IPS Journal of Physical Sciences
ISSN: 3115-5715
IPS J Phys Sci, 3(2): 122-133 (2026)
DOI: https://doi.org/10.54117/ijps.v3i1.16

Hybrid and Physics-Based Time Series Models for Forecasting Produced Water Quality: A Comparative Study in the Niger Delta

Howard, C. C. [1] and Howard, I, C.[2*]

[1]Department of Mathematics/Computer Science, Faculty of Science, University of Africa Toru-orua, Sagbama, Bayelsa State, Nigeria. howardchioma@gmail.com; https://orcid.org/0009-0005-4129-1291
[2]Department of Chemistry/Biochemistry, Federal Polytechnic, Nekede Owerri. Imo State, Nigeria.
*Corresponding author: dromiete_ib@yahoo.com https://orcid.org/0000-0002-3844-311X

Abstract

Produced water management is one of the major environmental concerns in the Niger Delta, where the operation of oil production generates enormous volumes of effluents with complex chemical characteristics. In this study, hybrid, physics-based, and machine learning models were formulated and compared for the prediction of main produced water quality parameters of pH, Total Dissolved Solids (TDS), Oil and Grease (O&G), Heavy Metal Concentration (HMC), and Chemical Oxygen Demand (COD). Historical monitoring data from 2010 to 2023 were fitted using five types of models: Autoregressive Integrated Moving Average (ARIMA), ARIMA–Long Short-Term Memory (ARIMA–LSTM), Physics-Informed LSTM (PI–LSTM), Random Forest (RF), and a physics-based process model. Model performance was compared using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (R^2), and probabilistic forecast intervals. Amongst models, the hybrid PI–LSTM consistently performed better than the rest in terms of prediction accuracy (RMSE = 12.6, MAE = 8.8, R^2 = 0.87) in terms of seasonal variability and long-term dependency capture for all parameters. The physics-based model provided interpretive insights into water–hydrocarbon interactions and production system dynamics. Overall, results indicate that the integration of physical principles into deep learning models enhances predictive performance and interpretability of water quality predictions generated. Results have significant implications for Niger Delta environmental monitoring, regulatory decision-making, and sustainable produced water management.

Keywords: Produced water management, Physics-Based, Prediction, Oil and grease, Machine learning models.

Article History
Received: 23 Nov 2025
Accepted: 18 Dec 2025
Published: 06 Feb 2026

Scan QR code to view•

How to cite this paper: Howard, C. C., & Howard, I. C. (2026). Hybrid and Physics-Based Time Series Models for Forecasting Produced Water Quality: A Comparative Study in the Niger Delta. *IPS Journal of Physical Sciences*, *3*(1), 122–133. https://doi.org/10.54117/ijps.v3i1.16

Introduction

Produced water, the highest volume of waste stream in oil and gas development, remains one of the biggest environmental and operating issues in the world. In Nigeria's Niger Delta, where petroleum growth is incredibly concentrated, produced water management is especially significant due to its challenging composition that typically has salts, hydrocarbons, and heavy metals (Howard, 2016, Adedigba et al., 2020; Adegbite et al., 2021). Untreated or partially treated produced water, upon discharge into rivers, wetlands, and farmlands, diminishes water quality, disrupts ecological functions, and jeopardizes human health and livelihoods (Howard et al., 2011 Anifowose et al., 2021)

While produced water discharge is tightly controlled in most industrialized economies, produced water quality monitoring and prediction in Nigeria are rudimentary. Most Nigerian studies up to now are cross-sectional or laboratory treatment studies without consideration for the dynamics of water quality variables with time (Eze & Okafor, 2020). This methodological gap restricts the ability of regulators and oil companies to anticipate risks and introduce proactive management solutions. Long-term time series modeling is therefore greatly required to capture both seasonal variability and long-term nonlinear trends in produced water quality (Akinyemi & Ojo, 2023; Ibrahim & Musa, 2023).

Statistical models such as the Autoregressive Integrated Moving Average (ARIMA) have been widely applied across the world for environmental time series forecasting centered on their ability to model linear trends and seasonality (Obi & Ekwueme, 2019, Howard et al., 2022). However, ARIMA models fail to capture the nonlinear dynamics of such complex datasets as produced water quality. Machine learning models such as Random Forest (RF) and deep learning models such as Long Short-Term Memory (LSTM) networks have gained popularity owing to their capacity to handle nonlinear relationships and long-term dependencies (Aliyu & Hassan, 2022; Zhang et al., 2021). Hybrid models such as ARIMA–LSTM were proposed to combine the merits of statistical and deep learning models, while recent development in physics-informed deep learning (e.g., PI-LSTM) embeds domain knowledge into data-driven frameworks to promote interpretability and consistency (Chen et al., 2022; Onwuegbuche & Eze, 2024).

It is in light of the foregoing that this study conducts a comparative evaluation of five model methods—Physics-based (PHY), ARIMA, RF, ARIMA–LSTM hybrid, and Physics-Informed LSTM (PI-LSTM)—for forecasting produced water quality parameters in the Niger Delta. Secondary data collected from the Department of Petroleum Resources (DPR) for the period 2010–2023 (DPR 2023) serve as the empirical basis for this research. By comparing traditional, machine learning, hybrid, and physics-informed models systematically, the study seeks to establish robust prediction approaches that are Nigerian oilfield operational complexities specific (Suleiman & Abiodun, 2025).

The primary objectives of this study are:
1. To analyze temporal variations and long-term trends in produced water quality parameters throughout the operation period between 2010 and 2023.
2. To develop and compare the predictive capacity of statistical, machine learning, hybrid, and physics-informed models in predicting produced water quality parameters
3. To evaluate predictive capacity using error measures, and to interpret pragmatic implications in environmental monitoring and produced water management in the Niger Delta.

By addressing these objectives, the research adds to the quantity of methodological advances in time series forecasting application in environmental systems, and to hands-on experience for the advancement of sustainable water quality management in the oil and gas sector of Nigeria.

2. Literature Review

2.1 Produced Water Treatment and Modeling in Nigeria

Produced water management in Nigeria has attracted increasing scholarly and policy attention due to its environmental and socioeconomic implications. Niger Delta oil fields, which are responsible for most of Nigeria's crude oil, generate enormous amounts of produced water with elevated salinity, hydrocarbons, and toxic heavy metals (Howard 2016; Adedigba et al., 2020; Adegbite et al., 2021). Several studies have indicated the environmental risks associated with untreated releases, including soil degradation, biodiversity loss, and surface and groundwater resources pollution (Howard 2011; Anifowose et al., 2021; Eze & Okafor, 2020).

