

Estimating Riverine Nitrate Concentrations from Discharge History with Bidirectional LSTM and GWO/WOA-Optimized XGBoost: The Choptank River, Chesapeake Bay Watershed

Omid Tajnia^{1✉}, Najmeh Khajeh Dargi²

¹ Technology Expert, Isfahan University of Technology, Isfahan, Iran.

² Ph.D. Student in Knowledge and Information Science, University of Isfahan, Isfahan, Iran.

Corresponding Author, email: (✉ tajnia@of.iut.ac.ir)

Received: 08 April 2025

Accepted: 14 March 2025

Available online: 18 March 2025

Abstract: Nitrate export from agricultural watersheds is a major cause of eutrophication in the Chesapeake Bay, but nitrate is usually sampled only a few times per month. This study estimates nitrate-plus-nitrite concentrations of the Choptank River near Greensboro, Maryland (USGS 01491000), from daily discharge history and seasonality, using 606 samples collected between 1979 and 2011. A LOADEST-type regression, random forest (RF), extreme gradient boosting (XGBoost), a multilayer perceptron (MLP), a bidirectional long short-term memory network (BiLSTM) fed with 90-day discharge sequences, and two hybrid models in which XGBoost was tuned by the grey wolf optimizer (XGB-GWO) and the whale optimization algorithm (XGB-WOA) were trained on water years 1981–2002, validated on 2003–2006 and tested on 2007–2011. The jack-knife estimates of the Weighted Regressions on Time, Discharge and Season (WRTDS) model were used as an external reference. Concentrations increased significantly over the record (Mann–Kendall $Z = 3.66$), so the test period lay above most training values. Among the causal models, the BiLSTM performed best ($R^2 = 0.785$, RMSE = 0.267 mg N/L, NSE = 0.645), followed by MLP (NSE = 0.488) and the two hybrids (NSE ≈ 0.43). Metaheuristic tuning reduced the test RMSE of XGBoost by 16%. The LOADEST-type regression failed to extrapolate its quadratic time trend (NSE = -0.30), and all tree-based models underestimated the test concentrations by about 0.2–0.3 mg N/L. The WRTDS jack-knife estimates (NSE = 0.709) remained the most accurate, but they use samples from before and after each estimated day. Antecedent discharge at daily to annual scales was the main predictor, which is consistent with the dilution of groundwater-derived nitrate during high flows.

Keywords: Nitrate, Water quality, Chesapeake Bay, BiLSTM, XGBoost, Grey wolf optimizer, Whale optimization algorithm

1. Introduction

Excess nitrogen delivered by rivers is a principal driver of eutrophication and summer hypoxia in estuaries worldwide. The Chesapeake Bay, the largest estuary in the United States, has been affected by nutrient enrichment for decades (Kemp et al., 2005). Nitrate is the dominant form of nitrogen in the agricultural rivers of the Delmarva Peninsula, and much of it reaches the streams through groundwater with residence times of years to decades. This delays the response of river concentrations to changes in management (Sanford and Pope, 2013; Van Meter and Basu, 2015). Tracking these slow changes requires reliable estimates of concentrations and loads between the typically sparse sampling dates.

Statistical regression models have long been the standard tool for this purpose. The seven-parameter regression of Cohn et al. (1992), implemented in the LOADEST program (Runkel et al., 2004), relates log-concentration to log-discharge, time and season. The Weighted Regressions on Time, Discharge and Season (WRTDS) method of Hirsch et al. (2010) relaxes the assumption of fixed coefficients by fitting locally weighted regressions and is now widely used by the U.S. Geological Survey through the EGRET package (Hirsch and De Cicco, 2015).

Machine learning offers an alternative that can learn non-linear concentration–discharge relationships and the effects of antecedent conditions directly from data. Tree ensembles such as random forest and XGBoost have been used to simulate nitrate in Mediterranean streams (Dorado-Guerra et al., 2022). Recurrent neural networks have been used to predict nitrate from discharge and other exogenous inputs (Race et al., 2022), to generate continuous daily nitrate series in Iowa (Saha et al., 2023, 2024), and to model nitrate export across the Mississippi River basin (Pandit et al., 2025). A related LSTM framework predicted dissolved oxygen in hundreds of US rivers from hydrometeorological data alone (Zhi et al., 2021). Saha et al. (2024) reported that deep learning compared favourably with LOADEST and WRTDS-Kalman at low-frequency monitoring sites. However, most of these studies used high-frequency sensor data. Much less is known about how machine learning performs with the sparse grab-sample records that make up most long-term monitoring, and whether metaheuristic hyperparameter tuning helps under such data limitations.

