



## Enhancing water quality assessment in Skikda, Algeria using the PCA-based weighted index (WQI\_P) and its predictive performance: a comparison with traditional WA\_WQI approaches

Title	Enhancing water quality assessment in Skikda, Algeria using the PCA-based weighted index (WQI_P) and its predictive performance: a comparison with traditional WA_WQI approaches
Author(s)	Mostefa, Benacherine;Hinda, Hafid;Martínez, Alejandro;Noua, Allaoua;Abdelatif, Satour;Fadila, Fertas;Diganta, Mir Talas Mahammad;Nadjib, Lyazid Mohamed;Debassi, Bouchra;Uddin, Md Galal
Publication Date	2026-02-04
Publisher	Elsevier
Repository DOI	<a href="https://doi.org/10.1016/j.jconhyd.2026.104875">https://doi.org/10.1016/j.jconhyd.2026.104875</a>

1 **Enhancing water quality assessment in Skikda, Algeria using the PCA-based**  
2 **weighted index (WQI\_P) and its predictive performance: a comparison with**  
3 **traditional WA\_WQI approaches**

4 **BENACHERINE MOSTEFA<sup>1,\*</sup>, Hafid Hinda<sup>1</sup>, Alejandro Garcia Martinez<sup>2</sup>, Allaoua**  
5 **Noua<sup>1</sup>, Satour Abdelatif<sup>3</sup>, Fertas Fadila<sup>4</sup>, Mir Talas Mahammad Diganta<sup>5,6,7</sup>, Lyazid**  
6 **Mohamed Nadjib<sup>8</sup>, Bouchra Debassi<sup>8</sup>, Md Galal Uddin<sup>5,6,7,9,\*</sup>**

7 <sup>1</sup>Laboratory of Natural Resources and Management of Sensitive Environments, Larbi  
8 Ben M'hidi University, Oum-El-Bouaghi 04000, Algeria

9 <sup>2</sup>Molecular Ecology Group (MEG), Water Research Institute (IRSA), National  
10 Research Council of Italy (CNR), Verbania, Italy

11 <sup>3</sup> Laboratoire de Conservation des Zones humides, Université 8 mai 1945, Guelma,  
12 Guelma 24000, Algeria.

13 <sup>4</sup>LARYSS Laboratory, Mohamed Khider University Biskra, Biskra 07000, Algeria.

14 <sup>5</sup>School of Engineering, College of Science and Engineering, University of Galway, Galway,  
15 Ireland

16 <sup>6</sup>Ryan Institute, University of Galway, Galway, Ireland.

17 <sup>7</sup>Eco-HydroInformatics Research Group (EHIRG), Civil Engineering, University of Galway,  
18 Galway, Ireland.

19 <sup>8</sup>Laboratory of Functional Ecology and Environment, University of Oum El Bouaghi, Oum-El-  
20 Bouaghi 04000, Algeria.

21 <sup>9</sup>Department of Civil, Structural and Environmental Engineering, and Sustainable Infrastructure  
22 Research & Innovation Group, Munster Technological University, Cork, Ireland

23 \*Corresponding author(s):

24 Benacherine Mostefa, Laboratory of Natural Resources and Management of Sensitive  
25 Environments, Larbi Ben M'hidi University, Algeria ([mostefa.benacherine@univ-oeb.dz](mailto:mostefa.benacherine@univ-oeb.dz))

26 Dr Md Galal Uddin, Postdoctoral researcher, Civil Engineering, College of Science and  
27 Engineering, University of Galway, Ireland([mdgalal.uddin@universityofgalway.ie](mailto:mdgalal.uddin@universityofgalway.ie))

## 28 **Abstract**

29 Ensuring reliable river-water quality assessment is increasingly important in North Africa,  
30 where pollution pressures and data limitations complicate monitoring. Therefore, the research  
31 developed a principal-component-analysis–based water quality index (WQI\_P) that is designed  
32 to address eclipsing, multicollinearity, and subjectively assigned weights that affect traditional  
33 indices such as the weighted-arithmetic WQI (WA\_WQI). The objective of the research is to  
34 evaluate whether PCA-derived weights and objective parameter selection improve reliability,  
35 uncertainty, and classification stability. A dataset of 159 river-water samples from the Skikda  
36 region (Algeria) was analyzed. After screening correlated variables and extracting PCA  
37 contributions, WQI\_P was constructed from the retained components. Eight machine-learning  
38 algorithms and a stacked ensemble were used under 10-fold cross-validation to compare the  
39 prediction performance and uncertainty of WQI\_P and WA\_WQI. Agreement metrics, PREI  
40 scores, confidence intervals, and class-transition analysis were used to assess the differences  
41 between the two indices, Predictive uncertainty was quantified using a Gaussian Monte Carlo  
42 simulation, which propagates variability by repeatedly perturbing model residuals to generate  
43 distributions of index predictions. The WQI\_P consistently produced lower prediction errors  
44 (stacked RMSE = 2.74; MAE = 1.75) than the WA\_WQI (RMSE = 3.16; MAE = 2.21), together  
45 with narrower 95% confidence intervals and reduced predictive uncertainty. The classification  
46 outcomes shifted toward a stricter and more balanced assessment: the proportion of samples  
47 classified as “Excellent” decreased (30 to 7), “Good” increased (55 to 88), and “Unsuitable”  
48 declined (40 to 12). These results indicated that grounding weights in the multivariate structure  
49 enhances stability and reduces dependence on a small set of dominant parameters. The findings  
50 demonstrated that the WQI\_P can improve transparency, objectivity, and monitoring efficiency

51 by focusing on the most informative variables. The index is applicable to data-scarce regions  
52 where objective weighting and uncertainty control are essential. Future work should test  
53 WQI\_P across larger and more heterogeneous basins, extend validation using spatial–temporal  
54 blocking, and explore its integration into operational monitoring frameworks.

55 **Keywords:** Principal component analysis, Machine learning, Skikda, Water quality index  
56 model.

## 57 **1. Introduction**

58 Water demand across agriculture, industry and public health continues to rise, which increases  
59 pressure not only on water availability but also on water quality. Global assessments show  
60 persistent degradation of rivers and lakes due to untreated wastewater, agricultural runoff, urban  
61 expansion and climate variability. According to the UN World Water Development Report  
62 2024, more than 2.2 billion people still lack safely managed drinking water and only 10 percent  
63 of global wastewater is safely treated (UNESCO 2024). UNEP evaluations indicate that many  
64 surface water bodies in Africa, Asia and Latin America do not meet basic environmental quality  
65 standards because of high loads of nutrients, organic matter and industrial pollutants (UNEP  
66 2021). Reports from WHO and UNICEF, OECD and IPCC confirm increasing nutrient  
67 enrichment, insufficient wastewater treatment and climate related hydrological stress (IPCC  
68 2023; OECD 2022; WHO and UNICEF 2022). These pressures highlight the need for reliable,  
69 transparent and scalable tools for evaluating surface water quality (Aljanabi et al. 2021; Chidiac  
70 et al. 2023; Singh et al. 2021).

71 A variety of approaches exist for water quality evaluation, including physico chemical and  
72 microbial regulatory standards and biological indicators (Edition 2011; Iyiola and Asiedu  
73 2020). Water Quality Indices (WQIs) aggregate multiple parameters into a single score to  
74 facilitate interpretation. However, traditional indices such as the weighted arithmetic WQI  
75 depend on fixed parameter sets and subjective weighting (Kachroud et al. 2019). Recent studies

76 propose statistical selection methods and uncertainty reduction algorithms to improve WQI  
77 construction (Fartas et al. 2022; Li et al. 2021). These contributions are valuable but remain  
78 insufficient because fully transparent and data-driven rules for indicator selection and weighting  
79 are still lacking.

80 Several newer frameworks attempt to overcome these limitations. The IEWQI introduces a  
81 structured approach to indicator selection, sub indexing, weighting and aggregation and has  
82 shown improved performance in transitional and coastal waters (Uddin et al. 2021a). The RMS  
83 WQI integrates machine learning workflows such as XGBoost with automated hyperparameter  
84 tuning and has demonstrated high predictive skill in groundwater and coastal environments (I.  
85 Khan et al. 2025a). Additional studies from North Africa and the Middle East combine WQIs  
86 with PCA, multivariate analysis and machine learning to assess groundwater quality, salinity  
87 hazards, toxic element contamination and health risks (Eid et al. 2025; El Osta et al. 2022; Gad  
88 et al. 2023, 2024; Gad and El-Hattab 2019; Hfaiedh et al. 2025; Salem et al. 2023). Although  
89 these works show substantial progress, The focus is mainly on groundwater systems, rely on  
90 fixed weighting and rarely evaluate uncertainty or compare results with recent indices such as  
91 IEWQI and RMS WQI. Together revealing gaps in transparent parameter selection, objective  
92 weighting, uncertainty quantification and reproducible comparative assessment.

93 From this body of literature, four main gaps emerge. First, the field still lacks objective and site  
94 independent rules for selecting parameter weights, especially when indicators are strongly  
95 correlated. Second, the uncertainty of weights and final scores is rarely quantified, which limits  
96 confidence in class assignments. Third, generalization across basins and seasons is seldom  
97 tested using rigorous data splitting strategies. Fourth, reproducible benchmarking against recent  
98 indices such as IEWQI and RMS WQI is uncommon (I. Khan et al. 2025b; Olbert et al. 2025;  
99 Uddin, Nash, Mahammad Diganta, et al. 2022). These limitations justify the need for a river

100 focused index that integrates objective indicator screening, data driven weighting, uncertainty  
101 quantification and reproducible comparative evaluation.

102 This study proposes a principal component analysis derived WQI (WQI\_P) for river water that  
103 incorporates objective variable screening, data driven weighting and explicit uncertainty  
104 estimation. The WQI\_P is compared with the traditional weighted arithmetic index (WA\_WQI)  
105 using agreement metrics, confidence intervals, PREI analysis and predictive evaluation with  
106 multiple machine learning algorithms. This comparison helps determine whether PCA based  
107 weighting improves classification reliability and predictive accuracy.

108 Principal Component Analysis was applied after correlation screening, multicollinearity checks  
109 and the Kaiser rule in order to identify a reduced set of informative variables (Nath Roy et al.  
110 2024; Tripathi and Singal 2019). PCA contribution values were then used as data driven weights  
111 to construct WQI\_P (Y. Khan and See 2016). Although PCA reduces collinearity and provides  
112 reproducible loadings, its limitations such as linear assumptions, sensitivity to extreme values  
113 and dependence on sample structure were mitigated through standardization and interpretability  
114 checks (Fartas et al. 2022; Ibrahim et al. 2023; Y. Khan and See 2016; Nath Roy et al. 2024;  
115 Tripathi and Singal 2019).

116 Predictive evaluation was performed using decision trees, random forests, artificial neural  
117 networks, XGBoost, KNN, SVM, linear models and LASSO under 10-fold cross validation.  
118 These models were used to assess the predictive stability of WQI\_P and WA\_WQI (Gao et al.  
119 2020; Gupta and Mishra 2023; Islam Khan et al. 2022; Uddin et al. 2022, 2023a, 2023b, 2023a;  
120 Xu et al. 2024)

121 The objective of this study is to determine whether a PCA based index can provide a more  
122 objective, reliable and predictive assessment of river water quality than WA\_WQI in the Skikda

123 region of Algeria. More specifically, (i) construct WQI\_P using correlation screening and PCA  
124 based weighting, (ii) compare WQI\_P and WA\_WQI using continuous scores, class  
125 assignments, confidence intervals and PREI and (iii) evaluate predictive performance using  
126 multiple machine learning algorithms. A quasi Gamma GLM with log link is used to interpret  
127 the influence of key chemical parameters. The central question is whether PCA derived weights  
128 can enhance transparency, predictive accuracy and class stability in river water quality  
129 assessment.

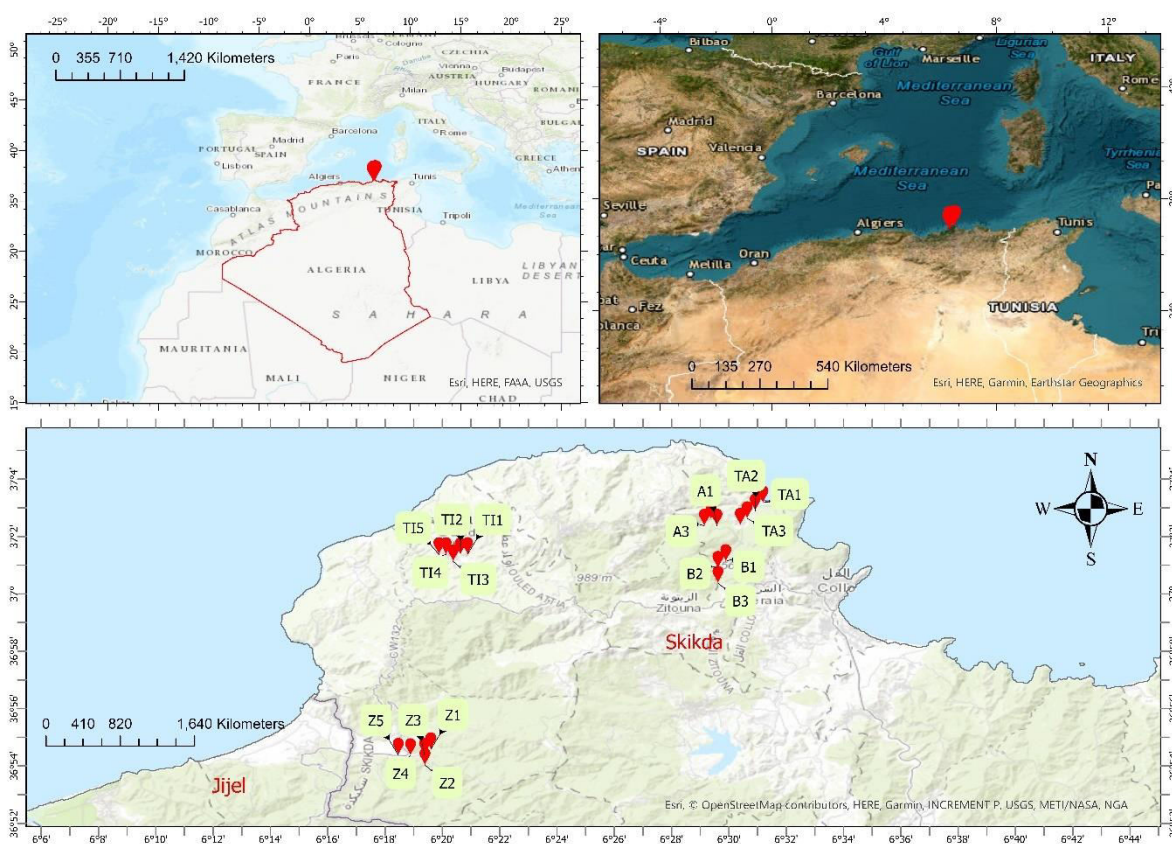
## 130 **2. Materials and methods**

### 131 ***2.1 Study area***

132 Twenty stations were sampled (Fig. 1) in the northwestern region of Skikda, an area with  
133 relatively stable climatic conditions but marked by local hydrological heterogeneity such as  
134 variations in flow velocity, channel width and depth, vegetation cover, and streambed  
135 morphology. The sites were distributed across five watercourses and their tributaries to reflect  
136 this variability. Specifically, four stations were placed along Oued Tamanart, three on Oued  
137 Afensou, three on Oued Boudoudeh (“Khraouet”), and five each on Oued Tizaghbane and Oued  
138 Z’hour. This spatial design allows for a consistent regional assessment while capturing  
139 meaningful hydrological variability relevant to water quality and biodiversity patterns  
140 (Boucenna et al. 2023; Satour et al. 2024).

141 This region offers several advantages for evaluating water quality indices. The watercourses  
142 show a relatively stable flow throughout the year, providing a reliable basis for temporal  
143 assessments. The 20 selected stations are located in an area with homogeneous and sustained  
144 rainfall, ensuring consistent environmental conditions (Boudiaf et al. 2022). Moreover, this  
145 agricultural region is under strong anthropogenic pressure, especially due to intensive irrigation,  
146 making it ideal for testing the indices' responsiveness to environmental stressors (Boucenna et  
147 al. 2023; Boudiaf et al. 2022; Kheloufi Attou et al. 2025). The compactness of the study area

148 also supports high-resolution analysis, helping us assess whether the newly developed WQI\_P  
 149 provides stable and confident classifications, or if minor variations in water quality can cause  
 150 major shifts in its outputs, especially when compared to the traditional WA\_WQI. These  
 151 northwestern basins, including Tamanart, Tizaghbane, and Zhor, receive up to 130 mm of  
 152 monthly precipitation (see Supplementary Materials, Fig. S1). The area supports intensive  
 153 agriculture, making local communities highly dependent on reliable and high-quality water  
 154 sources.



155  
 156 **Fig. 1.** Study area map with location of sampling sites, there are 20 stations in total: TI refers to stations at Oued  
 157 Tizaghbane, TA refers to Oued Tamanart, Z refers to Oued Z'Hour, B refers to Oued Boudoudeh, and A refers to  
 158 Oued Afensou.  
 159

160 **2.2. Sampling collection**

161 Four sampling campaigns were conducted during October–November 2022, March–April  
 162 2023, October–November 2023, and March–April 2024, targeting seasonal variations in water  
 163 levels linked to precipitation patterns. Low-flow dry periods were sampled in October–  
 164 November, while high-flow rainy periods were covered in March–April.

165 Quality assurance and quality control followed (Rodier et al. 2016). At each station, several  
166 field samples were collected in sterile 1.5 L glass bottles, stored in coolers at 4 °C, and  
167 transported with a documented chain of custody to the RNAMS laboratory (University of OEB,  
168 Algeria). In situ measurements of Temperature, pH, electrical conductivity, turbidity, total  
169 dissolved solids, salinity, dissolved oxygen, and oxygen saturation were obtained with a  
170 multiparameter meter (Multi 360i, WTW GmbH, Germany) after daily calibration against  
171 certified standards (Rodier et al. 2016). Laboratory analyses were performed in triplicate for  
172 each sample, and any result that did not meet predefined acceptance criteria was reanalyzed.

173 Chloride, sulfate, nitrate, bicarbonate (reported as alkalinity), calcium, magnesium, sodium,  
174 potassium, total hardness, calcium carbonate, ammonium, nitrite and orthophosphate were  
175 analyzed within three days following standard procedures (Rodier, 2016). Total hardness and  
176 the concentrations of calcium and magnesium were determined by complexometric titration  
177 with EDTA; chloride by the Mohr argentometric method; nitrate by the sodium-salicylate  
178 colorimetric method (AFNOR T90-012); ammonium and nitrite by UV-visible  
179 spectrophotometry (DR6000, Hach); sodium and potassium by flame photometry (PFP7,  
180 Jenway; ISO 9964-3); orthophosphate by the molybdenum-blue colorimetric method; sulfate  
181 by turbidimetry with barium chloride; and bicarbonate by acid titration of alkalinity. Calcium  
182 carbonate was reported as calcium-carbonate equivalents derived from the alkalinity and  
183 hardness determinations (Rodier et al. 2016).

184 In total, 21 physico-chemical variables were measured, representing all the parameters available  
185 for analysis. This comprehensive dataset was essential for testing our new variable selection  
186 strategy and assessing its effectiveness in identifying the most relevant indicators for water  
187 quality assessment; these parameters were selected based on their frequent use in regional and  
188 international water quality assessments.

### 189 2.3. Calculation of water quality indexes

190 Accordingly, the following sections detail the calculation methods for both WA\_WQI and  
191 WQI\_P, highlighting the differences in parameter selection, weighting attribution, and score  
192 aggregation—essential aspects for evaluating the robustness and simplicity of each index.  
193 These distinctions are particularly important given the known limitations in traditional WQI  
194 models, such as parameter eclipsing (where dominant variables can obscure the influence of  
195 others), subjectivity in weight assignment, and metaphoring or ambiguous interpretations of  
196 index values (Uddin et al. 2023b, 2023c, 2021b). By addressing these challenges, the WQI\_P  
197 aims to improve transparency, reduce uncertainty, and ensure a more representative assessment  
198 of water quality conditions(Uddin et al. 2023b).

#### 199 2.3.1. Calculation of the weighted arithmetic water quality indexes

200 This study focused on the weighted arithmetic water quality index for being the most widely  
201 use freshwater quality index in many of scientist in Algeria and in the study region (Allaoua et  
202 al. 2024; Fartas et al. 2022; Tyagi et al. 2013).This method is based on standard values and  
203 recommended permissible limits defined for each parameter, as specified by the SEQ-Eau water  
204 quality assessment system (Cadilhac 2003).

205 The Weighted Arithmetic Water Quality Index (WA\_WQI) is calculated through the following  
206 steps:

$$207 \quad K = \frac{1}{\sum_{i=1}^n C_i} \rightarrow w_i = \frac{K}{C_i} \rightarrow Q_i = \frac{V_i - C_{ideal}}{C_i - C_{ideal}} \times 100 \rightarrow WA\_WQI = \frac{\sum_{i=1}^N Q_i \times w_i}{\sum_{i=1}^N w_i} \quad (1)$$

208 K is a constant that ensures the sum of all weights equals 1; it is the inverse of the sum of the  
209 permissible limits of all considered parameters. The quality rating for each parameter, denoted  
210 as  $Q_i$ , was determined using the following formula: Where  $V_i$  represents the actual  
211 concentration of the nth parameter, and  $C_i$  corresponds to the permissible limit for that  
212 parameter.  $C_{ideal}$  Represents the ideal value of the parameter in pure water, where  $C_{ideal} = 0$   
213 (except for pH = 7 and DO = 14.6 mg/L).

