

# Evaluation of Nitrate Concentration Estimation using Machine Learning Models in Lake Erie Watersheds, USA

Saeyool Chloe Park

Seoul International School, 15 Seongnam-daero 1518beon-gil, Sujeong-gu, Seongnam-si, Gyeonggi-do, 13113 Republic of Korea;  
chloe.park.0405@gmail.com

**ABSTRACT:** This study analyzes machine-learning methods used to estimate daily nitrate concentrations by analyzing streamflow and water quality data from 18 watersheds in the Lake Erie area. Daily nitrate was predicted by using 10 input variables: Julian date, streamflow, total suspended solids, chloride, sulfate, total phosphorus, soluble reactive phosphorus, total Kjeldahl nitrogen, total silica, and conductivity. There are 27 machine learning models that are compared using K-fold cross-validation and analyzed with the mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination ( $R^2$ ). But agricultural land use also dominates 15 of the 18 analyzed watersheds, while two watersheds are urban-dominated, and one (Muskingum) watershed shows the highest forest land use. As a result, Rational Quadratic Gaussian Process Regression (RQ-GPR) and Random Forest (RF) among the machine learning models have the strongest overall performance. The correlation analysis of the variables shows that streamflow and Julian date are the most important variables when it comes to predicting nitrate concentration. This research supports the availability of machine learning models for estimating nitrate concentration, the water quality variable.

**KEYWORDS:** Julian Date, Lake Erie. Machine Learning, Nitrate Concentration, Streamflow.

## ■ Introduction

Nitrate is one of the major water quality variables because it is used in river impairment, eutrophication, and downstream ecological effects.<sup>1,2</sup> In agricultural watersheds, nitrate concentrations are significantly affected by hydrologic streamflow, land-use practices, and seasonal biogeochemical controls.<sup>3-6</sup> However, directly and continuously monitoring the nitrate concentration is more difficult and time-consuming than collecting routine hydrologic or supporting water quality measurements. So, the observed nitrate concentration during the monitoring period is usually collected on a weekly or monthly basis. But, the data were not enough to estimate the annual nitrate loads, leading to inconclusive results.<sup>7</sup> Therefore, it was determined that utilizing machine learning models to estimate daily nitrate concentration offers a faster and simpler approach to daily monitoring.<sup>8,9</sup> In addition, due to watershed heterogeneity, nonlinear models have been widely studied because nitrate concentrations are affected by land use and watershed characteristics.<sup>10,11</sup> As a result, it is necessary to apply machine learning models to estimate nitrate concentration because they estimate nonlinear relationships in data very well. This leads to the main objectives of this study: (1) to construct a nitrate concentration estimation method using selected observed nitrate concentrations with hydrologic streamflow, water-quality, and time variables; (2) to compare the estimated nitrate concentration of 27 machine learning models with the observed nitrate, where the concentration evaluates the performance of machine learning models with the experimental watersheds; and lastly (3) to identify the variables with the greatest effect on nitrate estimation.

## ■ Materials and Methods

### *Study area and watershed characteristics:*

18 watersheds from Lake Erie, in Ohio, USA, were used for this study, including Maumee, Wolf, Tiffin, Lost, West, Rock Creek, Beaver, Cuyahoga, Huron, Muskingum, Portage, Raisin, Scioto, Blanchard, Chickasaw, Potato Run, Shallow Run, and S. Turkeyfoot. Figure 1 shows the location of the applied watersheds, while Figure 2 presents the land-use characteristics of the watersheds used in this study. Agriculture was the dominant land use in 15 of the 18 watersheds. In Figure 2, the Wolf and Cuyahoga watersheds were the only urban-dominated basins, with urban fractions of 43.9% and 39.5%. Muskingum had the greatest forest fraction (43.0%), while West (92.0%) and S. Turkeyfoot (90.3%) had the highest agriculture fractions.

