scaled dot-product to calculate Attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where d_k was the vector dimension in both Q and K .

The encoder consisted of N same units (Fig. 2). Each unit consists of two sub-layers, the multi-head-self-attention layer, and the fully connected feed-forward network, where each sub-layer was processed with the residual connection “Add” and normalization “Norm”. The output of the sub-layer could be represented as:

$$\text{Output}_{\text{Sublayer}} = \text{Norm}(x + F(x))$$

Where $F(x)$ was a function of the sublayer itself, multi-head-self-attention, or fully connected feed-forward network.

The fully connected feed-forward network provided a non-linear transformation consisting of two linear transformations with the active function ReLu (Agarap, 2018). Compared with the encoder, the decoder added another MHSA layer (Fig. 2). A mask operation (Fan et al., 2021) was applied to this multi-head self-attention layer in order to prevent the model from being exposed to future information during training.

Because algal growth is a multi-factorial process, the determination of the driving factors of algal growth is a typical multivariate regression task. The key to solving this problem is to learn the spatial relationships to understand how the variables are related to each other. However, the standard Transformer is not designed for this because it treats the value of each variable at a given time period as a single marker on its graph: each variable cannot have its own view of the context it should prioritize (Grigsby et al., 2021). Therefore, we developed Bloomformer-1 for studying spatial relationships based on Transformer. The improved method first converted the context sequence in the database into a long spatial sequence. This sequence was also transposed to obtain the corresponding long spatial sequence. The sequence was then processed with a Transformer-based encoder-decoder architecture to obtain the predicted values for each variable. Finally, the predicted values were repackaged into their original format and trained to minimize prediction error metrics. The training framework of Bloomformer-1 consists of a reconstruction stage and a regression stage. The reconstruction task is an unsupervised pre-training and a reconstruction of the explanatory variables through the connected encoder and decoder stacks to extract their robust and compact features. The parameters of the encoder stack and position encoding obtained by the reconstruction task are shared with the corresponding part of the regression task. In this study, the number of units in encoder and decoder layer is 8, which represented the 7-dimensional water quality parameters and the 1-dimensional station location information. When performing the substation task, the station location information was the station number corresponding to each water quality parameter, from 1 to 9. When performing the whole MRP task, the station location information was set to 1. Mean square error (MSE, Supplementary material) was selected as the loss function both in the reconstructed stage and the regression stage. The framework and architecture of Bloomformer-1 is shown in Fig. 2. The MHSA mechanism of Bloomformer-1 allows the results of driving factor identification to be obtained during model training forward propagation direction and simultaneously derived.

2.3. Multiple linear regression

Multiple linear regression (MLR) is one of the typical traditional machine learning models that can be used to predict the result of an answer variable using a number of explanatory variables (Maulud and Abdulazeez, 2020). For the purpose of verifying performance, an MLR model was used in this study to compare with Bloomformer-1. The MLR

model was built by using the Scikit-learn function from the Python package. The parameter to be tuned was the degree of the polynomial features. The driving factor analysis methods for MLR was sensitivity analysis (SA) (Saltelli, 2002).

2.4. Support vector regression

Support vector regression (SVR) is a powerful traditional learning machine for searching the relationship between the answer variable and several explanatory variables, including linear and non-linear correlations. The SVM approach is to map the training data non-linearly into a high-dimensional feature space and then construct a separated hyper-plane there with maximum margin (Awad and Khanna, 2015). This study employed the SVR as a comparative model to assess the performance of Bloomformer-1. The SVR model was derived by calling the function in the Scikit-learn package in Python. Radial basis functions were selected as kernels because they provided better performance through the kernel test. The parameters that needed to be tuned in this study were the regularization parameter and the Kernel coefficient. The driving factor analysis methods for SVR was sensitivity analysis (SA) (Saltelli, 2002).

2.5. Gradient boosting regression tree

The gradient boosting regression tree (GBRT) algorithm is a combination of the classification and regression (CART) algorithm and the gradient boosting (GB) algorithm (He et al., 2013). CART allows for the modeling of non-linear relationships without requiring a priori information about the probability distribution of the variables (Nie et al., 2021). The gradient boosting algorithm combines weak learners by iteratively focusing on the error generated at each step until a suitable strong learner is obtained as a sum of successive weak learners (Friedman, 2001). The regression tree generated by the CART algorithm was used as the weak learner and was added to the model to correct errors in the previous model, thereby improving the accuracy of the model. This study employed GBRT as a comparative model to assess the performance of Bloomformer-1. The GBRT model was derived by calling the function in the Scikit-learn package in Python. The driving factor analysis methods for GBRT is to calculate the relative importance to the input variables, the idea being to score each input variable by estimating the reduction in relative variance (Su et al., 2022).

2.6. Extra trees regression

Extra trees regression (ETR) builds a collection of the unpruned decision or regression trees based on a classical top-down procedure that does not require a known underlying distribution of parameters or associated assumptions (Geurts et al., 2006). The main difference between this method and traditional tree ensemble methods is that it splits the nodes randomly and grows the tree based on the original training data set rather than using a bootstrap method. With these two features, ETR is able to produce outputs with lower variance and higher generalization than traditional tree-based models. In this study, the ETR was used to evaluate the performance of Bloomformer-1 as a comparative model. The ETR model was derived by calling the function in the Scikit-learn package in Python. As for GBRT, the driving factor analysis methods for ETR is to calculate the relative importance to the input variables (Su et al., 2022).