214 *2.3.2. Calculation of the new weighted arithmetic water quality index (WQI<sub>P</sub>)*

215 Principal Component Analysis (PCA) was first applied to reduce the dimensionality of the  
216 dataset. Components with eigenvalues greater than 1 were retained according to the Kaiser rule  
217 (Mishra et al. 2017; Rojas-Valverde et al. 2020).

218 To improve clarity, the parameter selection process for constructing WQI<sub>P</sub> followed the  
219 structured steps below:

220 Step 1 – PCA dimensionality reduction

221 PCA was applied to the full set of water quality parameters, and components with eigenvalues  
222 greater than 1 were retained.

223 Step 2 – Selection of candidate parameters for each PCA component

224 For each retained component, variables were ranked according to their PCA percentage  
225 contributions as shown in Table 2. The two variables with the highest contributions were  
226 selected as the initial candidate parameters

227 Step 3 – Redundancy control using correlation and VIF

228 Pearson correlation was computed among the initially selected variables. Pairs with an absolute  
229 correlation coefficient greater than 0.70 were considered redundant.

230 For redundant pairs:

231 If both variables had VIF values greater than 5, the variable with the lower PCA contribution  
232 was replaced by the third most contributing variable from the same component.

233 This procedure ensured that each component retained a representative variable without high  
234 multicollinearity.

235 Step 4 – Final multicollinearity check

236 VIF was recalculated for all remaining variables. Only variables with VIF values below 5 were  
237 retained in the final set.

238 Step 5 – Definition of contribution based weights ( $w_{ij}$ )

239 For every selected parameter  $i$  belonging to component  $j$ , the weight  $w_{i_j}$  was derived directly  
240 from its PCA percentage contribution to that component.  $w_{i_j}$  is not the PCA loading and not  
241 the squared loading. It corresponds to the normalized percentage contribution of variable  $i$  to  
242 component  $j$ , exactly as reported in Table 2 (FactoMineR output `var$contrib`).

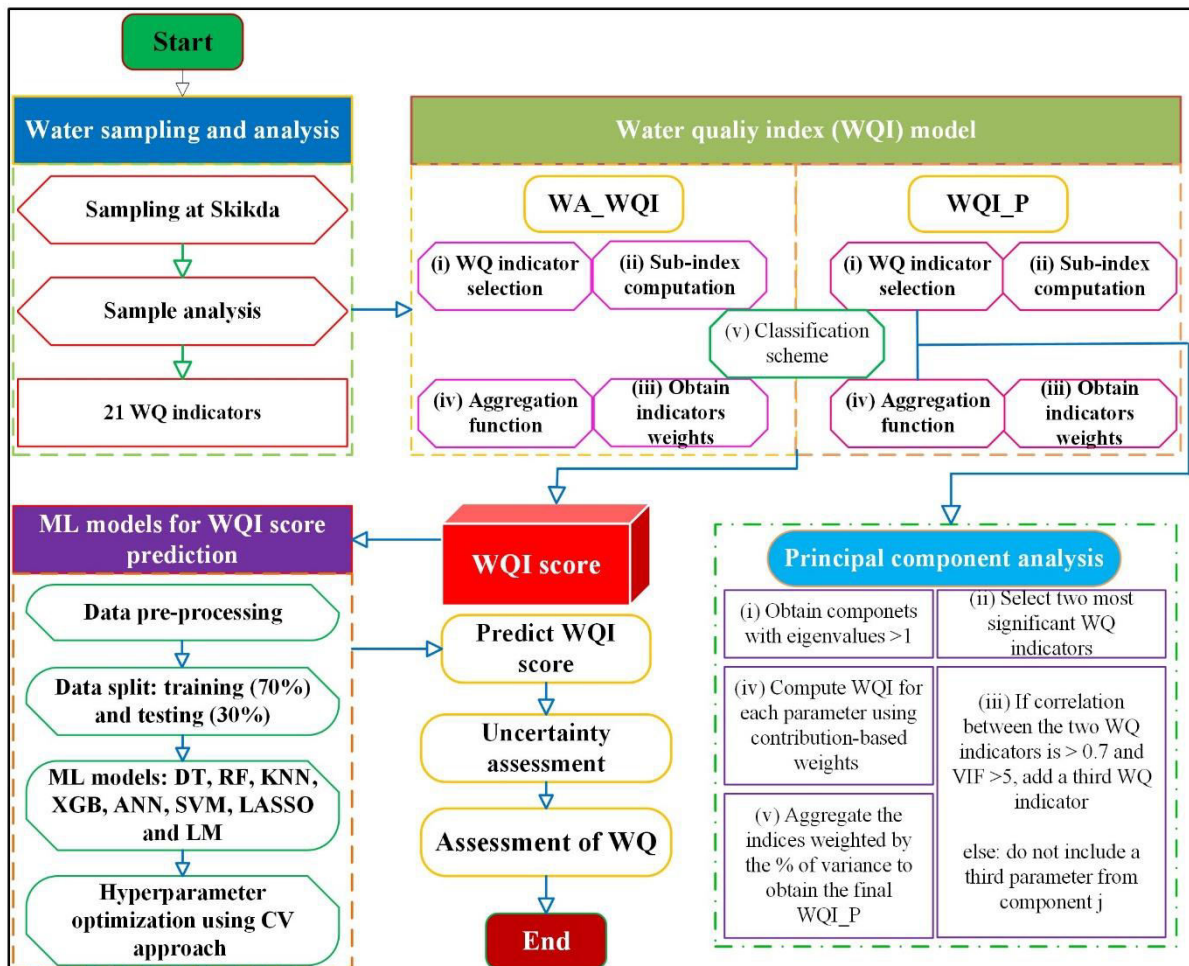
243 Weights within each component were normalized so that the sum of weights for the selected  
244 variables in component  $j$  equals 1.

245 This approach ensures that each parameter's influence reflects its relative statistical importance  
246 within its principal component.

247 After parameter selection, water quality indices were calculated for each retained PCA  
248 component using the selected parameters and their contribution-based weights. Unlike  
249 traditional approaches that rely on fixed or empirical weights—often based on regulatory limits  
250 or expert judgment—this method uses PCA contribution values as weights. This data-driven  
251 strategy ensures that each parameter's influence on water quality reflects its actual importance  
252 in the dataset (Parween et al. 2022).

253 Although the computation of  $WQI_P$  involves two aggregation levels, first within each PCA  
254 dimension ( $WQI_{D_j}$ , PCA Dimension  $j$ ) and then across dimensions weighted by explained  
255 variance, this structure is necessary to preserve essential hydrological and statistical  
256 information. Each PCA dimension represents a distinct hydrochemical gradient, grouping  
257 parameters that covary for different biogeochemical or physical reasons. Computing  
258  $WQI_{D_j}$  separately maintains the internal coherence of each gradient, prevents unrelated  
259 gradients from being mixed prematurely, and avoids distortion caused by extreme loadings. The  
260 second aggregation step, weighted by the proportion of variance explained by each component,  
261 ensures that dominant gradients influence the final index more than weaker ones. This  
262 hierarchical structure reduces eclipsing effects, enhances interpretability, and preserves the  
263 logic of PCA dimensionality reduction, which justifies the methodological complexity of

264 WQI\_P compared with simpler single step PCA integrations. A PCA-direct index is not suitable  
 265 because PCA scores are unitless and only represent variance, not water-quality classes or  
 266 regulatory thresholds. Combining all components into a single PCA score mixes different  
 267 hydro-chemical gradients and can hide important but low-variance parameters.  
 268 Finally, the developed WQI\_P was obtained by aggregating these indices, each weighted by the  
 269 percentage of variance explained by its corresponding component. This approach enhances  
 270 objectivity and reduces the biases associated with conventional weighting systems, providing a  
 271 more robust and accurate assessment of water quality (Nath Roy et al. 2024). All  
 272 of these steps are summarized in Fig. 2.



273 Fig. 2. Graphical summary of WQI\_P calculation using PCA-based parameter selection and weighting.  
 274  
 275

276 Calculation steps for WQI\_P as provided in Figure. 2:

277 Step 1:

278 The normalization constant for component  $j$  is defined as:

$$279 \quad K_j = \frac{1}{\sum_{i=1}^N C_{ij}} \quad (2)$$

280 The weight assigned to parameter  $i$  within component  $j$  is then calculated as:

$$281 \quad w_{ij} = \sum_{i=1}^N \frac{K}{C_{ij}} \quad (3)$$

282 The quality rating for parameter  $i$  is computed as:

$$283 \quad Q_i = \frac{V_i - C_{ideal}}{C_i - C_{ideal}} \times 100 \quad (4)$$

284 Finally, the water quality index for component  $j$  (PCA dimension  $j$ ) is calculated as:

$$285 \quad WQI_{Dj} = \frac{\sum_{i=1}^N Q_i \times w_{ij}}{\sum_{i=1}^N w_{ij}} \quad (5)$$

286 Where:  $C_{ij}$ : the contribution of parameter  $i$  to principal component  $j$ ,

287  $w_{ij}$ : the weight of parameter  $i$  for component  $j$ ,

288  $V_i$  is the observed value of parameter  $i$ ;

289  $C_i$  is the permissible standard value of parameter  $i$ ;

290  $C_{ideal}$  is the ideal value of the parameter;

291  $WQI_{Dj}$ : The water quality index calculated for principal component  $j$ . (that is, PCA Dimension  $j$ .)

293 Step 2: Aggregate component indices to compute overall  $WQI_P$ :

294 Compute the overall  $WQI_P$  by combining the  $WQI_{Dj}$  values using the proportion of variance explained by each component:

$$296 \quad WQI_P = \sum_{j=1}^R \left( \frac{Var_j}{\sum_{D=1}^R Var_D} \times WQI_{Dj} \right) \quad (6)$$

297  $Var_j$ : Variance explained by principal component  $j$ .

298  $\sum_{D=1}^R Var_D$ : Total variance explained by the components retained according to the Kaiser rule  
299 (eigenvalue > 1)

300 Finally, the calculated WQI is categorized into predefined ranges to evaluate water quality, as  
301 presented in Table 1.

302 Table 1. Classification Scheme of Water Quality Based on WQI Values (Tyagi et al. 2013).

WQI Range	Water quality class
$0 < \text{WQI} < 25$	Excellent
$25 \leq \text{WQI} < 50$	Good
$50 \leq \text{WQI} < 75$	Poor
$75 \leq \text{WQI} < 100$	Very poor
$\text{WQI} \geq 100$	Unsuitable for drinking

303

#### 304 **2.4. Supervised machine learning approach**

305 Supervised machine learning methods were implemented in three steps. First, all parameters  
306 were standardized. Second, machine learning algorithms were employed eight machine  
307 learning algorithms to predict the two water quality indices (WA\_WQI and WQI\_P) calculated  
308 for each of the 159 samples obtained from Algerian rivers. Third, eight algorithms were stacked  
309 the eight algorithms to combine and summarize their predictive outputs into a single model.  
310 Finally, the Root Mean Square Error (RMSE) were calculated, Mean Absolute Error (MAE),  
311 and the coefficient of determination ( $R^2$ ), and constructed a Taylor diagram to evaluate whether  
312 the newly developed WQI\_P improves predictive performance compared to the traditional  
313 WA\_WQI across individual algorithms and the stacked model.

##### 314 **2.4.1. Parameter standardization**

315 Standardizing variables is a critical step when predicting the water quality index (Banda and  
316 Kumarasamy 2020; Gazzaz et al. 2012; Khouri and Bashar Al-Moufti 2022; Uddin et al. 2022).  
317 The research utilized the z-score standardization, where:

318 
$$z = \frac{x_i - \bar{x}}{\sigma} \quad (7)$$

319 Where,  $z$  denotes the standardized score,  $\bar{x}$  represents the mean of the data,  $x_i$  refers to the  $i$ th  
320 data point, and  $\sigma$  signifies the standard deviation. All machine learning algorithms in  
321 subsequent sections were trained on these standardized variables.

#### 322 *2.4.2. Estimation of water quality indexes using machine learning algorithms*

323 Supervised learning algorithms were applied to estimate the Weighted Arithmetic Water  
324 Quality Index (WA\_WQI) and the newly developed PCA-based water quality index (WQI\_P),  
325 both calculated based on eight selected environmental parameters. The eight environmental  
326 parameters used as inputs for all machine learning models correspond to the final, non-  
327 redundant variables selected during the construction of WQI\_P (Section 2.3.2). These  
328 parameters were identified through PCA contribution ranking, correlation filtering, and VIF  
329 analysis. To ensure a fair comparison between indices, the same eight parameters were also  
330 used as predictors for WA\_WQI, even though WA\_WQI can theoretically incorporate all 21  
331 variables. Using an identical input set guarantees that differences in predictive performance  
332 arise from the index formulation itself and not from unequal model inputs.

333 The analysis utilized two datasets, corresponding respectively to each index. These indices  
334 were calculated for a total of 159 samples collected from three rivers in the Skikda region,  
335 providing a comprehensive dataset to evaluate the algorithms' performance (Bruce et al. 2020;  
336 Gazzaz et al. 2012; Hassan et al. 2021; Jayaraman et al. 2024; Kassambara 2018; Mahesh 2019;  
337 Uddin et al. 2023c; Uddin, Nash, Rahman, et al. 2022). We implemented eight different  
338 machine learning algorithms. Decision Tree algorithm (DT) was calculated using the `rpart`  
339 method included in the R package `rpart` (Bruce et al. 2020); The hyperparameters were fine-  
340 tuned, and particularly the complexity parameter (`cp`), to minimize RMSE and control tree size.  
341 Random Forest algorithm was calculated using the R package `randomForest` (Berrar 2018;  
342 Kassambara 2018; Mahesh 2019), employing a grid search to optimize the `mtry` parameter. The  
343 Artificial Neural Network (ANN) algorithm was built using the `nnet` method as implemented

344 in the R package caret(Bruce et al. 2020; Kassambara 2018), with a linear output specified for  
345 regression, and a grid search was performed to refine hyperparameters. XGBoost algorithm was  
346 trained using the xgbTree method, implemented in the R package xgboost(Bruce et al. 2020),  
347 testing various configurations for the hyperparameters nrounds, max\_depth, and eta. K-Nearest  
348 Neighbors (KNN) was optimised by determining the best kvalue through cross-validation  
349 (Bruce et al. 2020; Kassambara 2018), using the R package. In parallel, a Support Vector  
350 Machine (SVM) with a linear kernel was trained using the svmLinear method from the R  
351 package e1071, with data standardized using z-score scaling (center, scale) and  
352 hyperparameters optimized via tuneLength = 10. A Linear Regression (LM) model was also  
353 trained using the lm method under 10-fold cross-validation, serving as a classical regression  
354 baseline. Lastly, a Lasso Regression model was developed using the glmnet method from the  
355 R package glmnet, applying L1 regularization with automatic variable selection; preprocessing  
356 steps (center, scale) were used, and the penalty parameter (lambda) was optimized using  
357 tuneLength = 10. All eight models were trained on 70% of the dataset and tested on the  
358 remaining 30%, with performance evaluated using RMSE, MAE, and R<sup>2</sup> metrics.

### 359 2.4.3. Stacking

360 To further enhance predictive accuracy, a stacking ensemble was implemented after training  
361 the eight base learners (DT, RF, ANN, XGB, KNN, SVM, LM, and LASSO). For each index  
362 (WA\_WQI and WQI\_P), out-of-fold predictions from the base models were generated under  
363 the same 10-fold cross-validation scheme and assembled into a new matrix of meta-features.

364 A linear regression model was then used as the meta-learner to combine these predictions. This  
365 model learns an optimal weighted combination of the outputs of the eight algorithms, producing  
366 the final stacked prediction. The choice of a linear meta-learner follows common practice in  
367 regression-based stacking, as it provides a transparent and stable mechanism for aggregating  
368 heterogeneous model predictions.

369 The stacking procedure was applied identically to both indices to ensure that ensemble  
370 performance comparisons between WA\_WQI and WQI\_P were unbiased and strictly model-  
371 dependent rather than influenced by differences in the ensembling process.

#### 372 2.4.4. Cross-validation of the results

373 Model performance was evaluated using 10-fold cross-validation, a widely adopted method for  
374 assessing machine learning algorithms (Berrar 2018; Sweet et al. 2023). Performance was  
375 assessed based on Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the  
376 coefficient of determination ( $R^2$ ), which are widely recognized as standard metrics for  
377 evaluating regression models in water quality prediction tasks. These metrics provide  
378 complementary insights: RMSE penalizes large errors more heavily, MAE offers a direct  
379 interpretation of average prediction error, and  $R^2$  explains the proportion of variance captured  
380 by the model. Their combined use ensures a balanced and comprehensive assessment of model  
381 accuracy and robustness (Berrar 2018; Bruce et al. 2020; Uddin et al. 2023a, 2023b, 2023c,  
382 2022).

383 The criteria for algorithm evaluation are assessed as follows:

$$384 \quad RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (8)$$

$$385 \quad MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (9)$$

$$386 \quad R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (10)$$

387 Lower values of RMSE and MAE indicate higher algorithm accuracy, as this reflects smaller  
388 differences between the predicted and actual values, which correspond to better precision in the  
389 algorithm. On the other hand, the coefficient of determination  $R^2$  measures the proportion of  
390 variance in the dependent variable explained by the algorithm. A higher  $R^2$  value, closer to 1,  
391 indicates a better algorithm fit, while values closer to 0 suggest that the algorithm is less

392 effective in explaining the variance, signaling weaker performance. Therefore, an optimal  
393 algorithm is characterized by low RMSE and MAE values, along with a high  $R^2$  value,  
394 indicating both accuracy and a strong relationship between the predictors and the target  
395 variable. The RMSE measures prediction accuracy, with a lower RMSE indicating better  
396 performance. The correlation coefficient, close to 1, indicates a strong linear relationship  
397 between predictions and observations, while the standard deviation assesses the dispersion of  
398 prediction errors, with a lower standard deviation reflecting greater consistency.

399 Finally, the research plotted a Taylor diagram to comprehensively evaluate and visualize the  
400 predictive performance of different machine learning algorithms applied to the two water  
401 quality indices (WA\_WQI and WQI\_P)(Sajib et al. 2024; Singh et al. 2021). This diagram  
402 consolidates the three-performance metrics of RMSE, correlation coefficient, and standard  
403 deviation, providing an integrated view of both accuracy and consistency.

404 By comparing the predictive ability of the models across the two indices, the Taylor diagram  
405 facilitates the identification of the most accurate and consistent models, supporting a more  
406 comprehensive evaluation of water quality prediction methods.

407 Machine learning models were not employed to validate the ecological relevance of the indices,  
408 as both WA\_WQI and WQI\_P are deterministic functions of the same input variables.  
409 Consequently, high predictive performance was not interpreted as evidence of ecological  
410 validity. Instead, machine learning was deliberately used as a diagnostic framework to evaluate  
411 the internal numerical behavior of the indices, including stability, sensitivity to parameter  
412 perturbations, and error-propagation characteristics. Consistent predictability across  
413 heterogeneous algorithms is interpreted as an indicator of lower internal noise, reduced  
414 sensitivity to parameter fluctuations, and a more coherent input–output structure. In this  
415 context, high  $R^2$  values reflect numerical robustness and structural stability rather than

416 ecological meaning, consistent with recent studies that use machine learning to assess WQI  
417 reliability and uncertainty rather than environmental causality.

### 418 **2.5 Model(s) Uncertainty Analysis - Gaussian Monte Carlo Simulation approaches**

419 To quantify the propagation of predictive uncertainty across the WQI models, a Gaussian Monte  
420 Carlo simulation framework was employed. Because, several recent water research have  
421 reported that this approach could be effective compared to the other techniques (Diganta et al.,  
422 2025; Khan et al., 2025; Sajib et al., 2023; Uddin et al., 2024b). Details of the framework can  
423 be found in Uddin et al., (2023a). Model residuals were assumed to follow a normal probability  
424 distribution:

$$425 \quad \varepsilon \sim \mathcal{N}(\mu, \sigma^2) \quad (11)$$

426  $\varepsilon = \hat{y}_i - y_i$  represents the residual between the model prediction  $\hat{y}_i$  and the observed WQI  
427 value  $y_i$ ;

428  $\mu$  and  $\sigma$  are the empirical mean and standard deviation of residuals estimated from calibration  
429 data.

430 For each model,  $N = 10,000$  random perturbations were generated by Following fonction

$$431 \quad \varepsilon_i^{(k)} = \mu + \sigma Z_i^{(k)}, \quad Z_i^{(k)} \sim \mathcal{N}(0, 1) \quad (12)$$

432 where  $\varepsilon_i^{(k)}$  refers the simulated residual for the  $i^{\text{th}}$  observation in the  $k^{\text{th}}$  iteration;  
433  $\mu$  and  $\sigma$  are the empirical mean and standard deviation of residuals;  
434 and  $Z_i^{(k)}$  represents a standard-normal random variate.

