Water Supply

IWA PUBLISHING

Water Supply Vol 24 No 1, 204 doi: 10.2166/ws.2023.336

Optimization of the water quality monitoring network in a basin with intensive agriculture using artificial intelligence algorithms

Kimberly Mendivil-García[a], José Luis Medina[b], Héctor Rodríguez-Rangel[b], Adriana Roé-Sosa[c] and Leonel Ernesto Amábilis-Sosa[a,*]

[a] División de Estudios de Posgrado e Investigación, CONACYT-Tecnológico Nacional de México/IT de Culiacán, Av. Juan de Dios Batiz, No. 310, 80220, Culiacán, Sinaloa, México
[b] División de Estudios de Posgrado e Investigación, Tecnológico Nacional de México/IT de Culiacán, Av. Juan de Dios Batiz, No. 310, 80220, Culiacán, Sinaloa, México
[c] Universidad Tecnológica de Culiacán, Carretera Imala km 2, C.P. 80014, Culiacán, Sinaloa, México
*Corresponding author. E-mail: leonel.as@culiacan.tecnm.mx

LEA, 0000-0002-9020-1951

ABSTRACT

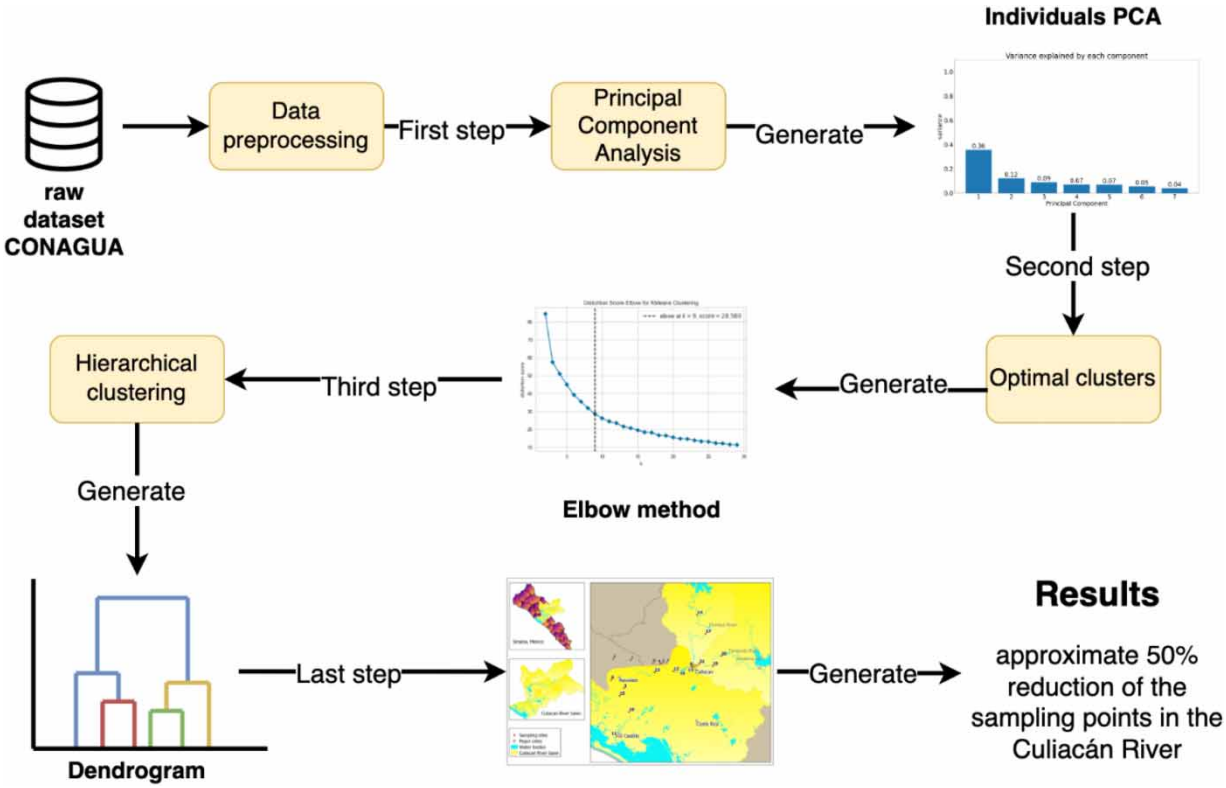
This research applies artificial intelligence algorithms for optimizing the water quality monitoring network in a representative basin with intensive agricultural and livestock activities. This study used the water quality database provided by the National Water Commission (CONAGUA). Bi-monthly monitoring was registered from 2013 to 2020 for 23 water quality parameters in 23 sampling locations in tributaries and the mainstream river. Therefore, it was necessary to apply principal component analysis to reduce the dimensionality of the data and thus identify the parameters that contribute most to the variation in the water quality. This artificial intelligence algorithm promoted the ease of clustering sampling sites with similar water quality characteristics by reducing the number of variables involved in the database. The reduction highlighted nutrients (TN and TP), parameters related to dissolved organic matter (NH_3-N and TOC), and pathogens such as fecal coliforms. The similarity of sampling sites was determined through hierarchical clustering using the Euclidean distance as a measure of dissimilarity and the Ward method as a grouping method. As a result, nine clusters were obtained for the rainy and dry seasons, reducing approximately 50% of the sampling sites and generating an optimized network of 11 sampling sites.

Key words: agricultural watershed, artificial intelligence, land uses, monitoring network, optimization, water quality

HIGHLIGHTS

- The monitoring network of a watershed with intensive agriculture was reduced by 50% using artificial intelligence algorithms.
- By applying principal component analysis, the variables that contribute the most significant variation to the water quality of the basin, highlighting nutrients, and pathogens were identified.
- It was possible to agglomerate sampling sites according to their similarity in terms of water quality.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Deterioration of the water quality in aquatic ecosystems has become a global concern as population growth generates a greater need for fresh water. The physicochemical and biological interactions of pollutants in the rivers and the diversity of diffuse sources make it difficult to implement specific action plans and mitigation measures that allow for the sustainable use of basins. This problem is more accentuated in developing countries due to a lack of environmental regulations and a need for more technical exchange between companies and environmental protection agencies to assess ecological impacts and develop sustainable practices in line with current global problems such as water quality (Giao *et al.* 2022). In addition, technical and scientific analyses are limited in developing countries compared to developed countries, and therefore, governmental monitoring programs and plans are necessary to protect the water quality of surface water bodies.

Among developing countries, Mexico stands out for its agricultural activity with approximately 21.6 million ha (SIAP 2020). Sinaloa is the state with the highest agricultural production, with 1.2 million ha and producing approximately 12,170 tonnes of food. The fertilization requires 200 kg N ha^{-1} and 50 kg P ha^{-1} year. The water requirement to carry out this economic activity amounts to 90 hM3 (CONAGUA 2020). On the other hand, deforestation due to land use change has caused a more significant transport of pollutants to the river due to soil impermeability. Consequently, eutrophication has been observed throughout the Culiacan River basin, proportional to the land use. Thus, the need to preserve the economic activity of the region makes it essential to maintain a good water quality and the easiest way to achieve it is by preserving the ecosystem services through the evaluation of physical–chemical parameters.

In developed countries, the monitoring of aquatic ecosystem's health consists of using smart environment monitoring systems through the Internet of Things (IoT) and sensors that allow obtaining real-time reliable results that permit the extensive automated database generation (Chowdury *et al.* 2019; Jan *et al.* 2021). In Mexico, through the National Water Quality Measurement Network (RENAMECA), the National Water Commission has established monitoring networks in the country's principal rivers and its procedures for analyzing water quality parameters such as heavy metals, nutrients, and pathogens. There is a record of at least 10 years of historical data for more than 20 parameters registered every 2 months

manually through fieldworks. However, this process generates a complex data matrix that makes it difficult to analyze, which in developing countries complicates the continuity of data analysis due to limited resources. Thus, the application of multivariate statistical techniques, such as principal component analysis (PCA), and artificial intelligence techniques, such as cluster analysis, helps to interpret complex data matrices to better understand the water quality and ecological status of the surface water bodies (Egbueri 2020; Yang *et al.* 2020). In addition, these techniques assist identifying possible pollution sources influencing water quality, as well as a quick solution to pollution problems.

PCA is one of the data exploration and analysis methods used in many fields, which belongs to dimensionality reduction methods. The PCA technique has the approach of representing the original data in lower dimensional data with as little loss of information as possible. The advantage of implementing PCA is the reduction of the original correlated variables into a set of principal components, a set of linearly uncorrelated variable values (Mechelli & Viera 2019).

In waterbody analysis, this method helps identify the parameters that mostly influence water quality and are possible sources of pollution. In addition, unsupervised learning by machine learning, offers tools such as clustering for the analysis of the similarity between data and creating groups, also techniques to determine the optimal number of clusters, such as elbow and silhouette methods (Krishnaraj & Deka 2020; Li *et al.* 2022). This grouping eliminates redundant monitoring sites on a network for more efficient and objective work.

Mamun *et al.* (2021) used PCA to evaluate the spatiotemporal variations in 12 water quality parameters of a reservoir for 23 years. The authors identified that nutrients and organic matter (anthropogenic) are responsible for water quality variation. Gyimah *et al.* (2021) used PCA and cluster analysis to analyze 16 water quality parameters in 10 sampling stations over four months, obtaining a classification of sampling sites that allowed identifying the leading cause of water quality is anthropogenic activity, so this suggested to pre-treat the effluents. Meanwhile, Kumar *et al.* (2020) studied three water bodies using cluster analysis and PCA using data from 4 years. The authors identified the source of pollution by heavy metals in every water bodies supported by the cluster analysis which grouped the heavy metals by similarity of concentrations.

Since the previous studies used cluster analysis and PCA as analytical tools, the present study proposes their combination enhanced by non-supervised algorithms for selecting the optimal number of groups in the clusters, such as the elbow method. In this way, it will go from being qualitative to a quantitative application because a poor selection of clusters can result in grouping very heterogeneous data or data that are very similar to each other and are grouped in different clusters. This combination may offer a methodology to extract information on similarities or dissimilarities between sampling sites, identification of water quality variables responsible for spatial and temporal variations in river water quality, and the influence of possible sources on water quality parameters by obtaining a different approach for clustering sampling sites in the Culiacan River that would be of great importance for quick intervention if necessary (Tung & Yaseen 2020; Wang *et al.* 2021; Zhu *et al.* 2022).

