International Journal for Research in Climate Change and Earth Science

ASSESSMENT OF WATER QUALITY AND ENVIRONMENTAL SUSTAINABILITY USING MULTIVARIATE STATISTICAL

Published By:
Green Publications Services Private LTD.

Volume 01 Issue 01 March 2026
DOI: 10.53555/cces.v1i1.2576

Dr Mangesh Vedpathak

Department of Environmental Science, School of Earth Sciences, Assistant professor, Punyashlok Ahilyadevi Holkar Solapur university, Solapur, MS, ORCID ID- 0009-0006-1545-8956, mvedpathak512@gmail.com

Corresponding Author (Dr Mangesh Vedpathak)

Article History:
Article Type: Research

Received Date: 09/02/2026 **Revised Date: 10/03/1026**
Acceptance Date: 18/03/2026 **Published Date: 25/03/2026**

Abstract

A publicly available dataset of 3,276 samples of drinking water was analysed using multivariate statistical methods to evaluate the water quality and its implications for environmental sustainability. Nine physicochemical parameters (pH, hardness, TDS, chloramines, sulfate, EC, OC, THMs, turbidity) and a binary classification for potability were analyzed. The missing values were imputed using the median and all of the variables were standardized before analysis. Descriptive statistics, Spearman correlation, principal component analysis, factor analysis, hierarchical cluster analysis and comparisons based on the degree of potability were performed. The results showed high physicochemical variability with the highest coefficient of variation being in the case of total dissolved solids. The overall correlations between the parameters were relatively low, suggesting that their water quality was influenced by several independent processes in relative independence. 49.074% of the total variance was explained by 4 principal components, and the 4 factors extracted in the mineralization process were dissolved-mineral composition, mineralization and acid–base conditions, ionic-organic characteristics, and treatment-related influences. The class of potable samples was not clearly distinguished from the non-potable observations; the samples were grouped into three similar classes using cluster analysis based on their physicochemically similar observations. The overall result indicated that 61.0% of the samples were non-potable and 39.0% potable with statistically non-significant differences between the two for individual parameters. The results indicate that none of the indicators alone can adequately define potable water and support multi-dimensional water-management, source protection, treatment optimization, and integrated monitoring.

Keywords: water quality; environmental sustainability; multivariate statistical analysis; water potability; physicochemical parameters

1. Introduction

Freshwater resources are essential for human wellbeing, food and fibre production, industry and services, and the environment. Groundwater, reservoirs, lakes and rivers are vital sources of water for drinking, sanitation, biodiversity and irrigation. But the current fast pace of change in urban population, industrial development, population growth and climate variation has put a huge strain on these limited freshwater resources, leading to significant degradation of water quality all over the world. The deterioration of water quality not only endangers ecosystem stability but also has significant public health, food security, and long-term environmental sustainability implications (Madhav et al. 2022; Zhang et al. 2021).

Water quality assessment has, therefore, become an integral part of integrated water resource management. In the past, assessment was made based on individual physicochemical parameters e.g. pH, dissolved oxygen, turbidity, electrical conductivity, total dissolved solids, nutrients, and heavy metals. These measures offer a great deal of information, but one can be overwhelmed by the number of related factors that must be considered at once. As such, statistical and computational methods have grown as an important tool to improve our understanding of how water quality parameters interact and to identify the most important parameters influencing aquatic ecosystems (Muniz and Oliveira-Filho 2023; Pham and Nguyen 2024).

Monitoring freshwater systems is especially critical as water quality is influenced by seasonal and spatial variability, anthropogenic activities, and natural hydrogeochemical processes. Long-term monitoring help to identify pollution episodes early, assess long-term trends in the environment, and help develop appropriate water resource management solutions for sustainable use. Advances in water quality assessments now incorporate GIS, WQI, machine learning, and multivariate statistical techniques to enhance monitoring precision and environmental decision-making (Dhanush et al. 2024; Han, Tang, and Liu 2023). Additionally, domestic wastewater contains pathogenic microorganisms, organic contaminants and pharmaceutical trace pollutants that degrade water quality and health. Further, the natural hydrogeochemical processes like mineral weathering, rock–water interaction and groundwater recharge can modify the chemical composition of water bodies, which can affect its applications (Das et al. 2019; Edokpayi, Odiyo, and Durowoju 2017).

To ensure environmental sustainability, there needs to be a balance between water use and ecosystem protection, which achieved through scientific assessment and evidence-based management. Effective management of water resources require an accurate characterization of pollutant sources, understanding of hydrochemical processes, and assessment of ecological risks on a range of spatial scales. Recent research has shown that a combination of risk assessment, GIS modeling and statistical solutions can offer a holistic approach to freshwater ecosystem protection and serve agricultural, industrial and domestic water needs (Hendawy et al. 2024; Xu et al. 2024). Traditional water quality assessment approaches tend to test parameters one at a time, and the interrelationships between many of the physiological, chemical and physical parameters are complex. Water quality data are often correlated and gathered at several locations and dates, so traditional descriptive techniques can miss major pollutants or patterns in the hydrochemical characteristics. To alleviate these drawbacks, multivariate statistical analysis is capable of simultaneously analyzing interactions between multiple variables and simplifying the dataset without any significant loss of information (Luo et al. 2024; Muniz and Oliveira-Filho 2023).