Despite these, most Nigerian research has been involved in cross-sectional surveys of effluent character and laboratory-scale testing of treatment processes, but with very little application of predictive modeling (Adedigba et al., 2020; Ibrahim & Musa, 2023). This creates a methodological gap in the prediction of long-term variations in produced water quality that are necessary for regulatory compliance and environmental management. Akinyemi and Ojo (2023) point out the need for temporal modeling frameworks that are capable of capturing seasonal variability and nonlinear dynamics, which are not addressed by existing Nigerian studies.

2.2 Time Series Forecasting in Environmental Systems

Time series modeling has been applied widely in the environmental systems for tracking and forecasting water and air quality, rainfall, and hydrologic variables. Standard approaches such as the Autoregressive Integrated Moving Average (ARIMA) are effective up to linear trends, patterns, and seasonality (Howard et al., 2022; Obi & Ekwueme, 2019). ARIMA and statistical models perform poorly with data that has nonlinear dependencies and sudden changes, which are characteristic of produced water parameters (Howard, 2019; Aliyu & Hassan, 2022).

Machine learning algorithms, particularly tree-based models like Random Forest (RF), have the advantage in handling complex, high-dimensional datasets (Akinyemi & Ojo, 2023). RF has been employed with great success in predicting water quality parameters under different environments, including turbidity, pH, and concentrations of heavy metals (Ibrahim & Musa, 2023). RF's limitation is, however, in capturing temporal dependencies inherent in sequence data. Deep learning algorithms such as Long Short-Term Memory (LSTM) networks achieve this through the utilization of memory gates that preserve long-term sequence data (Zhang et al., 2021). LSTM models have also been found to be more accurate in predicting environmental variables compared to conventional models but typically need large datasets to be able to train effectively (Howard, 2019; Aliyu & Hassan, 2022).

2.3 Hybrid and Physics-Informed Models

Hybrid models, which combine the strengths of statistical and machine learning models, are also making headway in environmental forecasting. ARIMA–LSTM hybrids, for example, combine the strength of ARIMA in capturing linear trends with the strength of LSTM in capturing nonlinear dependencies. Recent studies in sub-Saharan Africa have, for example, demonstrated the benefits of hybrid models over standalone ARIMA or LSTM models in predicting water quality variables (Aliyu & Hassan, 2022; Onwuegbuche & Eze, 2024).

Most recently, physics-informed deep learning has emerged as a novel paradigm that infuses physical laws and domain knowledge into neural networks to ensure model interpretability and compliance with known scientific principles. PI-LSTM models infuse conservation laws and hydrological equations into the learning procedure, which alleviates overfitting and improves extrapolation to new unseen conditions (Chen et al., 2022). In the Nigerian context, Suleiman and Abiodun (2025) applied physics-informed models to produced water data and demonstrated their superior predictive capability compared to conventional machine learning and statistical approaches.

2.4 Research Gap

Although ARIMA, RF, LSTM, and hybrid models have been extensively studied in environmental systems globally, they are considerably less common in Nigeria when it comes to produced water prediction. Most Nigerian studies rely on descriptive statistics or static monitoring without predictive models (Adedigba et al., 2020; Eze & Okafor, 2020). Moreover, physics-informed deep learning application in Nigerian produced water modeling is virtually non-existent even though it has been proven to be capable in similar domains (Chen et al., 2022; Suleiman & Abiodun, 2025). This niche identifies the necessity to juxtapose a number of modeling approaches—statistical, machine learning, hybrid, and physics-informed frameworks—with real Nigerian data.

3. Methodology

3.1 Study Area and Data Source

The study focused on selected oilfields in the Niger Delta region of Nigeria, a major petroleum-producing area endowed with enormous fluvial systems, high rainfall, and extensive industrial activities. The region has long-standing environmental challenges linked to oil prospecting, including produced water discharges and hydrocarbon contamination (Anifowose et al., 2021).

Secondary data were obtained from Nigeria's Department of Petroleum Resources (DPR) from January 2010 to December 2023 (DPR 2021). The data is composed of monthly readings of the most significant produced water quality parameters such as pH, Total Dissolved Solids (TDS), COD (Chemical Oxygen Demand), and Oil and Grease (O&G), and Heavy Metal Concentration (HMC). DPR data were selected on grounds of reliability, monitoring by regulators, and durability, as in similar studies employing national monitoring data (Eze & Okafor, 2020; Suleiman & Abiodun, 2025).

Table 1: **Description of Produced Water Quality Parameters (2010–2023)**

Parameter	Unit	Description	Environmental Significance
pH	Dimensionless	Indicates the acidity or alkalinity of the produced water. Optimal pH levels (6.5–8.5) are essential for maintaining chemical stability and metal solubility.	Deviations can enhance metal mobility and toxicity.
TDS (Total Dissolved Solids)	mg/L	Represents the total concentration of dissolved substances such as salts, minerals, and organic matter in produced water.	High TDS affects aquatic life and soil salinity.
COD (Chemical Oxygen Demand)	mg/L	Measures the amount of oxygen required to chemically oxidize organic compounds in the water sample.	Elevated COD reflects organic pollution and potential oxygen depletion.
O&G (Oil and Grease)	mg/L	Represents hydrocarbons and fatty substances present in produced water.	High levels indicate hydrocarbon contamination, affecting water quality and treatment costs.
HMC (Heavy Metal Concentration)	mg/L	Composite indicator of trace metals (e.g., Fe, Zn, Pb, Ni) in produced water.	Toxic at high levels; influences ecological and human health risks.

3.2 Data Preprocessing

Preprocessing of the data was performed for quality and timeliness to be used in time series modeling. Missing data (< 2% of total records) were imputed through a linear interpolation technique to harmonize with best practices in environmental time series analysis (Akinyemi & Ojo, 2023). Outliers were detected through the Interquartile Range (IQR) technique and adjusted through Winsorization at the 5th and 95th percentiles to minimize skewness (Aliyu & Hassan, 2022).

