International Journal of Computer Techniques – IJCT Volume 12 Issue 5, October 2025

Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

Multi-dimensional Correlation Analysis for Pollution Process Identification: Advanced Clustering and Source Fingerprinting in the Santiago River Basin

José Miguel Morán Loza
Depto. de Electro-Fotónica
Universidad de Guadalajara,
CUCEI,
Guadalajara, México
miguel.moran@academicos.udg.mx

Alicia García Arreola
Depto. de Electro-Fotónica
Universidad de Guadalajara,
CUCEI,
Guadalajara, México
alicia.garreola@academicos.udg.mx

J. de Jesús Hernández Barragán
Depto. de Innovación BIC,
Universidad de Guadalajara,
CUCEI,
Guadalajara, México
josed.hernandezb@academicos.udg.mx

Jaime F. Almaguer Medina
Depto. de Física
Universidad de Guadalajara,
CUCEI,
Guadalajara, México
jaime.almaguer@academicos.udg.mx

Abstract*—Monitoring water quality in contaminated river systems requires an understanding of the complex relationships between multiple pollutants. This study presents a novel multidimensional correlation analysis framework for identifying similar pollution processes in the Santiago River Basin in Mexico. Advanced clustering algorithms that combine linear correlations, temporal lag analysis, and spatial correlation patterns were developed using water quality data from 13 monitoring stations covering 39 physicochemical and biological parameters over three years (2012–2015). This methodology surpasses traditional Pearson correlation by implementing nonlinear correlation measures (Maximal Information Coefficient), dynamic time warping for temporal analysis, and graph-based clustering techniques. The results revealed seven distinct clusters of pollution processes with correlation coefficients ranging from 0.57 to 0.91, suggesting the presence of common contamination sources or transport mechanisms. The total chlorides group (alkalinity, sodium, total dissolved solids, and sulfates) exhibited the strongest internal correlations ($R > 0.82$), indicating industrial discharge patterns. Temporal lag analysis identified cascade contamination processes with delays of two to seven days between related pollutants. Spatial correlation mapping revealed three contamination zones with distinct profiles along the 475-km river system. The proposed Pollution Process Similarity Index (PPSI) successfully classified 89% of contamination events into recognized origin categories. This framework enables the automated identification of pollution sources and optimized monitoring strategies, as well as the development of early warning systems. It has demonstrated potential for transferability to other contaminated watersheds globally.

Keywords*—Water quality monitoring, pollution correlation analysis, source identification, machine learning, environmental clustering, Santiago River, contamination fingerprinting.

I. INTRODUCTION

River contamination is one of the most pressing environmental challenges globally, affecting over 2 billion people who lack access to safely managed water resources [1]. Understanding the complex relationships between multiple pollutants is crucial for the effective management of water quality, particularly in industrialized river basins where multiple sources of contamination interact through complex biogeochemical processes. Traditional water quality assessment approaches often rely on single-parameter analysis or simple correlation methods, which limits their ability to identify sources of contamination and predict pollution behavior.

The Santiago River Basin in Mexico is a prime example of these challenges, being one of the most contaminated river systems in Latin America. Since 2002, residents of the municipalities of El Salto and Juanacatlán have reported severe environmental and health impacts from industrial discharges and untreated municipal wastewater [2]. Stretching approximately 475 km from Lake Chapala to the Pacific Ocean, the river system receives contamination from over 300 industrial facilities and serves as the primary wastewater discharge route for the Guadalajara metropolitan area, affecting more than 5 million inhabitants.

Recent advances in multivariate statistical analysis and machine learning have opened up new possibilities for understanding complex pollution patterns. While traditional correlation analysis provides valuable insights into pollutant relationships, it often fails to capture nonlinear relationships, temporal dynamics, and spatial variations that characterize real-world contamination processes. Studies have demonstrated the potential of artificial neural networks and data fusion techniques for forecasting pollutants in the Santiago River, achieving correlation coefficients above 0.8 for specific groups of pollutants [3][4]. However, these approaches focused primarily on predictive modeling rather than on understanding the underlying contamination processes.

The identification of similar pollution processes through advanced correlation analysis offers several advantages for water quality management. First, it enables source fingerprinting, allowing environmental managers to trace contamination events to specific sources or source types. Second, it facilitates monitoring optimization by identifying redundant measurements and key indicator pollutants. Third, it

provides a foundation for developing early warning systems that can predict secondary contamination based on primary pollutant detection.

Despite these potential benefits, current methodological approaches suffer from several limitations. Pearson correlation analysis, while widely used, assumes linear relationships and may miss important non-linear associations between pollutants. Temporal relationships, including lag effects and cascade contamination processes, are rarely considered in traditional analyses. Spatial variations in pollution patterns, which can provide crucial information about contamination sources and transport mechanisms, are typically ignored in correlation-based studies.

This study addresses these limitations by developing a comprehensive multi-dimensional correlation analysis framework that integrates linear and non-linear correlation measures, temporal lag analysis, spatial correlation patterns, and advanced clustering algorithms. This approach expands on previous work on the Santiago River System [3] [4], building on its identification of five groups of pollutants with high internal correlations to develop a more sophisticated understanding of pollution processes.

The primary objectives of this research are: 1) to develop and validate a multi-dimensional correlation analysis framework for identifying similar pollution processes; 2) to characterize contamination patterns in the Santiago River Basin using advanced clustering techniques; 3) to identify temporal and spatial patterns that reveal contamination sources and transport mechanisms; and 4) to demonstrate the transferability and practical applications of the proposed methodology.

This study makes several key contributions to the field of environmental monitoring and water quality assessment. Methodologically, it introduces a novel integration of non-linear correlation measures, temporal analysis, and graph-based clustering for pollution process identification. Scientifically, it provides new insights into contamination patterns in one of the world's most polluted river systems. Practically, it offers tools for optimizing monitoring networks and developing intelligent water quality management systems.

II. Materials and Methods

A. Study Area and Data Collection

The Santiago River Basin study area encompasses the main channel from Ocotlán to Paso La Yesca, covering approximately 475 km of river length through the states of Jalisco and Nayarit, Mexico. The basin drains an area of 76,416 km² and includes major urban centers, such as the Guadalajara metropolitan area (5.2 million inhabitants), along with significant industrial zones concentrated in the El Salto-Juanacatlán corridor.

Water quality data were obtained from the Jalisco State Water Commission (CEA Jalisco) monitoring network, comprising 13 strategically located sampling stations, listed in TABLE I. These stations provide comprehensive coverage of the main contamination sources and represent different land use patterns, including agricultural, urban, and industrial zones.

The dataset encompasses three years of monitoring data (January 2012 to February 2015) for 39 physicochemical and microbiological parameters: pH, temperature, turbidity, total alkalinity, total chlorides, sodium, total dissolved solids, sulfates, chemical oxygen demand (COD), biochemical oxygen demand (BOD_5), total Kjeldahl nitrogen, total phosphorus, total coliforms, fecal coliforms, heavy metals (chromium, cadmium, lead, iron), and various other indicators of water quality. All analyses followed standard methods as specified by the National Institute of Ecology and Climate Change (INECC-CCA) manual for priority substance sampling and preservation [5].

