*K*irkuk University Journal for Agricultural Sciences
ISSN:2958-6585
https://kujas.uokirkuk.edu.iq

RESEARCH ARTICLE https://doi.org. 10.58928/ku25.16318

Machine Learning-Based Prediction of River Water Quality Using LSTM and RF Models with PCA and Stepwise Regression for Dimensionality Reduction: A Case Study of the Maroon River Basin.

Ahmad Majeed Muhammad Daloye[1] **Falah Hama Faraj Ali**[2]
Salim Neimat Azeez[3] **Hiwa Faraj Pana**[4]

[1]Surveying Department, Kalar Technical Institute, Garmian Polytechnic University,KGR, IRAQ,
[2]Surveying Department, Darbandikhan Technical Institute, Sulaimani Polytechnic University, Sulaimani, KGR, IRAQ,
[3]Protected Cultivation Department , Bakrajo Technical Institute, Sulaimani Polytechnic University, Sulaimani, KGR, IRAQ,
Civil Engineering and Architecture Faculty, Shahid Chamran University of Ahvaz, Ahvaz , IRAN

.*Corresponding Author: ahmad.majeed@gpu.edu.iq.

Received: 25/06/2025 **Revised:11/08/2025** **Accepted: 19/08/2025** **Published: 01/09/2025**

ABSTRACT

Monitoring and predicting river water quality is crucial for urban water management, agriculture, and environmental sustainability, especially in hot and arid regions. The presence of dams along the river course can significantly alter water quality by affecting flow regimes and salt accumulation, potentially leading to increased salinity and other related problems. However, such changes can be effectively managed through accurate modeling and forecasting. This study assesses the performance of two machine learning models, Random Forest (RF) and Long Short-Term Memory (LSTM), in predicting electrical conductivity (EC) and sodium absorption ratio (SAR) within the Maroon River Basin, Iran. Principal Component Analysis (PCA) and stepwise regression were employed to reduce input dimensionality and enhance model efficiency. Results indicate that the LSTM model consistently outperforms RF at both Idank station (upstream of Maroon Dam) and Tang-e-Tekab station (downstream of Maroon Dam) for both parameters, particularly in SAR prediction. At Idank station, the LSTM model combined with stepwise regression achieved the highest accuracy for EC prediction, with an R^2 of 0.96, RMSE of 61.56, and KGE of 0.96 on the test dataset. For SAR at the same station, LSTM again demonstrated exceptional performance, attaining an R^2 of 0.99, RMSE of 0.08, and KGE of 0.99. At Tang-e-Tekab station, LSTM with PCA yielded the most precise EC predictions (R^2 = 0.96, RMSE = 76.60, KGE = 0.97). Similarly, the best SAR predictions at this station were obtained using LSTM with PCA (R^2 = 0.96, RMSE = 0.18, KGE = 0.95). These findings underscore the effectiveness of combining LSTM networks with tailored input selection techniques based on site-specific conditions, highlighting their potential application in water resource decision support systems. Overall, this study demonstrates that although dam operations influence water quality, such impacts can be successfully managed through advanced predictive modeling to facilitate sustainable water resources management.

***Keywords*: **Maroon River, Water quality, EC, SAR, RF, LSTM, Dimensionality reduction

Introduction

Rivers are among the most vital sources of freshwater for both natural ecosystems and human societies. They play a critical role not only in supplying water for agricultural, industrial, and urban uses but also in maintaining ecological balance and ecosystem sustainability. However, increasing population growth and the expansion of anthropogenic activities have significantly impacted river water quality. Agricultural runoff, urban and industrial wastewater discharge, and land-use changes are among the significant contributing factors. These processes can lead to increased salinity and alterations in the ionic composition of river water, which may adversely affect human health, agricultural productivity, and environmental integrity [1].

Effective water resource management requires a comprehensive understanding of water quality variations and the ability to forecast future trends. In this regard, modeling water quality parameters such as Electrical Conductivity (EC) and Sodium absorption ratio (SAR) is of particular importance. These parameters are recognized as key indicators for assessing salinity and overall water quality. Fluctuations in EC and SAR values can signal the intrusion of pollutants or changes in water sources. For instance, elevated levels of EC and SAR may indicate the presence of dissolved salts in the river, often due to agricultural runoff or industrial effluents [2].

Numerous studies have shown that modeling EC based on the analysis of central cation and anion concentrations can contribute to understanding the chemical behavior of river systems and evaluating the influence of pollution sources. By examining the relationships between dissolved ions such as sodium (Na^+), chloride (Cl^-), and other salinity-related constituents, it is possible to assess their impact on key water quality parameters like EC and SAR [3]. For example, during the planning and siting of new industrial units, predictive models can be used to estimate potential pollutant loads to rivers. In such cases, data-driven models for estimating EC and SAR enable decision-makers to anticipate the consequences of human developments on water salinity and to implement timely management strategies [4].

The importance of this issue is especially pronounced in agricultural areas such as the Maroon River Basin, where the production of strategic crops like wheat, sugarcane, rice, and dates is affected by water quality. Increased salinity not only reduces crop yield and harms salt-sensitive plants but may also lead to the collapse of local ecosystems.

In this context, modeling water quality parameters under future urban, industrial, and agricultural development scenarios serves as an effective tool for predicting environmental changes without requiring frequent field measurements. Such modeling can function as a decision support system for water resource management and aid in the development of sustainable policies.

Data mining and artificial intelligence techniques provide powerful tools for predicting and managing water quality parameters. The Random Forest (RF) model, as one of the machine learning algorithms, has demonstrated strong performance in handling complex and nonlinear datasets and has been widely applied to predict water quality parameters [5,6]. Additionally, the Long Short-Term Memory (LSTM) neural network, due to its memory-based architecture, has shown high effectiveness in modeling time series and forecasting temporally dependent variables. In recent years, LSTM has been extensively used for predicting water quality parameters such as EC, dissolved oxygen (DO), and biochemical oxygen demand (BOD) [7].