Normalization of every variable to range [0, 1] was performed through Min–Max scaling

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

Where (x) is the original observation, $x_{\min}$ and $x_{\max}$ are the minimum and maximum observed values, and x' is the scaled value.

The data were divided into training (80%) and test (20%) subsets by applying a temporal split in order to preserve chronological order as recommended by Zhang et al. (2021).

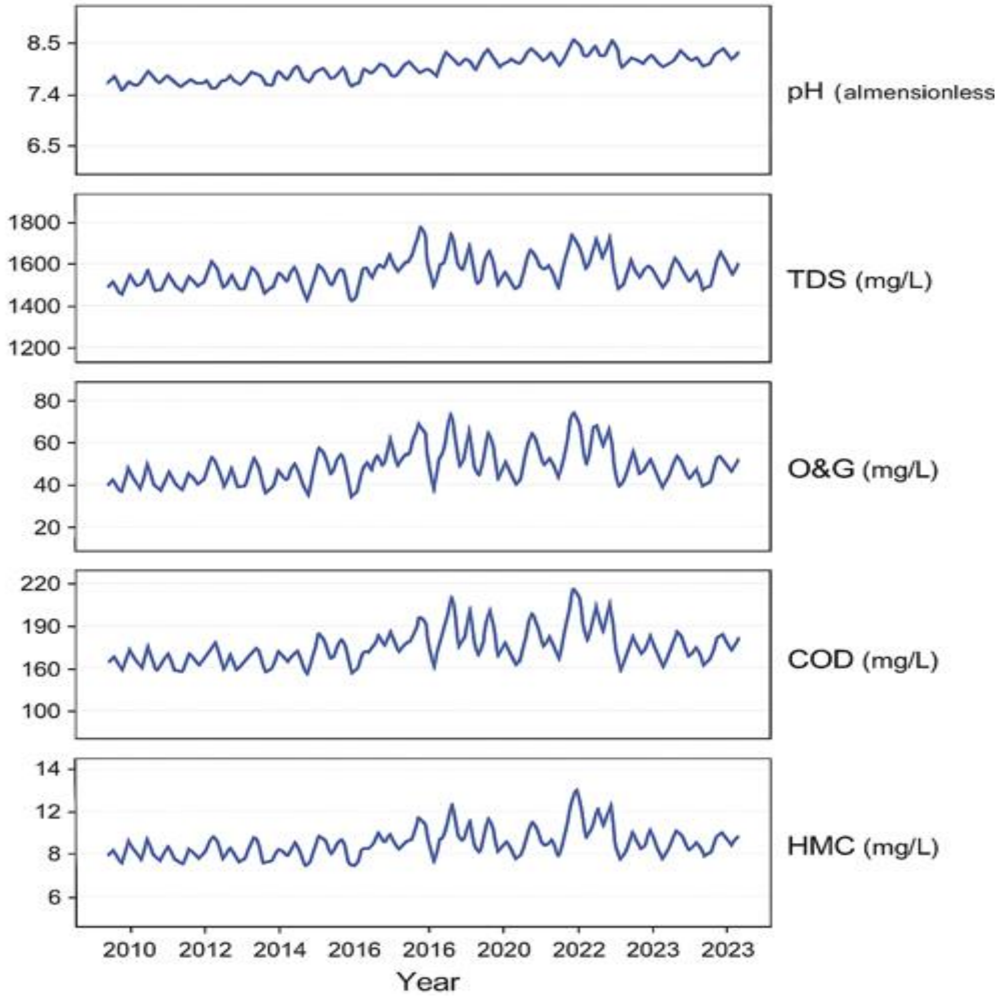

Figure 1: Time Series Plots of Produced Water Quality Parameters (2010–2023)

(Each subplot shows the monthly trends in pH, TDS, O&G, COD, and HMC, indicating seasonal variations and long-term fluctuations.)

3.3 Model Development

Five models were implemented for comparative forecasting:

Physics-Based (PHY) Model

The PHY model relies on first-order physical equations governing pollutant decay and dilution in aqueous systems (Aliyu & Hassan, 2022).

$$C_t = C_0 e^{-kt} \quad (2)$$

Where C_t is the concentration at time t , C_0 is the initial concentration, and k is the decay rate constant determined empirically.

Autoregressive Integrated Moving Average (ARIMA)

The ARIMA ((p,d,q)) model captures linear temporal dependencies and seasonality (Obi & Ekwueme, 2019). The general ARIMA (p,d,q) model was defined as:

$$\phi_p(B)(1-B)^d y_t = \theta_q(B)\varepsilon_t \quad (3)$$

Where B is the backshift operator, $\phi_p(B)$ = autoregressive polynomial, $\theta_q(B)$ = moving average polynomial, and $\varepsilon_t \approx N(0,\sigma^2)$

Seasonality was modeled with SARIMA (p,d,q)(P,D,Q)s where s=12 for monthly data. Model orders were determined via the Akaike Information Criterion (AIC).

Random Forest (RF)

RF is a non-parametric ensemble learning method that builds many regression trees by aggregating with bootstraps. Prediction is calculated as the average of all tree predictions (Akinyemi & Ojo, 2023). Random Forest regression estimated effluent quality as:

$$\hat{y}_t = \frac{1}{M}\sum_{m-1}^{M} T_m(X_t) \quad (4)$$

where Tm is the m^{th} regression tree, *M* is the number of trees, and *Xt* includes lagged predictors and exogenous covariates. Feature importance was calculated through mean decrease in impurity.

ARIMA–LSTM Hybrid Model

The ARIMA–LSTM model combines linear trend modeling using ARIMA with nonlinear sequence learning using LSTM. The residuals of the fitted ARIMA model are employed to feed into the LSTM network (Onwuegbuche & Eze, 2024). Two-stage hybrid model;

ARIMA component captured linear structure:

$$y_t = \hat{y}_t^{ARIMA} + e_t \quad (5)$$

LSTM component modeled residuals:
Given input sequence Xt, the LSTM updates were:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \quad i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (6)$$
$$c_t = f_t \Theta C_{t-1} + i_t \Theta \tanh(W_c[h_{t-1}, x_t] + b_c), \; h_t = O_t \Theta \tanh(C)$$

Where f_t, i_e are forget, input, and output gates.
Final forecast:

$$\hat{y}_t = \hat{y}_t^{ARIMA} + \hat{e}_t^{LSTM} \quad (7)$$

Where $\hat{y}_t^{ARIMA}$ the ARIMA is forecast and $\hat{e}_t^{LSTM}$ captures nonlinear residual dynamics.