435 Each perturbation was added to the deterministic prediction  $\hat{y}_i$  to form a synthetic realization as  
436 follows:

$$437 \quad y^{i(k)} = y^i + \varepsilon_i(k) \quad (13)$$

438 This process was repeated 10,000 times per model to produce a probabilistic ensemble of WQI  
439 estimates used for deriving uncertainty metrics ( $R^2$ , RMSE, MAE, Bias) and 95% confidence  
440 intervals.

### 441 3. Results

#### 442 3.1. Statistical summary of environmental parameters

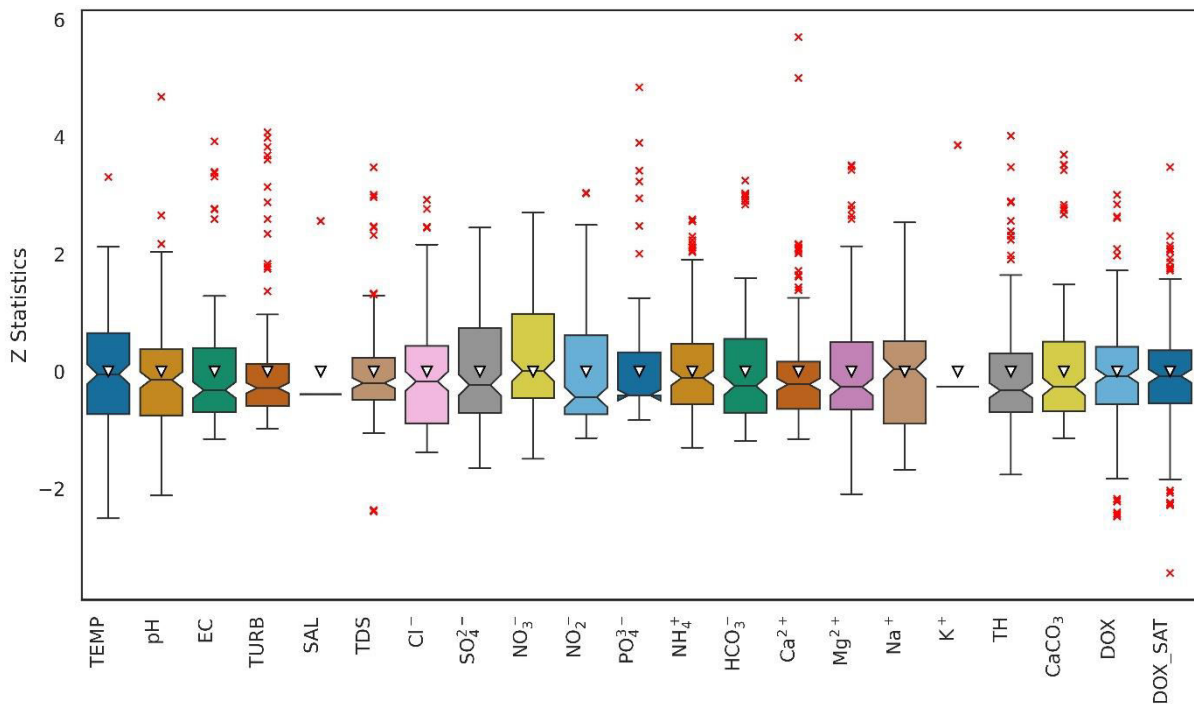
443 A total of 21 physico-chemical parameters were measured in three rivers of the northwest  
444 Skikda region (Nuzzo 2016; Wickham et al. 2024). Fig. 3 shows boxplots of standardized  
445 values, highlighting variability and enabling comparison among parameters, while descriptive  
446 statistics are presented in Figure S6 (supplementary material). The observed variability reflects  
447 the heterogeneous nature of river water quality and supports the application of PCA for data  
448 reduction and variable selection.

449 Water temperature ranged from 9 to 27.60 °C (mean = 17.01 °C), consistent with local climatic  
450 conditions. pH values (6.65–9.58; mean = 7.56) were generally within SEQ-Eau limits, with  
451 minor exceedances likely related to photosynthetic activity. Electrical conductivity remained  
452 low (181–672  $\mu\text{S}/\text{cm}$ ; mean = 292.9  $\mu\text{S}/\text{cm}$ ), indicating weak mineralization, whereas turbidity  
453 reached 4.3 NTU, exceeding guideline values and suggesting suspended matter inputs from  
454 runoff or anthropogenic sources (Uddin et al. 2023c, 2022).

455 Nitrate concentrations were well below SEQ-Eau standards (mean = 4.975 mg/L; max = 14  
456 mg/L), while nitrite occasionally exceeded the 0.03 mg/L threshold (max = 0.0792 mg/L),  
457 indicating possible recent organic pollution or microbial activity (Rodier et al. 2016).  
458 Orthophosphate and ammonium showed localized exceedances (max = 0.12 mg/L and 0.604  
459 mg/L, respectively), suggesting inputs from domestic, agricultural, or animal waste  
460 sources (Uddin et al. 2023b).

461 Dissolved oxygen conditions were favorable (7.17–9.55 mg/L; mean = 8.244 mg/L), with  
462 oxygen saturation consistently above 70%, reflecting good aeration and ecological conditions  
463 (Bouchra et al. 2025). Mineralization parameters remained within SEQ-Eau limits, indicating  
464 no significant salinization or industrial influence. Total hardness averaged 137.4 mg/L (13.74

465 °C) and peaked at 364 mg/L (36.4 °C), approaching but not exceeding critical thresholds, likely  
 466 reflecting the limestone-dominated geology of the region.

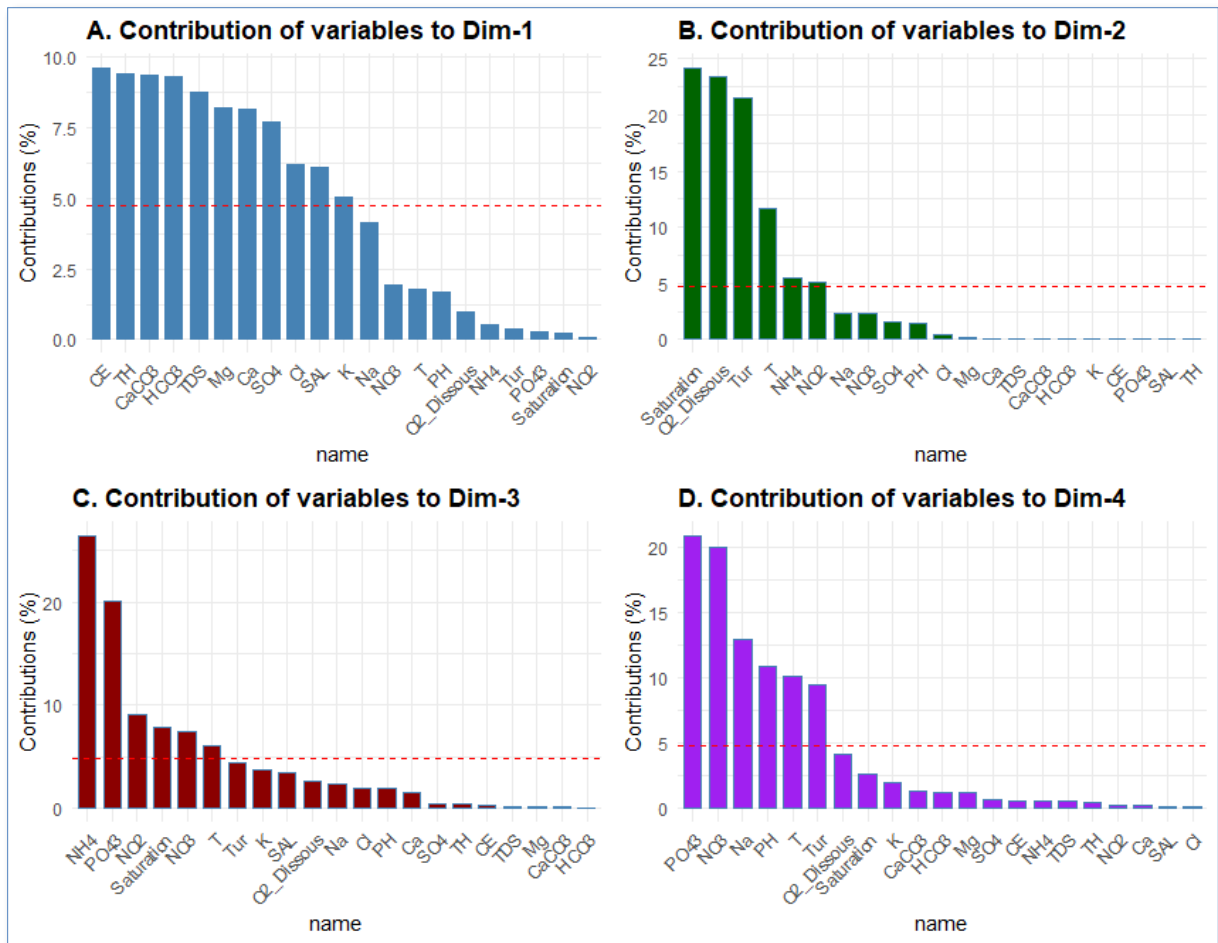


467  
 468 **Fig. 3.** Environmental parameters in Skikda river water samples.  
 469

470 **3.2. Results on the water quality indexes**

471 The research calculated six water quality indices for each 159 samples, including the traditional  
 472 water quality index (WA\_WQI), four water quality indexes corresponding to each PCA  
 473 dimensions (WQI\_Dim1-Dim4) and a summarized water quality index (WQI\_P).

474 After principal component analyses, and following Kaiser’s rule, four dimensions were  
 475 retained, hereafter referred to as Dim 1- Dim 4 (Fig. 4), and the following eight uncorrelated  
 476 parameters were retained for subsequent analyses: conductivity, turbidity, nitrates, nitrites,  
 477 orthophosphates, ammonium, saturation rate, and dissolved oxygen (Table 2; Figure S2), the  
 478 biplots of PCA for axes 1 and 2, axes 1 and 3, and axes 1 and 4 are provided in the supplementary  
 479 materials (Figures S03, S04, and S05, respectively).



480  
481 **Fig. 4.** Variable Contributions to Principal Components: Dim-1 to Dim-4.  
482

483 **Table 2.** Contribution value of each parameter to the formation of the axes.

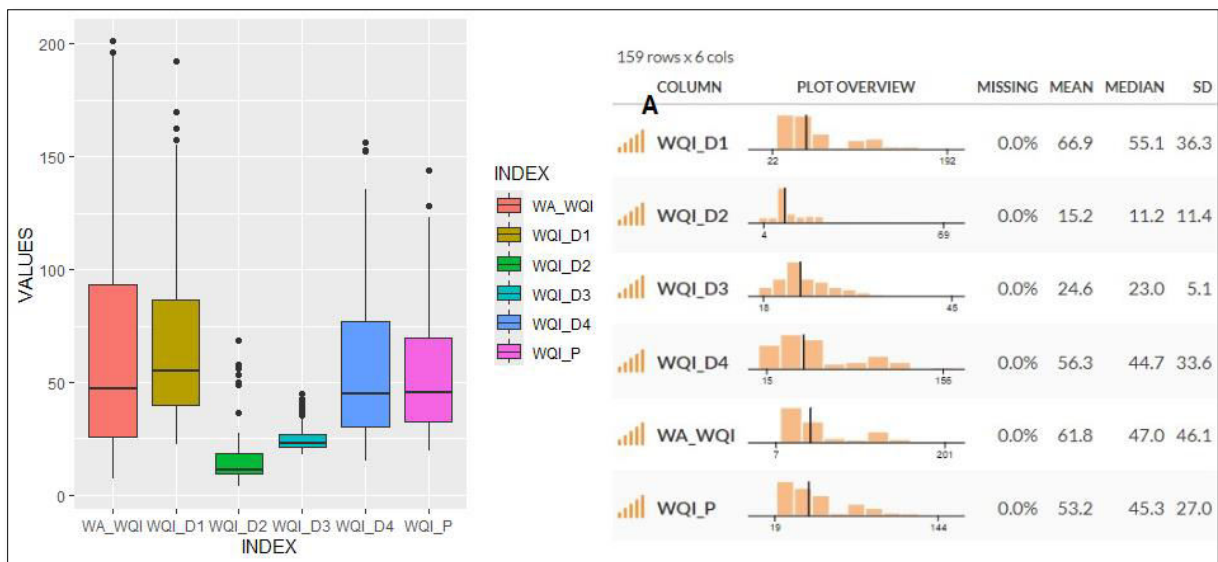
Paramètres	Dim-1	Dim-2	Dim-3	Dim-4
Conductivity	9,64	0,03	0,26	0,58
Turbidity	0,37	21,46	4,37	9,48
Nitrite	0,06	5,08	9,01	0,21
Nitrate	1,93	2,28	7,47	19,93
Ammonium	0,54	5,43	26,40	0,57
Saturation (%)	0,20	24,21	7,89	2,64
Dissolved O <sub>2</sub>	20,99	23,46	2,59	4,13
PO <sub>4</sub> <sup>3-</sup>	0,28	0,02	20,02	20,81

484

### 485 3.3. Statistical overview of water quality indices

486 The descriptive statistics graphs reveal a diverse distribution of values across the various Water  
487 Quality Indices (WQIs) (Fig. 5). WA\_WQI, WQI\_P, and WQI\_D1 exhibit relatively high mean  
488 values of  $61.8 \pm 46.1$ ,  $53.2 \pm 27.0$ , and  $66.9 \pm 36.3$ , respectively, indicating a generally poorer

489 water quality according to these particular measures. Conversely, the indices WQI\_D2,  
 490 WQI\_D3, and WQI\_D4 show lower values of  $15.2 \pm 11.4$ ,  $24.6 \pm 5.1$ , and  $56.3 \pm 33.6$ ,  
 491 respectively, which may suggest a comparatively better water quality. The spread of these  
 492 values, as measured by the standard deviation (SD), is particularly pronounced for WA\_WQI  
 493 ( $\pm 46.1$ ) and WQI\_D1 ( $\pm 36.3$ ), reflecting a greater variability around the mean. In contrast, the  
 494 indices WQI\_D2 ( $\pm 11.4$ ), WQI\_D3 ( $\pm 5.1$ ), and WQI\_D4 ( $\pm 33.6$ ) display more consistent  
 495 distributions of data points. Adjusting the parameter weights in these indices has led to more  
 496 stable and representative measures of overall water quality. For instance, the WA\_WQI index  
 497 is heavily influenced by nitrate and orthophosphate, with weights of 0.74 and 0.21, respectively,  
 498 making it highly sensitive to these two parameters and accounting for nearly 95% of its  
 499 variation. Consequently, WA\_WQI focuses primarily on these specific factors, whereas the  
 500 other indices provide a more comprehensive evaluation of water quality, encompassing  
 501 dimensions that WA\_WQI fails to adequately capture.

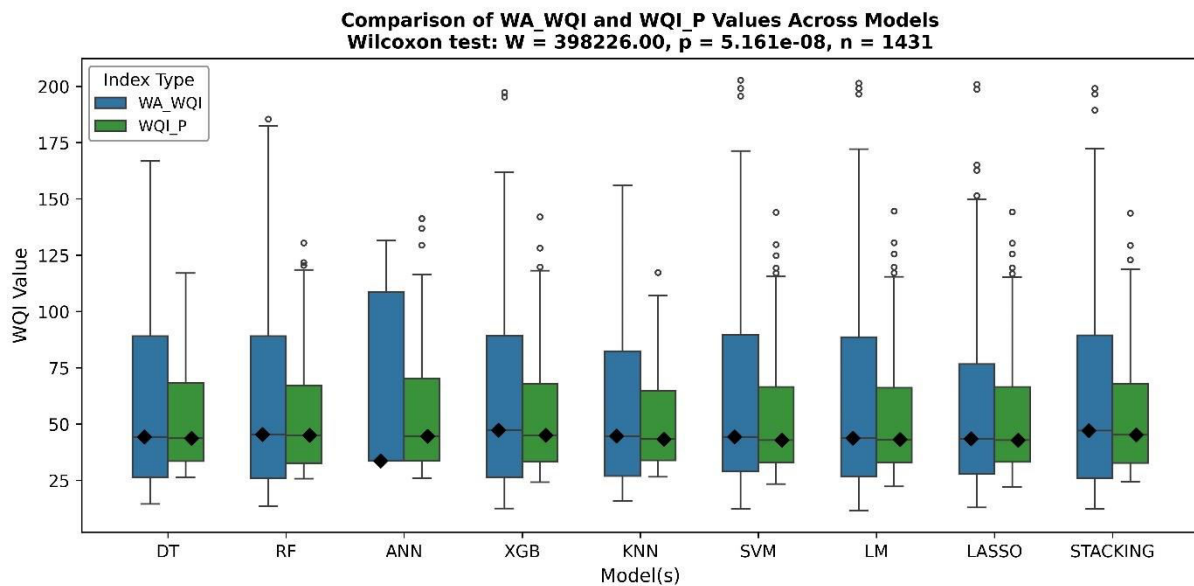


502  
 503 **Fig. 5.** Statistical Overview of Water Quality Indices (WA\_WQI, WQI\_D1, WQI\_D2, WQI\_D3, WQI\_D4,  
 504 WQI\_P) and Their Distribution.  
 505

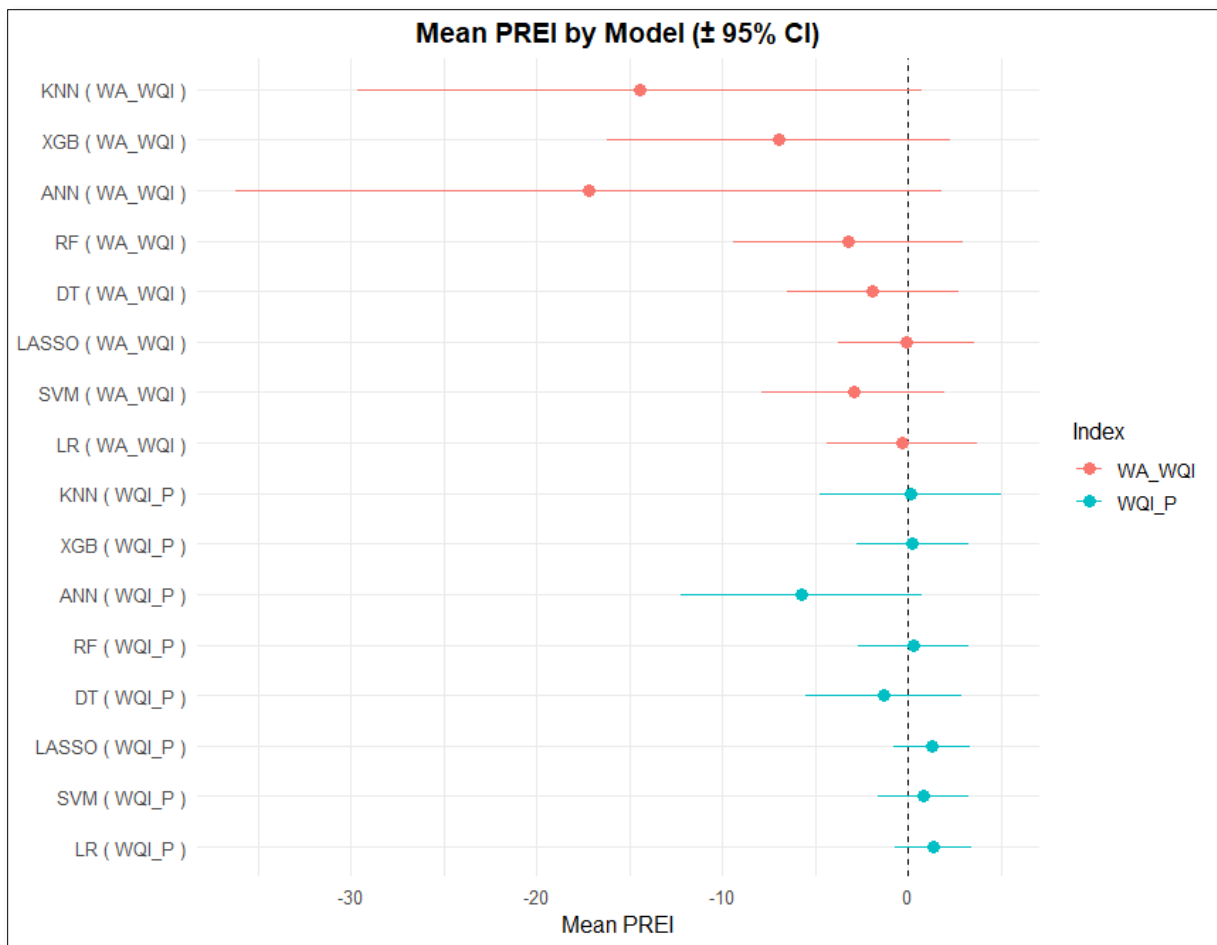
### 506 3.4. Distributional Comparison of WA\_WQI and WQI\_P Across Models

507 Fig. 6 shows boxplots comparing the WA\_WQI and WQI\_P values across all machine learning  
 508 models. Although both indices share a broadly overlapping range, the central tendency and