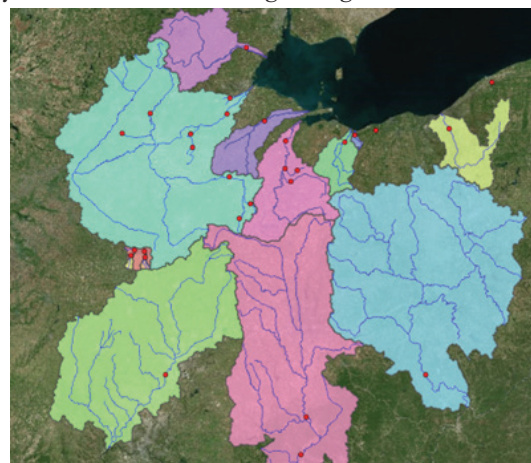


**Figure 1:** Locations of the 18 watersheds in the Lake Erie region, Ohio, USA, used in this study. Watershed boundaries and monitoring locations are shown for the selected basins.

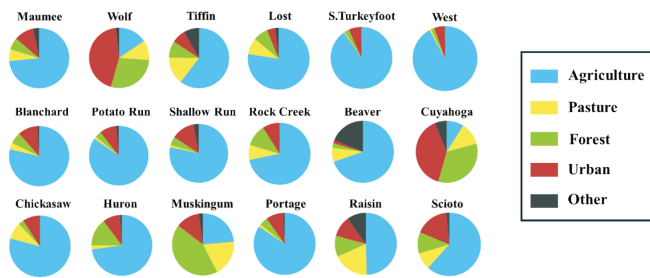


**Figure 2:** Land-use composition of the 18 watersheds used in this study. Agricultural, pasture, forest, urban, and other land-use categories are presented as percentages of total watershed areas.

### Data sources and predictors:

Daily streamflow data were collected from the U.S. Geological Survey (USGS) (<https://waterdata.usgs.gov>), and also the water-quality observations were collected from the National Center for Water Quality Research (NCWQR) (<https://ncwqr-data.org>). To match the time period of the streamflow and water quality data, the monitoring periods in the watershed datasets ranged from 1974-10-02 to 2023-05-15, depending on the watershed. The monitoring records were arranged into a daily database. The variables collected in this study, listed in Table 1, include Julian date (day of year), streamflow, total suspended solids (TSS), chloride, sulfate, total phosphorus (TP), soluble reactive phosphorus (SRP), total Kjeldahl nitrogen (TKN), total silica, and conductivity. In addition, Table 1 also presents the characteristics of the adopted variables. In the end, the target simulation variable to be measured was daily nitrate concentration.

**Table 1:** Input variables of machine learning models for daily nitrate estimation.

| Variables    | Unit/form         | Characteristics                           |
|--------------|-------------------|-------------------------------------------|
| Julian date  | day of year       | Seasonality and recurring annual controls |
| Streamflow   | m <sup>3</sup> /s | Hydrologic flow effects                   |
| TSS          | mg/L              | Soil water quality parameter              |
| Chloride     | mg/L              | Chemical water quality parameter          |
| Sulfate      | mg/L              | Chemical water quality parameter          |
| TP           | mg-P/L            | Nutrient water quality parameter          |
| SRP          | mg-P/L            | Nutrient water quality parameter          |
| TKN          | mg-N/L            | Nutrient water quality parameter          |
| Total silica | mg-Si/L           | Chemical water quality parameter          |
| Conductivity | Umho              | Soil water quality parameter              |

### Model comparison framework:

First of all, Figure 3 shows the overall study process. 10 variables were collected as observational data, including time, two soil, three chemical, and three nutrient water qualities, and hydrologic variables, and 27 machine learning models were used to estimate daily nitrate concentration. The 27 machine learning models were then further categorized into nine machine learning groups according to Table 2. The categories/groups are: linear regression models, regression trees, support vector machines (SVMs), Gaussian process regression (GPR), neural networks, kernel approximation methods, ensembles of trees, efficiently trained linear models, and random forests (RFs). Eventually, all 27 models and the categories were used to compare watershed performance. Furthermore, the created dataset was split into 80% for training and 20% for testing, with 10-

fold cross-validation used during model application. Then, the model performance was analyzed with MAE, RMSE, and R2. Finally, the performance of machine learning models was compared, and the optimal models to produce the best results were presented under the general conditions.

