Front. Environ. Sci. Eng. 2023, 17(6): 76
https://doi.org/10.1007/s11783-023-1676-2

RESEARCH ARTICLE

Approaching the upper boundary of driver-response relationships: identifying factors using a novel framework integrating quantile regression with interpretable machine learning

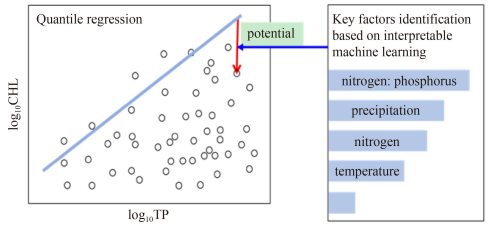
Zhongyao Liang[1,2,3], Yaoyang Xu[4], Gang Zhao[5], Wentao Lu[6,7], Zhenghui Fu (✉)[1], Shuhang Wang (✉)[1], Tyler Wagner[8*]

1 National Engineering Laboratory for Lake Pollution Control and Ecological Restoration, State Environment Protection Key Laboratory for Lake Pollution Control, Chinese Research Academy of Environmental Sciences, Beijing 100012, China
2 Fujian Provincial Key Laboratory for Coastal Ecology and Environmental Studies, Xiamen University, Xiamen 361102, China
3 College of the Environment & Ecology, Xiamen University, Xiamen 361102, China
4 Key Laboratory of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China
5 Department of Global Ecology, Carnegie Institution for Science, Stanford, CA 94305, USA
6 Institute of Strategic Planning, Chinese Academy of Environmental Planning, Beijing 100012, China
7 The Center for Beautiful China, Chinese Academy of Environmental Planning, Beijing 100012, China
8 U.S. Geological Survey, Pennsylvania Cooperative Fish and Wildlife Research Unit, Pennsylvania State University, University Park, PA 16802, USA

HIGHLIGHTS

- A novel framework integrating quantile regression with machine learning is proposed.
- It aims to identify factors driving observations to upper boundary of relationship.
- Increasing N:P and TN concentration help fulfill the effect of TP on CHL.
- Wetter and warmer decrease potential and increase eutrophication control difficulty.
- The framework advances applications of quantile regression and machine learning.

GRAPHIC ABSTRACT



ARTICLE INFO

Article history:
Received 5 June 2022
Revised 21 October 2022
Accepted 26 November 2022
Available online 11 January 2023

ABSTRACT

The identification of factors that may be forcing ecological observations to approach the upper boundary provides insight into potential mechanisms affecting driver-response relationships, and can help inform ecosystem management, but has rarely been explored. In this study, we propose a novel framework integrating quantile regression with interpretable machine learning. In the first stage of the framework, we estimate the upper boundary of a driver-response relationship using quantile regression. Next, we calculate "potentials" of the response variable depending on the driver, which are defined as vertical distances from the estimated upper boundary of the relationship to observations in the driver-response variable scatter plot. Finally, we identify key factors impacting the potential using a machine learning model. We illustrate the necessary steps to implement the framework using the total phosphorus (TP)-Chlorophyll *a* (CHL) relationship in lakes across the continental US. We found that the nitrogen to phosphorus ratio (N:P), annual average precipitation, total nitrogen (TN), and summer average air temperature were key factors impacting the potential of CHL depending on TP. We further revealed important implications of our findings for lake eutrophication management. The

✉ Corresponding authors
E-mails: fzh@pku.edu.cn (Z. Fu); wangsh@craes.org.cn (S. Wang)

*Disclaimer: This draft manuscript is distributed solely for purposes of scientific peer review. Its content is deliberative and predecisional, so it must not be disclosed or released by reviewers. Because the manuscript has not yet been approved for publication by the US Geological Survey (USGS), it does not represent any official finding or policy.

Special Issue—Artificial Intelligence/Machine Learning on Environmental Science & Engineering (Responsible Editors: Yongsheng Chen, Xiaonan Wang, Joe F. Bozeman III & Shouliang Yi)

Keywords:
Driver-response
Upper boundary of relationship
Interpretable machine learning
Quantile regression
Total phosphorus
Chlorophyll *a*

important role of N:P and TN on the potential highlights the co-limitation of phosphorus and nitrogen and indicates the need for dual nutrient criteria. Future wetter and/or warmer climate scenarios can decrease the potential which may reduce the efficacy of lake eutrophication management. The novel framework advances the application of quantile regression to identify factors driving observations to approach the upper boundary of driver-response relationships.

1 Introduction

Relationships between a predictor variable and an ecosystem response variable are widely used to illustrate the quantitative association between a driver and an ecological property of interest (Dillon and Rigler, 1974; Huo et al., 2013; Larned and Schallenberg, 2018; de Vries et al., 2021). Such driver-response relationships are often used to inform ecosystem management (Hunsicker et al., 2015; McDowell et al., 2018; Schallenberg, 2020). Quantile regression, which explores the effect of a driver on any interested quantile(s) of the response variable distribution (Koenker and Bassett 1978; Das et al., 2019), has been introduced as a useful alternative to models that focus on the mean of the response variable distribution to develop the driver-response relationship in environmental and ecological studies (Cade et al., 1999). Particularly, quantile regression can estimate the limiting effect of a driver on the response variable by fitting the upper boundary of the relationship (Cade et al., 1999; Xu et al., 2015). Here, the limiting effect reflects behaviors of the response variable when the driver is the only limiting factor, and under such conditions the driver-response relationship is recognized as the upper boundary of the relationship (Cade and Noon, 2003). For example, Sankaran et al. (2005) showed changes in maximum woody cover of African savannas with mean annual precipitation using a 99th quantile regression. Carvalho et al., (2013) applied quantile regression to characterize the relationship between the maximum biovolume of cyanobacteria and nutrient concentrations in lakes.

Exploring factors affecting or mediating driver-response relationships is an interest of ecologists (Freeman et al., 2009; Wagner et al., 2011). The identification of factors driving observations to approach the upper boundary of the relationship can help understand mechanisms and processes governing the driver's effect on the response variable (Liang et al., 2021a). A better understanding of such mechanisms can help inform management strategies that could inhibit or promote the effect of an ecological driver (Zou et al., 2020). However, identifying such factors has rarely been discussed. In particular, there is not an analytical framework available that outlines the necessary steps to identify key factors that may drive observations to approach the upper boundary of a driver-response relationship.

In this study, we propose a novel framework for identifying key factors driving observations to approach the upper boundary of a driver-response relationship. We achieve this framework by integrating quantile regression with interpretable machine learning (Murdoch et al., 2019; Rudin 2019). Interpretable machine learning has been introduced into environmental and ecological studies and has been shown to be an attractive approach for providing transparent and understandable associations (Lucas 2020; Ryo et al., 2020; Cha et al., 2021; Wang et al., 2021). In the framework, we firstly estimate the upper boundary relationship using quantile regression at an upper regression quantile (Sankaran et al., 2005; Carvalho et al., 2013; Fornaroli et al., 2018). Next, we define the potential of the response variable, depending on the drive, as the vertical distance from the upper boundary of the relationship to a specific observation in the driver-response variable scatter plot. This "potential" is calculated using the predicted value from the fitted quantile regression model minus the observed value of the response variable. Potentials of the response variable are calculated for all observations. Finally, we explore key factors impacting potentials using a machine learning model. As such, the factors identified as being important in predicting potentials are factors that are potentially important in driving observations to approach the upper boundary of the driver-response relationship.

