Wavelet-ICA Hybrid Spatial Data mining for Multiscale Source Apportionment of Water Pollution in the Latyan Dam Basin

Samira Hombari Amirabadi, Jalal Karami *

Remote Sensing and GIS Department, Tarbiat Modares University, Tehran, Iran - samira.amirabadi@yahoo.com, jl.karami@modares.ac.ir

Keywords: Wavelet–ICA hybrid, Water pollution source apportionment, Latyan Dam basin, Multiresolution analysis, Monthly water-quality time series

Abstract

This study aimed to elucidate the spatio-temporal, multi-scale latent sources of water pollution and conduct source apportionment in the Latyan Dam basin, Jajrud River, northeast Tehran, Iran, a vital drinking water supply for the Tehran metropolitan area. Over six years, from January 2002 through December 2007, nine monitoring stations recorded monthly time series for nine key water quality parameters: nitrate (NO_3^-), ammonia (NH_3), sulfate (SO_4^{2-}), chemical oxygen demand (COD), biological oxygen demand (BOD), dissolved oxygen (DO), total dissolved solids (TDS), nitrite (NO_2^-), and fluoride (F^-)—over six years from January 2002 through December 2007. The record for each parameter was decomposed into low- and high-frequency components using discrete wavelet multiresolution analysis (Daubechies-4). Subsequently, Independent Component Analysis (ICA) was applied to each sub-band to blindly separate independent pollutant source signatures. When applied to the Latyan Dam data, the method discerned multiscale signatures attributable to seasonal agricultural runoff (NO_3^-, NH_3), episodic industrial effluents (SO_4^{2-}, COD, BOD), and urban sewage spikes (TDS, NO_2^-, F^-). Clustering analysis divided the stations into three homogeneous regions based on source contributions. Agriculture was identified as the leading contributor across all stations, with secondary industrial influences observed in some regions. Quantitative evaluation demonstrated that Wavelet-ICA at level 5 yielded the largest correlation (0.37) and high mutual information (0.81). However, negative Signal-to-Interference Ratios (e.g., -3.31 at level 5) indicated significant problems with signal overlaps. This source apportionment framework, which is grounded in six years of monthly observations, provides water resource managers with enhanced temporal and spatial resolution of pollution inputs, thereby facilitating targeted mitigation and reservoir management strategies.

1. Introduction

Over the last two decades, rapid population growth and unplanned urban expansion have intensified pressure on Tehran's limited water resources. The Latyan Dam reservoir, which impounds the Jajrud River northeast of the city, provides roughly 25% of Tehran's potable supply. However, its water quality has deteriorated under the combined influence of agricultural intensification, industrial development, and urban wastewater discharge. Specifically, fertilizer applications in the upstream agricultural plains generate pulses of nitrate (NO_3^-) and ammonia (NH_3) during spring and autumn(Yu et al.,2010, Zhou et al.,2012), while the textile, chemical, and food-processing industries intermittently release sulfate (SO_4^{2-}), chemical oxygen demand (COD), and biological oxygen demand (BOD) loads during process upsets. At the same time, expanding peri-urban settlements contribute diffuse sewage rich in total dissolved solids (TDS), nitrite (NO_2^-), and fluoride (F^-), especially during storm-driven overflow events.(Karami et al.,2012, Karami et al.,2014). Standard monitoring at nine basin stations yields monthly time series for these nine parameters, along with dissolved oxygen (DO). However, aggregate trends obscure the distinct temporal signatures of overlapping sources.

Disentangling these latent pollution inputs is critical for targeted management: seasonal agricultural spikes necessitate adjusted fertilizer schedules and buffer-strip enforcement, while episodic industrial discharges require tighter process controls and emergency response plans. Traditional statistical regression or principal component methods can identify dominant modes of variability but require *a priori* assumptions about source profiles and often blend signals operating at different frequencies into single factors. By contrast, latent source separation techniques infer independent contributors directly from mixed observations, offering the promise of pinpointing pollutant origins without predefined source fingerprints.

Wavelet multiresolution analysis (MRA) provides a natural first step toward this goal by decomposing each water quality record into scale-specific sub-bands. Discrete wavelet transforms (DWT)—notably the Daubechies-4 family—have been used to isolate monsoon and seasonal cycles from high-frequency event pulses in hydrological and water-quality data. (Zhou et al., 2012) demonstrated that the wavelet spectra of Dianchi Lake's water indices reveal hidden seven-year periodicities, while Parmar and Bhardwaj (Parmar and Bhardwaj, 2013) employed DWT to distinguish monsoon versus dry–season behavior in multiple river parameters. Yu et al (Yu et al.,2010) further linked ammonia pollution peaks in the Minjiang River to agricultural fertilizer schedules when analyzed through wavelet decomposition. These studies confirm that wavelet MRA can disentangle temporal scales, yet they stop short of attributing each decomposed component to specific pollution sources.

In parallel, artificial intelligence (AI)–based hybrids have emerged to accurately predict and denoise water quality variables. Bhatt and Swain. (Bhatt and Swain., 2025) reviewed AI strategies—from support vector machines (SVM) to reinforcement learning controllers—for the detection and adaptive control of river pollution, underscoring the need for models that operate in real time under uncertainty. Huan et al. (Hua et al., 2025) integrated the SWAT hydrological model with interpretable deep learning networks to forecast non–point-source impacts, revealing key land-use and climatic drivers. Hybrid frameworks that combine wavelet denoising with neural networks, SVMs, or long short-term memory (LSTM) networks

* Corresponding author

have achieved 10–20 % improvements in forecasting accuracy for ammonia, phosphorus, and DO compared with standalone methods. (Nourani and Farboudfam, 2019; Song, et al.,2021; Nagaraju et al., 2023; Wang, et al., 2024; Wang et al., 2025). Yet, despite these advances in prediction and denoising, few studies have pursued latent source separation in water-quality contexts. Independent Component Analysis (ICA) provides a statistical mechanism to unmix observed signals into independent components. However, direct application to raw water-quality data often fails when source spectra overlap or noise levels are high. Recognizing this, Curceac et al. (Curceac et al., 2021) and Sankaran et al. (Sankaran, et al., 2024) applied multiscale coherence analyses to hydrological and air-quality data, demonstrating that wavelet–statistical hybrids better capture extreme events and cross-variable linkages. However, a dedicated framework that integrates DWT-based MRA with ICA for true source apportionment in a drinking-water reservoir remains lacking.

