American Journal of Applied Chemistry
2026, Vol. 14, No. 3, pp. 42–51
https://doi.org/10.11648/j.ajac.20261403.11

Research Article

Dissolved Oxygen Concentrations Modeling of the Tighen River Water Using Physicochemical Variables and Various Machine-Learning Algorithms in Guinea Republic

Abdoulaye Missira Bangoura[1], Noukpo Medard Agbazo[2,*], Saa Moussa Kamano[3], Mafory Bangoura[4], Kande Bangoura[5]

[1]Department of Chemistry, University of NZerekore, NZerekore, Guinea

[2]Department of Hydrology, University of NZerekore, NZerekore, Guinea

[3]Fishing and aquaculture, Higher Institute of Sciences and Veterinary Medicine of Dalaba, Dalaba, Guinea

4Dramatic Arts, Mory Kante Higher Institute of Arts of Dubreka, Conackry, Guinea

[5]Departement of Hydrology, Marine and Coastal Scientific Research Center, Conakry, Guinea

Abstract

Dissolved oxygen is an essential indicator of water pollution and the critical water quality constituent that impacts aquatic life. Thus, accurate modeling of its concentration is vital for freshwater resource management and protection. Despite this, in the African context, more specifically West Africa, there is virtually no scientific work that has focused on modeling dissolved oxygen concentrations in rivers and lakes. This preliminary work attempted to model and estimate, using others microbiological and physicochemical parameters and machine learning algorithms, the dissolved oxygen concentration of the Tighen River water in the Republic of Guinea. Based on two alternatives, three algorithms such as multiple linear regression (MLR), random forest (RF), and gradient boosting (GB) were employed to model and estimate dissolved oxygen concentrations. Alternative 1 referred to when microbiological and physicochemical parameters exhibiting correlations greater than + 0.1 or less than − 0.1 with dissolved oxygen are used for modeling its concentration, while alternative 2 referred to when variables exhibiting statistically significant correlations with dissolved oxygen are used. Results obtained from the models were evaluated using Nash-Sutcliffe efficiency coefficient (NSE), mean absolute error (MAE), Pearson correlation coefficient (RP), and root mean square error (RMSE) to identify the appropriate alternative and algorithm to model and estimate the dissolved oxygen. In the testing phase, the results showed that (1) among tested alternatives, alternative 2 quasi-systematically presents a smaller RMSE and MAE, and higher NSE and RP, indicating that it is significantly better than the alternative 1. (2) among the employed algorithms, under alternative 2, the RF algorithm exhibits the best performance in modeling dissolved oxygen, therefore, RF outperforms, MLR, and GB algorithm. These findings provide a scientific reference to enhance freshwater resource management and protection in Tighen river.

Keywords

Dissolved Oxygen, Modeling, Machine Learning, Water Quality, Applied Chemistry, Tighen River

*Correspondence: Noukpo Medard Agbazo (agbmednou@gmail.com)

Received: 1 May 2026; **Accepted:** 16 May 2026; **Published:** 27 May 2026

Copyright: © The Author(s), 2026. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

1. Introduction

Dissolved oxygen concentration is an important water quality parameter [1-5]. Moreover, it is known that accurate dissolved oxygen concentration modeling and predicting are crucial for river water quality monitoring, reservoir management and guaranting the sustainability of freshwater resources. For these reasons, numerous researches have focused on dissolved oxygen concentration prediction in the rivers by using data-driven models based on machine learning algorithms [6-10]. For instance, dissolved oxygen concentration in lake, river and reservoir water have been modeled or forecasted by artificial neural network model [11-19]; by feed-forward neural network model [18]; by decision tree models [20]; by Random Forest [21]; by AdaBoost, RF, and Gradient Boosting algorithms [22]; by Adaptive Neuro-Fuzzy Inference Systems [23]; by GA-XGCBXT algorithm [24]; by Kernel Ridge Regression, Elastic Net, and Light Gradient Boosting algorithm [25]; by K-Nearest Neighbors, and Support Vector Machines [26]; by fuzzy logic [27-29], and by support vector machine [30-32]. All of the aforementioned studies have demonstrated the potential of data-driven models based on machine learning algorithms for dissolved oxygen concentration modeling and prediction, including the requirement of less input data. On the other hand, none of the aforementioned studies has been carried out on rivers and streams in Africa and more specifically in West Africa such as Guinea, a country containing more than 1000 waterways and whose rivers are heavily polluted. This finding about West Africa countries could be explained by the unavailability of data on rivers necessary for carrying out such work and the lack of funding for research projects. This preliminary work attempted to model and estimate the dissolved oxygen concentration of the Tighen River in Guinea Republic by using others microbiological and physicochemical parameters and machine learning algorithms. The next section presents the dataset and methodology. The subsequent section deals with the results analysis, and finally, the conclusions are drawn.

2. Materials and Methods

2.1. Materials

Tighen River is situated in the Southern part of Guinea and located in the forest region, precisely in Lola's prefecture (Figure 1). It is bordered at North by the prefecture of Beyla, at Northwest and West by the prefecture of N'Zerekore, at East by the Republic of Ivory Coast, at South by the Liberia Republic. Five sites (Homeakoly1 upstream, Gotekoly, Tighen-mou 1 and Thieta within the city, Foulayapo downstream) distributed along the Tighen River were selected. Figure 1 shows the study area and sampling locations used in this study. Many physicochemical and microbiological parameters (water quality parameters) were measured monthly from April 2024 to March 2026 at each aforementioned site. Temperature, pH, conductivity, suspended solids, turbidity, TDS, dissolved oxygen, oxygen saturation, nitrates, nitrites, phosphate, iron, potassium, manganese, fecal coliforms, total coliforms, were the available physicochemical and microbiological parameters used.