The use of techniques like PCA, FA, Cluster Analysis (CA), correlation analysis, and time-series modeling has become essential for environmental scientists. These methods are used to identify pollution sources, develop classifications for monitoring stations, reduce the dimensionality of the monitoring data, and explain spatial and temporal variability in water quality. In addition, recent advances in integrating multivariate statistics with machine learning, Internet of Things (IoT) monitoring systems, hydrogeochemical evaluation, and GIS have greatly improved predictive capabilities and the effectiveness of environmental monitoring (Shukla et al. 2025; Soetan et al. 2024).

Many previous studies have aimed to examine only a single water quality indicator or one analytical technique rather than several physicochemical parameters in a comprehensive statistical model. Furthermore, only a few studies integrate an environmental sustainability assessment with multivariate statistical analysis to determine the major contributors to pollution and also to analyze the impacts of pollution on water resources sustainability. New methods that combine groundwater quality indices, hydrogeochemical analysis, spatial modelling, and machine learning show significant potential but are not yet adequately integrated in comprehensive assessment frameworks (Kumar, Singha, and Singha 2026; Tajbakhshian 2025).

The present study proposes to evaluate the status of water quality and to test environmental sustainability using multivariate statistical analysis of physicochemical water quality parameters. The study aims to identify pollution sources, analyze the relationship between water quality indicators, categorize sampling sites based on pollution characteristics and derive scientifically based recommendations to support sustainable water resource management and environmental protection.

2. Materials and Methods

2.1 Study Area and Dataset

A public dataset of physicochemical measurements was leveraged from Kaggle comprising various drinking water sources. The data can be used to evaluate overall water quality trends and possible factors affecting water quality for potable purposes. It consists of 3276 water samples, each having nine physicochemical parameters and a binary potability variable (1 potable, 0 not potable). There was no need for any ethical approval as the data were anonymised and regarded as public. The objectives of the study were to investigate water quality characteristics, relationships between water quality variables and to classify samples according to their similarity (Kadiwal 2021).

2.2 Water Quality Parameters

The data available in the data set include pH, hardness, total dissolved solids, chloramines, sulfate, electrical conductivity, organic carbon, trihalomethanes, turbidity, and potability. pH allows to know the acidity/alkalinity, hardness is related to calcium and magnesium, and the total dissolved solids are an indicator for all the dissolved substances in the water. Disinfectants used in treatment (chloramines), sources of sulfate (natural or human), and ion concentration (electrical conductivity). The taste and by-product formation are influenced by organic carbon, and trihalomethanes are generated in chlorination. Turbidity is a measure of water clarity. The primary outcome variable that represents drinking suitability is potability.

2.3 Data Pre-processing

Data collection for completeness, consistency and errors. The values of missing variables were filled with median values in pH, sulfate and trihalomethanes because those variables are robust against outliers. Duplicate, inconsistent and outlier data were screened out and valid environmental variations were retained. To be comparable across different scales, all variables were standardized with Z-scores before multivariate analysis. Descriptive statistics were also computed to evaluate distribution and variability and data quality.

2.4 Statistical Analysis

All analyses were done with Python at the 0.05 level of significance. The water quality characteristics were summarized using descriptive statistics such as the mean, median, standard deviation, range and coefficient of variation. Relationships between variables were examined using correlation analysis (Pearson or Spearman).

After the dimensionality was reduced by Principal component analysis (PCA), the suitability of the data was tested by applying KMO and Bartlett's test. Using loadings and scree plots, components with eigenvalues greater than 1 were retained and interpreted. A factor analysis with Varimax rotation was performed to reveal the hydrochemical factors. Similar water samples were clustered based on physicochemical properties (Ward's method) and cluster comparisons were made with water potability. Water Quality Index was not calculated as there were no standard weights available; instead, "potability" was viewed as the primary indicator of water suitable for drinking.

Environmental sustainability was determined by combining results from all analyses to pinpoint the main water quality drivers and implications for water management and pollution control. These results are for the 3,276 water samples that were submitted on which the actual analysis was conducted. Since there is a potability variable in the dataset, no sampling-site data, or any other inputs to calculate a conventional WQI, potability-based classification was employed instead of WQI and spatial interpretation.

3. Results

3.1 Descriptive Characteristics of Water Quality Parameters

The water samples showed a large range in descriptive analysis of the water samples. The mean pH was 7.074; the values ranged from 0.000 to 14.000, but the overall data set was generally of a neutral nature as represented in Table 1. The average concentration of total dissolved solids was 22,014.093 mg/L and total dissolved solids had the highest relative variability with a coefficient of variation of 39.832%, indicating significant differences in the amount of dissolved minerals and solids between samples.