Land-use dynamics further complicate the picture: Etemadi et al. (Etemadi et al., 2018) investigated land-use dynamics in coastal mangrove ecosystems of southern Iran—ecosystems valued for their carbon storage and other services, yet highly vulnerable to climate change and human impacts. Karami et al. (Karami et al., 2012; Karami et al., 2014; Karami et al., 2025) have demonstrated the power of combining remote sensing products, multivariate statistics, and data mining to trace spatio-temporal pollution patterns and subsidence impacts on contaminant pathways. Yet, these spatial analyses do not resolve the temporal scales at which different sources inject pollutants into the reservoir.

This study addresses these gaps by developing a wavelet–ICA hybrid algorithm for latent source apportionment of water pollution in the Latyan Dam basin, using monthly measurements from nine stations over six years (2002–2007). Our specific objectives are to:

1. Decompose each water-quality parameter via DWT (Daubechies 4) into low- and high-frequency sub-bands, isolating seasonal trends from event-driven fluctuations.
2. Unmix independent pollutant signatures by applying ICA within each wavelet sub-band, thereby enhancing noise suppression and scale-specific source recovery.
3. Apply clustering to group stations with similar pollution profiles, identifying spatially homogeneous regions.
4. Interpret real-world results to apportion latent sources—seasonal agricultural runoff (NO_3^-, NH_3), episodic industrial discharges (SO_4^{2-}, COD, BOD), and urban sewage spikes (TDS, NO_2^-, F^-)—providing reservoir managers with temporally resolved insights for targeted mitigation.

By integrating multiresolution signal processing with latent source separation, our work advances both methodological frontiers and practical tools for safeguarding Tehran's critical drinking water supply. The remainder of this paper is organized as follows: theoretical foundations, details of the hybrid algorithm, the study area and data, results and discussion, and recommendations for future research.

2. Materials and Methods

2.1 Study Area

The Latyan Dam basin is located northeast of Tehran, Iran, along the Jajrud River. It is a significant water supply source for the Tehran metropolitan area, providing approximately 25 % (25 million liters/day) of potable water for human consumption. The geographical latitudes and longitudes of the study area range from 35.7° to 35.9° N and 51.6° to 51.8° E, covering an area of roughly 3,300 km². The topography varies from highlands (mountains) in the upstream region to agricultural plains in the midstream and downstream areas. The climate in the region is Mediterranean, with seasonal precipitation occurring mainly during autumn and winter, which affects runoff responses and pollutant transport.

Land uses in the basin are diverse and include agricultural fields (mostly in upstream plain-highlands, with heavy fertilizer application), industrial areas (food production, textile, and chemical processing), and peri-urban settlements surrounding Tehran. Agricultural runoff (nitrates and ammonia) enters surface water in addition to industrial discharges (COD, BOD, sulfates), urban sewage (TDS, nitrites, and fluoride), and livestock waste (nitrates, organic matter). Seasonal precipitation mobilizes loads from agricultural runoff (both in volume and from fields) and highlights the impacts of urban growth and livestock effluent.

To assess water quality, we established nine sampling stations across the basin, covering upstream, midstream, and downstream locations to capture spatial and temporal variations in pollution loading. The locations of the Latyan Dam, Jajrud River, and sampling stations are shown in [Figure 1]. The overall study process and components, including data collection, signal processing, and source apportionment, are illustrated in [Figure 2].

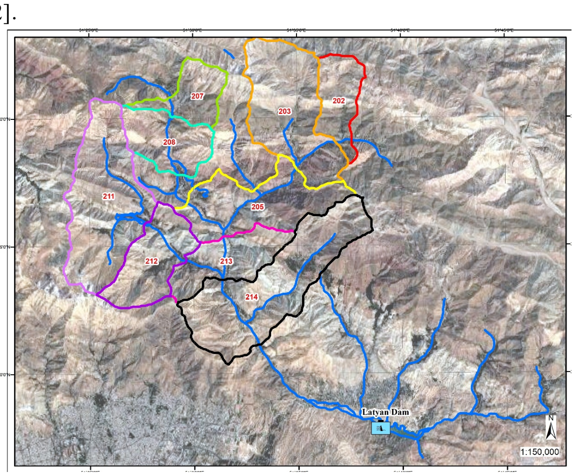

Figure 1. Map of the Latyan Dam basin, showing the location of the dam, Jajrud River, and nine water quality monitoring stations (202–214).

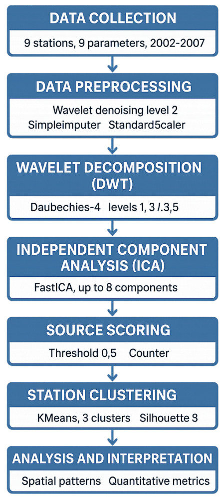

Figure 1. Flowchart of the study methodology, illustrating the steps from data collection to source apportionment using the Wavelet-ICA hybrid approach.