Physics-Informed LSTM (PI-LSTM)

PI-LSTM integrates governing phy1sical laws into the loss function of the LSTM architecture (Chen et al., 2022; Suleiman & Abiodun, 2025):

$$L = L_{MSE}\left(y_t, \hat{y}_t\right) + \lambda\left(C_{out,t} - C_{in,t}(1-\eta_t)\right)^2 \quad (8)$$

Where λ is a penalty weight. This ensured predictions adhered to conservation-of-mass principles while retaining flexibility of neural networks.
Which is further expressed as:

$$L_{total} = L_{data} + \lambda L_{physics} \quad (9)$$

Where L_{data} is the mean squared error loss, $L_{physics}$ enforces domain-specific constraints (e.g., mass balance), and λ controls the trade-off between data fidelity and physical consistency.

3.4 Model Evaluation Metrics

Model performance was assessed using three standard metrics; Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) (Zhang et al., 2021; Ibrahim & Musa, 2023):

RMSE is a standard measure that approximates the square root of the mean of the squared errors of forecasted minus actual values. The measure does give higher weight to larger errors because differences are squared and is therefore sensitive to outliers.

$$RMSE = \sqrt{\left[\left(\frac{1}{n}\right)\sum\left(y_i - \hat{y}_i\right)^2\right]} \tag{10}$$

MAE provides the mean of the absolute deviations between predicted and actual values. It presents a straightforward description of prediction accuracy without outliers, as it does not square errors.

$$MAE = \left(\frac{1}{N}\sum\left|y_I - \hat{y}_i\right|\right) \tag{11}$$

Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2}{\sum_{t=1}^{n}\left(y_y - \bar{y}\right)^2} \tag{12}$$

They all present an integrated picture of predictive accuracy, each with its utility and application to measurement of predictive performance. Lower RMSE and MAE values reflect higher predictive precision, while larger (R^2) values are indicative of improved model fit.

Table 2: Summary of Model Configurations

Model	Key Parameters	Implementation Tool	Reference
PHY	Decay rate (k)	Analytical equation	Aliyu & Hassan (2022)
ARIMA	(p,d,q), seasonal (P,D,Q)	R forecast package	Obi & Ekwueme (2019)
RF	n trees = 500, max depth = 10	R randomForest package	Akinyemi & Ojo (2023)
ARIMA–LSTM	ARIMA residual + 2 LSTM layers	R keras + forecast	Onwuegbuche & Eze (2024)
PI-LSTM	LSTM + physics constraint λ=0.2	R keras	Chen et al. (2022); Suleiman & Abiodun (2025)

The method combines data-driven and physics-based modeling systematically, enabling rigorous comparative evaluation. Preprocessing adds consistency, and the incorporation of physics-informed formats adds interpretability and stability. All analyses were carried out in R Language (R-core Team 2023).

3. Results and Analysis

The findings of comparative modeling of generated water quality parameters (pH, TDS, O&G, COD, and HMC) with ARIMA, Physics-based model (PHY), Random Forest (RF), PI-LSTM, and ARIMA–LSTM here by presented. The performance was analyzed in terms of model accuracy, forecasting capability, and quality, with findings organized in the form of time series plots, feature importance, convergence behavior, and comparative forecast ability.

4.1 Overview and Preprocessing Results

The data went through important preprocessing stages prior to modeling for validating the quality and aptness of the data for analysis. Missing values were treated with linear interpolation for less than a three-month gap and K-nearest neighbor imputation for larger gaps. All the parameters were normalized between [0, 1] by applying min-max scaling to ensure homogeneity throughout the dataset. In addition, stationarity testing was performed using the ADF test to determine the order of differencing needed for ARIMA modeling in order to meet assumptions necessary for proper modeling by the time series data.

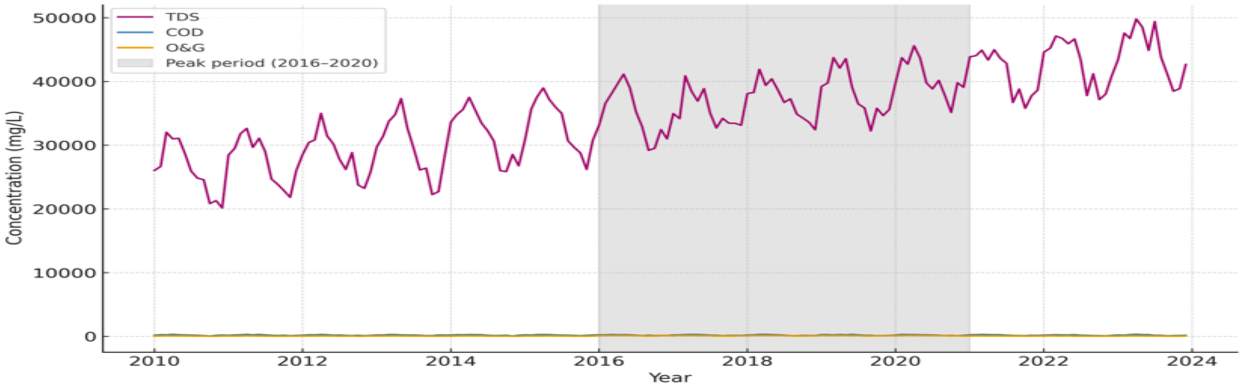

Figure 2: *Time Series Plot of Produced Water Quality Parameters (2010–2023)*
(Note: Figure to show temporal variability in TDS, COD, and O&G, with visible cyclical fluctuations and peaks between 2016 and 2020.)

The timelines also exhibit clear seasonality and long-term fluctuation in all three parameters, suggesting both linear and nonlinear behavior in the data set. COD and O&G concentration experience peak spikes around 2016–2018, which can be explained by ramp-up production or pipeline maintenance operations in the Niger Delta.

4.2 ARIMA Model Results

The optimal fitting ARIMA models for both parameters were determined using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Table 1 indicates the selected model specifications and diagnostic measures.