The objectives of this study were (i) to estimate nitrate-plus-nitrite concentrations of the Choptank River from daily discharge history and seasonality using a LOADEST-type regression, RF, XGBoost, MLP and a

bidirectional LSTM (BiLSTM); (ii) to evaluate whether tuning XGBoost with the grey wolf optimizer (GWO; Mirjalili et al., 2014) and the whale optimization algorithm (WOA; Mirjalili and Lewis, 2016) improves its generalization; and (iii) to compare the models with WRTDS estimates in a chronological evaluation that includes a long-term upward trend in concentrations.

2. Materials and Methods

2.1. Study Site

The Choptank River near Greensboro, Maryland (USGS station 01491000; 38.997° N, 75.786° W) drains 292.7 km² (113 mi²) of the Delmarva Peninsula on the Atlantic Coastal Plain (Fig. 1). The watershed is dominated by row-crop agriculture on permeable, sandy soils. Groundwater contributes a large share of streamflow, and nitrate from past fertilizer applications is stored in the surficial aquifer. The Choptank is a tributary of the Chesapeake Bay and one of the rivers monitored for Bay restoration programmes.

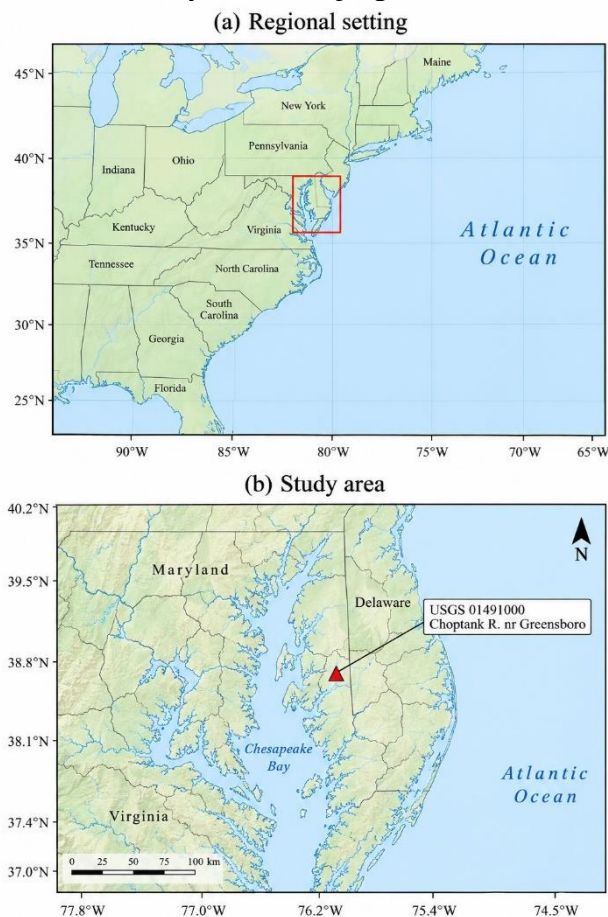


Figure 1. Location of the Choptank River station USGS 01491000 on the Delmarva Peninsula

2.2. Data

Daily mean discharge (Q , m³/s; 1 October 1979 – 30 September 2011) and filtered nitrate-plus-nitrite concentrations (USGS parameter code 00631, mg/L as N; 606 samples) were obtained from the example dataset Choptank_eList of the USGS EGRET package (Hirsch

and De Cicco, 2015; github.com/cran/EGRET). The original data come from the USGS National Water Information System. One sample was below the detection limit and was represented by the midpoint of its censoring interval. The EGRET object also contains WRTDS concentration estimates for every sample, which were obtained by jack-knife cross-validation (each sample is estimated with a model fitted without it). Descriptive statistics are given in Table 1.