2.2 Data collection

Water quality data were collected from nine monitoring sites within the Latyan Dam basin. This study focuses on nine water quality parameters: nitrate (NO_3^-), ammonia (NH_3), sulfate (SO_4^{2-}), chemical oxygen demand (COD), biological oxygen demand (BOD), dissolved oxygen (DO), total dissolved solids (TDS), nitrite (NO_2^-), and fluoride (F^-). According to the source identification framework, these parameters correspond to and help identify various pollution sources. Nitrate and ammonia are typically attributed to agricultural runoff and livestock practices involving fertilizers and manure. Sulfate, COD, and BOD are commonly found in industrial effluents, including food processing, chemical, and textile industries. TDS, nitrite, and fluoride typically result from urban sewage, particularly in peri-urban areas, while TDS and nitrate also represent surface runoff during rain events. DO indicates the removal of organic pollution and the overall health of the aquatic ecosystem. Data were collected monthly from January 2002 to December 2007.

3. Methodology

3.1 Discrete Wavelet Transform (DWT)

he Wavelet Transform (WT) is a valuable analytical tool for non-stationary time series, such as water quality data, because it preserves both temporal and frequency information, unlike the Fourier Transform, which provides only frequency information (Kumar et al., 2020; Mallat, 1989). The Discrete Wavelet Transform (DWT), widely used in environmental research, decomposes signals into approximation (low-frequency) and detail (high-frequency) coefficients, which describe trends and abrupt changes at various scales (Chen et al., 2016).
The Continuous Wavelet Transform (CWT) is defined as:

$$W(f,a,b) - \int_{-\infty}^{\infty} f(t)\psi*\left(\frac{t-b}{a}\right)dt \qquad (1)$$

Where:

f(t): The signal f(t) requires analysis,
ψ ( ): The mother wavelet is denoted $\psi(t)$,
a: the scale parameter a governs the wavelet's width,
b: The displacement parameter b governs the wavelet's position.
ψ*: A wavelet is the complex conjugate.
The DWT approximates by using high-pass ((g[k])) and low-pass ((h[k])) filters on a discrete signal (x[n]):
Approximation coefficient

$$A_n = \sum_{0}^{k}(h[k]x[n-k]) \qquad (2)$$

Detail coefficient

$$D_n = \sum_{0}^{k}(g[k]x[n-k]) \qquad (3)$$

These coefficients represent low-frequency trends and high-frequency fluctuations, respectively. For this study, five-level DWT decomposition based on the Daubechies-4 wavelet (db4) was used since it has compact support and is efficient at capturing characteristics of environmental signals (Zhang & Liu, 2017; Mallat, 1989). This multi-level decomposition separates short-term spikes of pollution and long-term trends, essential for water quality dynamics understanding and source identification of pollution in the Latyan Dam basin (Kumar et al., 2020).

3.2 Independent Component Analysis (ICA)

Independent Component Analysis (ICA) is a statistical approach that decomposes multivariate signals into statistically independent components and is particularly applicable to the source separation of hidden pollution sources in complex environmental data, such as water quality time series (Hyvärinen et al., 2001; Lee et al., 2007). In contrast to PCA, which produces uncorrelated components, ICA seeks maximum statistical independence, effectively separating latent sources such as agricultural runoff or industrial discharges without prior assumptions.
In this study, ICA is applied to the wavelet coefficients derived from water quality measurements to identify hidden sources of pollution affecting the Latyan Dam. The ICA model is given as:

$$X = AS + N \qquad (4)$$

where:
X is the observed data matrix (e.g., the wavelet coefficients),
A is the mixing matrix that describes how the independent sources combine to form the observed signals,
S is the matrix of independent components (hidden sources),
N is the noise term.

The objective is to estimate (A) and unmix (S) by maximization of non-Gaussianity based on higher-order statistics like kurtosis or negentropy.

The FastICA algorithm (Hyvärinen & Oja, 1997) is used, which encompasses:

Centering: Subtracting the mean to zero-center the data.

Whitening: Transforming data into uncorrelated components with unit variance based on the covariance matrix (C), with the whitening matrix being ($W=C^{-0.5}$)

Maximization of Non-Gaussianity: Iteratively applying a nonlinear function to maximize the independence of components.

Reconstruction: Independent component recovery representing sources of pollution.

By ICA processing on wavelet-decomposed water quality variables like COD, BOD, and TDS, the method unmixes sources like seasonal agricultural runoff or occasional industrial discharges, and extracts information regarding their time-domain input into water quality degradation in the Latyan Dam catchment (Hyvärinen et al., 2001).

3.3 Wavelet-ICA Hybrid

The Wavelet–ICA hybrid method integrates the Discrete Wavelet Transform (DWT) and Independent Component Analysis (ICA) to improve source separation in complex environmental datasets, such as water quality time series. (Hyvärinen & Oja, 2000; Mallat, 1999). DWT decomposes signals into frequency sub-bands, separating temporal patterns such as short-term pollution peaks, seasonal trends, and long-term dynamics, while ICA extracts statistically independent components that reflect latent pollution sources without any a priori assumptions. (Curceac et al., 2021).

The hybrid algorithm was applied to monthly data (2002–2007) from nine monitoring stations in the Latyan Dam basin, considering nine water quality parameters: nitrate (NO_3^-), ammonia (NH_3), sulfate (SO_4^{2-}), chemical oxygen demand (COD), biological oxygen demand (BOD), dissolved oxygen (DO), total dissolved solids (TDS), nitrite (NO_2^-), and fluoride (F^-). The procedure involves:

DWT Decomposition: Time series were decomposed at levels 1, 3, and 5 using the Daubechies-4 (db4) wavelet to retain high-frequency events (e.g., industrial effluent discharges), seasonal patterns (e.g., agricultural runoff), and multi-year trends (e.g., urban sewage), respectively. (Daubechies, 1992; Parmar and Bhardwaj, 2013).

ICA Processing**:** Five independent components were extracted using the FastICA algorithm from each level of detail and approximation coefficients, following centering and whitening via eigenvalue decomposition. (Comon, 1994; Hyvärinen, 1999).

Source Association: Independent components obtained from ICA applied to wavelet coefficients were associated with pollution sources based on parameter loadings in the mixing matrix (e.g., NO_3^- for agriculture, SO_4^{2-} for industry).