Given the significance of this issue, several studies have been conducted to address it.

For instance, Adib et al. (2020) estimated total dissolved solids (TDS), electrical conductivity (EC), and total hardness (TH) in the Sepidrood River using ANFIS, Gene Expression Programming (GEP), and Least-Squares Support Vector Machine (LS-SVM) models. Their findings revealed that LS-SVM yielded the highest accuracy for TDS, GEP for EC, and ANFIS for TH, underlining the power of intelligent systems for water quality forecasting [3].

Ubah et al. (2021) applied Artificial Neural Networks (ANNs) to forecast major irrigation-related parameters (pH, TDS, EC, and Na^+) in the Ele River. The models achieved high accuracy ($R^2 > 0.95$), capturing seasonal dynamics and indicating exceedance of FAO thresholds during dry seasons. The study confirmed ANN's effectiveness in water quality management [8].

Nouraki et al. (2021) employed Multiple Linear Regression (MLR), M5P, Support Vector Regression (SVR), and Random Forest Regression (RFR) to model TDS, SAR, and TH in the Karun River (1999–2019), using PCA for input variable selection. The best performance was observed with RFR (TDS), SVR (SAR), and MLR (TH), demonstrating ML capabilities under limited data conditions [9].

Trach et al. (2022) assessed and forecasted the Water Quality Index (WQI) in Ukrainian rivers using fuzzy logic and ANN models. A modified WQI framework emphasized hazardous chemicals, with the optimal ANN (Softmax activation and Adam optimizer) achieving $R^2 = 0.964$ and MAPE = 9.6%, confirming its predictive power [10].

Adib et al. (2022) compared MLP, RBF, ANFIS, LS-SVM, and GEP for estimating water quality at Pol-e-Astaneh station. LS-SVM, MLP, and GEP achieved the best accuracy for various parameters. Notably, models incorporating lag time underperformed, suggesting immediate input features are more informative [11].

Ibrahim et al. (2023) integrated PCA and ANN to model water quality in Malaysia. PCA effectively identified pollution sources, while ANN achieved near-perfect prediction of WQI ($R^2 = 0.9999$), supporting the use of hybrid models for environmental decision-making [12].

Adjovu et al. (2023) estimated TDS concentration in Lake Mead using EC and temperature through various machine learning models. They found that both simple models, such as linear regression, and advanced ensemble methods, like XGBoost and GBM, achieved high prediction accuracy. The study demonstrated the effectiveness of using surrogate variables and data-driven approaches for water quality monitoring [13].

Pourhosseini et al. (2023) developed hybrid SVM models optimized with metaheuristic algorithms (CA, HS, TLBO) to predict TDS in the Babolrood River, Iran. Using long-term monthly water quality data and input selection via Shannon's entropy and correlation analysis, the SVM-TLBO model achieved superior accuracy compared to baseline models, demonstrating its effectiveness in predicting TDS in river systems [14].

Pyo et al. (2023) reviewed the application of LSTM models for predicting water quality in inland environments. They highlighted LSTM's strengths in capturing temporal dependencies and discussed enhanced versions of LSTM using preprocessing, CNNs, attention mechanisms, and transfer learning. Their review confirms LSTM's robustness and adaptability for time-series water quality modelling [15].

Karbasi et al. (2024) applied a hybrid CNN-LSTM deep learning model, enhanced with Boruta-XGBoost feature selection, to forecast river EC up to 10 days ahead. Using data from two Australian rivers, the model outperformed other machine learning methods in both short- and medium-term predictions. Their results highlight the strong potential of deep learning and feature selection techniques for accurate EC forecasting in river systems [16].

Jaafer and Al-Mukhtar (2024) used ensemble learning (AdaBoost and Gradient Boosting) to predict DO and BOD in the Tigris River. Gradient Boosting achieved $R^2 > 0.99$, highlighting ensemble models' superior predictive capacity [17].

Ismail et al. (2024) developed ANN architectures (feedforward and radial basis networks) to estimate WQI in the Klang River. The best model feedforward ANN with a single hidden layer offered fast, cost-effective, and reliable prediction, serving as an early warning tool for pollution [18].

Satish et al. (2024) created a stacked ANN ensemble using geospatial, climatic, and land use data to forecast EC, BOD, nitrate, and DO in the Godavari River Basin. Ensemble methods (Bagging, Boosting) outperformed baseline FFNN models, and the stacked meta-model (XGBoost, RF, ET) achieved improved accuracy (e.g., BOD R^2 from 0.87 to 0.91) [19].

Adnan et al. (2025) proposed a hybrid ANN–Enhanced Runge Kutta (ANN-ERUN) model to predict BOD_5 in South Korea. The model achieved $R^2 = 0.857$ and RMSE = 1.24 mg/L using eight parameters. It outperformed traditional models and showed promise for water quality monitoring [20].

Khosravi et al. (2025) introduced a hybrid Bi-LSTM + BA-AMT model to forecast turbidity and DO in the Clackamas River. Bi-LSTM outperformed other methods, especially for turbidity, while BA-AMT effectively captured extreme events. The study emphasized the robustness of deep learning and metaheuristic optimization [21].

Abushandi (2025) applied ANN models to predict water quality in the Liffey (Ireland) and Andarax (Spain) rivers. The models achieved high performance (e.g., $R^2 = 0.98$ for EC), capturing complex spatiotemporal patterns and projecting notable trends, including a 20% drop in DO in the Liffey [22].

Al-Khuzaie et al. (2025) developed a GIS-integrated ANN model to predict the Heavy Metal Pollution Index (HPI) in Iraq's Euphrates River using data from 40 monitoring sites. The model attained excellent validation metrics ($R^2 = 0.999$, RSR = 1, NSE = 0.99), identifying major pollutants such as nickel and cadmium, which exceeded WHO standards [23].