2.7. Training and performance evaluation of model

Data from each of the nine water quality monitoring stations (P1 to P9) were fed into the appealing model for training to identify the drivers of algal growth at each water quality monitoring station. Chl-a and the other water quality parameters described previously were placed in the models as responses and explanatory variables, respectively. Before entering all data into the model, data normalization was performed to

ensure equality in model comparisons. Data normalization followed the Z-equation (See Supplementary material).

Evaluation of model performance is a critical step prior to practical application. The data set was divided into a training set and a test set according to the rule of randomly taking one step out of every five, which means 80% of the whole data set was used to train the model and 20% was used to test the model performance. A tenfold cross-validation was introduced to avoid over-fitting in the training phase. For the purpose of evaluating the accuracy and stability of each regression model, two indicators were used on the test set: coefficient of determination (R^2) and root mean square error (RMSE), following the equations:

$$R^2 = 1 - \frac{\sum_{i=0}^n (y_i - \hat{y}_i)^2}{\sum_{i=0}^n (y_i - \bar{y}_i)^2}$$

$$RMSE = \sqrt{\frac{\sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2}{n}}$$

2.8. Operation environment

The experiment was carried out on a PC with the following features: Hard-ware: CPU i7-6950X, RAM 64GB, dual GeForce RTX 3090, VRAM 24GB; Software: Ubuntu 20.04, Python3.6, Pytorch 1.10.0, Numpy 19.2.

3. Results

3.1. Model performance evaluation

The performance of regression models directly determined the accuracy and plausibility of the driver identification. After optimizing the proposed models, we compared the performance of five machine learning models across all monitoring stations using R^2 and RMSE in a tenfold cross-validation. The results for model performance are summarized in Table 1. The comparison between model simulation and the ground truth is shown in Figs. 3 and 4. In order to describe the training process of Bloomformer-1 more intuitively, the loss values during the training process are shown in Fig. S1.

The results of P1, P2, and P3 showed that Bloomformer-1 performed much better than the four traditional machine learning models because

Table 1
Results of model performance evaluation.

Stations	Indicator ^a	Bloomformer-1	ETR	GBRT	SVR	MLR
P1	R^2	0.85	0.75	0.72	0.63	0.42
	RMSE	0.32	0.56	0.57	0.60	0.73
P2	R^2	0.80	0.66	0.51	0.63	0.25
	RMSE	0.43	0.62	0.68	0.63	0.82
P3	R^2	0.83	0.70	0.39	0.58	0.39
	RMSE	0.40	0.59	0.69	0.64	0.79
P4	R^2	0.89	0.84	0.68	0.46	0.35
	RMSE	0.33	0.52	0.62	0.61	0.76
P5	R^2	0.90	0.89	0.78	0.88	0.49
	RMSE	0.30	0.50	0.58	0.51	0.71
P6	R^2	0.89	0.85	0.74	0.88	0.46
	RMSE	0.26	0.45	0.49	0.43	0.68
P7	R^2	0.94	0.94	0.85	0.92	0.68
	RMSE	0.23	0.43	0.47	0.45	0.66
P8	R^2	0.94	0.91	0.84	0.89	0.71
	RMSE	0.22	0.43	0.48	0.44	0.62
P9	R^2	0.93	0.91	0.89	0.86	0.62
	RMSE	0.28	0.46	0.48	0.49	0.68
Whole MRP	R^2	0.85	0.79	0.73	0.80	0.39
	RMSE	0.35	0.54	0.55	0.51	0.70

The bold values represent the best regression results.

^a Unit of RMSE is $\mu\text{g/L}$.

the difference in R^2 values between them was greater than 0.1. There was also a significant difference in RMSE values (e.g., in P1, Bloomformer-1 had an R^2 value of 0.85, while the four traditional machine learning models had R^2 values less than or equal to 0.75; Bloomformer-1 had an RMSE value of 0.32, while the other models had RMSE values greater than or equal to 0.56. The RMSE value of Bloomformer-1 was 0.32, while the RMSE values of the other models were all greater than or equal to 0.56).

According to the results of P4, P5, P6, P8, and P9, Bloomformer-1 showed relatively high performance. Although the difference with ETR in R^2 values was small (0.03–0.06), it still had a significant advantage in RMSE values (e.g., Bloomformer-1 had an RMSE value of 0.33 in P4, while ETR had the lowest RMSE of 0.52 among the four traditional machine learning models). In P7, except for MLR, the other three traditional machine learning models showed better performance, especially the R^2 value of ETR which was the same as Bloomformer-1 at 0.94. However, Bloomformer-1 still had a significant advantage in RMSE values (Bloomformer-1 0.23, ETR 0.43, GBRT 0.47, SVR 0.45, MLR 0.66). Consistent with the results from the individual stations, Bloomformer-1 showed superior performance on the whole MRP, as evidenced by the higher R^2 values (0.85) and lower RMSE values (0.35). In summary, Bloomformer-1 showed the highest R^2 with the lowest RMSE across all stations compared to traditional machine learning models and was, therefore, the best model in terms of performance to describe the relationship between Chl-a concentration and the water quality parameters.

3.2. Driving factors of algal growth

The driving factors of algal growth in the MRP based on the attention mechanism of Bloomformer-1 are shown in Fig. 5. In P1, P2 and the whole MRP, the most dominant driving factor of algal growth was TP, with 18.73%, 19.20% and 22.28%, respectively. It is noteworthy that $\text{PO}_4 - \text{P}$ also exhibited a very close occupancy rate in the whole MRP, at 16.09%. The results for P5, P6, P8, and P9 showed that the major driving factor of algal growth at these four stations was $\text{NO}_3 - \text{N}$ with 20.24%, 28.27%, 20.16%, and 17.16%, respectively. In P4 and P7, TN was the main driving factor of algal growth, with 22.16% and 17.96%, respectively. The results of P3 differed from the others, with 23.84% of $\text{NH}_3 - \text{N}$ as the most dominant driving factor of algal growth.