To demonstrate the steps of the proposed framework, we use the widely studied total phosphorus (TP)-Chlorophyll *a* (CHL) driver-response relationship in inland lakes (Dillon and Rigler, 1974; Jones et al., 1998; Havens and Nürnberg, 2004; Filstrup and Downing, 2017) as a case study. The TP-CHL relationship provides quantitative information that can help guide lake eutrophication management (Stow and Cha, 2013; Rowland et al., 2019) and receives continuous attention in the scientific literature (Yuan and Jones 2020; Quinlan et al., 2020). Although the limiting effect of TP on CHL has been explored in some studies (Jones et al., 2011; Chen and Li 2014; Xu et al., 2015), factors driving observations to approach the upper boundary of the TP-CHL relationship have not been investigated. We applied the proposed novel framework to further identify such factors.

2 Materials and methods

2.1 Study area

Lake water quality data were obtained from US Environmental Protection Agency's National Lakes Assessment sampled in 2007, 2012, and 2017. Water quality

indicators included TP, CHL, total nitrogen (TN), nitrogen to phosphorus ratio (N:P, calculated by TN/TP), ammonia nitrogen (NH_3), nitrate nitrogen (NO_3), water temperature (WT), dissolved organic carbon (DOC), acid neutralizing capacity (ANC), conductivity (COND), pH, lake surface area (AREA_HA), lake depth (Depth), and dissolved oxygen (DO). The Environmental Protection Agency's ecoregion (EPA_REG) was also included.

Meteorological and land use/land cover variables may also impact the growth of phytoplankton and influence lake eutrophication (Collins et al., 2019; Cheruvelil et al., 2022). Meteorological indicators were derived from the database of Monthly Climate and Climatic Water Balance for Global Terrestrial Surfaces (Abatzoglou et al., 2018). We considered six climate indicators, including summer average air temperature of the sampling year (TEMP), annual average air temperature of the past 30 years (TEMP_30), monthly average precipitation of the sampling summer (PRCP), monthly average precipitation of the past 30 years (PRCP_30), summer average wind speed of the sampling year (WS), and annual average wind speed of the past 30 years (WS_30). Land use data were derived from the US Geological Survey National Land Cover Database (Dewitz and U.S. Geological Survey, 2021). For the land use data, we calculated two indicators: the ratio of areas with a mixture of constructed materials and vegetation (Developed) and the ratio of areas dominated by trees generally greater than 5 m tall and greater than 20% of total vegetation cover (Forest). We provide details of data preparation and processing in the supplementary materials.

Observations of TP and CHL were used to develop the quantile regression model. The remaining variables (i.e., predictor variables) were used as factors impacting the potential of CHL depending on TP. We explored drivers of the potential at multiple spatial scales. For example, EPA_REG, PRCP, PRCP_30, TEMP, TEMP_30, WS, and WS_30 are regional scale drivers, while the other variables were derived at the lake scale. We restricted TP and CHL data to those that were collected during the summer period (from 15 June to 14 September). We removed observations with any missing values among the variables. We also removed 13 observations with extreme WT (i.e., WT < 5 °C or > 40 °C) or TP (i.e., TP < 1 μg/L or > 1500 μg/L) levels. The final sample size was 3230 observations included in the quantile regression and machine learning models. TP, CHL, TN, N:P, AREA_HA, Depth, COND, and DOC were log10 transformed prior to analyses.

2.2 Methodological framework

The analytical framework includes three steps (Fig. 1). It is worth noting that the framework is not constrained to a specific quantile regression or machine learning model. Despite displaying a linear quantile regression model in Fig. 1, nonlinear models (Koenker et al., 1994; Koenker and Park 1996) could also be used.

2.2.1 Estimating the upper boundary of a driver-response relationship using quantile regression

The first step of the framework is estimating the upper

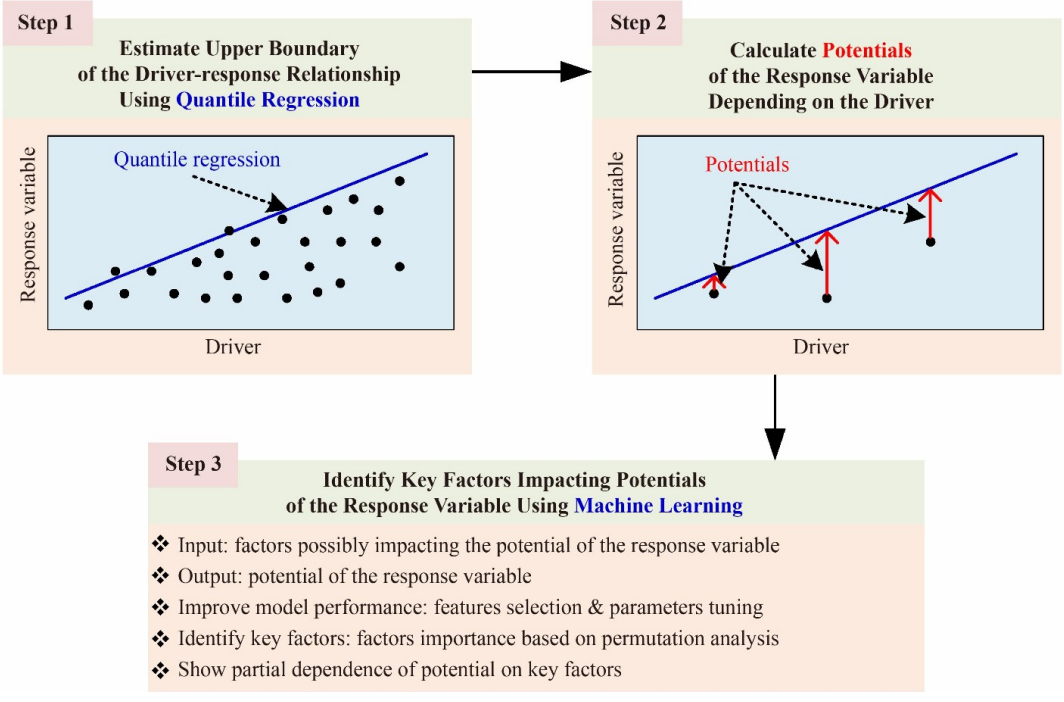


Fig. 1 Flowchart of the proposed methodology framework.

boundary of the driver-response relationship. Theoretically, when the response variable is only limited by the driver, the upper boundary of relationship can be directly obtained by illustrating the behavior of the response variable with the change of driver. However, this is not observed in nature because observations are always below the upper boundary relationship when other factors limit the response variable (Cade and Noon 2003; Zou et al., 2020).