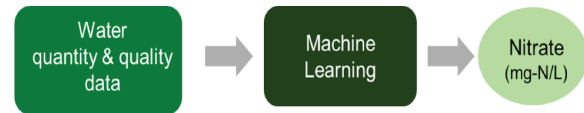


**Figure 3:** Overall framework of this study, including input variable data (Table 1) collection, machine learning model application (Table 2), and nitrate concentration estimation (Figures 4-7).

**Table 2:** Machine learning categories and models screened in the MATLAB-based application.

| Category                          | Representative models                                                                                 | Count |
|-----------------------------------|-------------------------------------------------------------------------------------------------------|-------|
| Linear regression models          | Robust, linear, interaction, and stepwise linear regressions                                          | 4     |
| Regression trees                  | Fine, medium, and coarse trees                                                                        | 3     |
| Gaussian Process Regression (GPR) | Rational quadratic GPR, squared exponential GPR, Matern 5/2 GPR, and exponential GPR                  | 4     |
| Support vector machines (SVMs)    | Linear SVM, quadratic SVM, cubic SVM, fine Gaussian SVM, medium Gaussian SVM, and coarse Gaussian SVM | 6     |
| Neural networks                   | Narrow, medium, and wide neural networks                                                              | 3     |
| Kernel approximation models       | SVM kernel and least-squares kernel regressions                                                       | 2     |
| Ensembles of trees                | Boosted trees and bagged trees                                                                        | 2     |
| Efficiently trained linear models | Efficient linear least squares and efficient linear SVMs                                              | 2     |
| Random forest (RF)                | Random forest                                                                                         | 1     |

### Performance evaluation methods:

Three standard regression error analysis methods were utilized: root mean square error (RMSE), mean absolute error (MAE), and coefficient of Determination (R2) to evaluate machine model performance. These methods compare the predicted nitrate concentrations with the observed values.

The RMSE quantifies the standard deviation between the estimated nitrate concentrations by machine learning models and the observed nitrate concentrations as follows:

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

where  $y_i$  is the observed nitrate concentration,  $\hat{y}_i$  is the simulated nitrate concentration by machine learning models, and  $n$  is the number of data points.

The MAE provides a measure of average nitrate estimation bias as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2)$$

The R2 quantifies the proportion of variance in observed and simulated nitrate concentrations by machine learning models, as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3)$$

where  $\bar{y}$  is the mean of the observed nitrate concentrations. While the ranges of the RMSE and MAE are data-dependent, ranging from zero to infinity, zero representing the optimal value, the R2 range ranges from 0 to 1, 1 representing the optimal value. RMSE presents the magnitude of the estimation errors. A smaller RMSE value indicates the machine learning models produce fewer large mistakes and create a more

accurate estimation. MAE shows the average absolute difference between the estimated and observed values. Lower MAE values demonstrate a better machine learning model performance because the estimations are closer to the actual data. R2 measures how well the model explains the changes in nitrate concentration. The R2 value ranges from 0 to 1, and values closer to 1 indicate that the machine learning model estimates the observed data more accurately. Overall, a well-performing model provides low RMSE and MAE and a high R2.

## Results

### *Watershed-wise machine learning models evaluation:*

Figure 4 shows the performance results of 27 machine learning models that estimate nitrate concentration, using 10 input variables per watershed, and also combining RMSE, MAE, and R2. The Rational Quadratic Gaussian Process Regression (RQ-GPR) had the best-performing model compared to all three evaluation methods in eight watersheds, including Cuyahoga, Potato Run, Muskingum, Scioto, Shallow Run, S. Turkeyfoot, West, and Wolf. Meanwhile, a small number of the bagged trees model and a medium neural network were selected. The RF dominated all three evaluation methods in the Raisin and Maumee watersheds. Lastly, in the Huron, Rock Creek, Portage, and Tiffin watersheds, the Exponential GPR and RQ-GPR models showed the best performance results. Overall, in Figure 5, five of 27 different machine learning models were shown as the optimal models depending on watersheds and performance methods. Among them, RQ-GPR was selected as the most optimal machine learning model in each performance method per watershed. RQ-GPR appeared in 30 out of the 54 optimal performance-error methods, and RF was the second most frequently selected machine learning model, with 14 selections, and, in addition, the exponential GPR with only 6 selections. These results suggest that RQ-GPR is the most powerful model for nitrate concentration, whereas RF performs especially well in some agricultural watersheds.