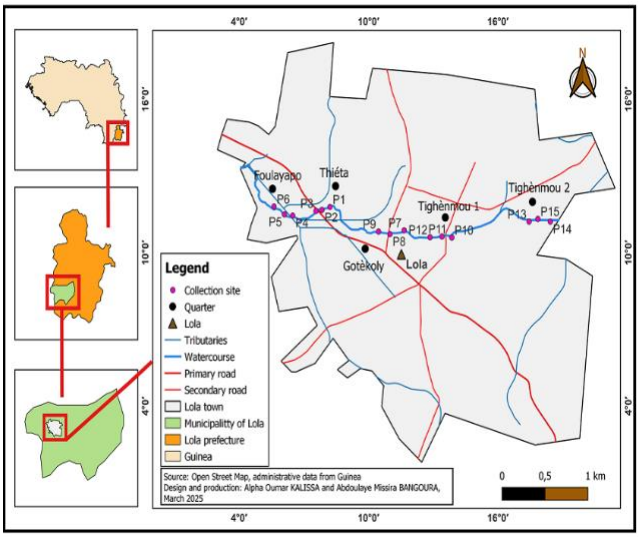

Figure 1. *Study area and sampling locations.*

2.2. Methods

This study employs the following algorithms for dissolved oxygen concentration modeling: multiple linear regression (MLR), random forest (RF), and gradient boosting (GB). The theoretical background of these methods is described below:

2.2.1. Multiple Linear Regression

Multiple Linear Regression (MLR) is used to assess the relationship between a dependent variable and multiple predictor variables. To be specific, MLR quantifies the extent to which each predictor variable contributes to variations in the dependent variable, providing a clearer understanding of the underlying relationships within the dataset [33]. MLR model can be expressed as:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon \quad (1)$$

Where: for i = n observations Y_i =The predicted value of the dependent variable; β_0=They-intercept, representing the value of y when all independent variables are zero; $\beta_1 x_{i1}$=The regression coefficient (β_1) of the first independent variable (x_{i1}), indicating the impact of x_1 on the predicted Y value; $\beta_p x_{ip}$=The regression coefficient of the last independent variable, showing its influence on Y; ε =The model error,

also known as the residuals, representing the unexplained variation in the predicted Y value.

2.2.2. Random Forest

Random Forest (RF) algorithm developed by [34], is used to predict a continuous dependent variable through complex interrelationship analyses. RF algorithm performs a random sampling of the original dataset using the decision tree as the basic random forest classifier resulting in n different sample datasets. RF algorithm is a variant of bagging that fits a multitude of decision trees on different sub-samples to find the output. Sampling features are termed column sampling and data points as row sampling. Trees are built with row and column samples. The advantage of building the model in such a way is that it is robust in estimating new data points [35, 36]. The functional equation for random forest can be expressed as:

$$R_F(x) = \frac{1}{t}\sum_{t=1}^{T} h_t(x) \quad (2)$$

Where t=1, 2,....., T, T is the number of decision trees, f(x) is the predicted output of the random forest for the input sample x, h_t is the predicted output of the t-th decision tree for the input sample x. T is the number of decision trees in the random forest.

2.2.3. Gradient Boosting

The gradient boosting algorithm (GB) [37] requires a set of n data points made of source (x_i) and target (y_i) variables: $(x_i, y_i)_{i=1}^{n}$. GB algorithm is described as follows [38]:

GB needs an initial estimate F_o, which is a leaf:

$$F_o = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma) \quad (3)$$

where L is the loss function, y_i are the true values, and γ is a constant. The first guess helps GB to build subsequent trees based on the previous trees.

After defining F_o, the iterative process can begin. For each iteration:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \text{ for i=1,…,n} \quad (4)$$

Let m be the current iteration (tree) and M the total number of trees, and m varies from 1 to M.

Fit a new decision tree $h_m(x)$ to the pseudo-residuals:

$$h_m(x) = argmin_h \sum_{i=1}^{n} (r_{im} - h(x_i))^2 \quad (5)$$

Update the model:

$$F_m(x) = F_{m-1}(x) - \nu h_m(x) \quad (6)$$

where ν is the learning rate, a free parameter.

Final model: After M iterations, the final model is:

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \nu h_m(x) \quad (7)$$

Gradient boosting regressor (GBR) approach can be seen as a functional gradient algorithm that aims at finding an additive model that minimizes the loss function. Thus, the GBR algorithm iteratively adds at each step a new decision tree that best reduces the loss function [39, 40]. A GBR with (n) number of trees can be stated as:

$$f_P(x_i) = \sum_{p}^{P} y_p h_p(x_i) \quad (8)$$

where h_p is a weak learner that performs poorly individually, y_p is a scaling factor adding the contribution of a tree to the model.

Based on the correlation between dissolved oxygen concentrations and others microbiological and physicochemical (temperature, pH, conductivity, suspended solids, turbidity, TDS, oxygen saturation ($SatO_2$), nitrates (NO_3), nitrites (NO_2), phosphate (PO_4), iron, potassium (K), manganese (Mn), fecal coliforms, total coliforms) parameters, two alternatives are adopted. Alternative 1 referred to when microbiological and physicochemical parameters exhibiting correlations greater than + 0.1 or less than − 0.1 with dissolved oxygen are used for modeling its concentration, while alternative 2 referred to when variables exhibiting statistically significant correlations with dissolved oxygen are used.