4. Discussion

4.1. Model performance

Inferring causation from correlation and determining the explanatory variables associated with the response variables is the basis for traditional model building, which requires a great deal of a priori knowledge and background information about the domain (Xia et al., 2019; Su et al., 2022). In traditional machine learning, feature extraction is manual-based and has limited learning capability, thus requiring the input terms (explanatory variables) have a clear one-way correlation with the response variables, which implies a high reliance on a priori knowledge. However, some explanatory variables are difficult to determine in practical applications, such as COD_{Mn} in this study. The relationship between COD_{Mn} and algal growth is bidirectional and complex (Li et al., 2020; Yan et al., 2016). The foundation of COD_{Mn} as an explanatory variable depends on which direction of the relationship is dominant, which requires a priori knowledge as well as prior experiments. Bloomformer-1 employs a combination of encoder and decoder structures as well as the MHSA mechanism to automatically extract features from raw data and to fully understand the raw data at the same time. This full understanding means that the complex relationship between COD_{Mn} and algal growth in the raw data is mined and quantified. In this way, a rigorous correlation analysis is not required before using Bloomformer-1. Moreover, building a model with excellent fitting performance is the first and most critical step to identify the driving factors

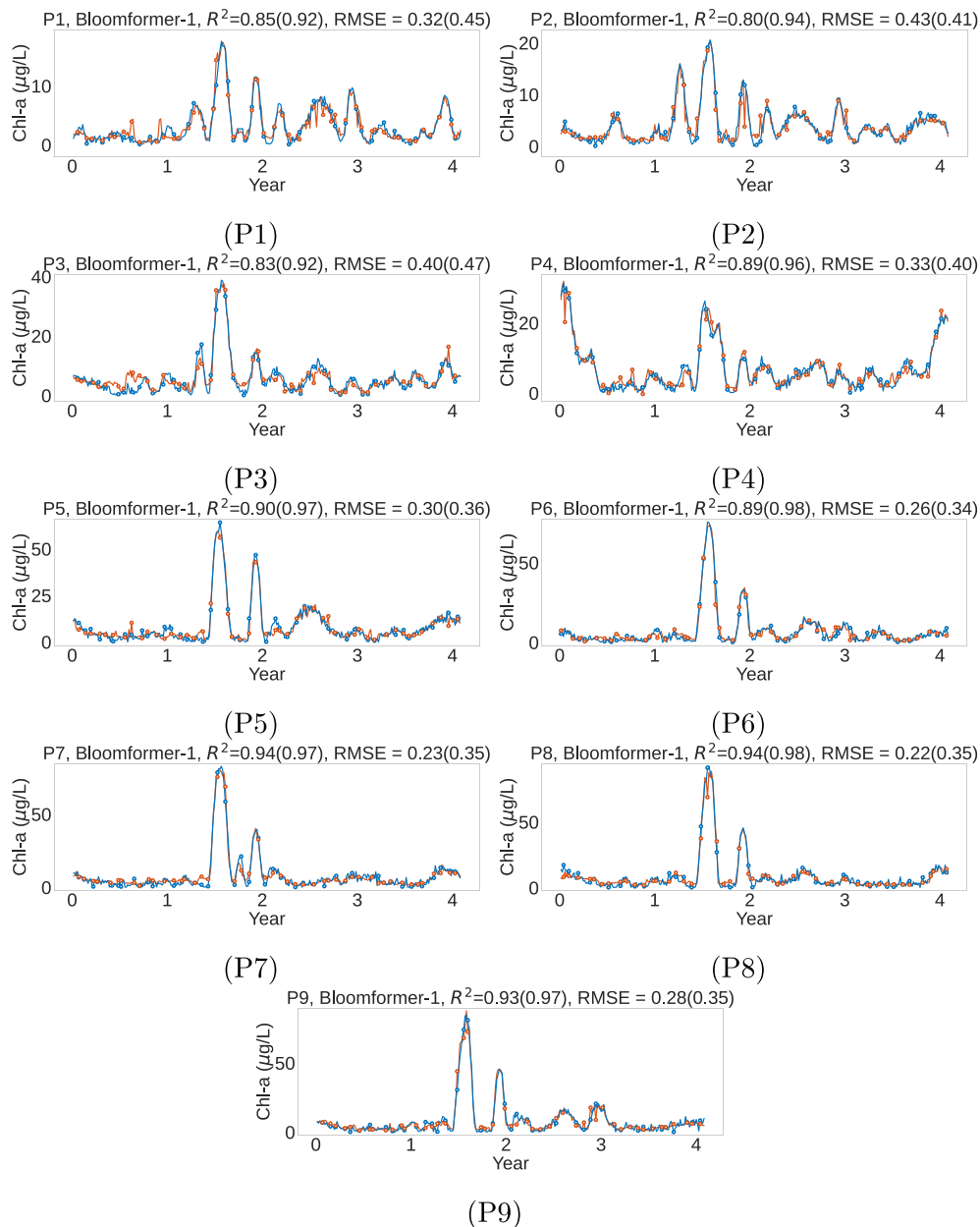


Fig. 3. Performance of Bloomformer-1 in P1–P9 (blue lines are observations, red lines are model simulations). The circles are the test set, where the blue circles are the true values and the red circles are the predicted values. The blue line, except for the blue circles, is the training set. Numbers show RMSE and R^2 for model prediction and training data (inside brackets).

of response variables. As a state-of-the-art deep learning model, Bloomformer-1 has an advantage in the accuracy of model fitting with R^2 (0.80–0.94). Compared with traditional machine learning, deep learning is more advanced and has a stronger learning ability to automatically extract, analyze and understand useful information from raw data to obtain better results (Chauhan and Singh, 2019; Janiesch et al., 2021).