Table 1. Descriptive statistics of nitrate samples and discharge (water years 1980–2011)

Variable	Min	Max	Mean	SD	Skew	Kurt
$\text{NO}_3^- + \text{NO}_2^-$ (mg N/L)	0.03	2.43	1.14	0.37	0.04	0.17
Q on sample days (m ³ /s)	0.11	158.57	9.03	15.97	4.10	22.71
Q , all days (m ³ /s)	0.01	246.36	4.09	7.18	10.43	213.47

Sampling intensity varied from 4–5 samples per year in 1982–1984 to 55 in 1989, and was about 15–20 per year after 2000 (Fig. 2). Mean concentration increased from 1.09 mg N/L in water years 1980–2002 to 1.27 mg N/L in 2003–2006 and 1.29 mg N/L in 2007–2011. The Mann–Kendall test on annual mean concentrations confirmed a significant upward trend ($Z = 3.66$, $p < 0.001$, Sen's slope = 0.09 mg N/L per decade). Concentrations decrease with increasing discharge (Fig. 3), which indicates dilution of nitrate-rich base flow by storm runoff. They are highest in winter and lowest in early summer.

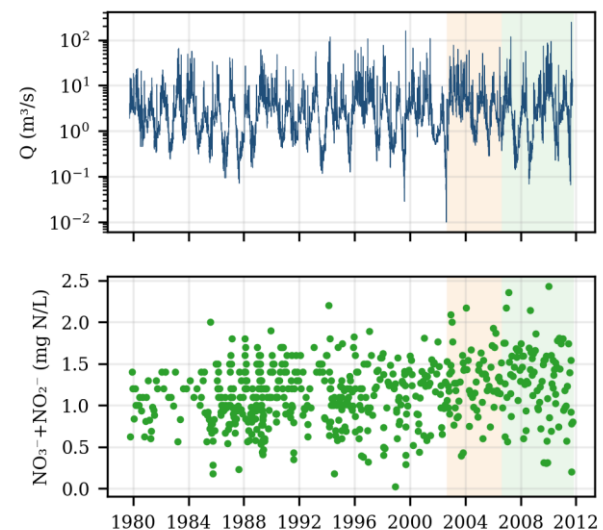


Figure 2. Daily discharge and nitrate samples of the Choptank River (shaded: validation and test periods)

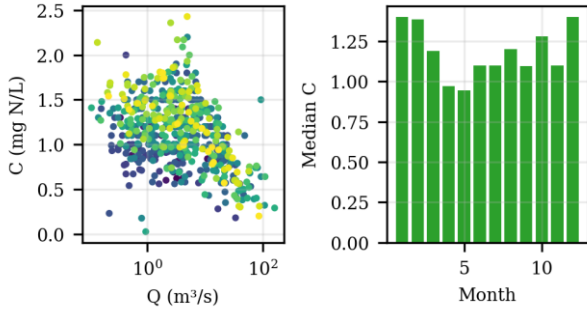


Figure 3. (a) Concentration–discharge relationship (colour: year) and (b) median concentration by month

2.3. Predictors and Data Partitioning

For each sample day t , 12 predictors were derived from the discharge record and the calendar: $\ln Q(t)$, $\ln Q(t-1)$, the daily change $\Delta \ln Q$, the logarithms of the 7-, 30-, 90- and 365-day mean discharge, the 30-day discharge anomaly $\ln Q(t) - \ln Q_{30}(t)$, the sine and cosine of the decimal year, and decimal time. All predictors use only information up to day t , so the models can be applied to estimate concentrations on unsampled days. The BiLSTM received the previous 90 days of $[\ln Q, \sin, \cos, \text{time}]$. All models were trained on $\ln C$, and their predictions were back-transformed with the exponential function.

The data were split by water year into training (1981–2002), validation (2003–2006) and test (2007–2011) periods. The first 455 days of the record were reserved for the 365-day mean discharge and the 90-day sequences, so that 592 samples were evaluated (training 435, validation 66, test 91).