2.2.4. Efficiency Criteria

Nash-Sutcliffe efficiency coefficient (NSE), mean absolute error (MAE), Pearson correlation coefficient (RP), and root mean square error (RMSE) are used to measure the efficiency of the model, and the formulas are calculated as shown below [41, 42]:

$$NSE = 1 - \frac{\sum_{i=1}^{N}\left(Do_{i_{expected}} - Do_{i_{predicted}}\right)^2}{\sum_{i=1}^{N}\left(Do_{i_{expected}} - \overline{Do}_{expected}\right)^2} \quad (9)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N} abs\left(Do_{i_{expected}} - Do_{i_{predicted}}\right) \quad (10)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[Do_{i_{expected}} - Do_{i_{predicted}}\right]^2} \quad (11)$$

$$R_P = \frac{\sum_{i=1}^{N}\left(Do_{i_{predicted}} - \overline{Do}_{predicted}\right)\left(Do_{i_{expected}} - \overline{Do}_{expected}\right)}{\sqrt{\sum_{i=1}^{N}\left(Do_{i_{predicted}} - \overline{Do}_{predicted}\right)^2}\sqrt{\sum_{i=1}^{N}\left(Do_{i_{expected}} - \overline{Do}_{expected}\right)^2}} \quad (12)$$

In Equations (9)-(12) where $Do_{i_{predicted}}$ and $Do_{i_{expected}}$ stand for the modeled and measured dissolved oxygen concentrations, respectively (with respective means of $\overline{Do}_{predicted}$ and $\overline{Do}_{expected}$). Moreover, N refers to the number of compared pairs.

The RMSE ranges from zero to infinity, with an RMSE of 0 indicating a perfect model. To be specific a smaller RMSE indicates a more accurate prediction by the model. The MAE ranges from zero to infinity, with MAE=0 denoting a perfect model. R_P value ranges from 0 to 1, and the closer it is to 1, the better the model is. These efficiency criteria were used for both the training and testing datasets. The dataset was divided into 80% of training and 20% of test datasets.

3. Results and Discussion

Figure 2 illustrates the correlation among various variables. To be specific, it delineates the correlation matrix between the physico-chemical, nutritive and microbiological parameters of the waters from the different study sites. The values of this matrix span from -1 to +1. In particular, a progression in the direction of +1 indicates a positive interrelation with the parameter under examination, while a progression toward -1 signifies an inverse association with the designated target. The correlation matrix highlights several interesting relationships between water quality parameters. Because this research was interested in finding a significant relationship between dissolved oxygen as a response variable and other predictors. Therefore, it can be observed that temperature shows relatively strong negative correlations with dissolved oxygen. Subsequently, TDS, K, NO2, PO4, N03 and Iron emerge as subsequent parameters in descending order of their correlational magnitude with dissolved oxygen. Thus, in the study area, temperature, TDS, K, NO2, PO4, NO3 and Iron were assumed to be the main predictors to model or predict dissolved oxygen. However, to confirm this assumption, statistical significance tests are necessary before building the model.

Figure 3 depicts the statistical significance test results of how these parameters are correlated with dissolved oxygen. The results show that for most of the aforementioned parameters, such as Temperature, TDS, K, NO2, and PO4, the upper and the lower bound of the 95% confidence limits of their Pearson coefficient with the dissolved oxygen have the same sign meaning that the correlation between each of them and the dissolved oxygen is statistically significant. These findings indicated that there exists a statistically significant relationship between dissolved oxygen and (Temperature, TDS, K, NO2, and PO4). Conversely, N03 and Iron exhibited statistically non-significant relationship with dissolved oxygen.

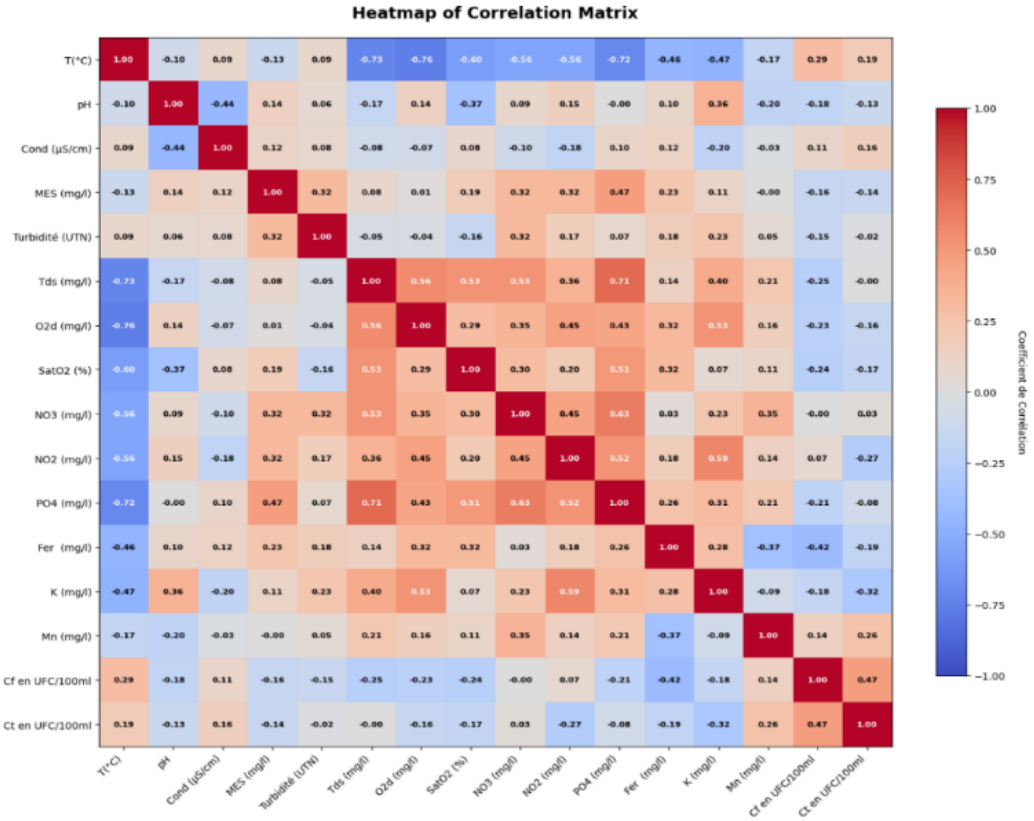

Figure 2. *Correlation Heat map between the physico-chemical, and microbiological parameters of the waters from the different study sites.*