The objectives of this research are (i) to propose and evaluate a methodology for the grouping of sampling sites, based on different multiparametric methods, including artificial intelligence, with indicators of precision and numerical sensitivity that allow corroboration based on the background generated from the study area given its economic importance and the agrifood industry including the expert's opinion and (ii) elucidate the clustering results at the hydrological basin level considering the hydrological, hydrodynamic and land use-water quality background of the study basin. Together, the objectives will optimize a complex monitoring network to streamline the assessment of water bodies and develop strategies according to current environmental needs in regions where land use analysis and hydrological information of water bodies are available.

2. METHODS

2.1. Study area and data collection

The Culiacan River basin is the second largest in Mexico regarding drained area and belongs to hydrological region 10 (RH10). It has a surface area of 19, 150.49 km^2, of which 9, 143.49 km^2 belong to Sinaloa. This portion has an average rainfall of 771 mm, with June to September being the region's rainy season (SEMARNAT 2021). The prevailing climate is warm and semi-dry, and the average annual temperature is 25 °C with minimum being near 10 °C, mainly in January, and maximum around 36 °C from May to July (INEGI 2021).

Figure 1 shows the Culiacan River basin. The mainstream is the Culiacan River, which is formed at the confluence of its two tributaries: Humaya and Tamazula, which converge in the city of Culiacan until it flows into the Altata-Ensenada del Pabellon lagoon system. It has an average slope of 0.09%, considered medium to low (INEGI 2019). The current monitoring

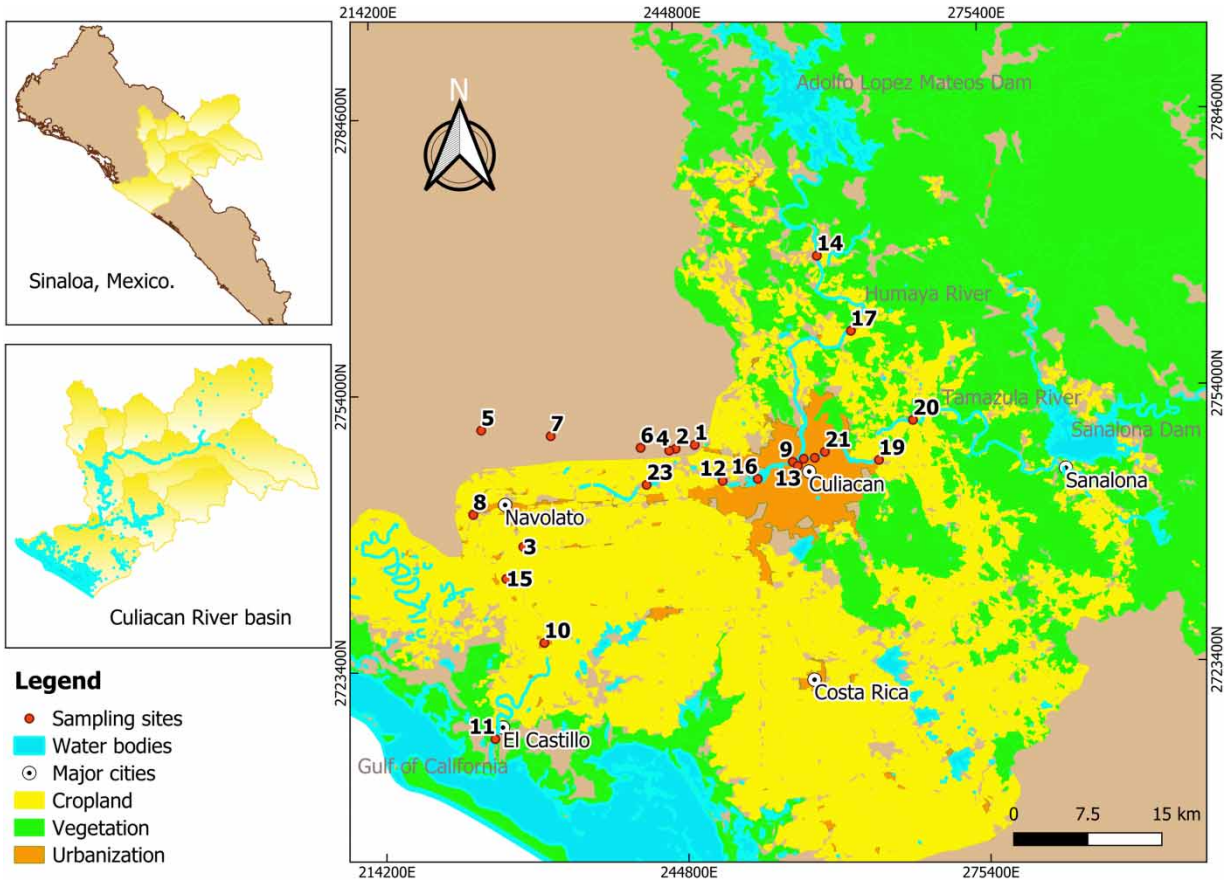


Figure 1 | The Culiacan River basin and its mainstreams.

network of the National Water Commission (CONAGUA) was used for the study. The network comprises 23 sampling sites distributed along the Humaya, Tamazula, and Culiacan Rivers, and the water quality in every sampling site is influenced by different land use.

Water quality parameters were analyzed bi-monthly, including nutrients, pathogens, and heavy metals in 2013–2020 (Table 1). Accredited laboratories carried out these determinations under international standards. Preserving and transporting the samples were conducted according to APHA (2005), establishing the procedure for sampling receiving bodies. An independent analysis was considered for the region's two main climatic seasons: the dry season from September to June and the rainy season from June to September.

2.2. Determination of land use in the Culiacan River basin

Land uses (agriculture, urbanization, and vegetation) were evaluated to elucidate the physicochemical characteristics of the sampling sites that are classified into the same groups. Land use vector layers provided by the National Institute of Statistics were analyzed by the geographic information system Qgis 3.4 (INEGI 2018). In addition, agriculture information was supported by data from the Agrifood and Fisheries Information Service (SIAP), as it presents a better-quality definition in the detail of the vector layers.

Circular buffers were created to determine the influence of each land use on the sampling sites with an area of influence of 10 km^2 (Figure 2). Table 2 shows the information extracted from the geographic information system through the 10-km^2 buffers. The percentage of land use around each sampling site of the river basin monitoring network can be observed. According to the site's location in the basin, land use is predominant. For example, at site 15 in 10 km^2 area, 83.31% is agriculture, and 7.02% is urbanization without native vegetation. This help to understand the clustering process.

Table 1 | Water quality parameters measured in the Culiacan River basin

Parameter	Reference
Chlorophyll, mg/m^3	SM 10200 H
NH_3-N, mg/L	EPA 350.1
NO_3-N, mg/L	EPA 353.2
N Org, mg/L	EPA 351.2
Total nitrogen (TN), mg/L	
PO_4-P	
Total phosphorus (TP), mg/L	EPA 365.1
E. coli, MPN/100 mL	NMX-AA-042-1987
Fecal coliforms, MPN/100 mL	NMX-AA-042-1987
Total coliforms, MPN/100 mL	NMX-AA-042-1987
Electrical conductivity., μS/cm	NMX-AA-093-SCFI-2000
TOC, mg/L	EPA 415.3
Dissolved organic carbon, mg/L	EPA 425.1
BOD, mg/L	NMX-AA-028-SCFI-2001
COD, mg/L	NMX-AA-030/2-SCFI-2011
Dissolved oxygen, mg/L	NMX-AA-012-SCFI-2001
TSS, mg/L	NMX-AA-034-SCFI-2001
CN, mg/L	NMX-AA-058-SCFI-2001
Cr, mg/L	EPA 6010 C
Hg, mg/L	EPA 7470 A

2.3. Singular value decomposition

The singular value decomposition (SVD) was implemented because this method generalizes the diagonalization of matrices. SVD is a method applicable to non-square matrices (Wall *et al.* 2003). Matrix diagonalization is represented by Equation (1):

$$M_{\text{mxm}} = P_{\text{mxm}} D_{\text{mxm}} P^{-1}_{\text{mxm}} \tag{1}$$

where M_{mxm} is the matrix of the sampling points with relation to the water quality parameters, P_{mxm} is an invertible matrix that is square, D_{mxm} is the diagonal matrix containing all zeros except along the diagonal, and P^{-1}_{mxm} is the transposed matrix.

The data implemented from CONAGUA with a matrix approach has 23 parameters captured since 2012, with more than 700 samples performed. Due to the characteristics described above, the data matrix is non-square, hence SVD is appropriate for the PCA. SVD is defined by Equation (2):

$$M_{\text{mxn}} = U_{\text{mxm}} S_{\text{mxn}} V^T_{\text{nxn}} \tag{2}$$

where M_{mxn} is the matrix of the sampling points with relation to the water quality parameters, and it is decomposed into three matrices denominated U, S y V^T_{nxn}. Where, U_{mxm} is an orthonormal matrix, integrated by sampling points captured since 2012, V^T_{nxn} is an orthonormal matrix represented by all parameters of water quality and both contain left and right singular vectors. Finally, the matrix D contains the singular values, these being non-negative numbers and of the same size as the rank of the matrix M.

SVD is since for any matrix M, the M^TM and MM^T matrices are symmetric. From this, U_{mxm} contains the eigenvectors of MM^T, V contains the eigenvectors of M^TM, and the diagonal of S contains the square root of the eigenvalues associated with M^TM and MM^T.

