International Journal of Clinical and Biomedical Research
Vol. 11, No. 2 (2026)
pISSN: 2521-0394 · eISSN: 2395-0471

ORIGINAL RESEARCH

ASSESSMENT OF PHYSICOCHEMICAL AND MICROBIAL DETERMINANTS OF GROUNDWATER QUALITY AND IDENTIFICATION OF EMERGING CONTAMINANT RISK CLUSTERS

Syed Shafi Ahmed[1], **Swati Yadav**[2], **Arun Kumar Yadav**[3], **Arshiya Masood Siddiqui**[1*]

[1] *Government Medical College, Budaun, Uttar Pradesh, India.*

[2] *Dr. Ram Manohar Lohia Institute of Medical Sciences, Lucknow, Uttar Pradesh, India.*

[3] *Hind Institute of Medical Sciences, Sitapur, Uttar Pradesh, India.*

*Correspondence: Arshiya Masood Siddiqui <arshiyamasood123@gmail.com>

ABSTRACT

Groundwater is the principal source of drinking water for the majority of India's population, but progressive deterioration of its physicochemical and microbiological quality poses an escalating public health risk. Multivariate statistical techniques can disentangle the dominant pollution sources and identify high-risk site clusters more effectively than univariate comparisons against regulatory limits. *Aim.* To identify the dominant factors driving variation in groundwater quality across four Indian states and to classify the constituent water-quality parameters into contamination-risk groupings using Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA). Method: Secondary data on groundwater quality, collected through the National Water Monitoring Program of the Central Pollution Control Board, for the years 2021–2023, were extracted for Uttar Pradesh, Kerala, Maharashtra and Madhya Pradesh. The dataset comprised physical (temperature), chemical (pH, conductivity, dissolved oxygen, biochemical oxygen demand, nitrate) and microbiological (total coliform, fecal coliform, fecal streptococci) indicators, each reported as minimum and maximum values. Variables were standardised (Z-scores), inspected for normality (Kolmogorov–Smirnov test) and submitted to PCA after verifying sampling adequacy using the Kaiser–Meyer–Olkin (KMO) measure and Bartlett's test of sphericity. Components with eigenvalues greater than 1 were retained and rotated using the Varimax procedure. Hierarchical cluster analysis (Ward's linkage, Euclidean distance) of standardised parameters was used to identify groups of co-varying indicators. All twenty parameters deviated significantly from a normal distribution (Kolmogorov–Smirnov $p < 0.001$). Mean total coliform counts (11,162–127,068 MPN/100 mL) and biochemical oxygen demand (BOD; 3.7–6.8 mg/L) were well above potable-water thresholds, and conductivity reached extreme values in localised sites. Spearman correlations showed strong inverse association between dissolved oxygen (DO) and BOD ($r = -0.60$ to -0.67; $p < 0.001$) and strong positive association between BOD and microbial indicators ($r = 0.40$–0.65; $p < 0.001$). The KMO measure was 0.683 and Bartlett's test was significant ($p < 0.001$), justifying PCA. Six components were extracted, cumulatively explaining 70.5% of total variance. The components were interpreted as faecal microbial contamination, organic load / oxygen demand, ionic and acid–base chemistry, physical thermal regime, nutrient enrichment, and residual variation. HCA identified three coherent groupings of parameters representing high-, moderate- and low-risk contamination categories.Microbial and organic pollution were the dominant determinants of groundwater quality in this multi-state dataset, with chemical and nutrient enrichment as secondary drivers. The findings point to widespread untreated sewage infiltration and agricultural runoff as the principal anthropogenic sources, and underscore the need for source protection, structured wastewater management, point-of-use disinfection and continuous monitoring. Integration of emerging organic contaminants into future surveillance is essential to strengthen risk-based groundwater governance in India.

Keywords: Groundwater quality; Principal component analysis; Hierarchical cluster analysis; Fecal coliform; Biochemical oxygen demand; Nitrate contamination; Public health risk.

INTRODUCTION

Groundwater is among the most important components of the global freshwater supply, providing all or part of the drinking water for almost half of the world's population and sustaining approximately 43% of irrigated agriculture

worldwide [1]. An estimated 2.5 billion people rely on groundwater for their basic water requirements [2]. India accounts for nearly 18% of the world's population but holds only about 4% of global freshwater resources, of which groundwater contributes roughly 30% of the national reserve [3]. More than 60% of irrigated agriculture and approximately 85% of the rural drinking-water supply in India are dependent on groundwater extraction [4]. Sustained over-abstraction, rapid urbanisation, industrial expansion and the intensification of agriculture have, however, progressively degraded both the quantity and the quality of these aquifers.

