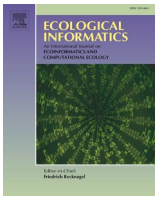
Ecological Informatics 82 (2024) 102655

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Improving river water quality prediction with hybrid machine learning and temporal analysis

Alberto Fernández del Castillo [a], Marycarmen Verduzco Garibay [a], Diego Díaz-Vázquez [a], Carlos Yebra-Montes [b], Lee E. Brown [c], Andrew Johnson [c], Alejandro Garcia-Gonzalez [d,*], Misael Sebastián Gradilla-Hernández [a,*]

[a] Tecnologico de Monterrey, Escuela de Ingenieria y Ciencias, Laboratorio de Sostenibilidad y Cambio Climático, Av. General Ramon Corona 2514, Nuevo México, CP 45138 Zapopan, Jalisco, Mexico
[b] ENES-León, Universidad Nacional Autónoma de México, Blvd. UNAM 2011, Predio el Saucillo y El Po-trero, CP, 37684 León, Guanajuato, Mexico
[c] School of Geography and water@leeds, University of Leeds, Leeds LS2 9JT, UK
[d] Tecnologico de Monterrey, Escuela de Medicina y Ciencias de la Salud, Av. General Ramon Corona 2514, Nuevo Mexico, CP, 45138 Zapopan, Jalisco, Mexico

ARTICLE INFO

Keywords:
Water Quality Index
Highly polluted river
Time series analysis
Cluster analysis
Monitoring network
Data Science

ABSTRACT

River systems provide multiple ecosystem services to society globally, but these are already degraded or threatened in many areas of the world due to water quality issues linked to diffuse and point-source pollutant inputs. Water quality evaluation is essential to develop remediation and management strategies. Computational tools such as machine learning based predictive models have been developed to improve monitoring network capabilities. The model's performance is reduced when datasets composed of reductant information are used for training, on the other hand, the selection of most representative and variable water quality scenarios could result in higher precision. This study analyzed historical water quality behavior in the Santiago River, Mexico, to identify the most variable and representative data available to train machine learning models (Adaptive Neuro Fuzzy Inference System – ANFIS, Artificial Neural Network – ANN, and Support Vector Machine - SVM). Thirteen monitoring sites were clustered according to their water quality variability from 2009 to 2022. Subsequently, a Time Series Analysis (TSA) was used to select the most representative monitoring station from each cluster. Data for 6/13 monitoring sites were retained for the Best Training Subset (BTS) used to train restricted models that performed with similar (ANN and SMV) or higher (ANFIS) prediction accuracy (in terms of RMSE, MAE, MSE and R^2) for both training and testing. This study provides evidence of water quality data containing redundant information that is not useful to improve machine learning model performance, in turn leading to overtraining. Combined analytical approaches can maximize the representativeness and variability of data selected for machine learning applications, leading to improved prediction.

1. Introduction

Anthropogenic activities including population growth, waste generation, changes in land use and climate change have driven major changes in river water quality worldwide (Li et al., 2022). Degradation of water quality leads to significant issues for biodiversity, agricultural crop growth and water consumption. The cost of degraded watersheds for water supply utilities alone has been estimated as >$5 billion annually (McDonald et al., 2016). Water quality monitoring programs (WQMP) aim to accurately assess the type and extension of water source pollution (Duan et al., 2016), through the constant and long-term measurement of biological, physical, and chemical parameters. However, non-specialists can find it challenging to evaluate, interpret and synthesize the emergent complex datasets (Gitau et al., 2016). Thus, water quality indices (WQIs) are used widely to summarize the information of multiple water quality parameters into a single dimensionless metric or category (Shil et al., 2019). WQIs are commonly used to evaluate surface and groundwater sources and to support decision-making around water management (Raman et al., 2009).

The measurement of water quality parameters (WQPs) required for WQI calculation can be expensive and time-consuming (Ho et al., 2019). However, computational approaches can be used for a more practical

* Corresponding authors.
E-mail addresses: alexgargo@tec.mx (A. Garcia-Gonzalez), msgradilla@tec.mx (M.S. Gradilla-Hernández).

https://doi.org/10.1016/j.ecoinf.2024.102655
Received 27 January 2024; Received in revised form 14 May 2024; Accepted 26 May 2024
Available online 6 June 2024

estimation and prediction of a WQI. The development of predictive models for WQI calculation offers multiple advantages for monitoring network improvement such as reducing the number of WQPs necessary for calculation and enlarging the monitoring networks to wider catchment areas (Zhu et al., 2022). Using predictive models also allows for real-time evaluation of water quality through the immediate interpretation of water quality data, which is impossible for traditional approaches relying on the analyst capacity (Ahmed et al., 2019). Additionally, WQI calculations based on predictive models offer the potential to reduce the time and cost of water quality monitoring programs by lowering the need for data collection in space and time. This approach is especially beneficial for underdeveloped or developing regions with limited financial resources and infrastructure to perform long-term monitoring programs (Lobato et al., 2015).

Machine learning (ML) approaches are ideal for predicting complex systems such as river systems where multiple surface water quality parameters are measured. Different machine learning algorithms have been evaluated in previous research to predict water quality parameters and water quality indices. Artificial Neural Network (ANN) is the most widely applied algorithm for water quality prediction in surface water, and results show a superior performance compared to conventional regression approaches (Rajaee et al., 2020). Another commonly used algorithm is Support Vector Machine (SVM), which in several cases has displayed higher prediction accuracy than ANN derived from its effectivity in reducing generalization error (Zhu et al., 2022). Recent research has focused on the application of hybrid models combining the capabilities of two different forecasting techniques to improve water quality prediction. For example, a WQI was predicted using a hybrid model combining Artificial Bee Colony and Back Propagation Neural Network algorithms, obtaining significantly higher prediction accuracy compared to SVM and Long Short-Term Memory (LSTM) neural networks (Chen et al., 2023). Other approaches include a previous data transformation to optimize the response of the predictive algorithm applied. For example, variational mode decomposition was applied to decompose water quality data into a series of relatively stable components used to train a LSTM neural network (Wang et al., 2023). This approach produced higher prediction accuracy compared to a single LSTM and Recurrent Neural Network.

The Adaptive Neuro-Fuzzy Inference System (ANFIS) is another hybrid model commonly applied to predict complex environmental systems behavior. This artificial intelligence (AI) approach couples two machine learning algorithms: an artificial neural network (ANN) and a fuzzy inference system (FIS). An ANN algorithm uses data and feedback to learn from the system behavior and estimate the output; however, the interpretability and understanding of the model are challenging (Sahu et al., 2011). In contrast, fuzzy logic is based on IF-THEN rules defining the relationships between variables expressed as linguistic terms, which is an easily interpretable approach. Fuzzy logic can incorporate existing knowledge and experience of a specific system or ecosystem into the algorithm. This advantage enables it to effectively analyze environmental issues where numerous interrelated variables significantly impact the results (Ellina et al., 2020). The ANFIS combines the capabilities of ANN and FIS, creating a hybrid intelligent system where the neural network is used to learn the fuzzy decision rules. This approach has the advantage of modeling nonlinear functions between input and output variables (Jang, 1994). ANFIS have been used increasingly together to predict surface water and groundwater quality (Dewanti and Abadi, 2019) and are efficient alternative tools for modeling and forecasting complex hydrological systems (Yan et al., 2010). However, the performance of machine learning models depends largely on the quality of the dataset used for training.

An ideal dataset would include a wide range of data representing all the possible states of the modeled system (Banadkooki et al., 2020). However, environmental processes are slow and seasonal, and can produce repetitive datasets which are inadequate for model training. Significant state changes are often observed only after several years of study and larger datasets do not necessarily produce more precise models as the use of too much data can lead to model overtraining and precision loss (Sjöberg and Ljung, 1995). Additionally, water quality datasets are composed of a large number of parameters that often represent redundant information (Haghiabi et al., 2018; Muharemi et al., 2019). Despite these known issues, detailed analysis of historical variation in a dataset is rarely employed for selecting training data subsets prior to developing predictive models. It is anticipated that their implementation may improve the training step and, therefore, lead to improved predictive outputs. For these reasons, analyzing historical trends in water quality is necessary to find the most variable and representative scenarios useful for machine learning model training.

The aim of this study was to develop and evaluate a novel approach based on TSA and CA to identify the most variable and representative scenarios within a water quality dataset collected in the Santiago River, Mexico. This reduced dataset was then used to train machine learning models, which were contrasted with outputs from the same approach but trained on the complete dataset. The results are evaluated to illustrate how machine learning models can predict water quality dynamics accurately, with a particular focus on temporal trends in the Santiago River. We also examine the extent to which water quality monitoring networks can be optimized by using advanced data analytics to identify representative zones for future data collection. In the following section, a description of the sampling site and water quality data collection is provided, then, we summarize the SR-WQI calculation and give details on the Cluster Analysis (CA), Time Series Analysis (TSA) and Machine Learning models applied. In Section 3 results from CA, TSA and Machine Learning models' prediction are displayed and these results are discussed in Section 4.