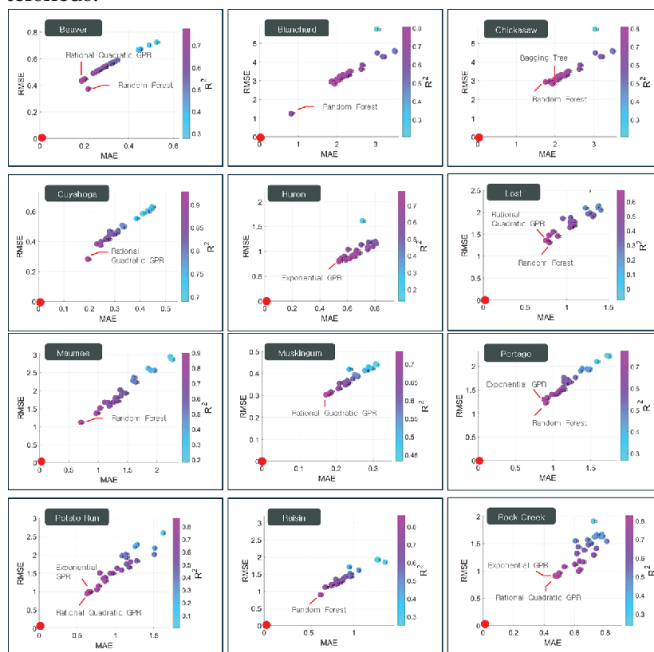
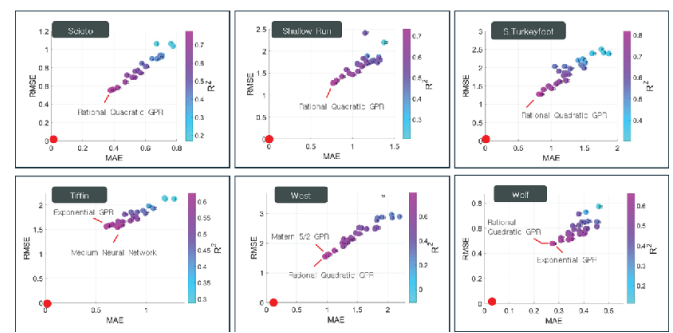


**Figure 4:** Selection of the basin-wise best-performing model for calculating the mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R2) data for each basin. Note that the red point is the optimal point. The closer the point is to the optimal point, the better the performance of the corresponding machine learning models. RQ-GPR and RF models were generally located closer to the optimal point than other machine learning models across most watersheds. This shows that these two models are among the most effective approaches for nitrate concentration estimation across the study watersheds.

|               | MAE     | RMSE    | R <sup>2</sup> |
|---------------|---------|---------|----------------|
| Beaver        | RQ-GPR  | RF      | RF             |
| Blanchard     | RF      | RF      | Bagging        |
| Chickasaw     | RF      | Bagging | RF             |
| Portage       | RQ-GPR  | RF      | Exp GPR        |
| Rock Creek    | Exp GPR | RQ-GPR  | RQ-GPR         |
| Huron         | Exp GPR | Exp GPR | Exp GPR        |
| Lost          | RQ-GPR  | RF      | RQ-GPR         |
| Tiffin        | Exp GPR | Med NN  | Med NN         |
| Cuyahoga      | RQ-GPR  | RQ-GPR  | RQ-GPR         |
| Potato Run    | RQ-GPR  | RQ-GPR  | RQ-GPR         |
| Muskingum     | RQ-GPR  | RQ-GPR  | RQ-GPR         |
| Scioto        | RQ-GPR  | RQ-GPR  | RQ-GPR         |
| Shallow Run   | RQ-GPR  | RQ-GPR  | RQ-GPR         |
| S. Turkeyfoot | RQ-GPR  | RQ-GPR  | RQ-GPR         |
| West          | RQ-GPR  | RQ-GPR  | RQ-GPR         |
| Wolf          | RQ-GPR  | RQ-GPR  | RQ-GPR         |
| Raisin        | RF      | RF      | RF             |
| Maumee        | RF      | RF      | RF             |