Finally, Isık and Akkan (2025) modeled WQI in the Southeastern Black Sea Basin using novel ANN models Single Multiplicative Neuron (SMN), Multilayer Perceptron (MLP), and Pi-Sigma ANN (PS-ANN). Based on monthly data from eight sites, the PS-ANN and SMN models introduced showed high accuracy and ability to capture complex nonlinear dynamics [24].

Although numerous studies have applied machine learning techniques to predict water quality, most have either focused on a limited set of models or targeted specific parameters under generic conditions. Furthermore, few of them have systematically categorized the modelling approaches based on prediction targets (e.g., EC, SAR), input selection strategies (e.g., PCA, feature ranking), or spatial scenarios (e.g., upstream vs. downstream monitoring). A critical synthesis highlighting methodological gaps and model suitability under varying riverine conditions is often lacking. This study addresses these gaps by combining dimensionality reduction techniques with state-of-the-art models (RF and LSTM), while explicitly accounting for spatial variability in water quality across different monitoring stations in the Maroon River Basin.

Moreover, predictive modelling plays a crucial role when real-time observational data are unavailable due to equipment malfunction, monitoring interruptions, or communication failures. By learning from historical patterns and upstream data, reliable models can serve as effective surrogates for estimating downstream water quality parameters under such conditions. Additionally, forecasting capabilities provide early warnings for salinity risks, which is particularly valuable during sensitive agricultural periods. These practical considerations underscore the need for robust predictive tools, even when observational data are available for both intake and out-take stations.

Materials and Methods

Study Area

The Maroon River watershed, a sub-basin of the Maroon–Jarrahi river system, is located in southwestern Iran. It lies between longitudes 50°05′ to 51°11′ E and latitudes 30°39′ to 31°21′ N. The river originates in the Zagros Mountains and flows approximately 120 km before entering the Maroon Dam reservoir. With a total length of about 422 km, the Maroon River is a vital water resource for Khuzestan Province, particularly in supplying irrigation water to southeastern agricultural lands.

This river is a perennial stream with a mixed hydrological regime driven by both rainfall and snowmelt. Most precipitation occurs as rainfall at lower elevations, while snowfall is dominant at higher altitudes. Annual precipitation in the basin varies significantly, ranging from about 150 mm in the plains to over 900 mm in the northern highlands. The total watershed area is approximately 3,824 km², with elevations spanning from 240 meters at the lowest point to 3,485 meters in the mountainous areas [25]. Figure 1 shows the major rivers of the Maroon watershed, the locations of two hydrometric stations (Idanak and Tang-e-Tekab), and the Maroon Dam.

According to regional environmental reports, extensive sand and gravel extraction has led to severe ecological disturbances in the riverbed. Additionally, wastewater discharge from 15 industrial and service facilities, along with domestic sewage from urban and rural settlements, has exacerbated water quality degradation. Particularly in the downstream section between the Idanak station and the Beheshtan plain, the river flows through gypsum- and salt-rich geological formations. Combined with intense evaporation caused by high regional temperatures, these conditions significantly elevate the salinity and mineral content of the river water [26].

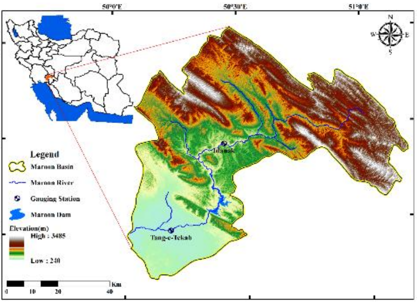


Figure 1: Maroon watershed and locations of stations and dam

Data and Preprocessing

This study focuses on modelling two critical water quality parameters, Electrical Conductivity (EC) and Sodium Adsorption Ratio (SAR), due to their importance in assessing water suitability for agriculture and environmental management. The dataset spans 21 years (1999–2019) and was collected monthly from two hydrometric stations, Idanak and Tang-e-Tekab, located within the Maroon River watershed. Parameters measured include discharge, temperature, TDS, EC, pH, $SO_4{}^{2-}$, $HCO_3{}^{-}$, Cl^{-} , Ca^{2+} , Mg^{2+} , Na^{+} , K^{+} , and SAR.

To enhance model efficiency and prevent overfitting, dimensionality reduction techniques were applied to eliminate multicollinearity and remove redundant or weakly correlated variables. Two methods—Stepwise Regression and Principal Component Analysis (PCA) were used to identify the most informative predictors for each target parameter.

Stepwise Regression

Stepwise Regression is a statistical method that iteratively adds or removes variables based on criteria such as the F-test and the coefficient of determination (R^2), resulting in a subset of predictors with the most significant influence on water quality parameters. This approach is particularly valuable in environmental modeling due to its transparency in variable selection and its direct interpretability [27].

Principal Component Analysis (PCA)

PCA is a widely used technique for dimensionality reduction, which transforms the original set of correlated variables into a new set of uncorrelated variables (principal components) that retain the most variance in the data. Components with low variance are deemed negligible, allowing for a compact representation of the dataset without significant information loss. PCA helps eliminate multicollinearity and reduce noise, thereby enhancing model performance [27].

The selected features from both methods were subsequently used to train machine learning models for predicting water quality parameters. This approach provides a unified modeling framework to compare the effectiveness of the two-dimensionality reduction techniques.

Random Forest (RF)

The Random Forest algorithm was introduced by Breiman in 2001 as an ensemble learning method designed for regression and classification problems based on the development of decision trees. A random forest is composed of an ensemble of unpruned decision trees, each generated through a recursive partitioning algorithm. In other words, the random forest combines multiple decision trees, each constructed from different self-organizing random samples of the data [28].