2. Materials and methods

2.1. Site description

The Santiago River (SR) is one of the largest rivers in Mexico; it originates in Lake Chapala (20°19′00.4"N, 102°47′30.9"W), and its mainstem has a length of 562 km ending in the Pacific Ocean (Rizo-Decelis & Andreo, 2016) (Fig. 1). A proportion (13%) of the Santigo River basin is in the State of Jalisco, covering an area of 9492.3 km^2. Jalisco is the third most populous state in the country, with a total of 8.3 M inhabitants in 2020 (INEGI, 2020), including the Metropolitan Area of Guadalajara (MAG). This portion of the basin encompasses extensive urban, agricultural, livestock, and industrial development. Since the 1950s, wastewater and runoff from crop fields, urban settlements, and landfills have been discharged into the SR. Additionally, there is a lack of sanitation infrastructure in the basin, affecting the river water quality severely, which in turn presents significant ecological and public health risks (McCulligh and Vega Fregoso, 2019). Since 2009, the Jalisco State Water Commission has monitored water quality in 10 points of the main channel (RS01 - RS10), two sites in the "El Ahogado" tributary (AA01 and AA02), and one site in the Zula River that flows through the highest part of the basin and crosses the most intensive agro-industrial region (RZ01). A total of 13 monitoring sites (MS) are used for water quality evaluation in Santiago River (Fig. 1). Water quality monitoring was performed according to standard methods (Rice et al., 2012) and the dataset used in this study was composed of 42 water quality parameters (WQPs) measured monthly from January 2009 to May 2023 ($n = 89{,}698$ observations) by the regional water agency "Comisión Estatal del Agua – CEA Jalisco", available at http://info.ceajalisco.gob.mx/sca/.

2.2. Data pre-processing

The complete dataset of 42 WQPs (Appendix A) for the 13 monitoring sites was subjected to an outlier detection and artificial data replacement process, as Casillas-García et al. (2021) reported. Processed data were used subsequently to train machine learning models, first with

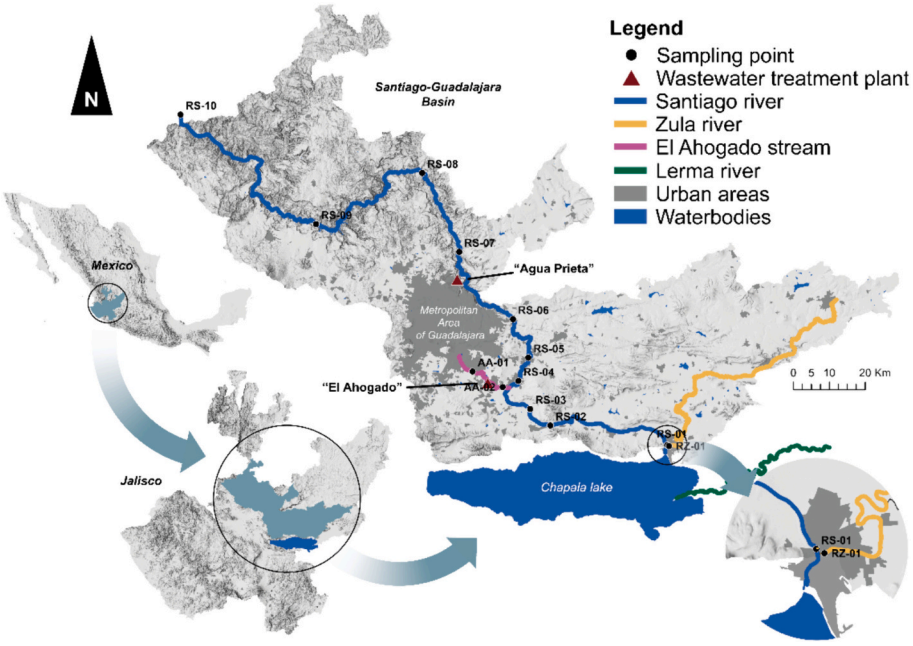


Fig. 1. Study area and geographical distribution of sampling points along the Santiago-Guadalajara River.

the full dataset of 13 monitoring sites and secondly with the best training subset (BTS) composed of sites selected by CA and TSA. Predictions from all models were evaluated by comparing the root-mean-square-error (RMSE), mean squared error (MSE), mean absolute error (MAE) and coefficient of determination (R^2) and finally, the BTS-ANFIS was tested against new data and individual data from each monitoring site. The general procedure for the training data selection and model development is described in Fig. 2.

2.3. Santiago River Water Quality Index Calculation

The Santiago River Water Quality Index (SR-WQI) was developed previously using a statistical multivariate approach, as reported by Casillas-García et al. (2021). The SR-WQI reduces the information provided by 42 WQP measured by the Jalisco State Water Commission into one dimensionless single number, facilitating the communication and analysis of river conditions for decision-makers and the general public. Through this approach, seventeen WQPs were identified as the most representative to be included in WQI calculation, and rating curves were developed for each using data from the last decade. Rating curves are used to relate the measurement range of each WQP with a determined water quality value, assigning a number from 0 (very poor quality) to 100 (excellent quality).

The water quality values assigned to each WQP measurement are called sub-indices (Qi). The rating curves developed by Casillas-García et al. (2021) include national legal limits for water quality regulation and provide high sensitivity for detecting measurements outside the thresholds established for protecting aquatic life.

The SR-WQI is calculated using a weighted average of each WQP (Eq. 1):

$$WQI = \sum_{k=1}^{17} w_k Qi_k \quad (1)$$

where $0 \leq Qi_k \leq 100$ is the rated sub-index obtained through the rating curves, and w_k is the weight of each k_{th} parameter (Table 1). Using the subindices of the selected parameters a PCA was performed to establish weights. The best set of weights was selected as the one that better distinguished between seasons and sampling points when using a linear discriminant analysis and ANOVA. Thus, the weight selected for each WQP provides high sensitivity to temporal and spatial variability.

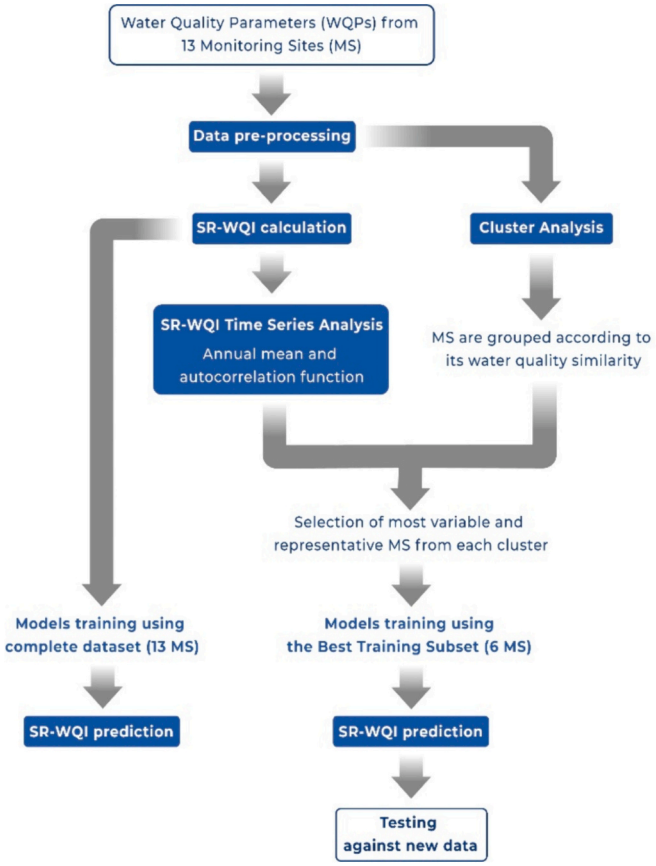


Fig. 2. Diagram of the data selection process used for ML models training.

Table 1
Water quality parameters and their assigned weights for the WQI calculation.

Number	Parameter	Abbreviation	Weight
1	Cadmium	Cd	0.057
2	Chromium	Cr	0.068
3	Biological oxygen demand	BOD_5	0.067
4	Dissolved oxygen	DO	0.064
5	Fecal coliforms	FC	0.045
6	Fluoride	FL	0.078
7	Fats, oils, and grease	FOG	0.045
8	Mercury	Hg	0.032
9	Ammonia	NH_3	0.072
10	Nitrates	NO_3	0.089
11	Lead	Pb	0.043
12	Hydrogen potential	pH	0.044
13	Total suspended solids	TSS	0.060
14	Sulfides	SULF	0.058
15	Total dissolved solids	TDS	0.080
16	Temperature	TEMP	0.046
17	Zinc	Zn	0.050

Table 2
Classification ranges.

WQI Range	Water Quality Class
0–25	Very bad
25–50	Bad
50–70	Medium
70–90	Good
90–100	Excellent

Finally, the SR-WQI calculation is assigned to a water quality class according to Table 2.

2.4. Cluster analysis

Cluster Analysis (CA) has the potential to identify and discard redundancies within datasets by grouping observations and thus minimizing within-group variance. The number of clusters is determined based on the analyzed dataset, allowing for the construction of robust and representative training datasets (Dincer and Yalçin, 2016). Clustering analysis has previously been applied to visualize similarities among analyzed datasets and identify redundancies in environmental monitoring networks (Dincer and Yalçin, 2016). Using the complete dataset of 42 WQPs (inputs), the 13 sites were grouped into clusters. The analysis was performed using the hierarchical clustering algorithm based on the principle that observations within the same cluster are as similar as possible, whereas objects from different clusters are as dissimilar as possible. In hierarchical clustering, each cluster is constructed based on the proximity matrix generated based on the observations and grouped to minimize the within-cluster variance, resulting in k clusters with similar characteristics within themselves (Murtagh and Contreras, 2012). The optimum number of clusters was determined based on the average silhouette index method, as reported by Kassambara (2017). The cluster number, which presented the maximum average silhouette width, was selected as the optimum cluster count (Appendix B). Based on the optimum number of clusters, a clustering analysis was performed using the "ward.D2" agglomeration method to compute the distance between clusters to minimize the total within-cluster variance (Kassambara, 2017). A dendrogram was constructed based on the identified clusters. The average silhouette index analysis and clustering analysis were conducted in RStudio® software version 4.1.2 employing the packages tidyverse, ggdendro, corrplot, ggcorrplot, FactoMineR, factoextra, multcomp, clustertend, NbClust, pvclust, fpc, lubridate, openxlsx, and ggpubr.