Abbreviations: RQ-GPR Rational Quadratic Gaussian Process, RF Random Forest, Exp GPR Exponential Gaussian Process, Bagging Bagging Tree, Med NN Medium Neural Network

**Figure 5:** Best-performing machine learning models identified for each watershed based on MAE, RMSE, and R2 evaluation criteria. The left column is the watershed list, and the three right columns with colors are the lists of best-performing machine learning models by evaluation method. Five of 27 machine learning models (RQ-GPR, RF, Exponential GPR, Med NN, and Bagging Tree) were shown as the optimal models based on watersheds and performance methods. The RQ-GPR is the most frequently selected optimal model for estimating nitrate concentration. It indicates that RQ-GPR is the most reliable nitrate concentration estimation result in any type of watershed, whereas RF performs well in large agricultural watersheds.

### *Scatter Plots of observed vs. estimated nitrate concentrations:*

The observed-versus-estimated nitrate concentration scatter plots for the best-performing machine learning models in each watershed are shown in Figures 6 and 7. Most strong coincidences occurred around the 1:1 line at low to moderate concentrations, therefore indicating that the trained machine learning models simulate the variation of daily nitrate concentration well. These results imply that most of the nitrate concentration data are low or medium nitrate concentrations.

However, at high nitrate concentrations, the point is far away, and deviations from the 1:1 line become more obvious. This result shows that high-concentration estimation is difficult for trained machine learning models because high concentrations are rare and have a smaller training dataset. Therefore, the accuracy of high-concentration estimation is strongly related to the overall accuracy of the machine learning model estimation.

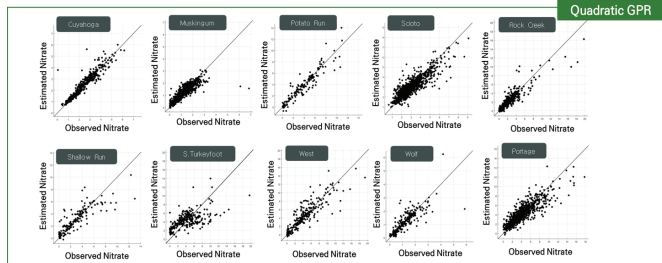


**Figure 6:** Observed-versus-estimated nitrate relationships using Rational Quadratic Gaussian Process Regression (RQ-GPR) for the optimal machine learning method in the 10 watersheds (Cuyahoga, Muskingum, Potato Run, Scioto, Rock Creek, Shallow Run, S. Turkeyfoot, West, Wolf, Portage). Most data points are distributed close to the 1:1 line. The high correlation between observed and estimated nitrate concentrations shows the model's accuracy, particularly within low-to-moderate nitrate concentration ranges. However, at high nitrate concentrations, the accuracy of the estimated nitrate concentration is low. This may be because of the relatively small number of training samples that represent high nitrate concentrations.

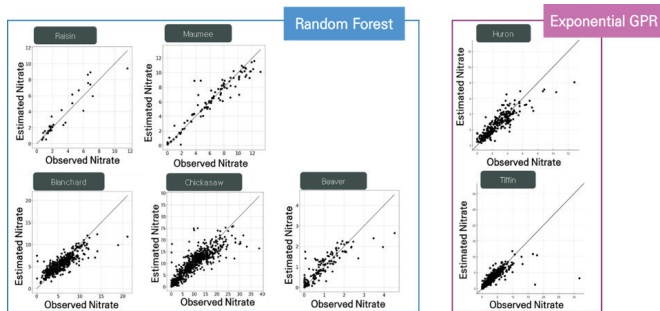


**Figure 7:** Observed-versus-estimated nitrate relationships using random forest for the optimal machine learning method in five watersheds (Raisin, Maumee, Blanchard, Chickasaw, and Beaver) and Exponential GPR for the optimal machine learning method in two watersheds (Huron and Tiffin). Most data points are distributed close to the 1:1 line, within low-to-moderate nitrate concentration ranges. Across all seven watersheds, deviations from the 1:1 line become clearer at high nitrate concentrations because of the small number of high-concentration training samples available for model training.