In the present study, when training traditional machine learning models, each explanatory variable was completely independent, for example, each decision tree that made up the ETR was unrelated to each other. This meant that the traditional machine learning models only focused on the logical relationship between each explanatory variable and the corresponding variable, ignored the additional effects of the interactions between explanatory variables on the corresponding variable. Consequently, the traditional machine learning models could only partially identify the drivers of algal growth, because algal growth is not only related to a single water quality parameter, but also to the

interactions between multiple water quality parameters in different spatial-temporal dimensions. The Transformer structure in Bloomformer-1 had the MHSA mechanism that could simultaneously focus on all relationship changes (Vaswani et al., 2017). Therefore, Bloomformer-1 can identify the drivers more reliably than traditional machine learning models.

4.2. Model interpretability

Model interpretability represents trustworthiness (Ridgeway et al., 1998), which can be expressed in terms of transferability and understandability (Lipton, 2016).

Transferability represents the ability to transfer learned skills to unfamiliar environments, especially in non-stationary environments (Lipton, 2016). In this study, Bloomformer-1 outperformed four traditional machine learning models on the test data set and was able to easily cope

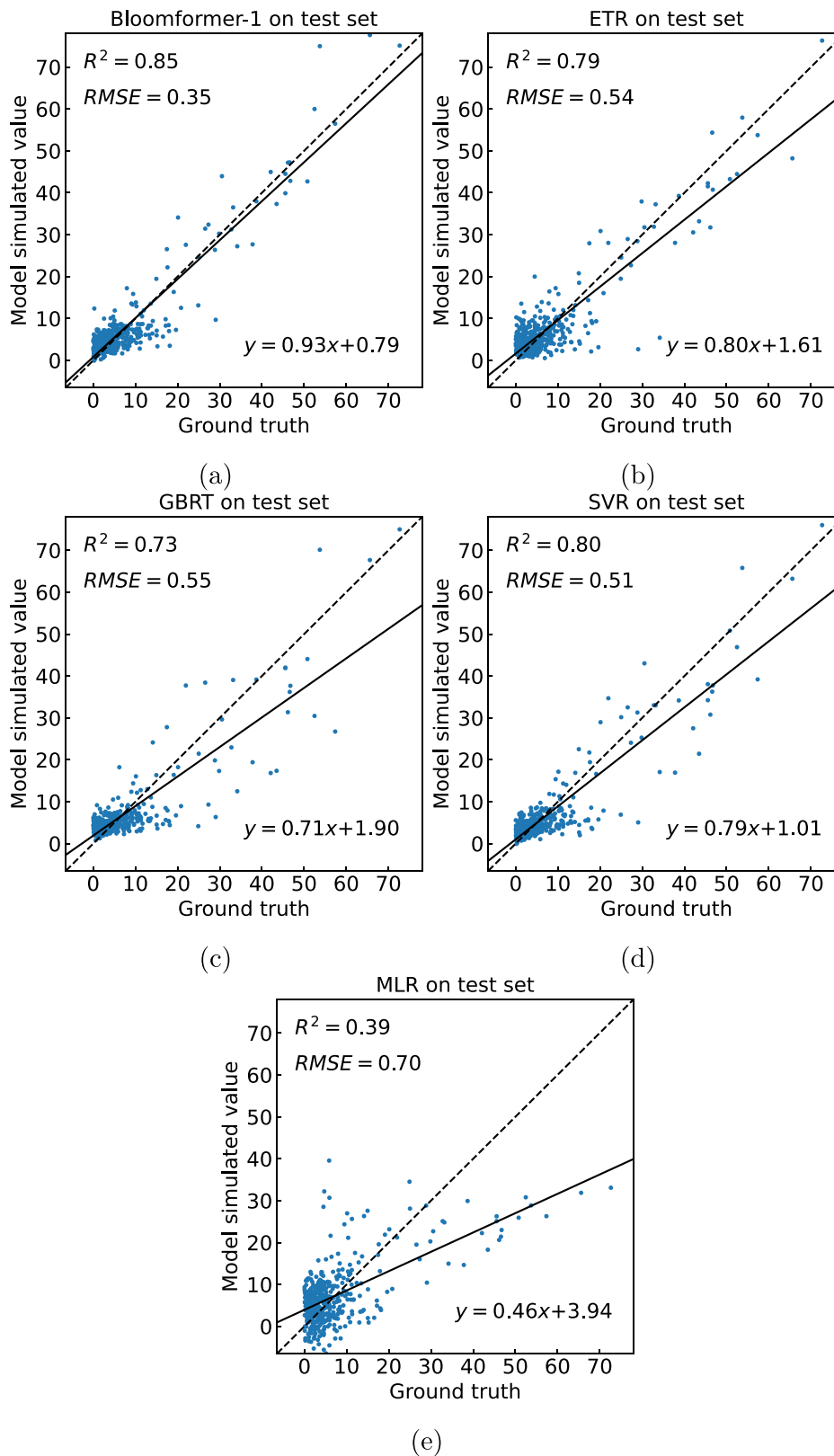


Fig. 4. Model performance evaluation in the whole MRP, where (a), (b), (c), (d) and (e) represent the test results of the Bloomformer-1, ETR, GBRT, SVR and MRL, respectively.

with abrupt changes in Chl-a concentration whereas traditional machine learning models were unable to do so (e.g., P3 in February 2020). These findings demonstrate that Bloomformer-1 has superior transferability.