509 overall distributions are clearly distinct: WQI\_P exhibits consistently lower medians, reduced  
 510 spread, and fewer extreme values, reflecting its stricter and more stable scoring pattern.  
 511 Because the Shapiro–Wilk test indicated non-normal distributions for both indices ( $p < 0.05$ ),  
 512 a non-parametric Wilcoxon paired test was applied. The Wilcoxon test result ( $W = 398226$ ,  $p$   
 513  $= 5.161 \times 10^{-8}$ ) shows an extremely significant statistical difference between WA\_WQI and  
 514 WQI\_P. This confirms that the two indices do not produce equivalent score distributions. This  
 515 significant difference is fully consistent with the reclassification outcomes, WQI\_P yields more  
 516 conservative and balanced water-quality classes, demonstrating that the PCA-based index  
 517 produces a systematically different and more constrained evaluation of water quality compared  
 518 with WA\_WQI.  
 519

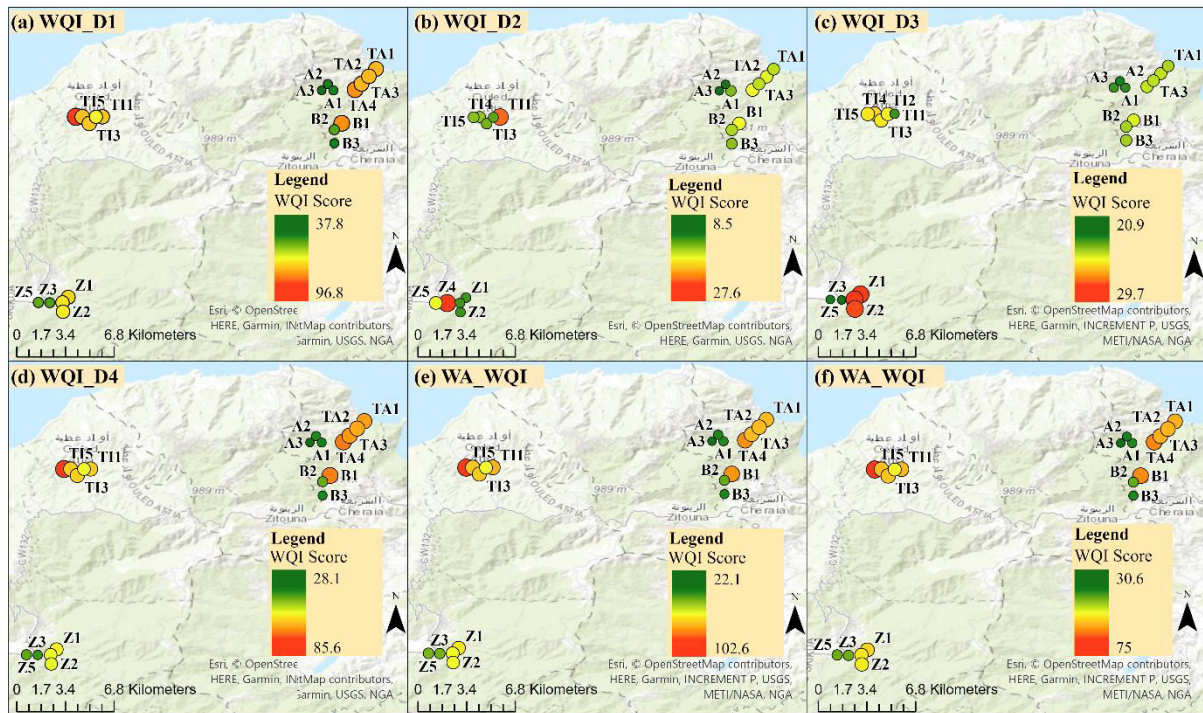


520  
 521 **Fig. 6.** Boxplot comparison of WA\_WQI and WQI\_P values using the Wilcoxon test.



522 **Fig. 7.** Mean prediction errors (PREI) of machine learning models for WQI\_P and WA\_WQI indices with 95%  
 523 confidence intervals.  
 524

525 Fig. 7 evaluates the predictive performance of a suite of machine learning models for both  
 526 indices. By presenting mean prediction errors together with 95% confidence intervals, this  
 527 figure highlights not only the relative accuracy of each model but also the degree of uncertainty  
 528 associated with its predictions. The results consistently show that WQI\_P is associated with  
 529 smaller errors and narrower confidence bounds across models, reflecting greater robustness and  
 530 stability. In contrast, WA\_WQI displays higher uncertainty and larger deviations. Taken  
 531 together, these findings suggest that WQI\_P provides a more reliable basis for further  
 532 applications, in line with recent recommendations that index selection should be guided by  
 533 predictive performance and uncertainty rather than by descriptive statistics alone (Uddin et al.  
 534 2023d, 2021a).  
 535



536  
537 **Fig. 8.** Spatial distribution of the WQI scores of various indices across the Skikda.  
538

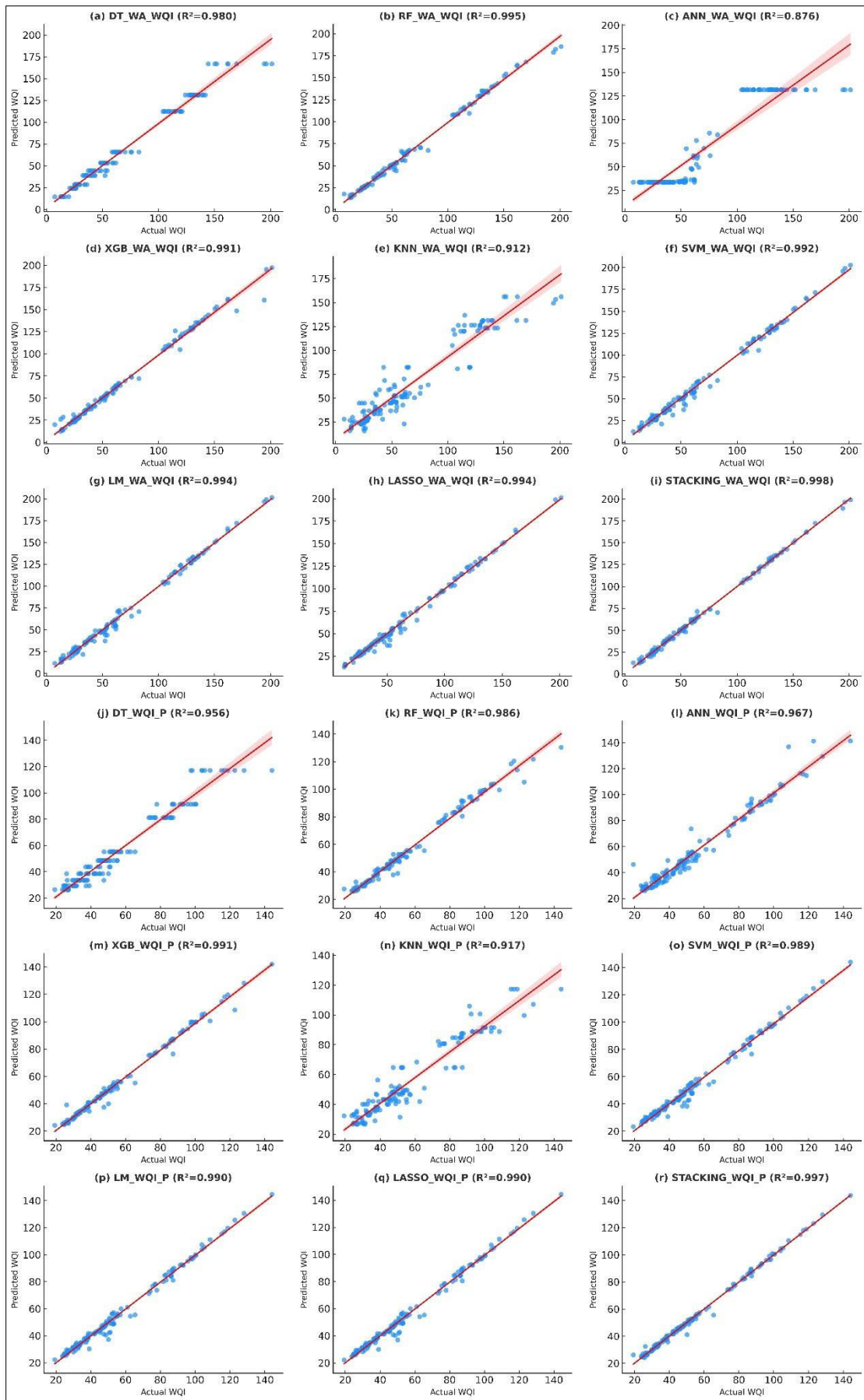
539 Figure 8 shows the spatial distribution of WQI scores at the different sampling sites in the study  
540 domain. As can be seen from fig. 8 that the WQI\_D1 showed maximum average WQI score  
541 (66.9) compared to the remaining models and the WQI\_D2 computed the minimum average  
542 WQI score (15.2). The average WQI score of 53.2 computed by the WQI\_P model was found  
543 lower when compared to the average WQI score (61.8) obtained from the WA\_WQI model  
544 (Fig. 8).

545 **3.5. Cross-validation of the predictive performance of WA\_WQI and WQI\_P indices**

546 Fig.9 compare predicted versus observed values for WA\_WQI and WQI\_P across eight base  
547 algorithms (DT, RF, ANN, XGB, KNN, SVM, LM, LASSO) and a stacked ensemble. In both  
548 panels, points cluster closely around the 1:1 line, indicating high predictive fidelity on the test  
549 data. Linear and margin-based approaches (LM, LASSO, SVM) show the tightest agreement  
550 with  $R^2$  typically around 0.98–0.99, while tree ensembles (RF, XGB) also perform strongly  
551 with only modest dispersion at the extremes. Single-tree DT and k-nearest neighbors display  
552 comparatively larger scatter and slightly lower  $R^2$ , though still within a high-performance range  
553 ( $>0.90$ ). The stacked ensemble tracks the best single algorithms, suggesting limited additional

554 gains. Taken together, the two figures show that both indices exhibit high predictive  
555 performance, with differences driven mainly by the choice of algorithm rather than the index  
556 itself.

557 However, these very high  $R^2$  values must be interpreted with caution. The dataset contains only  
558 159 samples and originally included 21 parameters, while several of the evaluated algorithms  
559 (ANN, XGB, RF) have high model capacity relative to the sample size. Such conditions create  
560 a substantial risk of overfitting, even under 10-fold cross-validation, especially if observations  
561 are not fully independent across seasons or sampling locations. Therefore, part of the strong  
562 predictive performance observed here likely reflects model optimism rather than purely  
563 generalizable signal. This highlights the need for validation using larger, more heterogeneous  
564 datasets in future studies.



566 **Fig. 9.** Predicted versus actual WA\_WQI values from nine machine learning models (DT, RF, ANN, XGB, KNN,  
567 SVM, LM, LASSO, STACKING) on the test data. The dashed black line shows the ideal 1:1 relationship, and the  
568 R<sup>2</sup> value indicates the prediction accuracy of each model.  
569

### 570 **3.6. Evaluation of Model Performance**

571 Fig.10 presents test-set results for eight algorithms (DT, RF, ANN, XGB, KNN, SVM, LM,  
572 LASSO) across two indices, WA\_WQI and WQI\_P, using RMSE, MAE and R<sup>2</sup>. Overall, the  
573 newly developed WQI\_P reduces prediction error for most models while maintaining high  
574 explained variance.

575 For RMSE, WQI\_P is lower than WA\_WQI for six of eight algorithms. The reductions are  
576 observed for ANN (7.93 vs 16.45), KNN (8.29 vs 14.81), XGB (4.58 vs 7.61), SVM (3.81 vs  
577 4.88), LM (3.41 vs 4.19) and LASSO (3.47 vs 4.12). DT and RF show slightly better RMSE  
578 with WA\_WQI, with DT at 6.90 vs 6.12 and RF at 4.75 vs 4.54.

579 For MAE, the pattern is consistent with RMSE. WQI\_P achieves lower MAE for ANN (4.69  
580 vs 12.06), KNN (5.84 vs 10.24), XGB (2.66 vs 4.10), SVM (2.58 vs 3.58), LM (2.15 vs 3.10)  
581 and LASSO (2.14 vs 2.98), while DT and RF are slightly better with WA\_WQI, with DT at  
582 5.32 vs 3.93 and RF at 3.05 vs 2.60.

583 For R<sup>2</sup>, both indices yield high values across all models, indicating strong predictive capability.  
584 WQI\_P is higher for ANN (0.93 vs 0.88) and KNN (0.92 vs 0.91), similar for XGB (0.97 vs  
585 0.97) and SVM (0.98 vs 0.98), and slightly lower than WA\_WQI for DT (0.93 vs 0.98), RF  
586 (0.97 vs 0.99), LM (0.98 vs 0.99) and LASSO (0.98 vs 0.99).

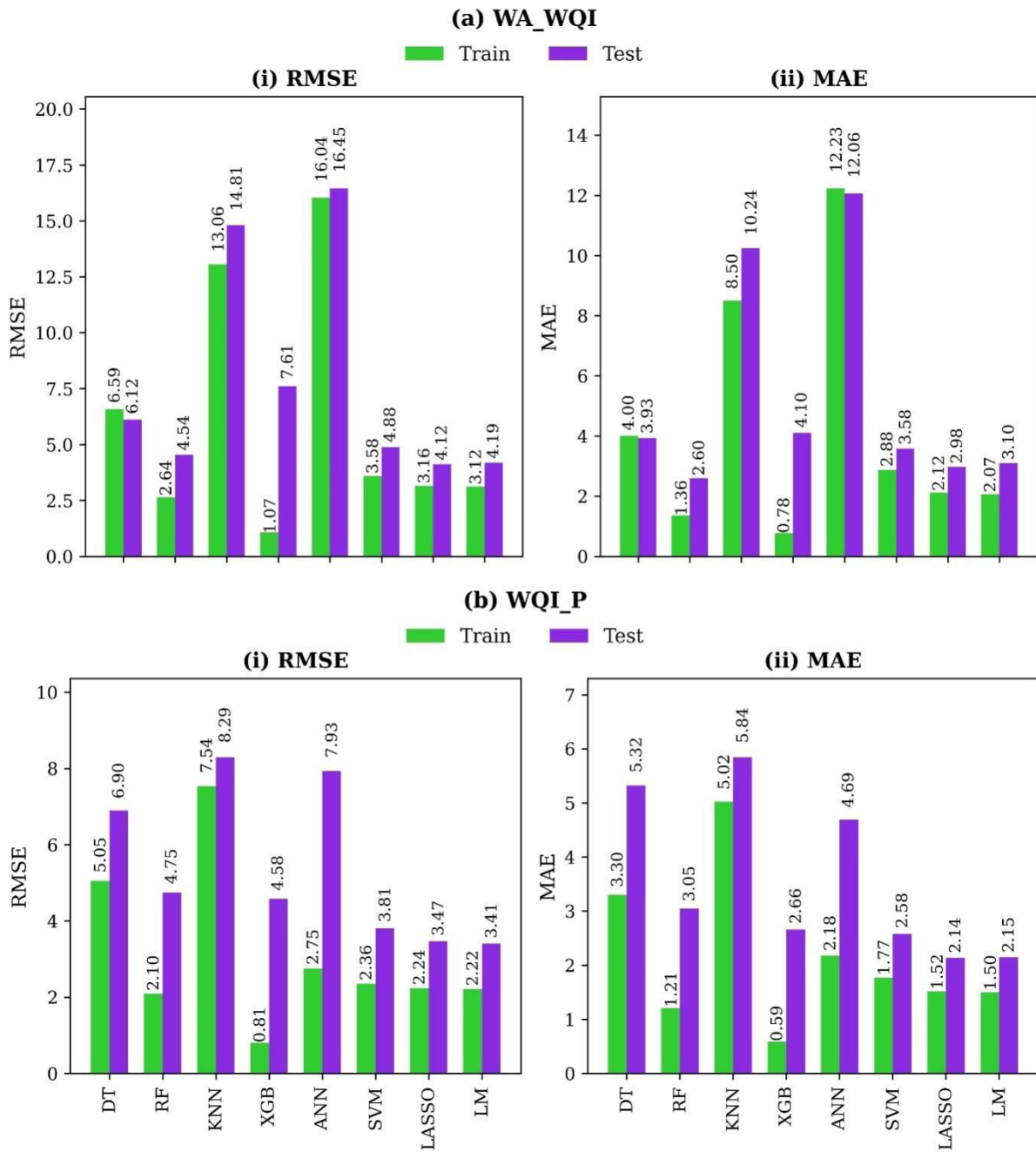
587 Taken together, the consistent reductions in RMSE and MAE for six of eight algorithms,  
588 combined with uniformly high R<sup>2</sup>, show that WQI\_P is overall the more performant and  
589 precise index on the test data.

590

591

592

593



594  
 595 **Fig. 10.** Comprehensive Evaluation of WQI Models: RMSE, MAE, and  $R^2$  on Test Data for WA\_WQI versus  
 596 WQI\_P Across DT, RF, ANN, XGB, KNN, SVM, LM, and LASSO.  
 597

598 **3.7. Model(s) uncertainty results**

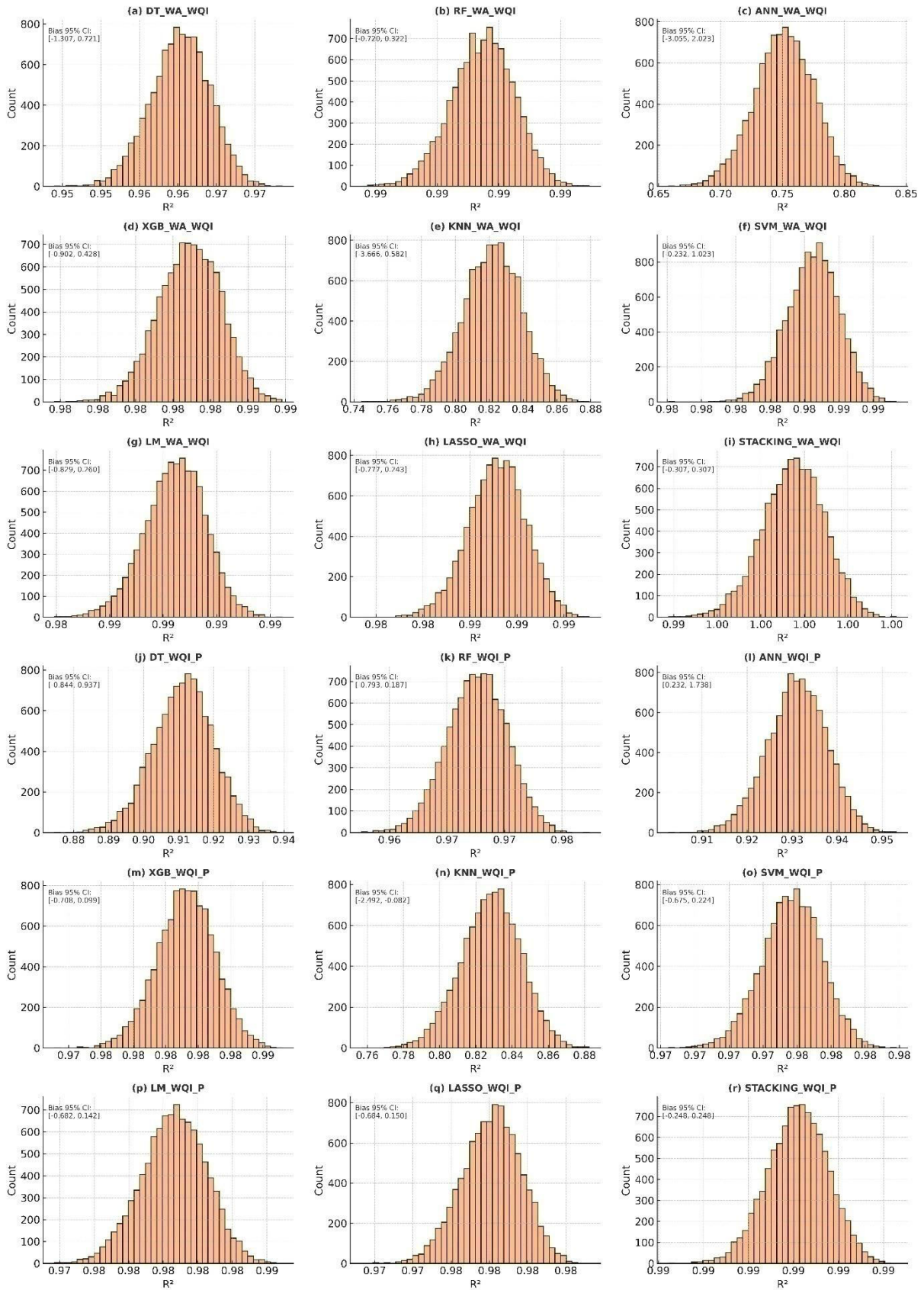
599 The research applied a comprehensive Gaussian Monte Carlo simulation framework  
 600 comprising 10,000 stochastic realizations to evaluate the predictive uncertainty and robustness  
 601 of the machine learning (ML) and ensemble models developed for Water Quality Index (WQI)  
 602 estimation. Model-specific residual distributions were used to generate stochastic perturbations,  
 603 enabling quantification of both random (aleatoric) and structural (epistemic) uncertainties. For

604 each realization, the  $R^2$ , RMSE, and bias metrics were recomputed to capture the propagation  
605 of observational noise and model error through the predictive  
606 workflow.