To build a regression tree, recursive partitioning, and multiple regression techniques are employed. The decision process at each internal node, starting from the root node, is repeated according to a tree-based rule until a predefined stopping criterion is met.

In the RF method, a random vector X_n, independent of the random vectors X_1, X_2,..., X_{n-1} is generated for the nth tree. All vectors follow the same distribution. The tree regression is computed using the training dataset and X_n, resulting in a set of n trees defined as follows [28]:

$$X_n = \{h_1(x), h_2(x), \dots, h_n(x)\} \quad (1)$$

$$h_n = h(x, X_n), x = \{x_1, x_2, \dots, x_p\} \quad (2)$$

The above p dimensional vector forms a forest, and the outputs for each tree are presented as follows:

$$\widehat{y_1} = h_1(x), \widehat{y_2} = h_2(x), \dots, \widehat{y_n} = h_n(x) \quad (3)$$

In the above equation, $\widehat{y_n}$ represents the output of the nth tree. To obtain the final output, the average of all the tree predictions is calculated.

Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are a specialized type of Recurrent Neural Network (RNN) designed to capture long-term dependencies in sequential data [29] effectively. Introduced by Hochreiter and Schmidhuber [30], LSTMs have demonstrated outstanding performance across various applications and are now widely regarded as a standard model for time-series prediction tasks.

Traditional RNNs often face challenges such as vanishing or exploding gradients during training, limiting their ability to learn long-range dependencies. LSTM overcomes these issues by incorporating a memory cell capable of preserving information over extended sequences. This feature makes LSTM particularly suitable for modeling time-series data with long lags and complex temporal dynamics.

The key component of the LSTM architecture is its gated mechanism, which controls the flow of information within the network. It consists of three main gates:

- Forget gate: decides which parts of the previous memory to discard;
- Input gate: controls how much new information should be added to the memory cell;
- Output gate: determines how much of the current memory state is passed to the next layer or time step.

These gates enable the network to selectively retain important information and discard irrelevant data over time, thereby enhancing learning efficiency and prediction accuracy. The internal operations of these gates and the update rules for the memory cell are mathematically described by Equations (4) to (8):

Input gate: $i_t = \sigma(W_l \cdot [h_{t-1}, x_t] + b_l)$ (4)

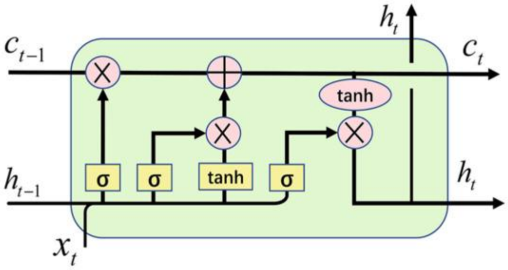
Forget gate: $f_t = \sigma\left(\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)\right), \quad \widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$ (5)

Output gate: $O_t = \sigma(W_O \cdot [h_{t-1}, x_t] + b_O)$ (6)

Long memory: $C_t = f_t * C_{t-1} + i_t * \widetilde{C_t}$ (7)

Short memory: $h_t = O_t * tanh(C_t)$ (8)

The matrix *W* denotes the weight parameters associated with the gates and memory cells in the LSTM architecture. The vector *X* represents the input data at each time step, while *h* denotes the hidden state, which is responsible for maintaining and updating the historical information across sequences. The functions σ and *tanh* correspond to the sigmoid and hyperbolic tangent activation functions, respectively. Once adequately trained, the LSTM network becomes capable of learning and extracting meaningful patterns from complex time-series data. These extracted features, represented by the hidden states, encapsulate the temporal dependencies in the input sequences. Based on this hidden information, the final fully connected layer of the model transforms the learned representations into accurately predicted outputs [31]. Figure 2 illustrates the general structure of the LSTM network.



Modeling Considerations and Evaluation Metrics

In the modeling process for Electrical Conductivity (EC), the Total Dissolved Solids (TDS) parameter was deliberately excluded as an input. This decision stems from the fact that EC can be approximately estimated using TDS through the empirical relationship EC ≈ 0.65 × TDS. Therefore, the objective of this study is to develop a model capable of predicting EC based solely on other water quality parameters without relying directly on TDS.

Similarly, for modeling Sodium Adsorption Ratio (SAR), the following equation is commonly used:

$$SAR = \frac{Na^{+}}{\sqrt{\frac{Ca^{+2} + Mg^{+2}}{2}}} \qquad (9)$$

If the concentrations of Na^{+} , Ca^{2+} , and Mg^{2+} ions are available for a given period, the Sodium Adsorption Ratio (SAR) can be directly calculated using Equation (9). However, in this study, it was assumed that at least one of these three parameters might be unavailable. Accordingly,Ca^{2+} , and Mg^{2+} were deliberately excluded from the input structure of the machine-learning models for SAR prediction. This decision ensures that the models are capable of estimating SAR independently of the empirical Equation, thereby enhancing their generalizability and robustness. This consideration was carefully applied to avoid the inclusion of derived variables, minimize data redundancy, and reduce the risk of statistical bias, ensuring that the models are developed based on realistic and practical assumptions.

For model training and validation, the dataset was randomly divided into two subsets: 70% for training and 30% for testing. To assess model performance, four statistical indicators were employed: the coefficient of determination (R^2), root mean square error (RMSE), the modified Kling-Gupta Efficiency (KGE), and the RSR index. These indicators are further explained in the following section.