2.4. Models

LOADEST-type regression. Following Cohn et al. (1992), a multiple regression of $\ln C$ on $\ln Q$, $(\ln Q)^2$, time, time², $\sin(2\pi t)$ and $\cos(2\pi t)$ was fitted by least squares:

$$\ln C = \beta_0 + \beta_1 \ln Q + \beta_2 (\ln Q)^2 + \beta_3 T + \beta_4 T^2 + \beta_5 \sin(2\pi T) + \beta_6 \cos(2\pi T) \quad (1)$$

WRTDS (Hirsch et al., 2010) estimates the same type of regression separately for every day, with weights that decrease with distance in time, discharge and season:

$$\ln C = \beta_0 + \beta_1 T + \beta_2 \ln Q + \beta_3 \sin(2\pi T) + \beta_4 \cos(2\pi T) + \varepsilon \quad (2)$$

The jack-knife WRTDS estimates provided with the dataset use samples from the whole 1979–2011 record, including the years after the estimated sample. They are therefore not causal forecasts, and they are reported as an external reference rather than as a competitor.

RF (Breiman, 2001) used 500 trees with a minimum leaf size of 2. **XGBoost** (Chen and Guestrin, 2016) used 300 trees of depth 4 and a learning rate of 0.05. The **MLP** had two hidden layers (32 and 16 ReLU units) and was trained with Adam and early stopping (Pedregosa et al., 2011).

The **BiLSTM** (Schuster and Paliwal, 1997) processes the 90-day input sequence in both temporal directions within the window and concatenates the final forward and backward hidden states:

$$h_t = [\text{LSTM}_{\rightarrow}(x_{1:t}); \text{LSTM}_{\leftarrow}(x_{t:1})] \quad (3)$$

Each direction had 32 LSTM units (Hochreiter and Schmidhuber, 1997), followed by dropout (0.2), a 16-unit dense layer and a linear output. Since the backward pass only covers days up to t , the model remains causal. Five networks were trained with Adam (learning rate 0.002, batch size 16, early stopping with patience 40) and averaged.

2.5. Hybrid XGBoost Models

In GWO the search agents move towards the three best solutions (α, β, δ) using $X(t+1) = (X_1 + X_2 + X_3)/3$, where $X_k = X_k^{\text{best}} - A_k \cdot |C_k \cdot X_k^{\text{best}} - X(t)|$ (Mirjalili et al., 2014). In WOA the agents either shrink around the best solution or follow a logarithmic spiral path towards it (Mirjalili and Lewis, 2016):

$$X(t+1) = |X^*(t) - X(t)| \cdot e^{bl} \cdot \cos(2\pi l) + X^*(t) \quad (4)$$

Seven XGBoost hyperparameters were optimized: number of trees (50–600), maximum depth (2–8), learning rate ($10^{-2.5}$ – $10^{-0.5}$), row and column subsampling (0.5–1), minimum child weight (1–10) and L2 regularization (10^{-2} – $10^{1.5}$). The fitness was the RMSE of the standardized $\ln C$ predictions in the validation period. Both algorithms used 20 agents and 40 iterations (MEALPY; Van Thieu and Mirjalili, 2023).

2.6. Evaluation

Estimates were evaluated in concentration units (mg N/L) with R^2 , RMSE, MAE, NSE (Nash and Sutcliffe, 1970), KGE (Gupta et al., 2009) and Willmott's index (Willmott, 1981), complemented by Taylor diagrams (Taylor, 2001) and error distributions.

$$\text{RMSE} = \sqrt{[(1/n) \sum (C_o - C_s)^2]} \quad (5)$$

$$\text{NSE} = 1 - \sum (C_o - C_s)^2 / \sum (C_o - \bar{C}_o)^2 \quad (6)$$

3. Results

3.1. Hyperparameter Optimization

Table 2. XGBoost hyperparameters (default and optimized) and validation fitness

Model	Tree s	Depth	η	Row w	Col	Min child	λ	Fitness
XGBoost	300	4	0.050 0	1.0 0	1.0 0	1.0	1.0 0	–
XGB-GWO	173	8	0.054 9	1.0 0	1.0 0	5.6	0.6 5	0.431
XGB-WOA	350	8	0.026 4	1.0 0	1.0 0	7.6	0.0 3	0.410

Both optimizers selected deep trees (depth 8) without row or column subsampling, and they compensated with a larger minimum child weight (5.6 and 7.6) and a moderate number of trees (Table 2). WOA reached a lower validation fitness (0.410) than GWO (0.431), starting from 0.483 and 0.462, respectively (Fig. 4). The searches took 2 and 4 min.

