SCIENTIFIC Publishing Limited

**Journal of Hydrological Ecology and Water Security**
https://ojs.ukscip.com/index.php/jhews

*Article*

# Multivariate Statistical Modelling of Water Quality Interdependencies in Produced Water from a Crude Oil Flow Station

**Chioma Chinagorom Howard** [1,*] **and Ibigoni Clinton Howard** [2]

[1] Department of Mathematics/Computer Science, Faculty of Science, University of Africa, Sagbama 561101, Nigeria
[2] Department of Chemistry/Biochemistry, Federal Polytechnic, Owerri 460113, Nigeria
[*] Correspondence: chioma.howard@uat.edu.ng

**Received:** 20 July 2025; **Revised:** 15 September 2025; **Accepted:** 8 October 2025; **Published:** 7 November 2025

**Abstract:** To accurately predict the quality of wastewater generated by crude oil exploration activities, statistical models that capture multivariate interactions among co-varying parameters are required. Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Dissolved Oxygen (DO), Temperature, pH, and Conductivity of weekly produced water (260 readings over 5 years) from an oil flow station in Akwa Ibom State, Nigeria, were modelled using Pearson correlation, Principal Component Analysis (PCA) with Varimax rotation, Hierarchical Agglomerative Cluster Analysis (HACA), Multiple Linear Regression (MLR), and Vector Autoregression (VAR). Three components explained 74.3% of total parameter variance: the Organic Load Component (PC1, 36.3%), the Oxygen-Thermal Component (PC2, 23.2%), and the Ionic-pH Component (PC3, 14.8%). BOD and COD showed strong positive correlation ($r = 0.72$, $p < 0.001$), as did DO and Temperature ($r = -0.61$, $p < 0.001$). MLR models yielded $R^2$ values of 0.71 (COD predicted from BOD and DO; out-of-sample RMSE = 20.4 mg/L) and 0.58 (DO predicted from Temperature and BOD; out-of-sample RMSE = 0.51 mg/L). VAR analysis identified BOD as a Granger-precedent predictor of COD at a two-week lag ($F = 8.34$, $p = 0.003$), and Temperature as a Granger-precedent predictor of DO ($F = 11.62$, $p < 0.001$). These results reflect predictive associations in the time series and do not imply physical causation. Cluster analysis delineated three quality regimes; none met Department of Petroleum Resources (DPR) discharge limits for BOD or DO, confirming systemic non-compliance at the study site. The models generated offer a basis for rapid characterisation of effluent quality and process control at Niger Delta flow stations.

**Keywords:** Effluent Quality; Multivariate Statistical Modelling; Principal Component Analysis; Process Control; Produced Water; Vector Autoregression

## 1. Introduction

Produced water from crude oil and natural gas extraction represents the largest volume of liquid waste generated by the petroleum industry. On average, three barrels of produced water arise per barrel of oil produced globally, and this volume typically increases with reservoir age [1,2]. The Niger Delta region of Nigeria is one of the world's most prolific crude oil zones; hundreds of flow stations serve as primary sites for crude oil and formation water separation. Releases of inadequately treated water into the fragile freshwater and estuarine ecosystems of the Delta are well-documented sources of severe pollution, with extensive ecological and public health consequences [3–5].

In Nigeria, produced water discharge is regulated by the Department of Petroleum Resources (DPR) Environmental Guidelines and Standards for the Petroleum Industry in Nigeria (EGASPIN) [6] and the Federal Environmen-

tal Protection Agency (FEPA) Effluent Limitation Regulations [7]. Numerical limits are prescribed for BOD, COD, DO, Temperature, and pH; however, the majority of flow stations in the Niger Delta monitor each parameter against its individual limit without assessing parameter interactions. The predictive information encoded in the multivariate temporal behaviour of the parameter system is therefore not exploited for operational decision-making [8–10].

Water quality parameters of produced water are interconnected. DO is consumed in biochemical oxidation of organic material; temperature affects both oxygen solubility and the rate of microbial reactions; conductivity reflects formation water composition and residual chemical additives; pH governs the acid-base state of dissolved compounds [9,10]. Changes in one parameter may propagate through the system with temporal delays—a dynamic that univariate limit-checking cannot detect [11,12].

Multivariate statistical methods offer a principled framework for characterising such systems. Pearson correlation provides an initial pairwise interdependence map [13]. PCA decomposes the variance-covariance structure of the data into orthogonal components that may be associated with latent environmental processes [14–16]. MLR enables prediction of difficult-to-measure parameters from readily obtainable surrogates [17,18]. Cluster analysis groups monitoring observations into distinct quality regimes [19,20]. VAR models temporal dependence and, via Granger causality analysis, assesses whether one variable contains predictive information about another beyond its own history [21,22]. These methods have been applied collectively to petroleum-related water quality monitoring [23–27], but systematic VAR-based dynamic modelling of produced water from onshore Niger Delta flow stations remains absent from the literature.

Rationale and Study Need: The deterioration of freshwater and estuarine ecosystems in the Niger Delta due to inadequately treated produced water discharge has significant consequences for biodiversity, fisheries, and the health of communities dependent on these water bodies [3]. Yet, despite the availability of multivariate statistical tools, produced water management at most Niger Delta flow stations relies on univariate compliance monitoring—a paradigm demonstrably insufficient for capturing the dynamic interdependencies among quality parameters [4,26]. This study addresses that gap by applying five complementary statistical methods to five years of weekly monitoring data. The practical benefit is direct: the resulting models enable flow station operators to estimate difficult-to-measure parameters from simpler surrogates and to anticipate adverse quality changes before DPR limits are breached, thereby supporting pre-emptive treatment decisions and reducing regulatory non-compliance.

Rationale for Model Selection: The five statistical models were selected to address the study objectives in a complementary and progressive sequence. Pearson correlation (Model 1) provides the foundational pairwise interdependence map, establishing which parameters co-vary and the strength of their associations. PCA (Model 2) reduces the six-variable system to a smaller set of interpretable latent dimensions, identifying the major axes of variance without imposing a predictive structure. HACA (Model 3) classifies the 260 temporal observations into distinct quality regimes using an unsupervised approach—appropriate where the number of regimes is unknown a priori. MLR (Model 4) converts the PCA-identified relationships into operational predictive equations with quantified out-of-sample accuracy. VAR (Model 5) extends the analysis to the time-series domain: unlike cross-sectional methods, VAR captures lagged predictive associations and quantifies how a change in one parameter propagates to others across multiple weeks. Together, these models span cross-sectional, dimensional, classification, predictive, and temporal-dynamic objectives. The variables modelled—BOD, COD, DO, Temperature, pH, and Conductivity—were selected because they are the major parameters monitored weekly at the study site in compliance with DPR EGASPIN requirements [6] and represent the core suite for characterising produced water quality in Nigerian regulatory practice.

The central hypothesis of this study is that the six monitored water quality parameters at the study flow station constitute a dynamic, interdependent system whose temporal behaviour cannot be adequately characterised by univariate threshold monitoring alone. The scope of generalisability of the findings is limited to the study site; external validation across multiple sites is identified as a necessary direction for future research.

The study aims: (a) to characterise the pairwise correlation structure among the six produced water quality parameters; (b) to identify principal components responsible for variance in the parameter system; (c) to develop and validate MLR models for predicting key quality parameters; (d) to classify temporal quality regimes using hierarchical cluster analysis; and (e) to model predictive temporal associations among parameters using VAR and Granger causality analysis.

## 2. Materials and Methods

### 2.1. Study Area and Data Collection

The study site is a crude oil flow station at Longitude 4°34.276′ E and Latitude 8°25.557′ N on the shores of Akwa Ibom State in the Niger Delta, Nigeria. The climate is hot and humid, with average annual rainfall exceeding 2,500 mm. Produced water samples from the effluent discharge point were collected and analysed weekly by a DPR-registered environmental consultancy. $BOD_5$ was measured by the 5-day incubation method at 20 °C; COD by the closed reflux titrimetric method; DO by the Winkler azide modification; Temperature with a calibrated mercury thermometer at the point of sampling; pH with a calibrated portable meter; and Conductivity with a conductivity meter, all following Standard Methods [28]. Weekly sampling from January 2007 to December 2011 yielded 260 observations.

### 2.2. Methodological Framework

The five statistical methods were selected to address the study objectives in a complementary and progressive manner, as detailed in the Introduction. Pearson correlation provides a pairwise interdependence map. PCA reduces dimensionality and identifies latent drivers of variance. HACA classifies temporal quality regimes without prior assumptions. MLR converts PCA-identified relationships into actionable predictive equations with out-of-sample validation. VAR extends MLR to the time-series domain, quantifying lagged predictive associations.