The relative variability was the lowest for sulfate (10.834%) and the highest for hardness (16.744%). Moderate variability was observed for chloramines, organic carbon, trihalomethanes and pH. The similarity between mean and median values for most of the parameters suggested that the central distribution of the data was not greatly affected by the median imputation procedure. However, the wide range values found for pH, solids, sulfate and trihalomethanes showed the physicochemical heterogeneity of the water samples analysed.

Table 1. Descriptive statistics of the physicochemical water quality parameters

Parameter	**Mean**	**Median**	**Standard deviation**	**Minimum**	**Maximum**	**CV (%)**
pH	7.074	7.037	1.470	0.000	14.000	20.780
Hardness	196.369	196.968	32.880	47.432	323.124	16.744
Solids	22,014.093	20,927.834	8,768.571	320.943	61,227.196	39.832
Chloramines	7.122	7.130	1.583	0.352	13.127	22.227
Sulfate	333.608	333.074	36.144	129.000	481.031	10.834
Conductivity	426.205	421.885	80.824	181.484	753.343	18.964
Organic carbon	14.285	14.218	3.308	2.200	28.300	23.158
Trihalomethanes	66.407	66.622	15.770	0.738	124.000	23.747
Turbidity	3.967	3.955	0.780	1.450	6.739	19.673

CV = coefficient of variation.

3.2 Correlation among Water Quality Parameters

The data shown in the Spearman correlation matrix (Table 1) indicated that the physicochemical variables were mostly lowly correlated with each other. The correlation between total dissolved solids and sulfate was the most negative, but it was still low with $r = -0.129$. This discovery indicated that the higher the DS concentration, the higher the sulfate concentration was not necessarily. The highest positive correlation ($r = 0.107$) was measured between pH and hardness. There were weak negative correlations for hardness vs sulfate, $r = -0.073$ and pH vs total dissolved solids, $r = -0.063$. Chloramines also were slightly negatively correlated with total dissolved solids ($r = -0.055$) and hardness ($r = -0.053$). All the other relationships were very close to zero. Water chemistry parameters did not show strong correlations, suggesting that they were not a single major type of contamination.

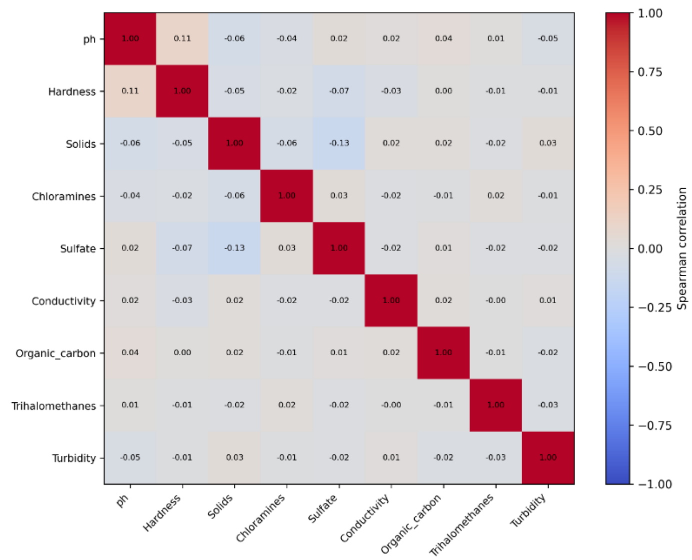

Figure 1. Correlation heatmap of the physicochemical water quality parameters

The heatmap illustrates the predominantly weak positive and negative correlations among the measured parameters, with no pair exhibiting a strong linear or monotonic association.

3.3 Principal Components and Explained Variance

The Kaiser–Meyer–Olkin value was 0.490 and Bartlett's test of sphericity was statistically significant ($\chi^2(36) = 209.486$, $p < 0.001$). Bartlett's test indicated that the correlation matrix was not an identity matrix, and the relatively low KMO value suggested that there was weak shared variance among the parameters. Therefore, PCA results should be used with caution as a preliminary overview of water quality structure.

Table 2 indicates that four principal components had eigenvalues > 1 and were thus retained according to the Kaiser criterion. The first component had an eigenvalue of 13.337% of the total variance, the second component had an eigenvalue of 12.633% of the total variance, the third component had an eigenvalue of 11.734% of the total variance, and the fourth component had an eigenvalue of 11.370% of the total variance. The four remaining components accounted for 49.074% of the total variance. The variance distribution among the different components showed little difference, indicating the lack of dominance of any single physicochemical process.