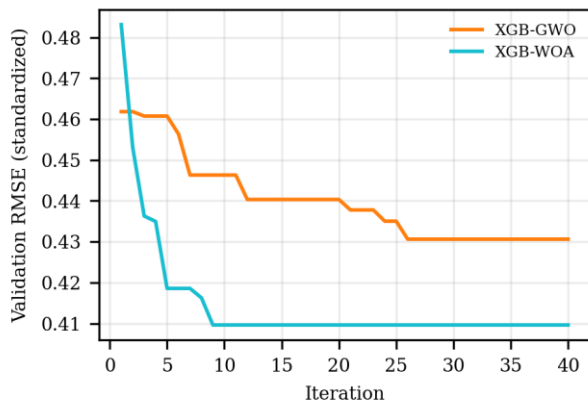


Figure 4. Convergence of GWO and WOA (validation RMSE of standardized $\ln C$)

3.2. Model Performance

Table 3. Performance of nitrate estimates (RMSE and MAE in mg N/L)

Model	NSE (train)	NSE (val)	R ² (test)	RMS E (test)	MA E (test)	NSE (test)	KG E (test)
LOADEST-type	0.313	0.097	0.569	0.510	0.423	0.299	0.324
RF	0.890	0.553	0.670	0.380	0.312	0.280	0.562
XGBoost	0.966	0.466	0.558	0.402	0.319	0.194	0.593
MLP	0.687	0.391	0.647	0.320	0.261	0.488	0.758
BiLSTM	0.763	0.696	0.785	0.267	0.212	0.645	0.607
XGB-GWO	0.970	0.674	0.628	0.338	0.259	0.430	0.659
XGB-WOA	0.956	0.697	0.634	0.337	0.265	0.432	0.644
WRTDS-LOO	0.400	0.458	0.715	0.241	0.178	0.709	0.764

WRTDS-LOO: jack-knife WRTDS estimates computed with the full 1979–2011 record (non-causal reference).

Table 3 summarizes the results. The tree ensembles fitted the training data closely (NSE = 0.89 for RF and 0.97 for XGBoost) but performed poorly in the test period (NSE = 0.28 and 0.19). Tuning by GWO and WOA improved XGBoost in both validation (NSE from 0.47 to about 0.70) and test periods (RMSE from 0.402 to 0.338 mg N/L, a 16% reduction). The two optimizers gave almost identical test scores.

Among the causal models, the BiLSTM was the best. It achieved the highest test R² (0.785) of all models, including WRTDS, and the lowest RMSE (0.267 mg N/L) and highest NSE (0.645) among the causal models. It was also the best causal model in the validation period (NSE = 0.696). The MLP ranked second (test NSE = 0.488). The LOADEST-type regression was the weakest model (test NSE = -0.30): its quadratic time terms, fitted to the flat 1981–2002 period, extrapolated a declining trend into the test years. The WRTDS jack-knife estimates had the lowest test RMSE (0.241 mg N/L) and highest NSE (0.709), but they benefit from samples collected after each estimated day.

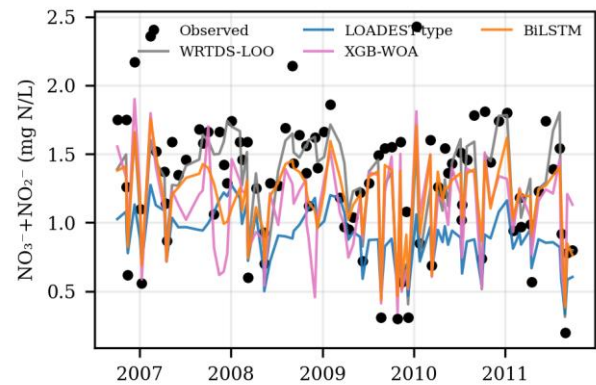


Figure 5. Observed and estimated nitrate concentrations in the test period (water years 2007–2011)

The test-period series (Fig. 5) show that all models captured the seasonal and flow-related variations, with low concentrations during storm events and high concentrations during winter base flow. The main difference was in the level of the estimates. The mean error was -0.39 mg N/L for the LOADEST-type model, -0.27 for RF and XGBoost, -0.20 for the hybrids, -0.17 for MLP, -0.12 for BiLSTM and -0.03 for WRTDS. The machine learning models thus learned the concentration–discharge–season relationship of the training period, but they could not fully reproduce the rise in concentrations after 2002. Tree-based models cannot extrapolate the time predictor beyond its training range, which explains their larger bias.