Quantile regression at an upper regression quantile is a proper tool to estimate the upper boundary of a driver-response relationship (Cade et al., 1999; Fornaroli et al., 2018). Since the regression quantile is less than one, the estimated relationship should be below the true upper boundary of the relationship. Generally, the use of a very large regression quantile (e.g., 99th percentile) will lead to a closer estimate to the true upper boundary of relationship, but it can be highly uncertain due to the small sample size that typically exists when estimating parameters at such a large quantile. Therefore, the selection of regression quantile is a trade-off between seeking a higher quantile to better represent the upper boundary of relationship and reducing the prediction uncertainty (Konrad et al., 2008). To distinguish the upper boundary relationship estimated using quantile regression with the true upper boundary, we refer to the former as the estimated upper boundary of the driver-response relationship.

For the TP–CHL relationship, following previous practices (Chen and Li 2014; Xu et al., 2015), both TP and CHL were $\log_{10}$ transformed prior to applying linear quantile regression to fit the upper boundary of the TP-CHL relationship. The main function of the linear quantile regression can be expressed by Eq. (1) (Koenker and Bassett, 1978):

$$y_i = \alpha x_i + \beta + \epsilon_i \tag{1}$$

where i is the index of observations (i = 1, 2, ..., N. N is the sample size). x and y represent $\log_{10}$ transformed TP and CHL (units: μg/L). α and β represent the regression slope and intercept. ϵ is the residual. Parameters estimation is based on the minimum of sum of the absolute residuals (Koenker and Bassett, 1978):

$$\min\left[\sum_{i\in\{i:y_i\geqslant\alpha x_i+\beta\}}\tau\,|y_i-\alpha x_i-\beta| + \sum_{i\in\{i:y_i<\alpha x_i+\beta\}}(1-\tau)\,|y_i-\alpha x_i-\beta|\right] \tag{2}$$

where τ ($0 < \tau < 1$) represents the regression quantile. Because the prediction uncertainty can be large for extreme regression quantiles (Das et al., 2019), we used the 0.95 regression quantile for estimating the upper boundary relationship. That is, τ equals to 0.95 in Eq. (2).

2.2.2 Calculating potentials of the response variable

The second step of the framework is calculating potentials of the response variable depending on the driver. As mentioned before, the definition of potential is the distance from the estimated upper boundary to a given observation in the driver-response variable scatter plot (Fig. 1). Supposing that the response variable is only limited by the driver under consideration, all the observations should be on the upper boundary. Because there more than one factor that limit ecological response variables (Cade and Noon, 2003), a single driver cannot fulfill its maximum effect on the response variable. Thus, the newly defined concept, the "potential", can be used to represent the joint effect of other limiting factors. Conversely, by mining the relationship between the potential and possible factors, we can reveal key factors impacting the potential and driving observations to approach the upper boundary of the relationship (See Section 2.2.3). Note that since some observations may be above the estimated upper boundary of relationship, their corresponding calculated potentials can be negative.

We emphasize two features of the newly defined concept, potential. 1) It is the difference of the response variable between the ideal (i.e., the stressor is the only limiting factor) and real (other factors may also limit the response variable) conditions. 2) It is driver dependent, because the calculation of potential relies on the driver-response relationship. Accordingly, in the TP-CHL relationship, potentials of CHL depending on TP were calculated by subtracting observed values of $\log_{10}$CHL from predicted $\log_{10}$CHL at the 0.95 regression quantile.

2.2.3 Identifying factors impacting the potential using machine learning

The third step of the framework is applying a machine learning approach to identify key factors impacting the potential of the response variable depending on the driver (Fig. 1). Machine learning models have been widely used in environmental and ecological studies (Sun and Scanlon, 2019; Castrillo and García 2020; Lucas, 2020; Tiyasha et al., 2020). One appealing aspect of many of these models is their ability to handle nonlinear and complicated relationships (Liang et al., 2020, Rousso et al., 2020). Moreover, the convenience in ranking the importance of input (predictor) variables (Chen et al., 2020; Dugan et al., 2020) makes machine learning particularly suitable in our framework.

In the framework, inputs (predictor variables) to the machine learning model were variables considered to possibly impact the potential, while the output (response) variable was the calculated potential values derived from the quantile regression. We conducted feature selection and hyperparameter tuning to obtain an optimized machine learning model (Niu et al., 2021). Feature selection is a process to select a subset of relevant input variables for model development, by which effects from noise or irrelevant variables are reduced (Chandrashekar

and Sahin 2014; Li et al., 2018). After the feature selection, we tuned several hyperparameters to seek an optimum set for the machine learning model (Araya and Ghezzehei 2019; Yang and Shami 2020). By using feature selection and hyperparameter tuning, the optimized model is expected to perform better with fewer input variables. Next, we ranked the importance of variables based on a permutation analysis (Altmann et al., 2010) for the optimized model and determined key factors impacting the potential. Finally, we explore the change of the response variable (the potential) as a function of each predictor while holding other predictors constant. We used the partial dependence profile (Biecek and Burzykowski, 2021) to show changes of potential predictions for a given predictor variable. The partial dependence profile is the average of ceteris paribus profiles showing how model predictions would change if the value of a single predictor variable changed (Becker et al., 2021).

For our case study, we used random forests (Breiman, 2001) to explore the relationship between the potential and possible predictor variables. There were 21 input variables before the feature selection, namely WT, pH, DO, DOC, ANC, COND, TN, N:P, NO_3, NH_3, AREA_HA, Depth, Developed, Forest, EPA_REG, PRCP, PRCP_30, TEMP, TEMP_30, WS, and WS_30. The output was the potential of CHL. We used 5-fold cross-validation to reduce the impact of overfitting of the random forest model (Yadav and Shukla, 2016). We used the average R^2 (RSQ) value of testing data sets as the measure of model performance. For the feature selection, we applied the sequential backward search algorithm (Yusta, 2009). For the hyperparameter tuning, we applied the hyperband algorithm (Li et al., 2017). According to prior knowledge on the importance of hyperparameters to the performance of random forests models (Probst et al., 2019), we selected four hyperparameters to tune (Table 1). Both feature selection and hyperparameter tuning aimed to maximize RSQ. We used the root mean square error (RMSE) loss after permutation to represent variables importance.

We used the R software (version 4.1.0, R Core Team, 2021) for all the computations. We developed the linear quantile regression model using the quantreg (version 5.85, Koenker, 2021) package. Random forests were fitted using the ranger (version 0.12.1, Wright and Ziegler, 2017) and mlr3verse (version 0.2.1, Lang and Schratz, 2021) packages.

Table 1 Search space for the four tuned hyperparameters in the random forests model

Hyperparameters	Abbreviation	Type	Range
Number of randomly drawn candidate variables	Mtry	Integer	1–6
Minimum number of observations in a terminal node	Min.node.size	Integer	1–5
Number of trees	Num.trees	Integer	200–1000
Sampling size controlled by sampling fraction	Sample.fraction	Double	0.75–1

3 Results

3.1 Upper boundary of TP-CHL relationship

Estimated average regression slope and intercept for the quantile regression (Eq. (1)) were 0.885 and 0.164, respectively. Because both TP and CHL were $\log_{10}$ transformed, the regression slope can be explained as the percent change in CHL concentration per 1% change in TP (Qian, 2009), that is, the 95th quantile of CHL distribution increases by 0.885% per 1% increases of TP. The regression intercept is the $\log_{10}$CHL value when $\log_{10}$TP is zero (where the TP concentration is 1 μg/L).