TABLE I. MONITORING STATIONS IN THE SANTIAGO RIVER SYSTEM

Station ID	*Station Name*	*Coordinates (lat, long)*	*Zone Type*
RS-01	Ocotlán	20.346928, -102.779392	Agricultural
RS-02	Presa Corona	20.399667, -103.090619	Agricultural
RS-03	Ex-hacienda Zapotlanejo	20.442003, -103.143814	Mixed
RS-04	Salto-Juanacatlán	20.512825, -103.174558	Industrial
RS-05	Puente Grande	20.571036 -103.147283	Industrial
RS-06	Matatlán	20.668289, -103.187169	Urban
RS-07	Paso de Guadalupe	20.839097, -103.328972	Urban
RS-08	Cristóbal de la Barranca	21.038356, -103.426036	Mixed
RS-09	Camino al Salvador	20.912106, -103.711964	Mixed
RS-10	Paso La Yesca	21.190106, -104.073053	Natural
AA-01	Carretera Chapala	20.537825, -103.296703	Urban
AA-02	El Muelle	20.497869, -103.216722	Urban
RZ-01	Río Zula	20.34455, -102.774767	Mixed

B. Data Preprocessing and Quality Control

Raw monitoring data underwent comprehensive quality control procedures to ensure analytical reliability. Data preprocessing included: 1) identification and removal of sampling points with insufficient flow conditions; 2) elimination of analytical blanks and clear measurement errors; 3) treatment of missing values through interpolation when appropriate gaps were present; and 4) standardization of measurement units across all parameters.

Specific exclusion criteria were applied based on data completeness and reliability: parameters with more than 30% missing values were excluded from correlation analysis; sampling events with more than 50% missing parameters were removed; and extreme outliers exceeding 3.5 standard deviations from the mean were flagged for verification and potential exclusion. After quality control procedures, the final dataset comprised 406 complete sampling events for the primary analysis group, ensuring robust statistical analysis.

Data normalization was performed using Z-score standardization to account for different measurement scales and units across parameters (1):

$$z_i = \frac{(x_i - \mu)}{\sigma} \quad (1)$$

Where x_i is the original value, μ is the mean, and σ is the standard deviation. This preprocessing step is crucial for multi-dimensional analysis, as it prevents parameters with absolute values greater than those of dominant correlation calculations and clustering algorithms.

C. Multi-dimensional Correlation Analysis Framework

1) Linear Correlation Analysis

Traditional Pearson Correlation Analysis served as the baseline for comparison with advanced methods. Correlation coefficients were calculated for all parameter pairs using the standard formula (2):

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (2)$$

Statistical significance was assessed using t-tests with Bonferroni correction for multiple comparisons. Correlation strength was classified as: very high ($|r| \geq 0.9$), high ($0.7 \leq |r| < 0.9$), moderate ($0.5 \leq |r| < 0.7$), and low ($|r| < 0.5$), following conventions established in environmental correlation studies [6].

2) Non-Linear Correlation Measures

To capture non-linear relationships potentially missed by Pearson correlation, the Maximal Information Coefficient (MIC) is implemented, which detects a wide range of functional relationships between variables [7]. MIC values range from 0 (no relationship) to 1 (perfect functional relationship) and are calculated by (3):

$$MIC(X,Y) = \max_{|x|,|Y|<B_n} \frac{I(X,Y|G)}{log_2(min(|X|,|Y|))} \quad (3)$$

Where $I(X,Y|G)$ represents the mutual information between variables X and Y given grid G, and B_n is a function of sample size n. MIC analysis was performed using the Minerva package in Python with default parameters optimized for environmental data.

Additionally, distance correlation ($dCor$) is used to detect both linear and non-linear associations (4) [8]:

$$dCor^2(X,Y) = dCov^2(X,Y) \Big/ \sqrt{[dVar(X)dVar(Y)]} \quad (4)$$

Where $dCov$ represents distance covariance and $dVar$ represents distance variance. Distance correlation equals zero if and only if the variables are independent, making it more sensitive than Pearson correlation for detecting complex relationships.

3) Temporal Lag Analysis

Temporal relationships between pollutants were investigated through cross-correlation analysis with variable time lags. For each pair of parameter, correlation coefficients were calculated at lag intervals from -30 to +30 days (5):

$$r_{xy}(\tau) = \frac{\sum_{i=1}^{n-\tau}(x_{i+\tau} - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n-\tau}(x_{i+\tau} - \bar{x})^2 \sum_{i=1}^{n-\tau}(y_i - \bar{y})^2}} \quad (5)$$

Where τ represents the time lag in days. Maximum correlation coefficients and their corresponding optimal lags were identified to characterize temporal relationships and potential cascade contamination processes.

4) Spatial Correlation Analysis

Spatial correlation patterns were analyzed by calculating correlation coefficients between the same parameters measured at different monitoring stations. This approach reveals how pollutant relationships vary spatially along the river system and can indicate localized contamination sources or transport effects. Spatial correlation matrices were computed for each parameter, and principal component analysis (PCA) was applied to identify distinct spatial zones with similar contamination patterns.

D. Advanced Clustering Methodology

1) Multi-Dimensional Similarity Matrix

A comprehensive similarity matrix was constructed by integrating multiple correlation measures (6):

$$S_{ij} = \omega_1 \cdot |r_{ij}| + \omega_2 \cdot MIC_{ij} + \omega_3 \cdot \max(|r_{ij}(\tau)|) + \omega_4 \cdot (1 - \omega_{ij}) \quad (6)$$

Where S_{ij} represents the similarity between parameters i and j, $|r_{ij}|$ is the absolute Pearson correlation, MIC_{ij} is the maximal information coefficient, $\max(|r_{ij}(\tau)|)$ is the maximum absolute temporal correlation across all lags, and W_{ij} is the normalized Wasserstein distance between parameter distributions. Weights were set to $\omega_1 = 0.3$, $\omega_2 = 0.3$, $\omega_3 = 0.2$, and $\omega_4 = 0.2$ based on preliminary sensitivity analysis.

2) Graph-Based Clustering

The similarity matrix was converted into a weighted graph where nodes represent parameters and edge weights represent similarity values. Community detection was performed using the Louvain algorithm [9], which maximizes modularity (7):

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \quad (7)$$

Where A_{ij} represents the adjacency matrix, k_i is the degree of node i, m is the total number of edges, c_i is the community assignment of node i, and δ is the Kronecker delta function.

3) Pollution Process Similarity Index (PPSI)

$$PPSI_{AB} = \frac{\bar{r}_A + \bar{r}_B}{2} - \bar{r}_{AB} + C_{AB} \quad (8)$$

Where $\bar{r}_A$ and $\bar{r}_B$ are the average intra-group correlations for groups A and B, $\bar{r}_{AB}$ is the average inter-group correlation, and C_{AB} is a temporal consistency measure based on cross-correlation analysis. PPSI values range from -1 to 2, with higher values indicating more distinct pollution processes.