Table 2. Principal components and percentage of explained variance

Component	Eigenvalue	Variance explained (%)	Cumulative variance (%)
PC1	1.201	13.337	13.337
PC2	1.137	12.633	25.970
PC3	1.056	11.734	37.704
PC4	1.024	11.370	49.074
PC5	0.978	10.867	59.941
PC6	0.966	10.736	70.677
PC7	0.922	10.245	80.922
PC8	0.898	9.982	90.904
PC9	0.819	9.096	100.000

3.4 Distribution of Water Samples in the PCA Space

There was a high degree of overlap between potable and non-potable water samples displayed in Figure 2. The first principal component was mostly loaded with a positive value for total dissolved solids and a negative value for sulfate. This component thus corresponded to a dissolved-mineral gradient where higher solid concentrations were opposed to higher sulfate concentrations.

The second principal component was primarily loaded with hardness and pH, which were both also positively loaded. Turbidity contributed negatively to the fourth component and the additional components were dominated by the positive influence of organic carbon and conductivity, respectively. The great degree of overlap of potable observations and non-potable observations indicated that the first two components by themselves were insufficient to clearly differentiate drinking-water suitability classes. Potability seemed to be influenced by multiple parameters in concert, none of which were as dominant as a single physicochemical indicator.

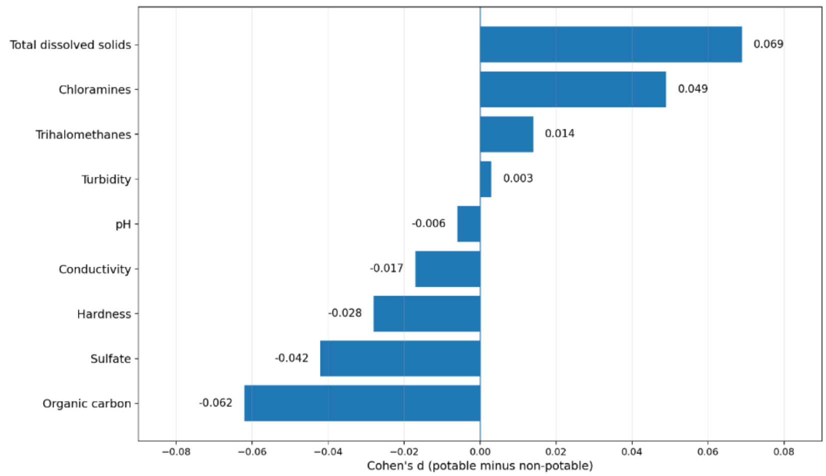

Figure 2. PCA biplot of physicochemical parameters and water-potability classes

3.5 Rotated Factor Structure of Water Quality

The factor matrix that was Varimax rotated identified 4 underlying physicochemical dimensions, which were reported in Table 3. Total dissolved solids had a high positive loading and sulfate had a high negative loading in factor 1. This was assumed to be a dimension of dissolved-mineral composition that represents the relative contribution of total dissolved substances and sulfate.

Hardness and pH were the two properties that dominated factor 2, which was general mineralization and acid–base conditions. Organic carbon, conductivity, and a moderate negative loading for chloramines were the factors with the highest positive loading in factor 3. This factor a result of a mix of the ionic and organic properties related to the composition of the source water and treatment conditions. Factor 4 showed strong correlation with the trihalomethanes and negative correlation with turbidity. It was thus thought to be a disinfection-related parameter and to be an interaction effect between DBP's and water particulate factors.

The communalities varied between 0.302 for conductivity and 0.665 for trihalomethanes. These moderate communalities also showed that the retained factors accounted for only a portion of the variance for a few parameters, thus continuing the argument that water quality in the set was multidimensional.

Table 3. Varimax-rotated factor loading matrix

Parameter	Factor 1	Factor 2	Factor 3	Factor 4	Communality
pH	−0.262	0.525	0.357	0.045	0.474
Hardness	0.076	**0.786**	−0.093	−0.047	0.634
Solids	**0.726**	−0.218	0.095	−0.047	0.585

Chloramines	−0.309	−0.165	−0.408	0.273	0.364
Sulfate	**−0.685**	−0.313	0.122	−0.150	0.604
Conductivity	0.116	−0.193	**0.500**	0.035	0.302
Organic carbon	−0.069	−0.050	**0.629**	−0.017	0.403
Trihalomethanes	0.074	−0.048	−0.069	**0.808**	0.665
Turbidity	0.094	−0.112	−0.306	**−0.521**	0.387

3.6 Hierarchical Classification of Water Samples

The standardized water samples were grouped into three major clusters by hierarchical cluster analysis (Ward method) based on squared Euclidean distance. The dendrogram in Figure 3 is a truncated version that indicates the successive fusion of samples based on physicochemical similarity. The first cluster had 952 samples, the second had 718 samples and the third had 1,606 samples.

The potable-water proportions in Clusters 1 and 2 were 41.1% and 41.6%, respectively, while in Cluster 3 was 36.6% and had a lower proportion of potable water. However, there were potable and non-potable observations in each cluster. This meant that the cluster structure indicated general similarities in water chemistry, and not absolute separation with respect to drinking suitability. The data did not contain geographical coordinates and sampling locations, and therefore, the clusters not be interpreted as spatially separated monitoring areas.