We used the quantile regression results at the 95th regression quantile (Fig. 2) to estimate the upper boundary of TP–CHL relationship. According to the parameter estimation algorithm (Eq. (2)), there are approximately 5% (the exact number is 4.96% in our case) of observations above or on the fitted curve (the black line in Fig. 2). Note that the 95% credible intervals for the predicted values are very small, indicating high reliability of potentials calculated in the second step of the framework.

3.2 Potentials of CHL depending on TP

Summary statistics of the calculated potentials are shown in Table 2 and Fig. S1. The average value was 0.59, which is higher than the median (0.52), indicating the distribution of the calculated potential is right skewed. The standard deviation was 0.45. It is not surprising that the minimum value was negative (−0.91), because there were some observations above the estimated upper boundary of TP-CHL relationship (Fig. 2).

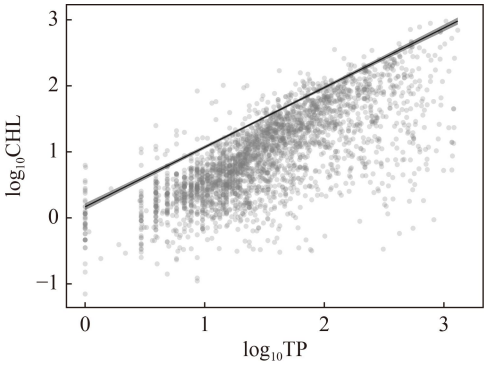


Fig. 2 Quantile regression results (τ = 0.95) for the TP-CHL relationship, representing the estimated upper boundary of TP-CHL relationship. Points are observations. The black line and gray shaded region represent the fitted line and 95% credible intervals, respectively.

Table 2 Basic statistics of calculated potentials of CHL depending on TP

Mean	Standard deviation	Minimum	Maximum	Quantiles		
				25%	50%	75%
0.59	0.45	−0.91	2.98	0.28	0.52	0.82

3.3 Key factors impacting the potential

The RSQ of the random forests model before feature selection was 0.498. After feature selection, 12 input variables remained, namely N:P, PRCP_30, TN, TEMP_30, DOC, NH_3, pH, NO_3, ANC, AREA_HA, COND, EPA_REG. Tuned hyperparameters values were 6, 2, 595, and 0.9804 for mtry, min.node.size, num.trees, and sample.fraction, respectively. The RSQ of the optimized random forests model increased to 0.522 (refer to Fig. S2 for the fitted plot). The optimized model had slightly better performance compared to the pre-tuned model. We also conducted a multivariate linear regression and found the RSQ was only 0.227, which was much less than that of the random forests model.

Results of variables importance are shown in Fig. 3. The variable N:P was the most important variable, followed by PRCP_30 and TN. Their RMSE losses were 0.335, 0.292, and 0.274, respectively, which were much higher compared with the rest of the predictor variables. TEMP_30 ranked fourth, with an RMSE loss of 0.240. RMSE losses for the other eight variables were relatively small ($\leqslant$ 0.20) compared with the aforementioned four factors.

Marginal effects of N:P, PRCP_30, TN, and TEMP_30 on the predicted potential of CHL are shown in Fig. 4. Generally, the predicted potential decreased with increasing of N:P, PRCP_30, TN, and TEMP_30, but at different rates. The predicted potential decreased the fastest with the increasing N:P. With increasing PRCP_30 and TN, the predicted potential decreased slower than that for N:P. The decreasing rate of predicted potential also appears smaller for TEMP_30 than those for the aforementioned three factors. As for the variation of the remaining factors, the corresponding predicted potentials were relatively constant (Fig. S3). Therefore, according to the rank of variable importance (Fig. 3) and marginal effects of factors (Fig. 4), we determined that key factors impacting the potential of CHL depending on TP were N:P, PRCP_30, TN, and TEMP_30.

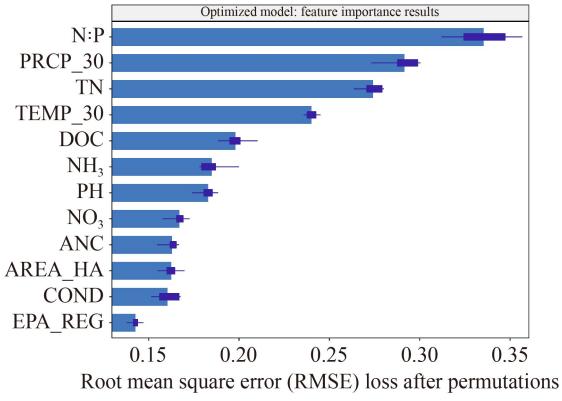


Fig. 3 Variables importance measured by the root mean square error loss from a random forest model based on permutation analysis. Bars charts and box plots show averages and distributions of root mean square error losses across the iterations of the algorithm.

An increase of each of these four factors reduces the potential of CHL depending on TP. Based on the definition of potential, the decrease of potential means that the observation approaches the upper boundary of TP-CHL relationship and thus leads to a higher CHL concentration with a given TP concentration value. Therefore, the increase in N:P, PRCP_30, TN, or TEMP_30 is likely to increase the CHL concentration.

4 Discussion

4.1 Implications for lake eutrophication management

The identification of factors that may be forcing ecological observations to approach the upper boundary of a driver-response relationship may help better understand system dynamics and inform ecosystem management (Zou et al., 2020; Liang et al., 2021a). Specifically, for the lake TP-CHL relationship, identifying key factors impacting the potential of CHL depending on TP has the following implications for lake eutrophication management. First, the results show factors governing the effect of TP on CHL, which is helpful to deepen our understanding of the TP-CHL relationship in lakes. N:P was identified as the most important factor effecting the potential, emphasizing the critical role of N:P on mediating the effect of TP on CHL. In addition, TN was also identified as an important factor impacting the potential of CHL, which is consistent with previous studies that have identified TN as an essential limiting nutrient for phytoplankton growth (Conley et al., 2009; Paerl et al., 2019). Because N:P is calculated using the $\log_{10}$ transformed TN to TP ratio, N:P is closely related to, but not identical to, TN. For example, if TP concentration doubles, to keep the same effect of N:P on the potential, TN concentration should also double. Otherwise, if the TN concentration remained unchanged, N:P would become smaller and the potential would become larger. Importantly, N:P is a commonly accepted indicator for nutrient limitation (Redfield 1958; Elser et al., 2007; Moon et al., 2021). A larger N:P indicates a greater possibility of TP limitation (Guildford and Hecky 2000; Liang et al., 2018), which strongly supports our finding in this study that a larger N:P leads to a smaller potential of CHL depending on TP.