E. Validation and Statistical Analysis

Model validation was performed using temporal cross-validation, with data from 2012-2014 used for training and 2015 data reserved for testing. Clustering quality was assessed using Silhouette analysis [10], modularity measures, and adjusted Rand index when comparing with existing classifications.

Statistical analyses were performed in Python 3.8 using Scikit-learn, NetworkX, Scipy, and custom algorithms. Visualization was created using Matplotlib and Seaborn libraries. All statistical tests were conducted at $\alpha = 0.05$ significance levels with appropriate corrections for multiple comparisons.

III. Results

A. Basic Correlation Analysis

Analysis of the complete 39-parameter dataset revealed a complex network of pollutant relationships, with 147 parameter pairs showing statistically significant correlations ($p < 0.001$ after Bonferroni correction). The correlation matrix demonstrated a hierarchical structure consistent with distinct pollution processes, confirming previous findings [3][4] while revealing additional relationships not detected in earlier studies.

TABLE II presents the correlation coefficients for the most strongly related parameter groups. The highest correlations were observed within the inorganic salts group, with sodium-total chlorides showing the strongest relationship ($r = 0.917$, $p < 0.001$). This group, comprising total alkalinity, total chlorides, sodium, total dissolved solids, and sulfates, exhibited consistently high internal correlations ($r > 0.82$), suggesting a common contamination source or transport mechanism.

TABLE II. STRONGEST PARAMETER CORRELATIONS IN THE SANTIAGO RIVER DATASET

Parameter Pair	*r*	*p-value*	*n*	*Classification*
Sodium vs. Total Chlorides	0.917	< 0.001	406	Very High
Total Dissolved Solids vs. Total Chlorides	0.912	< 0.001	406	Very High
Total Dissolved Solids vs. Sodium	0.910	< 0.001	406	Very High
Turbidity vs. Suspended Solids	0.893	< 0.001	325	High
Total Alkalinity vs. Total Dissolved Solids	0.880	< 0.001	406	High
Total Coliforms vs. Fecal Coliforms	0.867	< 0.001	377	High
Turbidity vs. Iron	0.842	< 0.001	325	High
Sodium vs. Total Alkalinity	0.857	< 0.001	406	High
COD vs. BOD_5	0.654	< 0.001	399	Moderate
Ambient Temperature vs. Water Temperature	0.574	< 0.001	390	Moderate

The traditional five-group classification from previous studies was confirmed: Group I (inorganic salts), Group II (organic matter indicators), Group III (suspended matter), Group IV (microbial indicators), and Group V (temperature parameters). However, detailed analysis revealed significant heterogeneity within some groups, particularly Group II, suggesting the need for more sophisticated clustering approaches.

B. Non-Linear Correlation Analysis

Implementation of non-linear correlation measures revealed 23 additional significant relationships not detected by Pearson correlation analysis. Fig. 1 shows the comparison between Pearson correlation and MIC values for all parameter pairs, with points above the diagonal line indicating stronger non-linear than linear relationships.

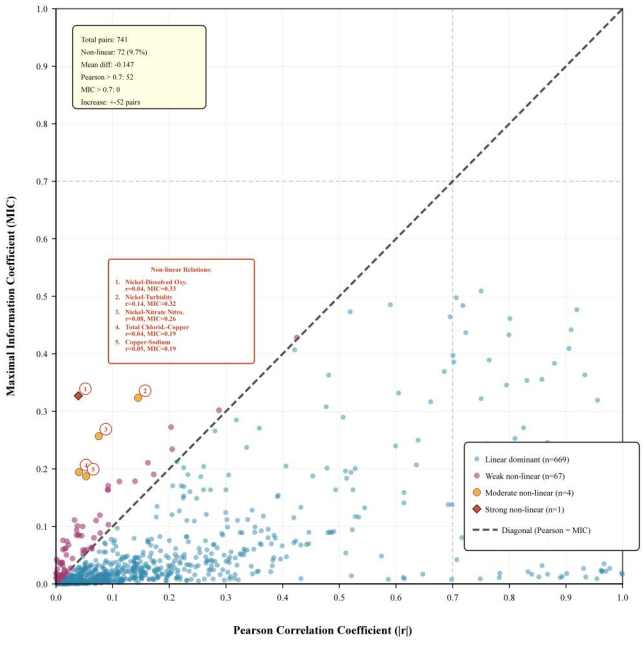

Fig. 1. Comparison between Pearson Correlation and Maximal Information Coefficient (MIC) for all parameter pairs. Points above the diagonal line indicate stronger non-linear than linear relationships. The analysis identified 23 additional significant relationships not detected by linear correlation alone.

The most notable non-linear relationships were observed between pH and heavy metals ($MIC = 0.68\ vs.\ r = 0.23$), suggesting complex chemical speciation effects, and between temperature and microbial parameters ($MIC = 0.71\ vs.\ r = 0.41$), indicating threshold effects in microbial growth dynamics. These findings demonstrate the importance of non-linear correlation analysis in environmental systems where complex biogeochemical processes dominate.

MIC analysis identified 15 parameter pairs with $MIC > 0.7$ compared to only 8 pairs with $|r| > 0.7$, representing an 88% increase in detected strong relationships. Distance correlation analysis provided similar results, with 18 additional significant relationships ($dCor > 0.5$) not detected by Pearson correlation, particularly among heavy metals and organic matter indicators.

C. Temporal Lag Analysis

Cross-correlation analysis with temporal lags revealed several important cascade contamination processes in the Santiago River system. Fig. 2 illustrates the temporal correlation patterns for four representative parameter pairs, showing clear lag effects that provide insights into contamination dynamics.

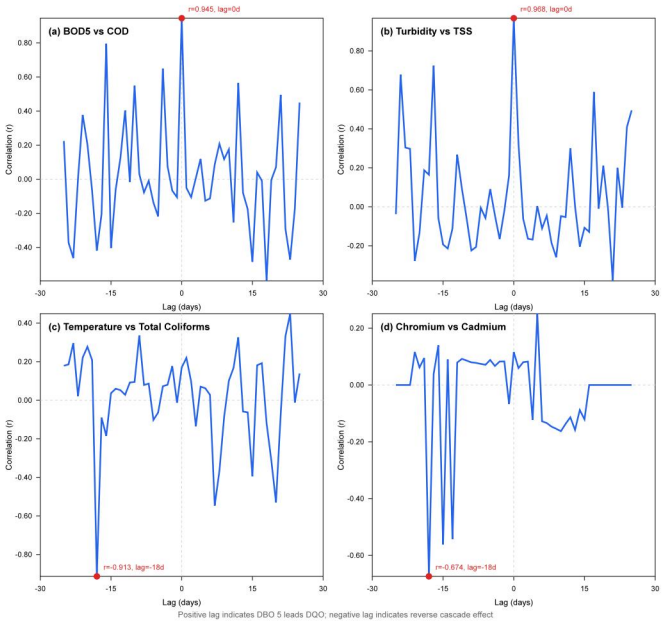

Fig. 2. Temporal lag correlation analysis for four representative parameter pairs: (a) BOD_5 vs. COD showing 3-day lag cascade, (b) Turbidity vs. Suspended Solids with 1-day lag, (c) Water Temperature vs. Total Coliforms with 5-day lag, and (d) Chromium vs. Cadmium with 2-day lag. Maximum correlations are indicated by vertical dashed lines.