2.5. Time series analysis of RS-WQI

Time Series Analysis (TSA) can be applied to water quality (Ghashghaie et al., 2022) to estimate future values of WQP or WQI based on historical data. TSA aims to understand and model the stochastic mechanism of hydrologic phenomena (Ghashghaie et al., 2022). This analysis can also detect long-term trends in water resources. Previous studies have used TSA to plan water resources management (Patle et al., 2015) and to predict WQPs (Elhag et al., 2021). Historical trends of the Santiago River Water Quality Index (SR-WQI) were analyzed using a TSA (TSA_{WQI}). Using the TSA_{WQI}, the most representative sites from each cluster were identified based on the variability of water quality. The predictive model's precision is improved when a dataset including various water quality scenarios is used for training. The annual mean and standard deviation of SR-WQI (output) were plotted to identify general trends and significant variations in water quality. The general behaviors were identified according to the following expression:

$$\alpha \leq \overline{WQI_t} \leq \beta \text{ or } \beta \geq \overline{WQI_t} \geq \alpha \text{ for } t \in \{2009, \cdots, 2022\} \quad (2)$$

where α and β are the minimum and maximum values observed in the SR-WQI annual mean ($\overline{WQI}$) respectively, in the time period t. Once described, the general behaviors of SR-WQI annual mean and the individual variations from these tendencies were also identified for each monitoring site. Sites with the most significant deviation from general tendencies were preferred for training of ML model.

Additionally, the statistical autocorrelation function (ACF) was calculated for the SR-WQI. The ACF was used to identify the "self-similarity" or periodicity of the water quality data from each site. The ACF has been applied widely in time series analysis to detect trends and analyze seasonal variation (Jiang and Adeli, 2004). In this research, ACF was used to identify the seasonal behavior of the SR-WQI and select the site with the minimum autoregressive trend. This function measures the similarity between an individual time series separated by different time lags (Arora and Keshari, 2021).ACF graphs were constructed using ggplot2 package in RStudio ® software version 4.1.2. The dashed lines in the ACF graph represent confidence bands at the 95% confidence level. Points outside this band indicate statistically significant different values (lag 0 is always 1). Since we are interested in using datasets with the greatest variability, sites presenting the least autocorrelation (least number of statistically significant different lags) were retained for models training.

2.6. Machine learning models

Machine Learning models were developed using MATLAB® 2021 software. The ANFIS model was developed using the Neuro-Fuzzy Designer toolbox, the ANN was created using the fitrnet function, Regression Learner toolbox was used for SVM, for all models default parameters were used. A Gaussian type (guessmf) was selected to create membership functions in ANFIS model. The bilayered neural network method was used for ANN development, while the medium Gaussian method for SVM was selected. The full dataset was composed of all sites (13 MS) water quality (WQPs and SR-WQI) data covering a period from 2009 to 2015 ($n = 894$), for the Best Training Subset (BTS) only data from sites selected by clustering and time series analyses (6 MS) were included, covering a period from 2009 to 2015. Only 12/17 parameters used for SR-WQI calculation were used for model development. These parameters were defined based on the results of Fernández del Castillo et al. (2022): Cd, BOD_5, DO, FC, FL, FOG, NH_3, Pb, pH, SST, TDS, TEMP. The other five parameters (Cr, Hg, NO_3, SULF, and Zn) were excluded because they were not significant for the development of previous predictive models for the SR-WQI, including multiple linear regression, lasso and ridge regressions and generalized additive model (Fernández del Castillo et al., 2022). For the validation step, the complete set of sites covering data from 2016 to 2020 ($n = 657$) was used, while the period

from 2021 to 2022 and all site data were used for the testing step (n = 372). Additionally, the ANFIS model was tested against each of the individual datasets from the 13 monitoring sites to test the model's ability to predict water quality in each site individually, covering the time from 2009 to 2022 ($n \approx 140$). The models' performance was evaluated using the root-mean-square-error (RMSE), mean squared error (MSE), mean absolute error (MAE) and coefficient of determination (R^2).

3. Results

3.1. Cluster analysis

Using the complete dataset of 42 WQPs (inputs), the average silhouette index method was applied to select the optimum cluster size, resulting in 6 clusters based on the maximum average silhouette width of all evaluated cluster counts. The results of clusters distribution and its location is shown in Fig. 3.

3.2. Time series analysis

Three different tendencies were identified for the SR-WQI annual mean (Fig. 4). Every site was similar to one of these; however, variations from these tendencies were found in several sites (Table 3). Sites with more significant variations from each cluster were preferred for BTS. Additionally, lower autocorrelation between sites included in the same cluster was also considered for selecting (Table 3) sites for BTS.

Cluster A was the only cluster that included two sites with different tendencies (RS09 and RS10). The site RS09 showed a trend like Tendency 1 (Fig. 4) but oscillating at higher values between 48 and 54. Site RS10 presented a similar behavior to Tendency 2 (Fig. 4) but was steadier, displaying only two phases instead of four. RS10 is located at the lowest part of the Santiago River basin, where urbanization is low, and vegetation is greater, causing a natural improvement in water quality. For these reasons and the lower autocorrelation displayed by RS09, this monitoring site was included in BTS.

Cluster B included sites RS07 and RS08, which displayed the trend of Tendency 2 (Fig. 4) characterized by an initial period where the average SR-WQI maintained between 31 and 37 from 2009 to 2012 followed by a steady increase during the years 2013, 2014, and 2015. This stage was followed by a second stable period where the SR-WQI varied between 42 and 46, then dropped for two years (2019 and 2020). Finally, a steady phase was observed in the range of 38 to 42 for the last two years, 2021 and 2022. No significant variation from Tendency 2 was observed in Cluster B, but RS07 was preferred for BTS because it displayed a slightly lower autocorrelation. The time series, annual mean, and the ACF for SR-WQI in RS07 are shown in Fig. 5b, e and h. The increase in SR-WQI observed in Tendency 2 was probably caused by the construction of the WWTP "Agua Prieta" (Fig. 1) in the year 2014, driving a water quality improvement in all the sites downstream (RS07, RS08, RS09, and RS10).

Tendency 1 was characterized by a SR-WQI with no clear trend to improve or worsen over time. The average SR-WQI oscillated between 43 and 48 in these sites, indicating an almost steady behavior. This trend is displayed in RS03 (Fig. 5a, d and g). This behavior was the most common within all the monitoring sites studied, including RS01, RS02, RS03, RS04, RS05, RS06, RS09, and RZ01, which were grouped in clusters A, C, D, and F (Table 3). Cluster C was composed only of RS06, as this site did not present a significant variation from the trend shown in Tendency 1; consequently, it was excluded from model training. Cluster D included the biggest group of sites; accordingly, two from this cluster were used for model training. Sites RS03 and RS04 were selected since they presented more variation (Table 3), and RS04 displayed the lowest autocorrelation from this cluster. Cluster E included sites AA01 and AA02, both located in the "El Ahogado" stream and displaying the trend shown in Tendency 3 (Fig. 4). This trend is characterized by a steady increase in water quality from 2009 to 2014 when the SR-WQI reached its maximum value. From 2015 to 2020, a steady decrease in water quality was observed until reaching values like those reported at the beginning of the time series. Fig. 5c, f, and i show the time series, the annual mean, and the ACF of the AA01 site, respectively. This cluster showed the lowest mean SR-WQI values ranging from 30 to 45. Additionally, the AA01 site displayed even lower values than AA02, probably because AA02 is located downstream of the WWTP "El Ahogado". Even if site AA02 presented a slight variation in 2016, data from AA01 was preferred for BTS since it displayed a lower autocorrelation.

Cluster F was composed of sites RS01 and RZ01. A slight decrease was displayed in RZ01 for the average SR-WQI value; however, it maintained in the Tendency 1 range. The RZ01 monitoring station is located at the end of Zula River, and its water quality can be influenced by the slow but steady agro-industrial development and cattle raising of the municipalities located at the higher zone of the basin. Additionally, RZ01 displayed the lowest autocorrelation from its cluster (Table 3).

The SR-WQI generally displayed a quasi-stationary behavior since only slight fluctuations were observed in most time series. Significant changes are only related to the construction of extensive sanitation

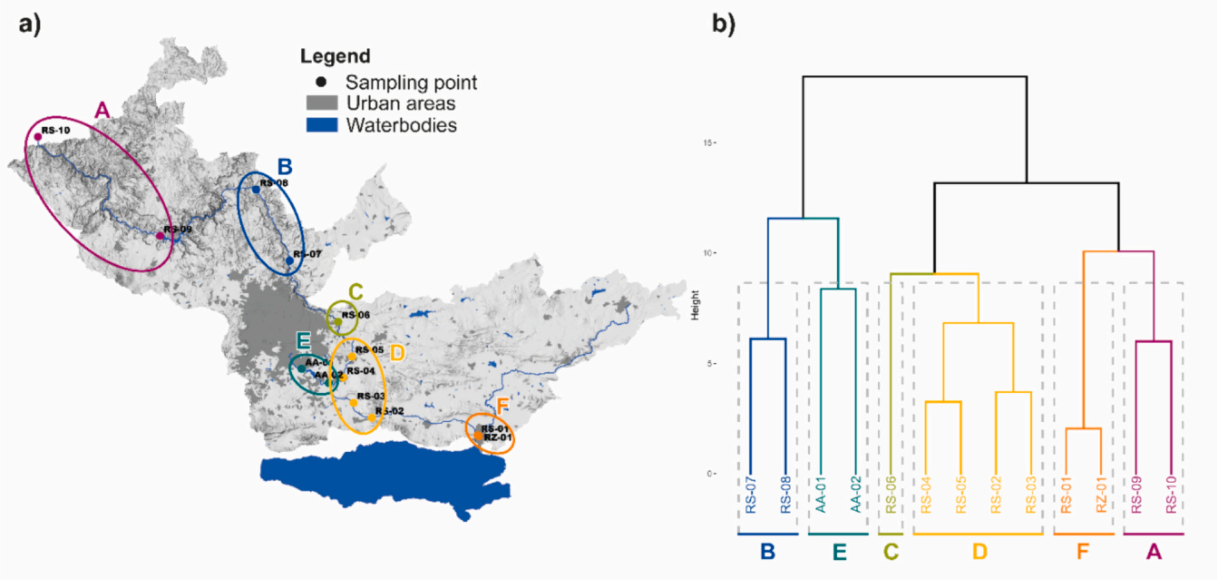


Fig. 3. a) Spatial distribution of clustered sampling points and b) resulting dendrogram from the cluster analysis.