Result Synthesis: A weighted summation index was used to combine results across decomposition levels, employing signal-to-noise ratios and correlations to quantify source contributions. The analysis yielded distinct pollution signatures: level 1 highlighted episodic industrial effluent discharges (SO_4^{2-}, COD spikes), level 3 reflected seasonal agricultural runoff (NO_3^-, NH_3 peaks), and level 5 captured long-term urban sewage trends (TDS, NO_2^-). Agricultural dominance was observed at all stations, with no industrial input and effectively zero contributions from urban sewage, runoff, or livestock, as indicated by basin land use. (Jie-xing et al., 2012; Curceac et al., 2021).

3.4 Score Assignment

We used the Wavelet–ICA hybrid method to construct a continuous grading system that quantifies the contribution of each pollution source—including industrial, agricultural, urban wastewater, surface runoff, and livestock—to water quality in the Latyan Dam basin. The system employed the FastICA mixing matrix to identify significant source contributions at wavelet decomposition levels 1 (short-term), 3 (seasonal), and 5 (long-term), based on monthly water quality data from nine monitoring stations collected from January 2002 to December 2007.

The first step in scoring was to examine the absolute loadings of the FastICA mixing matrix for each station and decomposition level. Loadings exceeding 0.5 times the maximum loading per independent component were considered significant, as this threshold provided the best balance of sensitivity and specificity while minimizing false positives caused by noise. (Hyvärinen & Oja, 2000). We found that significant loadings were associated with source-specific water quality characteristics adopted from Karami et al. (2012) and defined in the source identification key (Table 1). The primary parameters associated with each source are as follows: agricultural (NO_3^-, NH_3, TDS, BOD, COD), industrial (BOD, COD, TDS, SO_4^{2-}), urban sewage (BOD, COD, TDS, NO_2^-, F^-), surface runoff (TDS, NO_3^-), and livestock (NO_3^-, BOD, TDS). A Python pipeline analyzed the Wavelet–ICA outputs and assigned each source a score for each parameter that exceeded the threshold in any sub-band (i.e., detail or approximation coefficients). Before summing the scores across ranks per station, they were adjusted for differences in temporal resolution by scaling according to the number of sub-bands at each level. This approach quantified both the frequency and amplitude of source contributions over timeFor example, at Level 3, agriculture was the dominant source (scores: 13–31), which was consistent with the basin's upstream agricultural plains. Industrial contributions were minimal (scores: 0–5), as expected due to limited industrial activity. Urban sewage, surface runoff, and livestock contributions were negligible (scores: 0), consistent with the basin's land use patterns.

The scoring system effectively illustrated how pollution sources varied across time and space. High agricultural scores at higher-flow stations (e.g., station 211: agriculture 31 at Level 3) and low industrial contributions in outer-settlement areas (e.g., station 212: industry 5) demonstrated spatial differentiation. Agricultural dominance increased from Level 1 (score: 6) to Level 5 (score: 24), reflecting long-term patterns. In contrast, industrial signals were more pronounced in short-term (Level 1) evaluations (Curceac et al., 2021). These results, detailed in the data section (Tables 4–5, Figures 3–6), provide a solid foundation for implementing targeted anti-pollution measures in the Latyan Dam basin.

Pollution Source	Associated Parameters
Agriculture	NO_3^-, NH_3, TDS, BOD, COD
Industry	BOD, COD, TDS, SO_4^{2-}
Urban Sewage	BOD, COD, TDS, NO_2^-, F^-
Surface Runoff	TDS, NO_3^-
Livestock	NO_3^-, BOD, TDS

Table 1. Source identification key linking water quality parameters to pollution sources, adapted from Karami et al. (2012).

https://doi.org/10.5194/isprs-annals-X-4-W8-2025-359-2026

3.5 Station Clustering

We grouped nine monitoring stations in the Latyan Dam catchment based on source contribution scores derived from the Wavelet–ICA hybrid approach to identify spatially homogeneous domains. Stations were clustered according to standardized contributions from agricultural, industrial, urban sewage, surface runoff, and livestock sources at different decomposition levels (1, 3, and 5). Prior to clustering, StandardScaler was used to normalize the scores so that sources with different scales had equal weight, as agricultural scores (e.g., 13–31 at Level 3) were much higher than industrial scores (0–5).

The number of clusters was evaluated from 2 to min (nstations−1,10)\min(n_{\text{stations}} - 1, 10)min(nstations−1,10), with the optimal clustering selected using the Silhouette Score, which measures the cohesion within clusters and separation between clusters. At Level 3, three optimal clusters were identified, each exhibiting distinct spatial patterns:

- **Cluster 1 (202, 203, 205, 208, 211, 213):** High agricultural contributions (overall score: 172) and minimal industrial contributions (score: 9).
- **Cluster 2 (207, 212):** Moderate agricultural (score: 41) and industrial contributions (score: 8).
- **Cluster 3 (214):** High agricultural contributions (score: 13) and no industrial contributions.

This clustering approach demonstrated that pollutant profiles were spatially coherent across the basin, with upstream plain agricultural lands dominating and industrial activities being limited. The method was robust to noisy data, even under low signal-to-interference ratios (−3.04 at Level 3), owing to the use of standardized scores and multiscale source information.

4. Results

In this study, the Wavelet–ICA hybrid method was applied to identify sources of water pollution in the Latyan Dam basin using monthly water quality measurements from nine sites over six years (2002–2007). Source contributions derived from Wavelet–ICA at levels 1, 3, and 5 were compared with standard Independent Component Analysis (ICA) to determine dominant pollution sources, assess spatial distribution, and evaluate the performance of each method using key metrics. The findings provide clear insight into pollution patterns and establish a foundation for targeted water quality management.