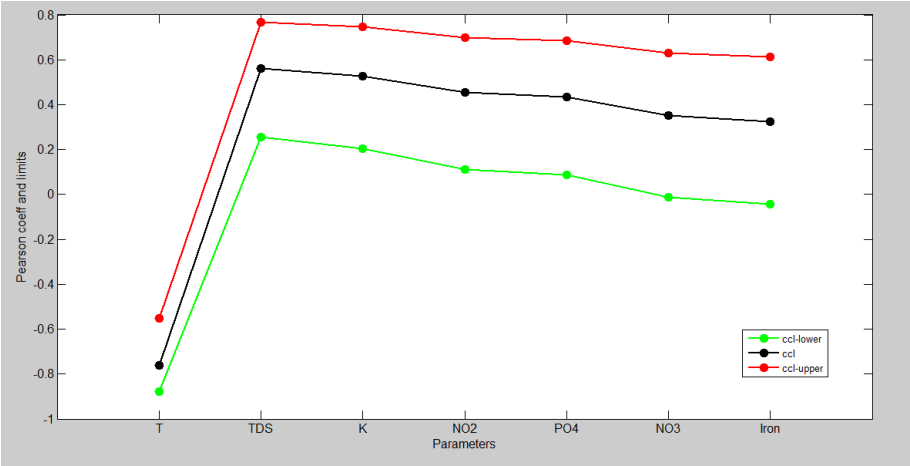

Figure 3. *Pearson coefficient between dissolved oxygen and other parameters. Red and green represent the upper bound and the lower bound of the 95% confidence interval. Black represents the Pearson coefficient.*

Figure 4 shows the time series for the dissolved oxygen prediction results, when variables exhibiting correlations greater than + 0.1 or less than − 0.1 with dissolved oxygen were used for predicting its concentration (alternative 1) using the train data set, test data set, and observed dissolved oxygen values. Figure 4a and Figure 4b plot respectively the fitting effectiveness of the training and predicted data under three modeling algorithms (multiple linear regression (MLR), random forest (RF), and gradient boosting (GB)). It can be seen that, during the training period (Figure 4a), the variations in the dissolved oxygen concentrations have been well-followed by the three algorithms. However, among the three algorithms, MLR sometimes poorly represented the pattern of dissolved oxygen concentration variation at certain points. This finding could reflect the limitations of the MLR algorithm in modeling dissolved oxygen concentrations Considering the testing period (Figure 4b), it is clearly observed that none of the three algorithms managed to accurately follow the pattern of dissolved oxygen concentration variation. This result could be related to the fact that all variables, even those not necessarily exhibiting a statistically significant relationship, were used to predict dissolved oxygen concentrations.

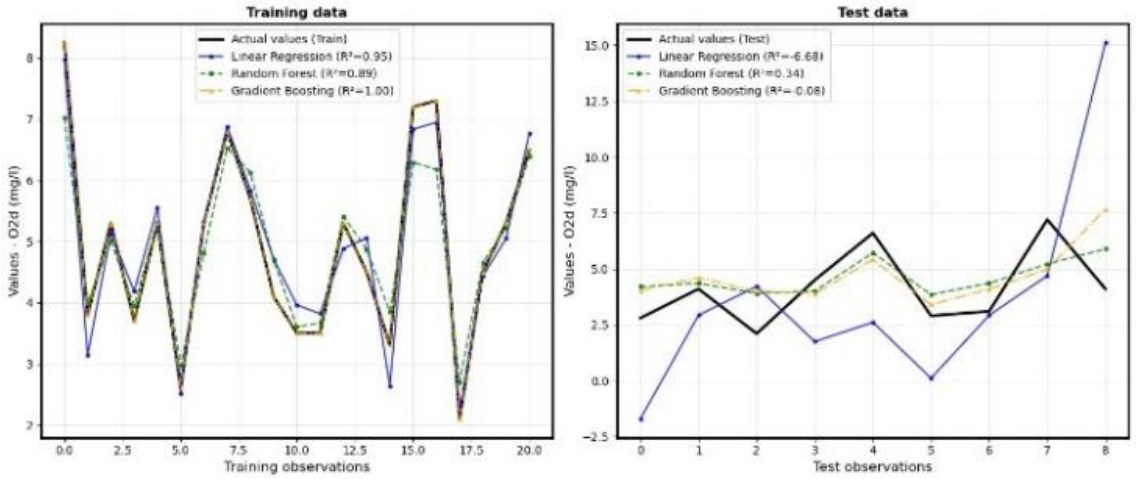

Figure 4. *Comparison of observed and predicted dissolved oxygen concentrations by the three models on the training (left) and test (right) sets on all variables.*

Figure 5 shows the time series for the dissolved oxygen prediction results, when variables exhibiting statistically significant correlations with dissolved oxygen were used for predicting its concentration (alternative 2) using the train data set, test data set, and observed dissolved oxygen values. To be specific this Figure presents the agreement between the modeled and measured dissolved oxygen concentrations. Figure 5a and Figure 5b plot respectively the fitting effectiveness of the training and predicted data under three modeling algorithms. It can be seen that, among the three algorithms, during the training period (Figure 5a), only gradient boosting better reproduces the

pattern of variations in dissolved oxygen concentration. Considering the testing period (Figure 5b), it appears that the three algorithms better reproduced the pattern compared to the case where all variables were used to predict dissolved oxygen concentration (Figure 4b). However, this result warrants further investigation based on more appropriate criteria.