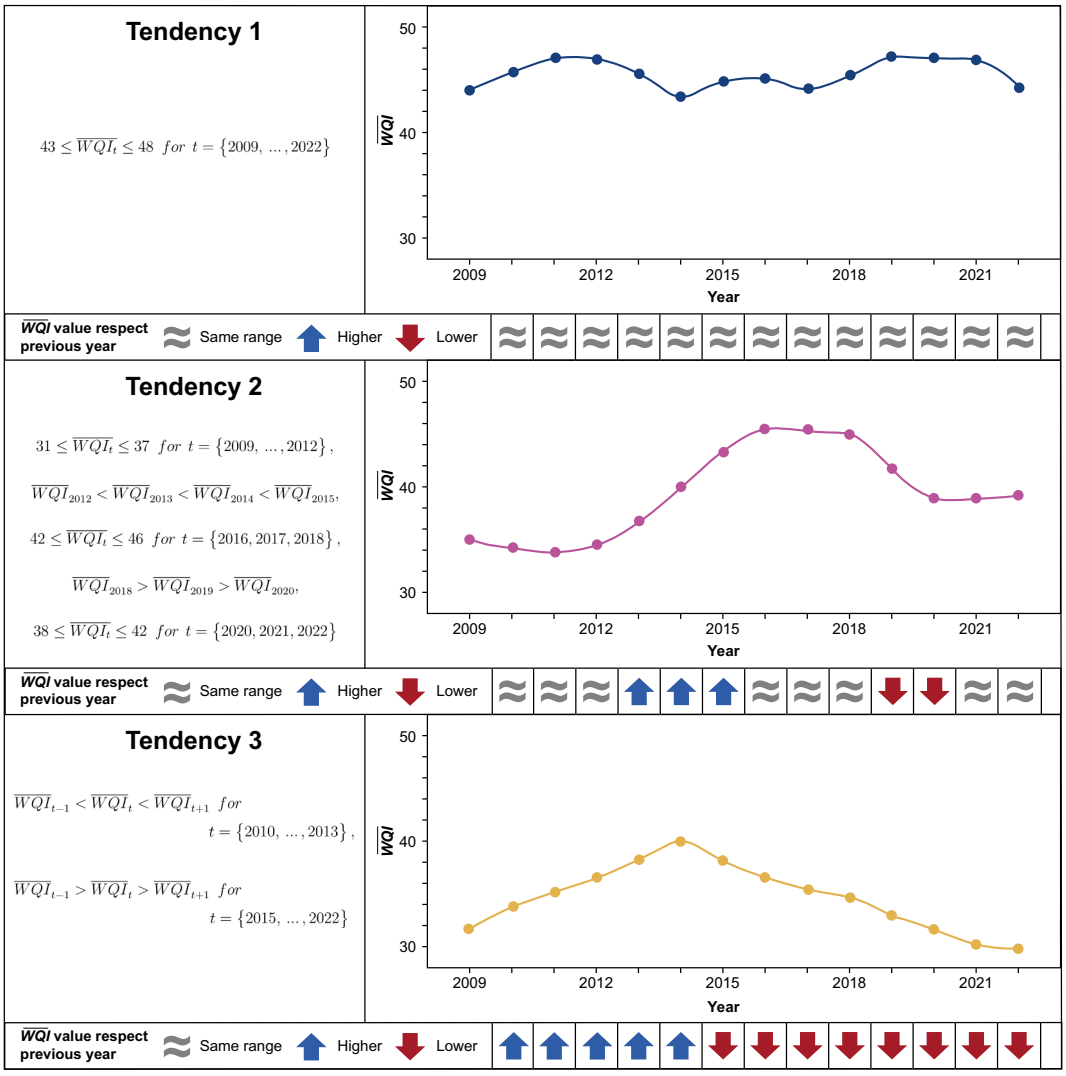


Fig. 4. General tendencies identified for the SR-WQI annual mean.

Table 3
Criteria for the selection of data included in Best Training Subset - BTS.

Cluster	Monitoring Stations	Tendency	Variation from tendency	Autocorrelation	BTS
A	RS09	1	*higher range*→$48 \leq \overline{WQI}_t \leq 54$ *Only two phases*	lower	Included
	RS10	2	$49 \leq \overline{WQI}_t \leq 53$ *for* $t = \{2009, 2010, 2011, 2012\}$ $57 \leq \overline{WQI}_t \leq 60$ *for* $t = \{2014, \cdots, 2020\}$	–	Not included
B	RS07	2	NA	lower	Included
	RS08	2	NA	–	Not included
C	RS06	1	NA	–	Not included
D	RS02	1	$\overline{WQI}_{2014} > 48$	–	Not included
	RS03	1	$\overline{WQI}_{2014} < 43$ *and* $\overline{WQI}_{2017} > 48$	–	Included
	RS04	1	$\overline{WQI}_t < 43$ *for* $t = \left\{ \begin{matrix} 2010, 2011, 2017, \\ 2018, 2019, 2020 \end{matrix} \right\}$	lower	Included
	RS05	1	NA	–	Not included
E	AA01	3	NA	lower	Included
	AA02	3	$\overline{WQI}_{2016} > \overline{WQI}_{2015}$	–	Not included
F	RS01	1	NA	–	Not included
	RZ01	1	$\overline{WQI}_{2018} < 43$	lower	Included

infrastructure such as WWTP. However, as in the case of monitoring stations RS07, RS08, and RS10, the system tends to recover stability after the initial change.

3.3. Machine learning models

Based on the results of clustering and times series analysis, data from six sites (RZ01, AA01, RS03, RS04, RS07, and RS09) were included in

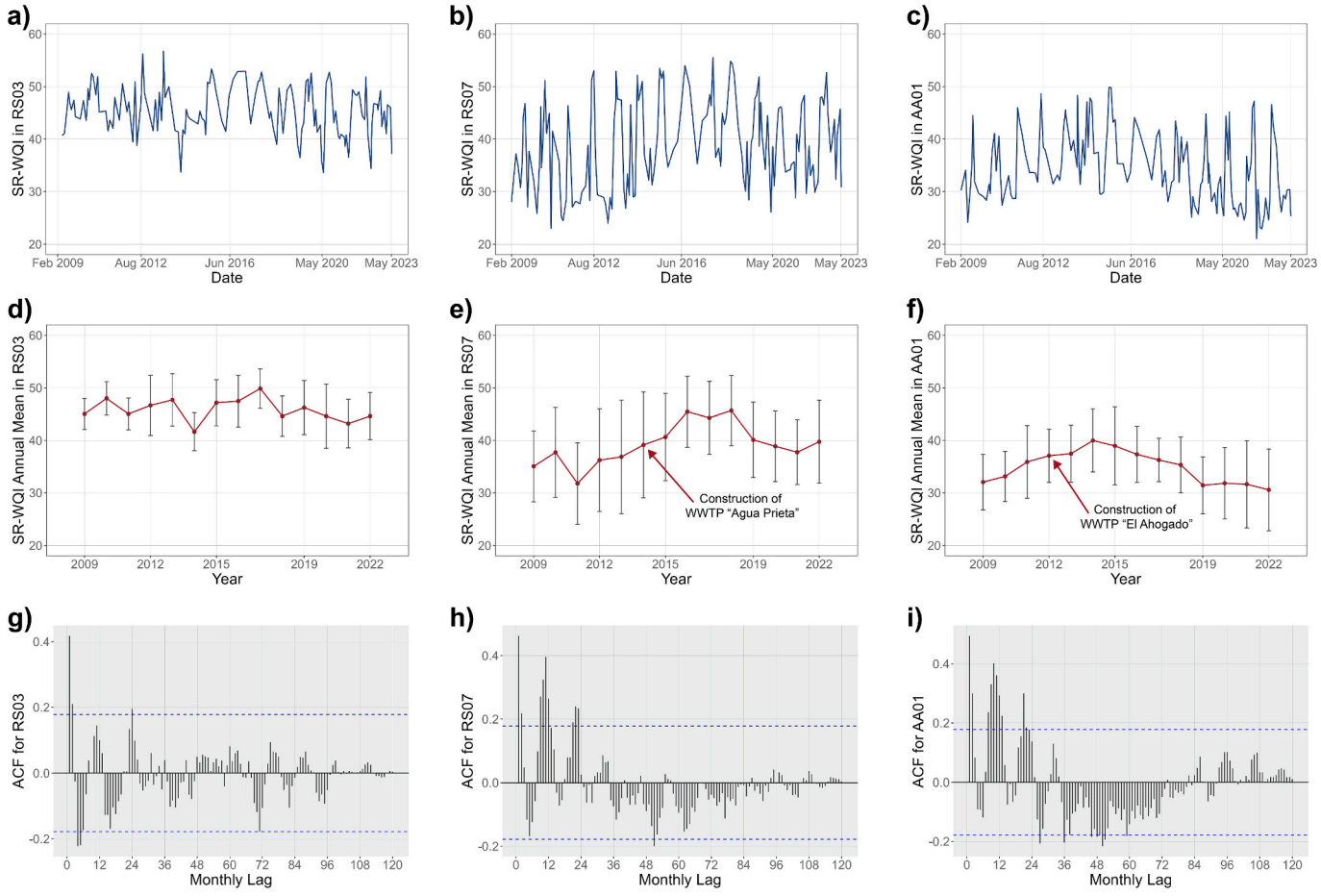


Fig. 5. SR-WQI time series of three monitoring sites a) RS03, b) RS07 and c) AA01, annual mean and standard deviation of SR-WQI for the monitoring sites d) RS03, e) RS07 and f) AA01 and autocorrelation function for g) RS03, h) RS07 and i) AA01 data.