Anthropogenic pressures introduce complex mixtures of physical, chemical and microbiological contaminants into groundwater, with consequences that span both acute infectious disease and chronic disease risk. Globally, unsafe drinking water remains one of the leading risk factors for diarrhoeal disease, contributing substantially to childhood mortality in low- and middle-income countries [5,6]. In India, fecal contamination of groundwater from untreated sewage continues to drive recurrent outbreaks of cholera, typhoid, dysentery and viral hepatitis [7]. In parallel, naturally occurring contaminants particularly arsenic in the Indo-Gangetic plain and fluoride across large tracts of central and southern India pose substantial chronic-disease burdens, including malignancy, skeletal and dental fluorosis, and neurological disease [4,8]. Anthropogenic inputs of fertilisers, pesticides and industrial effluents have further raised nitrate, phosphate and microbial concentrations beyond permissible regulatory limits, with elevated nitrate exposure causally linked to infantile methaemoglobinaemia ('blue baby syndrome') and implicated in the formation of carcinogenic N-nitroso compounds [9].

Superimposed on these conventional contaminants is a newer and rapidly evolving class of pollutants the so-called emerging contaminants (ECs) which include pharmaceuticals and their metabolites, personal-care product residues, microplastics, antibiotic residues and endocrine-disrupting chemicals [10,11]. These compounds typically enter aquifers via wastewater effluents, landfill leachates and unregulated industrial discharge and are of growing concern because of their persistence, potential for bioaccumulation and capacity to drive antimicrobial resistance [10]. Despite their public health relevance, ECs are not yet routinely incorporated into Indian groundwater surveillance programmes, leaving an important gap in national monitoring.

Because groundwater quality reflects the simultaneous interplay of geogenic, agricultural, industrial and hydrological influences, multivariate statistical approaches principally Principal Component Analysis (PCA), Factor Analysis and Hierarchical Cluster Analysis (HCA) are now established tools for reducing the dimensionality of complex hydrochemical datasets, identifying dominant pollution sources, and distinguishing natural from anthropogenic signals [12,13]. PCA has been used effectively to recover latent dimensions of microbial–organic load, nutrient enrichment and ionic salinisation in groundwater datasets from diverse settings [14,15], and HCA has been applied to group monitoring stations and parameters into coherent contamination-risk strata [12,16].

A persistent gap nonetheless exists in the Indian literature: large publicly available national datasets are underutilised in integrated multivariate analyses that simultaneously consider physicochemical and microbiological indicators within a public health framework [16,17]. Most published studies remain confined to traditional hydrochemical parameters and do not explicitly link statistical patterns to health-risk interpretation. The present study addresses this gap by applying a combined PCA and HCA framework to three years (2021–2023) of CPCB groundwater monitoring data drawn from four Indian states, with the dual objective of identifying the dominant determinants of groundwater quality variation and stratifying parameters into contamination-risk groupings of public health relevance.

The specific objectives of the study were: (i) to describe the distribution of physicochemical and microbiological parameters across monitoring stations between 2021 and 2023; (ii) to examine the structure of correlations among chemical and microbial indicators; (iii) to classify water-quality parameters into risk-based clusters using HCA; and (iv) to identify the principal latent factors driving water-quality variance using PCA.

DATA AND METHODS

Data source and study area

The study used publicly available secondary data on groundwater quality, collected and reported by the Central Pollution Control Board (CPCB) under the Ministry of Environment, Forest and Climate Change, Government of India, through its National Water Monitoring Program (NWMP). Annual NWMP datasets for the years 2021, 2022 and 2023 were downloaded from the CPCB portal. Monitoring stations from four geographically and hydrogeologically diverse Indian states Uttar Pradesh, Kerala, Maharashtra and Madhya Pradesh were retained for analysis, reflecting a range of climatic conditions, agricultural intensities and urbanisation patterns. The CPCB and the collaborating State Pollution Control Boards collect samples at standardised intervals through their regional laboratories using harmonised protocols for sampling and chemical and microbiological analysis [18].