Among these metrics, KGE holds particular importance. Unlike traditional indicators such as R^2 or RMSE, which evaluate limited aspects of performance, KGE simultaneously incorporates three critical components: bias, correlation, and variability ratio. In this study, the strong agreement between KGE and other metrics contributed to a more comprehensive and reliable evaluation of the overall model performance.

$$R^2 = 1 - \frac{\sum(O_i - P_i)^2}{\sum(O_i - \bar{O})^2} \tag{10}$$

$$RMSE = \sqrt{\frac{1}{n} \times \sum (O_i - P_i)^2} \tag{11}$$

$$KGE = 1 - \sqrt{(\Gamma - 1)^2 + (\beta - 1)^2 + (\gamma - 1)^2} \tag{12}$$

$$RSR = \frac{RMSE}{STDEV_{obs}} = \frac{\sqrt{\frac{1}{n} \times \sum(O_i - P_i)^2}}{\sqrt{\frac{1}{n-1} \times \sum(O_i - \bar{O})^2}} \tag{13}$$

Where:

O_i: Observed value

P_i: Predicted value

$\bar{O}$: Mean of observed values

n: Number of observations

Γ: Pearson correlation coefficient between observed and predicted values

β = mean (P) / mean (O): Ratio of predicted to observed means

γ = CV_P / CV_O: Ratio of predicted to observed coefficients of variation

2.6 TOPSIS

To comprehensively rank the predictive models, the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) was applied. This multi-criteria decision-making method evaluates each alternative's relative proximity to an ideal solution based on performance metrics. In this study, four statistical indicators (R², RMSE, KGE, and RSR) were used for both training and testing phases, and their weights were calculated using the Shannon entropy method [32].

The TOPSIS process involves constructing a decision matrix, normalizing the data, applying weights to each criterion, identifying the ideal and anti-ideal solutions, computing the Euclidean distances of each alternative to these ideal points, and finally calculating the relative closeness coefficient C_i^* for each alternative, defined as:

$$C_i^* = \frac{S_i^-}{S_i^- + S_i^+} \tag{14}$$

Where S_i^+ and S_i^- are the distances of the *i-th* alternative from the positive and negative ideal solutions, respectively. In the referenced study, the Shannon entropy method was used to determine the weights of the criteria objectively [32].

Figure 3 illustrates the main steps of the research process.

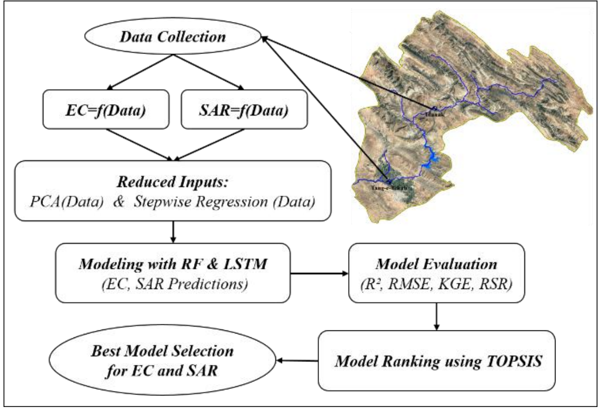


Figure 2: Flowchart of the research steps

3. Results

Descriptive statistical analysis for the two hydrometric stations, Idanak and Tang-e-Tekab, as presented in Table 1, reveals significant differences in both quantitative and qualitative characteristics of the Maroon River water. At the Idanak station, located in the upstream mountainous regions of the basin, the average discharge is 52.05 m^3/s with a standard deviation of 87.40, indicating considerable flow variability driven by seasonal precipitation and snowmelt. In contrast, at the Tang-e-Tekab station located downstream and influenced by the Maroon Dam the average discharge is lower (37.06 m^3/s), and the variability is reduced (standard deviation of 48.59), reflecting the regulated water releases from the dam reservoir.

Salinity-related parameters such as Total Dissolved Solids (TDS) and Electrical Conductivity (EC) are significantly higher at Tang-e-Tekab (1359.93 mg/l and 2039.12 µS/cm, respectively) compared to Idanak (637.03 mg/l and 972.16 µS/cm). This increase can be attributed to several factors, including the river's passage through evaporite formations rich in gypsum and salt, industrial and agricultural activities along the riverbanks, and the high rate of evaporation in the lower basin. The southern location and lower elevation of the downstream areas expose them to more intense evaporation, further exacerbated by substantial evaporative losses from the surface of the Maroon Dam reservoir, which concentrates the dissolved salts.

Moreover, higher concentrations of major ions such as Cl^- , Na^+ , Ca^{2+} , and Mg^{2+} at Tang-e-Tekab contribute to an elevated Sodium Adsorption Ratio (SAR) at this station (3.75 versus 1.92 at Idanak). These disparities underscore the significant influence of geographic location, geological features, and hydro-climatic conditions on water quality. They also highlight the importance of continuous spatiotemporal monitoring of water resources for effective water quality management.

Table 1: Descriptive statistics of Maroon River at two monitoring stations

Parameter	Unit	Station	Mean	Median	Std. Deviation	Minimum	Maximum
Q	m^3/s	Idanak	52.05	23.52	87.40	4.00	756.26
		Tang-e-Tekab	37.06	27.20	48.59	3.06	479.00
TDS	mg/L	Idanak	637.03	595.50	271.53	275.00	3127.00
		Tang-e-Tekab	1359.93	1373.00	219.51	685.00	1988.00
EC	µS/cm	Idanak	972.16	913.00	327.11	452.00	2212.00
		Tang-e-Tekab	2039.12	2060.00	350.24	845.00	2942.00
PH	---	Idanak	7.70	7.70	0.35	6.70	8.70
		Tang-e-Tekab	7.77	7.80	0.26	7.00	8.60
SO4	meq/L	Idanak	3.07	2.90	1.72	0.07	14.30
		Tang-e-Tekab	9.32	9.30	2.11	1.04	18.00
HCO3	meq/L	Idanak	3.01	3.06	0.66	0.49	6.03
		Tang-e-Tekab	2.41	2.39	0.51	0.79	4.00
Cl	meq/L	Idanak	3.72	3.29	2.01	0.58	11.71
		Tang-e-Tekab	9.23	9.13	2.48	2.94	16.13
Ca	meq/L	Idanak	4.75	4.68	1.30	1.65	11.71
		Tang-e-Tekab	9.86	9.80	1.74	4.39	15.00
Mg	meq/L	Idanak	1.63	1.56	0.68	0.10	4.29
		Tang-e-Tekab	2.16	1.94	0.95	0.22	6.45
Na	meq/L	Idanak	3.51	3.13	1.98	0.30	11.68
		Tang-e-Tekab	9.14	8.95	2.61	3.12	17.10
K	meq/L	Idanak	0.04	0.04	0.02	0.01	0.16
		Tang-e-Tekab	0.06	0.06	0.02	0.01	0.15
SAR	---	Idanak	1.92	1.79	0.95	0.19	5.47
		Tang-e-Tekab	3.75	3.62	1.05	1.52	7.71