One of the procedures executed for the implementation of SVD involves the standardization of the data set. Specifically, each input feature or column is adjusted to have a mean of zero and a standard deviation of one. Subsequently, the correlation

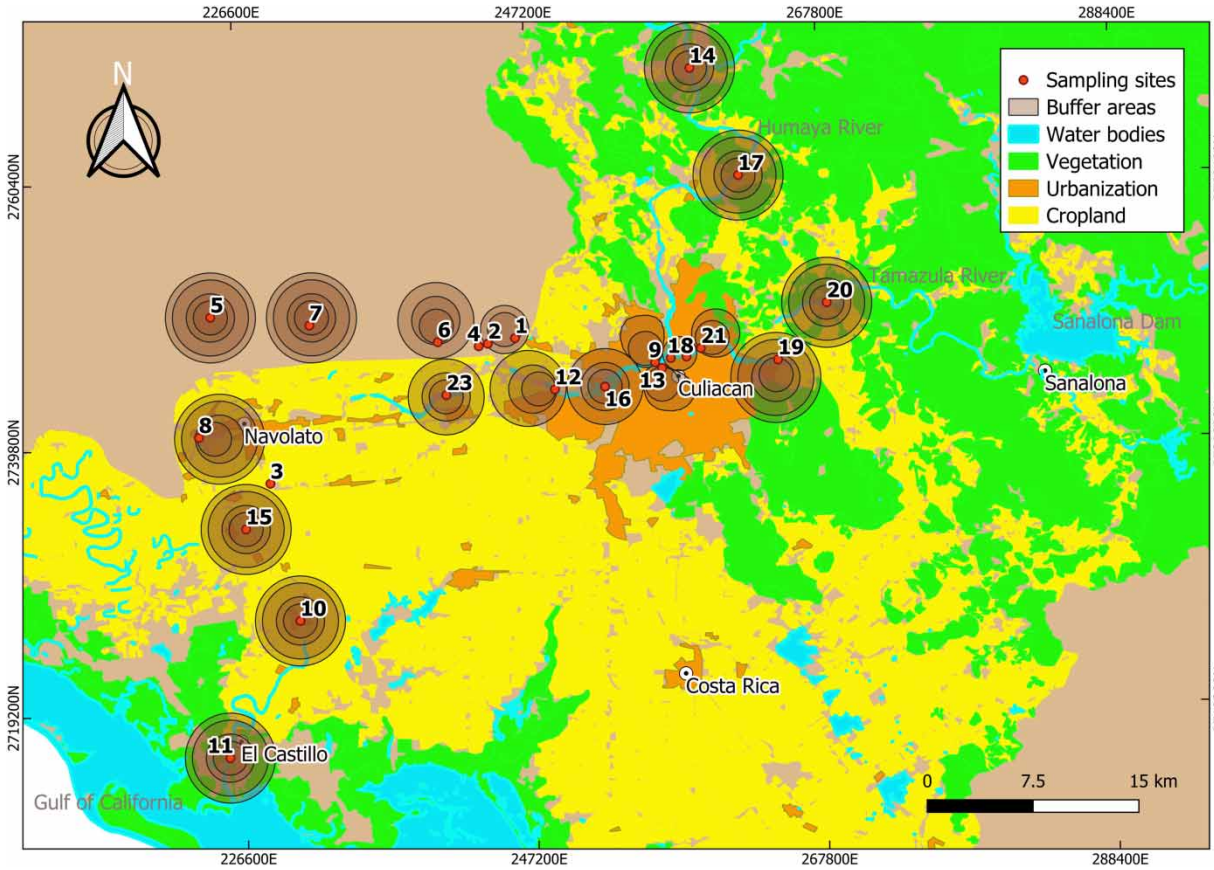


Figure 2 | Circular buffers on the monitoring network in the Culiacan River basin.

matrix is computed based on the standardized data, as delineated by Equation (3):

$$R = \frac{1}{n-1} M^{T}_{\text{estandardized}} M_{\text{estandardized}} \tag{3}$$

If the correlation matrix is substituted into the SVD equations (Equations (2) and (3)), the resulting formulation is described by Equation (4):

$$R = \frac{1}{n-1} VSV^{T} \tag{4}$$

where the R eigenvectors represent the columns of V, which represent the directions in the original feature space where the correlation is maximized. The R eigenvalues are proportional to the squares of the singular values in S, indicating the amount of correlation captured by each principal component. Consequently, the columns of V serve as the principal components, and the magnitude of the singular values in S delineates the significance of each principal component in terms of explained variance.

2.4. Cluster analysis for monitoring network sites

2.4.1. Determining the optimal number of clusters: the elbow method

K-means is one of the most widely used unsupervised learning algorithms to solve clustering problems, where it classifies unlabeled data into different given groups, denoted as the K value.

The elbow method was selected objectively to determine the number of clusters to provide a starting reference point to determine the number of clusters to be considered and, thus, to have an optimal cluster value before hierarchical clustering.

Table 2 | Monitoring network and land use influence at 10 km^2

No.	Sampling site	Land uses (%)		
		Agriculture	Urbanization	Vegetation
1, 2, 4	Culiacan Nte. Agua Arriba, Culiacan aguas abajo, Puente Ferrocarril	79.98	16.53	–
3	Aguas Arriba PTAR Navolato	–	–	–
5	Puente El Guamuchilito	85.01	–	–
6	Puente El Pinole	82.04	3.79	–
7	Puente La Palma	83.06	8.1	–
8	Puente Limoncito	54.03	20.56	–
9	Puente Negro	–	86.87	–
10	RH10-2 Sinaloa	82.84	10.67	–
11	Rio Culiacan 2	52.39	4.6	3.4
12	Río Culiacan 4	41.6	33.81	–
13	Río Culiacan 5	–	99.85	–
14	Río Culiacan 7	27.76	–	10.41
15	La Pipima	83.31	7.02	–
16	Puente USE	–	74.79	–
17	Río Humaya	40.09	2.97	34.43
18, 22	Río Tamazula 1, Puente Morelos	–	99.9	–
19	Río Tamazula 2	27.54	14.78	25.92
20	Río Tamazula 3	41.42	–	39.77
21	Puente Juárez	–	86.8	12.97
23	San Pedro	52.69	12.02	–

Using the total quadratic sum within the cluster defined by Equation (5), which is also the representation of '*K*-means', the aim is to obtain the number of clusters with the smallest quadratic sum:

$$J = \Sigma_{j=1}^{k}\Sigma_{i=1}^{n}||x_i^j - C_k||^2 \tag{5}$$

Each value obtained from clusters is associated with the sum of squares, where k is the number of clusters to explore, n is the total number of samples collected by CONAGUA over the years. In addition, $||x_i^j - C_k||^2$ represents the function of the distance to implement between the sample and the cluster centroid, where x represents the sample to be grouped, and c is the cluster centroid.

2.4.2. Data preprocessing

Before implementing clustering, it is important to process the data to adapt them and achieve the most efficient results. The implemented database contains information from the 23 sampling points of the Culiacan River, which has over 20 water quality parameters, collected from 2012 to 2020. Some values need to be included and provide a good estimation for clustering because they are considered null, and it is impossible to compare them to determine the similarity. The missing data are considered non-relevant information, so the entire row to which they belong is removed.

On the other hand, the variability between parameters, i.e., the ranges between them, change depending on the parameter. Data normalization is implemented for each parameter between 0 and 1 to ensure a good quality of input information for clustering.

Data were also separated according to the time of the year to perform the analysis by the dry season, rainy season, and annual season to determine the behavior of each one of them.

2.4.3. Similarity measurement: Euclidean distance

An important step in clustering is the measurement of the similarity of the grouped points to know the homogeneity of the data grouping and, in turn, to know how heterogeneous each of the groups created.

The measurement is performed by using the Euclidean distance, which works with the WQ data matrix. The Euclidean distance is defined by Equation (6):

$$D(x,\ y) = \sqrt{\Sigma_{i=1}^{n}(x_i - y_i)^2} \tag{6}$$

where $D(x,\ y)$ is the Euclidean distance between the cluster and the sampling point, y_i represents the input value to be grouped, in this case, the sampling points of the Culiacan River, and x_i is the centroid value of the different groups created. Each input value is iterated with the different groups to measure the similarity.

One of the purposes of the metric is to find the value closest to zero to determine to which group the input value belongs or, in its respective case, to determine how the water quality sampling points can be grouped for reduction.

2.4.4. Hierarchical clustering

Once the data have been preprocessed and the optimal number of clusters for this case has been determined, the hierarchical clustering technique is implemented.

The implementation of hierarchical clustering has the facility to represent the construction of the clusters from different levels of granularity. As a result, a binary tree is constructed, where the sampling points that are most closely linked to each other are joined together.

The graphical representation of hierarchical clustering is called a dendrogram. The dendrograms are plotted based on the metric implemented to differentiate the homogeneity of each cluster, in this case, the Euclidean distance. The *y*-axis corresponds to the Euclidean distance, and the *x* represents the sampling points.

This analysis was performed with the average values of the water quality parameters in each sampling site using the Ward method and the Euclidean distance as a measure of similarity.

3. RESULTS AND DISCUSSION

3.1. Reduction of the data matrix for the water quality analysis in the Culiacan River Basin

A PCA was applied to reduce the dimensionality of the water quality data matrix in the Culiacan River basin. This reduction was performed for the rainy and dry seasons. Figure 3(a) shows the component's sedimentation graph, where the variation percentage explained by each component (set of water quality parameters) is represented on the ordinate and expresses the components in decreasing order on the abscissa.

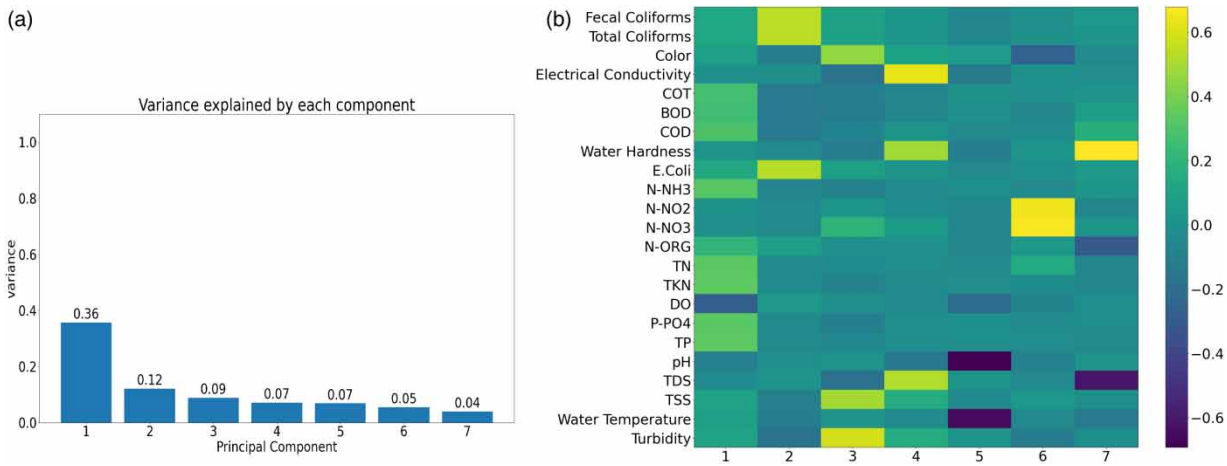


Figure 3 | PCA for the dry season: (a) component's sedimentation graph and (b) loads by component to water quality parameters in the rainy season.

The purpose of the analysis was to obtain a small number of linear combinations of the 17 variables that account for most of the variability in the data. In this case, seven components were extracted. Together, they account for a cumulative variance of 80% of the original water quality data variability. In an absolute value, parameters closest to 1.0 were selected according to the eigenvalue.