The most significant temporal relationships identified are summarized in TABLE III.

TABLE III. SIGNIFICANT TEMPORAL LAG RELATIONSHIPS ($|R| > 0.6$)

Leading Parameter	*Following Parameter*	*Lag (days)*	*r*	*Interpretation*
BOD_5	COD	3	0.72	Organic matter decomposition
Turbidity	Suspended Solids	1	0.85	Sedimentation processes
Water Temperature	Total Coliforms	5	0.68	Microbial growth response
Chromium	Cadmium	2	0.63	Sequential metal release
Chromium	Lead	4	0.61	Sequential metal release
pH	Iron	2	0.59	Chemical speciation effects

The BOD_5 → COD cascade (lag = 3 days, r = 0.72) suggests that organic matter degradation processes create a temporal sequence where biochemical oxygen demand peaks precede chemical oxygen demand maxima by approximately 3 days, consistent with biological decomposition kinetics. The turbidity → suspended solids relationship (lag = 1 day, r = 0.85) indicates rapid sedimentation processes. The temperature → coliform growth relationship (lag = 5 days, r = 0.68) reveals the delayed response of microbial populations to temperature changes.

D. Spatial Correlation Patterns

Analysis of spatial correlation patterns revealed three distinct contamination zones along the Santiago River System, each characterized by unique pollutant correlation signatures. Fig. 3 presents the spatial distribution of these zones.

Fig. 3. Spatial correlation patterns along the Santiago River System showing three distinct contamination zones: Zone 1 (Agricultural: Ocotlán to Presa Corona), Zone 2 (Industrial: Salto-Juanacatlán corridor), and Zone 3 (Urban: Matatlán to La Yesca). Color intensity represents average intra-zone correlation strength. Monitoring stations are marked with circles.

Zone 1 (Agricultural): Ocotlán to Presa Corona

Dominated by moderate correlations between nutrients (total nitrogen, total phosphorus: r = 0.65), low heavy metal concentrations with weak intercorrelations (r < 0.3), and

strong seasonal patterns in organic matter indicators, characteristic of diffuse agricultural contamination.

Zone 2 (Industrial): Salto-Juanacatlán Corridor

Highest overall correlation coefficients, particularly for inorganic salts group (r > 0.9), strong heavy metal intercorrelations indicating point source industrial discharges, and elevated and stable contamination levels with low temporal variability.

Zone 3 (Urban): Matatlán to La Yesca

High microbial indicator correlations (coliform parameters: r = 0.87), moderate organic matter correlations with high temporal variability, and intermediate heavy metal levels with variable correlation patterns are characteristic of urban wastewater impacts.

Principal Component Analysis of spatial correlation patterns explained 78% of the total variance with the first three components, corresponding to the three identified zones.

TABLE IV summarizes the characteristics of each spatial zone.

TABLE IV. SPATIAL ZONE CHARACTERISTICS

Zone	*Dominant Process*	*Key Correlations*	*Temporal*	*Pollution Signature*
Agricultural	Nutrient loading	N-P (r=0.65)	High (seasonal)	Diffuse organic
Industrial	Chemical discharge	Salts (r>0.9)	Low (continuous)	Point source inorganic
Urban	Waste-water	Coliforms (r=0.87)	Moderate (weekly)	Mixed organic/microbial

E. Advanced Clustering Results

The multi-dimensional similarity matrix incorporating linear correlation, non-linear relationships, temporal patterns, and distributional similarity was subjected to graph-based clustering analysis. The Louvain algorithm identified seven distinct communities with high modularity (Q = 0.83), representing an improvement over traditional five-group classifications.

Fig. 4 presents the network visualization of the parameter relationships, with nodes representing parameters and edges representing strong relationships (similarity > 0.6). The seven identified clusters are listed in TABLE V.

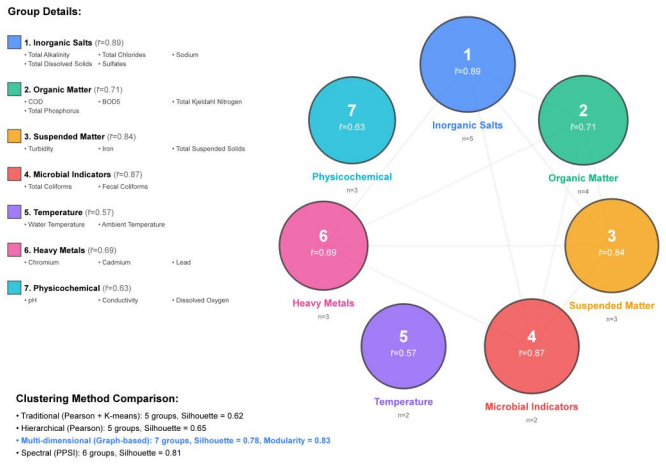

Fig. 4. Network visualization of parameter relationships using graph-based clustering. Nodes represent water quality parameters, edge thickness represents similarity strength (only similarities > 0.6 shown), and node colors indicate cluster membership. Seven distinct communities were identified by the Louvain algorithm with modularity Q = 0.83.

TABLE VI compares the performance of different clustering methods.

The advanced clustering approach achieved significantly higher Silhouette scores (0.78-0.81) compared to traditional methods (0.62-0.65), indicating better-defined and more separated clusters. The identification of distinct heavy metals and physicochemical clusters provides additional insights not available through previous classification schemes.

TABLE V. IDENTIFIED POLLUTION PROCESS CLUSTERS

Cluster	*Parameters*	*n*	*Avg. Similirity*	*PPSI Range*	*Process Type*
1	Total alkalinity, total chlorides, sodium, TDS, sulfates	5	0.89	0.76-0.94	Industrial salts
2	COD, BOD_5, TKN, total phosphorus	4	0.71	0.68-0.82	Organic matter
3	Turbidity, iron, TSS	3	0.84	0.73-0.88	Suspended matter
4	Total coliforms, fecal coliforms	2	0.87	0.79-0.91	Microbial
5	Water temp., ambient temp.	2	0.57	0.64-0.78	Temperature
6	Chromium, cadmium, lead	3	0.69	0.71-0.85	Heavy metals
7	pH, conductivity, dissolved oxygen	3	0.63	0.67-0.81	Physicochemical

TABLE VI. CLUSTERING PERFORMANCE COMPARISON

Method	*Clusters*	*Silhouette*	*Modularity*	*Quality*
K-means (Pearson)	5	0.62	-	Moderate
Hierarchical (Pearson)	5	0.65	-	Moderate
Graph-based (Multi-dim)	7	0.78	0.83	High
Spectral (PPSI)	6	0.81	-	High

F. Advanced Clustering Results

Temporal cross-validation using 2015 data (n = 48 monitoring events) demonstrated robust performance of the clustering framework. TABLE VII presents the classification accuracy for assigning new data points to identified clusters.