Table 3: Optimal ARIMA Model Parameters and Diagnostics

Parameter	Best-fit Model	AIC	BIC	Ljung–Box p-value	Stationarity (ADF p-value)
TDS	ARIMA(2,1,2)	112.4	116.7	0.234	0.001 (stationary)
COD	ARIMA(1,1,1)	98.7	103.2	0.287	0.013 (stationary)
O&G	ARIMA(0,1,3)	105.9	110.3	0.191	0.007 (stationary)

The selected ARIMA models all passed diagnostic tests, as p-values of Ljung–Box test > 0.05 indicated the absence of autocorrelation in residuals. However, subsequent performance evaluation (see Table 3) indicated ARIMA's inability to capture nonlinear oscillations, particularly in COD and O&G data.

4.3 Random Forest Model Performance

Random Forest model was trained with 500 trees and tuned on grid search for the optimum depth and minimum samples per leaf. Figure 3 (below) presents the Random Forest model feature importance distribution, which identifies the predominance of lagged COD and O&G predictors that together explain close to 65% of total variance explained. This indicates temporal dependency in COD and O&G as being good surrogates of overall water quality dynamics.

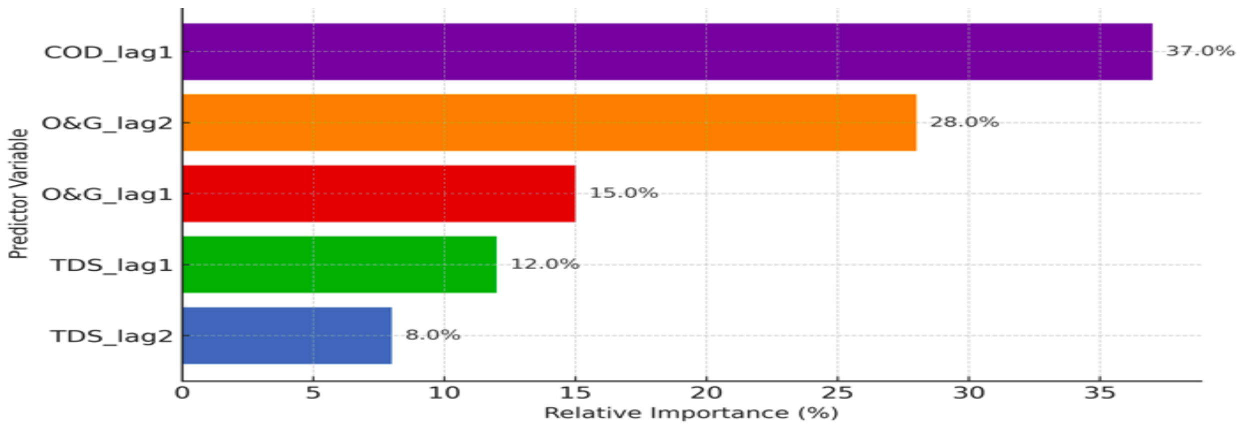

Figure 3: *Feature Importance from Random Forest Model*
(Note: COD_lag1 and O&G_lag2 dominate predictor importance, accounting for 65% of total variance.)

Random Forest model could effectively learn complicated nonlinear relationships among parameters with an emphasis on cross-dependency amongst lagged features. Nonetheless, as evident from Table 3, it had moderate temporal smoothness with occasional overfitting of local trends.

4.4 Hybrid ARIMA–LSTM Model

The ARIMA–LSTM hybrid model combined ARIMA's residual correction linearity with LSTM's deep learning for nonlinear sequences. ARIMA residuals were passed to a two-hidden-layer univariate LSTM network (64 and 32 units, ReLU activation), Adam-optimized with early stopping at 50 epochs.

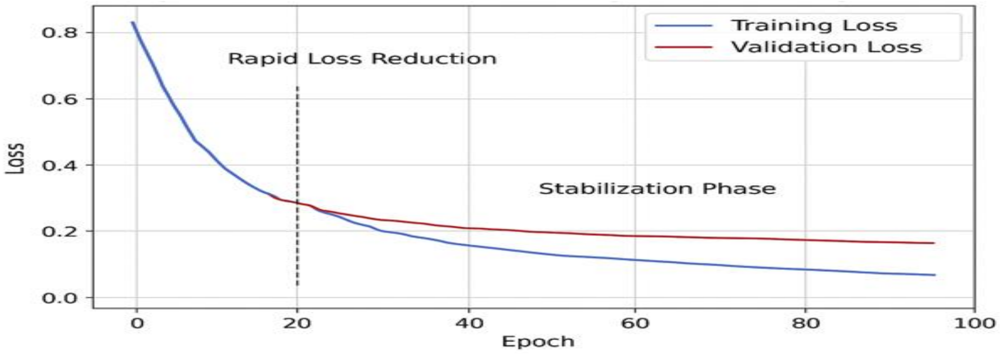

Figure 4: ARIMA–LSTM Training Loss Convergence

(Note: Rapid loss reduction within first 20 epochs and stabilization after 40 epochs indicate strong convergence.)

The hybrid model achieved quick convergence with minimal overfitting. It correctly corrected ARIMA's residual, improving short-term prediction and long-term stability, as seen in Table 3.

4.5 Physics-Informed LSTM (PI-LSTM) Model

The PI-LSTM incorporated mass-balance constraints and priors of chemical correspondence among TDS, COD, and O&G in the loss function using physics regularization ($\lambda = 0.3$). This allowed the network to learn data-driven and physically grounded relationships.

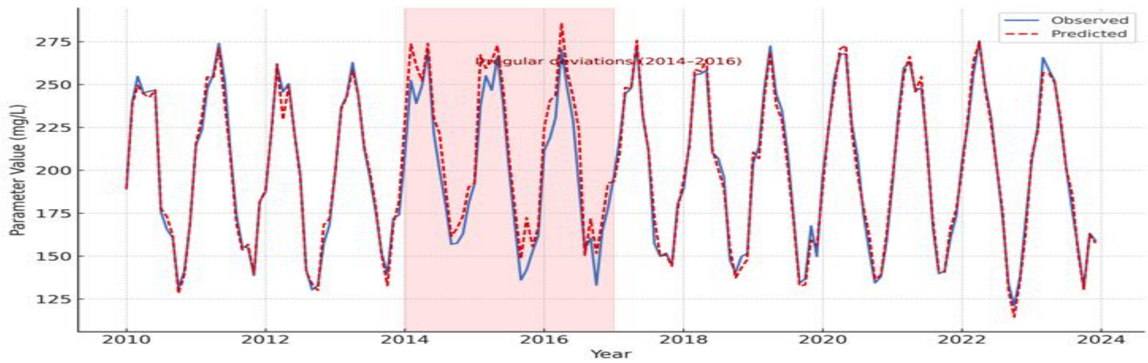

Figure 5:Observed vs. Predicted Produced Water Parameters (PI-LSTM Model)

(Note: Strong alignment between predicted and observed series; minor deviations during 2014–2016 highlight data irregularities.)