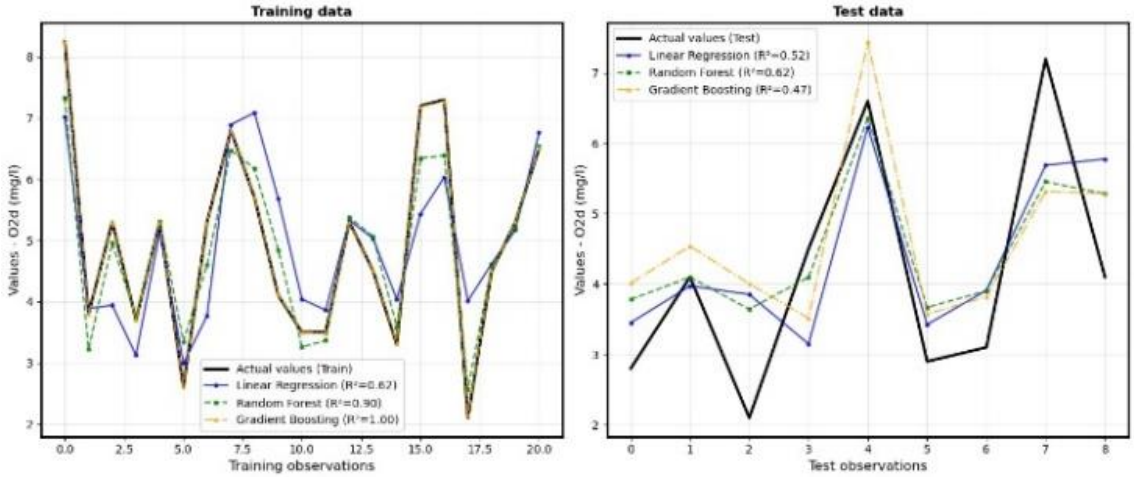

Figure 5. *Comparison of observed and predicted dissolved oxygen concentrations by the three models on the training (left) and test (right) sets on selected variables.*

Based on efficiency criteria such as MSE, RMSE, NSE, and RP, each model's effectiveness during training phase is presented in Figure 6. This Figure provides a comparative illustration during training phase of the performance of the models in modeling dissolved oxygen concentrations by using two alternatives. The first alternative (grey) is when variables exhibiting correlations greater than + 0.1 or less than − 0.1 with dissolved oxygen were used for predicting its concentration using, while the second alternative (orange) is when variables exhibiting statistically significant correlations with dissolved oxygen were used for predicting its concentration.

As shown in Figure 6a-d, in training phase, among all models, MLR performance varies strongly according to the alternative. This finding is mainly revealed in Figure 6b-d when RMSE and MAE values of the two alternatives are compared during the training phase. Indeed, in training phase, alternative 1 minimize MAE and RMSE compared to alternative 2 when MLR is used to model dissolved oxygen concentrations, while none significant difference is noted among the two alternatives when RF, and GB algorithm are used.

Comparing the performance of the used models in the training phase, based on the NSE and Rp values, a very satisfying correlation can be seen for all used models. Furthermore, it is clearly observed that RF, and GB algorithms exhibited the highest RP, NSE and the lowest MAE and RMSE. Thus, during the training phase, RF and GB present a tolerable level of error in modeling dissolved oxygen concentrations. Moreover, comparing RF and GB performance, it can be noted that GB achieves the lowest RMSE and MAE indicating its superior ability to minimize absolute errors and to capture the dissolved oxygen concentrations variation in training phase. Therefore, it presented the highest quality training. Overall, compared with MLR, RF and GB demonstrate superior performance. However, GB emerged as the most reliable model in the training phase, yielding the lowest values for RMSE and MAE and the highest NSE and RP.

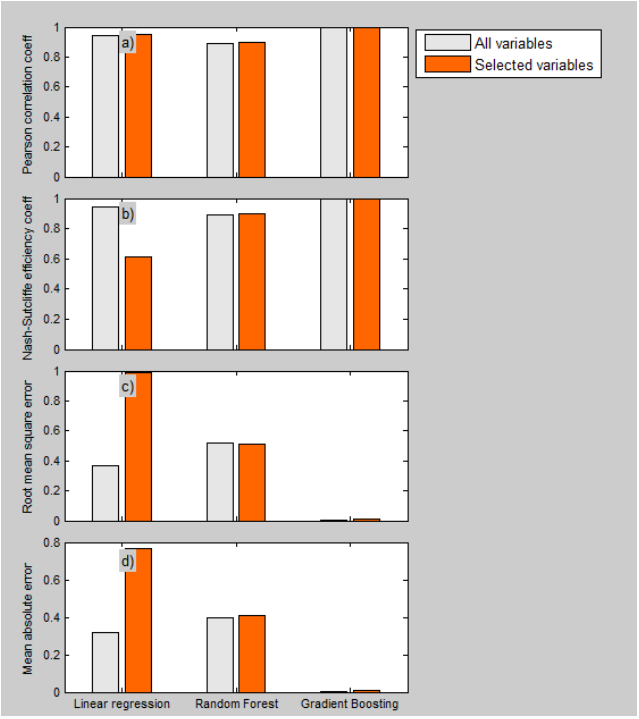

Figure 6. *Performance of the three models in modeling dissolved oxygen concentrations during the training phase by using variables exhibiting correlations greater than + 0.1 or less than − 0.1 with dissolved oxygen.*

Figure 7 depicts the comparison of the efficiency of the employed models according to the alternative adopted. As shown in Figure 7a-d, in terms of all accuracy criteria (RP, NSE RMSE and MAE), the alternative 2 emerged as the most reliable approach in the testing phase to improve the employed model in modeling (predicting) dissolved oxygen concentrations. It is important to remember that alternative 2 refers to when variables exhibiting statistically significant correlations with dissolved oxygen were used for predicting its concentration. From the results of the testing, it can be noted that, RF algorithm presents the largest RP (0.62) followed by the MLR (0.52) and GB (0.47). In addition, RF achieved the largest NSE (0.617), followed by MLR (0.523) and GB (0.465). Thus, except GB algorithm, MLR and RF algorithms show a satisfying correlation. Based on RMSE values, RF algorithm achieved the smallest RMSE (1.014) followed by MLR (1.130) and GB (1.197). Moreover, the smallest MAE (0.853) was obtained for the RF followed by MLR (0.975) and GB (0.853). By comparing these values it can be noted RF and MLR have predicted dissolved oxygen concentrations with a tolerable level of error. In other words, these values imply that among the employed models RF achieved the superior ability to minimize absolute errors and to capture the dissolved oxygen concentrations variation in testing phase. The RF presented the highest quality testing followed by the MLR and GB algorithm. Thus, the RF emerged as the most reliable model in the testing phase. Therefore, RF algorithm is feasible in dissolved oxygen concentrations prediction.