the Best Training Subset (BTS), covering years from 2009 to 2015 (n = 409). The full dataset (13 MS) included water quality data for the same period (n = 894). For the validation step, the complete set of sites covered data from 2016 to 2020 (n = 657), while the period from 2021 to 2022 and all site data were used for the testing step (n = 372). The evaluation measurements (RMSE, MSE, MAE and R^2) for ANFIS, ANN and SVM models are shown in Table 4, including calculations for each step (training, validation, and testing). The evaluation measurements were compared between models trained with BTS and the full dataset. For ANFIS models, all evaluation measurements were significantly higher when full dataset was used for training compared to BTS. In contrast, ANN and SVM models trained with the full dataset and BTS resulted in a slight improvement when the full dataset was used. ANN presented good prediction for training and validation steps, but for testing, ANN performance dropped. The best model performance was obtained by ANFIS trained with BTS and SVM trained with full dataset; however, only a small decrease was observed in SVM trained with BTS.

In Fig. 6 the predicted values from both ANFIS models are compared to the real SR-WQI calculation. Since BTS-ANFIS produced better results, it was used to compare predictions against each individual site data. The resulting RMSE is displayed in Table 5. The RMSE results for sites included in the BTS-ANFIS model were similar to those not included; however, the prediction performance of the SR-WQI at RS10 was slightly lower.

Table 4
Comparison of evaluation measurements of machine learning models trained with full dataset and the Best Training Subset.

		ANFIS using BTS	ANFIS (full dataset)	ANN using BTS	ANN (full dataset)	SVM using BTS	SVM (full dataset)
	Training	2.02	34.26	2.16	1.76	2.21	1.79
RMSE	**Validation**	3.67	62.55	3.45	2.65	3.82	3.49
	Testing	4.57	79.83	34.01	22.56	5.09	4.69
	Training	4.10	1173.96	4.67	3.10	4.87	3.19
MSE	**Validation**	13.52	3912.16	11.88	7.00	14.63	12.21
	Testing	20.97	6372.99	1156.39	508.95	25.95	21.98
	Training	1.61	20.08	1.57	1.38	1.64	1.31
MAE	**Validation**	2.77	31.99	2.46	1.95	2.91	2.57
	Testing	3.38	33.05	8.59	4.51	4.02	3.66
	Training	0.94	0.05	0.92	0.95	0.92	0.95
R^2	**Validation**	0.80	0.01	0.82	0.89	0.77	0.81
	Testing	0.73	0.01	0.02	0.08	0.66	0.71

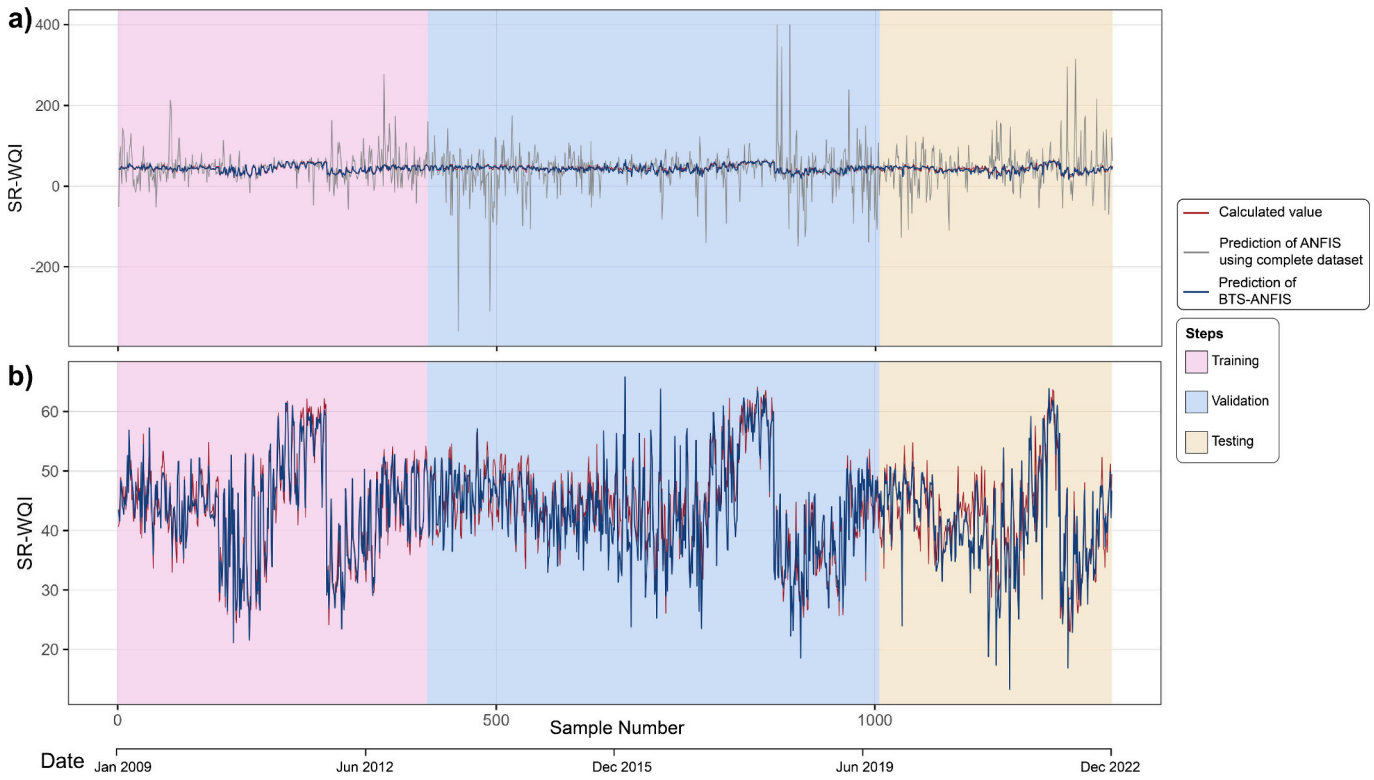


Fig. 6. Data fitting of SR-WQI prediction from a) ANFIS model developed using the complete dataset and b) BTS-ANFIS against SR-WQI real value.

Table 5
RMSE of BTS-ANFIS model tested against individual data from each monitoring site.

Included in the BTS							
Monitoring Station	AA01	RZ01	RS03	RS04	RS07	RS09	
Testing RMSE	3.05	2.82	2.82	2.28	3.26	2.86	
Not included in the BTS							
Monitoring Station	AA02	RS01	RS02	RS05	RS06	RS08	RS10
Testing RMSE	3.77	3.05	2.95	2.79	2.9	3.79	4.98

4. Discussion

In the present work, we developed a novel methodology based on CA and TSA for selecting the most representative scenarios in a multi-site, multi-year water quality dataset. Our results demonstrate that using reduced amounts of data (Best Training Subset – BTS) for training machine learning models can produce improved predictions compared to the complete dataset.

4.1. Time Series Analysis of RS-WQI

The results illustrated that SR-WQI historical behavior can be classified into three tendencies. Most monitoring sites followed a similar trend to Tendency 1, characterized by a steady behavior with no clear trend to improve or worsen over time. While several sites presented significant variations from this tendency, it is difficult to assess the causes of these variations as the Santiago River is an extensive system influenced by environmental and geographic changes as well as anthropogenic activities. TSA of different rivers have presented similar results where WQPs display seasonal steady behavior. The most significant differences are observed in time lapses of around 5–10 years and are frequently related to human interventions and urbanization (Ghashghaie et al., 2022; Parmar and Bhardwaj, 2014, 2015). Other studies have proved that hydrologic, geographic and anthropogenic features affect aquatic pollution, causing complex longitudinal patterns (Beckers et al., 2020). For Tendency 2 and Tendency 3, the reasons affecting its historical trend are more evident, where the construction of major sanitation facilities (Fig. 5e and f) led to clear improvements in the SR-WQI annual mean in both cases. Tendency 2 was likely influenced by the presence of the Wastewater Treatment Plant (WWTP) "Agua Prieta", which is located north of MAG on the limit of the urban zone (Fig. 1). The "Agua Prieta" WWTP construction ended in 2014. This WWTP has a design capacity of 8500 L/s but currently treats only an average of 6118 L/s, indicating that its capacity has not yet been exceeded. However, the improvement in river water quality was observed only until 2018, when a decrease in the SR-WQI was observed. This reduction in water quality was possibly caused by operational failures or inadequate management, but by 2020, the water quality index remained steady for the next two years (2021,2022), suggesting that some corrective actions were made. The influence of "Agua Prieta" WWTP was observed in all the monitoring sites downstream of the plant except for RS09. The site RS09 did not show the characteristic water quality improvement of Tendency 2, thus indicating that the water quality in this location is affected by an external factor, probably runoff or illegal discharges. Similar studies have demonstrated that significant river water quality changes are related to wastewater treatment facilities and their long-term capability to withstand the increases in wastewater loads caused by urban development and population growth (Barrenha et al., 2017; Qin et al., 2014; Zhang et al., 2015).