Parameters analysed

Ten core groundwater quality parameters were extracted: physical (temperature), chemical (pH, electrical conductivity, dissolved oxygen [DO], biochemical oxygen demand [BOD], nitrate) and microbiological (total coliform, fecal coliform, fecal streptococci). Each indicator was reported as a minimum and a maximum value across the calendar year for every station, yielding 20 derived variables. These parameters were selected because they collectively capture the principal physical, chemical and microbiological dimensions of drinking-water quality assessed against the World Health Organization (WHO) drinking-water guidelines and the Bureau of Indian Standards (BIS) IS 10500 specification [19,20].

Data pre-processing

Records with extensive missing values were excluded. The remaining dataset was checked for inconsistencies and entry errors. Descriptive statistics mean, standard deviation, skewness and kurtosis were computed for each variable. Normality was tested with the Kolmogorov–Smirnov (K–S) test; because all variables deviated significantly from a normal distribution ($p < 0.001$ for every parameter), non-parametric Spearman rank correlation was used in subsequent bivariate analyses. To remove the influence of differing units of measurement and to mitigate the leverage of outliers, all variables were converted to Z-scores before multivariate analysis.

Principal Component Analysis (PCA)

PCA was performed on the standardised dataset to identify the latent dimensions accounting for the variance in groundwater quality. The suitability of the dataset for factor extraction was assessed using the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity [21,22]. Principal components with eigenvalues greater than 1 were extracted in accordance with the Kaiser criterion, supplemented by visual inspection of the scree plot. Varimax orthogonal rotation was applied to facilitate interpretation by maximising the simple structure of loadings. Variables with extraction communalities greater than 0.6 were considered to be well represented by the retained components.

Hierarchical Cluster Analysis (HCA)

To complement the PCA, R-mode hierarchical cluster analysis was performed on the standardised variables using Ward's linkage method with squared Euclidean distance as the dissimilarity measure [12]. The output is presented as a dendrogram, and clusters of co-varying parameters were interpreted as representing distinct contamination signatures of differing potential public health risk.

Statistical analysis

All statistical analyses were performed using IBM SPSS Statistics version 25.0 (IBM Corp., Armonk, NY, USA). A two-sided p-value of less than 0.05 was considered statistically significant.

RESULTS

Descriptive characterisation of groundwater quality

Marked variability was observed across all measured parameters (Table 1). Groundwater temperature ranged from 1.6 °C to 46 °C, with mean minimum and maximum values of 20.05 °C and 28.69 °C, respectively; the relatively modest standard deviations and near-symmetric distributions indicate a stable thermal regime across most sites, although the upper end of the range can favour microbial proliferation and reduce DO solubility.

Dissolved oxygen ranged from 0 to 28 mg/L (mean 5.63–7.58 mg/L), with several samples falling below the 5 mg/L threshold consistent with high organic loading and oxygen depletion. The pH values ranged from 2.0 to 9.2 (mean 7.40–8.19), within the BIS desirable range of 6.5–8.5 in most cases, but with extreme outliers indicating localised acidification. Electrical conductivity varied widely, with maximum values reaching 71,212 µmhos/cm and high positive skewness (10.7) and kurtosis (139.3) indicative of a small number of saline or mineralised outlier sites. BOD ranged from 1 to 127 mg/L (mean 3.69–6.77 mg/L), substantially above the < 3 mg/L benchmark typically used to indicate potable water quality, while total coliform counts ranged from 2 to 2.8 × 10^7 MPN/100 mL with extreme skewness (17–21) and kurtosis (333–527), reflecting the coexistence of relatively clean sources with sites of severe faecal contamination. All variables deviated significantly from normality on the K–S test ($p < 0.001$), confirming the appropriateness of non-parametric correlation.

Table 1. Descriptive statistics of physicochemical and microbiological groundwater parameters across monitoring stations (2021–2023).