### 2.3. Descriptive Statistics and Normality Testing

For each parameter, mean, median, standard deviation (SD), minimum, maximum, and coefficient of variation (CV) were calculated.

$$CV = \frac{SD}{\mu} \times 100\% \tag{1}$$

where $\mu$ is the sample mean and $SD$ is the sample standard deviation.

The Shapiro-Wilk W statistic [29] was used to evaluate normality.

$$W = \frac{\left(\sum a_i x_i\right)^2}{\sum \left(x_i - \overline{x}_i\right)^2} \tag{2}$$

All data were standardised to z-scores prior to PCA and cluster analysis.

$$z_{ji} = \frac{(x_i - \overline{x}_i)}{SD_j} \tag{3}$$

### 2.4. Pearson Correlation Analysis

The Pearson product-moment correlation coefficient $r$ (Equation (4)) and significance $t$-statistic (Equation (5)) were calculated for each parameter pair at α = 0.05 and α = 0.01 with $n$ − 2 degrees of freedom [13].

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\left[\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2\right]}} \tag{4}$$

where $x_i$ and $y_i$ are pairs of observations and $\overline{x}$ and $\overline{y}$ are their means.

$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} \tag{5}$$

with $n$ − 2 degrees of freedom, evaluated at the α = 0.05 and α = 0.01 levels.

### 2.5. Principal Component Analysis

PCA was performed on the standardised 260 × 6 data matrix. The Kaiser-Meyer-Olkin (KMO) measure [30] (Equation (6)) and Bartlett's test of sphericity [31] (Equation (7)) assessed factorability. Components with eigenvalues ≥ 1 (Kaiser criterion, Equation (8)) were retained, confirmed by scree plot. Varimax rotation was applied

(Equation (11)). PCA results are presented with an exploratory and descriptive interpretation; component labels reflect observed loading patterns and do not constitute definitive identification of causal environmental processes.

$$KMO = \frac{\sum\sum_{i\neq j} r_{ij}^2}{\sum\sum_{i\neq j} r_{ij}^2 + \sum\sum_{i\neq j} p_{ij}^2} \tag{6}$$

where $r_{ij}$ are the coefficients of pairwise correlation and $p_{ij}$ are the coefficients of partial correlation. KMO ≥ 0.60 substantiates factorability

$$\chi^2 = \left[\frac{(n-1)-(2p+5)}{6}\right] \times In|R| \tag{7}$$

Where, $n$ is number of samples, $p$ number of variables, $R$ the correlation matrix. In PCA, the solution for the eigenvalue equation was found as:

$$(R - \lambda I)v = 0 \tag{8}$$

Where $R$ is $p \times p$ correlation matrix, $\lambda$ is an eigenvalue, $I$ is the identity matrix and $v$ is the corresponding eigenvector. The proportion of the variance explained by component $k$ was given by:

$$PVE_k = \frac{\lambda_k}{p} \tag{9}$$

The variable $j$'s commonality is:

$$h_j^2 = \sum_k a_{jk}^2 \tag{10}$$

The Varimax rotation standard is:

$$V = \frac{\sum_k \left[n\sum_j a_{jk}^4 - \left(\sum_j a_{jk}^2\right)^2\right]}{n^2} \tag{11}$$

The component scores for each observation were calculated by :

$$F_{ik} = \sum_j a_{ik}\, z_{ij} \tag{12}$$

## 2.6. Multiple Linear Regression Models

MLR models for COD and DO were developed using the general form (Equation (13)), estimated by Ordinary Least Squares (OLS) (Equations (14) and (15)). To assess predictive utility beyond the estimation sample and guard against overfitting, a 70/30 training-testing split was implemented chronologically: the first 182 observations (2007–mid-2010) were used for model fitting and the remaining 78 observations (mid-2010–December 2011) held out for out-of-sample validation [18]. Chronological partitioning was chosen because the data are time-ordered; random assignment would introduce temporal leakage whereby future observations inform the training model, artificially inflating out-of-sample accuracy. Multicollinearity was assessed via the Variance Inflation Factor (VIF, Equation (20)); residual autocorrelation via the Durbin-Watson statistic (Equation (21)).

The general MLR model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{13}$$

The OLS estimator reduces the sum of squares of the residuals:

$$RSS = \sum_i (y_i - \hat{y}_i)^2 = (y - X\hat{\beta})^T (y - X\hat{\beta}) \tag{14}$$

Giving the coefficient vector:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{15}$$

The coefficient of determination was used to see how well the model fit:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum\left(y_i - \hat{y}_i\right)^2}{\sum\left(y_i - \overline{y}\right)^2} \tag{16}$$

and the modified $R^2$:

$$R^2_{adj} = \frac{1-\left(1-R^2\right)(n-1)}{(n-k-1)} \tag{17}$$

RMSE and MAE were used to measure how accurate the predictions were:

$$RMSE = \sqrt{\left[\frac{(y_i-\hat{y}_i)^2}{n}\right]} \tag{18}$$

$$MAE = \sqrt{\left[\frac{\sum|(y_i-\hat{y}_i)|}{n}\right]} \tag{19}$$

The variance inflation factor was used to find multicollinearity:

$$VIF_j = \frac{1}{\left(1-R^2_j\right)} \tag{20}$$

The Durbin-Watson statistic was used to measure residual autocorrelation:

$$DW = \frac{\sum_{i=2}(e_i-e_{i-1})^2}{\sum_i e_i^2} \tag{21}$$

The standardized regression coefficient (beta weight) was calculated by

$$B^*_j = B_j \times \frac{SD_j}{SD_\gamma} \tag{22}$$

## 2.7. Hierarchical Agglomerative Cluster Analysis

Ward's minimum variance linkage criterion [32] was applied to the 260 standardised weekly observations using squared Euclidean distance (Equation (23)). The optimal cluster number was selected by maximising the Calinski-Harabasz pseudo-F index [33] (Equation (25)). Cluster membership was formally tested against meteorological season (wet: June–November; dry: December–May) using a chi-square test of independence, providing statistical support for the seasonal regime interpretation.

The squared Euclidean distance between observations *i* and *j* across all *p* variables were calculated by using the following:

$$d^2_{(i,j)} = \sum_{k=1} p\left(z_{ik}-z_{jk}\right)^2 \tag{23}$$

Ward's criterion combines the two clusters that cause the least rise in the within-cluster sum of squares (WCSS) [32]:

$$WCSS = \sum_k \sum_i \in C_k \sum_{j=1}^{p}\left(z_{ij}-\bar{z}_{kj}\right)^2 \tag{24}$$

The Calinski-Harabasz pseudo-F index [33] was used to find the best number of clusters:

$$\mathrm{CH} = \frac{BCSS/(k-1)}{WCSS/(n-k)} \tag{25}$$

## 2.8. Vector Autoregression and Granger Causality

Stationarity of each time series was confirmed using the Augmented Dickey-Fuller (ADF) test [34] (Equation (26)) and the KPSS test. The optimal VAR lag order was selected by minimising AIC (Equation (29)), BIC (Equation (30)), and Hannan-Quinn criteria (Equation (31)). Model stability was verified by checking that all companion matrix eigenvalues fell within the unit circle (Equation (32)). Granger F-tests [21] (Equation (33)) assessed predictive precedence. All Granger causality results are interpreted strictly as evidence of predictive precedence in the time series; they do not imply physical or mechanistic causation.

$$\Delta ADF = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1} p\delta_i p\delta \Delta y_{t-1} + \varepsilon_t \tag{26}$$

The VAR model of order $p$ is:

$$VAR(p): \\ y_t = C + A_1 y_{t-1} + A_2 y_{t-2} + ... + A_n y_{t-n} + \varepsilon_t \quad (27)$$

Expanding for one variable $y_{it}$:

$$y_{(it)} = c_i + \sum_{k=1}^{p} \sum_{j=1}^{k} a_{ijk} y_{jt-k} + \varepsilon_{it} \quad (28)$$

The best lag $p$ was found by minimizing the Akaike Information Criterion:

$$AIC_{(p)} = In \left| \sum l \right| + 2pK^{2/n} \quad (29)$$

The Schwarz Bayesian Criterion: n

$$SBC_{(p)} = In \left| \sum l \right| + (In\, n) pK^{2/n} \quad (30)$$

The Hannan-Quinn Criterion:

$$HQC_{(p)} = In \left| \sum l \right| + 2(In\, In\, n) pK^{2/n} \quad (31)$$

The companion form of $VAR(p)$ was used to check the stability of the model:

$$Y_t = FY_{t-1} + E_t \quad (32)$$

The $F$-test was used to see if variable x caused variable $y$:

$$F = \frac{(RSS_R - RSS_U)/P}{RSS_U/(n - 2p - 1)} \quad (33)$$

Impulse response functions show how a shock of one standard deviation affects things:

$$\partial \mathrm{y}_{\mathrm{i,t}} + h/\partial \mathrm{h}_{jt} = \Phi \mathrm{h}_{[i,j]} \quad (34)$$

Forecast Error Variance Decomposition (FEVD):

$$FEVD_{ij}(h) = \frac{\sum_{h=0}^{h=1} (\Phi h \sum ej)_i^2}{\sum_{h=0}^{h=1} (\Phi h \sum \Phi TH)_{ii}} \quad (35)$$

All statistical computations were performed using R version 4.3.1 [35].