Figure 3(b) shows the loads for every variable in each principal component in the dry season. Within component 1, TP has an eigenvalue of 0.33761, thus the most relevant for the higher variation to the entire data set. Similarly, TN has a load of 0.33206 and NH_3-N 0.32394, while TOC and DO show a load of 0.26834 and −0.28394, respectively. Thus, during the dry season, the parameter sets are composed of nutrients and the parameters are related to dissolved organic matter. This is explained by the fact that in the dry season, maize and beans are planted in the region using organophosphate agrochemicals and nitrogen fertilizers. When irrigated through a furrow, the unutilized nutrients by plants reach the body of water through runoff – making it easier to identify the probable source.

Also, the variable fecal coliforms were selected into component 2, which shows an eigenvalue of 0.54175. Although the eigenvalues of fecal and total coliforms are very close, the first one was selected because they are more closely related to the type of wastewater generated in the urban area and, therefore, provides greater objectivity in the study (Makuwa *et al.* 2020; Zhang *et al.* 2021). Although *Escherichia coli* shows a high weight, this is a specific type of fecal coliform, which the abovementioned parameter would represent. This result also agrees with Ahmed (2019), where parameters like NH_3 and fecal coliforms were the most dominant parameters for the PCA in a basin with agricultural and urban activities.

Figure 4(a) shows the sedimentation graph for the rainy season. The principal components are represented on the abscissa axis and the variation of each component in the original data set on the ordinate axis. In this case, seven components represent 79% of the variability in the original data, whereas component 1 represents 35% of the variation of the data set.

Figure 4(b) shows the eigenvalues of the water quality parameters analysis in the rainy season in every principal component. Within component 1, TP shows a load of 0.336479, the most relevant for the higher variation of the data set. TN has a load of 0.334654, NH_3-N has a load of 0.331582, while DO and TOC have loads of −0.27957 and 0.277802, respectively. Thus, during the rainy season, the parameter sets are also composed of nutrients and parameters related to dissolved organic matter, such as the dry season showed. This result agrees with Ustaoğlu *et al.* (2019), where the most relevant parameters in component 1 were TP and TN in the rainy season in an agricultural basin. On the other hand, the fecal coliform parameter was selected for the rainy season analysis within the second principal component, showing an eigenvalue of 0.520582. This parameter shows the greatest effect on component 2.

Based on the loads by component to water quality parameters in each season, a selection of parameters was made to define the variables sensitive to anthropogenic activity. The depuration provided that the parameters selected by components were the same for dry and rainy seasons, suggesting no significant difference in parameters between the seasons.

Indeed, the parameters with the most significant statistical relevance in the two principal components are those that explain the eutrophication of surface water quality integrated by nutrients. The Culiacan River basin has a 15% of the

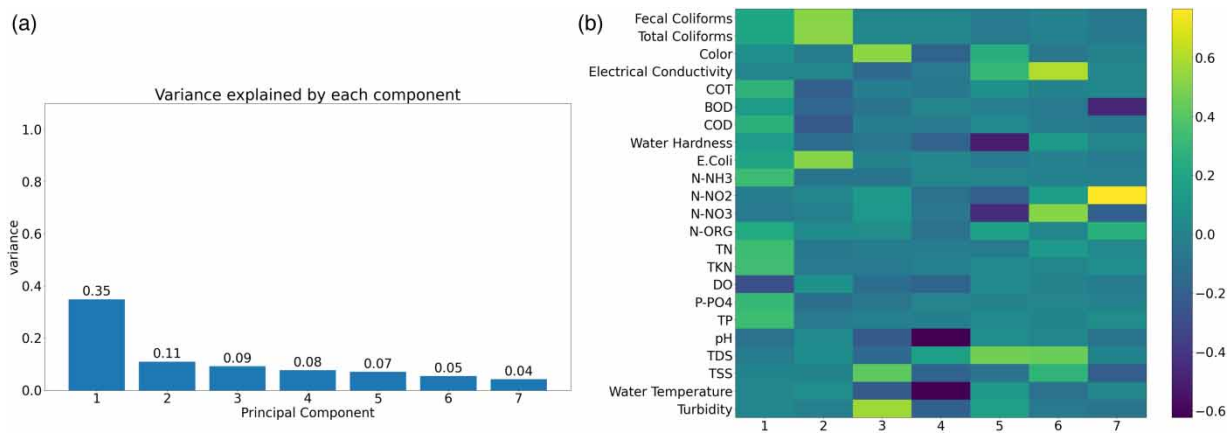


Figure 4 | PCA for the rainy season: (a) sedimentation graph for the rainy season and (b) loads by component to water quality parameters in the rainy season.

area occupied by cropland that uses organophosphorus and nitrogen agrochemicals. On the other hand, the parameters representing dissolved organic matter in the current also stand out. This idea may be associated with drains that receive discharges from the livestock and swine industry, increasing organic matter and directly affecting oxygenation. Finally, the parameter measures the presence of pathogenic contamination, which is strongly related to urban areas and wastewater discharges.

Therefore, the water quality behavior in the Culiacan River basin agrees with Dudley *et al.* (2020) and Yadav *et al.* (2019). These results can infer that the type of contamination is characteristic of surface bodies along with the combination of agricultural and urban areas. Therefore, the parameters that contribute the most significant variability to the water quality of the Culiacan River and that will serve for the generation of clusters are total organic carbon (TOC), ammonia nitrogen (NH_3-N), total nitrogen (TN), total phosphorus (TP), dissolved oxygen (DO), and fecal coliforms.

In Figures 3 and 4, there are parameters with similarity in terms of the eigenvalues of the selected parameters, such as in the case of PO_4 and TP, as well as TN and NTK, and only those parameters that consider all forms of nitrogen and phosphorus in this case, as these parameters are used in both Mexican and international regulations to quantify the water quality based on maximum permissible limits in surface waters.

3.2. Cluster analysis for monitoring network optimization

Descriptive statistics were performed on the parameters selected by the PCA as the most representative of the water quality in the monitoring network, and these are shown in Figures 5 and 6 for the dry and rainy seasons, respectively, and in Supplementary Material 1.

Figure 5 shows the distribution of pollutants in the Culiacan River basin, highlighting that mean concentrations of nutrients and pollutants are higher in sites 1, 2, 3, 4, 5, 6, and 7, reaching up to 37 mg/L COT; 897,320 MPN FC; 38.89 mg/L TN; 31.07 mg/L NH_3; and 6 mg/L TP. These sites are present in a drain where wastewater with and without treatment is

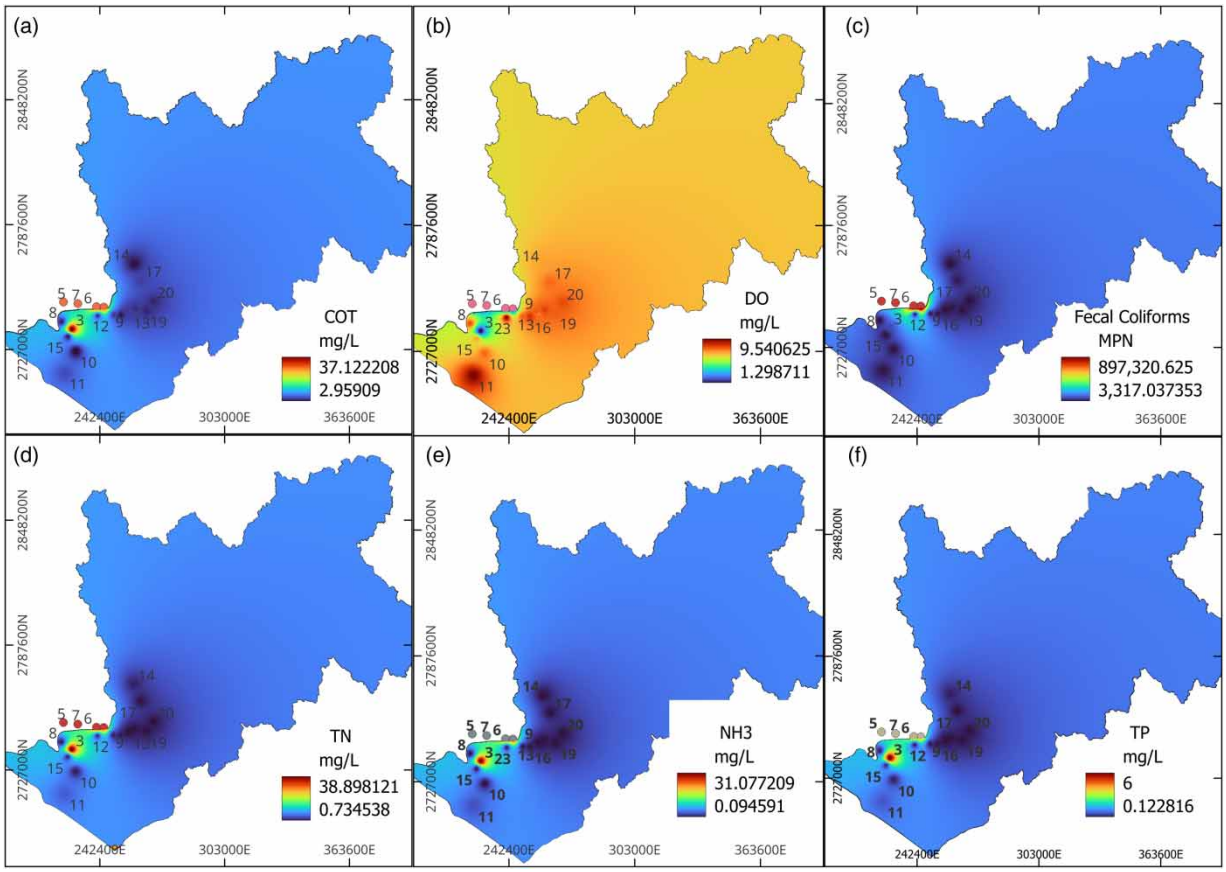


Figure 5 | Descriptive water quality parameters on the dry season in the Culiacan River basin.