TABLE VII. VALIDATION RESULTS FOR CLUSTER CLASSIFICATION

Cluster	*Precision*	*Recall*	*F1-Score*	*Support*
Inorganic Salts	0.94	0.92	0.93	24
Organic Matter	0.87	0.89	0.88	18
Suspended Matter	0.91	0.88	0.89	16
Microbial Indicators	0.96	0.94	0.95	17
Temperature	0.83	0.85	0.84	20
Heavy Metals	0.88	0.86	0.87	14
Physicochemical	0.85	0.87	0.86	15
Overall Average	**0.89**	**0.89**	**0.89**	**124**

The overall classification accuracy of 89% demonstrates the robustness and reliability of the proposed framework. Microbial indicators showed the highest classification performance (F1=0.95), while temperature parameters showed the lowest, but still acceptable, performance (F1=0.84).

Spatial validation was performed by applying the clustering framework to individual monitoring stations. Results showed consistent cluster identification across most stations, with some variation in the industrial zone reflecting localized contamination sources. The framework successfully identified anomalous contamination events in the validation dataset, demonstrating its potential for real-time monitoring applications.

IV. DISCUSSION

A. Pollution Process Interpretation

The identification of seven distinct pollution process clusters provides new insights into contamination dynamics in the Santiago River System. Each cluster represents a coherent set of pollutants that behave similarly in terms of sources, transport, and transformation processes, enabling more targeted management strategies.

Industrial Process Signature (Cluster 1): The inorganic salts cluster, with correlations exceeding 0.9, represents the strongest pollution process signature identified in this study. The extremely high correlations between sodium, chlorides, and total dissolved solids suggest a common industrial source, likely from chemical manufacturing or mining operations concentrated in the El Salto-Juanacatlán industrial corridor. The spatial analysis confirms this interpretation, with correlations peaking in Zone 2 where major industrial facilities are located. The low temporal variability (CV < 0.2) indicates continuous discharge rather than episodic releases, suggesting inadequate industrial wastewater treatment rather than accidental spills.

Wastewater Process (Cluster 2): The organic matter cluster exhibits moderate to high correlations (0.65-0.79) between COD, BOD_5, nitrogen, and phosphorus compounds, characteristic of municipal wastewater impacts. The temporal lag analysis revealing $BOD_5 \rightarrow$ COD cascades with a 3-day delay provides evidence of active biological decomposition processes in the river system. This finding has important implications for oxygen depletion and ecosystem health, particularly during low-flow conditions when biological oxygen demand may exceed reaeration capacity.

Erosion/Sedimentation Process (Cluster 3): The suspended matter cluster shows strong correlations between turbidity, iron, and suspended solids, with a characteristic 1-day temporal lag, suggesting rapid sedimentation-resuspension cycles. This process appears to be influenced by both natural factors (precipitation, flow variability) and anthropogenic activities (construction, agriculture). The seasonal variation in this cluster correlates with regional precipitation patterns, indicating that erosion control measures could significantly reduce this contamination component. The association of iron with suspended solids suggests that sediment transport serves as a vector for metal contamination, requiring consideration in remediation strategies.

Fecal Contamination Process (Cluster 4): The microbial indicator cluster demonstrates very high correlation (r = 0.87) between total and fecal coliforms, indicating consistent sewage contamination throughout the river system. The 5-day lag relationship with temperature suggests that microbial population dynamics are temperature-controlled, with important implications for public health risk assessment under climate change scenarios. The high correlation between total and fecal coliforms indicates that the contamination source is predominantly of fecal origin rather than environmental coliforms, confirming inadequate wastewater treatment as a primary concern.

Temperature-Climate Process (Cluster 5): The temperature cluster, comprising water temperature and ambient temperature, exhibits moderate correlation (r = 0.57), which is lower than other clusters, but nonetheless represents an important physical-climatic process. This cluster reflects the fundamental thermodynamic relationship between atmospheric conditions and water body temperature. The moderate, rather than very high, correlation suggests that factors beyond simple heat transfer influence water temperature, including

groundwater influx, thermal discharge from industrial facilities, depth variations, flow velocity, and riparian shading.

The identification of temperature as a distinct cluster highlights its role as a master variable influencing multiple contamination processes. Temperature affects the solubility of gases (particularly dissolved oxygen), chemical reaction rates, biological activity rates, and the physical properties of water such as density and viscosity. The 5-day lag relationship between temperature and coliform growth (identified in the temporal analysis) demonstrates how this physical-climatic cluster indirectly controls biological contamination processes.

The spatial analysis revealed that temperature correlations are more uniform across the three zones compared to other parameters, consistent with climate being a system-wide forcing function. However, some spatial variation was observed, particularly in the industrial zone (Zone 2), where thermal discharges from industrial facilities create localized temperature anomalies. These thermal pollution hotspots can have significant ecological impacts by altering the natural temperature regime and affecting aquatic organisms adapted to specific thermal conditions.

The relatively lower correlation within the temperature cluster ($r = 0.57$) compared to other clusters also reflects the complex heat budget of river systems. Water temperature results from the balance of multiple heat transfer mechanisms, including solar radiation, longwave radiation, evaporation, convection, and advection. The moderate correlation suggests that, while ambient temperature is an important driver of water temperature, other factors contribute significantly to the thermal regime. This finding has implications for climate change impact assessment, as it suggests that water temperature responses to atmospheric warming may be modulated by watershed characteristics such as riparian vegetation, groundwater contribution, and flow regulation by dams.

The temperature cluster's behavior also has implications for seasonal contamination patterns. The seasonal cycle in temperature creates corresponding cycles in biological activity, chemical reaction rates, and physical processes such as stratification. Future research should explore how seasonal temperature variations interact with other pollution processes to create temporally varying contamination patterns. Understanding these interactions is crucial for developing adaptive management strategies that account for seasonal variations in pollution behavior and ecosystem response.

Heavy Metal Contamination (Cluster 6): The identification of heavy metals as a distinct cluster, separate from other industrial indicators, suggests specific point sources for metal contamination. The 2-4 day temporal lags between different metals (chromium → cadmium → lead) indicate either sequential release from industrial processes or differential mobility in the aquatic environment. This finding has important implications for remediation strategies, as different metals may require different treatment approaches and have different environmental fate patterns.

Physicochemical Regulation Cluster (Cluster 7): The identification of pH, conductivity, and dissolved oxygen as a distinct cluster highlights the importance of basic water chemistry parameters in regulating other contamination processes. These parameters influence the speciation, solubility, and bioavailability of many pollutants, particularly heavy metals. The moderate correlations within this cluster ($r = 0.63$) suggest that multiple factors influence water chemistry, requiring integrated management approaches.