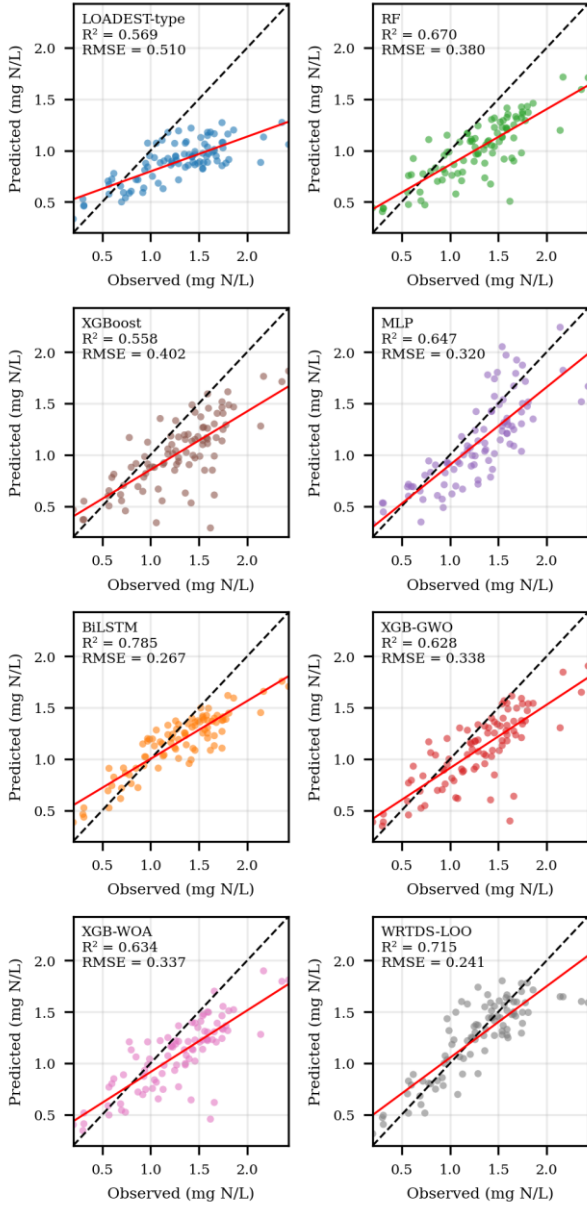


Figure 6. Observed versus estimated nitrate concentrations in the test period

The scatter plots (Fig. 6) and the Taylor diagram (Fig. 7) show that the BiLSTM had the highest correlation (0.89) but underestimated the variability (normalized standard deviation 0.64). The MLP reproduced the variability best (0.94) at a lower correlation (0.80), and WRTDS combined good correlation (0.85) with realistic variability (0.82). The error distributions (Fig. 8) confirm the negative bias of the tree-based and LOADEST-type models.