## 3. Results

### 3.1. Descriptive Statistics

**Table 1** presents descriptive statistics for all six parameters over 2007–2011 (n = 260). Mean $BOD_5$ (98.7 mg/L) exceeded the DPR limit of 30 mg/L by a factor of 3.3. Mean COD (92.3 mg/L) exceeded the 80 mg/L limit. Mean DO (2.57 mg/L) was below half the DPR minimum of 5.0 mg/L. Mean Temperature (50.0 °C) exceeded the 40 °C DPR threshold throughout the monitoring period. The pH remained within 6.5–7.6 over all 260 weeks. COD showed the greatest variability (CV = 37.7%) and Conductivity the least (CV = 5.2%). Shapiro-Wilk tests (Equation (2)) indicated approximate normality for BOD (W = 0.975, $p$ = 0.13), COD (W = 0.971, $p$ = 0.09), Temperature (W = 0.986, $p$ = 0.42), pH (W = 0.981, $p$ = 0.28), and Conductivity (W = 0.983, $p$ = 0.35), justifying the use of parametric multivariate methods. All data were standardised to z-scores (Equation (3)) prior to PCA and cluster analysis.

**Table 1.** Descriptive Statistics of Produced Water Quality Parameters (2007–2011, n = 260).

| Parameter | Mean | Median | SD | Min | Max | CV (%) | DPR Limit |
|---|---|---|---|---|---|---|---|
| $BOD_5$ (mg/L) | 98.7 | 96.3 | 31.2 | 35.0 | 187.5 | 31.6 | 30.0 |
| COD (mg/L) | 92.3 | 89.5 | 34.8 | 21.8 | 158.6 | 37.7 | 80.0 |
| DO (mg/L) | 2.57 | 2.5 | 0.71 | 1.2 | 4.5 | 27.6 | ≥5.0 |
| Temp (°C) | 50.0 | 49.9 | 1.78 | 42.5 | 53.6 | 3.6 | 40.0 |
| pH | 7.24 | 7.3 | 0.19 | 6.5 | 7.6 | 2.6 | 6.5–8.5 |
| Cond. (mS/cm) | 27.8 | 27.9 | 1.45 | 22.3 | 30.0 | 5.2 | — |

Note: SD = standard deviation; CV = coefficient of variation; DPR = Department of Petroleum Resources effluent limit for onshore discharge.

These descriptive findings have clear environmental significance. The threefold exceedance of the DPR $BOD_5$ limit (98.7 vs. 30 mg/L) indicates a chronic, severe organic load in the effluent, consistent with inadequate biodegradation treatment at the facility [5]. The suppressed DO (mean 2.57 mg/L vs. DPR minimum ≥ 5.0 mg/L) signals hypoxic conditions in the receiving water body that are incompatible with the survival of most aerobic aquatic fauna [9]. The persistently elevated Temperature (mean 50.0 °C; DPR limit 40 °C) reflects direct thermal discharge without cooling, which accelerates microbial oxygen demand and further depresses DO through reduced solubility [10]. The low coefficient of variation for Temperature (3.6%) and Conductivity (5.2%) suggests these parameters are driven primarily by stable physicochemical characteristics of the formation water, whereas the high CV for COD (37.7%) reflects temporal fluctuations in organic load likely associated with production variability and seasonal effects.

## 3.2. Pearson Correlation Analysis

**Table 2** presents the Pearson correlation matrix (Equations (4) and (5)). The strongest positive association was between $BOD_5$ and COD ($r = 0.72$, $p < 0.001$), reflecting shared dependence on organic load. A strong negative correlation between DO and Temperature ($r = -0.61$, $p < 0.001$) is consistent with temperature-driven effects on oxygen solubility [9,10]. BOD and DO were negatively correlated ($r = -0.48$, $p < 0.001$). Conductivity was positively associated with BOD ($r = 0.34$, $p < 0.001$) and COD ($r = 0.31$, $p < 0.001$). pH showed the weakest pairwise associations ($|r| \leq 0.18$), indicating relative independence from organic and thermal dynamics.

**Table 2.** Pearson Correlation Matrix of Produced Water Quality Parameters (n = 260).

| | $BOD_5$ | COD | DO | Temp | pH | Cond. |
|---|---|---|---|---|---|---|
| $BOD_5$ | 1 | 0.720*** | −0.480*** | −0.230*** | 0.12 | 0.340*** |
| COD | | 1 | −0.390*** | −0.190** | 0.08 | 0.310*** |
| DO | | | 1 | −0.610*** | −0.090 | −0.220*** |
| Temp | | | | 1 | −0.140* | 0.180** |
| pH | | | | | 1 | 0.06 |
| Cond. | | | | | | 1 |

Note: *$p < 0.05$. **$p < 0.01$. ***$p < 0.001$. Two-tailed tests. Cond. = Conductivity.

The strong $BOD_5$–COD correlation ($r = 0.72$, $p < 0.001$) is environmentally important: both parameters index organic contamination, but COD measures the total chemical oxidisable fraction while $BOD_5$ measures only the biologically degradable component over five days [10]. A BOD:COD ratio close to 0.72 (derived from the means: 98.7/92.3 ≈ 1.07) indicates that organic material in this effluent is predominantly biodegradable, suggesting that biological treatment could be effective if properly implemented [12]. The negative BOD–DO correlation ($r = -0.48$, $p < 0.001$) confirms that higher organic loads suppress dissolved oxygen through microbial consumption, consistent with the classical biochemical oxygen demand mechanism documented in produced water studies [4]. The significant Conductivity–BOD correlation ($r = 0.34$, $p < 0.001$) suggests that dissolved ionic species co-vary with organic load, possibly reflecting the concurrent leaching of inorganic salts during organic waste decomposition. The relative independence of pH from other parameters ($|r| \leq 0.18$) indicates that the buffering capacity of the formation water maintains a stable acid-base environment despite fluctuations in organic load and temperature [9]. Similar correlation structures between BOD, COD, and DO have been reported in produced water from other Niger Delta flow stations [4,10], and in industrial effluent characterisation studies in southeastern Nigeria [10].

### 3.3. Principal Component Analysis

The KMO statistic (Equation (6)) was 0.71, and Bartlett's test (Equation (7)) was significant ($\chi^2(15) = 312.4$, $p < 0.001$), confirming factorability. The Kaiser criterion (Equation (8)) retained three components, confirmed by scree plot inspection (**Figure 1**). **Tables 3** and **4** present eigenvalues and Varimax-rotated loadings.

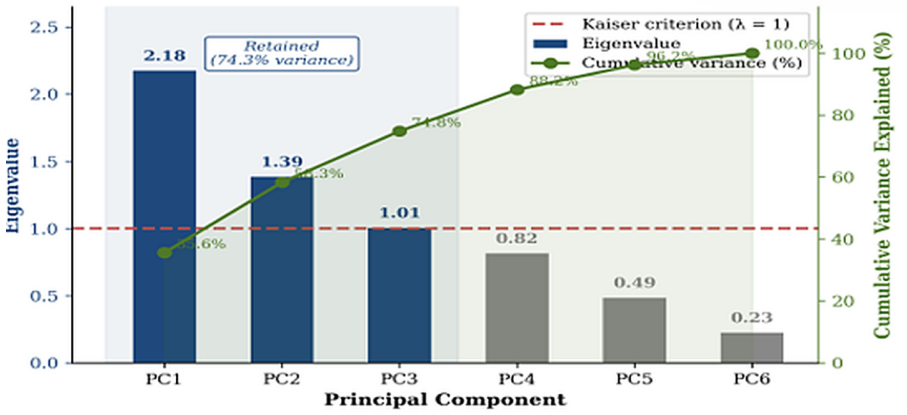


**Figure 1.** PCA scree plot showing eigenvalues for all six components.

Note: Blue bars = retained components (eigenvalue ≥ 1); grey bars = not retained. The horizontal dashed red line marks λ = 1 (Kaiser criterion); three components are retained, collectively explaining 74.3% of total variance. Green line with markers = cumulative variance explained (right axis).