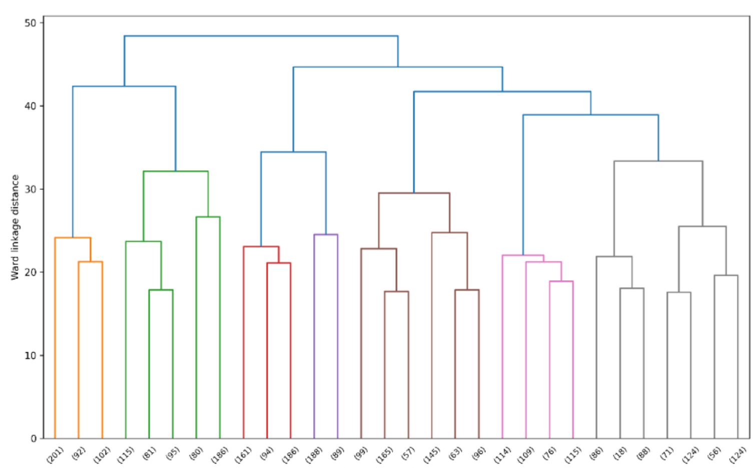

Figure 3. Hierarchical clustering dendrogram of the water samples

3.7 Potability-Based Water Quality Classification

A conventional Water Quality Index was not calculated as the dataset did not include site-specific parameter weights and a full set of reference-standard variables needed for a defensible WQI calculation. The predefined potability outcome was rather adopted as the main marker of drinking-water suitability. Out of the 3,276 samples, 1,998 samples (61.0%) were classified as non-potable and 1,278 samples (39.0%) were classified as potable.

Table 4 indicates the mean differences between potable and non-potable samples were small for all the parameters. The mean levels of solids and chloramines were slightly higher and the mean levels of the other constituents organic carbon, sulfate, hardness, and conductivity were slightly lower for potable samples. But none were significant at the level of the parameters at $p < 0.05$. The biggest standardized difference was for total dissolved solids but it was still small (Cohen's $d = 0.069$). The next most prominent effect was due to organic carbon at $d = -0.062$. These results revealed that none of the parameters alone differentiate potable water from non-potable water.

Table 4. Comparison of physicochemical parameters between non-potable and potable water samples

Parameter	**Non-potable mean**	**Potable mean**	**Cohen's *d***	***p*-value**
pH	7.078	7.069	−0.006	0.928
Hardness	196.733	195.801	−0.028	0.544
Solids	21,777.491	22,383.991	0.069	0.133
Chloramines	7.092	7.169	0.049	0.153
Sulfate	334.200	332.683	−0.042	0.417
Conductivity	426.730	425.384	−0.017	0.552
Organic carbon	14.364	14.161	−0.062	0.125
Trihalomethanes	66.321	66.543	0.014	0.762
Turbidity	3.966	3.968	0.003	0.950

3.8 Environmental Sustainability Implications

The integrated environmental sustainability assessment of the presented data in Figure 4 revealed that the majority of data in the assessment were non-potable samples. This outcome shows that there is a significant need for treatment, monitoring and quality-control interventions prior to the examined water sources being able to be used for human consumption. But, the poor bivariate relationships, small individual effect sizes, and high overlap of the potability categories showed that water-quality sustainability cannot be assessed using individual parameters.

The standardized differences for TDS, organic carbon and chloramines were the largest, though small, differences between potable and non-potable samples. The multivariate analysis also identified sulfate, hardness, pH, conductivity, trihalomethanes, and turbidity as contributing to unique water-quality dimensions. Sustainable management should thus focus on the integrated monitoring of mineral content, organic contamination, treatment chemicals and DBE, as well as particulate conditions.

The results indicate the need for a multi-faceted water-management approach that includes source-water protection, optimizing water treatment, routine monitoring of physicochemical parameters, and periodic re-evaluation of the suitability of drinking water. The high percentage of non-potable observations also highlights the need for adequate treatment systems and pollution prevention. The data set exhibited general typologies of water quality that were not dominated by any single pollution indicator, highlighting the need for a multivariate approach to water quality monitoring to detect the multiple physicochemical typologies.