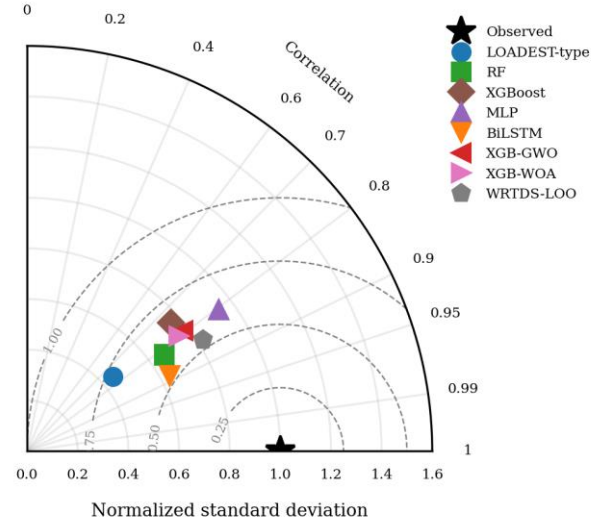


Figure 7. Taylor diagram of the test-period estimates

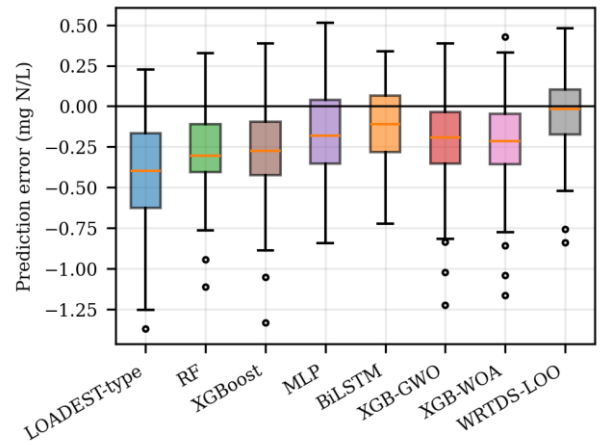


Figure 8. Distribution of estimation errors in the test period

3.3. Importance of Predictors

The permutation importance of XGBoost (Fig. 9) shows that discharge information dominated: the most important predictors were $\ln Q(t-1)$, $\ln Q(t)$, $\ln Q_{365}(t)$ and $Q_{anom30}(t)$, followed by $\ln Q_{30}(t)$. Seasonality ($\sin$) had a smaller effect, and the time predictor had almost no importance for the tree model. The strong role of the one-year mean discharge indicates that wet years, which recharge the surficial aquifer and flush stored nitrate, are followed by higher concentrations. The 30-day anomaly captures the dilution of base flow by storm runoff.

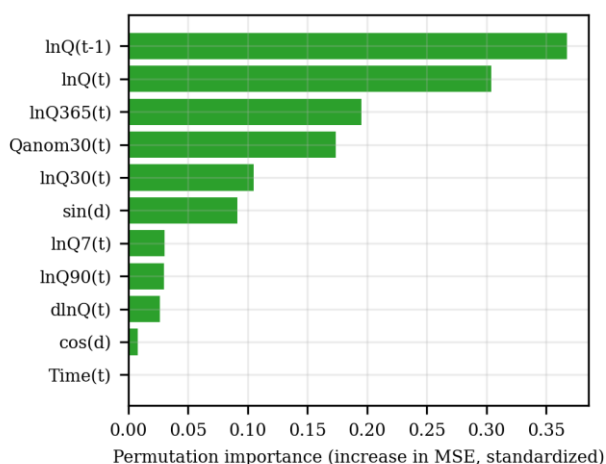


Figure 9. Permutation importance of the XGBoost predictors (test period)

4. Discussion

The main challenge in this dataset was non-stationarity. Nitrate concentrations in the Choptank River increased by about 18% between the training and test periods, most likely because of the delayed arrival of nitrate stored in groundwater (Sanford and Pope, 2013). Models that rely on a time predictor with a fixed functional form (LOADEST-type) or that cannot extrapolate (RF, XGBoost) were biased. Models that use smooth functions of the discharge history (MLP, BiLSTM) were less affected. WRTDS avoids the problem by re-estimating its coefficients locally in time, which is only possible when samples on both sides of the estimated day are available. For real-time or forward estimation, the BiLSTM offers the best compromise in this study.

Metaheuristic tuning was clearly useful for XGBoost: with only 440 training samples, the default model overfitted, and both GWO and WOA found configurations that generalized better. However, the tuned models still did not reach the accuracy of the neural networks, because tuning cannot remove the structural inability of trees to extrapolate trends. GWO and WOA gave practically the same results, so the choice between them is of secondary importance compared with the choice of model structure and predictors.

Several limitations should be noted. Only discharge and calendar information were used. Adding water temperature, specific conductance or land-management data could improve the estimates, as shown by Race et al. (2022) and Pandit et al. (2025). The test period covered only five water years with 91 samples, and the WRTDS reference was not computed in a strictly forward manner. Future work should evaluate a forward-moving WRTDS, combine WRTDS residuals with machine learning corrections, and extend the analysis to several Chesapeake Bay tributaries.

5. Conclusions

Nitrate-plus-nitrite concentrations of the Choptank River were estimated from daily discharge history and seasonality with a LOADEST-type regression, RF, XGBoost, MLP, BiLSTM and two metaheuristic-tuned XGBoost models, and were compared with WRTDS. The main conclusions are: (1) the BiLSTM was the most accurate causal model (test $R^2 = 0.785$, $NSE = 0.645$); (2) GWO and WOA reduced the test RMSE of XGBoost by 16%, but the hybrids remained biased because trees cannot extrapolate the upward concentration trend; (3) a fixed-form LOADEST-type regression failed under the trend ($NSE = -0.30$); (4) the non-causal WRTDS jack-knife estimates were the most accurate overall ($NSE = 0.709$); and (5) antecedent discharge at daily to annual scales was the main control of concentrations. Deep sequence models are therefore recommended for estimating nitrate on unsampled days in trending agricultural rivers, preferably together with periodic recalibration.