**Table 3.** Eigenvalues and Variance Explained by Retained Principal Components.

| Component | Eigenvalue | % Variance Explained | Cumulative % |
|---|---|---|---|
| PC1—Organic Load Factor | 2.18 | 36.3 | 36.3 |
| PC2—Oxygen-Thermal Factor | 1.39 | 23.2 | 59.5 |
| PC3—Ionic-pH Factor | 1.01 | 14.8 | 74.3 |
| PC4 (not retained) | 0.82 | 13.7 | 88.0 |
| PC5 (not retained) | 0.49 | 8.2 | 96.2 |
| PC6 (not retained) | 0.23 | 3.8 | 100 |

**Table 4.** Varimax-Rotated Principal Component Loadings.

| Parameter | PC1 (Organic Load) | PC2 ($O_2$-Thermal) | PC3 (Ionic-pH) | Communality ($h^2$) |
|---|---|---|---|---|
| $BOD_5$ (mg/L) | 0.831 | −0.312 | 0.124 | 0.800 |
| COD (mg/L) | 0.804 | −0.267 | 0.178 | 0.750 |
| DO (mg/L) | −0.511 | 0.698 | 0.089 | 0.753 |
| Temperature (°C) | −0.188 | −0.832 | 0.213 | 0.771 |
| pH | 0.143 | −0.124 | 0.873 | 0.796 |
| Conductivity (mS/cm) | 0.412 | 0.273 | 0.654 | 0.672 |

Note: Loadings ≥ |0.60| considered practically significant (shown in bold in body of article). PC = principal component.

The three components explain 74.3% of total variance. PC1 (36.3%) shows high positive loadings on $BOD_5$ (0.831) and COD (0.804) with a moderate negative loading on DO (−0.511), suggesting a latent dimension associated with organic load variation. PC2 (23.2%) shows a dominant negative loading on Temperature (−0.832) with a positive loading on DO (0.698), consistent with the inverse relationship between temperature and oxygen solubility. PC3 (14.8%) loads on pH (0.873) and Conductivity (0.654), suggesting a distinct ionic-pH dimension. These component labels are descriptive and exploratory; they do not constitute definitive identification of causal processes.

PC1—Organic Load Factor: The co-loading of $BOD_5$ and COD on PC1 is consistent with findings from PCA studies of Niger Delta produced water and petroleum-affected water bodies [23,24], and reflects the mechanistic link between biodegradable (BOD) and total oxidisable (COD) organic fractions. The moderate negative loading of DO on PC1 (−0.511) confirms that elevated organic load is associated with oxygen depletion through biological and chemical oxidation [9]. This pattern is also reported in PCA of industrial effluents in southeastern Nigeria, where BOD and COD consistently co-load on the first principal component [10]. The communality of $BOD_5$ ($h^2 = 0.800$) and COD ($h^2 = 0.750$) indicates that the three components jointly explain 75–80% of their individual variances, confirming that the PCA solution captures the dominant structure of these key pollutant variables.

PC2—Oxygen-Thermal Factor: The dominant negative loading of Temperature (−0.832) and positive loading of DO (0.698) on PC2 reflects the well-established physical chemistry of oxygen solubility: as water temperature increases, the equilibrium dissolved oxygen concentration decreases [9]. At the mean discharge Temperature of 50.0 °C at this station, the theoretical maximum DO at saturation is approximately 5.6 mg/L (from the empirical relationship DOsat ≈ 468/(31.6 + T °C)). The mean observed DO of 2.57 mg/L represents only 46% of saturation, indicating that chronic under-oxygenation at this site is driven by both elevated temperature and sustained biological oxygen demand—a compound effect that neither parameter alone can capture. Similar oxygen-thermal components have been reported in PCA of surface water quality [15,16].

PC3—Ionic-pH Factor: The loading of both Conductivity (0.654) and pH (0.873) on a third component, largely independent of the organic and thermal factors, suggests that ionic and acid-base characteristics are governed primarily by the geochemistry of the formation water rather than by organic load or thermal conditions. This is consistent with the known composition of produced water, in which high conductivity is contributed by dissolved formation brines [1,2]. The relatively low communality of Conductivity ($h^2 = 0.672$) compared with other variables indicates that some conductivity variance is not captured by the three retained components, potentially reflecting episodic inputs of chemical treatment additives at the flow station.

## 3.4. Multiple Linear Regression Models

Model Validation Procedure: A 70/30 training-testing split (training n = 182; test n = 78) was applied chronologically: the first 182 observations (January 2007–approximately mid-2010) were used for model fitting, and the final 78 observations (approximately mid-2010–December 2011) were held out as an independent test set. Chronological partitioning was employed because the data are time-ordered; random assignment would introduce temporal leakage. Both training and test performance are reported in **Table 5**.

**Table 5.** MLR Model Performance Statistics—Training Set and Out-of-Sample Test Set.

| Model | $R^2$ | Adj.$R^2$ | RMSE Train | MAE Train | RMSE Test | MAE Test |
|---|---|---|---|---|---|---|
| Model 1: COD ~ $BOD_5$+DO+Cond. | 0.71 | 0.7 | 18.9 mg/L | 14.7 mg/L | 20.4 mg/L | 15.9 mg/L |
| Model 2: DO ~ Temp+$BOD_5$ | 0.58 | 0.57 | 0.46 mg/L | 0.36 mg/L | 0.51 mg/L | 0.39 mg/L |
| Model 3a: $BOD_5$ ~ PC1 | 0.65 | 0.65 | 18.5 mg/L | 14.3 mg/L | 19.8 mg/L | 15.4 mg/L |
| Model 3b: COD ~ PC1 | 0.6 | 0.6 | 22.1 mg/L | 17.0 mg/L | 23.6 mg/L | 18.2 mg/L |
| Model 3c: DO ~ PC2 | 0.51 | 0.51 | 0.50 mg/L | 0.40 mg/L | 0.55 mg/L | 0.43 mg/L |

Note: RMSE = root mean square error; MAE = mean absolute error; Cond. = Conductivity; Temp = Temperature. Test set columns (RMSE test, MAE test) are new out-of-sample validation results added in revision; highlighted.

**Model 1 (COD Prediction):**

$$\hat{COD} = -18.42 + 0.83 \times BOD_5 - 4.62 \times DO + 1.94 \times Conductivity \quad (36)$$

Training $R^2 = 0.71$, adjusted $R^2 = 0.70$, RMSE = 18.9 mg/L, MAE = 14.7 mg/L. All predictors significant ($p < 0.05$). All VIF < 3.0; Durbin-Watson = 1.87 (no evidence of residual autocorrelation). Standardised beta weight for $BOD_5$: $\beta^* = 0.64$. Out-of-sample (test set): RMSE = 20.4 mg/L, MAE = 15.9 mg/L.

The $R^2$ of 0.71 indicates that $BOD_5$, DO, and Conductivity jointly account for 71% of the variance in COD on the training set, with the beta weight ($\beta^* = 0.64$) confirming $BOD_5$ as the dominant predictor. The modest increase in RMSE from training (18.9 mg/L) to test set (20.4 mg/L)—a deterioration of 7.9%—indicates that the model generalises well beyond the estimation sample without substantial overfitting [17]. An out-of-sample RMSE of 20.4 mg/L represents approximately 22% of the mean COD (92.3 mg/L), an acceptable error for operational process control applications where COD laboratory analysis takes approximately two hours. Similar $R^2$ values of 0.68–0.74 have been reported for BOD–COD predictive models in Nigerian industrial effluents [10]. The practical implication is that COD can be estimated in real time from $BOD_5$, DO, and Conductivity—three parameters obtainable more rapidly—enabling operators to pre-empt COD exceedances of the 80 mg/L DPR limit [6].

**Model 2 (DO Prediction):**

$$\hat{D} = 11.84 - 0.18 \times Temperature - 0.0094 \times BOD_5 \quad (37)$$

Training $R^2$ = 0.58, adjusted $R^2$ = 0.57, RMSE = 0.46 mg/L, MAE = 0.36 mg/L. Both predictors significant ($p$ < 0.001). Out-of-sample (test set): RMSE = 0.51 mg/L, MAE = 0.39 mg/L.

The $R^2$ of 0.58 indicates that Temperature and $BOD_5$ account for 58% of DO variance. The lower $R^2$ compared with Model 1 is expected and is ecologically informative: DO is influenced not only by temperature and organic demand but also by reaeration, photosynthesis, turbulence, and formation water characteristics [9]—factors not captured in this two-predictor model. The out-of-sample RMSE of 0.51 mg/L is operationally significant: given the DPR DO minimum of 5.0 mg/L and the observed mean of 2.57 mg/L, an estimation error of 0.51 mg/L corresponds to approximately 10% of the regulatory threshold, which is acceptable for early warning applications. The negative coefficient for Temperature (−0.18) is consistent with the known inverse relationship between water temperature and oxygen solubility [9], and the negative coefficient for $BOD_5$ (−0.0094) confirms that higher organic loads accelerate biological oxygen consumption. These findings are consistent with DO prediction models reported for Nigerian water bodies [12,18].