Agriculture remained the largest contributor to water pollution across all time scales, with its impact increasing from short-term to long-term analyses. As shown in Table 2, agricultural scores rose from 6 at Level 1 (short-term) to 24 at Level 5 (long-term), reflecting its consistent influence. Station-specific scores further emphasized this predominance, ranging from 4 (station 214, ICA only) to 44 (station 211, Level 5), with maximum scores observed at Level 5 for stations 202 (43) and 208 (42). Parameters such as nitrate (NO_3^-), ammonia (NH_3), and total dissolved solids (TDS) were strongly associated with agricultural runoff, confirming its dominant role in water quality degradation.

Low industrial contributions were recorded, with scores ranging from 0 to 6 across levels and stations. For example, station 212 had an industry score of 6 at Level 5, reflecting localized effects from episodic releases of sulfate (SO_4^{2-}), chemical oxygen demand (COD), and biological oxygen demand (BOD). Urban wastewater, surface runoff, and livestock showed no detectable inputs (scores of 0), consistent with the predominance of agriculture and limited urban or livestock activity in the basin. This pattern is evident in Fig. 3, where, at Level 3, agricultural scores (13–31) markedly exceeded industrial scores (0–5), highlighting the basin's strong agricultural influence.

Level	Source	Score
1	Agriculture	6
3	Agriculture	13
5	Agriculture	24

Table 2. Total Agriculture Source Scores Across Decomposition Levels

The ICA Component Matrix, derived from the FastICA mixing matrix, had further information on pollution patterns. For instance, at station 202 (ICA Only), COD (-0.571) and DO (-0.531) were of high weights in IC1, reaffirming the dominant role of agriculture since the parameters are related to nutrient runoff (Table 3).

	IC1	IC2	IC3	IC4	IC5	IC6	IC7	IC8	Station
COD	-0.57113	0.164218	0.042207	0.12234	-0.66726	-0.25994	0.208004	0.26648	202
BOD	0.431935	0.392759	-0.31156	0.351326	-0.3874	0.146769	-0.50009	-0.12778	202
DO	-0.53133	0.523642	-0.13387	-0.21491	-0.24283	0.407274	-0.38755	-0.05837	202
BR	0.332119	-0.10235	-0.13917	0.097417	-0.22425	0.234195	-0.84816	0.160371	202
F	0.035262	0.454204	-0.51055	0.25901	0.25042	-0.4612	-0.09008	0.424631	202
TDS	0.320777	0.042818	-0.13456	0.907499	0.097058	-0.08426	-0.14029	0.129113	202
SO4	-0.11345	0.11948	-0.03586	-0.55611	-0.3582	0.37273	0.269384	0.566282	202
NH3	0.453137	0.092896	0.226551	0.307977	-0.10001	0.165808	-0.77081	-0.08998	202
NO3	-0.03089	-0.15046	0.15748	-0.51302	0.345685	0.697282	0.279411	0.018622	202

Table 3. ICA Component Matrix (ICA method)

The KMeans clustering algorithm, with Silhouette scoring to determine the optimal number of clusters, was employed to group the nine monitoring stations based on their pollution source scores. For this analysis, ICA only and Wavelet–ICA at Levels 1, 3, and 5—representing baseline, short-term, cyclic, and long-term pollution dynamics, respectively—were considered (refer to Table 4).

With ICA only, three clusters were identified: Cluster 1 (stations 203, 205, 207, 208, 211, 213) had the highest score of 47, dominated by agriculture with no industrial contribution. Cluster 2 (stations 202, 212) exhibited a slight industrial contribution (score: 2) alongside agricultural contributions (score: 13). Cluster 3 (station 214) contained the remaining agricultural score (4), indicating a distinct profile.

For Wavelet–ICA Level 1, two clusters were observed. Cluster 2 (stations 202, 203, 205, 207, 211, 212, 213) showed high agricultural contributions (score: 94) and low industrial contributions (score: 8), likely reflecting short-term pollution events. Cluster 1 (stations 208, 214) consisted solely of agricultural contributions (score: 20).

Wavelet-ICA Level 3: we formed three clusters. Cluster 1 (stations 202, 203, 205, 208, 211, 213) had the highest agricultural score (172) and minimal industry (score: 9). Cluster 2 (stations 207, 212) showed moderate contributions (agriculture: 41, industry: 8). Cluster 3 (station 214) was

https://doi.org/10.5194/isprs-annals-X-4-W8-2025-359-2026 |

agriculture only (score: 13) showing seasonal patterns quite effectively.

Three clusters were identified using Wavelet–ICA Level 5. Cluster 1 (stations 202, 203, 205, 207, 208, 211, 213) had the highest agricultural score (283) and an industrial score of 14, reflecting long-term trends. Cluster 2 (station 212) and Cluster 3 (station 214) had lower agricultural scores (32 and 24, respectively). The persistent agriculture-only clustering at station 214 likely reflects its unique, downstream location.

These results illustrate the pervasive dominance of agriculture at these scales, while industrial impacts are limited to specific stations, such as 202 and 212. Figures 3–6 show the geographic arrangement of stations with their cluster labels, demonstrating the consistency of the results.

The increase in agricultural scores from Level 1 (6) to Level 5 (24) reflects the method's enhanced ability to capture long-term pollution trends at higher decomposition levels. At the station level, agricultural scores ranged from 4 (station 214, ICA only) to 44 (station 211, Level 5), with Level 5 records showing the highest scores (e.g., station 202: 43; station 208: 42). Industrial scores were low, ranging from 0 to 6 (e.g., station 212: 6 at Level 5), indicating localized industrial effects. Figure 3 presents station-specific source scores at Level 3, where agricultural scores (13–31) greatly exceed industrial scores (0–5), reflecting the basin's strong agricultural influence.