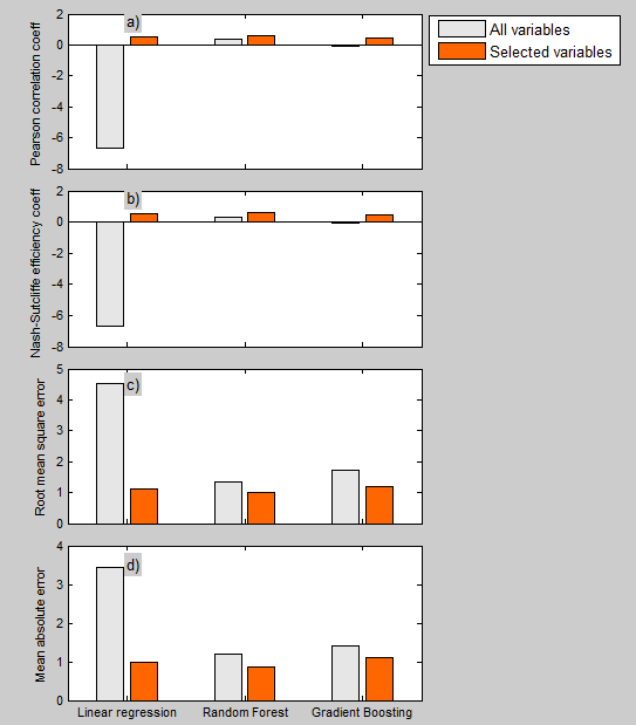

Figure 7. *Performance of the three models in modeling dissolved oxygen concentrations during the training phase by using variables exhibiting statistically significant correlations with dissolved oxygen.*

By comparing Figure 6a-d (training phase) and Figure 7a-d (testing phase), it can be noted that the evaluation results of the training phase are significantly better than those of the testing phase for all of the employed models (algorithms), especially for GB algorithm. This finding means GB model (algorithm) is prone to overfitting.

Our finding related to the superiority of random forest over multiple linear regression and gradient boosting for dissolved oxygen concentrations modeling are consistent with the results reported in other rivers of the world by [43-47].

4. Conclusions

Accurate dissolved oxygen concentration estimation is of significant importance, because it is one of the significant indicators for quality of river water monitoring. In this study, based on two alternatives, the random forest, gradient boosting, and multiple linear regression algorithm were employed to model and estimate dissolved oxygen concentrations of the Tighen River in the Republic of Guinea. Based on the correlation between dissolved oxygen concentrations and others microbiological and physicochemical parameters the two alternatives adopted are defined as: Alternative 1 referred to when microbiological and physicochemical parameters exhibiting correlations greater than + 0.1 or less than − 0.1 with dissolved oxygen are used for modeling its concentration, while alternative 2 referred to when variables exhibiting statistically significant correlations with dissolved oxygen are used. The efficiency criteria used to identify the appropriate alternative and algorithm to model and estimate the dissolved oxygen included Nash-Sutcliffe efficiency coefficient (NSE), mean absolute error (MAE), Pearson correlation coefficient (RP), and root mean square error (RMSE), which were calculated for both the training and testing datasets.

This preliminary work arrives at the following main conclusions:

The correlation matrix analysis highlights that temperature is the most influential factor in estimating dissolved oxygen concentrations. Furthermore, TDS, K, NO2 and PO4 also have substantial impacts in descending order in determining dissolved oxygen concentrations. Therefore, Temperature, TDS, K, NO2, and PO4 emerge as the main predictors to model or predict dissolved oxygen in the study area.

Among the tested alternatives, alternative 2 is significantly better than the alternative 1.

Among the employed algorithms, in the testing phase and under alternative 2, the Random Forest algorithm exhibits superior performance with the highest goodness of fit and the lowest errors in modeling and estimating dissolved oxygen concentrations than gradient boosting and multiple linear regression. Therefore, Random Forest is the is best-suited algorithm, it performs strong capability in modeling dissolved oxygen concentrations. Furthermore, the order of performance among the three algorithms is (Gradient Boosting) < (multiple linear regression) < (Random Forest).

The weakness of the implemented algorithms can be explained by the length of the data. The limitations of this preliminary work on modeling dissolved oxygen concentration are: (i) that it was not based on long-term data (spanning decades) and (ii) that it did not focus on all the streams and rivers of Guinea, even though Guinea has the most streams and rivers of any country in West Africa. This is because such long-term data do not exist and are not available for most of the country's streams and rivers.

Abbreviations

MLR	Multiple Linear Regression
RF	Random Forest
GB	Gradient Boosting
NSE	Nash-Sutcliffe Efficiency
MAE	Mean Absolute Error
RP	Pearson Correlation
RMSE	Root Mean Square Error
GA-XGCBXT	Genetic Algorithm (GA) with the High-performance Gradient Boosting
DO	Dissolved Oxygen
TDS	Total Dissolved Solid
$SatO_2$	Oxygen Saturation
NO_3	Nitrates
NO_2	Nitrites
PO_4	Phosphate
K	Potassium
Mn	Manganese

Acknowledgments

The authors thank the referees for their valuable suggestions and helpful to improve this work.

Author Contributions

Abdoulaye Missira Bangoura: Data curation, Formal Analysis, Methodology, Writing – original draft

Noukpo Medard Agbazo: Conceptualization, Data curation, Investigation, Software, Writing – original draft, Writing – review & editing

Saa Moussa Kamano: Formal Analysis, Investigation

Mafory Bangoura: Supervision, Validation

Kande Bangoura: Supervision, Validation

Conflicts of Interest

The authors declare that there is no conflict of interest.