Understandability represents our ability to understand how a model

works (Lipton, 2016). When dealing with multidimensional variables, SVR is difficult to understand because the human brain is unable to visualize the hyperplane when the number of variables have more than three dimensions. Both GBRT and ETR also showed low

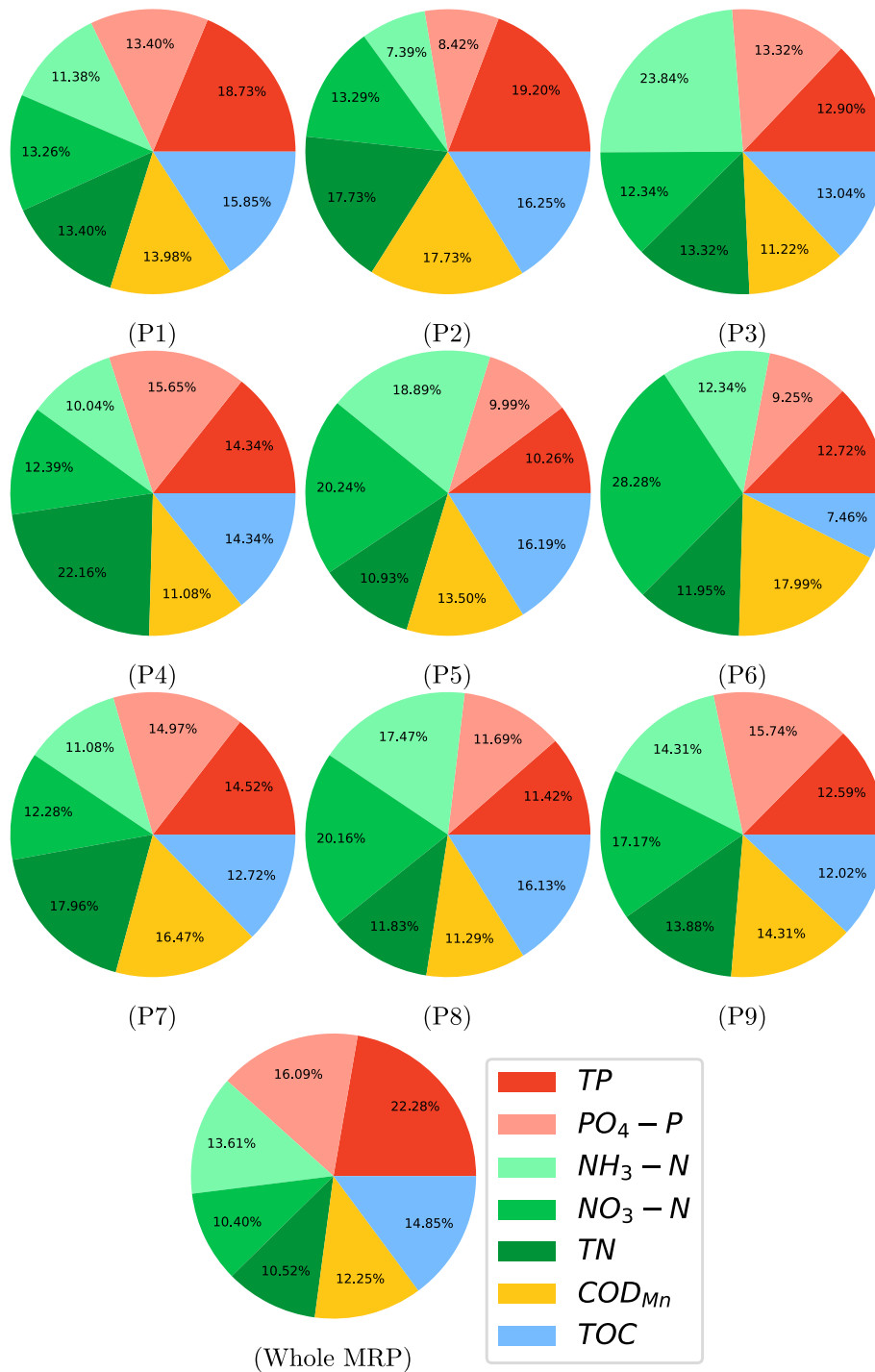


Fig. 5. Driving factors of algal growth at each of the sampling stations (P1–P9), and in the whole MRP, based on Bloomformer-1 modelling.

understandability. The direction and/or shape of covariate effects usually cannot be obtained by the simple interpretation of GBRT (Welchowski et al., 2022). ETR uses the same principles as Random Forest, except that the selection of attributes and cut points is strongly randomized when splitting the tree nodes (Geurts et al., 2006). Random forest is considered as a black box model in many studies (Wright, 2018), so ETR based on the same principle can also be considered as a black box model. On the contrary, Bloomformer-1 possessed a relatively high level of understandability. First, Bloomformer-1 worked by synthesizing the computational resources derived from the analysis and continuously adjusting the weights of each computational resource to obtain the desired results. This work pattern imitates that of humans and is therefore easy to

understand. Secondly, the attribution algorithm (Hao et al., 2021) of the self-attentive mechanism could provide an interpretable description of the information interactions within Bloomformer-1 and construct attribution trees to visualize the direct information interactions in different layers. As a result, Bloomformer-1 has a high degree of interpretability, and the obtained results are highly applicable to real-world situations.