#### Variable significance evaluation:

Figure 8 shows the correlation heatmap of nitrate concentrations versus applied input variables for each watershed. The variable significance analysis revealed that streamflow and the Julian date were the most effective predictors of nitrate concentrations. This result is possible because streamflow summarizes dilution, runoff generation, storm transport, and tile-drain response, while the Julian date captures recurring seasonal variables such as crop cycles and fertilizer timing of agricultural activity. In contrast, other water-quality variables provide little simulation information. These results further indicate hydrology and seasonality as dominant features, which suggest that much of the nitrate concentration is explained by streamflow through the watershed and by the seasonal conditions under which that transport occurs.

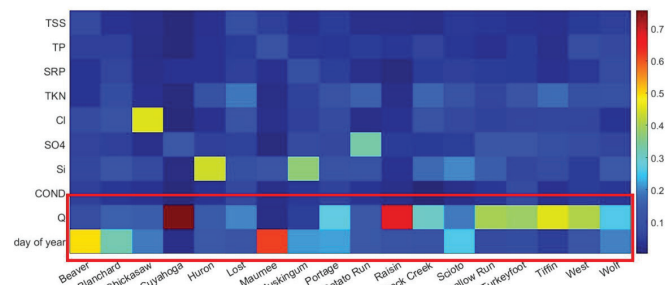


**Figure 8:** 10 input variable correlation heatmap with nitrate concentrations in 18 watersheds. The right color bar shows the magnitude of correlation coefficients between 0 and 1. Streamflow and Julian date exhibit stronger correlations with nitrate concentration than the other input variables. In contrast, other water-quality variables show weak correlations with nitrate concentration. This shows that water quantity and seasonality are the primary variables to estimate nitrate concentration.

## Discussion

Based on nitrate concentration estimation results from the 27 machine learning models applied to 18 watersheds, the optimal models are different depending on the watersheds. However, RQ-GPR and RF models show the most reliable results for daily nitrate estimation in overall watersheds. The RQ-GPR model provides more reliable nitrate concentration estimation results in both urban and agricultural watersheds. This indicates that kernel-based models, such as GPRs, achieved good estimation results on nonlinear relational data in any watershed characteristics. In addition, the RF model shows good results in the Raisin and Maumee. These two watershed characteristics are that the watershed area is larger than 15,000 km<sup>2</sup>. Thus, I propose that the RF model is suitable for large watersheds.

Regarding correlation results, streamflow and Julian date were shown to be strongly significant variables for estimating nitrate concentration. These results are consistent with previous studies showing a strong correlation between nitrate concentration and streamflow or seasonal time.<sup>12,13</sup> This is because streamflow integrates water-quality transport processes, storm flushing of nutrients, drainage activation, and the travel times of water-quality indicators, which produce a high correlation between streamflow and nitrate concentration; whereas seasonality, represented by Julian date, is one of the main factors affecting nitrate concentration as agricultural activities, such as cropping patterns and fertilizer application, are strongly correlated with nitrate concentration in streamflow.

Also, this study presents a limitation regarding machine learning model applications. The high nitrate concentrations are difficult to simulate because such concentrations are not observed very often, and machine learning models have few chances to train. This indicates that high nitrate concentrations should be sufficiently monitored or accumulated for accurate estimation.

## Conclusion

This study estimated daily nitrate concentrations using 27 machine learning models with 10 input variables across 18 watersheds near Lake Erie in the United States and reached the following conclusions. RQ-GPR showed the most consistent performance across different watershed types, while

RF worked especially well in large-scale watersheds. But also, streamflow and Julian date turned out to be the most important input variables because of the hydrology and seasonality affecting nitrate transport. Estimating high concentrations was still difficult, and improving this should be a focus for future work. Overall, this study suggested that machine learning models are an applicable method that could help with nitrate concentration estimation without requiring continuous daily monitoring.

## ■ Acknowledgments

I want to thank my mentor, Sunghoon Jang, at Johns Hopkins University, for his constant support and guidance throughout my endeavors in this project.