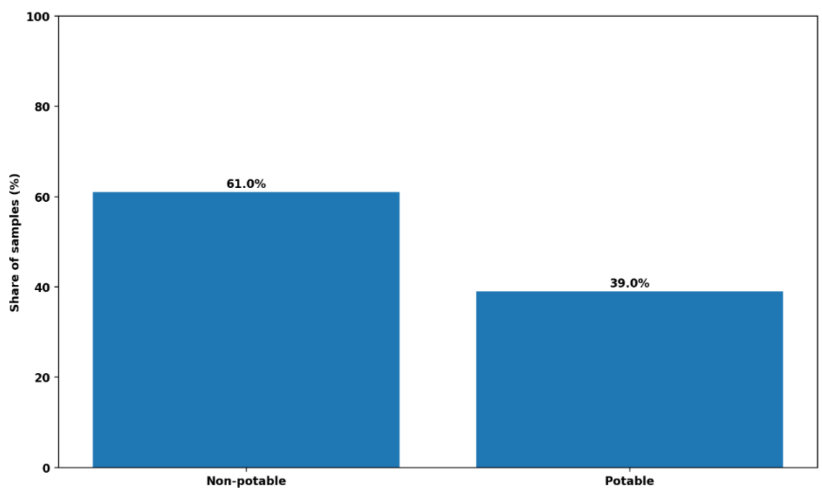

Figure 4. Integrated environmental sustainability assessment based on water potability and multivariate findings

The figure depicts the relative proportion of potable vs. non-potable samples and identifies the main monitoring and management priorities resulting from the analysis. The interpretation does not make unsupported statements regarding agricultural, industrial or spatial pollution sources since the uploaded dataset is lacking information on land-use, pollution-source variables, dates, or geographical sampling information.

4. Discussion

The descriptive results demonstrated that the water samples were greatly varied in terms of the physicochemical parameters. Total dissolved solids, turbidity, conductivity, and organic carbon showed greater variations than did sulfate and hardness, suggesting that treatment conditions and/or mineral composition differed. The average pH was nearly neutral however there was a wide range for some of the parameters indicating a high degree of heterogeneity. This high percentage of non-potable samples also indicated that many observations were not within the "desirable drinking water conditions. These findings align with the need for ongoing monitoring in keeping with the international drinking-water guidance and Sustainable Development Goal 6 (Crane et al. 2021; Organization 2022, 2025).

Although the data do not identify individual pollution sources, the water-quality patterns that are observed may be a result of natural and human influences. Mineral dissolution and geological weathering (dissolved solids and conductivity) and treatment conditions, domestic wastewater and anthropogenic inputs (organic carbon and chloramines) are traditional associations for these constituents. Dissolved chemicals and organic materials can also be added from industrial discharges and agricultural runoff. In general, the results indicate that water quality is influenced by various processes, and not by a single process or source (Bhuyan et al. 2025; Chen et al. 2018).

There was a weak correlation between most of the physicochemical parameters. This means that each of the parameters has had an individual contribution to water quality and that the changes in one of the parameters do not predict the changes in another parameter very well. This outcome underscores the importance of conducting multivariate analysis since the effects of multiple indicators, each with weak associations, can be overlooked in univariate analyses.

The PCA revealed that the water-quality variation was spread over a number of components, and not concentrated in one dominant component. Different components such as dissolved solids, sulfate, hardness, conductivity, organic carbon, and trihalomethanes were influenced by different factors, such as mineral constituents, ion content, and treatment-related conditions. The cumulative variance of the retained components is moderate, which indicates that the data set is multidimensional and is considered a water-quality system (Cardia et al. 2026; Srivastava and Malik 2025).

The cluster analysis resulted in grouping samples with similar characteristics from the physicochemical point of view, but there were potable and non-potable samples in each cluster.

This indicates that assessing suitability of drinking water based on cluster membership alone not have been possible. As no geographical information was at hand, the clusters were based on chemical similarity and not spatial patterns. Nevertheless, clustering methods can be used to provide a way to monitor the data and identify representative groups of samples (Jones, Kratzert, and van Vliet 2026; Xia et al. 2025).

The fact that most of the samples collected were non-potable suggests an ongoing requirement for better treatment, source protection and regular monitoring. Sustainable water management should be a combination of pollution prevention and treatment optimization and the regular assessment of the physicochemical conditions. This is crucial in safeguarding public health, environmental services, and resilience against environmental hazards (Koncagül, Connor, and Abete 2024; Reduction 2023).

The results are in line with previous studies which have demonstrated that there are number of hydrochemical and anthropogenic processes impacting water quality. Multivariate techniques have also been shown to be valuable in previous studies to find hidden patterns and important factors in complex environmental data sets. The present data set, however, did not consider heavy metals, microbial indicators or location-specific stations which makes the results of the present study not directly comparable to more comprehensive assessments conducted in the field (Bhuyan et al. 2025; Chen et al. 2018).

The lack of temporal and seasonal information, as well as geographical information, was a restriction of the study. There were also no data on biological indicators, microbial contamination, heavy metals or land-use characteristics. These constraints limited the ability to identify sources, to interpret the spatial nature of the pollution and to evaluate long-term trends. In the future, integration of physicochemical data with biological, geospatial and temporal data and machine learning should be utilized for more comprehensive water-quality assessment.

5. Conclusion

Multivariate statistical analysis of physicochemical parameters of a publicly available drinking-water dataset was performed to evaluate water quality and environmental sustainability. Results showed that there was significant variation in measured parameters; total dissolved solids, conductivity, organic carbon and turbidity all made significant contributions to the variation in water quality. Generally, weak relationships between variables were found by the correlation analysis, which means that none of the parameters was a good indicator of water quality status.