607 The Monte Carlo results revealed pronounced variability in uncertainty behaviour across model  
608 classes, emphasizing the fundamental influence of algorithmic design on stability under  
609 perturbation. Figure 11 illustrates the probabilistic distribution of  $R^2$  values across all models,  
610 while fig. 12 visualizes the comparative spread of RMSE, bias confidence width, and  $R^2$   
611 variance across various models.

612 From the fig. 11, it can be seen clearly the ensemble-based approaches, particularly  
613 STACKING\_WQI\_P, STACKING\_WA\_WQI, and RF\_WA\_WQI, exhibited superior  
614 statistical reliability and minimal uncertainty dispersion. Their Monte Carlo distributions were  
615 sharply peaked (median  $R^2 > 0.99$ , variance  $< 0.005$ ), with narrow RMSE confidence intervals  
616 ( $< 5\%$ ) and negligible mean bias ( $|\text{Bias}| \approx 0$ ), confirming the excellent predictive stability (Fig.  
617 11). This performance reflects the combined benefit of ensemble averaging and Principal  
618 Component Analysis (PCA)-based feature orthogonalization, which jointly suppress correlated  
619 noise and attenuate bias propagation. Compared to the single-learner models such as  
620 ANN\_WA\_WQI and KNN\_WA\_WQI demonstrated wider, asymmetric uncertainty envelopes,  
621 indicating high sensitivity to input perturbations, overfitting tendencies, whereas the models  
622 have limited generalization under nonstationary conditions.

623 However, the Monte Carlo uncertainty analysis substantiates that ensemble and hybrid  
624 architectures substantially outperformed compared to the single-learner models in both  
625 predictive precision and reliability.



626  
 627  
 628  
 629  
 630  
 631

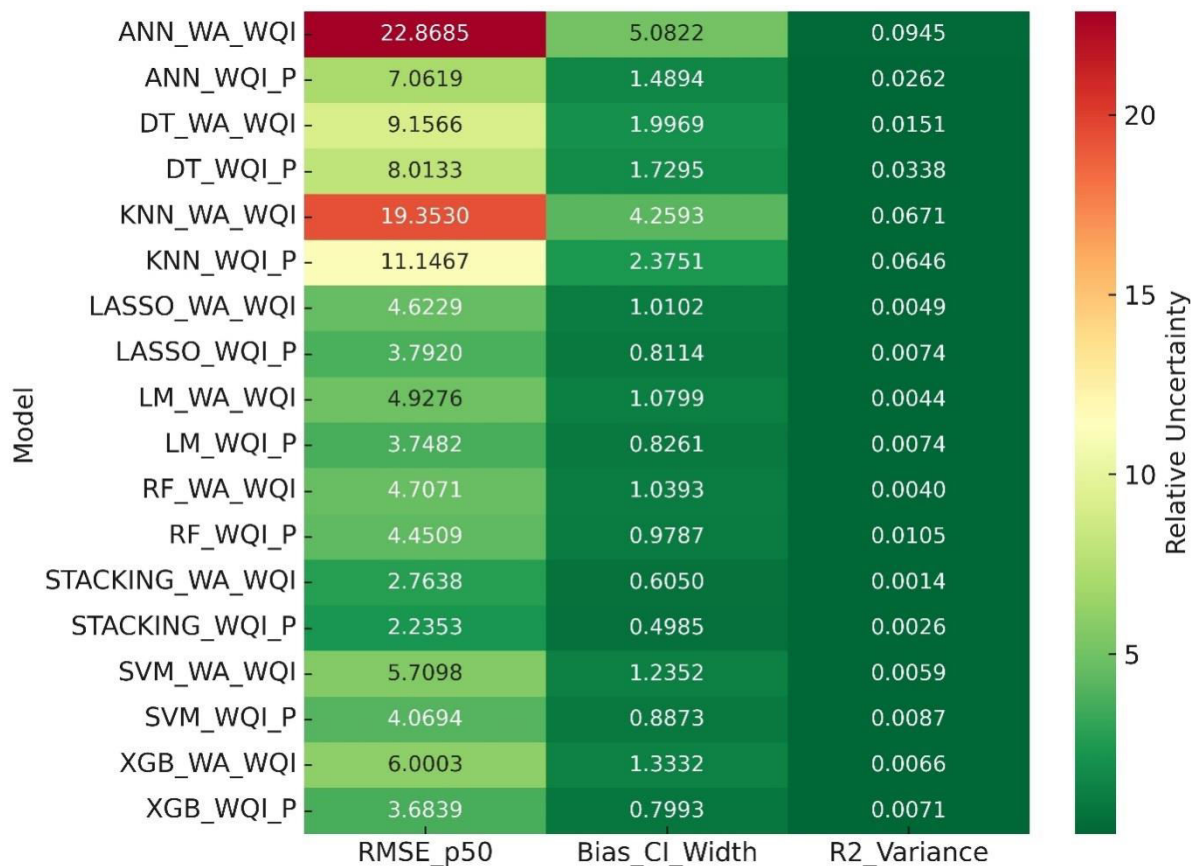
**Fig. 11.** Uncertainty distribution of  $R^2$  Values from 10,000 Gaussian Monte Carlo Simulations across WQI Models at parametric t-interval (95% CI) for the model's mean bias.

632 **Table 3.** WQI scores predicting performance of various model(s) at 95% confidence level.

Model	R <sup>2</sup> (Median ±95% CI)	RMSE (Median ±95% CI)**	Bias 95% CI	Interpretation
RF_WA_WQI	0.990 (± 0.002)	Low (< 10)	± 0.22	Very stable; minimal spread; highest precision.
XGB_WA_WQI	0.983 (± 0.003)	Low	± 0.28	Robust; strong generalization under perturbation.
STACKING_WA_WQI	0.998 (± 0.001)	Very low	± 0.31	Best overall consistency; negligible sensitivity.
DT_WA_WQI	0.960 (± 0.01)	Moderate	± 0.71	Sensitive to noise; higher variance.
KNN_WA_WQI	0.82 (± 0.04)	Moderate-high	± 0.58	Unstable with noise; lower predictive reliability.
ANN_WA_WQI	0.75 (± 0.05)	High	± 1.02	Strongly data-sensitive; needs larger samples.
RF_WQI_P	0.974 (± 0.01)	Low	± 0.32	PCA-based WQI improves robustness lightly.
STACKING_WQI_P	0.997 (± 0.002)	Very low	± 0.26	Most stable and highest-performing hybrid.
LM/LASSO_WQI_P	0.99 (± 0.004)	Low	± 0.20	Linear variants perform consistently.

633 \*\*This classification defined based on the empirical distribution of RMSE values obtained from the 10000 MCS for each model. Categories to  
 634 the following numerical thresholds : (i) very low—the RMSE median< 0.01 and CI width< 0.02 ; (ii) low – if the RMSE< 0.1 and CI  
 635 width<0.20 ; (iii) moderate – if the RMSE median 0.10 to 0.30 or CI width 0.20 to 0.50 ; (iv) moderate-high – when the RMSE median was  
 636 found 0.30 to 0.50 or CI width 0.50 to 0.80 ; and (v) high –if the RMSE median>0.50 or CI width>1.0 ; whereas the CI width computed from  
 637 the RMSE<sub>97.5</sub>– RMSE<sub>2.5</sub>of<sup>th</sup> percentiles, respective.

638



639 **Fig. 12.** Comparison of holistic uncertainty across various models.  
 640  
 641

642 It can be seen from the fig. 12, the comparative uncertainty heatmap further stratified models  
643 into three stability clusters:

644 (i) High-stability group – *STACKING\_WQI\_P*, *STACKING\_WA\_WQI*, and *RF\_WA\_WQI*;

645 (ii) Moderate-stability group – *XGB\_WA\_WQI*, *RF\_WQI\_P*, and *LASSO\_WQI\_P*;

646 (iii) Low-stability group – *ANN\_WA\_WQI* and *KNN\_WA\_WQI*.

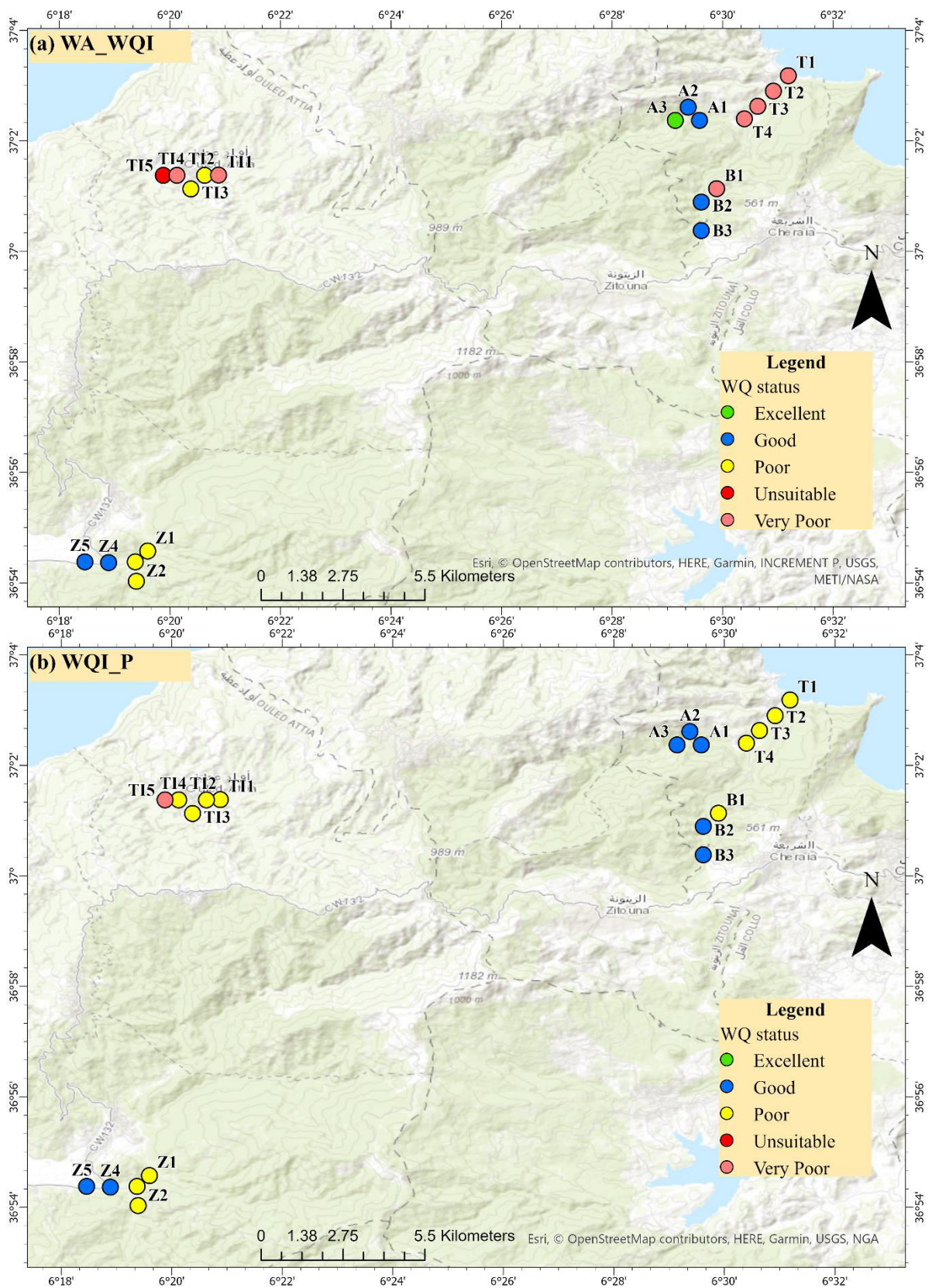
647 It can be seen from the table 3, The *STACKING\_WQI\_P* shows the remarkable consistency of  
648 the PCA-integrated stacking model (highlights the synergistic gains of dimensionality reduction  
649 and ensemble learning, effectively mitigating inherent uncertainties (Table 3). Similarly, the  
650 *RF\_WA\_WQI* achieved strong performance with low residual dispersion, presenting a  
651 computationally efficient yet resilient alternative for operational-scale environmental  
652 deployment of the model.

653 Based on the combined evaluation of  $R^2$  variance, RMSE confidence width, and bias  
654 propagation, also considering other uncertainty metrics from the Table 3 and figure 12, the  
655 *RF\_WQI\_P* model shows the relatively the lowest uncertainty these are statistically comparable  
656 to the highest-performing *WA\_WQI* ensemble models, with the smallest RMSE confidence  
657 interval and minimal bias. Conversely, the ANN- and the KNN-based models should be applied  
658 with a few generalised conditions like Bayesian dropout or posterior calibration. However, the  
659 model(s) uncertainty results indicates that the *RF\_WQI\_P* model performance not only strong  
660 central accuracy of the model but also the best outperformed model under stochastic  
661 perturbations, that could be utilized the reliable assessment of WQ with higher CI more  
662 accurately within any geographical extend.

### 663 **3.8. Assessment of WQ status**

664 Figure 13 shows the spatial distribution of water quality status obtained from both WQI models  
665 (*WA\_WQI* and *WQI\_P*). In the case of WA-WQI model, the WA-WQI model rated the water  
666 quality of the sampling sites into five classes (excellent, good, poor, very poor and unsuitable

667 for drinking) (Fig. 13a). Unlike WA\_WQI model, the WQI\_P model classified the water quality  
668 of the sampling sites into three categories (good, poor and very poor) (Fig. 13b). Both models  
669 demonstrated the “good” and “poor” water quality at all the five sampling sites in the Oued  
670 Z'Hour (Fig. 13). However, the WA\_WQI model showed the “poor”, “very poor” and  
671 “unsuitable for drinking” water status at the sampling sites in Oued Tizaghbane, whereas the  
672 WQI\_P model rated the water quality as “poor” and “very poor” (Fig. 13). In the Oued  
673 Tamanart, the water quality of the sampling sites showed “very poor” status with the WA\_WQI  
674 model, whereas the WQI\_P rated the water quality of the sampling sites as “poor” status (Fig.  
675 13). Moreover, both models showed different water state for the sampling sites in Oued Afensou  
676 (WA\_WQI: A3, “excellent”; WQI\_P: A3, “good”), and Oued Boudoudeh (WA\_WQI: B1,  
677 “very poor”; WQI\_P: B1, “poor”) (Fig. 13).

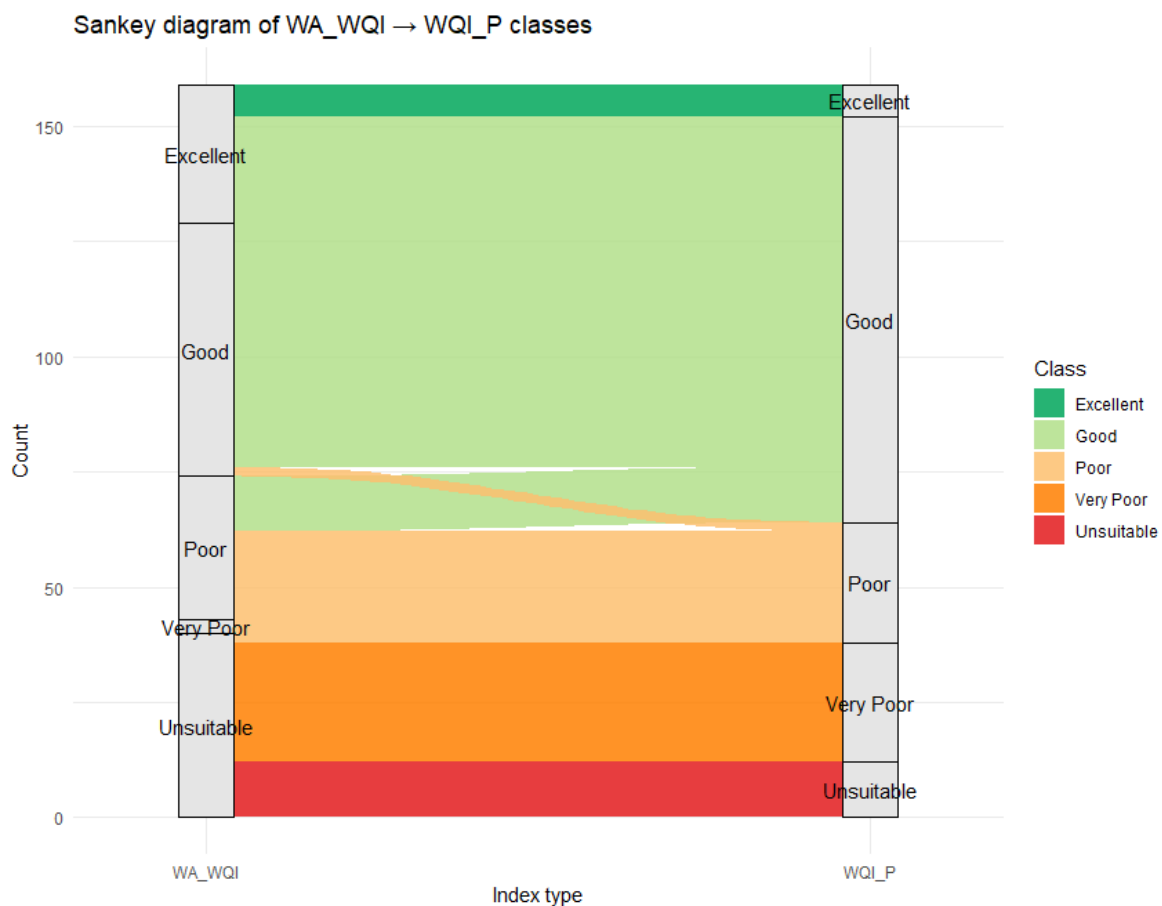


678  
679

Fig. 13. Assessing water quality status using WQ\_WQI and WQI\_P across various sites in Skikda.

680 **4. Discussion**

681 This study introduces a more objective and reliable water quality index (WQI\_P) based on  
682 principal component analysis, in which parameter weights are derived from variable  
683 contributions to PCA axes and aggregated using variance-based weighting. This data-driven  
684 framework reduces redundancy, addresses multicollinearity, and mitigates the eclipsing effect  
685 while maintaining a transparent weighting structure (Nath Roy et al. 2024). Compared with the  
686 traditional WA\_WQI, which relies on fixed permissible limits and simple weights, WQI\_P  
687 provides a more balanced representation of overall water quality (Uddin et al. 2022).



688 **Fig. 14.** Sankey diagram of transitions between WA\_WQI and WQI\_P classes.  
689

690 Fig. 14 illustrates substantial reclassification between WA\_WQI and WQI\_P. Under WQI\_P,  
691 the Excellent class decreases markedly, while the Good and Very Poor classes increase,  
692 indicating stricter upper thresholds and a redistribution toward more conservative

693 classifications. These shifts suggest improved discrimination across quality categories and  
694 reduced overestimation of water quality compared with WA\_WQI.

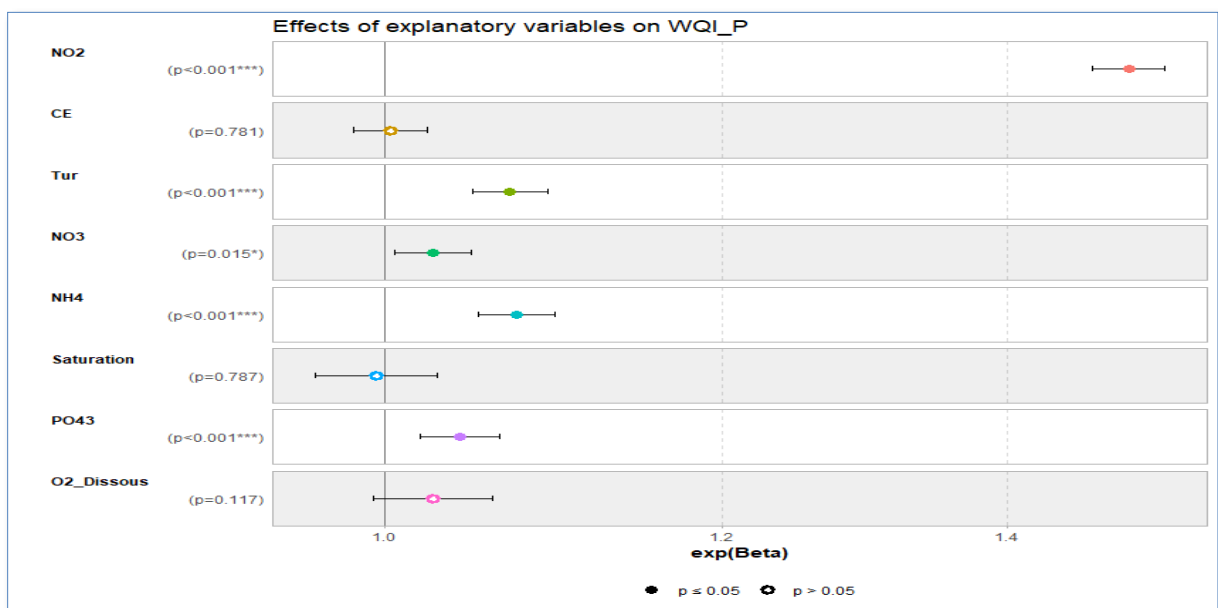
695 Such behavior aligns with well-documented limitations of conventional indices, where  
696 dominance by a small number of parameters masks broader water quality signals (Kachroud et  
697 al. 2019; Uddin et al. 2023d, 2023c, 2024). In this study, WA\_WQI was largely controlled by  
698 nitrate and orthophosphate, together accounting for approximately 95% of the total weight,  
699 making class assignments highly sensitive to minor nutrient fluctuations. By contrast, PCA-  
700 based weighting in WQI\_P distributes influence more evenly across retained components,  
701 effectively reducing eclipsing and yielding more interpretable classifications, in line with data-  
702 driven frameworks such as IEWQI (Uddin et al. 2022).