Monthly average precipitation of the past 30 years

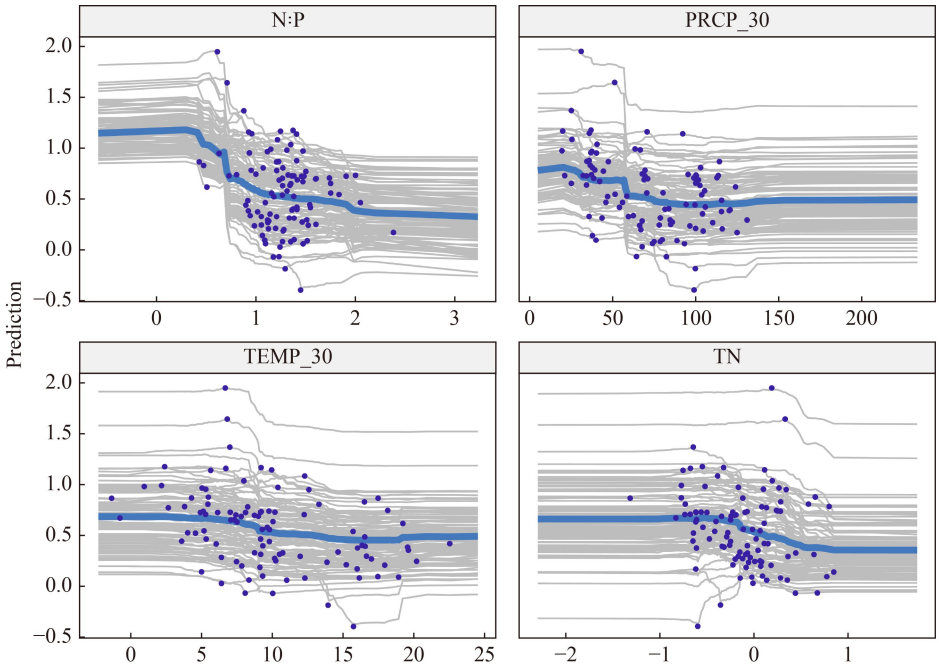


Fig. 4 Partial dependence profiles (thick blue lines) showing changes of potential predictions with N:P, PRCP_30, TEMP_30, and TN. For each factor, narrow gray lines are ceteris paribus profiles given a set of observations and the corresponding partial dependence profile is the average of ceteris paribus profiles. Ceteris paribus profiles show how a model's prediction would change if the value of a single exploratory variable changed (Biecek and Burzykowski 2021). Dots are 100 randomly sampled observations for the profiles calculation. N:P and TN are $\log_{10}$ transformed. Units for PRCP_30, TEMP_30, and TN are mm, °C, and μg/L, respectively.

(PRCP_30) may not directly affect phytoplankton growth; however, it can indirectly impact lake ecosystems through direct effects on watershed runoff (Nyenje et al., 2010; Stockwell et al., 2020). Increasing watershed runoff can lead to an increase in bioavailable nutrients, such as dissolved inorganic nitrogen or soluble reactive phosphorus, in the waterbody (Motew et al., 2018; Woolway et al., 2020) and promote the phytoplankton growth. Therefore, lakes in wetter climates (a larger PRCP_30) are likely to experience smaller potentials of CHL. Note that PRCP_30, rather than PRCP, was identified as a key factor, indicating that PRCP_30 was a more robust indicator of a lake's climate than PRCP. A possible reason is that summer nutrients concentrations in lakes can be affected by load input in previous months or years (Obenour et al., 2014; Collins et al., 2019) due to the internal processing of nutrients (Søndergaard et al., 2003; Tong et al., 2021).

Monthly average temperature of the past 30 years (TEMP_30) and WT were positively correlated with one another and both are important for phytoplankton growth (Paerl and Paul, 2012). TEMP_30 is also related to other factors (such as sunshine duration and air pressure) impacting phytoplankton growth (Zhang et al., 2018), which may be the reason that TEMP_30, rather than WT, was identified as a key factor. Compared with TEMP_30, TEMP was not identified as a key factor, reflecting the long-term effects of temperature on the phytoplankton growth. The importance rank of TEMP_30 falls behind those of N:P and PRCP_30, indicating a less important role of summer temperature in limiting the phytoplankton growth when compared to N concentrations.

Second, our framework and findings are useful for informing management actions aimed at curbing lake eutrophication. The results highlight critical roles of N:P and TN for lake eutrophication management from the new perspective of CHL potential depending on TP. Traditionally, N:P is used to help indicate a shift in nutrient limitation of phytoplankton (Guildford and Hecky 2000; Liang et al., 2018). Our results broaden the impact of N:P on phytoplankton growth by revealing the effect of N:P on the potential of CHL, and highlights that a decrease in N:P can help the reduction of CHL concentration via the decrease of CHL potential depending on TP. The reduction of TN can directly and indirectly (by the reduction of N:P) lead to a higher potential of CHL and be conducive to the reduction of CHL concentration, which possibly provides two additional explanations on how TN impacts CHL in lakes. Besides, lower N:P can lead to the dominance of nitrogen-fixing cyanobacteria (Havens et al., 2003), which makes lake eutrophication management more difficult in many cases. Therefore, the simultaneous control of N and P concentrations is likely necessary for effective lake eutrophication management (Elser et al., 2007, Paerl et al., 2016; Liang et al., 2021b), while solely reducing TP concentrations may lead to an

increase in N:P which can partially offset the direct effect of TP concentration reduction on CHL concentrations.

Our findings also reveal that lakes at wetter and warmer climates are likely to have lower potentials and thus higher CHL concentrations given the same TP concentration. Future wetter and/or warmer climate scenarios (Sinha et al., 2017; Kalcic et al., 2019) might decrease the potential of CHL depending on TP and reduce the effect of TP concentration reduction on CHL. Such a scenario could result in difficulties implementing traditional lake eutrophication management strategies.

Third, it is worth noting that N:P and TN are lake-specific variables, while PRCP_30 and TEMP are regional factors. Our findings show that factors at both site-specific and regional scales can impact the potential of CHL, emphasizing the need to consider factors at multiple scales when identifying forcing variables of driver-response relationships for informing lake eutrophication management (Soranno et al., 2014).

Last, our findings help inform future studies on lake eutrophication management. We recognize that the RSQ of the optimized model is not extremely high, indicating important factors may have been excluded as input variables for the random forests model. For example, carbon is also treated as an essential nutrient for phytoplankton growth (Kragh and Sand-Jensen, 2018). However, DOC does not reflect bioavailable carbon sources well (Mette 1997; Hammer et al., 2019; Zagarese et al., 2021) and the inclusion of inorganic carbon could improve model performance. Light has also been identified as a limiting factor of phytoplankton growth (Loiselle et al., 2007; Chen et al., 2015). Unfortunately, we did not find a reliable indicator of light limitation. Although we did obtain transparency and turbidity data for the study lakes, exploratory analysis showed a negative correlation between CHL and transparency and turbidity, indicating that the variation in CHL was more likely to be the cause, instead of the effect, of changes in transparency and turbidity. To further improve model performance and better inform lake eutrophication management, future efforts could incorporate additional factors impacting the light environment and the potential of CHL.