4.3. Driving factors of algal growth

Nutrients play a vital role in algal growth, in particular their supply and its variability affect algal biomass and net productivity (Yang et al., 2016; Koeller et al., 2009). Among them, nitrogen (N) and phosphorus

(P) are essential elements for algae (Hecky and Kilham, 1988). Nitrogen to phosphorus ratios (N:P) are often used to determine the nutrient limitation status of water bodies (Redfield, 1963), but difficulties remain because the optimal N:P ratio varies considerably, i.e., from 4 to 133, for different water bodies (Klausmeyer et al., 2004). Previous studies on MRP have recognized phosphorus as the limiting factor for algal growth, but it was not definitive that it was the most critical nutrient limitation (Nong et al., 2020). The results of this study indicated that TP was the most critical factor in the whole MRP. These results agreed with other studies on algal growth and further confirmed the driving role of nutrients on algal growth.

Although the water quality of the MRP has been good and stable since 2014, the nutrient load has been increasing. Besides the increasing nutrient load of Danjiangkou reservoir, the rain runoff, dry and wet deposition along the channel were the important mechanisms of nutrient input (Wang et al., 2021; Nong et al., 2020). Inundation of farmland and mountainous areas led to the release of nitrogen, phosphorus and other nutrients from the soil into the water, resulting in increased nutrient concentrations in the Danjiangkou Reservoir. In recent years, rainfall along the MRP has increased and this, coupled with dry and wet deposition, has resulted in more nutrients, both from the land and the air, being deposited into the MRP, which made the rich material basis for algae rapid growth. It could be deduced that nutrient control, especially phosphorus, should be important strategy for controlling algal growth and maintaining water quality stability.

5. Future work

Bloomformer-1, as an advanced deep learning model, has obvious performance advantages over traditional machine learning models in processing high volume as well as high dimensional data (Fig. S2). As the database used in this study has medium capacity and dimensionality, the potential of Bloomformer-1 was not fully realized, which was also why traditional machine learning models were able to perform well in some scenarios. In addition, due to the complexity and size of the MRP, a deeper understanding of the relationship between algal growth and water quality is necessary. Therefore, future work should focus on building databases with higher data capacity and dimensionality (including collecting physical and hydrological data), increasing the density of monitoring stations, and using automated monitoring equipment. Using such databases, Bloomformer-1, with its excellent self-learning capability, could make more relevant and timely conclusions regarding the management of algal growth in the MRP.

6. Conclusion

Bloomformer-1, a deep learning-based Transformer model for end-to-end identification of the drivers of algal growth without the need for extensive prior knowledge and prior experiments, achieved the highest R^2 (0.80–0.94) and lowest RMSE (0.22–0.43 $\mu\text{g/L}$) on both individual subsites and full-line simulations in the MRP compared with traditional machine learning models, namely ETR, GBRT, SVR and MLR. Bloomformer-1 also had higher interpretability, implying that Bit was trustworthy and that the results obtained from this model could be directly applied to real-world scenarios. TP was the most important driver for the MRP. Phosphorus control and reduction would be an important strategy for controlling algal growth and maintaining water quality stability in the MRP.

Funding

This research was Jointly funded by National Key R&D plan (No.2021YFC3200900) and National Natural Science Foundation of China (No.31971477).

Credit author statement

Conceptualization: Jing Qian.
 Data curation: Jing Qian, Li Qian.
 Formal analysis: Jing Qian and Nan Pu.
 Funding acquisition: Yonghong Bi and Stefan Norra.
 Investigation: Jing Qian, Nan Pu, Li Qian and Xiaobai Xue.
 Methodology: Jing Qian.
 Project administration: Yonghong Bi and Stefan Norra.
 Resources: Yonghong Bi.
 Software: Jing Qian, Nan Pu and Li Qian.
 Supervision: Stefan Norra and Yonghong Bi.
 Visualization: Jing Qian, Nan Pu and Li Qian.
 Writing – original draft: Jing Qian.
 Writing – review & editing: Stefan Norra and Yonghong Bi.

Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

We appreciate the help from Yuxuan Zhu and Gang Ruan with the experiments. We would also like to thank Di Wang for proofreading.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.watbs.2023.100184>.