Parameter	Mean	SD	Skewness	Kurtosis	K-S p-value
Temperature (°C), min	20.05	3.72	-0.37	1.13	< 0.001
Temperature (°C), max	28.69	4.20	0.45	1.20	< 0.001
Dissolved oxygen (mg/L), min	5.63	1.77	-1.42	1.96	< 0.001
Dissolved oxygen (mg/L), max	7.58	2.22	0.27	10.63	< 0.001
pH, min	7.40	0.47	- 3.85	42.72	< 0.001
pH, max	8.19	0.35	- 0.25	-0.38	< 0.001
Conductivity (µmhos/cm), min	337.34	211.71	2.01	6.17	< 0.001
Conductivity (µmhos/cm), max	1411.12	4307.88	10.66	139.28	< 0.001
BOD (mg/L), min	3.69	6.60	5.66	35.07	< 0.001
BOD (mg/L), max	6.77	10.63	4.89	32.92	< 0.001
Total coliform (MPN/100 mL), min	11,161.7	83,035.4	21.35	526.65	< 0.001
Total coliform (MPN/100 mL), max	127,068	1,329,536	17.47	332.94	< 0.001

Correlation structure

Spearman correlations among the key parameters are summarised in Table 2. DO and BOD were strongly inversely correlated ($r = -0.597$ for DO minimum vs. BOD minimum; $r = -0.671$ for DO minimum vs. BOD maximum; both $p < 0.001$), consistent with biochemical oxygen consumption by aerobic decomposition of organic matter. BOD was strongly and positively correlated with both fecal coliform ($r = 0.53–0.64$) and fecal streptococci ($r = 0.37–0.42$; all $p < 0.001$), indicating that organic load and microbial contamination co-occur. Fecal coliform and fecal streptococci were themselves strongly inter-correlated ($r = 0.55–0.73$; $p < 0.001$), reinforcing their joint utility as indicators of faecal contamination. Nitrate (maximum) showed moderate inverse correlation with DO ($r = -0.51$; $p < 0.001$) and positive correlation with BOD ($r = 0.40–0.48$; $p < 0.001$), but only weak correlations with microbial indicators ($r = 0.05–0.20$), suggesting a partly distinct pollution pathway most plausibly agricultural fertiliser leaching.

Table 2. Spearman rank correlations among key groundwater quality parameters.

	DO min	DO max	BOD min	BOD max	NO_3 max	FC min	FC max
DO min	1.00						
DO max	0.636**	1.00					
BOD min	-0.597**	-0.528**	1.00				
BOD max	-0.671**	-0.507**	0.850**	1.00			
Nitrate max	-0.506**	-0.381**	0.396**	0.475**	1.00		
Fecal coliform (FC) min	-0.263**	-0.013	0.642**	0.580**	0.107**	1.00	
Fecal coliform (FC) max	-0.252**	0.101**	0.533**	0.568**	0.200**	0.860**	1.00
Fecal streptococci (FS) min	-0.180**	-0.071	0.390**	0.402**	0.052	0.579**	0.558**
Fecal streptococci (FS) max	-0.197**	-0.053	0.367**	0.420**	0.077	0.485**	0.587**

** Correlation significant at the 0.01 level (two-tailed); * significant at the 0.05 level. DO, dissolved oxygen; BOD, biochemical oxygen demand; NO_3, nitrate; FC, fecal coliform; FS, fecal streptococci.

Hierarchical cluster analysis of parameters

Hierarchical cluster analysis performed on the standardised parameters generated a dendrogram (Figure 1) showing three coherent groupings of indicators. The first cluster comprised BOD, DO and the coliform group, representing a high-risk microbial–organic pollution signature consistent with sewage intrusion. A second cluster brought together electrical conductivity and nitrate, indicating a chemical and nutrient pollution signature consistent with agricultural runoff and industrial discharge. A third cluster contained temperature and pH, representing physical and acid–base background variation with comparatively limited direct contamination significance, though both can modulate the bioavailability and biological activity of pollutants. Early-stage merges between BOD minimum and maximum, and between DO minimum and maximum, confirmed the close internal coherence of these organic-load indicators.