To optimize the input structure of machine learning models and improve prediction accuracy, two-dimensionality reduction techniques Principal Component Analysis (PCA) and Stepwise Regression were employed. The PCA method aimed to retain 95% of the total variance by extracting a reduced number of principal components, which are linear combinations of the original parameters. In contrast, Stepwise Regression identified the most statistically significant subset of original variables by evaluating their individual and collective contributions to the predictive models.

As shown in Table 2 and Figure 4, at the Idanak station, PCA reduced the number of input components for EC modeling to six (PCA1 to PCA6), while for SAR modeling, five principal components were retained. Using the Stepwise Regression method, the selected variables for EC modeling included $SO_4{}^{2-}$, Cl^- , $HCO_3{}^-$, and Mg^{2+} , whereas for SAR modeling, Na^+ , $SO_4{}^{2-}$, $HCO_3{}^-$, Cl^- , and discharge (Q) were chosen as inputs.

Similarly, at the Tang-e-Tekab station located downstream of the Maroon Dam PCA reduced the number of principal components to six for EC and five for SAR modeling. Meanwhile, Stepwise Regression selected Cl^- , Ca^{2+} , Mg^{2+} , and Na^+ for EC modeling and Na^+ , EC, $SO_4{}^{2-}$, $HCO_3{}^-$, and Cl^- for SAR modeling.

The differences in selected input variables between the two stations highlight the influence of geographic location on model input structure. At the upstream Idanak station, which is situated in the Zagros mountain range before the dam, the river water is less affected by intense evaporation, agricultural and industrial activities, and geological interactions. Under such conditions, ions such as $SO_4{}^{2-}$, Cl^- , $HCO_3{}^-$, and Mg^{2+} were identified as the most influential inputs for EC modeling, primarily reflecting natural mineral dissolution processes. For SAR modeling, variables including Na^+ , $SO_4{}^{2-}$, $HCO_3{}^-$, Cl^- , and Q were selected, with the inclusion of discharge being significant due to its impact on ion concentrations driven by seasonal flow variability caused by precipitation and snowmelt.

In contrast, the downstream Tang-e-Tekab station is more heavily influenced by controlled dam releases, higher evaporation from the reservoir surface, anthropogenic activities, and increased water-rock interactions. At this location, the selected variables for EC modeling Cl^{-} , Ca^{2+} , Mg^{2+} , and Na^{+} highlight the dominant role of dissolved ions in downstream salinity increase. The inclusion of Ca^{2+} and Na^{+} is particularly relevant due to the potential ion exchange and dissolution of mineral formations. In SAR modeling, the input set included Na^{+} , EC, $SO_4^{\ 2-}$, $HCO_3^{\ -}$, and Cl^{-} . The presence of EC as an input in the SAR model suggests that the overall concentration of dissolved salts plays a significant role in sodium adsorption under stable, dam-regulated flow conditions.

Overall, the results of input variable selection at the Idanak and Tang-e- Tekab stations reflect the distinct hydrological, geological, and locational characteristics of these sites. The stepwise regression method, by considering the local features of each station, identified the most influential parameters for accurate water quality modeling while preserving the interpretability of the original variables. This facilitates more precise conceptual, managerial, and decision-making analyses regarding the governing physical and chemical processes. In contrast, Principal Component Analysis (PCA) aims to reduce dimensionality and eliminate multicollinearity by focusing on combinations that capture the most significant variance in the data. Through effective information compression, PCA primarily enhances model efficiency and is, therefore, better suited for datasets with high internal correlation. A detailed comparison of the performance of these two approaches in modeling will be presented in the following section of this paper.

Table 2: Model Input Variables Selected for EC and SAR Prediction at Idanak and Tang-e- Tekab Stations Using PCA and Stepwise Regression Methods

Station	Dimensionality Reduction Method	Inputs	Target
Idanak	PCA	PCA1, PCA2, PCA3, PCA4, PCA5, PCA6	EC
		PCA1, PCA2, PCA3, PCA4, PCA5	SAR
	Stepwise Regression	SO4, Cl, HCO3, Mg	EC
		Na, SO4, HCO3, Cl, Q	SAR
Tang-e- Tekab	PCA	PCA1, PCA2, PCA3, PCA4, PCA5, PCA6	EC
		PCA1, PCA2, PCA3, PCA4, PCA5	SAR
	Stepwise Regression	Cl, Ca, Mg, Na	EC
		Na, EC, SO4, HCO3, Cl	SAR