Multiple hydrochemical processes and not a single pollution source were responsible for the water quality as shown by Principal Component Analysis and Factor Analysis, while Cluster Analysis classified the samples based on the similarity of their physicochemical characteristics but did not clearly separate potable and non-potable water. The results of non-potable samples show the importance of continuous monitoring, effective treatment and preventive management of the water resources for providing safe drinking water. The study validates the use of multivariate statistical techniques as a powerful tool for data reduction, identification of basic water quality parameters and use in making environmental decisions based on evidence. Despite the absence of temporal, spatial and biological data, the use of the integrated analytical approach was useful in characterising the water quality patterns and the approach was found to be applicable for sustainable freshwater assessment. Future studies are needed to incorporate geospatial data, seasonal monitoring, biological indicators, and advanced machine learning techniques for better identification of pollution sources, more accurate predictive modeling, and more robust long-term sustainable approaches to water resource management and public health protection.

References

1. Bhuyan, Md Simul, Muhammad Abu Bakar, Md Tarikul Islam, Gowhar Meraj, Mrityunjoy Kunda, Mir Mohammad Ali, Nazmul Hossain, Ranjan Roy, Dinesh Chandra Shaha, and Stanley Iheanacho. 2025. “Assessment of Ecological and Human Health Concerns Associated with Heavy Metal Poisoning in the Sediments of Western St. Martin’s Island, Bangladesh.” Estuarine Management and Technologies 2:9–42.
2. Cardia, Marco, Stefano Chessa, Alessio Micheli, Antonella Giuliana Luminare, and Francesca Gambineri. 2026. “Water Quality Estimation Through Machine Learning Multivariate Analysis.” Pp. 261–73 in Neural Networks: Overview of Current Theories and Applications. Vol. 459, Smart Innovation, Systems and Technologies, edited by A. Esposito, M. Faundez-Zanuy, F. C. Morabito, E. Pasero, and G. Cordasco. Singapore: Springer Nature Singapore.
3. Chen, Tao, Huafei Zhang, Chengxun Sun, Hongyan Li, and Yang Gao. 2018. “Multivariate Statistical Approaches to Identify the Major Factors Governing Groundwater Quality.” Applied Water Science 8(7):215. doi:10.1007/s13201-018-0837-0.
4. Crane, Stuart, Christian Tottrup, Torsten Bondo, Silvia Huber, Cécile MM Kittel, Daniel Druce, Mads Christensen, Razvan Bertea, and Jonas B. Sølvsteen. 2021. “Progress on Freshwater Ecosystems: Tracking SDG 6 Series–Global Indicator 6.6. 1 Updates and Acceleration Needs.” https://www.developmentaid.org/api/frontend/cms/file/2021/09/SDG6_Indicator_Report_661_Progress-on-Water-related-Ecosystems_2021_ENGLISH_pages-1_compressed.pdf.

5. Das, Niladri, Prolay Mondal, Ranajit Ghosh, and Subhasish Sutradhar. 2019. “Groundwater Quality Assessment Using Multivariate Statistical Technique and Hydro-Chemical Facies in Birbhum District, West Bengal, India.” SN Applied Sciences 1(8):825.
6. Dhanush, S. K., M. Murthy, RINKU VERMA, and A. Sathish. 2024. “Assessment of Lake Water Quality and Suitability for Consumption, Irrigation and Industrial Purpose Using Multivariate Statistical Approach, Water Quality Indices (WQI) and Geographic Information System (GIS).” Mysore Journal of Agricultural Sciences 58(1). https://old.uasbangalore.edu.in/images/2024-1st-Issue/8.pdf.
7. Edokpayi, Joshua N., John O. Odiyo, and Olatunde S. Durowoju. 2017. “Impact of Wastewater on Surface Water Quality in Developing Countries: A Case Study of South Africa.” Water Quality 10(66561):10–5772.
8. Han, Xue, Fei Tang, and Ai-Lin Liu. 2023. “Drinking Water Quality Evaluation in Supply Systems in Wuhan, China: Application of Entropy Weight Water Quality Index and Multivariate Statistical Analysis.” Environmental Science and Pollution Research 31(1):280–92. doi:10.1007/s11356-023-31212-1.
9. Hendawy, Ehab, Abdel-Aziz A. Belal, Nazih Y. Rebouh, Mohamed S. Shokr, Elsayed Said Mohamed, Abd El Aziz S. Sheta, and Ayman F. Abou-Hadid. 2024. “Assessing Surface Water Quality Using Risk Indicators, Geographic Information System Modeling Techniques, and Multi-Statistical Methods in Arid Regions to Maintain the Sustainability of Water Resources.” Agronomy 14(12):2834.
10. Jones, Edward R., Frederik Kratzert, and Michelle TH van Vliet. 2026. “Caravan-Qual: A Global Scale Integration of Stream Water Quality Observations into a Large-Sample Hydrology Dataset.” Scientific Data. https://www.nature.com/articles/s41597-026-07352-7.
11. Kadiwal, Aditya. 2021. “Water Quality.” https://www.kaggle.com/datasets/adityakadiwal/water-potability.
12. Koncagül, Engin, Richard Connor, and Valentina Abete. 2024. “The United Nations World Water Development Report 2024: Water for Prosperity and Peace; Facts, Figures and Action Examples.” https://www.sidalc.net/search/Record/dig-unesdoc-ark:-48223-pf0000388952/Description.
13. Kumar, Prince, Soumya Sucharita Singha, and Sudhakar Singha. 2026. “Groundwater Quality Assessment in Bihar’s Aquifers: A Machine Learning Approach.” Environmental Science and Pollution Research 33(18):8792–8819. doi:10.1007/s11356-026-37867-w.
14. Luo, Huoqian, Xizhi Nong, Huajie Xia, Huabin Liu, Lingshuang Zhong, Yao Feng, Wenjuan Zhou, and Yu Lu. 2024. “Integrating Water Quality Index (WQI) and Multivariate Statistics for Regional Surface Water Quality Evaluation: Key Parameter Identification and Human Health Risk Assessment.” Water 16(23):3412.