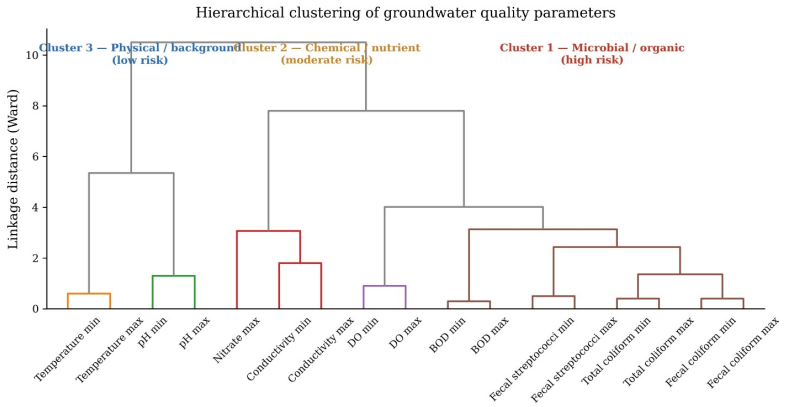

Figure 1. *Dendrogram of R-mode hierarchical cluster analysis*

Note: Ward's linkage, squared Euclidean distance of standardised groundwater quality parameters, showing three coherent groupings: a high-risk microbial–organic cluster (BOD, DO, coliforms, streptococci), a moderate-risk chemical–nutrient cluster (conductivity, nitrate) and a low-risk physical–background cluster (temperature, pH).

Principal Component Analysis

The KMO measure of sampling adequacy was 0.683, within the acceptable range of 0.6–0.8, and Bartlett's test of sphericity was statistically significant ($p < 0.001$) (Table 3 caption note), confirming that the dataset was suitable for PCA. Communalities after extraction exceeded 0.60 for most variables and were particularly high for total coliform (0.97), fecal coliform (0.82–0.97), nitrate maximum (0.88) and BOD maximum (0.86), indicating that these variables were strongly represented by the extracted components.

Six components with eigenvalues greater than 1 were retained, cumulatively explaining 70.53% of total variance. The scree plot showed a clear inflection at the sixth component, supporting this retention (Figure 2).

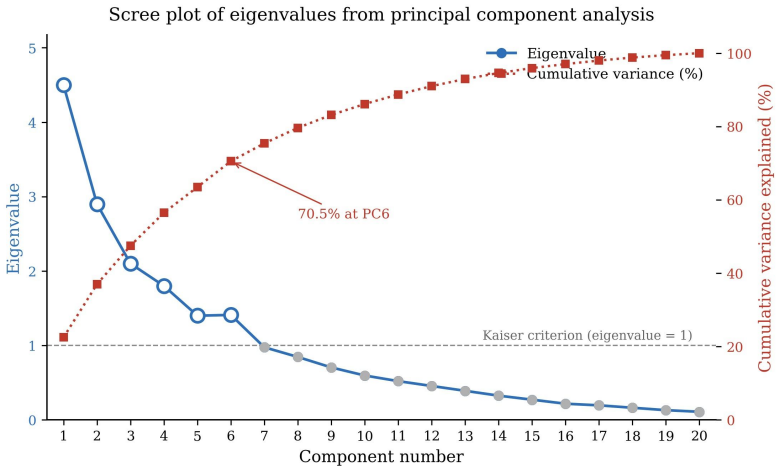

Figure 2. *Scree plot of eigenvalues from principal component analysis of standardised groundwater quality parameters.*

Note: The marked inflection beyond the sixth component supports retention of six principal components, which together explained 70.5% of total variance.

Following Varimax rotation, the components were interpretable as follows: Component 1 carried strong loadings of total and fecal coliforms and fecal streptococci, representing faecal microbial contamination; Component 2 was dominated by BOD with an inverse loading of DO, representing organic pollution and oxygen demand; Component 3 was loaded by pH and conductivity, capturing acid–base and ionic chemistry; Component 4 reflected temperature, representing the physical thermal regime; Component 5 was driven by nitrate maximum, capturing nutrient enrichment from agricultural sources; and Component 6 carried smaller, mixed loadings interpreted as residual variation (Table 3).

Table 3. Summary of Varimax-rotated principal components extracted from groundwater quality parameters and their interpretation.

Parameter	PC1	PC2	PC3	PC4	PC5	PC6
Total coliform (min/max)	strong					
Fecal coliform (min/max)	strong					
Fecal streptococci (min/max)	strong					
BOD (min/max)		strong				
Dissolved oxygen (min/max)		strong (−)				
pH (min/max)			strong			
Conductivity (min/max)			strong			
Temperature (min/max)				strong		
Nitrate max					strong	
Mixed parameters (minor)						loading
Interpreted as	Faecal microbial contamination	Organic pollution / oxygen demand	Ionic / acid-base chemistry	Physical thermal regime	Nutrient enrichment	Residual variation

Note: KMO measure of sampling adequacy = 0.683; Bartlett's test of sphericity, χ^2 test statistic significant at $p < 0.001$. Six components extracted with eigenvalues > 1, cumulatively explaining 70.53% of total variance. 'Strong' denotes a Varimax-rotated factor loading of high absolute magnitude; '(-)' indicates an inverse loading on the component.