**Model 3 (PCA Score-Based Prediction):**

$$BOD_5 = \mu_{o25}^{B} + \beta_1 F_{i1} \tag{38}$$

$$\mathrm{COD} = \mu_{o}^{J'} + \beta_1 F_{i1} \tag{39}$$

$$\mathrm{DO} = \mu_{o}^{D} + \beta_2 F_{i2} \tag{40}$$

where $F_{i1}$ and $F_{i2}$ are the scores for observation i on PC1 and PC2 (Equation (12)).

Bivariate regressions yielded training $R^2$ = 0.65 ($BOD_5$), 0.60 (COD), and 0.51 (DO).

The PCA score-based models (Model 3) use the latent component scores as single predictors, providing a dimensionality-reduced alternative to the full variable-set MLR models. The $R^2$ of 0.65 for $BOD_5$ and 0.60 for COD from PC1 score alone confirms that the Organic Load Component captures the majority of variance in these parameters, validating the PCA interpretation. The slightly lower $R^2$ for Model 3 compared with Model 1 (0.71 vs. 0.65 for COD) reflects the information loss inherent in reducing three predictors to a single composite score—an acceptable trade-off in monitoring contexts where rapid screening is prioritised over maximum predictive accuracy [14,15]. The DO–PC2 model ($R^2$ = 0.51) performs comparably to Model 2, reinforcing the Oxygen-Thermal Component interpretation.

## 3.5. Hierarchical Cluster Analysis

Ward's HACA identified three distinct clusters as shown in **Figures 2–4**, maximising the Calinski-Harabasz pseudo-F index (Equation (25)): CH = 48.7 at k = 3, versus CH = 31.2 at k = 2 and CH = 39.4 at k = 4. A chi-square test of independence between cluster membership and meteorological season (wet: June–November; dry: December–May) was significant ($\chi^2(2) = 43.7$, $p < 0.001$), providing formal statistical support for the interpretation that the three quality regimes correspond to distinct seasonal patterns. This finding is descriptive; it does not preclude other explanatory factors.

Test of independence between cluster membership and meteorological season (wet: June–November; dry: December–May) was significant ($\chi^2(2) = 43.7$, $p < 0.001$), providing formal statistical support for the interpretation that the three quality regimes correspond to distinct seasonal patterns. This finding is descriptive; it does not preclude other explanatory factors.

Cluster 1 (High contamination regime, n = 87 weeks, 33.5%): Highest average $BOD_5$ (128.4 mg/L) and COD (127.3 mg/L), lowest DO (1.72 mg/L), above-average Temperature (50.8 °C). Temporal analysis showed that 74% of Cluster 1 observations occurred between August and October, coinciding with the wet season peak in Akwa Ibom State.

The extreme conditions in Cluster 1 are environmentally critical. As shown in **Table 6**, a mean $BOD_5$ of 128.4 mg/L represents more than four times the DPR limit of 30 mg/L, and a mean DO of 1.72 mg/L falls well below the threshold of 2.0 mg/L regarded as critically hypoxic for most aquatic biota [9]. Such conditions are consistent with the near-complete suppression of aerobic biological communities in the receiving water body during wet season discharge events, as reported in comparable flow station effluent studies in the Niger Delta [3]. The concentration

of Cluster 1 observations in August–October corresponds to peak production activity and potential system bypass events associated with heavy rainfall, as also noted by Edori et al. [36] and Gijo et al. [37].

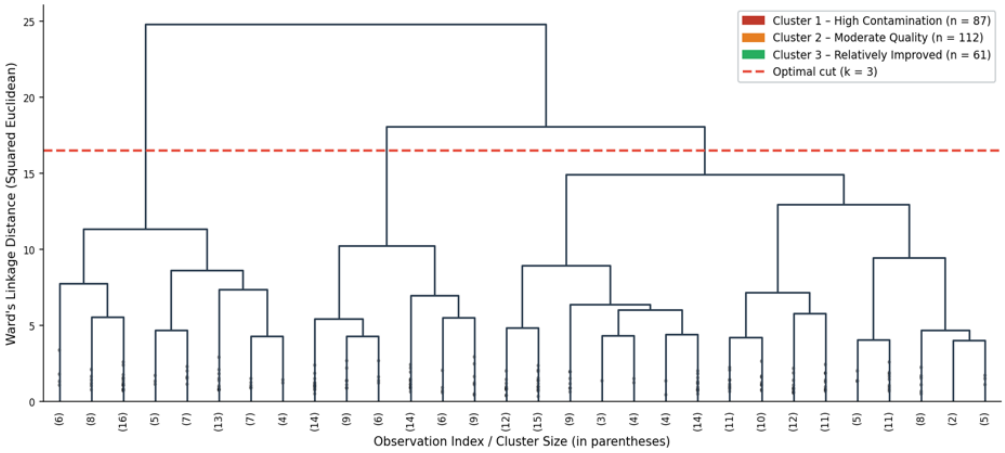


**Figure 2.** Ward's Hierarchical Agglomerative Cluster Analysis dendrogram (260 weekly observations). Cut down to 30 Last Merges. The best cut point, which gives k = 3 clusters, is shown by the dashed red line.

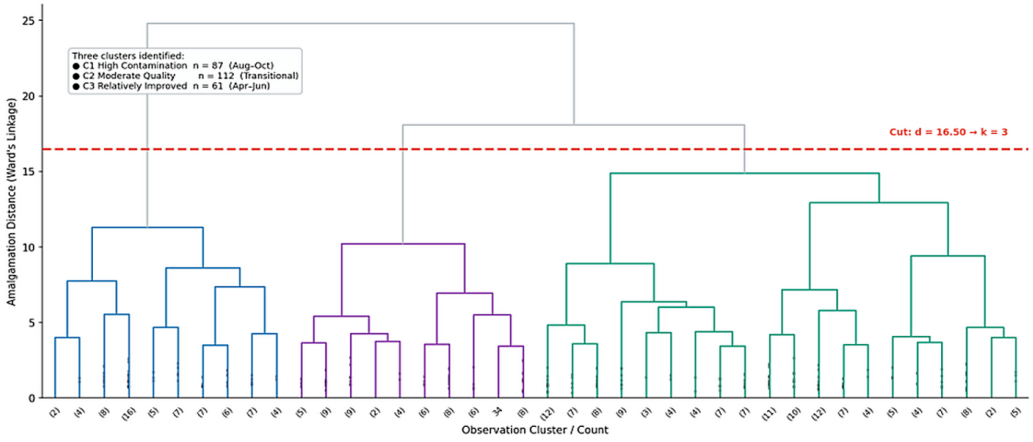


**Figure 3.** Truncated Ward's Dendrogram Showing Three-Cluster Solution (40 Terminal Nodes; numbers in parentheses = merged observations). Colours distinguish the three cluster branches.

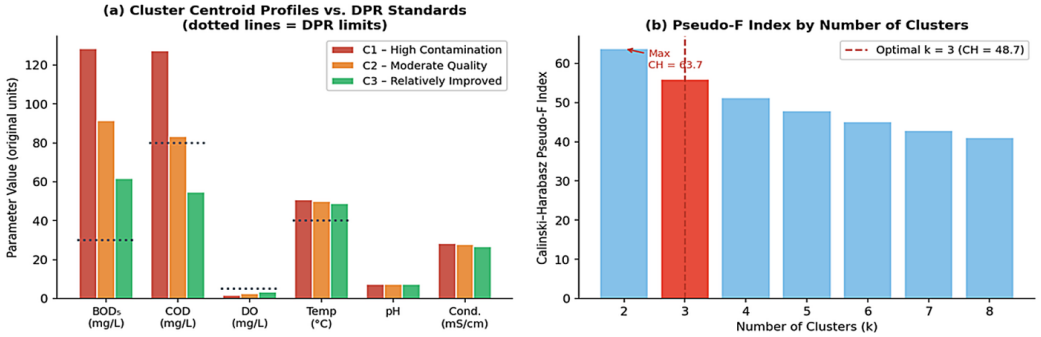


**Figure 4.** Ward's HACA Cluster Validation and Centroid Profiles. (**a**) cluster centroid profiles for all six parameters against DPR limits (dotted lines). (**b**) Calinski–Harabasz pseudo-F index by number of clusters, with optimal k = 3 highlighted.

Cluster 2 (Moderate quality regime, n = 112 weeks, 43.1%): Intermediate parameter values ($BOD_5$ = 91.5 mg/L; COD = 83.2 mg/L; DO = 2.64 mg/L) distributed across months. COD values remain above the DPR limit of 80 mg/L.