703 Predictive performance further supports the robustness of WQI\_P. As shown in Figure 7,  
704 WQI\_P exhibits lower mean prediction error and narrower 95% confidence intervals than  
705 WA\_WQI across all eight algorithms, indicating reduced uncertainty and improved reliability.  
706 The stacking meta-learner achieved superior performance for WQI\_P (RMSE = 2.74; MAE =  
707 1.75;  $r = 0.995$ ) compared with the WA\_WQI baseline, with more stable error distributions  
708 across models. These results align with recommendations to evaluate water quality indices  
709 based on predictive accuracy and uncertainty, rather than descriptive fit alone (Parween et al.  
710 2022; Tripathi and Singal 2019; Uddin et al. 2023c, 2024, 2021b). The stacking meta-learner  
711 provides the most stable performance, although the very high  $R^2$  values should be interpreted  
712 cautiously given the limited sample size and the high capacity of several machine-learning  
713 models.

714 The quasi-Gamma GLM results provide mechanistic insight into the drivers of water-quality  
715 variation. Increases in  $\text{NO}_2^-$ ,  $\text{NO}_3^-$ ,  $\text{NH}_4^+$ ,  $\text{PO}_4^{3-}$ , and turbidity are associated with multiplicative  
716 decreases in expected WQI\_P ( $\text{IRR} < 1$ ), whereas dissolved oxygen and oxygen saturation show

717 positive effects ( $IRR > 1$ ) (Fig. 15). Because predictors were centered and scaled, effect sizes  
 718 are directly comparable, and confidence intervals not crossing unity confirm the direction of  
 719 effects. These findings align with established evidence identifying nutrients and turbidity as  
 720 principal degraders of water quality(Bouchra et al. 2025; Chawaka et al. 2024; Koue 2024;  
 721 Mostefa et al. 2025).

722 Observed exceedances of nitrite, nitrate, ammonium, and orthophosphate are consistent with  
 723 known pollution pressures in the Skikda region. Intensive vegetable farming along Oued  
 724 Z'Hour, Oued Tamanart, and Oued Tizaghbaner relies on synthetic nitrogen and phosphate  
 725 fertilizers, which are mobilized by runoff and shallow subsurface flow during early rainfall  
 726 events(Das 2025; Mostefa et al. 2025; Zhang et al. 2025).Urban expansion further contributes  
 727 through poorly treated domestic effluents and greywater discharges, particularly during dry-  
 728 season low-flow conditions(Abidi Saad et al. 2024; Akinnigbagbe et al. 2025; Das 2025; Zhang  
 729 et al. 2025).Land–water continuum studies support efficient transfer of nutrients into river  
 730 networks following episodic storms, explaining the elevated concentrations observed at several  
 731 stations(Mellander et al. 2025; Mostefa et al. 2025; Rodríguez-Cardona et al. 2020).



732 **Fig. 15.** Exponentiated coefficients ( $\exp(\beta)$ ) with p-values showing the effects of physico-chemical variables on  
 733 WQI\_P using GLM with quasi-Gamma distribution.  
 734  
 735

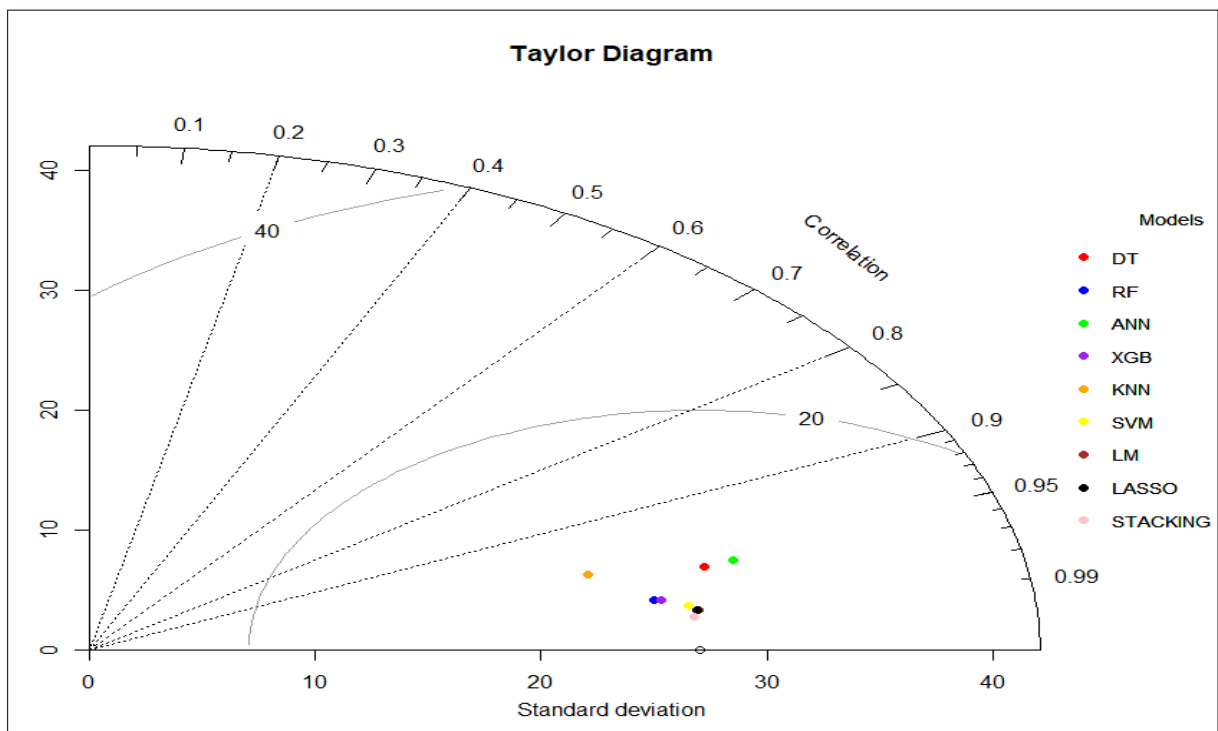
736 From a regulatory perspective, several parameters exceeded guideline limits, including  $\text{NO}_2^-$ ,  
737  $\text{PO}_4^{3-}$ , and  $\text{NH}_4^+$ , while nitrate remained below the regulatory threshold but at levels conducive  
738 to eutrophication (Table S1) (Chidiac et al. 2023). These exceedances are consistent with  
739 contamination from agricultural runoff and wastewater inputs documented in comparable  
740 studies (An et al. 2024; Chawaka et al. 2024; İsmail Akçay and Özgür Özbay 2024; Koue 2024;  
741 Tesoriero et al. 2024).

742 Conceptually, WQI\_P aligns with the IEWQI framework that promotes data-driven indicator  
743 selection, weighting, and aggregation to limit eclipsing and quantify uncertainty (Uddin et al.  
744 2023c, 2024). However, WQI\_P differs in deriving weights directly from PCA contribution  
745 values and aggregating retained components through variance-based weighting. RMS-WQI  
746 integrates machine learning into index construction, whereas WQI\_P employs machine learning  
747 solely for predictive evaluation.

748 Relative to WQIR formulations that combine PCA with ANN-based prediction, WQI\_P  
749 formalizes weighting directly from PCA contribution values and aggregates retained  
750 components using variance-based weighting (Fartas et al. 2022). Nevertheless, PCA-derived  
751 weights remain sample dependent, a limitation shared with IEWQI and RMS-WQI. Addressing  
752 this limitation will require future work on cross-basin recalibration or transfer-learning  
753 strategies.

754 Comparison with other PCA-informed indices, including weighted arithmetic WQIs and  
755 RIWQI, confirmed that reducing eclipsing leads to stricter and more balanced classification  
756 (Ghoderao et al. 2022). The advantage of PCA based approaches is reduced exposure to  
757 eclipsing when a small subset of variables carries most of the weight. The reclassification  
758 patterns here, including the contraction of the Excellent class and redistribution across Good  
759 and lower categories, are consistent with a stricter and more balanced assignment under WQI\_P  
760 (Ghoderao et al. 2022).

761 The Revised Iranian WQI (RIWQI) similarly incorporates multicollinearity checks and PCA to  
 762 refine parameter sets and weights, illustrating the broader trend toward data informed indices.  
 763 That rationale matches the weighting logic of WQI\_P and helps explain the observed  
 764 improvements in predictive reliability and class stability (Fathi et al. 2022).  
 765 In this dataset, stacking over eight algorithms achieves RMSE 2.73833, MAE 1.74824, and  
 766 correlation 0.9948635 for WQI\_P, surpassing the WA\_WQI stacking baseline (RMSE  
 767 3.163572, MAE 2.211848). A prudent phrasing follows: in this dataset, WQI\_P reduces RMSE  
 768 and MAE more than WA\_WQI, while remaining consistent with IEWQI, WQIR, and RIWQI  
 769 reports that link improvements to mitigation of eclipsing and better calibration of uncertainty.  
 770 These comparisons indicate that WQI\_P addresses several gaps identified in the introduction,  
 771 notably eclipsing, collinearity, and predictive uncertainty, while the sample dependence of PCA  
 772 weights remains an open issue for future refinement.



773  
 774 **Fig. 16.** Taylor diagrams comparing WQI prediction algorithms (WQI\_P).  
 775

776 Fig. 16 shows that all algorithms achieve high correlation with observations ( $\approx 0.95-0.99$ ) and  
 777 comparable variance representation, with the stacking ensemble providing the best overall

778 performance by jointly minimizing RMSE and variance mismatch. Among single learners,  
779 ANN, KNN, and XGB perform more favorably than DT and RF, consistent with the lower  
780 PREI values and tighter 95% confidence intervals reported earlier for WQI\_P.

781 Potential overfitting likely reflects the limited sample size and temporal dependence among  
782 observations. Mitigation requires expanded sampling coverage, constrained model complexity,  
783 and principled feature reduction and preprocessing, as recommended in similar studies (Uddin  
784 et al. 2023c).

## 785 **5. Conclusion**

786 This study introduced a PCA based water quality index named WQI\_P that aims to reduce  
787 eclipsing, multicollinearity, and subjective weighting, which are common limitations in  
788 classical indices such as the weighted arithmetic WA\_WQI. By deriving weights from PCA  
789 contribution values and combining component level subindices using explained variance,  
790 WQI\_P links the internal covariance structure of the data to the final index scores. This  
791 approach provides a more objective and data informed representation of river water quality.

792 Key findings from this research are as follows:

- 793 • Among the studied water quality parameters,  $\text{NO}_2^-$ , TURB,  $\text{NH}_4^+$  and  $\text{PO}_4^{3-}$  breached  
794 their guidelines limit in 25.2%, 23.3%, 8.8% and 0.6% samples, respectively.
- 795 • Applied to 159 river samples from the Skikda region, WQI\_P produced stricter and  
796 more balanced classifications. The Excellent category was considerably reduced, and  
797 observations were more evenly redistributed across Good, Poor, and Very Poor classes.
- 798 • The predictive evaluations using eight machine learning models demonstrated that the  
799 WQI\_P achieved lower prediction errors, smaller uncertainty ranges, and more stable  
800 performance than WA\_WQI.

801 • These findings indicate that PCA based weighting improves interpretability and  
802 predictive reliability, which makes WQI\_P a useful alternative for practical water  
803 quality assessment.

804 However, several limitations should be acknowledged, including the number of samples is  
805 modest and confined to a single region, and repeated measurements may increase the risk of  
806 overfitting when using high-capacity models. Additionally, weights derived from PCA remain  
807 dependent on the dataset used for their calculation, which means that the WQI\_P is not entirely  
808 site independent. These aspects must be considered before extending the results beyond the  
809 study area. Despite these limitations, the research provides valuable insights for local aquatic  
810 managers and policy makers for the sustainable management of water resources, and for  
811 achieving Sustainable Development Goal target 6.3 (by 2030, improve water quality by  
812 reducing pollution, eliminating dumping and minimizing release of hazardous chemicals and  
813 materials, halving the proportion of untreated wastewater and substantially increasing recycling  
814 and safe reuse globally). Looking ahead, future research should include the validation of  
815 WQI\_P in larger and more diverse basins, assessment of its stability under seasonal and climatic  
816 variation, and systematic comparison with recently developed indices such as IEWQI, RMS  
817 WQI, WQIR, and RIWQI. Additional research on cross basin recalibration or transfer learning  
818 could be helpful for applying PCA based indices to new regions while preserving the shared  
819 hydro-chemical patterns. This could ultimately strengthen the scientific and practical value of  
820 WQI\_P and support its integration into the operational water quality monitoring programs.

### 821 **Funding Declaration**

822 This research did not receive any specific grant from funding agencies in the public,  
823 commercial, or non-profit sectors. Open access funding is provided by University of Galway.

## 824 **Declarations**

825 All authors have read, understood, and have complied as applicable with the statement on  
826 “Declaration of competing interests” as found in the “Instructions for Authors.”

## 827 **Acknowledgement**

828 The authors sincerely acknowledge the Eco-Hydro Informatics Research Group (EHIRG), Civil  
829 Engineering, School of Engineering, College of Science and Engineering, University of  
830 Galway, Ireland, for providing computational laboratory facilities to complete this research.  
831 The authors also extend their gratitude to the Lead Researcher of the EHIRG for supervising  
832 this research and providing significant input that contributed to its successful completion. The  
833 authors would also like to express their deep appreciation to Professor Khiari, Director of the  
834 (RNAMS), as well as to Professor Saadali, members of the RNAMS, for their valuable  
835 guidance, support, and encouragement throughout the research process.

## 836 **References**

- 837 Abidi Saad, M., Seghir, K., Touahri, A., Bendekkoum, M., & Bellaoueur, A. (2024). Applying the water quality  
838 indices and geographic information system approach to assessing the suitability of groundwater quality  
839 for drinking and irrigation purposes in the semi-arid region of Tebessa-Ain Chabro, Northeastern  
840 Algeria. *Applied Water Science*, 14(10), 221. <https://doi.org/10.1007/s13201-024-02275-3>
- 841 Akinnigbagbe, A. E., Popoola, S. O., Oyatola, O. O., Oghenede, E. K., & Nubi, O. A. (2025). Spatiotemporal  
842 patterns of water quality, nutrient dynamics and chlorophyll a concentrations in Five Cowries Creek  
843 from 2022 to 2024. *Discover Water*, 5(1), 63. <https://doi.org/10.1007/s43832-025-00257-3>
- 844 Aljanabi, Z. Z., Al-Obaidy, A.-H. M. J., & Hassan, F. M. (2021). A brief review of water quality indices and  
845 their applications. *IOP Conference Series: Earth and Environmental Science*, 779(1), 012088.  
846 <https://doi.org/10.1088/1755-1315/779/1/012088>
- 847 Allaoua, N., Hafid, H., & Chenchouni, H. (2024). Exploring groundwater quality in semi-arid areas of Algeria:  
848 Impacts on potable water supply and agricultural sustainability. *Journal of Arid Land*, 16(2), 147–167.  
849 <https://doi.org/10.1007/s40333-024-0004-4>
- 850 An, L., Li, Q., Wu, P., Lu, W., Li, X., Zhang, C., & Zhang, R. (2024). Potential impacts of coal mining activities  
851 on nitrate sources and transport in a karst river basin in southwest China. *Environmental Science and  
852 Pollution Research*, 31(10), 15412–15423. <https://doi.org/10.1007/s11356-024-32167-7>
- 853 Banda, T. D., & Kumarasamy, M. V. (2020). Development of Water Quality Indices (WQIs): A Review. *Polish  
854 Journal of Environmental Studies*, 29(3). [https://www.pjoes.com/pdf-110526-  
855 48363?filename=Development%20of%20Water.pdf](https://www.pjoes.com/pdf-110526-48363?filename=Development%20of%20Water.pdf). Accessed 12 December 2024
- 856 Berrar, D. (2018). Cross-Validation. <https://doi.org/10.1016/B978-0-12-809633-8.20349-X>
- 857 Boucenna, H., Satour, A., Hezil, W., Taferghoust, M., Samraoui, F., & Samraoui, B. (2023). Diversity,  
858 distribution, and conservation of the Trichoptera and their habitats in north-eastern Algeria. *Aquatic  
859 Conservation*, 33(5), 502–516. <https://doi.org/10.1002/aqc.3931>
- 860 Bouchra, D., Allaoua, N., Ghanem, N., Hafid, H., Benacherine, M., & Chenchouni, H. (2025). Assessment of  
861 water quality of groundwater, surface water, and wastewater using physicochemical parameters and  
862 microbiological indicators. *Science Progress*, 108, 1–35. <https://doi.org/10.1177/00368504251348544>
- 863 Boudiaf, B., Şen, Z., & Boutaghane, H. (2022). North coast Algerian rainfall monthly trend analysis using  
864 innovative polygon trend analysis (IPTA). *Arabian Journal of Geosciences*, 15(21), 1626.  
865 <https://doi.org/10.1007/s12517-022-10907-8>