DISCUSSION

This multivariate analysis of three years of national groundwater monitoring data from four geographically diverse Indian states identifies microbial and organic pollution as the dominant determinants of groundwater quality variation, with chemical and nutrient enrichment emerging as a secondary, partly distinct, signature. Six principal components together accounted for 70.5% of total variance, with Component 1 (faecal microbial contamination) and Component 2 (organic load/oxygen demand) carrying the largest share. Hierarchical clustering of parameters produced three coherent groupings microbial–organic, chemical–nutrient and physical–background that mirror the PCA structure and offer a parsimonious framework for risk-based interpretation of groundwater data.

The dominance of the microbial–organic signal in our analysis is consistent with several recent multivariate studies of Indian groundwater. Panghal and Bhateria, using factor and cluster analysis of groundwater data from Beri block, Haryana, identified microbial and organic pollutants as the leading drivers of variance and attributed them principally to untreated sewage and septic infiltration [16]. Masood and colleagues, who integrated multivariate analysis with a water quality index in a comparable Indian setting, similarly found that microbial and organic indicators co-clustered with elevated electrical conductivity and were most plausibly attributable to anthropogenic activity [17]. The strong inverse correlation between DO and BOD observed in our data ($r = -0.60$ to -0.67) reproduces a consistent ecological pattern of oxygen depletion driven by aerobic microbial decomposition of organic matter, recently described in peri-urban aquifers in Bangladesh by Farzana et al. [23] and predicted by classical biogeochemical theory.

The total coliform counts observed in our data, with mean maxima exceeding 10^5 MPN/100 mL, sit well above the BIS and WHO standards, which specify the absence of detectable coliforms in drinking water [19,20]. Comparable patterns of widespread faecal contamination have been reported in groundwater from the Anambra Basin in Nigeria [24] and in Indian shallow aquifers in Agra [25], where untreated sewage discharge, poor sanitation and septic infiltration were similarly implicated. Sustained exposure to such water is associated with diarrhoeal disease, typhoid, dysentery, cholera and viral hepatitis [5,7], and the elevated BOD found in our dataset is consistent with ongoing decomposition of biodegradable organic waste, which both depletes oxygen and supports microbial proliferation.

The second principal component in our analysis dominated by electrical conductivity and acid–base chemistry, and supplemented in our clustering by nitrate captures a partially distinct chemical and nutrient signature. Local outliers with extreme conductivity values (up to 71,212 µmhos/cm) and the moderate positive correlation between nitrate maximum and BOD ($r = 0.40$–0.48) suggest mixed sources: ionic enrichment likely reflecting saline intrusion or geogenic mineralisation, and nitrate likely reflecting agricultural fertiliser leaching superimposed on partial sewage contributions. Similar dual signatures have been reported by Alsubih et al. in Delhi industrial areas [26] and by Barad et al. in Western Odisha [27], who attributed elevated nitrate principally to fertiliser leaching and septic infiltration. The relatively weak correlation between nitrate and microbial indicators in our dataset further supports the inference of multiple, partly independent pollution pathways, consistent with the observations of Mathew and Kanmani [10]. Nitrate concentrations above the BIS limit of 45 mg/L, observed at a number of stations in our data, are of particular public health concern given the established causal link between nitrate ingestion and infantile methaemoglobinaemia and the suggestive association with adult gastrointestinal cancers mediated by endogenous nitrosamine formation [9].

From a methodological standpoint, our six-factor PCA solution explaining 70.5% of total variance is broadly consistent with published applications of PCA to groundwater datasets. Muniz and Oliveira-Filho, in a systematic review of multivariate statistical analyses of water quality between 2001 and 2020, reported that the majority of published PCA solutions extract three to seven components and cumulatively explain between 60% and 80% of variance [12]. The KMO value of 0.683 obtained here is within the acceptable range, and the dominance of microbial–organic factors among the leading components is a feature shared with the studies cited above [14–17,23–27]. The convergence of PCA and HCA findings both methods independently isolating a microbial–organic, a chemical–nutrient and a physical–background grouping lends internal consistency to the interpretation and supports the use of such combined chemometric frameworks as a routine analytical approach for national groundwater datasets.