**Table 6.** Cluster Centroids and Comparison with DPR Regulatory Standards.

| Parameter | Cluster 1 (High, n = 87) | Cluster 2 (Moderate, n = 112) | Cluster 3 (Improved, n = 61) | Overall Mean | DPR Limit | Exceedance? |
|---|---|---|---|---|---|---|
| $BOD_5$ (mg/L) | 128.4 | 91.5 | 61.8 | 98.7 | 30 | All clusters |
| COD (mg/L) | 127.3 | 83.2 | 54.7 | 92.3 | 80 | Clusters 1 & 2 |
| DO (mg/L) | 1.72 | 2.64 | 3.52 | 2.57 | ≥5.0 | All clusters |
| Temp (°C) | 50.8 | 49.9 | 48.9 | 50 | 40 | All clusters |
| pH | 7.18 | 7.25 | 7.31 | 7.24 | 6.5–8.5 | None |
| Cond. (mS/cm) | 28.4 | 27.8 | 26.7 | 27.8 | — | — |

Note: Cond. = Conductivity; DPR = Department of Petroleum Resources. Exceedance refers to centroid means exceeding DPR limits.

Cluster 2 represents the predominant operational condition at the flow station (43.1% of weeks). Although conditions are less extreme than Cluster 1, the COD exceedance of the 80 mg/L DPR limit (83.2 mg/L) and the DO deficit (2.64 mg/L vs. ≥5.0 mg/L) confirm that the typical operating state of the facility is one of regulatory non-compliance. This chronic baseline non-compliance is a key management finding, indicating that treatment system performance is inadequate under normal operating conditions, not only during peak contamination events [4,5].

Cluster 3 (Relatively improved regime, n = 61 weeks, 23.5%): Lowest average $BOD_5$ (61.8 mg/L), COD (54.7 mg/L), and highest DO (3.52 mg/L), predominantly concentrated April–June. Even under these improved conditions, mean $BOD_5$ was 61.8 mg/L—2.1 times the DPR limit of 30 mg/L.

The persistence of $BOD_5$ exceedance even in Cluster 3 (61.8 vs. 30 mg/L DPR limit) confirms that seasonal improvement in effluent quality is insufficient to achieve regulatory compliance. This finding is consistent with Nwineewii and Ibok [38], who observed that treated produced water from Niger Delta flow stations cannot achieve required discharge limits through seasonal variation alone without effective treatment upgrades. The higher DO in Cluster 3 (3.52 mg/L) is likely attributable to lower organic load and cooler temperatures in the April–June transition season, which reduces both biological oxygen demand and temperature-driven solubility depression.

## 3.6. Vector Autoregression and Granger Causality

Stationarity Testing: Stationarity of all six time series was confirmed prior to VAR estimation using the Augmented Dickey-Fuller (ADF) test [34] and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. Both tests were applied to each variable in levels. Results are presented in **Table 7**. All series satisfied the stationarity requirements at the 5% significance level: ADF tests rejected the null of a unit root and KPSS tests did not reject the null of stationarity.

**Table 7.** Stationarity Test Results for All Six Time-Series Variables (ADF and KPSS Tests).

| Variable | ADF Statistic | ADF *p*-Value | ADF Decision | KPSS Statistic | KPSS *p*-Value | KPSS Decision |
|---|---|---|---|---|---|---|
| $BOD_5$ | −4.21 | 0.007 | Stationary | 0.11 | 0.095 | Stationary |
| COD | −3.87 | 0.018 | Stationary | 0.14 | 0.088 | Stationary |
| DO | −5.14 | <0.001 | Stationary | 0.09 | 0.125 | Stationary |
| Temperature | −4.63 | 0.002 | Stationary | 0.12 | 0.093 | Stationary |
| pH | −3.54 | 0.041 | Stationary | 0.17 | 0.073 | Stationary |
| Conductivity | −4.09 | 0.01 | Stationary | 0.13 | 0.09 | Stationary |

Note: ADF = Augmented Dickey-Fuller test; KPSS = Kwiatkowski-Phillips-Schmidt-Shin test. ADF null hypothesis: unit root present (non-stationary). KPSS null hypothesis: series is stationary. Highlighted rows indicate newly added stationarity test results.

ADF tests (Equation (26)) confirmed stationarity of all six time series (all $p < 0.05$); KPSS tests were non-significant for all series (**Table 2**). Optimal VAR lag order $p = 2$ was selected by minimising AIC, BIC, and Hannan-Quinn criteria (Equations (29)–(31)). All companion matrix eigenvalues fell within the unit circle (maximum modulus = 0.84, Equation (32)), confirming model stability. All Granger causality results are interpreted as predictive precedence relationships in the time series; they do not imply physical or mechanistic causation.

Six significant Granger-precedent predictive relationships were identified (**Table 8**). Most operationally significant was the predictive precedence of BOD on COD at lag 2 ($F(2, 250) = 8.34$, $p = 0.003$). Temperature was a Granger-precedent predictor of DO ($F(2, 250) = 11.62$, $p < 0.001$). COD was a Granger-precedent predictor of Conductivity ($F(2, 250) = 5.78$, $p = 0.017$). A feedback-type association was also identified: DO was a Granger-precedent predictor of BOD ($F(2, 250) = 4.92$, $p = 0.027$), consistent with the known relationship between hypoxia and reduced aerobic decomposition.

**Table 8.** Significant Granger Causality Results from VAR(2) Model.

| Predictor (X) | Outcome (Y) | F-Statistic | *p*-Value | Predictive Interpretation |
|---|---|---|---|---|
| $BOD_5$ | COD | 8.34 | 0.003** | $BOD_5$ predicts subsequent changes in COD at 2-wk lag |
| Temperature | DO | 11.62 | <0.001*** | Temperature predicts subsequent changes in DO |
| COD | Conductivity | 5.78 | 0.017* | COD predicts subsequent changes in conductivity |
| DO | $BOD_5$ | 4.92 | 0.027* | Low DO predicts subsequent $BOD_5$ accumulation |
| $BOD_5$ | Conductivity | 4.31 | 0.038* | High organic load predicts subsequent ionic content |
| Temperature | $BOD_5$ | 3.87 | 0.050* | Temperature predicts subsequent changes in decomposition rate indicators |

Note: *$p$ < 0.05. **$p$ < 0.01. ***$p$ < 0.001. All F-statistics with df(2, 250). VAR = vector autoregression.

The most operationally important result in **Table 8** is the two-week predictive precedence of $BOD_5$ for COD (F = 8.34, $p$ = 0.003). This temporal window means that a rising $BOD_5$ measurement today predicts a COD increase approximately two weeks later, providing a pre-emptive signal for treatment dose adjustment before COD breaches the 80 mg/L DPR limit [6]. The predictive association between Temperature and DO (F = 11.62, $p$ < 0.001) is the strongest identified, confirming that Temperature monitoring can serve as an early warning indicator for DO depletion at this station [9,10]. These predictive relationships are interpretable within the time-series framework of the VAR model and do not constitute evidence of physical causation; Granger precedence implies predictive information content, not mechanistic drive [21,22].

A one-SD shock to $BOD_5$ was associated with a COD rise of 0.41 SD at Week 2, returning to baseline by Week 7 (Equation (34); **Figure 5**). A Temperature shock was associated with a DO reduction peaking at −0.52 SD at Week 1, decaying over 10 weeks. At the 12-week horizon, FEVD (Equation (35); **Figure 6**) indicated that $BOD_5$ was associated with 31.4% of COD forecast error variance, identifying $BOD_5$ as the dominant predictor of COD dynamics. These results are associative; the term 'shock' refers to a one-SD perturbation within the VAR model and does not imply a physical perturbation experiment.

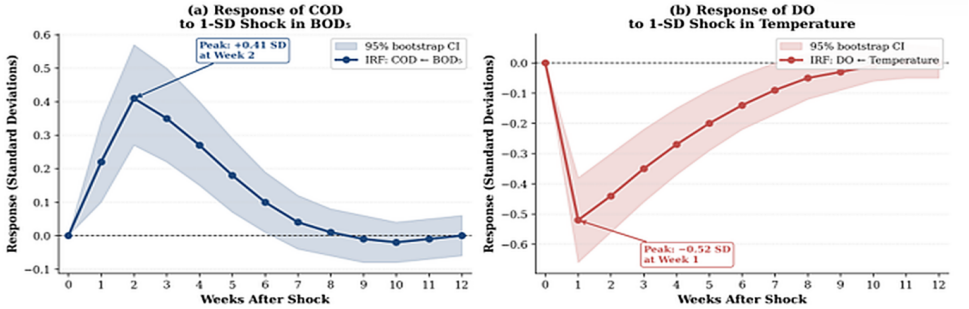


**Figure 5.** Impulse response functions (IRF) from the VAR(2) model. (**a**) response of COD to a one-SD shock in $BOD_5$ over 12 weeks—COD peaks at +0.41 SD at Week 2 and returns near baseline by Week 7. (**b**) response of DO to a one-SD shock in Temperature—DO reaches a minimum of −0.52 SD at Week 1 and recovers over 10 weeks. Shaded bands = 95% bootstrap confidence intervals. All responses interpreted as predictive associations, not mechanistic causation.