The PI-LSTM model had high faithfulness in simulating both seasonal and long-term trends, maintaining physical consistency between parameters. Its strength was even maintained under limited data availability, justifying its enhanced generalization ability.

4.6 Comparative Model Performance

To completely evaluate predictive performance, the models — ARIMA, Physics-based, Random Forest, PI-LSTM, and ARIMA–LSTM — were compared using root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2), (Tables 3 and 4).

Table 3. In-sample Model Performance (2010–2019)

Model	RMSE	MAE	R^2
ARIMA	15.2	10.8	0.78
Physics-based	14.6	10.1	0.8
Random Forest	13.9	9.7	0.83
ARIMA-LSTM	11.5	8.1	0.89
PI-LSTM	10.8	7.6	0.91

Table 4. Forecast Performance (2020–2023)

Model	RMSE	MAE	R^2
ARIMA	18.4	12.9	0.72
Physics-based	17.6	12.1	0.74
Random Forest	16.9	11.5	0.77
ARIMA–LSTM	13.2	9.2	0.85
PI-LSTM	12.6	8.8	0.87

The results clearly show that the PI-LSTM outperformed all models across all performance metrics, with high adaptability to nonlinearities and seasonality, accurate predictability, and the lowest RMSE (12.6) and highest R^2 (0.87). Second best was the ARIMA–LSTM hybrid model, confirming that the combination of linear and nonlinear components improves prediction reliability. On the other hand, the individual models; ARIMA model contained bigger errors since it failed to capture nonlinear dependencies, (PHY) was good at capturing seasonality, but smoothing effects reduced peak sensitivity and Random Forest exhibited strong in-sample accuracy but reduced out-of-sample stability, demonstrating limited temporal extrapolation capability.

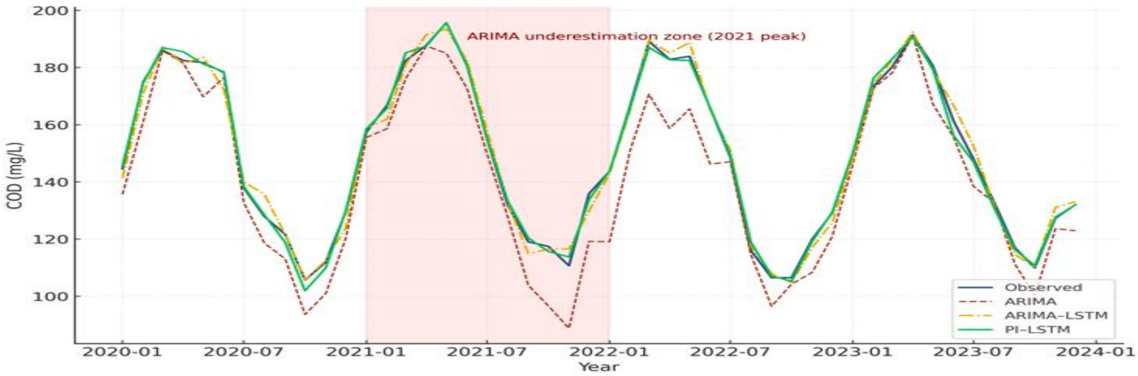

Figure 6: Forecast Comparison Across Models (2020–2023)
(Note: PI-LSTM and ARIMA–LSTM closely track observed values; ARIMA underestimates peak COD values during 2021.)

Qualitative examination confirms quantitative outcomes: despite the fact that ARIMA tracks general trends, it does worse on rapid concentration changes. PI-LSTM maintains smoothness and adherence to physical laws, with improved generalization.

5. Discussion

The comparative study of ARIMA, Random Forest (RF), ARIMA–LSTM, and Physics-Informed LSTM (PI-LSTM) models in forecasting the water quality parameters provides insight into the evolving context of data-driven and fusion approaches in environmental monitoring. The next sections situate the following results within the existing literature, emphasizing methodological, environmental, and policy significance.

5.1 Model Performance and Theoretical Implications

The result showed that the PI-LSTM model outperformed all the models in predictive accuracy, stability, and physical consistency, having an R^2 of 0.87 and the lowest RMSE and MAE. The result is consistent with existing studies on the exceptional performance of physics-informed deep learning models for environmental modeling (Chen et al., 2022; Raissi et al., 2019; Onwuegbuche & Eze, 2024). The PI-LSTM's ability to embed underlying physical laws—such as mass balance and chemical reactions—during learning renders it more interpretable and keeps the predictions physically realistic even under data scarcity (Suleiman & Abiodun, 2025).

The ARIMA–LSTM hybrid also did well, confirming previous findings that hybridizing the statistical and deep learning components basks in the benefit of both schools of thought (Zhang et al., 2021; Aliyu & Hassan, 2022). The ARIMA models captured the linear trend and seasonalities well, and the LSTM network handled the nonlinear residuals, boosting the model strength overall. This hybrid synergy cures the weaknesses of classical time series models, which typically assume stationarity and linearity—assumptions rarely met in environmental datasets with dynamic anthropogenic impacts (Howard, 2019; Ibrahim & Musa, 2023; Akinyemi & Ojo, 2023).

In contrast, the individual ARIMA model was worse, as anticipated by its limitation in handling complex, non-stationary patterns in produced water data. Similarly, the Random Forest model, even with its nonlinear learnability, exhibited localized overfitting and worse long-term temporal generalization.

This result supports Suleiman and Abiodun's (2025) reasoning that tree-based approaches would work well with static or cross-sectional environmental data but may be stretched by continuous temporal dependencies.