Tendency 3 was influenced by WWTP "El Ahogado". This WWTP was constructed in 2012 and is probably why water quality improved in "El Ahogado" stream in 2013–2014; however, since then, the water quality in the zone has been dropping. The "El Ahogado" WWTP has a design treatment capacity of 2500 L/s; however, the plant has received an average of 3947 L/s in recent years. Accordingly, the capacity of this plant is exceeded by around 58%, and consequently, wastewater is not treated efficiently. Events of discharges of untreated wastewater have

been reported, causing a reduction in the water quality of the "El Ahogado" stream. The poor capacity of WWTP to withstand overflows is a common situation in many regions around the world, which is directly affecting the water quality of receiving water bodies (Boënne et al., 2014; Owolabi et al., 2022; Teklehaimanot et al., 2015).

In both cases ("El Ahogado" and "Agua Prieta"), the improvement in SR-WQI annual mean after WWTP construction was around five units. The total inversion cost of "Agua Prieta" WWTP was calculated as USD $94 million in 2004 (Comisión Estatal de Agua y Saneamiento Gobierno del Estado de Jalisco, 2004), while for "El Ahogado" WWTP, the total inversion cost was calculated as around USD $335 million in 2008 (Comisión Estatal del Agua, Gobierno del Estado de Jalisco, 2008). Considering Mexican inflation, the cost of both plants with the current exchange rate is around USD $861 million. This inversion cost was necessary to improve WQI by five units in five of the 13 monitoring stations. Using this cost as a reference, we estimated the inversion cost necessary to improve the SR-WQI to a good classification (which means reaching a value of at least 70) in all the 13 monitoring sites of the network starting from the 2022 annual mean. This cost was estimated to be approximately USD $12.57 billion.

While water quality in the SR is clearly affected significantly by anthropogenic inputs, changes in water quality following actions taken to improve water quality would not be immediate. Ecosystem restoration can be slow after corrective measures are implemented, with a recent study from China illustrating mixed levels of recovery among biotic and abiotic indicators even after ten years (Fu et al., 2021). Since the beginning of the water quality monitoring period in the SR, the local government has taken action to mitigate the contamination problem, including the construction of sanitation infrastructure. As shown by the time series analysis, in the last ten years, the water quality has been stable in most of the sites, indicating that these actions have avoided a general decrease in water quality for most sites; however, they have not been sufficient to improve the river conditions as seen in other studies where effects of other stressors such urbanization (hydro morphology and hydrological alteration, runoff pollution) were suggested to have suppressed recovery expected from point-source pollutant restoration. Additionally, a decreasing trend can be observed in the last years of the time series, indicating that ecological degradation in RS continues. Consequently, efficient river remediation and contamination control strategies are necessary and urgent.

4.2. Machine learning models

The performance of BTS-ANFIS was significantly higher for all evaluation measurements (Table 4) than the one developed using the complete dataset (13 MS), attributed to model overtraining when thefull dataset was used for training. The CA and TSA_{WQI} are efficient techniques to identify tendencies in historical data and select those providing greater variability for predictive model training. This research has demonstrated that using reduced amounts of data (composed of the most representative scenarios) for training an ANFIS model significantly improved prediction performance. Additionally, the data reduction did not greatly affect ANN and SVM models as only a slight decrease in evaluation measurements was observed when BTS was used, even though this subset contained <50% of the full dataset ($n = 409$ vs $n = 894$). This finding is remarkable since the general approach in machine learning is that the bigger the dataset used for training, the greater the precision of the model generated (Al-Jarrah et al., 2015). Our finding provides evidence that using a larger dataset for training machine learning models does not necessarily result in better model performance because river quality data often contain reductant information that produces overtraining (Bilbao and Bilbao, 2017). Our CA and TSA results clearly demonstrate this tendency with several monitoring stations providing similar water quality measurements(Sections 3.1 and 3.2). Consequently, if data from all monitoring sites is used for training it could lead to overtraining, as was the case for the ANFIS model trained with the full dataset where significantly higher values for all evaluation measurements (RMSE, MSE, MAE and R^2) were observed. Addionally, the three models (ANFIS, ANN, and SVM) performed similar evalution measurements when were trained using the BTS, which also prove that the selection of most representative and variable scenarios for training is effective to improve machine learning models performance.

Furthermore, for ANN and SVM, similar evaluation measurements were observed between models trained with BTS and those trained with the full dataset. As the full dataset contains a highly similar behavior to the data already provided in BTS, this information is not useful to increase model accuracy. This also provides evidence that a larger dataset does not lead to a proportional accuracy increment. Moreover, our results clearly demonstrate that higher precision is achieved when predicting water quality if only the most representative scenarios are included in the dataset used for training. Previous studies for modeling water quality behavior have presented overfitting problems (Rodriguez-Perez et al., 2020).

The BTS-ANFIS efficiently predicted SR-WQI even in monitoring sites that were excluded from the model development. This finding demonstrates the strength of this methodology because these data can be considered as completely 'new' water quality information in terms of model assessment. Since the BTS-ANFIS performed strongly for these sites, it could also be anticipated to be highly transferable to additional sites monitored in the future. This advantage impacts monitoring network extensions because the BTS-ANFIS provides a good water quality prediction; thus, extensive validation and analyst interpretation may be unnecessary. This model could also be helpful for real-time monitoring since the models developed could estimate water quality immediately using sensor readings (Chowdury et al., 2019). This model is a powerful tool since the viability of having an analyst providing a real-time interpretation of water quality parameters across multiple river sites, each with many environmental parameters, is impossible.

One limitation of this approach is that the model estimation is limited to the range of historical data; consequently, for higher or lower SR-WQI the model is unlikely to provide a precise prediction. The lower prediction performance was observed in RS10. This site is far from the MAG, in a less urbanized and forested area; therefore, SR-WQI values in RS10 are higher compared with the rest of the sites. Since this site reports most of its values on the upper limit of the dataset used for training, model prediction was expected to be less accurate. However, even for this site, the prediction results are similar to previous reports.

Results displayed by the BTS-ANFIS were similar to previous reports of WQI prediction using ANFIS. Sahu et al. (2011) used ANFIS and PCA to predict the WQI of a heavily polluted groundwater zone adjacent to mines. In this model, the authors used an initial dataset of twelve WQPs; after PCA, the resulting eight principal components were used for model development, displaying an average absolute percentage relative error of 7.31 for training and 9.33 for testing. Sahoo et al. (2015) used a similar approach to predict the WQI of River Brahmani, India. The model was developed from an initial dataset of eleven WQPs and reduced to four principal components after PCA. This model displayed a mean absolute percentage error of 0.37 and 1.09 for training and testing data, respectively. Yan et al. (2010) also developed an ANFIS model to predict water quality status using data from 100 water quality monitoring stations located along almost all the major river basins in China. The training dataset included three WQPs, and the resulting RMSE varied from 0.3338 to 2.5564 and 0.3704 to 2.4341 for training and testing data, respectively. Similar results have been reported for the other models, where SVM outperformed ANN for water quality prediction (Haghiabi et al., 2018; Liu and Lu, 2014). Notably, in most previous reports, few input variables were used compared to our study. The complexity of a system depends on the number of inputs used to model it. As the number of inputs increases, it becomes difficult to estimate the output precisely. The high number of input variables used for the models developed could cause an accuracy decrease in model prediction. Therefore, WQIs have been assessed with other modeling approaches

but using a similar number of inputs, resulting in similar model efficiency. Intelligent models based on extreme learning machine were proposed to predict WQI at the Kinta River basin using seven WQPs as inputs, obtaining a RMSE = 3.606 for training and RMSE = 3.816 for testing (Abba et al., 2020). The efficiency of the machine learning models developed in this research can be attributed to the precise data selection for model training using clusters and time series analysis.

5. Conclusions

The performance of BTS-ANFIS (six sites) was significantly higher than that of the model trained with a complete dataset (13 sites), while in the case of ANN and SVM similar results were observed for BTS and full dataset. The approach developed in this research efficiently selected the most variable and representative data subset. This methodology proved effective in improving the machine learning models performance and provided evidence that using a larger dataset for training does not necessarily result in enhanced model performance. The selection of the most variable data representing all possible states of the system is of greater importance when aiming to maximize model precision. Machine learning models present a promising option to extend the monitoring network to new sites and perform real-time monitoring since they can be developed to provide an immediate evaluation of water quality. Additionally, this approach could be replicated using data from different water bodies to produce regional or even national models.

The analysis proposed was helpful for studying the water quality behavior in Santiago River, providing evidence of improvements and drops caused by, for example, sanitation infrastructure construction or exceedance of wastewater treatment capacity. On the other hand, the time series analysis opens a framework for generating artificial data beyond the limits of historical reports. Developing synthetic data scenarios will help predict water quality behavior in response to future contamination control strategies, such as the construction of wastewater treatment plants or more restrictive wastewater discharge policies.