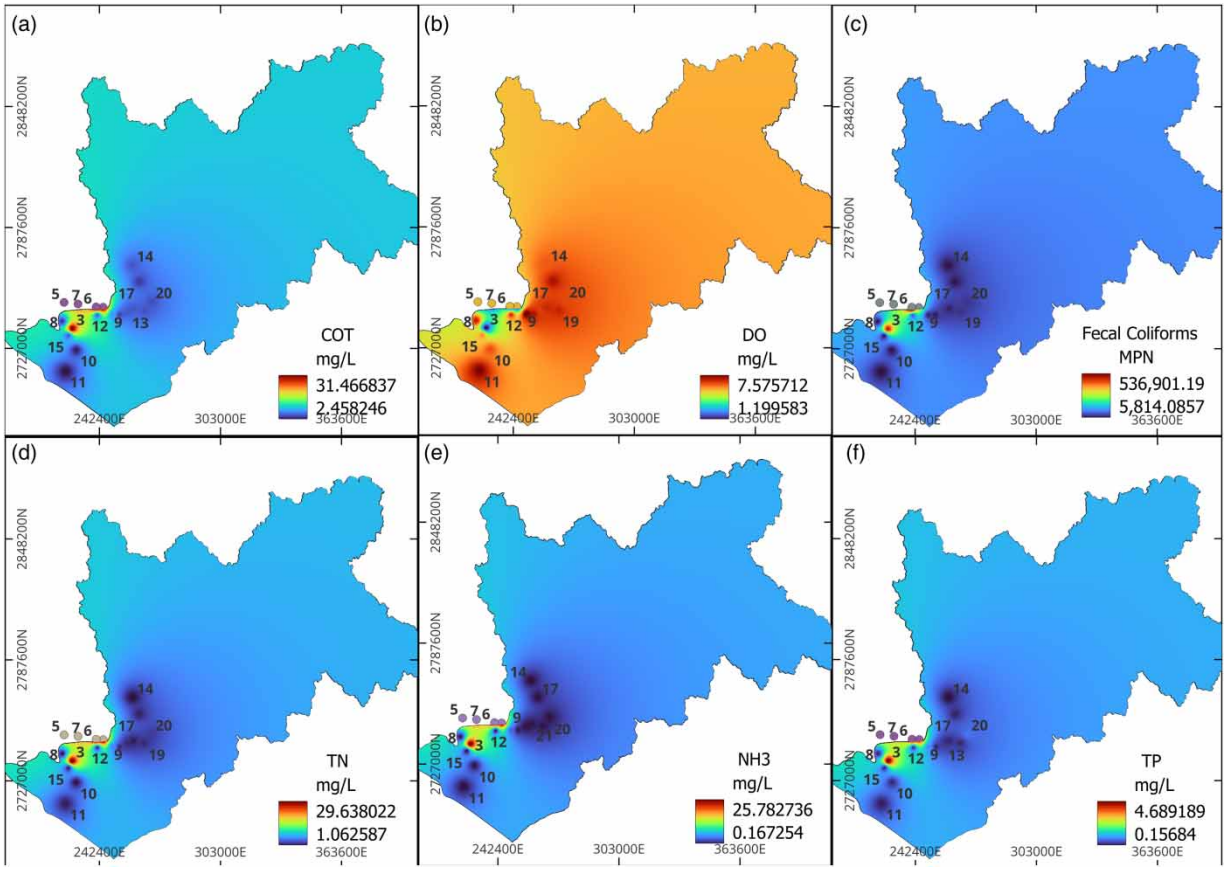


Figure 6 | Descriptive water quality parameters on the rainy season in the Culiacan River basin.

discharged, as evidenced by the high concentrations. On the other hand, among the sites on the mainstream, site 11 is the second with the highest concentrations in all the parameters, and it is located in the river mouth. There are several point sources in this place, such as discharges from fishing cooperatives, aquaculture farms, and domestic wastewater discharges, so it is easy to differentiate the water quality for the rest of the basin. For the rest of sites, there are high concentrations of TN and TP in sites where agricultural land predominates, making evident the agrochemical runoff (0.73–38.89 mg/L TN and 0.12–6 mg/L TP). This result agrees with Pei *et al.* (2022), since they relate the high concentrations of TN and TP with the runoff generated by the croplands, showing a similar nutrient concentration range (1.74–21 mg/L TN and 0.04–1.09 mg/L TP). In the case of urbanization land, fecal coliforms is the significant parameter. The monitoring network reported concentrations ranging from 3,317 MPN to 897,320 MPN. This result agrees with Yuan *et al.* (2019), where fecal coliforms associated with urbanization land ranged from 90,000 to 480,000 MPN.

In Figure 6, the spatial distribution of the concentration of the parameters in the sampling sites does not change. The highest concentrations continue occurring in the sites on the drainage drain, reaching up to 31.46 mg/L COT; 536,901 MPN FC; 29.63 mg/L TN; 25.78 mg/L NH_3; and 4.68 mg/L TP. TN and TP predominate in agricultural land. In addition, as can be seen in both figures, the DO behavior in both climatic seasons is inversely proportional to the organic matter content in the sample sites, which is why sites such as the drains present the worst oxygenation conditions, presenting higher concentrations of the other parameters.

The elbow method offers 29 possible groupings with the data assigned to the hierarchical cluster. Then, the selection of nine clusters for the rainy season and nine for the dry season coincides with the asymptotic value to determine the optimal number of clusters (Supplementary Material 2). Although the results in rainy and dry seasons are very similar, in the rainy season, there is a more significant difference in distortion score among the optimal number of clusters, and this may be since in the rainy season, precipitation and runoff cause homogenization between sampling sites due to the combination of diffuse and point sources of contamination, making the distortion score smaller between clusters. On the other hand, in

the dry season, there is more significant differentiation between concentrations of nutrients and pollutants as there are no contributions from diffuse sources that alter the characteristics of the area where they are found, making the distortion score higher among clusters. This phenomenon agrees with Molekoa *et al.* (2021), which report higher concentrations of pollutants in the dry season and inferred that rainfall significantly impacts water quality parameters by dilution and attenuation during rainy and dry seasons, respectively.

Figure 7 shows the hierarchical cluster for the rainy season. One of the groups is constituted by sites 12, 13, 20, 18, 22, 21, and 9. All these sites are in the middle basin, where the predominant land use is urbanization, ranging from 33 to 99% in an area of influence of 10 km^2. Therefore, the pollution sources observed in this area refer to point sources. According to water quality parameters, these sites have a higher presence of fecal coliforms, especially during the rainy season, registering up to 70,000 MPN/100 mL, since the rainwater system in some sections is irregularly connected to the drainage system, which, being obsolete, causes sewage overflows and runoffs into the river (Freeman *et al.* 2019; Chen *et al.* 2020), in addition to effluents from recreational areas.

These concentrations have the same order of magnitude as those reported by Zhang *et al.* (2021), where fecal coliform contamination in a watershed with agricultural activity averaged 143,127 MPN/100 mL. Although, in that study, the watershed also has contamination due to urbanization. Only 31% of its annual fecal coliform records are attributed to this land use, while more than 50% refer to contamination from agriculture. This similarity may refer to the lax regulations that exist in the region and for which the environmental standards for national waters (NOM-001-SEMARNAT-2021) have been updated to ensure non-significant environmental impact.

Another group is formed by sites 14, 17, 10, 15, 23, 16, and 8. These sites are in agricultural zones, with ratios between 40 and 80% observed in an area of 10 km^2. It is observed that some sites in this group are in a higher zone or do not share the

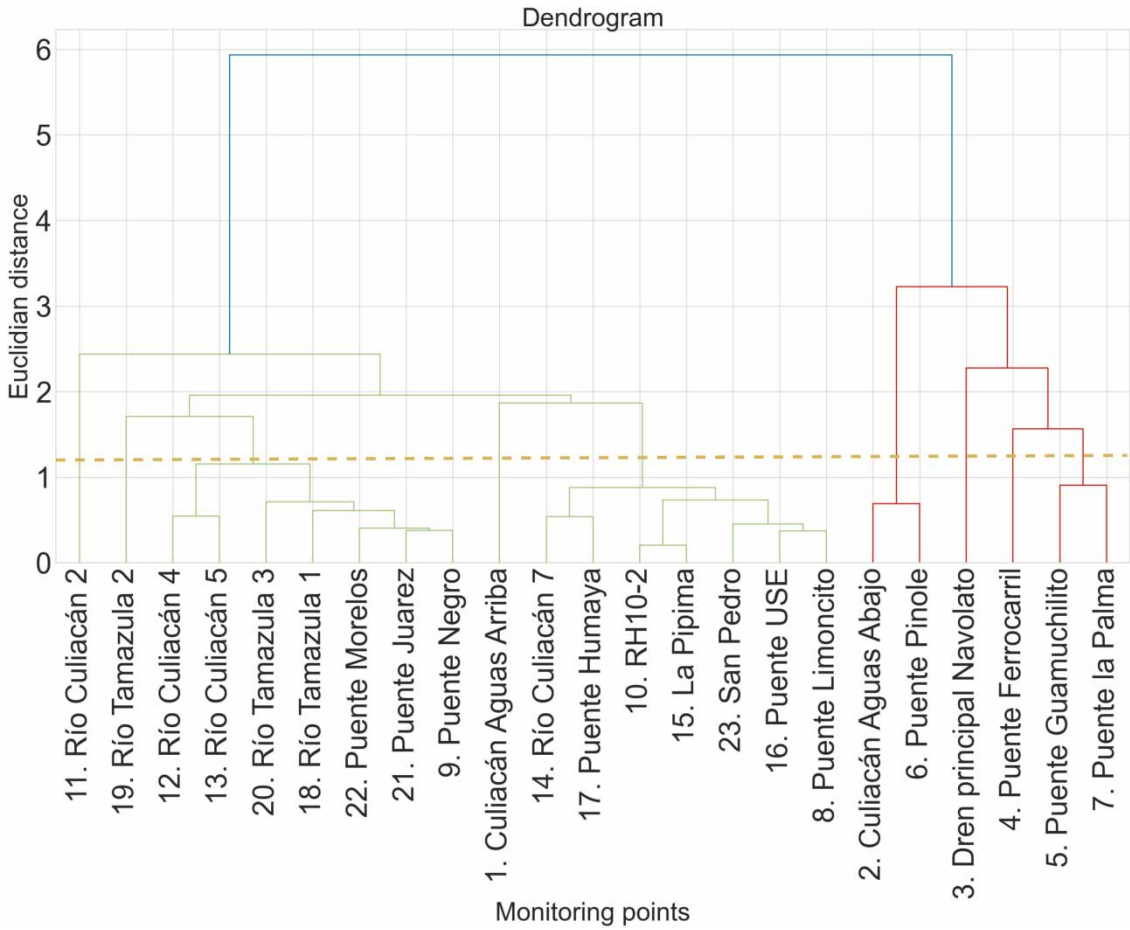


Figure 7 | Hierarchical cluster for the rainy season.

exact geographic location. However, the effect of agricultural land use on these sites can be seen in their water quality. The concentrations of TP and TN are similar in each of the sites, reaching records of 0.29 and 1.96 mg/L, respectively, with these concentrations being much higher than what was reported by Wu & Lu (2019) in a watershed with agricultural activity (TP of 0.05 mg/L and TN of 2.0 mg/L), while in a particular case of TN, the concentrations are shown to be very close to the limits established by the European Union directives, which establishes that the nitrogen standard is 2.5 mg/L to prevent eutrophication of water bodies (European Parliament 2022). However, they exceed what is established by the Environmental Protection Agency (EPA) and SEMARNAT, which suggest that concentrations above 0.04 mg/L led to eutrophication in water bodies and the proposed maximum limit of 0.1 mg/L of this nutrient in fresh waters, respectively. Therefore, concentrations like those reported in this study are attributed to anthropogenic activities such as agriculture due to the use of organophosphorus and nitrogen fertilizers that infiltrate the river (Stackpoole *et al.* 2019; Zhu *et al.* 2019).