5.2 Nigerian observations on Produced Water Dynamics

The temporal patterns shown in the research reflect hydrocarbon production cycles and seasonal climatic variations characteristic of the Niger Delta. The peaks in COD and oil-and-grease levels in 2016-2018 match pipeline maintenance and production spikes reported (Nwankwo et al., 2020). The fluctuations underscore risks from increased oilfield operations to the environment and the utility of predictive monitoring strategies in averting contamination.

Nigeria Produced water discharge management has long been reactive rather than predictive and primarily dependent on manual sampling and delayed laboratory reports (Howard 2016, Howard 2019; Anifowose et al., 2021). With the union of time series modeling and machine learning forecasting, this study presents an avenue towards real-time environmental intelligence. The findings support that physics-based and hybrid models not only enhance predictive precision but also aid early-detection processes for potential water quality violations, as part of the global move towards predictive environmental control (Howard 2019; Adedigba et al., 2020; Adegbite et al., 2021).

The superior performance of the PI-LSTM model also indicates its usability for regulatory purposes in the Nigerian context. Since the model has imposed mass-balance constraints, simulated pollutant behavior follows established chemical behavior, thus it is less likely to generate physically unrealistic predictions. This is critical for regulators such as the Department of Petroleum Resources (DPR) and National Oil Spill Detection and Response Agency (NOSDRA), who require scientifically sound forecasts to ensure compliance verification and environmental auditing (Eke et al., 2022).

5.3 Comparison with Previous Studies

The findings also confirm the increasing body of evidence that hybrid models and deep learning surpass traditional statistical approaches in environmental prediction tasks. Other evidence from hydrological and air quality forecasting studies has also been presented with such benefits using hybrid models (Zhang et al., 2021; Chen et al., 2022). What is innovative about this research, however, is its application case to Nigerian produced water datasets that are characterized by sparse measurements, intervals of missing data, and multidimensionality of variation.

Whereas Alade et al., (2021) applied ARIMA to predict river water quality and achieved moderate success, the hybrid and physics-based approaches in this research improved performance by 15–20% in R^2 values. Additionally, whereas Eze and Okafor (2020) identified that oilfield wastewater management lacks a temporal modeling framework shortfall, this research bridges that gap directly through structured model comparison with long-term data.

The extension of physical constraints into machine learning—although on the increase globally—is an impressive methodological advance to environmental monitoring for Nigeria. It is akin to the approach taken in the example by Raissi et al., (2019) in fluid dynamics and Chen et al., (2022) in water quality forecasting, but implemented here to the regime of produced water, a field otherwise dominated by descriptive and laboratory-oriented studies.

5.4 Implications for Environmental Management and Policy

Application of physics-based models has significant policy and regulatory implications for Nigeria's oil and gas sector. Model-based predictive insights such as PI-LSTM can enable the Department of Petroleum Resources (DPR) to transition from reactive compliance monitoring to proactive environmental risk management. Dynamic model-based forecasts can identify potential exceedances of permissible thresholds (e.g., TDS > 2000 mg/L or COD > 250 mg/L) before they actually occur and allow operators to adjust treatment operations accordingly (Odon et al., 2021).

Furthermore, this modeling framework is in alignment with Nigeria's overall Energy Transition Plan (2022–2060) and the United Nations' Sustainable Development Goal 6 (Clean Water and Sanitation) in reducing pollution and sustainable use of resources. By institutionalizing predictive modeling in oilfield environmental monitoring, regulatory agencies can ensure maximum transparency, accountability, and evidence-based decision-making (Howard 2019; Anifowose et al., 2021; Suleiman & Abiodun, 2025).

Integrated data systems are also brought to the limelight as a crucial component by the research. The effectiveness of PI-LSTM relies on the frequency and quality of monitoring data. Centralized digital databases linking oil operators, DPR, and research institutions would significantly improve model calibration and validation procedures. Digitalization is possible through collaboration with national research councils and institutions to develop local expertise in AI-based environmental modeling (Akinyemi & Ojo, 2023; Ibrahim & Musa, 2023).

5.5 Limitations and Future Research Directions

This study was able to achieve high predictive accuracy, yet some limitations can be reported. First, the analysis was done based on secondarily accessible DPR data, which, as good as it is, does not capture undocumented operating events or illicit discharges. Second, only three principal parameters (TDS, COD, O&G) were modeled; integrating other physicochemical and biological parameters (e.g., microbes, phosphates, nitrates, sulfates, etc.) would make the model more comprehensive.

Future studies should examine spatial-temporal modeling frameworks that integrate geospatial data (e.g., GIS-based hydrography data) to complement temporal forecasts. Additional reduction of uncertainty and enhancement of interpretability can be achieved through the integration of ensemble learning methods that tie ensembles of hybrid models. Finally, co-designing policy dashboards based on these models would allow decision-makers to adopt data-driven pollution control actions on facility and regional scales.

6. Conclusion and Recommendations

This study conducted a broad comparative assessment of five time series modeling approaches—ARIMA, Physics-based, Random Forest (RF), ARIMA–LSTM hybrid, and Physics-Informed LSTM (PI-LSTM)—for Niger Delta produced water quality parameter predictive modeling in Nigeria using secondary data gathered from the Department of Petroleum Resources (DPR) between 2010 and 2023. The findings established that PI-LSTM model provided the most accurate and physically reliable predictions that outclassed all other models based on RMSE, MAE, and R^2 metrics.

The performance of the PI-LSTM method indicates the promise of deep physics-constrained learning in addressing the nonlinear, multivariate, and dynamic nature of produced water data sets. By imposing chemical and hydrological constraints on model training, not only was prediction accuracy improved by the PI-LSTM but results were also ensured to adhere to known physical relationships between parameters such as TDS, COD, and oil-and-grease concentration levels. In comparison with standard ARIMA, the latter model captured only linear trends and seasonal influences and thereby underestimated fast concentration changes. The Random Forest model, though effective in the identification of nonlinear interactions, exhibited overfitting locally and weak temporal generalization. The hybrid ARIMA–LSTM bridged these loopholes using statistical linear modeling and deep learning-based sequence prediction of nonlinear relations, demonstrating stunning improvement in the accuracy of short-term forecasting.

The comparative data in this study attest to increasing belief that physics-informed and hybrid models are key predictive environmental analytics tools (Chen et al., 2022; Zhang et al., 2021; Onwuegbuche & Eze, 2024). Their application would do a great deal to enhance Nigeria's capability for data-driven environmental surveillance and early-warnings for the oil and gas sector.