- 866 Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical statistics for data scientists: 50+ essential concepts using R*  
867 *and Python*. O'Reilly Media.  
868 [https://books.google.com/books?hl=fr&lr=&id=k2XcDwAAQBAJ&oi=fnd&pg=PP1&dq=practical+St](https://books.google.com/books?hl=fr&lr=&id=k2XcDwAAQBAJ&oi=fnd&pg=PP1&dq=practical+Statistics+for+Data+Scientist+by+Peter+Bruce,+Andrew+Bruce&ots=dELhlg0nD0&sig=dweF6d7fgnVsNjzdBuO6LrJW9Yg)  
869 [atistics+for+Data+Scientist+by+Peter+Bruce,+Andrew+Bruce&ots=dELhlg0nD0&sig=dweF6d7fgnVs](https://books.google.com/books?hl=fr&lr=&id=k2XcDwAAQBAJ&oi=fnd&pg=PP1&dq=practical+Statistics+for+Data+Scientist+by+Peter+Bruce,+Andrew+Bruce&ots=dELhlg0nD0&sig=dweF6d7fgnVsNjzdBuO6LrJW9Yg)  
870 [NjzdBuO6LrJW9Yg](https://books.google.com/books?hl=fr&lr=&id=k2XcDwAAQBAJ&oi=fnd&pg=PP1&dq=practical+Statistics+for+Data+Scientist+by+Peter+Bruce,+Andrew+Bruce&ots=dELhlg0nD0&sig=dweF6d7fgnVsNjzdBuO6LrJW9Yg). Accessed 17 August 2025
- 871 Cadilhac, L. (2003). Le système d'évaluation de la qualité des eaux souterraines " SEQ - Eaux souterraines ". *La*  
872 *Houille Blanche*, 89(2), 125–128. <https://doi.org/10.1051/lhb/2003040>
- 873 Chawaka, S. N., Boets, P., Mereta, S. T., Goethals, P. L. M., & Ancha, V. R. (2024). Effect of agriculture on  
874 surface water quantity and quality in Gilgel Gibe watershed, southwestern Ethiopia. *Environmental*  
875 *Monitoring and Assessment*, 196(6), 578. <https://doi.org/10.1007/s10661-024-12732-w>
- 876 Chidiac, S., El Najjar, P., Ouaini, N., El Rayess, Y., & El Azzi, D. (2023). A comprehensive review of water  
877 quality indices (WQIs): history, models, attempts and perspectives. *Reviews in Environmental Science*  
878 *and Bio/Technology*, 22(2), 349–395. <https://doi.org/10.1007/s11157-023-09650-7>
- 879 Das, A. (2025). An optimization based framework for water quality assessment and pollution source  
880 apportionment employing GIS and machine learning techniques for smart surface water governance.  
881 *Discover Environment*, 3(1), 117. <https://doi.org/10.1007/s44274-025-00327-2>
- 882 Edition, F. (2011). Guidelines for drinking-water quality. *WHO chronicle*, 38(4), 104–8.  
883 [https://apps.who.int/iris/bitstream/handle/10665/204411/9789241547611\\_eng.pdf](https://apps.who.int/iris/bitstream/handle/10665/204411/9789241547611_eng.pdf). Accessed 17 August  
884 2025
- 885 Eid, M. H., Saeed, O., Szűcs, P., Kovács, A., Székács, A., Mörtl, M., et al. (2025). Impacts and sources of  
886 potential toxic elements on water quality and optimizing machine learning models for sustainable  
887 management. *Modeling Earth Systems and Environment*, 11(5), 375. [https://doi.org/10.1007/s40808-](https://doi.org/10.1007/s40808-025-02548-z)  
888 [025-02548-z](https://doi.org/10.1007/s40808-025-02548-z)
- 889 El Osta, M., Masoud, M., Alqarawy, A., Elsayed, S., & Gad, M. (2022). Groundwater Suitability for Drinking  
890 and Irrigation Using Water Quality Indices and Multivariate Modeling in Makkah Al-Mukarramah  
891 Province, Saudi Arabia. *Water*, 14(3), 483. <https://doi.org/10.3390/w14030483>
- 892 Fartas, F., Remini, B., Sekiou, F., & Marouf, N. (2022). The use of PCA and ANN to improve evaluation of the  
893 WQIclassic, development of a new index, and prediction of WQI, Coastel Constantinois, northern coast  
894 of eastern Algeria. *Water Supply*, 22(12), 8727–8749. <https://doi.org/10.2166/ws.2022.389>
- 895 Fathi, P., Ebrahimi Dorche, E., Zare Shahraki, M., Stribling, J., Beyraghdar Kashkooli, O., Esmaili Ofogh, A.,  
896 & Bruder, A. (2022). Revised Iranian Water Quality Index (RIWQI): a tool for the assessment and  
897 management of water quality in Iran. *Environmental Monitoring and Assessment*, 194(7), 504.  
898 <https://doi.org/10.1007/s10661-022-10121-9>
- 899 Gad, M., & El-Hattab, M. (2019). Integration of water pollution indices and DRASTIC model for assessment of  
900 groundwater quality in El Fayoum depression, western desert, Egypt. *Journal of African Earth*  
901 *Sciences*, 158, 103554. <https://doi.org/10.1016/j.jafrearsci.2019.103554>
- 902 Gad, M., Gaagai, A., Agrama, A. A., El-Fiqy, W. F. M., Eid, M. H., Szűcs, P., et al. (2024). Comprehensive  
903 evaluation and prediction of groundwater quality and risk indices using quantitative approaches,  
904 multivariate analysis, and machine learning models: An exploratory study. *Heliyon*, 10(17).  
905 <https://doi.org/10.1016/j.heliyon.2024.e36606>
- 906 Gad, M., Gaagai, A., Eid, M. H., Szűcs, P., Hussein, H., Elsherbiny, O., et al. (2023). Groundwater Quality and  
907 Health Risk Assessment Using Indexing Approaches, Multivariate Statistical Analysis, Artificial Neural  
908 Networks, and GIS Techniques in El Kharga Oasis, Egypt. *Water*, 15(6), 1216.  
909 <https://doi.org/10.3390/w15061216>
- 910 Gao, Y., Qian, H., Ren, W., Wang, H., Liu, F., & Yang, F. (2020). Hydrogeochemical characterization and  
911 quality assessment of groundwater based on integrated-weight water quality index in a concentrated  
912 urban area. *Journal of Cleaner Production*, 260, 121006. <https://doi.org/10.1016/j.jclepro.2020.121006>
- 913 Gazzaz, N. M., Yusoff, M. K., Aris, A. Z., Juahir, H., & Ramli, M. F. (2012). Artificial neural network modeling  
914 of the water quality index for Kinta River (Malaysia) using water quality variables as predictors.  
915 *Marine Pollution Bulletin*, 64(11), 2409–2420. <https://doi.org/10.1016/j.marpolbul.2012.08.005>
- 916 Ghoderao, S. B., Meshram, S. G., & Meshram, C. (2022). Development and evaluation of a water quality index  
917 for groundwater quality assessment in parts of Jabalpur District, Madhya Pradesh, India. *Water Supply*,  
918 22(6), 6002–6012. <https://doi.org/10.2166/ws.2022.174>
- 919 Gupta, D., & Mishra, V. K. (2023). Development of entropy-river water quality index for predicting water  
920 quality classification through machine learning approach. *Stochastic Environmental Research and Risk*  
921 *Assessment*, 37(11), 4249–4271. <https://doi.org/10.1007/s00477-023-02506-0>
- 922 Hassan, Md. M., Hassan, Md. M., Akter, L., Rahman, Md. M., Zaman, S., Hasib, K. Md., et al. (2021). Efficient  
923 Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms. *Human-Centric*  
924 *Intelligent Systems*, 1(3), 86–97. <https://doi.org/10.2991/hcis.k.211203.001>

- 925 Hfaiedh, E., Gaagai, A., Petitta, M., Ben Moussa, A., Mlayah, A., Eid, M. H., et al. (2025). Hydrogeochemical  
 926 characterization and water quality evaluation associated with toxic elements using indexing approaches,  
 927 multivariate analysis, and artificial neural networks in Morang, Tunisia. *Environmental Earth Sciences*,  
 928 *84*(13), 361. <https://doi.org/10.1007/s12665-025-12165-9>
- 929 Ibrahim, A., Ismail, A., Juahir, H., Iliyasu, A. B., Wailare, B. T., Mukhtar, M., & Aminu, H. (2023). Water  
 930 quality modelling using principal component analysis and artificial neural network. *Marine Pollution*  
 931 *Bulletin*, *187*, 114493. <https://doi.org/10.1016/j.marpolbul.2022.114493>
- 932 IPCC. (2023). *Climate Change 2023: Synthesis Report*.  
 933 [https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC\\_AR6\\_SYR\\_LongerReport.pdf](https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC_AR6_SYR_LongerReport.pdf)
- 934 Islam Khan, Md. S., Islam, N., Uddin, J., Islam, S., & Nasir, M. K. (2022). Water quality prediction and  
 935 classification based on principal component regression and gradient boosting classifier approach.  
 936 *Journal of King Saud University - Computer and Information Sciences*, *34*(8, Part A), 4773–4781.  
 937 <https://doi.org/10.1016/j.jksuci.2021.06.003>
- 938 İsmail Akçay & Özgür Özbay. (2024). Assessment of Ecological and Potential Health Risk Caused by Nitrate  
 939 Pollution of the Berdan and Göksu River Basins, Turkey. *Journal of Water Chemistry and Technology*,  
 940 *46*(6), 645–651. <https://doi.org/10.3103/S1063455X24060018>
- 941 Iyiola, A. O., & Asiedu, B. (2020). Benthic macro-invertebrates as indicators of water quality in Ogunpa River,  
 942 South-Western Nigeria. *West African Journal of Applied Ecology*, *28*(1), 85–95.  
 943 <https://www.ajol.info/index.php/wajae/article/view/199528>. Accessed 15 April 2025
- 944 Jayaraman, P., Nagarajan, K. K., Partheeban, P., & Krishnamurthy, V. (2024). Critical review on water quality  
 945 analysis using IoT and machine learning models. *International Journal of Information Management*  
 946 *Data Insights*, *4*(1), 100210. <https://doi.org/10.1016/j.ijime.2023.100210>
- 947 Kachroud, M., Trolard, F., Kefi, M., Jebari, S., & Bourrié, G. (2019). Water Quality Indices: Challenges and  
 948 Application Limits in the Literature. *Water*, *11*(2), 361. <https://doi.org/10.3390/w11020361>
- 949 Kassambara, A. (2018). *Machine Learning Essentials: Practical Guide in R*. STHDA.
- 950 Khan, I., Nizam, S., Bamal, A., Sajib, A. M., Mahammad Diganta, M. T., Shaida, M. A., et al. (2025a).  
 951 Optimized intelligent learning for groundwater quality prediction in diverse aquifers of arid and semi-  
 952 arid regions of India. *Cleaner Engineering and Technology*, *26*, 100984.  
 953 <https://doi.org/10.1016/j.clet.2025.100984>
- 954 Khan, I., Nizam, S., Bamal, A., Sajib, A. M., Mahammad Diganta, M. T., Shaida, M. A., et al. (2025b).  
 955 Optimized intelligent learning for groundwater quality prediction in diverse aquifers of arid and semi-  
 956 arid regions of India. *Cleaner Engineering and Technology*, *26*, 100984.  
 957 <https://doi.org/10.1016/j.clet.2025.100984>
- 958 Khan, Y., & See, C. S. (2016). Predicting and analyzing water quality using Machine Learning: A  
 959 comprehensive model. In *2016 IEEE Long Island Systems, Applications and Technology Conference*  
 960 *(LISAT)* (pp. 1–6). Presented at the 2016 IEEE Long Island Systems, Applications and Technology  
 961 Conference (LISAT), Farmingdale, NY, USA: IEEE. <https://doi.org/10.1109/LISAT.2016.7494106>
- 962 Kheloufi Attou, A., Baba-Hamed, K., & Bouanani, A. (2025). A quantitative study of extreme rainfall intensity  
 963 and occurrence in northern Algeria. *Atmosfera*, *39*. <https://doi.org/10.20937/ATM.53442>
- 964 Khouri, L., & Bashar Al-Moufti, M. (2022). Selection of suitable aggregation function for estimation of water  
 965 quality index for the Orontes River. *Ecological Indicators*, *142*, 109290.  
 966 <https://doi.org/10.1016/j.ecolind.2022.109290>
- 967 Koue, J. (2024). Impact of riverine inputs on nutrient dynamics and water quality in enclosed water bodies.  
 968 *Terrestrial, Atmospheric and Oceanic Sciences*, *35*(1), 1–9. <https://doi.org/10.1007/s44195-024-00081-7>
- 969
- 970 Li, L., Rong, S., Wang, R., & Yu, S. (2021). Recent advances in artificial intelligence and machine learning for  
 971 nonlinear relationship analysis and process control in drinking water treatment: A review. *Chemical*  
 972 *Engineering Journal*, *405*, 126673. <https://doi.org/10.1016/j.cej.2020.126673>
- 973 Mahesh, B. (2019). *Machine Learning Algorithms -A Review*. *International Journal of Science and Research*  
 974 *(IJSR)* (Vol. 9). <https://doi.org/10.21275/ART20203995>
- 975 Mellander, P.-E., Bol, R., Bieroza, M., Burgess, E., Ezzati, G., Glendell, M., et al. (2025). Achieving agricultural  
 976 and environmental targets in a changing climate requires a whole-system based approach. *Discover*  
 977 *Geoscience*, *3*(1), 205. <https://doi.org/10.1007/s44288-025-00321-4>
- 978 Mishra, S. P., Sarkar, U., Taraphder, S., Datta, S., Swain, D., Saikhom, R., et al. (2017). Multivariate statistical  
 979 data analysis-principal component analysis (PCA). *International Journal of Livestock Research*, *7*(5),  
 980 60–78. [https://www.academia.edu/download/54658288/1\\_Multivariate\\_Statistical\\_Data\\_Analysis-Principal\\_Component\\_Analysis\\_PCA.pdf](https://www.academia.edu/download/54658288/1_Multivariate_Statistical_Data_Analysis-Principal_Component_Analysis_PCA.pdf). Accessed 26 April 2025
- 981
- 982 Mostefa, B., Nadjib, L. M., Noua, A., Hana, S. A., El Latif, S. A., Norhane, C., et al. (2025). From fragmented to  
 983 integrated: Advancing surface water quality assessment through WQPI in the upper Medjerda River,

984 Northeastern Algeria. *Desalination and Water Treatment*, 323, 101342.  
985 <https://doi.org/10.1016/j.dwt.2025.101342>

986 Nath Roy, B., Roy, H., Rahman, K. S., Mahmud, F., Bhuiyan, M. M. K., Hasan, M., et al. (2024). Principal  
987 component analysis incorporated water quality index modeling for Dhaka-based rivers. *City and*  
988 *Environment Interactions*, 23, 100150. <https://doi.org/10.1016/j.cacint.2024.100150>

989 Nuzzo, R. L. (2016). The Box Plots Alternative for Visualizing Quantitative Data. *PM&R*, 8(3), 268–272.  
990 <https://doi.org/10.1016/j.pmrj.2016.02.001>

991 OECD. (2022). *Water Quality and Agriculture: Meeting the Policy Challenge*.  
992 [https://www.oecd.org/content/dam/oecd/en/publications/reports/2012/03/water-quality-and-](https://www.oecd.org/content/dam/oecd/en/publications/reports/2012/03/water-quality-and-agriculture_g1g196a1/9789264168060-en.pdf)  
993 [agriculture\\_g1g196a1/9789264168060-en.pdf](https://www.oecd.org/content/dam/oecd/en/publications/reports/2012/03/water-quality-and-agriculture_g1g196a1/9789264168060-en.pdf)

994 Olbert, A. I., Diganta, M. T. M., Bamal, A., Burke, W., Sajib, A. M., Abioui, M., et al. (2025). Developing river  
995 water quality prediction model incorporating reliable indexing approach. *Journal of Environmental*  
996 *Sciences*. <https://doi.org/10.1016/j.jes.2025.07.038>

997 Parween, S., Siddique, N. A., Mahammad Diganta, M. T., Olbert, A. I., & Uddin, M. G. (2022). Assessment of  
998 urban river water quality using modified NSF water quality index model at Siliguri city, West Bengal,  
999 India. *Environmental and Sustainability Indicators*, 16, 100202.  
1000 <https://doi.org/10.1016/j.indic.2022.100202>

1001 Rodier, J., Legube, B., & Merlet, N. (2016). *L'analyse de l'eau - 10e éd.* Dunod.

1002 Rodríguez-Cardona, B. M., Coble, A. A., Wymore, A. S., Kolosov, R., Podgorski, D. C., Zito, P., et al. (2020).  
1003 Wildfires lead to decreased carbon and increased nitrogen concentrations in upland arctic streams.  
1004 *Scientific Reports*, 10(1), 8722. <https://doi.org/10.1038/s41598-020-65520-0>

1005 Rojas-Valverde, D., Pino-Ortega, J., Gómez-Carmona, C. D., & Rico-González, M. (2020). A systematic review  
1006 of methods and criteria standard proposal for the use of principal component analysis in team's sports  
1007 science. *International Journal of Environmental Research and Public Health*, 17(23), 8712.  
1008 <https://www.mdpi.com/1660-4601/17/23/8712>. Accessed 2 May 2025

1009 Sajib, A. M., Diganta, M. T. M., Moniruzzaman, Md., Rahman, A., Dabrowski, T., Uddin, M. G., & Olbert, A. I.  
1010 (2024). Assessing water quality of an ecologically critical urban canal incorporating machine learning  
1011 approaches. *Ecological Informatics*, 80, 102514. <https://doi.org/10.1016/j.ecoinf.2024.102514>

1012 Salem, S. B. H., Gaagai, A., Ben Slimene, I., Moussa, A. B., Zouari, K., Yadav, K. K., et al. (2023). Applying  
1013 Multivariate Analysis and Machine Learning Approaches to Evaluating Groundwater Quality on the  
1014 Kairouan Plain, Tunisia. *Water*, 15(19), 3495. <https://doi.org/10.3390/w15193495>

1015 Satour, A., Wissem, H., Taferghoust, M., Hayat, B., Samraoui, F., & Samraoui, B. (2024). Land use and beyond:  
1016 unraveling environmental determinants of odonate assemblages in northeastern Algeria. *International*  
1017 *Journal of Odonatology*, 27, 172–186. <https://doi.org/10.48156/1388.2024.1917289>

1018 Singh, B., Sihag, P., Singh, V. P., Sepahvand, A., & Singh, K. (2021). Soft computing technique-based  
1019 prediction of water quality index. *Water Supply*, 21(8), 4015–4029. <https://doi.org/10.2166/ws.2021.157>

1020 Sweet, L., Müller, C., Anand, M., & Zscheischler, J. (2023). Cross-Validation Strategy Impacts the Performance  
1021 and Interpretation of Machine Learning Models. <https://doi.org/10.1175/AIES-D-23-0026.1>

1022 Tesoriero, A. J., Robertson, D. M., Green, C. T., Böhlke, J. K., Harvey, J. W., & Qi, S. L. (2024). Prioritizing  
1023 river basins for nutrient studies. *Environmental Monitoring and Assessment*, 196(3), 248.  
1024 <https://doi.org/10.1007/s10661-023-12266-7>

1025 Tripathi, M., & Singal, S. K. (2019). Use of Principal Component Analysis for parameter selection for  
1026 development of a novel Water Quality Index: A case study of river Ganga India. *Ecological Indicators*,  
1027 96, 430–436. <https://doi.org/10.1016/j.ecolind.2018.09.025>

1028 Tyagi, S., Sharma, B., Singh, P., & Dobhal, R. (2013). Water quality assessment in terms of water quality index.  
1029 *American Journal of water resources*, 1(3), 34–38.  
1030 [https://www.academia.edu/download/68284301/Water\\_Quality\\_Assessment\\_in\\_Terms\\_of\\_Wat202107](https://www.academia.edu/download/68284301/Water_Quality_Assessment_in_Terms_of_Wat20210724-4216-1hpdqkq.pdf)  
1031 [24-4216-1hpdqkq.pdf](https://www.academia.edu/download/68284301/Water_Quality_Assessment_in_Terms_of_Wat20210724-4216-1hpdqkq.pdf). Accessed 10 December 2024

1032 Uddin, M. G., Nash, S., Mahammad Diganta, M. T., Rahman, A., & Olbert, A. I. (2022). Robust machine  
1033 learning algorithms for predicting coastal water quality index. *Journal of Environmental Management*,  
1034 321, 115923. <https://doi.org/10.1016/j.jenvman.2022.115923>

1035 Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2022). A comprehensive method for improvement of water  
1036 quality index (WQI) models for coastal water quality assessment. *Water Research*, 219, 118532.  
1037 <https://doi.org/10.1016/j.watres.2022.118532>

1038 Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2023a). Performance analysis of the water quality index  
1039 model for predicting water state using machine learning techniques. *Process Safety and Environmental*  
1040 *Protection*, 169, 808–828. <https://doi.org/10.1016/j.psep.2022.11.073>

1041 Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2023b). A novel approach for estimating and predicting  
1042 uncertainty in water quality index model using machine learning approaches. *Water Research*, 229,  
1043 119422. <https://doi.org/10.1016/j.watres.2022.119422>

1044 Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2023c). A sophisticated model for rating water quality.  
1045 *Science of The Total Environment*, 868, 161614. <https://doi.org/10.1016/j.scitotenv.2023.161614>

1046 Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2023d). A novel approach for estimating and predicting  
1047 uncertainty in water quality index model using machine learning approaches. *Water Research*, 229,  
1048 119422. <https://doi.org/10.1016/j.watres.2022.119422>

1049 Uddin, M. G., Rahman, A., Rosa Taghikhah, F., & Olbert, A. I. (2024). Data-driven evolution of water quality  
1050 models: An in-depth investigation of innovative outlier detection approaches-A case study of Irish  
1051 Water Quality Index (IEWQI) model. *Water Research*, 255, 121499.  
1052 <https://doi.org/10.1016/j.watres.2024.121499>

1053 Uddin, Md. G., Nash, S., & Olbert, A. I. (2021a). A review of water quality index models and their use for  
1054 assessing surface water quality. *Ecological Indicators*, 122, 107218.  
1055 <https://doi.org/10.1016/j.ecolind.2020.107218>

1056 Uddin, Md. G., Nash, S., & Olbert, A. I. (2021b). A review of water quality index models and their use for  
1057 assessing surface water quality. *Ecological Indicators*, 122, 107218.  
1058 <https://doi.org/10.1016/j.ecolind.2020.107218>

1059 UNEP. (2021). *UNEP World Water Quality Assessment: A Global Overview*.  
1060 <https://wedocs.unep.org/rest/api/core/bitstreams/943473b7-ee3f-46ae-8e91-ddc9c9f8866e/content>

1061 UNESCO. (2024). *The United Nations World Water Development Report 2024: Water for Prosperity and Peace*.  
1062 Paris. <https://unesdoc.unesco.org/ark:/48223/pf0000388948/PDF/388948eng.pdf.multi>

1063 WHO and UNICEF. (2022). *Progress on Household Drinking Water, Sanitation and Hygiene: 2000-2022*.

1064 Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., et al. (2024, April 23). ggplot2:  
1065 Create Elegant Data Visualisations Using the Grammar of Graphics. [https://cran.r-](https://cran.r-project.org/web/packages/ggplot2/index.html)  
1066 [project.org/web/packages/ggplot2/index.html](https://cran.r-project.org/web/packages/ggplot2/index.html). Accessed 8 October 2024

1067 Xu, J., Mo, Y., Zhu, S., Wu, J., Jin, G., Wang, Y.-G., et al. (2024). Assessing and predicting water quality index  
1068 with key water parameters by machine learning models in coastal cities, China. *Heliyon*, 10(13),  
1069 e33695. <https://doi.org/10.1016/j.heliyon.2024.e33695>

1070 Zhang, X., Yu, B., Xin, Z., Cong, M., & Zhang, C. (2025). Spatial–temporal variations of river water quality  
1071 under human-induced land use changes in large river basins. *Scientific Reports*, 15(1), 36955.  
1072 <https://doi.org/10.1038/s41598-025-20876-z>

1073

Supplementary material

## **Enhancing water quality assessment in Skikda, Algeria using the PCA-based weighted index (WQI\_P) and its predictive performance: a comparison with traditional WA\_WQI approaches**

**BENACHERINE MOSTEFA<sup>1,\*</sup>, Hafid Hinda<sup>1</sup>, Alejandro Garcia Martinez<sup>2</sup>, Allaoua Noua<sup>1</sup>, Satour Abdelatif<sup>3</sup>, Fertas Fadila<sup>4</sup>, Mir Talas Mahammad Diganta<sup>5,6,7</sup>, Lyazid Mohamed Nadjib<sup>8</sup>, Bouchra Debassi<sup>8</sup>, Md Galal Uddin<sup>5,6,7,9,\*</sup>**

<sup>1</sup>Laboratory of Natural Resources and Management of Sensitive Environments, Larbi Ben M'hidi University, Oum-El-Bouaghi 04000, Algeria

<sup>2</sup>Molecular Ecology Group (MEG), Water Research Institute (IRSA), National Research Council of Italy (CNR), Verbania, Italy

<sup>3</sup> Laboratoire de Conservation des Zones humides, Université 8 mai 1945, Guelma, Guelma 24000, Algeria.