4.2 Advancing the application of quantile regression

The novelty of this study lies in the proposal of the analytical framework, by which we make a step forward in the application of quantile regression. Conventionally, quantile regression has been widely used to estimate the upper boundary of driver-response relationships (Cade et al., 1999; Xu et al., 2015), as we did in the first step of the framework. To obtain a deeper understanding of the driver-response relationship, it is also important to identify factors driving observations to approach the upper boundary of the relationship. As we demonstrated using the TP-CHL relationship for inland lakes, the proposed framework was indeed capable of revealing such factors, which can help inform ecosystem management. As such, the proposed novel framework is expected to broaden the application of quantile regression in environmental and ecological studies.

Here, we also emphasize the merit of interpretable machine learning. Because there are few studies exploring factors impacting the potential of the response variable, we have limited information on which variables were important for effecting CHL potential and what the relationship between the potential and possible drivers was. Under such conditions, machine learning methods are advantageous because of their ability in handling complicated nonlinear relationships (Liang et al., 2020; Rousso et al., 2020) and ranking variable importance (Chen et al., 2020; Dugan et al., 2020).

4.3 Generalization of the proposed framework

In this study, we used the TP-CHL relationship of inland lakes as a case study to illustrate the necessary steps to implement the proposed framework. The novel framework is flexible and can accommodate more complicated driver-response relationships, such as a nonlinear relationship (Sankaran et al., 2005) or a relationship with a change point (Liang et al., 2021a). In addition, a hierarchical structure can be used for parameter estimation based on partial data pooling (Fornaroli et al., 2014). Moreover, multiple drivers can be used in quantile regression models (Zou et al., 2020). Also, the framework is applicable when the response variable is categorical (Benoit and den Poel, 2010), counts (Lee and Neocleous, 2010), or left-censored (Alhamzawi and Ali, 2020). The proposed framework can also be applicable when joint potentials - potentials of more than one driver - are of interest. In this case, a multivariate quantile regression model could be developed. As such, we expect that the proposed framework can be generalized to other environmental and ecological studies.

Acknowledgements This research was funded by the National Natural Science Foundation of China (Nos. 71761147001 and 42030707), the International Partnership Program by the Chinese Academy of Sciences (No. 121311KYSB20190029), the Fundamental Research Fund for the Central Universities (No. 20720210083), and the National Science Foundation (Nos. EF-1638679, EF-1638554, EF-1638539, and EF-1638550). Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the US Government.

Electronic Supplementary Material Supplementary material is available in the online version of this article at https://doi.org/10.1007/s11783-023-1676-2 and is accessible for authorized users.

Data Accessibility Statement The data supporting the findings of this study are available within the article and its supplementary materials. The code that support the findings of this study are available from the first author (Z. Liang, liangzhongyao@xmu.edu.cn), upon reasonable request.

References

Abatzoglou J T, Dobrowski S Z, Parks S A, Hegewisch K C (2018). TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. Scientific Data, 5(1): 170191

Alhamzawi R, Ali H T M (2020). Brq: an R package for Bayesian quantile regression. Metron, 78(3): 313–328

Altmann A, Toloşi L, Sander O, Lengauer T (2010). Permutation importance: a corrected feature importance measure. Bioinformatics (Oxford, England), 26(10): 1340–1347

Araya S N, Ghezzehei T A (2019). Using machine learning for prediction of saturated hydraulic conductivity and its sensitivity to soil structural perturbations. Water Resources Research, 55(7): 5715–5737

Becker M, Binder M, Bischl B, Lang M, Pfisterer F, Reich N G, Richter J, Schratz P, Sonabend R (2021). mlr3 book

Benoit D F, den Poel D V (2010). Binary quantile regression: a Bayesian approach based on the asymmetric Laplace distribution. Journal of Applied Econometrics, 27(7): 1174–1188

Biecek P, Burzykowski T (2021). Explanatory Model Analysis. New York: Chapman and Hall/CRC

Breiman L (2001). Random forests. Machine Learning, 45(1): 5–32

Cade B S, Noon B R (2003). A gentle introduction to quantile regression for ecologists. Frontiers in Ecology and the Environment, 1(8): 412–420

Cade B S, Terrell J W, Schroeder R L (1999). Estimating effects of limiting factors with regression quantiles. Ecology, 80(1): 311–323

Carvalho L, McDonald C, de Hoyos C, Mischke U, Phillips G, Borics G, Poikane S, Skjelbred B, Solheim A L, Wichelen J V, et al. (2013). Sustaining recreational quality of European lakes: minimizing the health risks from algal blooms through phosphorus control. Journal of Applied Ecology, 50(2): 315–323

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

Cha Y, Shin J, Go B, Lee D S, Kim Y, Kim T, Park Y S (2021). An interpretable machine learning method for supporting ecosystem management: application to species distribution models of freshwater macroinvertebrates. Journal of Environmental Management, 291: 112719

Chandrashekar G, Sahin F (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1): 16–28

Chen K, Chen H, Zhou C, Huang Y, Qi X, Shen R, Liu F, Zuo M, Zou X, Wang J, Zhang Y, Chen D, Chen X, Deng Y, Ren H (2020). Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water Research, 171: 115454

Chen M, Fan M, Liu R, Wang X, Yuan X, Zhu H (2015). The dynamics of temperature and light on the growth of phytoplankton. Journal of Theoretical Biology, 385: 8–19

Chen X, Li X (2014). Using quantile regression to analyze the stressor-response relationships between nutrient levels and algal biomass in three shallow lakes of the lake Taihu Basin, China. Chinese Science Bulletin, 59(28): 3621–3629

Cheruvelil K S, Webster K E, King K B S, Poisson A C, Wagner T (2022). Taking a macroscale perspective to improve understanding of shallow lake total phosphorus and chlorophyll *a*. Hydrobiologia, 849(17–18): 3663–3677

Collins S M, Yuan S, Tan P N, Oliver S K, Lapierre J F, Cheruvelil K S, Fergus C E, Skaff N K, Stachelek J, Wagner T, et al. (2019). Winter precipitation and summer temperature predict lake water quality at macroscales. Water Resources Research, 55(4): 2708–2721

Conley D J, Paerl H W, Howarth R W, Boesch D F, Seitzinger S P, Havens K E, Lancelot C, Likens G E (2009). Controlling eutrophication: nitrogen and phosphorus. Science, 323(5917): 1014–1015

Das K, Krzywinski M, Altman N (2019). Quantile regression. Nature Methods, 16(6): 451–452

de Vries J, Kraak M H, Skeffington R A, Wade A J, Verdonschot P F (2021). A Bayesian network to simulate macroinvertebrate responses to multiple stressors in lowland streams. Water Research, 194: 116952

Dewitz J, U.S. Geological Survey (2021). National land cover database (NLCD) 2019 products (Ver. 2.0, June 2021). Washington, DC: U.S. Geological Survey Data Release

Dillon P J, Rigler F H (1974). The phosphorus-chlorophyll relationship in lakes. Limnology and Oceanography, 19(5): 767–773