References

- Agarap, A.F., 2018. Deep learning using rectified linear units (ReLU). arXiv e-prints arXiv: 1803.08375arXiv:1803.08375.
- ASTM, 1993. Standard Practices for Measurement of Chlorophyll Content of Algae in Surface Waters, pp. 3731–3787.
- Awad, M., Khanna, R., 2015. Support vector regression. In: *Efficient Learning Machines*. Springer, pp. 67–80.
- Bierman, P., Lewis, M., Ostendorf, B., Tanner, J., 2011. A review of methods for analysing spatial and temporal patterns in coastal water quality. *Ecol. Indicat.* 11 (1), 103–114. <https://doi.org/10.1016/j.ecolind.2009.11.001>.
- Chauhan, N.K., Singh, K., 2019. A review on conventional machine learning vs deep learning. In: 2018 International Conference on Computing, Power and Communication Technologies, GUCON, vol. 2018, pp. 347–352. <https://doi.org/10.1109/GUCON.2018.8675097>.
- Deng, T., Chau, K.W., Duan, H.F., 2021. Machine learning based marine water quality prediction for coastal hydro-environment management. *J. Environ. Manag.* 284 (December 2020), 112051. <https://doi.org/10.1016/j.jenvman.2021.112051>.
- Du, X., Cai, Y., Wang, S., Zhang, L., 2017. Overview of deep learning. In: *Proceedings - 2016 31st Youth Academic Annual Conference of Chinese Association of Automation*, vol. 2016. YAC, pp. 159–164. <https://doi.org/10.1109/YAC.2016.7804882>.
- Fan, Z., Gong, Y., Liu, D., Wei, Z., Wang, S., Jiao, J., Duan, N., Zhang, R., Huang, X., 2021. Mask attention networks: rethinking and strengthen trans-former. In: *NAACL-HLT 2021-2021 Conference of the North American Chap-Ter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pp. 1692–1701. <https://doi.org/10.18653/v1/2021.naacl-main.135> arXiv:2103.13597.
- Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29 (5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>.
- Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. *Mach. Learn.* 63 (1), 3–42. <https://doi.org/10.1007/s10994-006-6226-1>.
- Gökçe, D., 2016. Algae as an indicator of water quality. In: Thajuddin, N., Dhanasekaran, D. (Eds.), *Algae*, IntechOpen, Rijeka, Ch. 4, pp. 81–101. <https://doi.org/10.5772/62916>.
- Grigsby, J., Wang, Z., Qi, Y., 2021. Long-range Transformers for Dynamic Spatiotemporal Forecasting arXiv preprint arXiv:2109.12218.
- Hao, Y., Dong, L., Wei, F., Xu, K., 2021. Self-attention attribution: interpreting information interactions inside transformer. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 12963–12971.
- He, Q., Kamarianakis, Y., Jintanukul, K., Wynter, L., 2013. Incident duration prediction with hybrid tree-based quantile regression. In: *Advances in Dynamic Network Modeling in Complex Transportation Systems*. Springer, pp. 287–305.
- Hecky, R.E., Kilham, P., 1988. Nutrient limitation of phytoplankton in freshwater and marine environments: a review of recent evidence on the effects of enrichment.