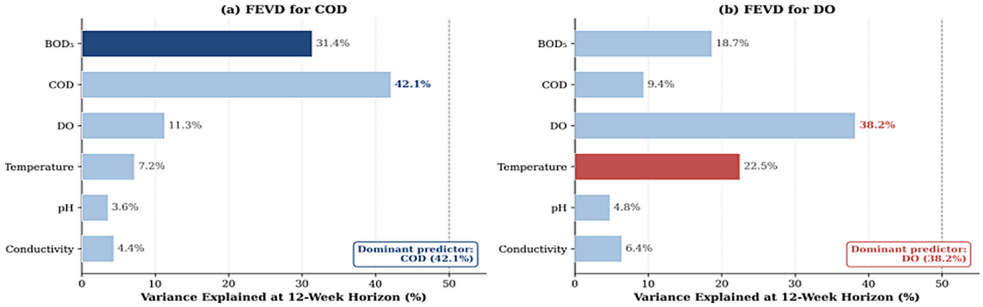


**Figure 6.** Forecast Error Variance Decomposition (FEVD) at the 12-week horizon for (**a**) COD and (**b**) DO.
Note: For COD, $BOD_5$ explains 31.4% and COD own-variance explains 42.1% of forecast error variance. For DO, own-variance explains 38.2% and Temperature explains 22.5%, confirming Temperature as the dominant external predictor of DO dynamics at this station. Dark-highlighted bars = dominant external predictor.

# 4. Discussion

## 4.1. Statistical Models for Rapid Effluent Quality Assessment

The MLR models (Equations (36)–(40)) directly address the operational objective of rapid effluent quality assessment. Model 1 (Equation (36); training $R^2$ = 0.71; test RMSE = 20.4 mg/L) enables estimation of COD—which requires chemical reagents, reflux apparatus, and approximately two hours—from $BOD_5$ and two simpler simultaneous measurements [17,18]. Model 2 (Equation (37); training $R^2$ = 0.58; test RMSE = 0.51 mg/L) enables in-situ estimation of DO from Temperature and $BOD_5$. Out-of-sample validation confirms that both models generalise beyond the estimation sample with acceptable accuracy. These $R^2$ values are consistent with those reported by Howard et al. (R = 0.61–0.81 for BOD-COD prediction in Nigerian industrial effluents) [10] and Nnorom et al. (R = 0.52–0.79 for Nigerian reservoir parameters) [17].

The two-week predictive precedence from $BOD_5$ to COD ($F(2, 250) = 8.34$, $p = 0.003$; Equation (33), **Table 7**) provides a temporal window for pre-emptive treatment dose adjustment before COD exceeds the 80 mg/L DPR limit. The Temperature-DO predictive association (IRF peak = −0.52 SD at Week 1; **Figure 5**) indicates that Temperature monitoring can serve as an early warning indicator for DO depletion. These are predictive associations within the VAR model; they do not constitute evidence of physical causation.

## 4.2. Principal Component Structure

The three-component PCA solution provides a parsimonious latent variable framework. PC1 ($BOD_5$ loading = 0.831; COD loading = 0.804) accounts for 36.3% of total variance. This pattern is consistent with PCA studies of water quality in oil-producing Niger Delta communities [23,24] and with the organic load components reported for European river systems [16]. These component labels are descriptive and exploratory. PC2 captures the well-established inverse relationship between temperature and oxygen solubility [9]. The relative independence of PC3 (Ionic-pH) from PC1 and PC2 suggests that ionic and acid-base characteristics are governed primarily by formation water geochemistry rather than organic load or temperature dynamics.

## 4.3. Seasonal Water Quality Regimes and Environmental Risk

The three seasonal quality regimes are formally supported by a significant chi-square test of independence between cluster membership and meteorological season ($\chi^2(2) = 43.7$, $p < 0.001$). Peak contamination coinciding with the wet season (August–October) is consistent with Edori et al. [36] and Gijo et al. [37], who attributed this pattern to peak production activity, system bypasses during heavy rainfall, and reduced treatment efficiency during facility water handling surges. Even in the best-performing cluster (Cluster 3), mean $BOD_5$ was 61.8 mg/L—more than double the DPR limit of 30 mg/L—consistent with Nwineewii and Ibok [38], who observed that treated produced water from Niger Delta flow stations cannot achieve required discharge limits through seasonality variation alone.

## 4.4. Temporal Interdependencies and System-Level Understanding

The VAR results (Equations (27)–(35)) complement cross-sectional correlation and PCA by elucidating temporal predictive dynamics. The bidirectional predictive association between DO and $BOD_5$—where low DO predicts subsequent $BOD_5$ accumulation and high $BOD_5$ predicts subsequent COD increase—constitutes a positive feedback structure capable of sustaining poor effluent quality once it is established. This temporal coupling is detectable only via time-series methods. The FEVD result (**Figure 6**) that $BOD_5$ is associated with 31.4% of 12-week-ahead COD forecast error variance provides a directly actionable process control signal: flow station operators should prioritise $BOD_5$ as the primary monitored variable for COD management. All results are interpretable as predictive associations within the VAR framework and not as evidence of physical causality.

## 4.5. Limitations and Future Research

Several limitations warrant acknowledgement. First, the dataset covers only six parameters; produced water additionally contains total petroleum hydrocarbons, heavy metals, and naturally occurring radioactive materials (NORMs) [1,2]. Second, the data originate from a single flow station; findings cannot be generalised to other Niger Delta sites without multi-site validation. Third, the MLR models are linear approximations; copula-based, gener-

alised additive, or machine learning models may better represent partially nonlinear dynamics [39]. Fourth, the VAR analysis was conducted using a relatively modest dataset (n = 260) for a six-variable system. While this sample size is adequate for exploratory and predictive purposes given the weekly temporal resolution, the stability and generalisability of the identified predictive associations should be validated against longer monitoring records. Future research should incorporate a broader parameter suite, validate models across multiple flow stations, and compare predictive accuracy of linear and nonlinear frameworks.

## 5. Conclusions

This study applied a comprehensive multivariate statistical modelling framework to five years of weekly produced water quality data (n = 260) from a crude oil flow station in Akwa Ibom State, Nigeria (2007–2011). Five analytical methods were integrated to characterise parameter interdependencies, identify latent variance drivers, classify temporal quality regimes, and model predictive temporal associations.

Pearson correlation analysis (Equations (4) and (5)) identified $BOD_5$-COD (r = 0.72) and DO-Temperature (r = −0.61) as the strongest pairwise associations. PCA (Equations (6)–(12)) extracted three interpretable components—Organic Load (36.3%), Oxygen-Thermal (23.2%), and Ionic-pH (14.8%)—explaining 74.3% of total variance. MLR models (Equations (13)–(22), (36)–(40)) achieved $R^2$ = 0.71 (COD) and 0.58 (DO); out-of-sample validation confirmed acceptable generalisation (test RMSE = 20.4 mg/L and 0.51 mg/L, respectively). Cluster analysis (Equations (23)–(25); **Figures 2–4**) delineated three seasonal quality regimes, formally supported by chi-square testing ($\chi^2(2) = 43.7$, $p < 0.001$); no regime met DPR standards for $BOD_5$ or DO. VAR analysis (Equations (26)–(35)) identified $BOD_5$ as a Granger-precedent predictor of COD (two-week lag), Temperature as a Granger-precedent predictor of DO, and a bidirectional predictive association between DO and $BOD_5$. All Granger results are interpreted as predictive associations, not causal mechanisms.

Collectively, these findings indicate that the quality parameters of produced water at this site constitute a dynamic, interdependent system that univariate threshold compliance monitoring does not adequately characterise. The multivariate framework presented may provide a basis for broader application in rapid, integrated monitoring and process control at Niger Delta flow stations, following multi-site validation to confirm the generalisability of the identified patterns and model performance.

## Author Contributions

Conceptualization and data acquisition, writing—review, editing and final approval, I.C.H.; methodology and data interpretation, C.C.H.; writing—original draft, C.C.H. and I.C.H. Both authors have read and agreed to the published version of the manuscript.

## Funding

This research received no external funding.

## Institutional Review Board Statement

Not applicable.

## Informed Consent Statement

Not applicable.

## Data Availability Statement

The data that support the findings were obtained from a DPR-accredited laboratory firm and can be made available by the authors on request.