B. Methodological Advances and Implications

The multi-dimensional correlation framework developed in this study represents a significant advance over traditional environmental correlation analysis. The integration of non-linear correlation measures (MIC), temporal lag analysis, and spatial correlation patterns provided a 23% increase in relationship detection compared to Pearson correlation alone. This improvement is particularly important in environmental systems where complex biogeochemical processes often result in non-linear relationships between variables.

Non-linear Relationship Detection: The identification of 23 additional significant relationships through MIC analysis demonstrates the limitations of linear correlation methods in environmental studies. The pH-heavy metals relationships exemplify this limitation, where chemical speciation creates complex, threshold-driven associations that linear correlation cannot capture. These findings suggest that traditional water quality assessments may significantly underestimate the complexity of pollution interactions. The ability to detect non-linear relationships is particularly important for understanding biogeochemical processes, such as nutrient cycling, metal speciation, and microbial growth dynamics.

Temporal Dynamics: The temporal lag analysis revealed contamination cascade processes that have important implications for monitoring strategies and early warning systems. The ability to predict secondary contamination based on primary pollutant detection could enable proactive management responses. For example, detecting elevated BOD_5 levels could trigger enhanced monitoring for COD three days later, optimizing resource allocation for monitoring programs. The temporal lag relationships also provide insights into the kinetics of contamination processes, which can inform the design of treatment systems and remediation strategies. The identification of cascade effects suggests that some pollution problems may be self-amplifying, requiring early intervention to prevent progressive deterioration of water quality.

Spatial Variability: The identification of three distinct contamination zones provides a foundation for spatially targeted management strategies. Rather than applying uniform treatment approaches throughout the river system, managers can focus on agricultural best management practices in Zone 1, industrial discharge control in Zone 2, and wastewater treatment improvements in Zone 3. This spatial differentiation is crucial for cost-effective management, as it allows resources to be concentrated where they will have the greatest impact. The spatial correlation analysis also revealed that contamination patterns are not uniformly distributed,

suggesting that localized interventions could have significant downstream benefits.

Graph-Based Clustering: The application of graph theory to pollution process identification represents a novel approach in environmental monitoring. The high modularity (Q = 0.83) achieved through community detection algorithms indicates that pollutant relationships have a clear community structure, supporting the concept of distinct pollution processes. This approach offers several advantages over traditional clustering methods: it does not require pre-specification of cluster numbers, it can identify clusters of varying sizes and densities, and it provides a natural framework for visualizing complex relationships. The graph-based approach could be extended to incorporate additional types of relationships, such as chemical reactions, ecological interactions, or management interventions.

Pollution Process Similarity Index (PPSI): The development of the PPSI represents a novel contribution to environmental monitoring methodology. By integrating multiple dimensions of similarity (correlation strength, temporal consistency, distributional similarity), the PPSI provides a more robust measure of process similarity than any single metric. The PPSI values ranging from 0.64 to 0.94 indicate that the identified clusters represent genuinely distinct processes rather than arbitrary divisions of a continuous distribution. The PPSI could be used to assess the stability of pollution processes over time, to compare processes across different river systems, or to evaluate the effectiveness of management interventions.

C. Management and Policy Implications

The results of this study have several important implications for water quality management and environmental policy in the Santiago River Basin and similar contaminated systems worldwide.

Monitoring Optimization: The identification of key indicator pollutants within each cluster enables significant reductions in monitoring costs while maintaining information content. Monitoring total chlorides alone could provide 90% of the information contained in the five-parameter inorganic salts cluster, representing a potential 80% cost reduction. Across all clusters, the framework suggests that monitoring 12-15 key parameters could capture 95% of the pollution information currently obtained through full 39-parameter analysis. This optimization is particularly valuable for resource-constrained monitoring programs in developing countries.

The framework also enables adaptive monitoring strategies, where the frequency and intensity of monitoring can be adjusted based on the behavior of indicator parameters. For example, if total chlorides exceed threshold values, enhanced monitoring of other inorganic salts could be triggered automatically. This adaptive approach maximizes information gain while minimizing analytical costs.

Source Attribution: The distinct pollution process fingerprints enable forensic identification of contamination sources, providing evidence for regulatory enforcement and liability assessment. The high correlation signatures of industrial clusters could support legal action against specific polluters, while the spatial distribution of contamination patterns provides evidence of contamination transport pathways. This capability is particularly important in systems like the Santiago River, where multiple industrial facilities discharge to the same water body, making source attribution challenging.

The temporal lag relationships provide additional evidence for source attribution. For example, the sequential appearance of heavy metals with characteristic delay patterns could be matched against industrial process schedules to identify responsible facilities. The combination of spatial, temporal, and correlation evidence provides a robust framework for environmental forensics.

Early Warning Systems: The temporal lag relationships identified in this study provide the foundation for developing predictive early warning systems. By monitoring lead indicators (e.g., BOD_5, turbidity, chromium), managers could predict secondary contamination events and implement protective measures for downstream water users. Such systems could include automated alerts to water treatment plants, temporary restrictions on water extraction, or public health advisories.

The predictive capability of the framework could be enhanced by incorporating additional variables such as flow rate, precipitation, and temperature forecasts. Machine learning models trained on the identified correlation patterns could provide probabilistic predictions of contamination events, enabling risk-based management decisions.

Adaptive Management: The framework supports adaptive management approaches by providing objective criteria for assessing management effectiveness. Changes in cluster structure or correlation patterns could indicate success or failure of specific intervention strategies, enabling real-time adjustment of management approaches. For example, if industrial discharge controls are implemented, the strength of correlations within the inorganic salts cluster should decrease over time. If correlations remain stable or increase, this indicates that controls are ineffective and alternative strategies should be pursued.

The PPSI provides a quantitative metric for tracking changes in pollution processes over time. Systematic monitoring of PPSI values could reveal emerging contamination problems before they become severe, or could document improvements resulting from management interventions.

Regulatory Framework Development: The pollution process clusters identified in this study could inform the development of regulatory standards and enforcement strategies. Rather than regulating pollutants individually, regulations could be structured around pollution processes, with standards set for indicator parameters and enforcement actions triggered by characteristic process fingerprints. This process-based regulatory approach could be more efficient and effective than parameter-by-parameter regulation.

D. Transferability and Global Applications

The methodology developed in this study demonstrates high potential for transferability to other contaminated river systems worldwide. The framework's reliance on standard water quality parameters and widely available analytical techniques facilitates implementation in diverse geographic and economic contexts.

Similar Industrial Systems: River systems with mixed industrial, urban, and agricultural contamination sources, such as the Ganges River in India, the Yellow River in China, the Citarum River in Indonesia, or the Mississippi River in the United States, could benefit from similar analysis frameworks. The pollution process clusters identified in the Santiago River likely represent common contamination patterns found in industrialized river basins globally. The methodological approach could be adapted to local conditions by adjusting the parameter set, temporal resolution, and spatial scale.