References

[1] Schmid, B. H., and Koskiaho, J. (2006). Artificial neural network modeling of dissolved oxygen in a wetland pond: the case study Hovi, Finland.

[2] Singh, K. P.; Basant, A.; Malik, A.; Jain, G. Artificial neural network modeling of the river water quality-A case study. Ecol. Model. 2009, 220, 888–895. https://doi.org/10.1016/j.ecolmodel.2009.01.004

[3] Rankovic, V., Radulovic, J., Radojevic, I., Ostojic, A., and Comic, A. (2010). Neural network modeling of dissolved oxygen in the Gruza reservoir, Serbia. Ecol. Model., 221, 1239-1244.

[4] Ay, M., and Kisi, O. (2012). Modeling of dissolved oxygen concentration using different neural network techniques in Foundation Creek, El Paso County, Colorado, USA. J. Environ. Eng., 138(6), 654-662. http://dx.doi.org/10.1061/(ASCE)EE.1943-7870.0000511

[5] Li, Q.; He, J.; Mu, D.; Liu, H.; Li, S. Dissolved Oxygen Modeling by a Bayesian-optimized Explainable Artificial Intelligence Approach. Appl. Sci. 2025, 15, 1471. https://doi.org/10.3390/app15031471

[6] Olyaie, E.; Abyaneh, H. Z.; Mehr, A. D. A comparative analysis among computational intelligence techniques for dissolved oxygen prediction in Delaware River. Geosci. Front. 2017, 8, 517–527.

[7] Dodig, A.; Ricci, E.; Kvascev, G.; Stojkovic, M. A novel machine learning-based framework for the water quality parameters prediction using hybrid long short-term memory and locally weighted scatterplot smoothing methods. J. Hydroinform. 2024, 26, 1059–1079.

[8] He, H.; Boehringer, T.; Schäfer, B.; Heppell, K.; Beck, C. Analyzing spatio-temporal dynamics of dissolved oxygen for the River Thames using superstatistical methods and machine learning. Sci. Rep. 2024, 14, 21288.

[9] Macêdo, B. d. S.; Lima, L.; Fonseca, D. L.; Boratto, T. H. A.; Saporetti, C. M.; Fetoshi, O.; Hajrizi, E.; Bytyçi, P.; Aires, U. R. V.; Yonaba, R.; et al. Evolutionary-Assisted Data Driven Approach for Dissolved Oxygen Modeling: A Case Study in Kosovo. Earth2025, 6, 81. https://doi.org/10.3390/earth6030081

[10] Zhao, Y.; Chen, M. Prediction of river dissolved oxygen (DO) based on multi-source data and various machine learning coupling models. PLoS ONE 2025, 20, e0319256.

[11] Kisi, O., Ozkan, C., and Akay, B. (2012). Modeling Discharge-Sediment Relationship Using Neural Networks with Artificial Bee colony Algorithm. J. Hydrol. 428-429, 94–103. https://doi.org/10.1016/j.jhydrol.2012.01.026

[12] Chen, W.-B.; Liu, W.-C. Artificial neural network modeling of dissolved oxygen in reservoir. Environ. Monit. Assess. 2014, 186, 1203–1217.

[13] He, Z., Wen, X., Liu, H., and Du, J. (2014). A Comparative Study of Artificial Neural Network, Adaptive Neuro Fuzzy Inference System and Support Vector Machine for Forecasting River Flow in the Semiarid Mountain Region. J. Hydrol. 509, 379–386. https://doi.org/10.1016/j.jhydrol.2013.11.054

[14] Zhang, Y., Fitch, P., Vilas, M. P., and Thorburn, P. J. (2019). Applying MultiLayer Artificial Neural Network and Mutual Information to the Prediction of Trends in Dissolved Oxygen. Front. Environ. Sci. 7, 46. https://doi.org/10.3389/fenvs.2019.00046

[15] Chen, Lh., Zhang, Xy. (2009). Application of Artificial Neural Networks to Classify Water Quality of the Yellow River. In: Cao, By., Zhang, Cy., Li, Tf. (eds) Fuzzy Information and Engineering. Advances in Soft Computing, vol 54. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-88914-4_3

[16] Ahmed, A. M. Prediction of dissolved oxygen in Surma River by biochemical oxygen demand and chemical oxygen demand using the artificial neural networks (ANNs). J. King Saud. Univ.-Eng. Sci. 2017, 29, 151–158.

[17] Selim, A.; Shuvo, S. N. A.; Islam, M.; Moniruzzaman, M.; Shah, S.; Ohiduzzaman, M. Predictive models for dissolved oxygen in an urban lake by regression analysis and artificial neural network. Total Environ. Res. Themes 2023, 7, 100066.

[18] Areerachakul, S.; Junsawang, P.; Pomsathit, A. Prediction of dissolved oxygen using artificial neural network. Int. Conf. Comput. Commun. Manag. 2011, 5, 524–528.

[19] Zhu, S.; Heddam, S. Prediction of dissolved oxygen in urban rivers at the Three Gorges Reservoir, China: Extreme learning machines (ELM) versus artificial neural network (ANN). Water Qual. Res. J. 2020, 55, 106–118.

[20] Gorgan-Mohammadi, F.; Rajaee, T.; Zounemat-Kermani, M. Decision tree models in predicting water quality parameters of dissolved oxygen and phosphorus in lake water. Sustain. Water Resour. Manag. 2023, 9, 1.

[21] Krivoguz, D.; Semenova, A.; Malko, S. Performance of machine learning algorithms in predicting dissolved oxygen concentration. In Proceedings of the International Scientific Conference on Agricultural Machinery Industry “Interagromash”, Rostov-on-Don, Russia, 25–27 May 2022; Springer: Cham, Switzerland, 2022; pp. 1137–1144.

[22] Moon, J.; Lee, J.; Lee, S.; Yoon, H. Urban River Dissolved Oxygen Prediction Model Using Machine Learning. Water 2022, 14, 1899.