CRediT authorship contribution statement

Alberto Fernández del Castillo: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Marycarmen Verduzco Garibay:** Formal analysis, Data curation. **Diego Díaz-Vázquez:** Methodology, Data curation. **Carlos Yebra-Montes:** Formal analysis, Data curation. **Lee E. Brown:** Writing – review & editing, Methodology, Formal analysis, Data curation. **Andrew Johnson:** Writing – review & editing, Methodology. **Alejandro Garcia-Gonzalez:** Writing – review & editing, Supervision, Project administration, Investigation, Formal analysis, Data curation, Conceptualization. **Misael Sebastián Gradilla-Hernández:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This research was supported by the “Challenge-Based Research Funding Program” from the Tecnológico de Monterrey (grant ID: E049 – EIC - GI01 - A-T9 - D) "Assessing the ecological and public health risks caused by the presence of persistent and emerging pollutants in Santiago River sediments". The authors thank the “Comisión Estatal del Agua de Jalisco” who kindly provided the water quality data used in this research, and the Tecnológico de Monterrey campus Guadalajara that provided installations and equipment used in this research.

Appendix A. List of 42 Water Quality Parameters and its units

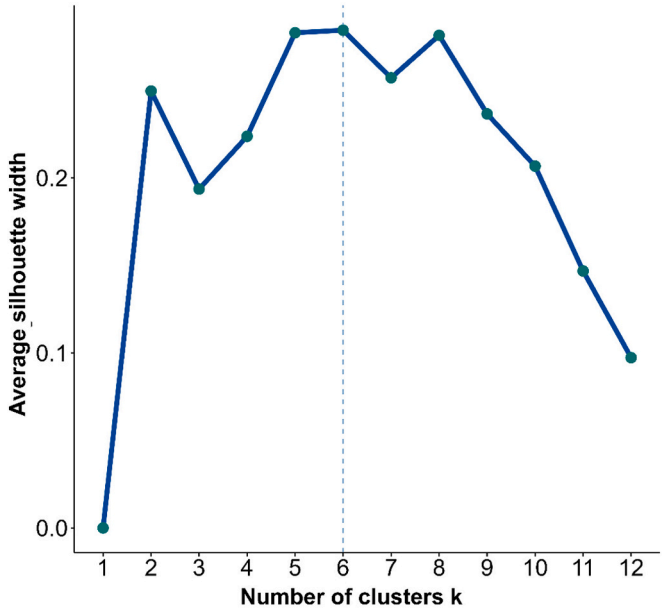
Abbreviation	Parameter	Units
Al	Aluminium	mg/L
ALK	Total alkalinity	mg CaCO3
As	Arsenic	mg/L
Ba	Barium	mg/L
Cd	Cadmium	mg/L
COND	Electric conductivity	µS/cm
Cr	Cromium	mg/L
Cu	Copper	mg/L
DBO5	Biological Oxigen Demand	mg/L
DO	Dissolved Oxygen	mg/L
DQO	Chemical Oxygen Demand	mg/L
FC	Fecal Coliforms	MPN/100 mL
Fe	Iron	mg/L
FL	Fluorides	mg/L
FOG	Fats, oils and grease	mg/L
Hg	Mercury	mg/L
Mn	Manganese	mg/L
Na	Sodium	mg/L
NH3	Ammoniacal nitrogen	mg/L
Ni	Nickel	mg/L
NKJ	Total Kjeldahl Nitrogen	mg/L
NO2	Nitrogen from Nitrites	mg/L
NO3	Nitrogen from Nitrates	mg/L
Pb	Lead	mg/L
pH	Hydrogen potential	
SAAM	Methylene Blue Active Substances	mg/L
SO4	Sulfates	mg/L
SS	Settleable solids	mg/L
SST	Total suspended solids	mg/L

(*continued on next page*)

(*continued*)

Abbreviation	Parameter	Units
SULF	Sulfides	mg/L
TC	Total coliforms	MPN/100 mL
TCL	Total chlorides	mg/L
TDS	Total dissolved solids	mg/L
TEMP	Temperature	°C
TEMPA	Environmental temperature	°C
TH	Total Hardness	mg CaCO3
TN	Total Nitrogen	mg/L
TP	Total Phosphorus	mg/L
TS	Total Solids	mg/L
TURB	Turbidity	NTU
Zn	Zinc	mg/L
DTEMP	Temperature Difference	°C

Appendix B. Average Silhouette Index for Water Quality dataset



References

Abba, S.I., Hadi, S.J., Sammen, S.Sh., Salih, S.Q., Abdulkadir, R.A., Pham, Q.B., Yaseen, Z.M., 2020. Evolutionary computational intelligence algorithm coupled with self-tuning predictive model for water quality index determination. J. Hydrol. 587, 124974 https://doi.org/10.1016/j.jhydrol.2020.124974.

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A.A., Irfan, R., García-Nieto, J., 2019. Efficient water quality prediction using supervised machine learning. Water 11, 2210. https://doi.org/10.3390/w11112210.

Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K., Taha, K., 2015. Efficient Machine Learning for Big Data: A Review. Big Data Res., Big Data, Analytics, and High-Performance Computing, 2, pp. 87–93. https://doi.org/10.1016/j.bdr.2015.04.001.

Arora, S., Keshari, A.K., 2021. ANFIS-ARIMA modelling for scheming re-aeration of hydrologically altered rivers. J. Hydrol. 601, 126635 https://doi.org/10.1016/j.jhydrol.2021.126635.

Banadkooki, F.B., Ehteram, M., Panahi, F., Sh. Sammen, S., Othman, F.B., EL-Shafie, A., 2020. Estimation of total dissolved solids (TDS) using new hybrid machine learning models. J. Hydrol. 587, 124989 https://doi.org/10.1016/j.jhydrol.2020.124989.

Barrenha, P.I.I., Tanaka, M.O., Hanai, F.Y., Pantano, G., Moraes, G.H., Xavier, C., Awan, A.T., Grosseli, G.M., Fadini, P.S., Mozeto, A.A., 2017. Multivariate analyses of the effect of an urban wastewater treatment plant on spatial and temporal variation of water quality and nutrient distribution of a tropical mid-order river. Environ. Monit. Assess. 190, 43. https://doi.org/10.1007/s10661-017-6386-4.

Beckers, L.-M., Brack, W., Dann, J.P., Krauss, M., Müller, E., Schulze, T., 2020. Unraveling longitudinal pollution patterns of organic micropollutants in a river by non-target screening and cluster analysis. Sci. Total Environ. 727, 138388 https://doi.org/10.1016/j.scitotenv.2020.138388.

Bilbao, I., Bilbao, J., 2017. Overfitting problem and the over-training in the era of data: Particularly for Artificial Neural Networks. In: 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS). Presented at the 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS), pp. 173–177. https://doi.org/10.1109/INTELCIS.2017.8260032.

Boënne, W., Desmet, N., Looy, S.V., Seuntjens, P., 2014. Use of online water quality monitoring for assessing the effects of WWTP overflows in rivers. Environ Sci Process Impacts 16, 1510–1518. https://doi.org/10.1039/C3EM00449J.

Casillas-García, L.F., de Anda, J., Yebra-Montes, C., Shear, H., Díaz-Vázquez, D., Gradilla-Hernández, M.S., 2021. Development of a specific water quality index for the protection of aquatic life of a highly polluted urban river. Ecol. Indic. 129, 107899 https://doi.org/10.1016/j.ecolind.2021.107899.

Chen, L., Wu, T., Wang, Z., Lin, X., Cai, Y., 2023. A novel hybrid BPNN model based on adaptive evolutionary Artificial Bee Colony Algorithm for water quality index prediction. Ecol. Indic. 146, 109882 https://doi.org/10.1016/j.ecolind.2023.109882.

Chowdury, M.S.U., Emran, T.B., Ghosh, S., Pathak, A., Alam, Mohd M., Absar, N., Andersson, K., Hossain, M.S., 2019. IoT based Real-Time River water quality monitoring system. In: Procedia Comput. Sci., The 16th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2019),The 14th International Conference on Future Networks and Communications (FNC-2019),The 9th International Conference on Sustainable Energy Information Technology, 155, pp. 161–168. https://doi.org/10.1016/j.procs.2019.08.025.

Comisión Estatal de Agua y Saneamiento Gobierno del Estado de Jalisco, 2004. Manifestación de Impacto Ambiental Modalidad Particular Proyectos Hidráulicos para el Proyecto: Planta de Tratamiento de Aguas Residuales Municipales "Agua Prieta".

Comisión Estatal del Agua, Gobierno del Estado de Jalisco, 2008. Manifestación de Impacto Ambiental Modalidad Particular Proyectos Hidráulicos para el Proyecto: Planta de Tratamiento de Aguas Residuales de la Cuenca del Ahogado y sus Obras Asociadas.

Dewanti, N.A., Abadi, A.M., 2019. Fuzzy logic application as a tool for classifying water quality status in Gajahwong River, Yogyakarta, Indonesia. IOP Conf. Ser. Mater. Sci. Eng. 546, 032005 https://doi.org/10.1088/1757-899X/546/3/032005.

Dincer, N.G., Yalçin, M.O., 2016. Revealing information and equipment redundancies in air pollution monitoring networks in Turkey. Int. J. Environ. Sci. Technol. 13, 2927–2938. https://doi.org/10.1007/s13762-016-1118-9.

Duan, W., He, B., Nover, D., Yang, G., Chen, W., Meng, H., Zou, S., Liu, C., 2016. Water quality assessment and pollution source identification of the eastern Poyang Lake Basin using multivariate statistical methods. Sustainability 8, 133. https://doi.org/10.3390/su8020133.

Elhag, M., Gitas, I., Othman, A., Bahrawi, J., Psilovikos, A., Al-Amri, N., 2021. Time series analysis of remotely sensed water quality parameters in arid environments, Saudi Arabia. Environ. Dev. Sustain. 23, 1392–1410. https://doi.org/10.1007/s10668-020-00626-z.