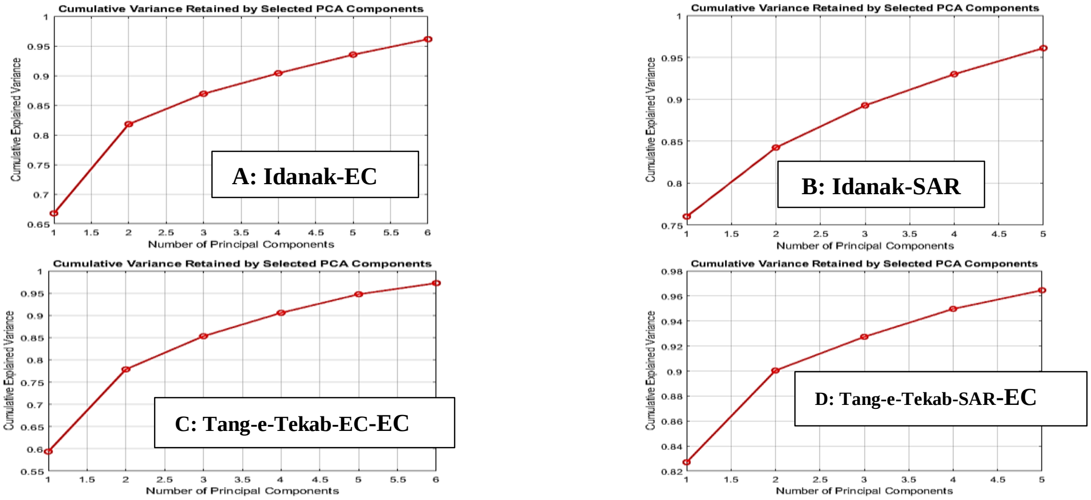


Figure 3: Percentage of Variance Explained by Principal Components Extracted from Model Inputs

Results Analysis

Based on the modeling outcomes presented in Table 3, a comprehensive and multi-dimensional assessment can be made regarding the performance of the models, the dimensionality reduction techniques employed, and the influence of station-specific characteristics.

Table 3

Station	Parameter	Dim. Reduction Method	Model	Train				Test				TOPSIS (Ranking)
				R^2	RMSE	KGE	RSR	R^2	RMSE	KGE	RSR	
Idanak	EC	PCA	RF	0.89	106.64	0.90	0.34	0.84	140.61	0.81	0.40	0.41 (3)
			LSTM	0.92	93.58	0.94	0.28	0.94	80.87	0.93	0.26	0.54 (2)
			RF	0.92	127.74	0.85	0.39	0.89	111.65	0.86	0.34	0.38 (4)

		Stepwise Regression	LSTM	0.91	99.17	0.94	0.29	0.96	61.56	0.96	0.21	0.55 (1)
	SAR	PCA	RF	0.92	0.27	0.92	0.29	0.90	0.32	0.86	0.33	0.32 (4)
			LSTM	0.97	0.16	0.98	0.17	0.94	0.21	0.97	0.24	0.41 (2)
		Stepwise Regression	RF	0.95	0.21	0.96	0.22	0.95	0.22	0.93	0.23	0.35 (3)
			LSTM	0.99	0.03	1.00	0.03	0.99	0.08	0.99	0.08	0.56 (1)
Tang-e-Tekab	EC	PCA	RF	0.89	125.31	0.86	0.34	0.85	123.14	0.88	0.39	0.42 (3)
			LSTM	0.96	76.60	0.97	0.21	0.91	99.97	0.94	0.33	0.56 (1)
		Stepwise Regression	RF	0.90	116.03	0.86	0.33	0.85	131.48	0.85	0.39	0.41 (4)
			LSTM	0.94	85.77	0.96	0.24	0.92	90.37	0.96	0.27	0.55 (2)
	SAR	PCA	RF	0.94	0.28	0.90	0.26	0.87	0.37	0.85	0.37	0.42 (4)
			LSTM	0.98	0.14	0.99	0.13	0.96	0.18	0.95	0.21	0.64 (1)
		Stepwise Regression	RF	0.95	0.24	0.94	0.23	0.94	0.26	0.92	0.25	0.49 (3)
			LSTM	0.99	0.04	1.00	0.04	0.94	0.29	0.95	0.25	0.61 (2)

Comparison of Model Performance (RF vs. LSTM)

The LSTM model consistently outperformed the RF model in most scenarios, particularly in the prediction of the SAR parameter. This superiority is evident in lower RSR and RMSE values, alongside higher R² and KGE scores across various cases. For instance, at the Idanak station, the LSTM model, when fed with input variables selected through the Stepwise method, achieved outstanding accuracy in SAR prediction, with an R² of 0.99 and a remarkably low RMSE of 0.08 in the testing phase representing the best performance across all evaluated scenarios.

In contrast, although the RF model demonstrated acceptable results in certain instances, its overall performance was consistently ranked lower than that of the LSTM model based on the TOPSIS method. This discrepancy can be attributed to fundamental differences in the structural capabilities of the two models when handling complex, nonlinear relationships among input variables. The RF model, which is based on decision tree ensembles, generally performs well for classification or problems with discrete and relatively simple structures. However, it tends to face limitations when dealing with datasets characterized by continuous interactions and multivariate dependencies [33].

On the other hand, the LSTM model, due to its deep and flexible neural network architecture, is better equipped to capture hidden nonlinear patterns among temporal variables. Its superior capability in adjusting the weights of dependent inputs enhances its predictive accuracy. This characteristic enables LSTM to deliver more precise modeling of EC and SAR parameters, especially under hydrologically and geologically complex conditions at the studied stations [34].

Role of Dimensionality Reduction Techniques (PCA vs. Stepwise Regression)

Principal Component Analysis (PCA), as a dimensionality reduction method, transforms the original input variables into a new set of independent components by eliminating multicollinearity while retaining the maximum possible variance from the original data. Although this transformation enhances computational efficiency, it comes at the cost of interpretability, as the original physical meaning of the variables is lost. Nevertheless, PCA demonstrated strong performance in specific scenarios particularly in predicting EC at the Tang-e- Tekab station using the LSTM model where it achieved the top rank with an R² of 0.96 and an RMSE of 76.60.

This success can be attributed to the natural compatibility between deep neural networks and decor-related, compressed datasets. The hierarchical architecture of LSTM networks enables them to uncover hidden nonlinear patterns embedded within the principal components, allowing for effective learning despite the abstraction of the original input features.