Dugan H A, Skaff N K, Doubek J P, Bartlett S L, Burke S M, Krivak-Tetley F E, Summers J C, Hanson P C, Weathers K C (2020). Lakes at risk of chloride contamination. Environmental Science & Technology, 54(11): 6639–6650

Elser J J, Bracken M E, Cleland E E, Gruner D S, Harpole W S, Hillebrand H, Ngai J T, Seabloom E W, Shurin J B, Smith J E (2007). Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. Ecology Letters, 10(12): 1135–1142

Filstrup C T, Downing J A (2017). Relationship of chlorophyll to phosphorus and nitrogen in nutrient-rich lakes. Inland Waters, 7(4): 385–400

Fornaroli R, Cabrini R, Sartori L, Marazzi F, Vracevic D, Mezzanotte V, Annala M, Canobbio S (2015). Predicting the constraint effect of environmental characteristics on macroinvertebrate density and diversity using quantile regression mixed model. Hydrobiologia, 742(1): 153–167

Fornaroli R, Ippolito A, Tolkkinen M J, Mykra H, Muotka T, Balistrieri L S, Schmidt T S (2018). Disentangling the effects of low pH and metal mixture toxicity on macroinvertebrate diversity. Environmental Pollution, 235: 889–898

Freeman A M, Lamon E C III, Stow C A (2009). Nutrient criteria for lakes, ponds, and reservoirs: a Bayesian TREED model approach. Ecological Modelling, 220(5): 630–639

Guildford S J, Hecky R E (2000). Total nitrogen, total phosphorus, and nutrient limitation in lakes and oceans: is there a common relationship? Limnology and Oceanography, 45(6): 1213–1223

Hammer K J, Kragh T, Sand-Jensen K (2019). Inorganic carbon promotes photosynthesis, growth, and maximum biomass of phytoplankton in eutrophic water bodies. Freshwater Biology, 64(11): 1956–1970

Havens K E, James R, East T L, Smith V H (2003). N:P ratios, light limitation, and cyanobacterial dominance in a subtropical lake impacted by non-point source nutrient pollution. Environmental Pollution, 122(3): 379–390

Havens K E, Nürnberg G K (2004). The phosphorus-chlorophyll relationship in lakes: potential influences of color and mixing regime. Lake and Reservoir Management, 20(3): 188–196

Hunsicker M E, Kappel C V, Selkoe K A, Halpern B S, Scarborough C, Mease L, Amrhein A (2015). Characterizing driver-response relationships in marine pelagic ecosystems for improved ocean management. Ecological Applications, 26(3): 651–663

Huo S, Xi B, Ma C, Liu H (2013). Stressor-response models: a practical application for the development of lake nutrient criteria in China. Environmental Science & Technology, 47(21): 11922–11923

Jones J R, Knowlton M F, Kaiser M S (1998). Effects of aggregation on chlorophyll-phosphorus relations in Missouri Reservoirs. Lake and Reservoir Management, 14(1): 1–9

Jones J R, Obrecht D V, Thorpe A P (2011). Chlorophyll maxima and chlorophyll: total phosphorus ratios in Missouri reservoirs. Lake and Reservoir Management, 27(4): 321–328

Kalcic M M, Muenich R L, Basile S, Steiner A L, Kirchhoff C, Scavia D (2019). Climate change and nutrient loading in the western Lake Erie basin: warming can counteract a wetter future. Environmental Science & Technology, 53(13): 7543–7550

Koenker R (2021). Quantreg: Quantile Regression. R Package Version 5.85

Koenker R, Bassett G (1978). Regression quantiles. Econometrica, 46(1): 33–50

Koenker R, Ng P, Portnoy S (1994). Quantile smoothing splines. Biometrika, 81(4): 673–680

Koenker R, Park B J (1996). An interior point algorithm for nonlinear quantile regression. Journal of Econometrics, 71(1–2): 265–283

Konrad C P, Brasher A M D, May J T (2008). Assessing streamflow characteristics as limiting factors on benthic invertebrate assemblages in streams across the western United States. Freshwater Biology, 53(10): 1983–1998

Kragh T, Sand-Jensen K (2018). Carbon limitation of lake productivity. Proceedings of the Royal Society B. Biological Sciences, 285(1891): 20181415

Lang M, Schratz P (2021). mlr3verse: Easily Install and Load the 'mlr3' Package Family. R Package Version 0.2.1

Larned S T, Schallenberg M (2019). Stressor-response relationships and the prospective management of aquatic ecosystems. New Zealand Journal of Marine and Freshwater Research, 53(4): 489–512

Lee D, Neocleous T (2010). Bayesian quantile regression for count data with application to environmental epidemiology. Applied Statistics, 59(5): 905–920

Li J, Cheng K, Wang S, Morstatter F, Trevino R P, Tang J, Liu H (2018). Feature selection. ACM Computing Surveys, 50(6): 1–45

Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A (2017). Hyperband: a novel bandit-based approach to hyperparameter optimization. Journal of Machine Learning Research, 18(1): 1–52

Liang Z, Liu Y, Xu Y, Wagner T (2021a). Bayesian change point quantile regression approach to enhance the understanding of shifting phytoplankton-dimethyl sulfide relationships in aquatic ecosystems. Water Research, 201: 117287

Liang Z, Soranno P A, Wagner T (2020). The role of phosphorus and nitrogen on chlorophyll *a*: evidence from hundreds of lakes. Water Research, 185: 116236

Liang Z, Wu S, Chen H, Yu Y, Liu Y (2018). A probabilistic method to enhance understanding of nutrient limitation dynamics of phytoplankton. Ecological Modelling, 368: 404–410

Liang Z, Xu Y, Qiu Q, Liu Y, Lu W, Wagner T (2021b). A framework to develop joint nutrient criteria for lake eutrophication management in eutrophic lakes. Journal of Hydrology (Amsterdam), 594: 125883

Loiselle S A, C'ozar A, Dattilo A, Bracchini L, G'alvez J A (2007). Light limitations to algal growth in tropical ecosystems. Freshwater Biology, 52(2): 305–312

Lucas T C D (2020). A translucent box: interpretable machine learning in ecology. Ecological Monographs, 90(4): e01422

McDowell R W, Schallenberg M, Larned S (2018). A strategy for optimizing catchment management actions to stressor-response relationships in freshwaters. Ecosphere, 9(10): e02482

Hein M (1997). Inorganic carbon limitation of photosynthesis in lake phytoplankton. Freshwater Biology, 37(3): 545–552

Moon D L, Scott J T, Johnson T R (2021). Stoichiometric imbalances complicate prediction of phytoplankton biomass in U.S. lakes: implications for nutrient criteria. Limnology and Oceanography, 66(8): 2967–2978

Motew M, Booth E G, Carpenter S R, Chen X, Kucharik C J (2018). The synergistic effect of manure supply and extreme precipitation on surface water quality. Environmental Research Letters, 13(4): 044016

Murdoch W J, Singh C, Kumbier K, Abbasi-Asl R, Yu B (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences of the United States of America, 116(44): 22071–22080