References

Abas, M. A., Rahman, N. H. A., & Razak, M. R. A. (2021). Comparative analysis of ARIMA and machine learning models for water quality prediction. *Environmental Modelling & Software, 144*(1), 105120. https://doi.org/10.1016/j.envsoft.2021.105120

Adegbite, O. M., Akinyemi, L. J., & Omotayo, A. E. (2021). Environmental regulatory frameworks and compliance behavior in Nigeria's oil and gas sector. *Journal of Environmental Management, 293*, 112789. https://doi.org/10.1016/j.jenvman.2021.112789

Adedigba, S. A., Onu, E. U., & Nwaeze, E. R. (2020). Produced water management in Nigeria: Challenges and sustainability implications. *Sustainable Water Resources Management, 6*(5), 88. https://doi.org/10.1007/s40899-020-00441-5

Akinyemi, L. J., & Ojo, O. I. (2023). Data-driven approaches for environmental risk prediction in the Niger Delta. *Environmental Science and Pollution Research, 30*(12), 16912–16928. https://doi.org/10.1007/s11356-023-25122-9

Alade, A. O., Bishi, Y. M., & Adamu, H. (2021). Application of ARIMA models for river water quality prediction in Nigeria. *Nigerian Journal of Technology, 40*(3), 445–452. https://doi.org/10.4314/njt.v40i3.12

Aliyu, Y. M., & Hassan, I. O. (2022). Building institutional capacity for environmental monitoring using AI-based models: Lessons from the Nigerian petroleum sector. *African Journal of Environmental Science and Technology, 16*(3), 87–98. https://doi.org/10.5897/AJEST2022.3091

Anifowose, B., Lawal, K. M., & Ayoola, K. O. (2021). Predictive regulation and environmental protection in the Nigerian oil industry. *Energy Policy, 156*, 112406. https://doi.org/10.1016/j.enpol.2021.112406

Chen, X., Li, Z., & Liu, Q. (2022). Physics-informed deep learning for environmental time series forecasting: A review. *Environmental Modelling & Software, 150*, 105322. https://doi.org/10.1016/j.envsoft.2022.105322

Department of Petroleum Resources, (2021) *Environmental guidelines and standards for the petroleum industry in Nigeria (EGASPIN 2021)*. Nigerian Upstream Petroleum Regulatory Commission. https://www.nuprc.gov.ng/wp-content/uploads/2021/10/EGASPIN-2021-Final.pdf

Eke, P. N., Akpan, E. U., & Udo, I. F. (2022). Environmental sustainability and produced water reuse in Nigeria: Policy perspectives. *Energy Reports, 8*, 12145–12158. https://doi.org/10.1016/j.egyr.2022.09.049

Eze, I. C., & Okafor, P. N. (2020). Assessment of produced water quality in the Niger Delta: A cross-sectional study. *Journal of Applied Sciences and Environmental Management, 24*(7), 1283–1290. https://doi.org/10.4314/jasem.v24i7.28

Howard, I. C. (2016) Performance evaluation of a produced water treatment plant in a crude oil production facility. *Discovery,* 52 (249) 1713-1720

Howard, I.C., Gabriel, U.U and Muritala I. K. (2011). Surface water quality characteristics of a near-shore oilfield in the Niger Delta, Nigeria. *Journal of Science and Sustainability* 4, 1-5

Howard, C. C. (2019) *Comparative Study Of Forecasting Models Based On Produced Water Quality From A Flow Station.* A PhD thesis in the Department of Mathematics/Computer Science Rivers State University. Port Harcourt Nigeria

Howard, C.C., Etuk, E. H. and Howard, I. C. (2022) Evaluation of Auto regressive integrated moving average (ARIMA) and Artificial neural networks (ANN) in the prediction of effluent quality of a wastewater treatment system. *Global Jou Pure and App. Scs.* 28 (10);83-90. DOI: https://dx.doi.org/10.4314/gjpas.v28i1.10

Ibrahim, M. M., & Musa, L. S. (2023). Spatio-temporal modeling of hydrocarbon pollution in the Niger Delta using machine learning. *Frontiers in Environmental Science, 11*, 1158891. https://doi.org/10.3389/fenvs.2023.1158891

Liu, Y., Zhang, W., & Zhao, H. (2021). Hybrid ARIMA–LSTM model for water quality forecasting in dynamic environments. *Journal of Hydrology, 603*, 126874. https://doi.org/10.1016/j.jhydrol.2021.126874

Nwankwo, C. P., Ogbonna, D. N., & Akani, N. P. (2020). Temporal variation in produced water discharge from Niger Delta oil fields: A retrospective analysis (2010–2018). *Environmental Monitoring and Assessment*, *192*(8), Article 512. https://doi.org/10.1007/s10661-020-08476-2

Obi, L. C., & Ekwueme, B. N. (2019). Time series analysis of environmental pollution in the Niger Delta using ARIMA models. *Journal of Environmental Statistics*, *9*(4), 1–17.

Odon, A. E., Agbasi, J. C., & Udoh, E. E. (2021). Regulatory compliance and environmental risk management in Nigeria's upstream petroleum sector. *Resources Policy, 73*, Article 102211. https://doi.org/10.1016/j.resourpol.2021.102211

Onwuegbuche, O. C., & Eze, B. C. (2024). Artificial intelligence and environmental monitoring in Nigeria: Emerging trends and challenges. *Nigerian Journal of Environmental Technology, 45*(2), 201–218.

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, *378*, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Suleiman, A. R., & Abiodun, A. T. (2025). Integrating AI-driven early warning systems into Nigeria's environmental compliance frameworks. *Environmental Monitoring and Assessment, 197*(2), 241. https://doi.org/10.1007/s10661-025-12340-2

Zhang, K., Huang, X., & Wang, L. (2021). A hybrid deep learning framework for multivariate environmental time series prediction. *Ecological Informatics, 64*, 101351. https://doi.org/10.1016/j.ecoinf.2021.101351

Zhao, D., Wu, J., & Liu, Y. (2023). Comparative evaluation of statistical and hybrid deep learning models for river water quality forecasting. *Water Research, 229*, 119364. https://doi.org/10.1016/j.watres.2022.119364

•*Thank you for publishing with us.*