<sup>4</sup>LARYSS Laboratory, Mohamed Khider University Biskra, Biskra 07000, Algeria.

<sup>5</sup>School of Engineering, College of Science and Engineering, University of Galway, Galway, Ireland

<sup>6</sup>Ryan Institute, University of Galway, Galway, Ireland.

<sup>7</sup>Eco-HydroInformatics Research Group (EHIRG), Civil Engineering, University of Galway, Galway, Ireland.

<sup>8</sup>Laboratory of Functional Ecology and Environment, University of Oum El Bouaghi, Oum-El-Bouaghi 04000, Algeria.

<sup>9</sup>Department of Civil, Structural and Environmental Engineering, and Sustainable Infrastructure Research & Innovation Group, Munster Technological University, Cork, Ireland

\*Corresponding author(s):

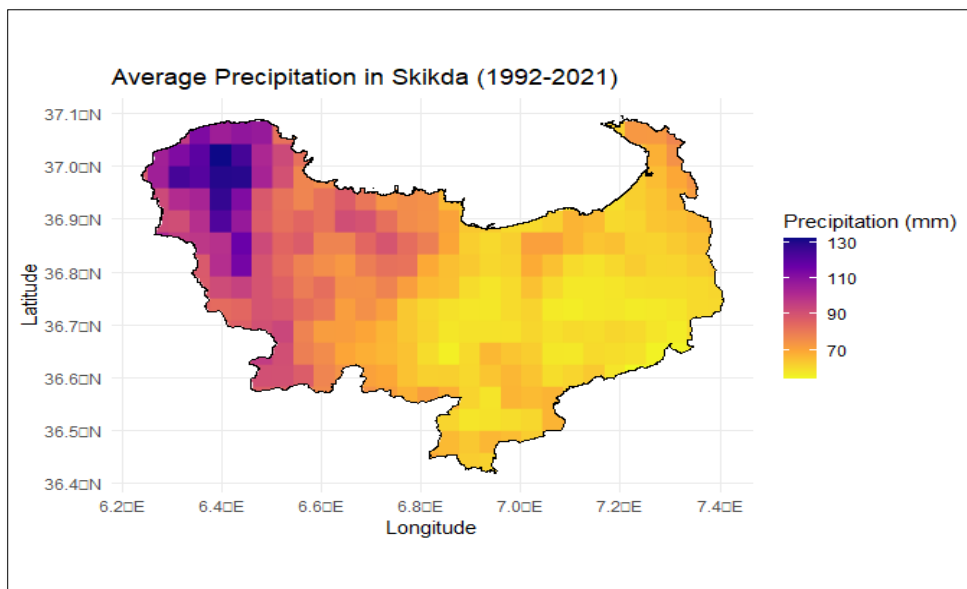
Benacherine Mostefa, Laboratory of Natural Resources and Management of Sensitive Environments, Larbi Ben M'hidi University, Algeria ([mostefa.benacherine@univ-oeb.dz](mailto:mostefa.benacherine@univ-oeb.dz))

Dr Md Galal Uddin, Postdoctoral researcher, Civil Engineering, College of Science and Engineering, University of Galway, Ireland([mdgalal.uddin@universityofgalway.ie](mailto:mdgalal.uddin@universityofgalway.ie))

The code used for this study is provided. The complete file is available under the name "Feuille 23" (<https://github.com/mosetfa/mosetfa/blob/main/feuille%2023>).

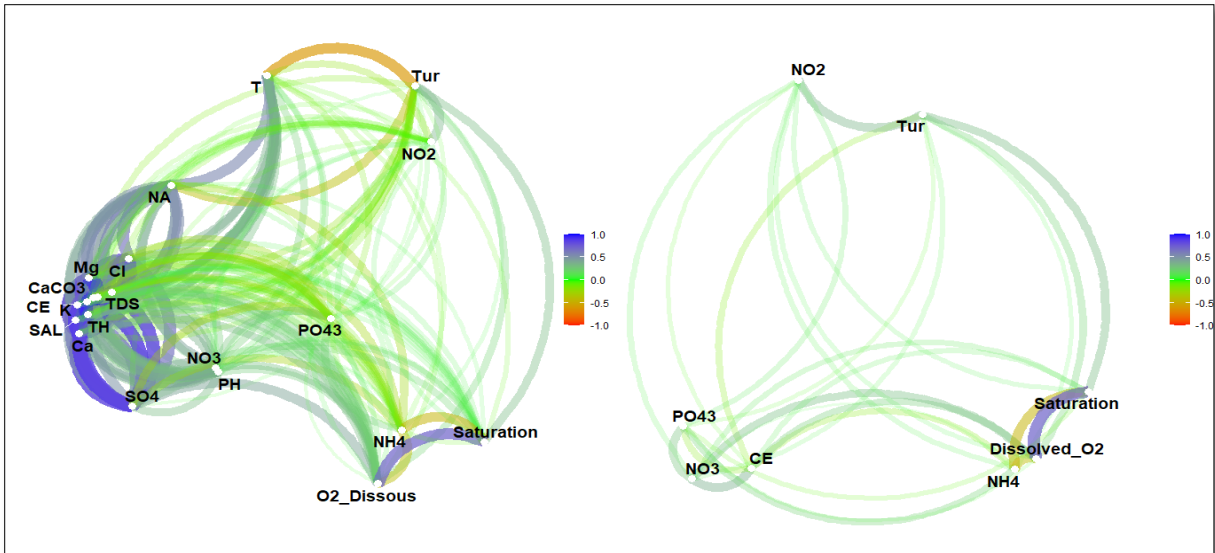
Table 01: Permissible Limits of the Parameters Used in Calculating the WQI. (RAPPORT DE PRESENTATION SEQ-EAU)

Parameters	Permissible thresholds
Conductivity	2500 $\mu\text{s}/\text{cm}$
Turbidity	1 NTU
Nitrite	0,03 mg/l
Nitrate	50 mg/l
Ammonium	0,5 mg/l
Saturation %	70 %
O <sub>2</sub> , mg/l	7 mg/l
Orthophosphate	0,1 mg/l



**FIGURE 01:** Heatmap of average monthly precipitation in Skikda from 1992 to 2021.

We utilized R to systematically download and process climate data from the TerraClimate dataset. The map (figure 2) depicting average monthly precipitation in the Skikda region from 1992 to 2021 reveals an uneven distribution of rainfall, with the highest values predominantly concentrated in the northwestern area. The dark blue zones on the map indicate regions where average monthly precipitation can reach up to 130 mm, highlighting these areas as particularly wet. This is where the Tamanart, Tizaghbane, and Zhor wadis are located, selected for their accessibility and favorable climatic conditions that support high precipitation levels, as illustrated by the map.



**Figure 02: Comparative Correlation Network Graphs** (Full Set of 21 Parameters vs. Representative Subset of 8 Parameters). This figure illustrates the correlation matrix between the 21 physicochemical variables, while the second figure shows the correlation only between the 8 selected parameters. This is done to highlight that these selected parameters capture a range of information from the first figure, effectively summarizing it by focusing on the 8 parameters. The associations between the parameters indicate the correlation coefficient (either positive or negative), with the color also reflecting the intensity of the correlation.

The network matrix reveals a strong correlation between electrical conductivity (CE) and most major ions, such as Na, Cl, Ca, and Mg, indicating that an increase in the concentration of these ions results in higher conductivity. Additionally, a significant correlation is observed between dissolved oxygen and oxygen saturation, which reflects their direct relationship. However, notable negative correlations are present, particularly between ammonium (NH<sub>4</sub>) and both Dissolved\_O<sub>2</sub> and saturation (%). Nutrients such as nitrate (NO<sub>3</sub>) and phosphate (PO<sub>4</sub>) also show significant correlations, suggesting crucial interactions within the biogeochemical cycles of the water.

The two correlation network graphs presented illustrate, on one hand, the correlations among 21 parameters and, on the other, those within a subset of 8 parameters carefully selected to minimize informational redundancy. A comparative analysis of these graphs reveals a marked visual similarity, suggesting that the 8 chosen variables effectively capture the essential correlation patterns present in the full set of 21 variables. This similarity confirms that the selected variables are both representative and sufficient for further analyses without any significant loss of information. The strongest correlations, indicated by thicker and more vividly colored lines, remain evident in the reduced graph, demonstrating that the most significant relationships have been preserved despite the reduction in the number of variables. This process of variable selection proves its effectiveness in simplifying the model while maintaining data integrity, thus making the analysis more concise and interpretable without compromising the quality of crucial information.

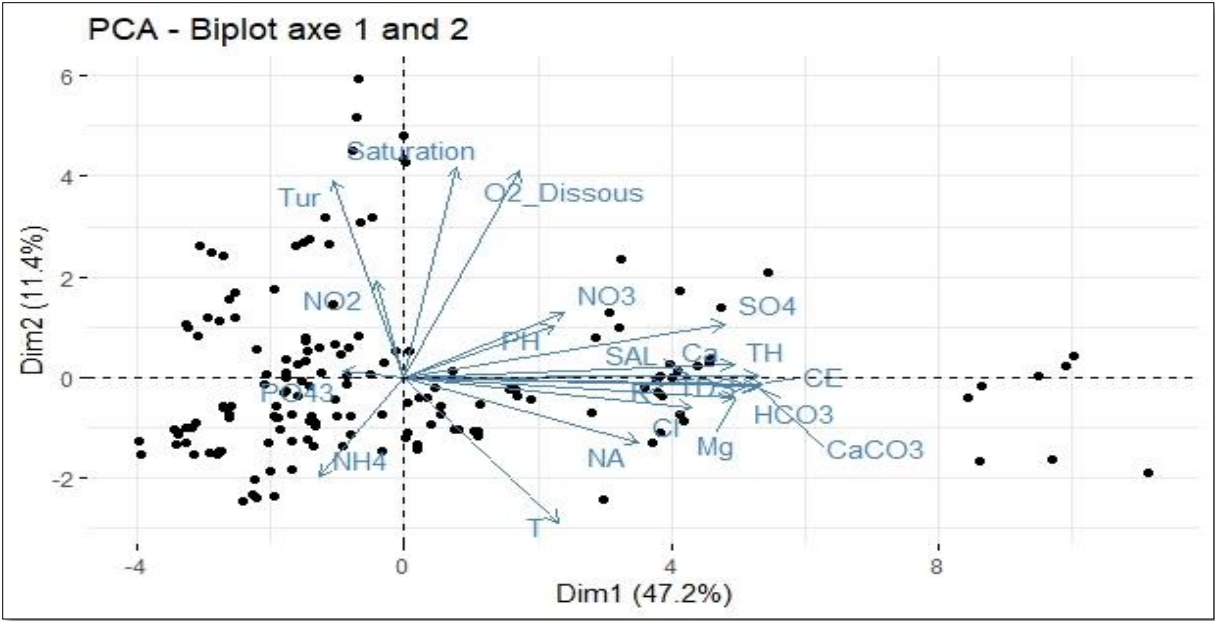


Figure03: PCA biplot on axes 1 and 2, showing only variable vectors.

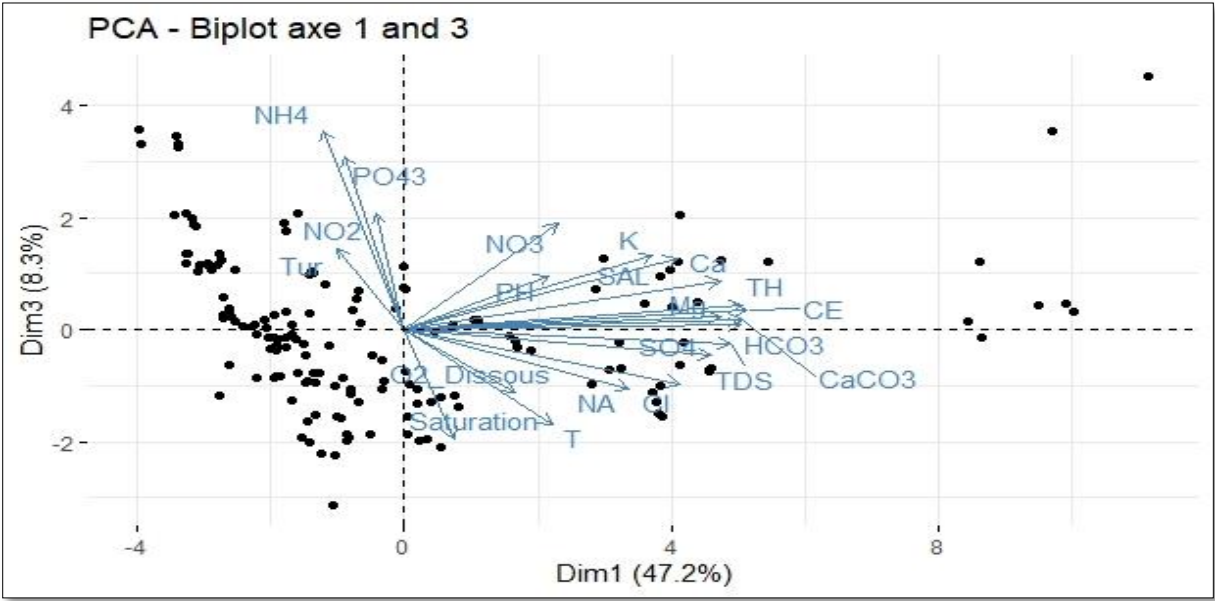


Figure04: PCA biplot on axes 1 and 3, showing only variable vectors.

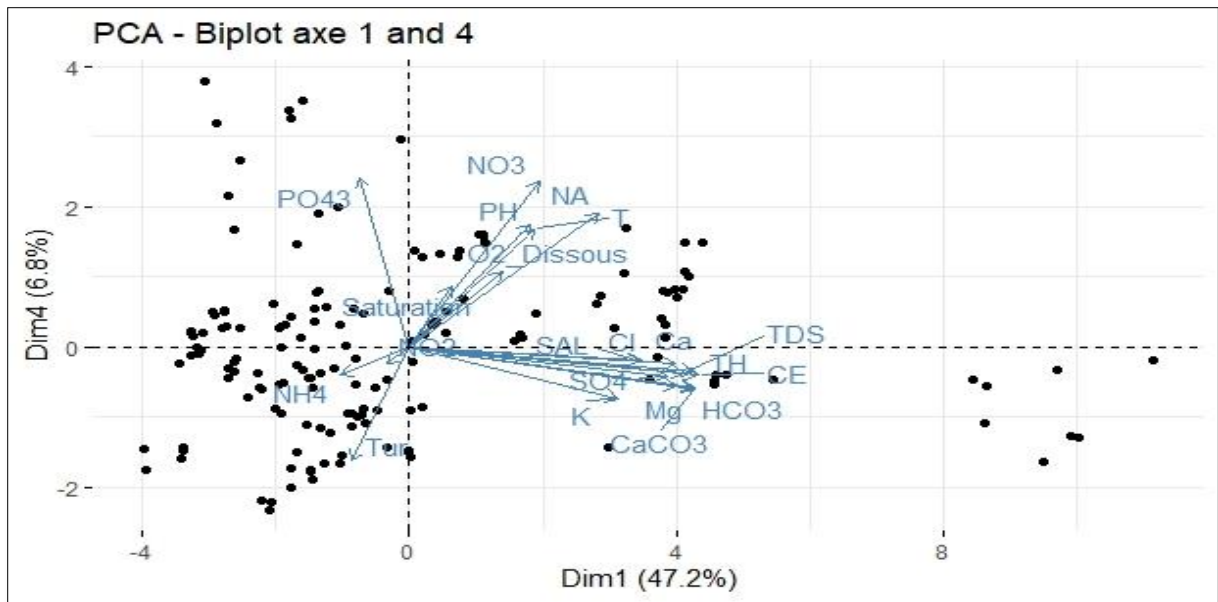


Figure05: PCA biplot on axes 1 and 4, showing only variable vectors.

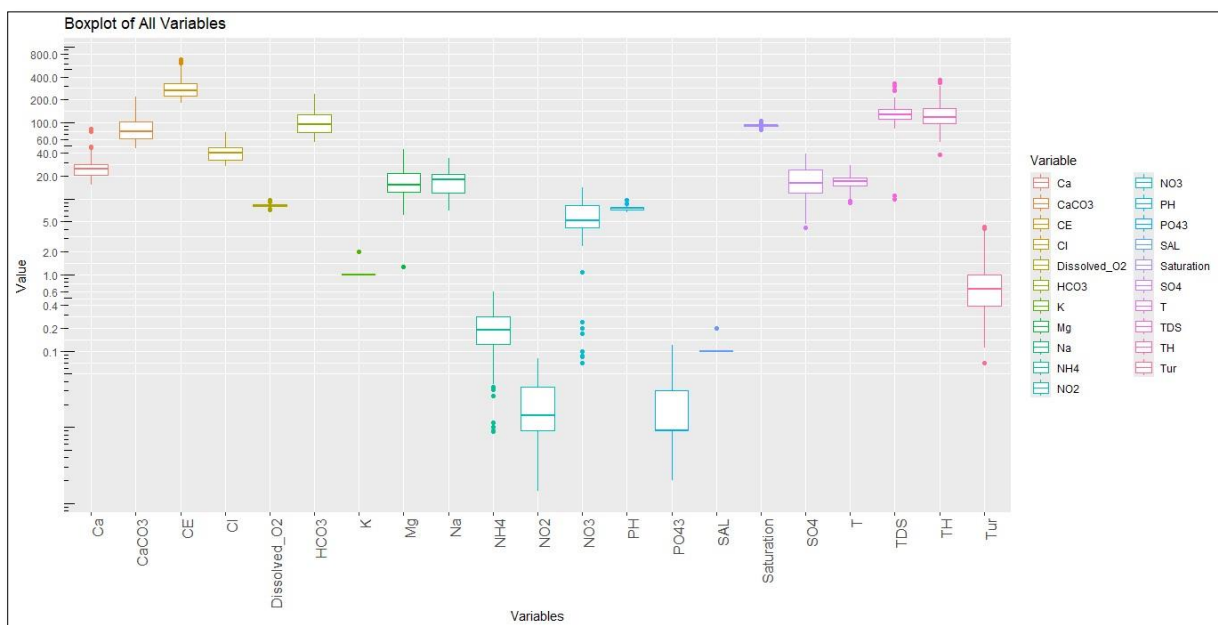


Figure 05.Values of environmental parameters in Skikda river water samples.

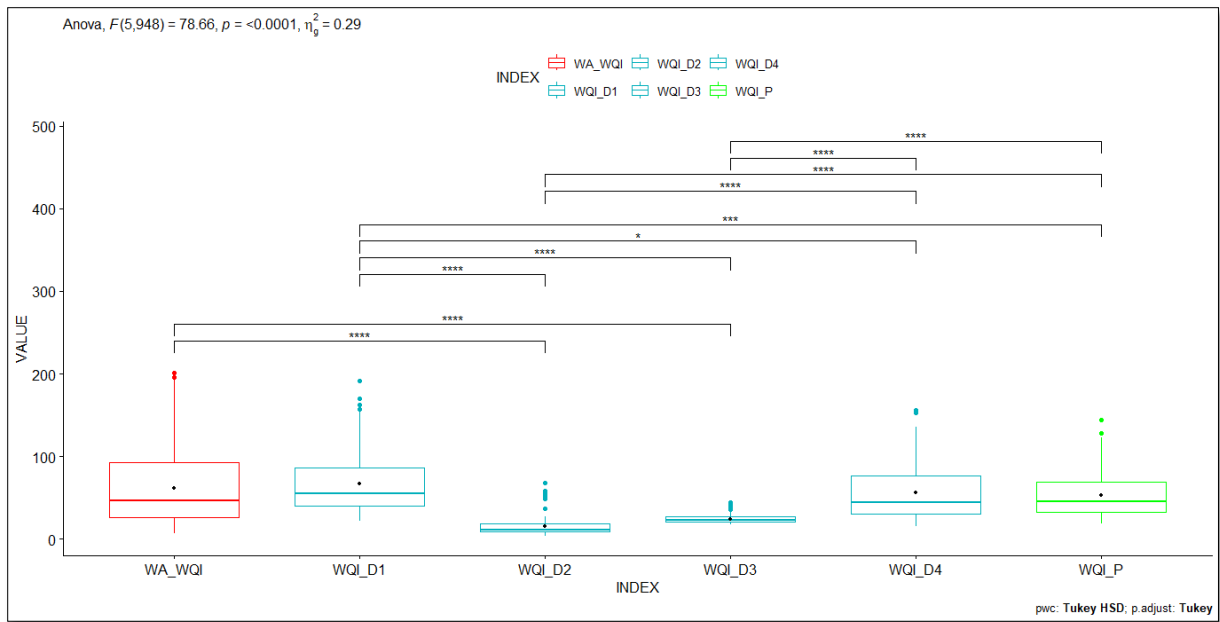


FIGURE 06: ANOVA Analysis with Post-Hoc Tukey HSD: Comparative Boxplots of Water Quality Indices (WQI).