On the other hand, sites 2 and 6 form another cluster with a distance between them as 0.8. In contrast, sites 5 and 7 form a new cluster with a 0.9 distance between them. Both clusters are characterized by being located on a main drain where effluents from the livestock and swine industry are received, in addition to contamination from diffuse sources from agricultural areas adjacent to the receiving body. Although the two clusters are on the same receiving body, the one formed by sites 6 and 2 has the point discharge of livestock activity immediately, while the cluster formed by sites 5 and 7 is 20 km away from each other, which causes degradation of pollutants on their way downstream.

As a result of these discharges, at sites 2, 5, 6, and 7, the TN concentrations range between 22 and 38 mg/L. Specifically, sites 2 and 6 exceed the maximum permissible limits established by NOM-001-SEMARNAT-2021 (DOF 2022), presenting average concentrations of 32 and 38 mg/L TN, respectively. Although the concentrations reported at sites 5 and 7 are within the maximum permissible limits of the Mexican standard, the concentrations registered are much higher than the rest of the sites in the basin (between 3.2 and 5.2 mg/L), even exceeding that are reported by Hirt *et al*. (2018), where TN concentrations range between 4 and 5 mg/L in an agricultural watershed. This may be associated with the fact that adjacent to the agricultural drain of the present study, a meat product processing plant is installed, so the number of surfactant substances used and animal fat in the effluent generates a high concentration of this nutrient (Harrison *et al.* 2019; Wang *et al.* 2020). These excessive nutrient levels generate hypoxia in the receiving water body, experiencing DO concentrations of up to 1 mg/L (Siriwardana *et al.* 2019; Testa *et al.* 2021).

Sites 11, 19, 1, 3, and 4 did not show statistical similarity with any other site of the monitoring network. In particular, the site 11 had the most significant statistical difference with the rest of the clusters in the dendrogram, which may be related to its location. Site 11 is only 3 km from the mouth of the river, where there are additional agricultural influences, and there are also communities where 70% of the population is engaged in fishing and aquaculture. This characteristic causes the parameters to be affected, so the concentrations are much higher than at other sites due to the composition of the discharges. To support this statement, Mendivil-Garcia *et al.* (2022) identified nine-point sources of pollution in the last 6 km of the Culiacan River that causes oxygen deficits after every shrimp farm wastewater discharge. Although there is currently good oxygenation in this area, the deficit generated after each discharge may increase considering the gradual intensification in temperature due to climate change.

Figure 8 shows the hierarchical cluster in the dry season. At a 1.4 Euclidean distance, the nine clusters, suggested by the elbow method, were generated. At 0.5, a first group is formed, with sites 5 and 7 being the most negligible statistical difference in the dendrogram. However, this cluster originates from the dry and the rainy seasons, with a more significant statistical difference in the second case. The repetition of this cluster in both climatic seasons may be because they are on a receiving body surrounded by crops planted in spring–summer and autumn–winter and irrigated by furrow. Even though the parameters analyzed are the same, the concentrations are notably different between the two seasons of the year, in addition to receiving discharges from the pork and livestock industry throughout the year.

Sites 4 and 6 are other clusters that originated during the dry season, with a 1.4 distance between them. Like sites 5 and 7, they are located on a wastewater receiving body adjacent to an agricultural zone, so the similarity of pollutants generates the clustering. Also, this cluster is approximately 10 km upstream of the sites 5 and 7, so despite being on the same water body, the hydrological and physical-chemical processes generate changes in the composition of the water, showing degradation/lower concentration downstream, and arranging them in two different clusters.

A third group is formed at 1.4 Euclidean distance with sites 13, 21, 19, 9, 16, and 20. This group is characterized by sampling sites located in the city's urban area, and its composition is similar to that presented by the dendrogram in the rainy season except for three sampling sites, 22, 14, and 17 generating a new cluster with a difference between them being

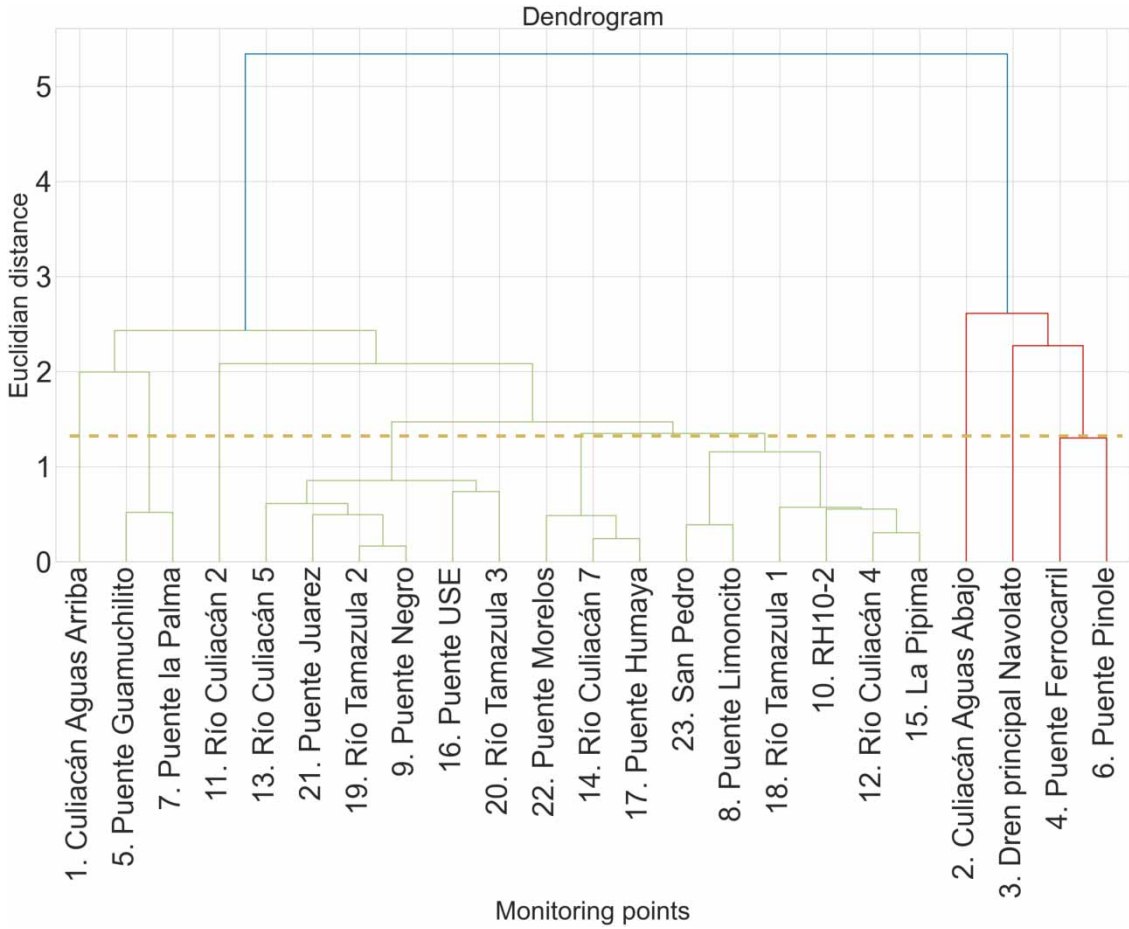


Figure 8 | Hierarchical cluster for the dry season.

0.5. This division of the cluster may be since, during the dry season, no runoff could diffusely contribute to pollutants, accentuating only the point sources in each sampling site, evidencing the homogenization that is generated between sampling sites in the rainy season (Lee *et al.* 2019; Darbandsari & Coulibaly 2020). The grouping of these sites for being within an urban area agrees with Dabgerwal & Tripathi (2016), where the optimization of the monitoring network with similarities in parameters related to domestic discharges were grouped in the same cluster.

A fourth cluster is integrated by sites 23, 8, 18, 10, 12, and 15. These sites are characterized by being in the lower basin, where the agricultural zone represents 15% of the total basin and approximately 80% of the basin section is in the state of Sinaloa, so the type of pollution in this area is mainly diffuse, especially during the dry season because irrigation of crops is done through hydraulic infrastructure such as irrigation canals, so that nutrients and pollutants reach the river by infiltration (Sorando *et al.* 2019; Zhu *et al.* 2019; Hoffmann *et al.* 2020). This accentuates the similarity between the sites due to the type of contamination they receive.

Sites 1, 11, 2, and 3 showed no statistical similarity with other sites in the monitoring network. Site 2 shows the most significant dissimilarity, with 2.7 over the other clusters.

Among these sampling sites, site 3, dren principal Navolato, stands out, which registered significantly higher concentrations ($P < 0.05$) than the other sites in the network, reaching average values of 45 mg/L of TOC, 29 mg/L of NH_3-N, 42 mg/L of TN, 5.6 mg/L of TP and 530,000 MPN/100 mL of fecal coliforms. No buffer was created to analyze the land use of that sampling site as it was less than 5 km away between sites 8 and 15, which would cause an overlap between zones of influence. However, referring to the predominant soil of these two sites, it can be inferred that the proportion of agricultural land in an area of influence of 10 km^2 is between 60 and 80%. There are no records of nearby point sources. However, the site is in an area where the population has typical practices such as having home pens, washing clothes over the

river, and lacking drainage and sewerage, among others. These are without environmental regulation due to the region's conflicts of uses and customs. This grouping agrees with Shrestha & Kazama (2007) where the sites with agricultural influence and those that showed urban influence were grouped, respectively, by the type of contamination.