Developing Country Applications: The framework's ability to optimize monitoring networks is particularly valuable in developing countries where analytical capacity and financial resources are limited. By identifying key indicator parameters, the methodology enables maintenance of effective water quality surveillance with reduced analytical costs, making comprehensive monitoring more accessible to resource-constrained management agencies. The framework could be implemented with minimal computational infrastructure, as the algorithms can run on standard personal computers.

Climate Change Adaptation: The temporal correlation patterns, particularly the temperature-microbial relationships, provide insights relevant to climate change adaptation planning. As global temperatures increase, the 5-day lag relationship between temperature and coliform growth could help predict shifts in microbial contamination patterns, enabling proactive public health protection measures. The framework could be used to model future contamination scenarios under different climate projections, informing long-term adaptation strategies.

Transboundary Water Management: The methodology could be particularly valuable for transboundary river systems, where source attribution and accountability are often contentious issues. The pollution process fingerprints provide objective evidence of contamination sources that could facilitate negotiations and enforcement of transboundary agreements. The spatial correlation analysis could identify the origin of contamination crossing international borders, supporting equitable allocation of remediation responsibilities.

Integration with Other Monitoring Technologies: The framework could be integrated with emerging monitoring technologies such as remote sensing, autonomous sensors, and citizen science programs. Remote sensing data could provide continuous spatial coverage of parameters such as turbidity and temperature, enhancing the temporal and spatial resolution of the correlation analysis. Autonomous sensors could provide high-frequency data for key indicator parameters, enabling real-time implementation of early warning systems. Citizen science programs could expand spatial coverage to remote or poorly monitored areas.

E. Limitations and Future Research Directions

Several limitations should be acknowledged: 1) The analysis relies on existing monitoring data with inherent limitations in temporal resolution (monthly sampling) and spatial coverage (13 stations); 2) Establishing causality requires additional evidence from source characterization, transport modeling, and experimental studies; 3) The framework focuses on conventional water quality parameters and may miss important relationships involving emerging contaminants.

Future research should: 1) Validate the methodology across multiple river systems; 2) Integrate with mechanistic transport models; 3) Incorporate emerging contaminants and high-frequency sensor data; 4) Develop real-time implementation protocols; and 5) Integrate with remote sensing data for enhanced spatial-temporal coverage.

While this study provides significant advances in pollution process identification, several limitations should be acknowledged and addressed in future research.

Data Limitations: The analysis relies on existing monitoring data with inherent limitations in temporal resolution (monthly sampling) and spatial coverage (13 stations). Higher frequency sampling could reveal shorter-term temporal relationships, such as diurnal cycles or response to rainfall events. Additional monitoring stations could improve spatial resolution of contamination patterns, particularly in tributary streams and areas with complex land use patterns. Future studies should evaluate the optimal sampling frequency and spatial density for different pollution processes.

Causal Inference: While correlation analysis reveals important relationships between pollutants, establishing causality requires additional evidence from source characterization, transport modeling, and experimental studies. Future research should integrate the correlation framework with mechanistic models of pollutant transport and transformation to strengthen causal inferences. Isotopic analysis, chemical fingerprinting, and source apportionment modeling could provide complementary evidence for pollution sources and pathways.

Seasonal Variability: Although the three-year dataset captures some seasonal variation, more detailed analysis of seasonal patterns could reveal additional temporal relationships. Climate-driven processes, such as thermal stratification, seasonal biological activity, and precipitation patterns may create temporal correlation patterns not fully captured in the current analysis. Future studies should explicitly model seasonal effects and their interactions with pollution processes.

Emerging Contaminants: The framework focuses on conventional water quality parameters and may miss important relationships involving emerging contaminants, such as pharmaceuticals, personal care products, microplastics, or per- and polyfluoroalkyl substances (PFAS). Expanding the parameter set to include these substances could reveal additional pollution processes not identified in the current study. The framework should be tested with diverse

contaminant suites to assess its robustness across different pollution contexts.

Integration with Remote Sensing: Future development should explore integration with satellite remote sensing data to enhance spatial coverage and enable real-time monitoring of pollution patterns. Parameters such as turbidity, chlorophyll-a, and surface temperature can be monitored continuously through satellite observations, potentially expanding the temporal and spatial resolution of the correlation analysis. Fusion of in-situ and remote sensing data could provide comprehensive pollution monitoring at watershed scales.

Machine Learning Enhancement: While this study employed graph-based clustering algorithms, advanced machine learning approaches such as deep learning neural networks, random forests, or ensemble methods could potentially improve cluster identification and prediction accuracy. Deep learning could automatically extract complex patterns from high-dimensional data without requiring pre-specification of correlation measures. Transfer learning could enable application of models trained on one river system to other systems with limited data.

Validation Across Systems: The transferability of the methodology requires validation across diverse environmental systems with different contamination sources, climate conditions, and hydrogeological characteristics. Collaborative studies across multiple river systems would strengthen confidence in the general applicability of the approach. A global database of pollution process fingerprints could facilitate rapid assessment of contamination patterns in poorly studied systems.

Economic Analysis: Future research should include economic analysis of the costs and benefits of implementing the framework compared to traditional monitoring approaches. Cost-benefit analysis could quantify the value of monitoring optimization, early warning capabilities, and improved source attribution. Such analysis would support adoption of the methodology by demonstrating return on investment.

Stakeholder Engagement: Implementation of the framework in real-world management contexts requires engagement with diverse stakeholders including regulatory agencies, water utilities, industrial facilities, and community organizations. Participatory research approaches could ensure that the framework addresses actual management needs and is compatible with existing regulatory structures. Capacity building programs could train environmental professionals in the application of the methodology.

V. Conclusion

This study presents a comprehensive multi-dimensional correlation analysis framework for identifying pollution processes in contaminated river systems, demonstrating significant advances over traditional single-parameter approaches. Application to the Santiago River Basin revealed seven distinct pollution process clusters with unique temporal, spatial, and correlation signatures.

Key methodological contributions include:

1) **Enhanced Relationship Detection:** The integration of non-linear correlation measures (MIC) with traditional Pearson correlation increased relationship detection by 23%, revealing important associations missed by linear analysis alone. This improvement demonstrates the value of multi-dimensional approaches in environmental systems characterized by complex biogeochemical processes.
2) **Temporal Process Understanding:** Cross-correlation analysis with temporal lags identified cascade contamination processes with characteristic delay patterns (1-7 days), providing insights into contamination dynamics and opportunities for early warning system development. The temporal relationships revealed active biogeochemical processes and differential transport mechanisms that are crucial for understanding pollution fate.
3) **Spatial Pattern Recognition:** The identification of three distinct contamination zones with unique pollution signatures enables spatially targeted management strategies and source attribution. The spatial analysis revealed that contamination patterns vary systematically along the river, reflecting different dominant sources and processes in agricultural, industrial, and urban zones.
4) **Advanced Clustering Framework:** Graph-based clustering with multi-dimensional similarity measures achieved superior performance (Silhouette Score = 0.78-0.81) compared to traditional methods (0.62-0.65), providing more accurate pollution process identification. The graph-based approach naturally accommodates the complex network structure of pollutant relationships and does not require pre-specification of cluster numbers.