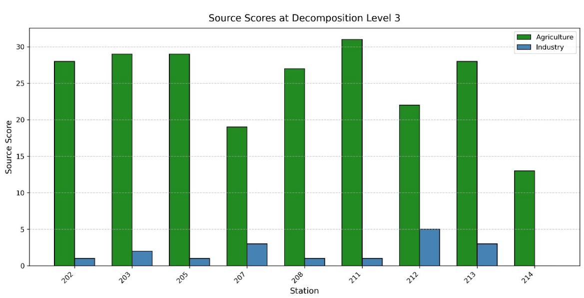

Figure 3. Bar chart displaying agriculture (scores: 13–31) and industry (scores: 0–5) contributions for each station at decomposition level 3

	IC1	IC2	IC3	IC4	IC5	IC6	IC7	IC8	Station
COD	-0.57	0.16	0.04	0.12	-0.67	-0.26	0.21	0.27	202
BOD	0.43	0.39	-0.31	0.35	-0.39	0.15	-0.50	-0.13	202
DO	-0.53	0.52	-0.13	-0.21	-0.24	0.41	-0.39	-0.06	202
BR	0.33	-0.10	-0.14	0.10	-0.22	0.23	-0.85	0.16	202
F	0.04	0.45	-0.51	0.26	0.25	-0.46	-0.09	0.42	202
TDS	0.32	0.04	-0.13	0.91	0.10	-0.08	-0.14	0.13	202
SO4	-0.11	0.12	-0.04	-0.56	-0.36	0.37	0.27	0.57	202
NH3	0.45	0.09	0.23	0.31	-0.10	0.17	-0.77	-0.09	202
NO3	-0.03	-0.15	0.16	-0.51	0.35	0.70	0.28	0.02	202

Table 3. ICA Component Matrix (ICA method)

To identify the spatial patterns of pollution in the Latyan Dam basin, the nine monitoring stations were grouped based on their source pollution scores using the KMeans algorithm, with the optimal number of clusters determined via the Silhouette Score. Clustering was performed for four methods: ICA alone and Wavelet–ICA at decomposition levels 1, 3, and 5, representing baseline, short-term, seasonal, and long-term pollution scenarios, respectively. As shown in Table 4, the results reveal clear spatial clusters in which agricultural inputs dominate, with minimal industrial contributions.Wavelet-ICA level 1 results: we have two clusters to work with. Cluster 2 (stations 202, 203, 205, 207, 211, 212, 213) had high marks for agricultural contributions (score: 94) and low for industrial credits (score: 8) most likely due to some short-term pollution events. Cluster 1 (stations 208, 214) was agricultural only (score: 20).

Wavelet-ICA Level 3: we formed three clusters. Cluster 1 (stations 202, 203, 205, 208, 211, 213) had the highest agricultural score (172) and minimal industry (score: 9). Cluster 2 (stations 207, 212) showed moderate contributions (agriculture: 41, industry: 8). Cluster 3 (station 214) was agriculture only (score: 13) showing seasonal patterns quite effectively.

Three clusters were identified using Wavelet-ICA level 5. Cluster 1 (stations 202, 203, 205, 207, 208, 211, 213) had the highest agricultural score (283) and industry (14), reflecting long-term trends. Cluster 2 (station 212) and Cluster 3 (station 214) had lower agricultural scores (32 and 24 respectively). The reason for the consistent agriculture only clustering in 214 is most likely its unique, and likely downstream, site.

The results illustrate the pervasive dominances of agriculture at these scales while the industrial impacts are limited to some specific stations 202 and 212. We show in Figure 3-6 where the geographic arrangement of stations and their cluster labels are illustrated showing the uniformity of results.

Method	Cluster	Stations	Dominant Sources (Aggregate Scores)
ICA Only	1	203, 205, 207, 208, 211, 213	Agriculture (47), Industry (0)
ICA Only	2	202, 212	Agriculture (13), Industry (2)
ICA Only	3	214	Agriculture (4), Industry (0)
Wavelet-ICA Level 1	1	208, 214	Agriculture (20), Industry (0)
Wavelet-ICA Level 1	2	202, 203, 205, 207, 211, 212, 213	Agriculture (94), Industry (8)
Wavelet-ICA Level 3	1	202, 203, 205, 208, 211, 213	Agriculture (172), Industry (9)
Wavelet-ICA Level 3	2	207, 212	Agriculture (41), Industry (8)
Wavelet-ICA Level 3	3	214	Agriculture (13), Industry (0)
Wavelet-ICA Level 5	1	202, 203, 205, 207, 208, 211, 213	Agriculture (283), Industry (14)
Wavelet-ICA Level 5	2	212	Agriculture (32), Industry (6)
Wavelet-ICA Level 5	3	214	Agriculture (24), Industry (0)

Table 4. Station Clustering Across Methods

https://doi.org/10.5194/isprs-annals-X-4-W8-2025-359-2026

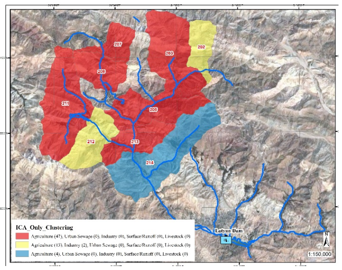

Figure 3. Spatial Clustering of Monitoring Stations Based on Pollution Source Contributions Using ICA Only in the Latyan Dam Basin

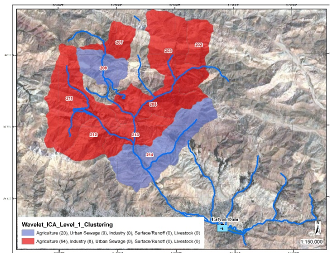

Figure 4. Spatial Clustering of Monitoring Stations Based on Pollution Source Contributions Using Wavelet-ICA at Decomposition Level 1 in the Latyan Dam Basin