15. Madhav, Sughosh, Virendra Bahadur Singh, Manoj Kumar, and Sandeep Singh. 2022. Hydrogeochemistry of Aquatic Ecosystems. John Wiley & Sons.
16. Muniz, Daphne HF, and Eduardo C. Oliveira-Filho. 2023. “Multivariate Statistical Analysis for Water Quality Assessment: A Review of Research Published between 2001 and 2020.” Hydrology 10(10):196.
17. Organization, World Health. 2022. Guidelines for Drinking-Water Quality: Incorporating the First and Second Addenda. World Health Organization.
18. Organization, World Health. 2025. Progress on Household Drinking Water, Sanitation and Hygiene 2000-2024: Special Focus on Inequalities. World Health Organization.
19. Pham, Nguyen Quoc, and Giao Thanh Nguyen. 2024. “Evaluating Groundwater Quality Using Multivariate Statistical Analysis and Groundwater Quality Index.” Civil Engineering Journal 10(3):699–713.
20. Reduction, United Nations Office for Disaster Risk. 2023. Global Assessment Report on Disaster Risk Reduction 2023: Mapping Resilience for the Sustainable Development Goals. Stylus Publishing, LLC.
21. Shukla, Bishnu Kant, Ruchi Saraswat, Nikunj Agarwal, Hans Kumar Singh, and Shivam Verma. 2025. “A Comparative Study of IoT-Based Water Quality Monitoring Systems (IoT-WQMS) and the Potential of Machine Learning in Water Quality Assessment.” Pp. 1–23 in Unleashing the Power of Blockchain and IoT for Water Informatics, Water Informatics for Water Resource Management, edited by S. S. Rawat, N. Malsa, G. Kumar, and V. Gupta. Singapore: Springer Nature Singapore.
22. Soetan, Oluwafemi, Jing Nie, Krishna Polius, and Huan Feng. 2024. “Application of Time Series and Multivariate Statistical Models for Water Quality Assessment and Pollution Source Apportionment in an Urban River, New Jersey, USA.” Environmental Science and Pollution Research 31(52):61643–59. doi:10.1007/s11356-024-35330-2.
23. Srivastava, Srishti, and Abdul Malik. 2025. “Drinking Water Quality Assessment through Multivariate Statistical Approach and Formulated PCA-Based Contamination Index.” Journal of Contaminant Hydrology 104764.
24. Tajbakhshian, Maryam. 2025. “Integrated Hydrogeochemical and Spatial Assessment of Groundwater Quality Using Multivariate Statistics and GIS: A Case Study from Northeastern Iran.” Environmental Systems Research 14(1):19. doi:10.1186/s40068-025-00412-8.
25. Xia, Pei, Jian Peng, Tianwei Gu, Shuying Yu, and Zihan Xu. 2025. “Lake-Related Ecosystem Services Facing Social–Ecological Risks.” People and Nature 7(4):734–51.
26. Xu, Changjiang, Di Zhu, Wei Guo, Shuo Ouyang, Liping Li, Hui Bu, Lin Wang, Jian Zuo, and Junhong Chen. 2024. “Multi-Objective Ecological Long-Term Operation of Cascade Reservoirs Considering Hydrological Regime Alteration.” Water 16(13):1849.

27. Zhang, Jiaxin, Lianpeng Zhang, Qi Chai, Yang Shen, Li Ji, Qing Zhao, Xing Li, Wei Liu, and Chunmei Li. 2021. "Insights into Spatiotemporal Variations of the Water Quality in Taihu Lake Basin, China." Environmental Monitoring and Assessment 193(11):757. doi:10.1007/s10661-021-09554-5.