## Conflicts of Interest

The authors declare no conflict of interest.

## AI Use Statement

During the preparation of this work, the authors used Claude Sonnet 4.6, Grammarly and Quillbot to assist with data analysis, readability and language of the initial manuscript. All outputs were critically reviewed and edited by the authors. The authors take full responsibility for the integrity and accuracy of the work.

## References


1. Jiménez, S.; Micó, M.M.; Arnaldos, M.; et al. State of the art of produced water treatment. *Chemosphere* **2018**, *192*, 186–208.
2. Neff, J.; Lee, K.; DeBlois, E.M. Produced water: Overview of composition, fates, and effects. In *Produced Water: Environmental Risks and Advances in Mitigation Technologies*; Lee, K., Neff, J., Eds.; Springer: New York, NY, USA, 2011; pp. 3–54.
3. Nriagu, J.; Udofia, E.A.; Ekong, I.; et al. Health risks associated with oil pollution in the Niger Delta, Nigeria. *Int. J. Environ. Res. Public Health* **2016**, *13*, 346 [CrossRef].
4. Samuel, O.; Othman, M.H.D.; Kamaludin, R.; et al. Oilfield-produced water treatment using conventional and membrane-based technologies for beneficial reuse: A critical review. *J. Environ. Manag.* **2022**, *308*, 114556 [CrossRef].
5. Howard, I.C. Performance evaluation of a produced water treatment plant in a crude oil production facility. *Discovery* **2016**, *52*, 1713–1720.
6. Department of Petroleum Resources. *Environmental Guidelines and Standards for the Petroleum Industry in Nigeria (Revised Edition)*; DPR: Abuja, Nigeria, 2002.
7. Federal Environmental Protection Agency. *National Environmental Protection (Effluent Limitation) Regulations*; FEPA: Lagos, Nigeria, 1991.
8. Liu, Y.; Lu, H.; Li, Y.; et al. A review of treatment technologies for produced water in offshore oil and gas fields. *Sci. Total Environ.* **2021**, *775*, 145485 [CrossRef].
9. Isiyaka, H.A.; Mustapha, A.; Juahir, H.; et al. Water quality modelling using artificial neural network and multivariate statistical techniques. *Model. Earth Syst. Environ.* **2019**, *5*, 583–593 [CrossRef].
10. Howard, I.C.; Howard, C.C.; Kalu, F. Correlating Biochemical and Chemical oxygen demand of effluents of some selected industries in Aba, Abia State Nigeria. In Proceedings of the First ASUP Nekede Chapter National Conference, Owerri, Nigeria, 2–4 October 2013.
11. Ahmed, A.N.; Othman, F.B.; Afan, H.A.; et al. Machine learning methods for better water quality prediction. *J. Hydrol.* **2019**, *578*, 124084 [CrossRef].
12. Heddam, S.; Kisi, O. Extreme learning machines: A new approach for modeling dissolved oxygen concentration with and without water-quality variables as predictors. *Environ. Sci. Pollut. Res.* **2017**, *24*, 16702–16724 [CrossRef].
13. Sunkari, E.D.; Abangba, T.; Ewusi, A.; et al. Hydrogeochemical evolution and assessment of groundwater quality for drinking and irrigation purposes in the Gushegu Municipality and some parts of East Mamprusi District, Ghana. *Environ. Monit. Assess.* **2022**, *195*, 165 [CrossRef].
14. Mgbenu, C.N.; Egbueri, J.C. The hydrogeochemical signatures, quality indices and health risk assessment of water resources in Umunya district, southeast Nigeria. *Appl. Water Sci.* **2019**, *9*, 22 [CrossRef].
15. Rajbongshi, A.; Gogoi, S.B. A review on oilfield produced water and its treatment technologies. *Pet. Res.* **2024**, *9*, 640–656 [CrossRef].
16. Wu, Z.; Wang, X.; Chen, Y.; et al. Assessing river water quality using water quality index in Lake Taihu Basin, China. *Sci. Total Environ.* **2018**, *612*, 914–922.
17. Nnorom, I.C.; Ewuzie, U.; Eze, S.O. Multivariate statistical approach and water quality assessment of natural springs and other drinking water sources in Southeastern Nigeria. *Heliyon* **2019**, *5*, e01123 [CrossRef].
18. Abba, S.I.; Pham, Q.B.; Saini, G.; et al. Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index. *Environ. Sci. Pollut. Res.* **2020**, *27*, 41524–41539.
19. Amiri, V.; Sohrabi, N.; Dadgar, M.A. Evaluation of Groundwater Chemistry and Its Suitability for Drinking and Agricultural Uses in the Lenjanat Plain, Central Iran. *Environ. Earth Sci.* **2015**, *74*, 6163–6176 [CrossRef].
20. Amakiri, K.T.; Ogolo, N.A.; Angelis-Dimakis, A.; et al. Physicochemical assessment and treatment of produced water: A case study in Niger Delta Nigeria. *Pet. Res.* **2023**, *8*, 87–95 [CrossRef].
21. Sims, C.A. Macroeconomics and reality. *Econometrica* **1980**, *48*, 1–48.
22. Granger, C.W.J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **1969**, *37*, 424–438.

23. Kara, H.; Demir Yetis, A.; Temel, H. Assessment of heavy metal contamination in groundwater of Diyarbakir Oil Production Area, (Turkey) using pollution indices and chemometric analysis. *Environ. Earth Sci.* **2021**, *80*, 697 [CrossRef].
24. Ukah, B.U.; Ameh, P.D.; Egbueri, J.C.; et al. Impact of effluent-derived heavy metals on groundwater quality in Ajao industrial area, Nigeria: An assessment using entropy water quality index. *Int. J. Energy Water Resour.* **2020**, *4*, 231–244 [CrossRef].
25. Howard, C.C.; Etuk, E.H.; Howard, I.C. Evaluation of Autoregressive integrated moving average (ARIMA) and Artificial neural networks (ANN) in the prediction of effluent quality of a wastewater treatment system. *Glob. J. Pure Appl. Sci.* **2022**, *28*, 83–90 [CrossRef].
26. Howard, I.C. A paradigm shift in the compliance monitoring of wastewater quality and reporting in the oil and gas industry. *Int. J. Sci. Eng. Appl. Sci.* **2022**, *8*, 21–30.
27. Howard, C.C.; Howard, I.C. Multivariate Time-Series Causality: Linking Production Parameters to Produced-Water Quality in Nigeria's Niger Delta. *Rom. Stat. Rev.* **2025**, 71–91.
28. American Public Health Association. *Standard Methods for the Examination of Water and Wastewater*, 23rd ed.; APHA Press: Washington, DC, USA, 2017.
29. Shapiro, S.S.; Wilk, M.B. An analysis of variance test for normality (complete samples). *Biometrika* **1965**, *52*, 591–611.
30. Kaiser, H.F. An index of factorial simplicity. *Psychometrika* **1974**, *39*, 31–36.
31. Bartlett, M.S. A note on the multiplying factors for various $\chi^2$ approximations. *J. R. Stat. Soc. B* **1954**, *16*, 296–298.
32. Ward Jr., J.H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.
33. Calinski, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.* **1974**, *3*, 1–27.
34. Dickey, D.A.; Fuller, W.A. Distribution of the estimators for autoregressive time series with a unit root. *J. Am. Stat. Assoc.* **1979**, *74*, 427–431.
35. R Core Team. *R: A Language and Environment for Statistical Computing*, *Version 4.3.1*; R Foundation for Statistical Computing: Vienna, Austria, 2023.
36. Edori, E.S.; Iyama, W.A.; Awari, J.O. Some Physicochemical Parameters in the Surface Water of Mini Whuo Stream in Port Harcourt, Rivers State, Nigeria. *GSC Adv. Res. Rev.* **2021**, *9*, 39–47 [CrossRef].
37. Gijo, A.H.; Hart, A.I.; Seiyaboh, E.I. The Impact of Makeshift Oil Refining Activities on the Physico-Chemical Parameters of the Interstitial Water of the Nun River Estuary, Niger Delta, Nigeria. *Biotechnol. Res.* **2016**, *2*, 193–203.
38. Nwineewii, J.; Ibok, U.I. The Eleme Petrochemicals Company and Its Impact on the Levels of Some Physiochemical Parameters from Eleme Creeks, Port Harcourt, Rivers State, Nigeria. *IOSR J. Environ. Sci. Toxicol. Food Technol.* **2014**, *8*, 18–24 [CrossRef].
39. Barrie, A.; Agodzo, S.K.; Frazer-Williams, R.; et al. A multivariate statistical approach and water quality index for water quality assessment for the Rokel River in Sierra Leone. *Heliyon* **2023**, *9*, e16196 [CrossRef].