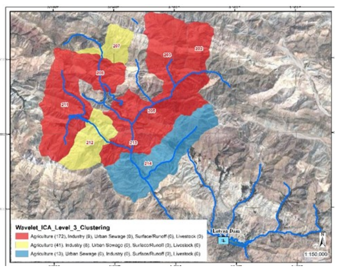

Figure 5. Spatial Clustering of Monitoring Stations Based on Pollution Source Contributions Using Wavelet-ICA at Decomposition Level 3 in the Latyan Dam Basin

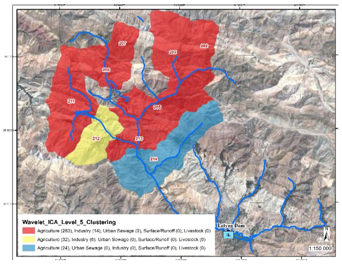

Figure 6. Spatial Clustering of Monitoring Stations Based on Pollution Source Contributions Using Wavelet-ICA at Decomposition Level 5 in the Latyan Dam Basin

Evaluation of the Wavelet-ICA hybrid method was done by measuring SIR, Correlation, Mutual Information, and other parameters (Table 5). These parameters check whether the method in question is capable of source separation, estimation of source-to-reference relations, and dependency estimation between the given sources and the estimated sources.

Method	Total Source Scores	Normalized Source Scores	Average SIR	Average Correlation	Average Mutual Information
ICA Only	66.00	7.33	-4.16	0.30	0.84
Wavelet + ICA (Level 1)	6.00	0.67	-2.37	0.27	0.79
Wavelet + ICA (Level 3)	13.00	1.44	-3.04	0.27	0.84
Wavelet + ICA (Level 5)	24.00	2.67	-3.31	0.37	0.81

Table 5. Performance Metrics Across Methods

The SIR values across all methods (e.g. -3.04 at Level 3) show pollution signal separation difficulties because of dirty environmental data. Still, Wavelet-ICA was better than ICA Only, with Level 1 achieving the least negative SIR of -2.37, indicating greater noise reduction for short-term events. Level 3 equaled ICA Only's Mutual Information of 0.84 and performed well at capturing seasonal patterns, while Level 5 had the highest Correlation of 0.37 showing good fit for long-term trends. These results are further illustrated in Figure 7, that depicts the trade-offs of these levels and shows those of noise reduction, attenuation, and strength dependency, as done by Level 1 and Level 3 respectively.

5. Discussion

The Latyan Dam basin was used in this study to examine the impacts of water pollution on agriculture. Agriculture was found to be the primary contributor to water quality degradation across short-term, seasonal, and long-term periods. Upstream sources (Figure 1) consistently exhibited high pollutant levels, which gradually decreased at downstream stations. Peaks in nitrate (NO_3^-) and ammonia (NH_3) concentrations were also observed. These findings highlight the potential of agricultural best practices—such as precision farming, cover crops, and buffer zones—to reduce nutrient runoff and help maintain a vital water supply that provides Tehran with 25% of its drinking water.

Small industrial contributions were observed at stations 202 and 212, indicating localized hotspots with periodic increases in sulfate (SO_4^{2-}) and Chemical Oxygen Demand (COD). These areas and similar sources require closer attention and stricter management measures. The absence of significant effects from urban sewage, surface runoff, and livestock further highlights the distinct agricultural character of the basin, with peri-urban development being the secondary contributor after farming. This reinforces the conclusion that agricultural pollution is the primary concern.

Grouping the nine monitoring stations revealed distinct spatial patterns across the basin. Station 214 consistently exhibited an agriculture-only profile, likely due to its downstream location.

This emphasizes the basin's unique environmental characteristics and the need for targeted management strategies. These insights can inform the strategic placement of monitoring stations and the design of region-specific interventions to enhance water quality management.

The Wavelet-ICA hybrid method outperformed standard ICA, particularly in denoising at Level 1 (SIR: -2.37) and identifying seasonal patterns at Level 3 (Mutual Information: 0.84). However, consistently low SIR values across all techniques indicate the challenges of working with noisy environmental data. Future studies should explore more advanced denoising approaches or incorporate additional variables, such as heavy metals or biological indicators, to improve the accuracy of source separation.

This study advances environmental science by providing a robust framework for identifying pollution sources, which can be applied to other water bodies facing similar issues. Accurately detecting pollution sources across different temporal scales is essential for managing complex environmental systems. The findings support data-driven strategies to safeguard critical water resources like the Latyan Dam by providing actionable insights. They also highlight the need for further methodological improvements to handle noisy data and enhance applicability in diverse contexts.

6. Conclusion

This study applied a novel Wavelet-ICA hybrid method to identify the sources of water pollution in the Latyan Dam basin and quantify their contributions over time and space. By integrating wavelet decomposition with Independent Component Analysis (ICA), the method successfully disentangled complex pollution signals across short-term (Level

1), seasonal (Level 3), and long-term (Level 5) timescales. The analysis revealed that agriculture was the dominant source of pollution, with scores ranging from 4 to 44 and reaching up to 283 in clustering results. In contrast, industry contributed minimally, with scores ranging from 0 to 6 (up to 14 in clusters). Urban runoff, wastewater, and livestock showed negligible impacts, reflecting the basin's primarily agricultural character with limited industrial influence. The nine monitoring stations were grouped using KMeans and validated with Silhouette Scores, revealing distinct pollution patterns across the basin. Using only Independent Component Analysis (ICA), three clusters were identified: an agriculture-dominated cluster (Cluster 1: stations 203, 205, 207, 208, 211, 213; score: 47) and a cluster with minor industrial inputs (Cluster 2: stations 202, 212; industry score: 2). Wavelet-ICA at Level 1 identified two clusters reflecting short-term variations (Cluster 2: agriculture 94; industry 8). At Level 3, three clusters emerged, highlighting seasonal patterns (Cluster 1: agriculture 172; industry 9). Level 5 clustering, capturing long-term trends, revealed a dominant agriculture cluster (Cluster 1: agriculture 283; industry 14). Station 214 consistently formed an agriculture-only cluster in most analyses, likely due to its unique downstream location. These results provide critical insights for targeted pollution mitigation, such as reducing agricultural runoff in upstream areas corresponding to Cluster 1 stations.