Niu W, Feng Z, Li S, Wu H, Wang J (2021). Short-term electricity load time series prediction by machine learning model via feature selection and parameter optimization using hybrid cooperation search algorithm. Environmental Research Letters, 16(5): 055032

Nyenje P, Foppen J, Uhlenbrook S, Kulabako R, Muwanga A (2010). Eutrophication and nutrient release in urban areas of sub-Saharan Africa: a review. Science of the Total Environment, 408(3): 447–455

Obenour D R, Gronewold A D, Stow C A, Scavia D (2014). Using a Bayesian hierarchical model to improve Lake Erie cyanobacteria bloom forecasts. Water Resources Research, 50(10): 7847–7860

Paerl H W, Havens K E, Xu H, Zhu G, McCarthy M J, Newell S E, Scott J T, Hall N S, Otten T G, Qin B (2020). Mitigating eutrophication and toxic cyanobacterial blooms in large lakes: the evolution of a dual nutrient (N and P) reduction paradigm. Hydrobiologia, 847(21): 4359–4375

Paerl H W, Paul V J (2012). Climate change: Links to global expansion of harmful cyanobacteria. Water Research, 46(5): 1349–1363

Paerl H W, Scott J T, McCarthy M J, Newell S E, Gardner W S, Havens K E, Hoffman D K, Wilhelm S W, Wurtsbaugh W A (2016). It takes two to tango: when and where dual nutrient (N & P) reductions are needed to protect lakes and downstream ecosystems.

Environmental Science & Technology, 50(20): 10805–10813

Probst P, Wright M N, Boulesteix A L (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery, 9(3): e1301

Qian S S (2009). Environmental and Ecological Statistics with R. New York: Chapman and Hall/CRC

Quinlan R, Filazzola A, Mahdiyan O, Shuvo A, Blagrave K, Ewins C, Moslenko L, Gray D K, O'Reilly C M, Sharma S (2021). Relationships of total phosphorus and chlorophyll in lakes worldwide. Limnology and Oceanography, 66(2): 392–404

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Redfield A C (1958). The biological control of chemical factors in the environment. American Scientist, 46(3): 205–221

Rousso B Z, Bertone E, Stewart R, Hamilton D P (2020). A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes. Water Research, 182: 115959

Rowland F E, Stow C A, Johengen T H, Burtner A M, Palladino D, Gossiaux D C, Davis T W, Johnson L T, Ruberg S (2020). Recent patterns in Lake Erie phosphorus and chlorophyll *a* concentrations in response to changing loads. Environmental Science & Technology, 54(2): 835–841

Rudin C (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5): 206–215

Ryo M, Angelov B, Mammola S, Kass J M, Benito B M, Hartig F (2021). Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. Ecography, 44(2): 199–205

Sankaran M, Hanan N P, Scholes R J, Ratnam J, Augustine D J, Cade B S, Gignoux J, Higgins S I, Roux X L, Ludwig F, et al. (2005). Determinants of woody cover in African savannas. Nature, 438(7069): 846–849

Schallenberg M (2021). The application of stressor-response relationships in the management of lake eutrophication. Inland Waters, 11(1): 1–12

Sinha E, Michalak A M, Balaji V (2017). Eutrophication will increase during the 21st century as a result of precipitation changes. Science, 357(6349): 405–408

Søndergaard M, Jensen J P, Jeppesen E (2003). Role of sediment and internal loading of phosphorus in shallow lakes. Hydrobiologia, 506–509(1–3): 135–145

Soranno P A, Cheruvelil K S, Bissell E G, Bremigan M T, Downing J A, Fergus C E, Filstrup C T, Henry E N, Lottig N R, Stanley E H, et al. (2014). Cross-scale interactions: quantifying multi-scaled cause-effect relationships in macrosystems. Frontiers in Ecology and the Environment, 12(1): 65–73

Stockwell J D, Doubek J P, Adrian R, Anneville O, Carey C C, Carvalho L, Domis L N D S, Dur G, Frassl M A, Grossart H P, et al. (2020). Storm impacts on phytoplankton community dynamics in lakes. Global Change Biology, 26(5): 2756–2784

Stow C A, Cha Y (2013). Are chlorophyll a–total phosphorus correlations useful for inference and prediction? Environmental Science & Technology, 47(8): 3768–3773

Sun A Y, Scanlon B R (2019). How can big data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. Environmental Research Letters, 14(7): 073001

Tiyasha, Tung T M, Yaseen Z M (2020). A survey on river water quality modelling using artificial intelligence models: 2000–2020. Journal of Hydrology, 585: 124670

Tong Y, Xu X, Qi M, Sun J, Zhang Y, Zhang W, Wang M, Wang X, Zhang Y (2021). Lake warming intensifies the seasonal pattern of internal nutrient cycling in the eutrophic lake and potential impacts on algal blooms. Water Research, 188: 116570

Wagner T, Soranno P A, Webster K E, Cheruvelil K S (2011). Landscape drivers of regional variation in the relationship between total phosphorus and chlorophyll in lakes. Freshwater Biology, 56(9): 1811–1824

Wang R, Kim J H, Li M H (2021). Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach. Science of the Total Environment, 761: 144057

Woolway R I, Kraemer B M, Lenters J D, Merchant C J, O'Reilly C M, Sharma S (2020). Global lake responses to climate change. Nature Reviews. Earth & Environment, 1(8): 388–403

Wright M N, Ziegler A (2017). Ranger: a fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software, 77(1): 1–17

Xu Y, Schroth A W, Isles P D F, Rizzo D M (2015). Quantile regression improves models of lake eutrophication with implications for ecosystem-specific management. Freshwater Biology, 60(9): 1841–1853

Yadav S, Shukla S (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In: 2016 IEEE 6th International Conference on Advanced Computing (IACC). IEEE

Yang L, Shami A (2020). On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing, 415: 295–316

Yuan L L, Jones J R (2020). Rethinking phosphorus–chlorophy Ⅱ relationships in lakes. Limnology and Oceanography, 65(8): 1847–1857

Yusta S C (2009). Different metaheuristic strategies to solve the feature selection problem. Pattern Recognition Letters, 30(5): 525–534

Zagarese H E, de los Ángeles Gonźalez Sagrario M, Wolf-Gladrow D, Nõges P, Nõges T, Kangur K, Matsuzaki S I S, Kohzu A, Vanni M J, Özkundakci D, et al. (2021). Patterns of CO_2 concentration and inorganic carbon limitation of phytoplankton biomass in agriculturally eutrophic lakes. Water Research, 190: 116715

Zhang Y, Qin B, Zhu G, Shi K, Zhou Y (2018). Profound changes in the physical environment of Lake Taihu from 25 years of long-term observations: implications for algal bloom outbreaks and aquatic macrophyte loss. Water Resources Research, 54(7): 4319–4331

Zou W, Zhu G, Cai Y, Xu H, Zhu M, Gong Z, Zhang Y, Qin B (2020). Quantifying the dependence of cyanobacterial growth to nutrient for the eutrophication management of temperate-subtropical shallow lakes. Water Research, 177: 115806