[23] Arora, S.; Keshari, A. K. Dissolved oxygen modelling of the Yamuna River using different ANFIS models. Water Sci. Technol. 2021, 84, 3359–3371.

[24] Khan, P. W.; Byun, Y. C. Optimized Dissolved Oxygen Prediction Using Genetic Algorithm and Bagging Ensemble Learning for Smart Fish Farm. IEEE Sens. J. 2023, 23, 15153–15164.

[25] Kozhiparamban, R. A. H.; Swetha, P.; Harigovindan, V. Accurate Dissolved Oxygen Prediction for Aquaculture Using Stacked Ensemble Machine Learning Model. Natl. Acad. Sci. Lett. 2023, 46, 203–207.

[26] Guo, P.; Liu, H.; Liu, S.; Xu, L. Numeric Prediction of Dissolved Oxygen Status Through Two-Stage Training for Classification Driven Regression. In Proceedings of the 2019 International Conference on Machine Learning and Cybernetics (ICMLC), Kobe, Japan, 7–10 July 2019; IEEE Computer Society: Washington, DC, USA, 2019.

[27] Altunkaynak, A., Özger, M., and Çakmakcı, M. (2005). Fuzzy Logic Modeling of the Dissolved Oxygen Fluctuations in Golden Horn. Ecol. Model. 189, 436–446. https://doi.org/10.1016/j.ecolmodel.2005.03.007

[28] Giusti, E., and Marsili-Libelli, S. (2009). Spatio-Temporal Dissolved Oxygen Dynamics in the Orbetello Lagoon by Fuzzy Pattern Recognition. Ecol. Model. 220, 2415–2426. https://doi.org/10.1016/j.ecolmodel.2009.06.007

[29] Heddam, S. (2014). Modeling Hourly Dissolved Oxygen Concentration (Do) Using Two Different Adaptive Neuro-Fuzzy Inference Systems (ANFIS): A Comparative Study. Environ. Monit. Assess. 186, 597–619. https://doi.org/10.1007/s10661-013-3402-1

[30] Tarmizi, A., Ahmed, A. N., and El-Shafie, A. (2014). Dissolved Oxygen Prediction Using Support Vector Machine in Terengganu River Middle-East. J. Sci. Res. 21 (11), 2182–2188. https://doi.org/10.5829/idosi.mejsr.2014.21.11.21844

[31] Yu, H., Chen, Y., Hassan, S., and Li, D. (2016). Dissolved Oxygen Content Prediction in Crab Culture Using a Hybrid Intelligent Method. Sci. Rep. 6, 27292. https://doi.org/10.1038/srep27292

[32] Ji, X., Shang, X., Dahlgren, R. A., and Zhang, M. (2017). Prediction of Dissolved Oxygen Concentration in Hypoxic River Systems Using Support Vector Machine: A Case Study of Wen-Rui Tang River, china. Environ. Sci. Pollut. Res. 24, 16062–16076. https://doi.org/10.1007/s11356-017-9243-7

[33] Tsakiri, K., Marsellos, A., & Kapetanakis, S. (2018). Artificial Neural Network and Multiple Linear Regression for Flood Prediction in Mohawk River, New York. Water, 10(9), 1158. https://doi.org/10.3390/w10091158

[34] Breiman, L. Random forests. Mach. Learn. 2001, 45, 5–32.

[35] Prajwala, T. R. A comparative study on decision tree and random forest using R tool. Int. J. Adv. Res. Comput. Commun. Eng. 4, 196–199 (2015);

[36] Chen, Y. T. (2021). Analytical comparison of random forest and gradient boosting decision trees for integrated learning algorithms. J. Comput. Knowl. Technol. 17 (15), 32–34. https://doi.org/10.14004/j.cnki.ckt.2021.1441

[37] Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. Ann. Stat. 2001, 29.

[38] Bentéjac, C.; Csörgo, A.; Martínez-Muñoz, G. A Comparative Analysis of XGBoost. arXiv 2019, arXiv: abs/1911.01914.

[39] Jerome, H. F. (2001). Greedy function approximation: a gradient boosting machine. J. Ann. Statistics 11 (10), 877–884.

[40] Otchere, D. A., Ganat, T. O. A., Ojero, J. O., Tackie-Otoo, B. N. & Taki, M. Y. Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterization predictions. J. Pet. Sci. Eng. 208, 109244 (2022).

[41] Chicco, D., Warrens, M. J. & Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. Peerj Comput. Sci. 7, e623 (2021).

[42] Althoff, D. & Rodrigues, L. N. Goodness-of-fit criteria for hydrological models: model calibration and performance assessment. J. Hydrol. 600, 126674 (2021).

[43] Garabaghi, F. H., Benzer, S., & Benzer, R. (2023). Modeling dissolved oxygen concentration using machine learning techniques with dimensionality reduction approach. Environmental Monitoring and Assessment, 195(7), 879. https://doi.org/10.1007/s10661-023-11492-3

[44] Li, S., Qasem, S. N., Band, S. S., Ameri, R., Pai, H. T., & Mehdizadeh, S. (2024). Explainable machine learning models for estimating daily dissolved oxygen concentration of the Tualatin River. Engineering Applications of Computational Fluid Mechanics, 18(1). https://doi.org/10.1080/19942060.2024.2304094

[45] Krivoguz, D., Semenova, A., & Malko, S. (2022, May). Performance of machine learning algorithms in predicting dissolved oxygen concentration. In A. Beskopylny, M. Shamtsyan, & V. Artiukh (Eds.), International scientific conference on agricultural machinery industry "interagromash" (pp. 1137–1144). Springer International Publishing.

[46] Ahmed, M. H. Prediction of the Concentration of Dissolved Oxygen in Running Water by Employing A Random Forest Machine Learning Technique. J. Hydrol. 2021.

[47] Ay, M.; Ki¸si, Ö. Estimation of dissolved oxygen by using neural networks and neuro fuzzy computing techniques. J. Civil Eng. 2017, 21, 1631–1639.