Performance indicators showed what was good and bad about the method. Level 1 Wavelet-ICA had the worst signal-to-interference ratio (SIR = -2.37), which means it got rid of the most noise. Level 3 had the most Mutual Information (0.84), which matched the seasonal pollution patterns that are essential to the basin. However, keeping a negative SIR (like -3.04 at Level 3) shows that noisy environment observations aren't being used correctly, which means that better methods are needed. The Wavelet-ICA hybrid has done better than ICA alone (SIR = -4.16), showing that it is better at separating sources at different scales.

The study helps to analyze water quality by giving a way to find pollution sources that may be conveniently used for sustainable water resource management in the Latyan Dam basin. More work might make advanced noise reduction methods better to get the best SIR, add more factors like heavy metals, and change the findings of field observation measurements so that they can be used easily in real life. These efforts would make it easier to use the technology to help solve problems with worldwide water contamination.

References

Bhatt, D and Swain, M., 2025: Artificial intelligence based detection and control strategies for river water pollution: A comprehensive review. *Journal of Contaminant Hydrology*, 271, 104541. https://doi.org/10.1016/j.jconhyd.2025.104541.

Curceac, S., Milne, A., Atkinson, P.M., Wu, L., Harris, P., 2021: Elucidating the performance of hybrid models for predicting extreme water flow events through variography and wavelet analyses. *Journal of Hydrology*, 598, 126442. https://doi.org/10.1016/j.jhydrol.2021.126442

Etemadi, H., Smoak, J.M., Karami, J., 2018: Land use change assessment in coastal mangrove forests of Iran utilizing satellite imagery and CA–Markov algorithms to monitor and predict future change. *Environmental Earth Sciences*, 77, 208. doi.org/10.1007/s12665-018-7392-8.

Huan, J., Fan, Y., Xu, X., Zhou, L., Zhang, H., Zhang, C., Hu, Q., Cai, W., Ju, H., Gu, S., 2025: Deep learning model based on coupled SWAT and interpretable methods for water quality prediction under the influence of non-point source pollution. *Computers and Electronics in Agriculture*, 231, 109985. doi.org/10.1016/j.compag.2025.109985.

Zhou, J., Wang, J., Wang, P., Yang, H., 2012: Wavelet analysis of water quality changes in Dianchi Lake during the past 7 years. *Procedia Earth and Planetary Science*, 5, 280–288. doi.org/10.1016/j.proeps.2012.01.048.

Karami, J., Alimohammadi, A., Modabberi, S., 2012: Analysis of the spatio-temporal patterns of water pollution and source contribution using the MODIS sensor products and multivariate statistical techniques. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(4), 1243–1255. doi.org/10.1109/JSTARS.2012.2187273.

Karami, J., Alimohammadi, A., Seifouri, T., 2014: Water quality analysis using a variable consistency dominance-based rough set approach. C*omputers, Environment and Urban Systems*, 43, 25–33. doi.org/10.1016/j.compenvurbsys.2013.09.005

Karami, J., Babaee, F., Mahmoudnia, P., Sharifi Kia, M., 2025: Integrated spatiotemporal data mining and DInSAR for improved understanding of subsidence related to groundwater depletion impacts. *Journal of Geographical Sciences*, 35(3), 598–618. doi.org/10.1007/s11442-025-2336-1.

Nagaraju, T.V., Sunil, B.M., Chaudhary, B., Prasad, C.D., Gobinath, R., 2023: Prediction of ammonia contaminants in the aquaculture ponds using soft computing coupled with wavelet analysis. *Environmental Pollution*, 331(Pt 2), 121924. doi.org/10.1016/j.envpol.2023.121924.

Nourani, V and Farboudfam, N., 2019: Rainfall time series disaggregation in mountainous regions using hybrid wavelet-artificial intelligence methods. *Environmental Research*, 168, 306–318. doi.org/10.1016/j.envres.2018.10.012.

Parmar, K.S and Bhardwaj, R., 2013: Wavelet and statistical analysis of river water quality parameters. *Applied Mathematics and Computation*, 219(20), 10172–10182. doi.org/10.1016/j.amc.2013.03.109.

Sankaran, A., Plocoste, T., Rajakrishnan, A., Thodika, R.C., Rema, R.A., Dharman, S., Madhukumar, M., 2024: Multiscale coherence analysis of PM2.5 and PM10 concentrations of four Indian urban areas using wavelet transform. *Urban Climate*, 56, 102056. doi.org/10.1016/j.uclim.2024.102056.

Song, C., Yao, L., Hua, C., Ni, Q., 2021: A novel hybrid model for water quality prediction based on synchrosqueezed wavelet transform technique and improved long short-term memory. *Journal of Hydrology*, 603, 126879. doi.org/10.1016/j.jhydrol.2021.126879.

Wang, C., Xiong, W., Zhang, G., 2025: Application of deep learning models with spectral data augmentation and denoising for predicting total phosphorus concentration in water pollution. *Journal of the Taiwan Institute of Chemical Engineers*, 167(1), 105852. doi.org/10.1016/j.jtice.2024.105852

Wang, L., Huang, Y., Ren, J., Zhou, H., Chen, L., Zhou, H., 2024: Double-transfer learning-based least squares wavelet support vector machine for water quality classification. *Desalination and Water Treatment*, 320, 100714. doi.org/10.1016/j.dwt.2024.100714.

Yu, B., Huang, C.-M., Liu, Z., Zhang, B., Qin, F., 2010: Wavelet analysis on NH3-N pollution index changes of the Middle-Upper Minjiang River during last 6 years. *Procedia Environmental Sciences*, 2, 9–14. doi.org/10.1016/j.proenv.2010.10.003.