The present study did not directly measure emerging organic contaminants such as pharmaceuticals, personal-care products and microplastics. Banerjee et al. have argued that statistical frameworks of the kind used here are well suited to integration of EC monitoring data once such measurements become routinely available [11], and recent multivariate work has begun to incorporate selected ECs into hydrochemical analyses [10]. Extension of the present framework to include these contaminants is an important priority and is consistent with the broader One Health agenda for water-related antimicrobial resistance surveillance.

Taken together, the findings reaffirm that microbial, organic and nutrient contamination are the principal determinants of groundwater degradation in the studied states, that these signals can be disentangled by routine multivariate techniques, and that the resulting risk groupings provide an evidence base for targeted public health action. Source protection (well-head and aquifer protection zones), structured wastewater management, source-level and household-level disinfection (chlorination, UV treatment), and continuous parametric and microbiological monitoring are the most plausible high-yield interventions and align with the requirements of Sustainable Development Goal 6 (clean water and sanitation) and Sustainable Development Goal 3 (good health and wellbeing) [28].

CONCLUSION

Multivariate analysis of three years of national groundwater monitoring data from four Indian states identified six latent factors that together explained more than 70% of total variation in water quality. Faecal microbial contamination and organic pollution emerged as the two most prominent factors, followed by ionic chemistry, physical thermal variation, nutrient enrichment and residual variability. Hierarchical clustering of parameters independently produced congruent groupings interpretable as high-, moderate- and low-risk contamination signatures. Elevated coliform counts and biochemical oxygen demand, well above BIS and WHO benchmarks at a substantial proportion of stations, indicate widespread untreated sewage infiltration with attendant risk of diarrhoeal and enteric infectious disease; localised excess nitrate signals fertiliser leaching with risk of methaemoglobinaemia and potential carcinogenicity. Routine integration of PCA and HCA into national groundwater surveillance and prospectively the inclusion of emerging organic contaminants offers a scalable, evidence-based framework for risk-based water governance and targeted public health intervention in India.

Limitations: The analysis is based on secondary data and is therefore constrained by the parameters and reporting practices of the CPCB monitoring framework; emerging contaminants, heavy metals and several geogenic species (arsenic, fluoride, iron) were not included in the present dataset. While the PCA structure provides a parsimonious description of pollution sources, it does not establish causal attribution; integration with land-use data, sanitation indicators and meteorological information would be needed to formally infer source contributions.

DECLARATIONS

Ethical approval: Not applicable. The study used publicly available, de-identified secondary data on environmental water quality and did not involve human or animal participants.

Data source: Central Pollution Control Board, Ministry of Environment, Forest and Climate Change, Government of India: https://cpcb.nic.in/nwmp-data-2023/.

Conflicts of interest: The authors declare no conflicts of interest.

References

1. Margat J, van der Gun J. Groundwater around the world: a geographic synopsis. Boca Raton (FL): CRC Press; 2013.
2. United Nations. The United Nations World Water Development Report 2022: Groundwater Making the Invisible Visible. Paris: UNESCO; 2022.
3. Central Ground Water Board (CGWB). National Compilation on Dynamic Ground Water Resources of India, 2022. New Delhi: Ministry of Jal Shakti, Government of India; 2022.
4. Mukherjee A, Saha D, Harvey CF, Taylor RG, Ahmed KM, Bhanja SN. Groundwater systems of the Indian sub-continent. J Hydrol Reg Stud. 2015;4:1–14.

5. Prüss-Ustün A, Wolf J, Bartram J, Clasen T, Cumming O, Freeman MC, et al. Burden of disease from inadequate water, sanitation and hygiene for selected adverse health outcomes: an updated analysis with a focus on low- and middle-income countries. Int J Hyg Environ Health. 2019;222(5):765–777.

6. GBD 2017 Diarrhoeal Disease Collaborators. Quantifying risks and interventions that have affected the burden of diarrhoea among children younger than 5 years: an analysis of the Global Burden of Disease Study 2017. Lancet Infect Dis. 2020;20(1):37–59.