## ■ References

1. Carpenter, S. R., Caraco, N. F., Correll, D. L., Howarth, R. W., Sharpley, A. N., & Smith, V. H. (1998). Nonpoint pollution of surface waters with phosphorus and nitrogen. *Ecological Applications*, 8(3), 559–568. [https://doi.org/10.1890/1051-0761\(1998\)008\[0559:N-POSWW\]2.0.CO;2](https://doi.org/10.1890/1051-0761(1998)008[0559:N-POSWW]2.0.CO;2)
2. Howarth, R. W., Swaney, D. P., Billen, G., Garnier, J., Hong, B., Humborg, C., Johnes, P., Mörth, C. M., & Marino, R. (2011). Coupled biogeochemical cycles: Eutrophication and hypoxia in coastal marine ecosystems. *Frontiers in Ecology and the Environment*, 9(1), 18–26. <https://doi.org/10.1890/100008>
3. Meals, D. W., Dressing, S. A., & Davenport, T. E. (2010). Lag time in water quality response to best management practices: A review. *Journal of Environmental Quality*, 39(1), 85–96. <https://doi.org/10.2134/jeq2009.0108>
4. Wollheim, W. M., Bernal, S., Burns, D. A., Gareis, J. A. L., Hensley, R. T., Kaushal, S. S., Newcomer, T. A., Rosi, E. J., & Zarnetske, J. P. (2017). River network saturation concept: Factors influencing nitrate removal in river networks. *Biogeosciences*, 14(15), 4547–4562. <https://doi.org/10.5194/bg-14-4547-2017>
5. Smith, R. A., Alexander, R. B., & Schwarz, G. E. (2018). Nutrient yields and eutrophication of U.S. streams and rivers. *Environmental Science & Technology*, 52(8), 4189–4204. <https://doi.org/10.1021/acs.est.8b00239>
6. Song, J. H., Her, Y., & Guo, T. (2022). Quantifying the contribution of direct runoff and baseflow to nitrogen loading in the Western Lake Erie Basins. *Scientific reports*, 12(1), 9216. <https://doi.org/10.1038/s41598-022-12740-1>
7. Pellerin, B. A., Saraceno, J. F., Shanley, J. B., Sebestyen, S. D., Aiken, G. R., Wollheim, W. M., & Bergamaschi, B. A. (2014). Taking the pulse of snowmelt: In situ sensors reveal seasonal, event, and diurnal patterns of nitrate and dissolved organic matter variability in an upland forest stream. *Biogeochemistry*, 116(1–3), 275–291. <https://doi.org/10.1007/s10533-013-9857-8>
8. Rode, M., Wade, A. J., Jakeman, A. J., Hrachowitz, M., Kristensen, E. A., Samaniego, L., & Sprague, L. A. (2021). Machine learning for water quality modeling: Review and perspective. *Hydrology and Earth System Sciences*, 25(12), 6507–6529. <https://doi.org/10.5194/hess-25-6507-2021>
9. Shen, C., Karangizi, S., Lawson, K., & Li, D. (2020). Machine learning in hydrology: A glimpse of the past and a preview of the future. *Water Resources Research*, 56(11), e2019WR026704. <https://doi.org/10.1029/2019WR026704>
10. Qian, S., Zhang, Y., Wu, Y., & Li, X. (2022). Capturing nonlinear relationships in watershed nitrate dynamics using advanced machine learning models. *Journal of Hydrology*, 612, 128200. <https://doi.org/10.1016/j.jhydrol.2022.128200>
11. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
12. Vogel, R. M., Rudolph, B. E., & Hooper, R. P. (2005). Probabilistic behavior of water-quality loads. *Journal of Environmental Engineering*, 131(7), 1081–1089. [https://doi.org/10.1061/\(ASCE\)0733-9372\(2005\)131:7\(1081](https://doi.org/10.1061/(ASCE)0733-9372(2005)131:7(1081)
13. Vogel, R. M., Stedinger, J. R., & Hooper, R. P. (2003). Discharge indices for water quality loads. *Water Resources Research*, 39(10). <https://doi.org/10.1029/2002WR001872>

## ■ Author

Saeyool Chloe Park is a junior attending Seoul International School in South Korea. She is interested in water quality and environmental issues, but she also loves studying environmental challenges through anthropology, data science, and statistics.