Ellina, G., Papaschinopoulos, G., Papadopoulos, B.K., 2020. Research of fuzzy implications via fuzzy linear regression in data analysis for a fuzzy model. J. Comput. Methods Sci. Eng. 20, 879–888. https://doi.org/10.3233/JCM-194015.

Fernández del Castillo, A., Yebra-Montes, C., Verduzco Garibay, M., de Anda, J., Garcia-Gonzalez, A., Gradilla-Hernández, M.S., 2022. Simple prediction of an ecosystem-specific water quality index and the water quality classification of a highly polluted river through supervised machine learning. Water 14, 1235. https://doi.org/10.3390/w14081235.

Fu, H., Gaüzère, P., García Molinos, J., Zhang, P., Zhang, H., Zhang, M., Niu, Y., Yu, H., Brown, L.E., Xu, J., 2021. Mitigation of urbanization effects on aquatic ecosystems by synchronous ecological restoration. Water Res. 204, 117587 https://doi.org/10.1016/j.watres.2021.117587.

Ghashghaie, M., Eslami, H., Ostad-Ali-Askari, K., 2022. Applications of time series analysis to investigate components of Madiyan-rood river water quality. Appl Water Sci 12, 202. https://doi.org/10.1007/s13201-022-01693-5.

Gitau, M.W., Chen, J., Ma, Z., 2016. Water quality indices as tools for decision making and management. Water Resour. Manag. 30, 2591–2610. https://doi.org/10.1007/s11269-016-1311-0.

Haghiabi, A.H., Nasrolahi, A.H., Parsaie, A., 2018. Water quality prediction using machine learning methods. Water Qual. Res. J. 53, 3–13. https://doi.org/10.2166/wqrj.2018.025.

Ho, J.Y., Afan, H.A., El-Shafie, A.H., Koting, S.B., Mohd, N.S., Jaafar, W.Z.B., Lai Sai, H., Malek, M.A., Ahmed, A.N., Mohtar, W.H.M.W., Elshorbagy, A., El-Shafie, A., 2019. Towards a time and cost effective approach to water quality index class prediction. J. Hydrol. 575, 148–165. https://doi.org/10.1016/j.jhydrol.2019.05.016.

INEGI, 2020. Información poblacional [WWW Document]. Demogr. Jalisco. URL. https://cuentame.inegi.org.mx/monografias/informacion/jal/poblacion/ (accessed 1.9.23).

Jang, J.-S.R., 1994. Structure determination in fuzzy modeling: a fuzzy CART approach. In: Proceedings of 1994 IEEE 3rd International Fuzzy Systems Conference. Presented at the Proceedings of 1994 IEEE 3rd International Fuzzy Systems Conference, vol.1, pp. 480–485. https://doi.org/10.1109/FUZZY.1994.343738.

Jiang, X., Adeli, H., 2004. Wavelet packet-autocorrelation function method for traffic flow pattern analysis. Comput.-Aided Civ. Infrastruct. Eng. 19, 324–337. https://doi.org/10.1111/j.1467-8667.2004.00360.x.

Kassambara, A., 2017. Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning. STHDA.

Li, Y., Fang, L., Yuanzhu, W., Mi, W., Ji, L., Guixiang, Z., Yang, P., Chen, Z., Bi, Y., 2022. Anthropogenic activities accelerated the evolution of river trophic status. Ecol. Indic. 136, 108584 https://doi.org/10.1016/j.ecolind.2022.108584.

Liu, M., Lu, J., 2014. Support vector machine—an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river? Environ. Sci. Pollut. Res. 21, 11036–11053. https://doi.org/10.1007/s11356-014-3046-x.

Lobato, T.C., Hauser-Davis, R.A., Oliveira, T.F., Silveira, A.M., Silva, H.A.N., Tavares, M. R.M., Saraiva, A.C.F., 2015. Construction of a novel water quality index and quality indicator for reservoir water quality evaluation: a case study in the Amazon region. J. Hydrol. 522, 674–683. https://doi.org/10.1016/j.jhydrol.2015.01.021.

McCulligh, C., Vega Fregoso, G., 2019. Defiance from down river: deflection and dispute in the urban-industrial metabolism of pollution in Guadalajara. Sustainability 11, 6294. https://doi.org/10.3390/su11226294.

McDonald, R.I., Weber, K.F., Padowski, J., Boucher, T., Shemie, D., 2016. Estimating watershed degradation over the last century and its impact on water-treatment costs for the world's large cities. Proc. Natl. Acad. Sci. 113, 9117–9122. https://doi.org/10.1073/pnas.1605354113.

Muharemi, F., Logofătu, D., Leon, F., 2019. Machine learning approaches for anomaly detection of water quality on a real-world data set*. J. Inf. Telecommun. 3, 294–307. https://doi.org/10.1080/24751839.2019.1565653.

Murtagh, F., Contreras, P., 2012. Algorithms for hierarchical clustering: an overview. WIREs Data Min. Knowl. Discov. 2, 86–97. https://doi.org/10.1002/widm.53.

Owolabi, T.A., Mohandes, S.R., Zayed, T., 2022. Investigating the impact of sewer overflow on the environment: a comprehensive literature review paper. J. Environ. Manag. 301, 113810 https://doi.org/10.1016/j.jenvman.2021.113810.

Parmar, K.S., Bhardwaj, R., 2014. Water quality management using statistical analysis and time-series prediction model. Appl Water Sci 4, 425–434. https://doi.org/10.1007/s13201-014-0159-9.

Parmar, K.S., Bhardwaj, R., 2015. Statistical, time series, and fractal analysis of full stretch of river Yamuna (India) for water quality management. Environ. Sci. Pollut. Res. 22, 397–414. https://doi.org/10.1007/s11356-014-3346-1.

Patle, G.T., Singh, D.K., Sarangi, A., Rai, A., Khanna, M., Sahoo, R.N., 2015. Time series analysis of groundwater levels and projection of future trend. J. Geol. Soc. India 85, 232–242. https://doi.org/10.1007/s12594-015-0209-4.

Qin, H., Su, Q., Khu, S.-T., Tang, N., 2014. Water quality changes during rapid urbanization in the Shenzhen River catchment: an integrated view of socio-economic and infrastructure development. Sustainability 6, 7433–7451. https://doi.org/10.3390/su6107433.

Rajaee, T., Khani, S., Ravansalar, M., 2020. Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: a review. Chemom. Intell. Lab. Syst. 200, 103978 https://doi.org/10.1016/j.chemolab.2020.103978.

Raman, B.V., Bouwmeester, R., Mohan, S., 2009. Fuzzy logic water quality index and importance of water quality parameters. Air Soil Water Res. 2 https://doi.org/10.4137/ASWR.S2156. ASWR.S2156.

Rice, E.W., Baird, R.B., Eaton, A.D., Clesceri, L.S., et al., 2012. Standard Methods for the Examination of Water and Wastewater. American Public Health Association Washington, DC.

Rizo-Decelis, L.D., Andreo, B., 2016. Water quality assessment of the Santiago River and attenuation capacity of pollutants downstream Guadalajara City, Mexico. River Research and Applications 32 (7), 1505–1516.

Rodriguez-Perez, J., Leigh, C., Liquet, B., Kermorvant, C., Peterson, E., Sous, D., Mengersen, K., 2020. Detecting technical anomalies in high-frequency water-quality data using artificial neural networks. Environ. Sci. Technol. 54, 13719–13730. https://doi.org/10.1021/acs.est.0c04069.

Sahoo, M.M., Patra, K.C., Khatua, K.K., 2015. Inference of water quality index using ANFIA and PCA. Aquat. Procedia, International Conference on Water Resources, Coastal and Ocean Engineering (ICWRCOE'15) 4, pp. 1099–1106. https://doi.org/10.1016/j.aqpro.2015.02.139.

Sahu, M., Mahapatra, S.S., Sahu, H.B., Patel, R.K., 2011. Prediction of water quality index using neuro fuzzy inference system. Water Qual Expo Health 3, 175–191. https://doi.org/10.1007/s12403-011-0054-7.

Shil, S., Singh, U.K., Mehta, P., 2019. Water quality assessment of a tropical river using water quality index (WQI), multivariate statistical techniques and GIS. Appl Water Sci 9, 168. https://doi.org/10.1007/s13201-019-1045-2.

Sjöberg, J., Ljung, L., 1995. Overtraining, regularization and searching for a minimum, with application to neural networks. Int. J. Control. 62, 1391–1407. https://doi.org/10.1080/00207179508921605.

Teklehaimanot, G.Z., Kamika, I., Coetzee, M.A.A., Momba, M.N.B., 2015. Population growth and its impact on the design capacity and performance of the wastewater treatment plants in Sedibeng and Soshanguve, South Africa. Environ. Manag. 56, 984–997. https://doi.org/10.1007/s00267-015-0564-3.

Wang, Z., Wang, Q., Wu, T., 2023. A novel hybrid model for water quality prediction based on VMD and IGOA optimized for LSTM. Front. Environ. Sci. Eng. 17, 88. https://doi.org/10.1007/s11783-023-1688-y.

Yan, H., Zou, Z., Wang, H., 2010. Adaptive neuro fuzzy inference system for classification of water quality status. J. Environ. Sci. 22, 1891–1896. https://doi.org/10.1016/S1001-0742(09)60335-1.

Zhang, D., Tao, Y., Liu, X., Zhou, K., Yuan, Z., Wu, Q., Zhang, X., 2015. Spatial and temporal variations of water quality in an artificial urban river receiving WWTP effluent in South China. Water Sci. Technol. 73, 1243–1252. https://doi.org/10.2166/wst.2015.592.

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B., Ye, L., 2022. A review of the application of machine learning in water quality evaluation. Eco-Environ. Health 1, 107–116. https://doi.org/10.1016/j.eehl.2022.06.001.