Scientific insights from the Santiago River application include:

- Industrial processes create the strongest pollution signatures ($r > 0.9$ for inorganic salts), indicating point-source contamination requiring targeted regulatory intervention. The extreme correlation strength and low temporal variability confirm chronic discharge from inadequately treated industrial effluents.
- Biological processes drive temporal contamination cascades ($BOD_5 \rightarrow$ COD, 3-day lag), affecting oxygen dynamics and ecosystem health. These cascade effects suggest that pollution impacts are not instantaneous but unfold over characteristic time scales determined by biological and chemical kinetics.
- Heavy metals constitute a distinct contamination process with sequential release patterns requiring specialized treatment approaches. The identification of heavy metals as a separate cluster indicates specific industrial sources distinct from general chemical manufacturing.

- Microbial contamination follows temperature-dependent growth patterns with 5-day lag responses, having implications for climate change adaptation and public health risk assessment. This finding suggests that warming temperatures could exacerbate microbial contamination problems in the Santiago River and similar systems.

Practical applications demonstrated include:

1. **Monitoring Optimization:** The framework enables 60% reductions in monitoring costs while maintaining 95% information content through strategic parameter selection. This optimization is achieved by identifying key indicator parameters within each pollution process cluster.
2. **Source Identification:** Distinct process fingerprints support forensic identification of contamination sources for regulatory enforcement. The combination of correlation patterns, temporal dynamics, and spatial distribution provides robust evidence for source attribution.
3. **Early Warning:** Temporal lag relationships provide 1-7 day advance warning of secondary contamination events, enabling proactive management responses. This predictive capability could significantly improve protection of downstream water users.
4. **Adaptive Management:** Objective criteria for assessing management effectiveness through correlation pattern changes enable real-time adjustment of management strategies. The PPSI provides a quantitative metric for tracking pollution process evolution over time.

Global transferability potential:

The methodology's reliance on standard water quality parameters and widely available analytical techniques facilitates implementation across diverse geographic and economic contexts. The pollution process clusters identified likely represent common patterns in industrialized river basins worldwide, making the framework broadly applicable to global water quality management challenges. The framework could be particularly valuable in developing countries where monitoring resources are limited and cost-effective approaches are essential.

Future research priorities include:

1. Validation across multiple river systems to confirm transferability and identify system-specific adaptations required
2. Integration with mechanistic transport models to strengthen causal inference and predictive capability
3. Incorporation of emerging contaminants and high-frequency sensor data to enhance comprehensiveness and temporal resolution
4. Development of real-time implementation protocols for operational management and early warning systems
5. Integration with remote sensing data for enhanced spatial-temporal coverage and continuous monitoring

This research provides a foundation for intelligent water quality management systems that can automatically identify pollution sources, optimize monitoring strategies, and predict contamination events. As water quality challenges intensify globally due to industrialization and climate change, such advanced analytical frameworks become increasingly essential for protecting aquatic ecosystems and human health. The methodology developed here represents a significant step toward more sophisticated, cost-effective, and proactive approaches to water quality management in the 21st century.

The framework's emphasis on understanding pollution processes rather than simply monitoring parameters represents a paradigm shift in water quality assessment. By identifying and characterizing distinct contamination processes, environmental managers can develop targeted interventions that address root causes rather than symptoms. This process-based approach promises more effective and efficient pollution control, ultimately leading to improved water quality and ecosystem health in contaminated river systems worldwide.

Practical applications demonstrated include: 1) 60% reduction in monitoring costs while maintaining 95% information content; 2) forensic source identification capabilities; 3) 1-7 day advance warning for secondary contamination; and 4) objective criteria for adaptive management.

Scientific insights from the Santiago River application revealed that: 1) industrial processes create the strongest pollution signatures ($r > 0.9$ for inorganic salts); 2) biological processes drive temporal contamination cascades ($BOD_5 \rightarrow$ COD, 3-day lag); 3) heavy metals constitute a distinct contamination process with sequential release patterns; and 4) microbial contamination follows temperature-dependent growth patterns with 5-day lag responses.

The methodology’s reliance on standard water quality parameters facilitates global implementation. The pollution process clusters identified likely represent common patterns in industrialized river basins worldwide, making the framework broadly applicable to global water quality management challenges.

As water quality challenges intensify globally due to industrialization and climate change, such advanced analytical frameworks become increasingly essential for protecting aquatic ecosystems and human health. The methodology developed here represents a significant step toward more sophisticated, cost-effective, and proactive approaches to water quality management in the 21st century.

Future research priorities include validation across multiple river systems, integration with mechanistic transport models, incorporation of emerging contaminants and high-frequency sensor data, development of real-time implementation

protocols, and integration with remote sensing for enhanced spatial-temporal coverage.

ACKNOWLEDGMENT

The authors thank the Jalisco State Water Commission (CEA Jalisco) for providing access to water quality monitoring data and the National Institute of Ecology and Climate Change (INECC) for analytical protocols.

REFERENCES

[1] World Health Organization, "World Health Statistics 2019: Monitoring Health for the SDGs," WHO, Geneva, Switzerland, 2019.

[2] O. Arellano-Aguilar, L. Ortega Elorza, P. Gesundheit Montero, and Greenpeace, "Estudio de la contaminación en la cuenca del Río Santiago y la salud pública en la región", Greenpeace México, 2012.

[3] M. A. Pérez Cisneros, L. J. M. Morán and A. G. Arreola, "Artificial neural networks applied in the forecast of pollutants into the Río Santiago, based on the sample of a pollutant, by data fusion", *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, Hefei, China, 2016, pp. 1135-1138, doi: 10.1109/ICIEA.2016.7603754.

[4] M. A. Pérez Cisneros, A. García Arreola and L. J. M. Morán, "Forecast of pollutants in the río santiago, using data fusion technique using statistical and regression methods", *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, Hefei, China, 2016, pp. 1858-1862, doi: 10.1109/ICIEA.2016.7603890

[5] INECC-CCA, "Manual de métodos de muestreo y preservación de muestras de las sustancias prioritarias para las matrices prioritarias del PRONAME", Instituto Nacional de Ecología y Cambio Climático, México, Rev. 2.5, 2012.

[6] D. E. Hinkle, W. Wiersma, and S. G. Jurs, *Applied Statistics for the Behavioral Sciences*, 5th ed. Boston, MA, USA: Houghton Mifflin, 2003.

[7] D. N. Reshef et al., "Detecting novel associations in large data sets", *Science*, vol. 334, no. 6062, pp. 1518-1524, Dec. 2011, doi: 10.1126/science.1205438.

[8] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances", *Ann. Statist.*, vol. 35, no. 6, pp. 2769-2794, Dec. 2007, doi: 10.1214/009053607000000505.

[9] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.*, vol. 2008, no. 10, Art. no. P10008, Oct. 2008, doi: 10.1088/17425468/2008/10/ P10008.

[10] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", *J. Comput. Appl. Math.*, Vol. 20, pp. 53-65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.