6. Data Availability

Discharge, nitrate samples and WRTDS estimates for USGS station 01491000 were obtained from the Choptank_eList dataset of the EGRET R package (Hirsch and De Cicco, 2015; github.com/cran/EGRET), originally from the USGS National Water Information System. The code used for the analysis is available from the authors on request.

7. References

- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).
- Cohn, T. A., Caulder, D. L., Gilroy, E. J., Zynjuk, L. D., & Summers, R. M. (1992). The validity of a simple statistical model for estimating fluvial constituent loads: An empirical study involving nutrient loads entering Chesapeake Bay. *Water Resources Research*, 28(9), 2353–2363.
- Dorado-Guerra, D. Y., et al. (2022). Machine learning models to predict nitrate concentration in a river basin. *Environmental Research Communications*, 4.
- Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1–2), 80–91.
- Hirsch, R. M., & De Cicco, L. A. (2015). *User guide to Exploration and Graphics for RivEr Trends (EGRET) and dataRetrieval: R packages for hydrologic data* (version 2.0). U.S. Geological Survey Techniques and Methods, book 4, chap. A10.
- Hirsch, R. M., Moyer, D. L., & Archfield, S. A. (2010). Weighted regressions on time, discharge, and season (WRTDS), with an application to Chesapeake Bay river

- inputs. *Journal of the American Water Resources Association*, 46(5), 857–880.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Kemp, W. M., Boynton, W. R., Adolf, J. E., et al. (2005). Eutrophication of Chesapeake Bay: Historical trends and ecological interactions. *Marine Ecology Progress Series*, 303, 1–29.
- Mirjalili, S., & Lewis, A. (2016). The whale optimization algorithm. *Advances in Engineering Software*, 95, 51–67.
- Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. *Advances in Engineering Software*, 69, 46–61.
- Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10(3), 282–290.
- Pandit, A., et al. (2025). Deep learning prediction and interpretation of riverine nitrate export across the Mississippi River Basin. *Water Resources Research*.
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Race, M., et al. (2022). A nonlinear autoregressive exogenous (NARX) model to predict nitrate concentration in rivers. *Environmental Science and Pollution Research*.
- Runkel, R. L., Crawford, C. G., & Cohn, T. A. (2004). *Load Estimator (LOADEST): A FORTRAN program for estimating constituent loads in streams and rivers*. U.S. Geological Survey Techniques and Methods, book 4, chap. A5.
- Saha, G., et al. (2023). A deep learning-based novel approach to generate continuous daily stream nitrate concentration for nitrate data-sparse watersheds. *Science of the Total Environment*.
- Saha, G., et al. (2024). Performance evaluation of deep learning based stream nitrate concentration prediction model to fill stream nitrate data gaps at low-frequency nitrate monitoring basins. *Journal of Environmental Management*.
- Sanford, W. E., & Pope, J. P. (2013). Quantifying groundwater's role in delaying improvements to Chesapeake Bay water quality. *Environmental Science & Technology*, 47(23), 13330–13338.
- Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673–2681.
- Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, 106(D7), 7183–7192.
- Van Meter, K. J., & Basu, N. B. (2015). Catchment legacies and time lags: A parsimonious watershed model to predict the effects of legacy storage on nitrogen export. *PLoS ONE*, 10(5), e0125971.
- Van Thieu, N., & Mirjalili, S. (2023). MEALPY: An open-source library for latest meta-heuristic algorithms in Python. *Journal of Systems Architecture*, 139, 102871.
- Willmott, C. J. (1981). On the validation of models. *Physical Geography*, 2(2), 184–194.
- Zhi, W., Feng, D., Tsai, W.-P., Sterle, G., Harpold, A., Shen, C., & Li, L. (2021). From hydrometeorology to river water quality: Can a deep learning model predict dissolved oxygen at the continental scale? *Environmental Science & Technology*, 55(4), 2357–2368.