7. World Health Organization. Drinking-water [fact sheet]. Geneva: WHO; 2023.

8. Chakraborti D, Singh SK, Rahman MM, Dutta RN, Mukherjee SC, Pati S, et al. Groundwater arsenic contamination in the Ganga River Basin: a future health danger. Int J Environ Res Public Health. 2018;15(2):180.

9. Ward MH, Jones RR, Brender JD, de Kok TM, Weyer PJ, Nolan BT, et al. Drinking water nitrate and human health: an updated review. Int J Environ Res Public Health. 2018;15(7):1557.

10. Mathew RA, Kanmani S. A review on emerging contaminants in Indian waters and their treatment technologies. Nat Environ Pollut Technol. 2020;19(2):549–562.

11. Banerjee A, Singh S, Ghosh A. Detection and removal of emerging contaminants from water bodies: a statistical approach. Front Anal Sci. 2023;3:1115540.

12. Muniz DH, Oliveira-Filho EC. Multivariate statistical analysis for water quality assessment: a review of research published between 2001 and 2020. Hydrology. 2023;10(10):196.

13. Singh KP, Malik A, Mohan D, Sinha S. Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India): a case study. Water Res. 2004;38(18):3980–3992.

14. Alshahrani M, Ahmad M, Laiq M, Nabi M. Geostatistical analysis and multivariate assessment of groundwater quality. Sci Rep. 2025;15(1):7435.

15. Egbueri JC, Ezugwu CK, Unigwe CO, Onwuka OS, Onyemesili OC, Mgbenu CN. Multidimensional analysis of the contamination status, corrosivity and hydrogeochemistry of groundwater from parts of the Anambra Basin, Nigeria. Anal Lett. 2021;54(13):2126–2156.

16. Panghal V, Bhateria R. A multivariate statistical approach for monitoring of groundwater quality: a case study of Beri block, Haryana, India. Environ Geochem Health. 2021;43(7):2615–2629.

17. Masood A, Aslam M, Pham QB, Khan W, Masood S. Integrating water quality index, GIS and multivariate statistical techniques towards a better understanding of drinking water quality. Environ Sci Pollut Res. 2022;29(18):26860–26876.

18. Central Pollution Control Board. Guide manual: water and wastewater analysis. New Delhi: CPCB, Ministry of Environment, Forest and Climate Change, Government of India; 2017.

19. World Health Organization. Guidelines for drinking-water quality: fourth edition incorporating the first and second addenda. Geneva: WHO; 2022.

20. Bureau of Indian Standards. Drinking water specification (second revision), IS 10500:2012, amended 2015. New Delhi: BIS; 2012.

21. Kaiser HF. An index of factorial simplicity. Psychometrika. 1974;39(1):31–36.

22. Bartlett MS. Tests of significance in factor analysis. Br J Math Stat Psychol. 1950;3(2):77–85.

23. Farzana F, Roy TK, Hossain SA, Mazrin M, Islam MS, Mahiddin NA, et al. Assessment of groundwater quality and potential health risks related to heavy metals in a peri-urban area of a developing country. Sci Rep. 2025;15(1):27970.

24. Egbueri JC. Groundwater quality assessment using pollution index of groundwater (PIG), ecological risk index (ERI) and hierarchical cluster analysis (HCA): a case study. Groundw Sustain Dev. 2020;10:100292.

25. Yadav KK, Gupta N, Kumar V, Choudhary P, Khan SA. GIS-based evaluation of groundwater geochemistry and statistical determination of the fate of contaminants in shallow aquifers from different functional areas of Agra city, India: levels and spatial distributions. RSC Adv. 2018;8(29):15876–15889.

26. Alsubih M, El Morabet R, Khan RA, Khan NA, ul Haq Khan M, Ahmed S, et al. Occurrence and health risk assessment of arsenic and heavy metals in groundwater of three industrial areas in Delhi, India. Environ Sci Pollut Res. 2021;28(44):63017–63031.

27. Barad S, Thakur RR, Nandi D, Bera DK, Sahu PC, Mishra P, et al. Hydrogeochemical and geospatial insights into groundwater contamination: fluoride and nitrate risks in Western Odisha, India. Water. 2025;17(10):1514.

28. United Nations. Transforming our world: the 2030 Agenda for Sustainable Development. New York: UN; 2015.