Once the sampling sites grouped by similarity of water quality had been identified in every season, a descriptive analysis was carried out, calculating each parameter's global means and standard deviations. Table 3 shows the global means and standard deviation for every parameter in the rainy season. Supplementary Material 1 shows the mean values and standard deviation for every sampling sites by season. Under spatial criteria, redundant monitoring sites were discarded to establish an optimized monitoring network. The monitoring sites close to spatially important sites were eliminated, due to being at the confluence of mainstreams or close-to-point source of pollution. Discarded sites for the rainy season were 6, 7, 9, 10, 13, 15, 16, 17, 21, and 18/22. On the other hand, discarded monitoring sites on the dry season were 6, 7, 8, 9, 10, 12, 15, 16, 21, and 18/22.

The statistical analysis generates an optimized network for each climatic season (dry and rainy season). This result is obtained by selecting sites within the nine clusters formed in each climatic season. Although there are nine clusters, the sites of the new monitoring networks are 11 and 12 sites for the rainy and dry seasons, respectively. However, for this type of environmental study, it is essential to consider the hydrodynamics and the order of the water bodies, including all those tributaries of the main streams (Madsen & Wersal 2017). That is why the differences regarding the Euclidean distance within each cluster were considered to differentiate subgroups that would allow selecting sampling sites on the tributary rivers.

In addition, it is essential to consider that these sites are located on streams very close to the population. This suggests that they could be used for some service with human contacts, such as domestic plots, so they must continue to be monitored. In addition, this is of greater interest because each tributary has dams receiving streams from upstream mining areas.

Therefore, the suggested optimized network for the rainy season includes sites 1, 2, 3, 5, 8, 11, 12, 14, 19, 20, and 23. On the other hand, for the dry season, the sites considered are 1, 2, 3, 4, 5, 11, 13, 14, 19, 20, and 23 (Figure 9). To verify that the water quality from the optimized network represents the original network, the mean and standard deviation of each parameter in the optimized network were calculated. Table 3 shows the statistics for rainy and dry seasons, respectively. After this, an analysis of means, medians, and standard deviations of both networks was carried out, validating that the new network has no significant difference ($P < 0.05$) according to the original network (Supplementary Material 3).

Table 3 | Global average and standard deviation for the original and optimized water quality monitoring network

	Rainy season		Dry season	
Parameter	**AVG**	**SD**	**AVG**	**SD**
Original monitoring network				
TOC (mg/L)	11.023	6.876	9.667	6.1715
TN (mg/L)	8.701	3.192	8.1865	3.1975
NH_3 (mg/L)	6.594	2.399	5.9425	2.7585
TP (mg/L)	1.313	0.5325	1.1685	0.4485
DO (mg/L)	5.42	1.427	6.4135	1.448
Fecal coliform (MPN)	110994	194128	152400	383240
Optimized monitoring network				
TOC (mg/L)	10.561	5.72	10.132	6.895
TN (mg/L)	8.57333333	3.1675	8.7733	3.4633
NH_3 (mg/L)	6.66916667	2.5875	6.4892	3.1883
TP (mg/L)	1.33083333	0.5508	1.2892	0.495
DO (mg/L)	5.48166667	1.55	6.4275	1.6608
Fecal coliform (MPN)	83840	161881	128487	351370

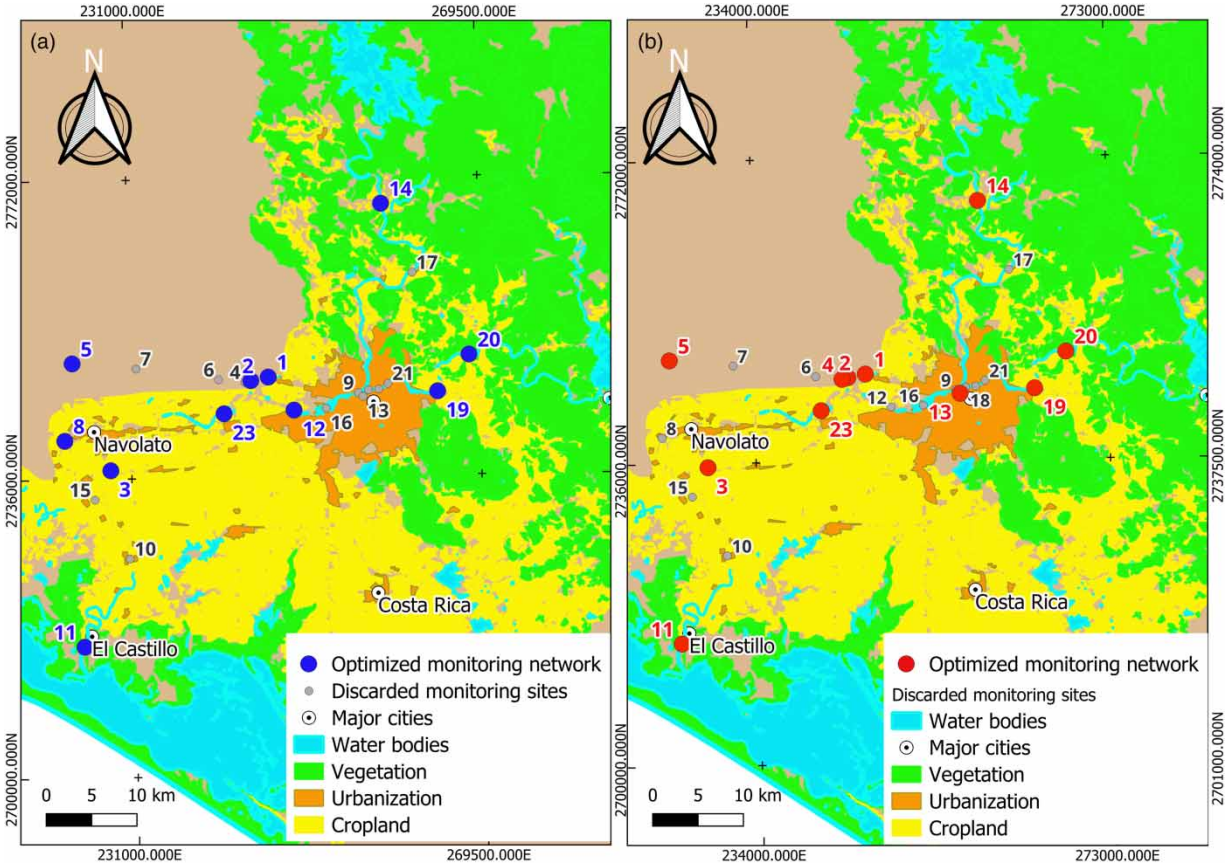


Figure 9 | Optimized monitoring network in the Culiacan River basin: (a) rainy season and (b) dry season.

4. CONCLUSIONS

Based on the results, multiparametric statistics and artificial intelligence provided the grouping of sampling sites that would reduce the current monitoring network by approximately 50%, guaranteeing the spatial distribution of water quality. The relationship between water chemistry and land use in each suggested cluster supports this reduction. Although the dendrograms of dry and rainy seasons offer nine clusters, each based on the intercluster difference, the selection of sampling sites also considers the intracluster differentiation, so based on the existence of subgroups within each cluster, the number of monitoring sites that would fulfill the purpose of monitoring the water quality of the Culiacan river was selected, going from a network of 23 sites to one with 11 in its optimized form.

On the other hand, according to the variables obtained from PCA, it was possible to select the primary water quality variables in the Culiacan River through eigenvalues, in which TN and TP nutrients are associated with the agricultural activity, that is representative of the region, showing higher concentrations in the rainy season due to the runoff caused by precipitation bringing with it pollutants from various sources. In contrast, in the dry season, the contribution of runoff caused by irrigation in the furrow of the crops is evident. In contrast, the values of NH_3 and TOC remained at high concentrations in sites with discharges from point sources. Therefore, as organic matter parameters are associated with residual water, it is inferred that these bodies of water receive discharges without prior treatment or with ineffective treatment because the registered concentrations exceed the maximum permissible limits established by Mexican regulations (TOC $<$ 38 mg/L). Furthermore, in these same sites, DO concentrations reach hypoxia. This phenomenon is related to the fact that this parameter is inversely proportional to the concentration of dissolved organic matter.

Thus, this methodology has the advantage of providing confidence to the results of multiparametric statistics, such as the elbow method that helped define the optimal number of clusters to select in the dendrograms of each season. Meanwhile, the dendrogram grouped sites in areas that receive point source pollution, areas where cropland predominates, and urban areas.

It is possible to understand data from the monitoring network and know the origin of the pollution in the basin. This allowed to discard redundant sites in the monitoring network and generate an alternative that, in addition to providing technical knowledge of the region, reduces the costs and operating time of the managing entities of the basin. Nevertheless, it is convenient to incorporate more grouping methods, besides the hierarchical cluster, for the analysis of a larger data set that gives equally reliable results and reinforces this study's results.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information. All the data required in this study is available in: https://docs.google.com/spreadsheets/d/1rlGN26LseBtsElCRPkewqmzJNTeFk6sP/edit#gid=1425574258.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Ahmed, M. A. R. 2019 Application of surface water quality classification models using principal components analysis and cluster analysis. Available at SSRN: http://dx.doi.org/10.2139/ssrn.3364401.

APHA 2005 *Standard Methods for the Examination of Water and Wastewater*. 21st edn. American Public Health Association. Washington, DC, USA.

Chen, Q., Chen, H., Zhang, J., Hou, Y., Shen, M., Chen, J. & Xu, C. 2020 Impacts of climate change and LULC change on runoff in the Jinsha River Basin. *J. Geogr. Sci.* **30**, 85–102.

Chowdury, M. S. U., Emran, T. B., Ghosh, S., Pathak, A., Alam, M. M., Absar, N., Andersson, K. & Hossain, M. S. 2019 IoT based real-time river water quality monitoring system. *Procedia Comput. Sci.* **155**, 161–168.

CONAGUA 2020 Actualización de la disponibilidad media anual de agua en el acuífero río Culiacan (2504), estado de Sinaloa. Available from: https://sigagis.conagua.gob.mx/gas1/sections/Edos/sinaloa/sinaloa.html (accessed 30 may 2023).

Dabgerwal, D. K. & Tripathi, S. K. 2016 Assessment of surface water quality using hierarchical cluster analysis. *Int. J. Environ.* **5** (1), 32–44.

Darbandsari, P. & Coulibaly, P. 2020 Inter-comparison of lumped hydrological models in data-scarce watersheds using different precipitation forcing data sets: Case study of Northern Ontario, Canada. *J. Hydrol. Reg. Stud.* **31**, 100730.