- Limnol. Oceanogr. 33 (4part2), 796–822. <https://doi.org/10.4319/lo.1988.33.4part2.0796>.
- Janiesch, C., Zschiech, P., Heinrich, K., 2021. Machine learning and deep learning. *Electron. Mark.* 31 (3), 685–695. <https://doi.org/10.1007/s12525-021-00475-2>.
- Klausmeyer, C.A., Litchman, E., Daufreshna, T., Levin, S.A., 2004. Optimal nitrogen-to-phosphorus stoichiometry of phytoplankton. *Nature* 429 (6988), 171–174. <https://doi.org/10.1038/nature02454>.
- Knüsel, B., Baumberger, C., 2020. Understanding climate phenomena with data-driven models. *Stud. Hist. Philos. Sci.* 84, 46–56. <https://doi.org/10.1016/j.shpsa.2020.08.003>.
- Koeller, P., Fuentes-Yaco, C., Platt, T., Sathyendranath, S., Richards, A., Ouellet, P., Orr, D., Skúladóttir, U., Wieland, K., Savard, L., Aschan, M., 2009. Basin-scale coherence in phenology of shrimps and phytoplankton in the North Atlantic Ocean. *Science* 324 (5928), 791–793. <https://doi.org/10.1126/science.1170987>.
- Li, Y., Nwankwegu, A.S., Huang, Y., Norgbey, E., Paerl, H.W., Acharya, K., 2020. Evaluating the phytoplankton, nitrate, and ammonium interactions during summer bloom in tributary of a subtropical reservoir. *J. Environ. Manag.* 271 (May), 110971. <https://doi.org/10.1016/j.jenvman.2020.110971>.
- Lindsay, G.W., 2020. Attention in psychology, neuroscience, and machine learning. *Front. Comput. Neurosci.* 14 (April), 1–21. <https://doi.org/10.3389/fncom.2020.00029>.
- Liping, W., Binghui, Z., 2013. Prediction of chlorophyll-a in the Daning River of Three Gorges Reservoir by principal component scores in multiple linear regression models. *Water Sci. Technol.* 67 (5), 1150–1158. <https://doi.org/10.2166/wst.2013.679>.
- Lipton, Z.C., 2016. The Mythos of Model Interpretability, arXiv E-Prints arXiv:1606.03490arXiv:1606.03490.
- Long, Y., Feng, M., Li, Y., Qu, J., Gao, W., 2022. Comprehensive risk assessment of algae and shellfish in the middle route of South-to-North Water Diversion Project. *Environ. Sci. Pollut. Control Ser.* 29 (52), 79320–79330. <https://doi.org/10.1007/s11356-022-21210-0>.
- Ly, Q.V., Nguyen, X.C., Lê, N.C., Truong, T.D., Hoang, T.H.T., Park, T.J., Maqbool, T., Pyo, J.C., Cho, K.H., Lee, K.S., Hur, J., 2021. Application of Machine Learning for eutrophication analysis and algal bloom prediction in an urban river: a 10-year study of the Han River, South Korea. *Sci. Total Environ.* 797, 149040. <https://doi.org/10.1016/j.scitotenv.2021.149040>.
- Maulud, D., Abdulazeez, A.M., 2020. A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends* 1 (4), 140–147. <https://doi.org/10.38094/jastt1457>.
- Nelson, N.G., Munoz-Carpena, R., Philips, E.J., Kaplan, D., Sucusy, P., Hendrickson, J., 2018. Revealing biotic and abiotic controls of harmful algal blooms in a shallow subtropical lake through statistical machine learning. *Environ. Sci. Technol.* 52 (6), 3527–3535. <https://doi.org/10.1021/acs.est.7b05884>.
- Nie, P., Roccotelli, M., Fanti, M.P., Ming, Z., Li, Z., 2021. Prediction of home energy consumption based on gradient boosting regression tree. *Energy Rep.* 7, 1246–1255. <https://doi.org/10.1016/j.egy.2021.02.006>.
- Nong, X., Shao, D., Zhong, H., Liang, J., 2020. Evaluation of water quality in the South-to-North Water Diversion Project of China using the water quality index (WQI) method. *Water Res.* 178, 115781. <https://doi.org/10.1016/j.watres.2020.115781>.
- 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. <https://doi.org/10.1016/j.scitotenv.2014.09.005>.
- Qian, L., Plant, C., Böhm, C., 2021. Density-based clustering for adaptive density variation. In: 2021 IEEE International Conference on Data Mining (ICDM), IEEE, pp. 1282–1287.
- Qian, J., Liu, H., Qian, L., Bauer, J., Xue, X., Yu, G., He, Q., Zhou, Q., Bi, Y., Norra, S., 2022. Water quality monitoring and assessment based on cruise monitoring, remote sensing, and deep learning: a case study of Qingcaosha Reservoir. *Front. Environ. Sci.* 10 (October), 1–13. <https://doi.org/10.3389/fenvs.2022.979133>.
- Ralston, D.K., Moore, S.K., 2020. Modeling harmful algal blooms in a changing climate. *Harmful Algae* 91 (November), 101729. <https://doi.org/10.1016/j.hal.2019.101729>.
- Redfield, A.C., 1963. The influence of organisms on the composition of seawater. *Sea* 2, 26–77.
- Ridgeway, G., Madigan, D., Richardson, T., O’Kane, J., 1998. Interpretable boosted naive bayes classification. In: *The 4th International Conference on Knowledge Discovery and Data Mining (KDD-1998)*, pp. 101–104. URL citeseer.ist.psu.edu/ridgeway98interpretable.html.
- Saltelli, A., 2002. Sensitivity analysis for importance assessment. *Risk Anal.* 22, 579–590. <https://doi.org/10.1111/0272-4332.00040>.
- Su, Y., Hu, M., Wang, Y., Zhang, H., He, C., Wang, Y., Wang, D., Wu, X., Zhuang, Y., Hong, S., Trolle, D., 2022. Identifying key drivers of harmful algal blooms in a tributary of the Three Gorges Reservoir between different sea-seasons: causality based on data-driven methods. *Environ. Pollut.* 297 (August 2021), 118759. <https://doi.org/10.1016/j.envpol.2021.118759>.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: *Advances in Neural Information Processing Systems*, pp. 5998–6008.
- Wang, Y., Li, Y., Liang, J., Bi, Y., Wang, S., Shang, Y., 2021. Climatic changes and anthropogenic activities driving the increase in nitrogen: evidence from the south-to-north water diversion project. *Water (Switzerland)* 13 (18). <https://doi.org/10.3390/w13182517>.
- Welchowski, T., Maloney, K.O., Mitchell, R., Schmid, M., 2022. Techniques to improve ecological interpretability of black-box machine learning models: case study on biological health of streams in the United States with gradient boosted trees. *J. Agric. Biol. Environ. Stat.* 27 (1), 175–197. <https://doi.org/10.1007/s13253-021-00479-7>.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al., 2020. Transformers: state-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45.
- Wright, R., 2018. Interpreting Black-Box Machine Learning Models Using Partial Dependence and Individual Conditional Expectation Plots, Exploring SAS® Enterprise Miner Special Collection, pp. 1950–2018. URL <https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/1950-2018.pdf>.
- Xia, R., Zhang, Y., Wang, G., Zhang, Y., Dou, M., Hou, X., Qiao, Y., Wang, Q., Yang, Z., 2019. Multi-factor identification and modelling analyses for managing large river algal blooms. *Environ. Pollut.* 254, 113056. <https://doi.org/10.1016/j.envpol.2019.113056>.
- Yan, H., Huang, Y., Wang, G., Zhang, X., Shang, M., Feng, L., Dong, J., Shan, K., Wu, D., Zhou, B., Yuan, Y., 2016. Water eutrophication evaluation based on rough set and petri nets: a case study in Xiangxi-River, Three Gorges Reservoir. *Ecol. Indic.* 69, 463–472. <https://doi.org/10.1016/j.ecolind.2016.05.010>.
- Yang, Z., Zhang, M., Shi, X., Kong, F., Ma, R., Yu, Y., 2016. Nutrient reduction magnifies the impact of extreme weather on cyanobacterial bloom formation in large shallow Lake Taihu (China). *Water Res.* 103, 302–310. <https://doi.org/10.1016/j.watres.2016.07.047>.
- Yu, P., Gao, R., Zhang, D., Liu, Z.P., 2021. Predicting coastal algal blooms with environmental factors by machine learning methods. *Ecol. Indic.* 123, 107334. <https://doi.org/10.1016/j.ecolind.2020.107334>.
- Zhu, Y., Mi, W., Tu, X., Song, G., Bi, Y., 2022. Environmental factors drive periphytic algal community assembly in the largest long-distance water diversion channel. *Water* 14 (6). <https://www.mdpi.com/2073-4441/14/6/914>.
- Zhu, J., Lei, X., Quan, J., Yue, X., 2019. Algae growth distribution and key prevention and control positions for the middle route of the south-to-northwater diversion project. *Water (Switzerland)* 11 (9), 1–18. <https://doi.org/10.3390/w11091851>.