DOF 2022 *Diario Oficial de la Federación: Official Mexican STANDARD NOM-001-SEMARNAT-2021, Which Establishes the Permissible Limits of Pollutants in Wastewater Discharges in Receiving Bodies Owned by the Nation*. Available from: https://dof.gob.mx/nota_detalle.php?codigo=5645374&fecha=11/03/2022#gsc.tab=0 (accessed: 4 january 2023).

Dudley, B. D., Burge, O. R., Plew, D. & Zeldis, J. 2020 Effects of agricultural and urban land cover on New Zealand's estuarine water quality. *N. Z. J. Mar. Freshwater Res.* **54** (3), 372–392.

Egbueri, J. C. 2020 Groundwater quality assessment using pollution index of groundwater (PIG), ecological risk index (ERI) and hierarchical cluster analysis (HCA): A case study. *Groundwater Sustainable Dev.* **10**, 100292.

European Parliament 2022 *Monitoring of Nitrogen in Water in the EU*, European Union, Netherlands.

Freeman, L. A., Corbett, D. R., Fitzgerald, A. M., Lemley, D. A., Quigg, A. & Steppe, C. N. 2019 Impacts of urbanization and development on estuarine ecosystems and water quality. *Estuaries Coasts* **42**, 1821–1838.

Giao, N. T., Dan, T. H., Ni, D. V., Anh, P. K. & Nhien, H. T. H. 2022 Spatiotemporal variations in physicochemical and biological properties of surface water using statistical analyses in Vinh Long Province, Vietnam. *Water* **14** (14), 2200.

Gyimah, R. A. A., Gyamfi, C., Anornu, G. K., Karikari, A. Y. & Tsyawo, F. W. 2021 Multivariate statistical analysis of water quality of the Densu River, Ghana. *Int. J. River Basin Manage.* **19** (2), 189–199.

Harrison, J. A., Beusen, A. H., Fink, G., Tang, T., Strokal, M., Bouwman, A. F., Metson, G. S. & Vilmin, L. 2019 Modeling phosphorus in rivers at the global scale: Recent successes, remaining challenges, and near-term opportunities. *COSUST* **36**, 68–77.

Hirt, U., Venohr, M., Kreins, P. & Behrendt, H. 2008 Modelling nutrient emissions and the impact of nutrient reduction measures in the Weser river basin, Germany. *Water Sci. Technol.* **58** (11), 2251–2258.

Hoffmann, C. C., Zak, D. & Kronvang, B. 2020 An overview of nutrient transport mitigation measures for improvement of water quality in Denmark. *Ecol. Eng.* **155**, 105863.

INEGI 2018 *INEGI: Land use and Vegetation*. Available from: https://www.inegi.org.mx/temas/usosuelo/#Descargas (accessed 4 january 2023).

INEGI 2019 *Informe técnico de la cuenca hidrologica del rio Culiacán. Humedales*. Instituto Nacional de Estadística y Geografía, Aguascalientes.

INEGI 2021 *Información de México para niños*. Available from: https://www.cuentame.inegi.org.mx/monografias/informacion/sin/territorio/clima.aspx?tema=meINEGI (accesed 4 january 2023).

Jan, F., Min-Allah, N. & Düştegör, D. 2021 IoT based smart water quality monitoring: Recent techniques, trends and challenges for domestic applications. *Water* **13** (13), 1729.

Krishnaraj, A. & Deka, P. C. 2020 Spatial and temporal variations in river water quality of the Middle Ganga Basin using unsupervised machine learning techniques. *Environ. Monit. Assess.* **192** (12), 744.

Kumar, V., Sharma, A., Kumar, R., Bhardwaj, R., Kumar Thukral, A. & Rodrigo-Comino, J. 2020 Assessment of heavy-metal pollution in three different Indian water bodies by combination of multivariate analysis and water pollution indices. *HERA* **26** (1), 1–16.

Lee, M.-H., Im, E.-S. & Bae, D.-H. 2019 Impact of the spatial variability of daily precipitation on hydrological projections: A comparison of GCM- and RCM-driven cases in the Han River basin, Korea. *Hydrol. Processes* **33** (16), 2240–2257.

Li, D., Sun, Y., Sun, J., Wang, X. & Zhang, X. 2022 An advanced approach for the precise prediction of water quality using a discrete hidden Markov model. *J. Hydrol.* **609**, 127659.

Madsen, J. D. & Wersal, R. M. 2017 A review of aquatic plant monitoring and assessment methods. *J. Aquat. Plant Manage.* **55**, 1–12.

Makuwa, S., Tlou, M., Fosso-Kankeu, E. & Green, E. 2020 Evaluation of fecal coliform prevalence and physicochemical indicators in the effluent from a wastewater treatment plant in the North-West Province, South Africa. *IJERPH* **17** (17), 6381.

Mamun, M., Kim, J. Y. & An, K.-G. 2021 Multivariate statistical analysis of water quality and trophic state in an artificial dam reservoir. *Water* **13** (2), 186.

Mechelli, A. & Viera, S. 2019 *Machine Learning: Methods and Applications to Brain Disorders*. Academic Press, London, UK.

Mendivil-García, K., Amabilis-Sosa, L. E., Salinas-Juárez, M. G., Pat-Espadas, A., Rodríguez-Mata, A. E., Figueroa-Pérez, M. G. & Roé-Sosa, A. 2022 Climate change impact assessment on a tropical river resilience using the Streeter-Phelps dissolved oxygen model. *Front. Environ. Sci.* **10**, 1–15.

Molekoa, M. D., Avtar, R., Kumar, P., Thu Minh, H. V., Dasgupta, R., Johnson, B. A. & Yunus, A. P. 2021 Spatio-temporal analysis of surface water quality in Mokopane area, Limpopo, South Africa. *Water* **13** (2), 220.

Pei, L., Wang, C., Zuo, Y., Liu, X. & Chi, Y. 2022 Impacts of land Use on surface water quality using self-organizing map in middle region of the Yellow River Basin, China. *IJERPH* **19** (17), 10946.

SEMARNAT 2021 *SEMARNAT: Historical Average Precipitation by State*. Available from: http://dgeiawf.semarnat.gob.mx:8080/ibi_apps/WFServlet?IBIF_ex=D3_AGUA01_01&IBIC_user=dgeia_mce&IBIC_pass=dgeia_mce&NOMBREENTIDAD=*&NOMBREANIO=* (accessed 4 January 2023).

Shrestha, S. & Kazama, F. 2007 Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environ. Modell. Software* **22** (4), 464–475.

SIAP 2020 *Agrifood and Fisheries Information Service: Agricultural Production Statistics*. Available from: http://infosiap.siap.gob.mx/gobmx/datosAbiertos.php (accessed: 4 january 2023).

Siriwardana, C., Cooray, A. T., Liyanage, S. S. & Koliyabandara, S. M. P. A. 2019 Seasonal and spatial variation of dissolved oxygen and nutrients in Padaviya Reservoir, Sri Lanka. *J. Chem.* **2019**, 1–14.

Sorando, R., Comín, F. A., Jiménez, J. J., Sánchez-Pérez, J. M. & Sauvage, S. 2019 Water resources and nitrate discharges in relation to agricultural land uses in an intensively irrigated watershed. *Sci. Total Environ.* **659**, 1293–1306.

Stackpoole, S. M., Stets, E. G. & Sprague, L. A. 2019 Variable impacts of contemporary versus legacy agricultural phosphorus on US river water quality. *Biol. Sci.* **116** (41), 20562–20567.

Testa, J. M., Basenback, N., Shen, C., Cole, K., Moore, A., Hodgkins, C. & Brady, D. C. 2021 Modeling impacts of nutrient loading, warming, and boundary exchanges on hypoxia and metabolism in a shallow estuarine ecosystem. *JAWRA* **58** (6), 876–897.

Tung, T. M. & Yaseen, Z. M. 2020 A survey on river water quality modelling using artificial intelligence models: 2000–2020. *J. Hydrol.* **585**, 124670.

Ustaoğlu, F., Tepe, Y. & Taş, B. 2019 Assessment of stream quality and health risk in a subtropical Turkey river system: A combined approach using statistical analysis and water quality index. *Ecol. Indic.* **113**, 105815.

Wall, M. E., Rechtsteiner, A., Rocha, L. M., 2003 Singular value decomposition and principal component analysis. In: *A Practical Approach to Microarray Data Analysis* (Berrar, D. P., Dubitzky, W. & Granzow, M., eds). Springer, Boston, MA.

Wang, Y., Xie, X., Liu, C., Wang, Y. & Li, M. 2020 Variation of net anthropogenic phosphorus inputs (NAPI) and riverine phosphorus fluxes in seven major river basins in China. *Sci. Total Environ.* **742**, 140514.

Wang, R., Kim, J. H. & Li, M. H. 2021 Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach. *Sci. Total Environ.* **761**, 144057.

Wu, J. & Lu, J. 2019 Landscape patterns regulate non-point source nutrient pollution in an agricultural watershed. *Sci. Total Environ.* **669**, 377–388.

Yadav, S., Babel, M., Shresta, S. & Deb, P. 2019 Land use impact on the water quality of large tropical river: Mun River Basin, Thailand. *Environ. Monit.* **191** (614), 1–22.

Yang, W., Zhao, Y., Wang, D., Wu, H., Lin, A. & He, L. 2020 Using principal components analysis and IDW interpolation to determine spatial and temporal changes of surface water quality of Xin'anjiang river in Huangshan, China. *IJERPH* **17** (8), 2942.

Yuan, T., Vadde, K. K., Tonkin, J. D., Wang, J., Lu, J., Zhang, Z. & Sekar, R. 2019 Urbanization impacts the physicochemical characteristics and abundance of fecal markers and bacterial pathogens in surface water. *IJERPH* **16** (10), 1739.

Zhang, X., Chen, L. & Shen, Z. 2021 Impacts of rapid urbanization on characteristics, sources and variation of fecal coliform at watershed scale. *J. Environ. Manage.* **286**, 112195.

Zhu, X., Liu, W., Chen, J., Bruijnzeel, L. A., Mao, Z., Yang, X. & Jiang, X. J. 2019 Reductions in water, soil and nutrient losses and pesticide pollution in agroforestry practices: A review of evidence and processes. *Plant Soil* **453**, 45–86.

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B. & Ye, L. 2022 A review of the application of machine learning in water quality evaluation. *EEH* **1** (2), 107–116.

First received 14 August 2023; accepted in revised form 4 December 2023. Available online 22 December 2023