In contrast, Stepwise Regression preserves the original structure and physical interpretability of the input variables, which is especially beneficial for managerial and conceptual analyses. By selectively identifying the most influential predictors, this method has yielded outstanding results in specific cases for example, SAR modeling at the Idanak station using the LSTM model, where it achieved the highest rank with R² = 0.99 and RMSE = 0.08.

Hence, in contexts where interpretability and parameter traceability are prioritized, particularly for policy or management applications, the Stepwise Regression method proves more advantageous.

Influence of Station Location on Model Performance

The geographical location of monitoring stations plays a crucial role in shaping the complexity of water quality dynamics and, consequently, the performance of predictive models. The Idanak station, located upstream and before the Maroon Dam, is characterized by more natural hydrological and geochemical conditions. Although the data in this region may exhibit higher fluctuations due to natural variability, they are relatively less affected by anthropogenic interventions such as dam operations or surface evaporation. This relative simplicity allows models, especially those based on interpretable input selections such as Stepwise Regression, to perform effectively.

In contrast, the Tang-e- Tekab station is situated downstream of the Maroon Dam, where the stored water is subject to extensive evaporation, leading to increased ionic concentrations. Additionally, the prolonged residence time of water within the reservoir enhances opportunities for chemical reactions and interactions with geological formations. These factors introduce additional complexity into the relationships between water quality variables, necessitating models capable of capturing intricate, nonlinear patterns. In this regard, the PCA+LSTM model owing to its advanced learning capacity and temporal feature extraction capabilities demonstrated superior performance, particularly in handling the complex hydrochemical behaviors observed downstream.

Final Evaluation Using the TOPSIS Method

The TOPSIS multi-criteria decision-making (MCDM) approach was utilized to provide a comprehensive ranking of the models by simultaneously considering four statistical indicators across both training and testing phases. Based on this integrated assessment:

- For EC prediction at both stations, the LSTM model regardless of whether PCA or Stepwise Regression was used for input selection outperformed the RF model.
- Similarly, in SAR modeling, LSTM consistently ranked first or second across all scenarios, indicating a clear and consistent advantage in performance.

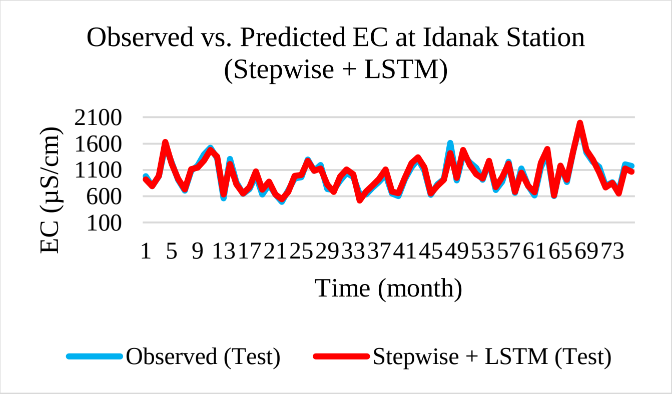
Overall, at the Idanak station, located upstream of the Maroon basin and prior to the dam, hydrological and geological conditions are relatively more stable and less influenced by anthropogenic interventions or long-term accumulation effects. In such a context, the relationships among water quality variables tend to be more direct and closely tied to local characteristics such as geological formations, precipitation patterns, and surface runoff dynamics. Consequently, Stepwise Regression, which selects core variables through an interpretable and targeted process, was effective in identifying the truly influential predictors of water quality. When coupled with the LSTM model capable of learning complex nonlinear relationships this resulted in a highly accurate and simultaneously interpretable model suitable for practical water resource management.

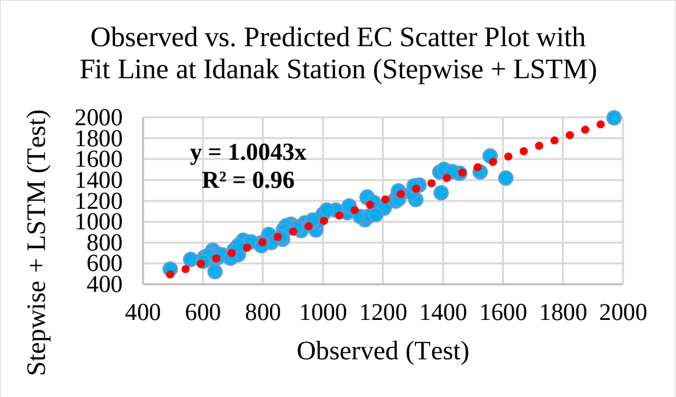
In contrast, the Tang-e- Tekab station, located downstream of the Maroon Dam at the basin's outlet, is subject to more complex influences, including dam releases, long-term solute accumulation, significant evaporation, and extended interactions with geological formations and riverine sediments. These conditions lead to stronger multicollinearity among water quality parameters, making direct analysis more challenging. Under such circumstances, PCA proved effective in reducing collinearity and compressing the data into uncorrelated principal components, optimizing the input space for modeling. The LSTM model, with its powerful capacity to learn complex patterns, was able to leverage these components to deliver more accurate predictions.

In essence, the success of the Stepwise + LSTM combination at Idanak stemmed from the match between the basin's relatively analyzable and straightforward characteristics and the model's ability to identify and exploit key inputs. Conversely, the superior performance of the PCA + LSTM combination at Tang-e- Tekab can be attributed to the need for dimensionality reduction and collinearity mitigation in a highly complex hydrogeochemical environment, as well as the LSTM's strength in extracting latent patterns from transformed features. These findings are consistent with those of similar studies, which highlight that model performance is strongly influenced by site-specific hydrological conditions, data quality, and input structure [35].

The plotted graphs in Figures 5 and 6 based on the testing dataset illustrate the comparison between observed and predicted values of EC and SAR parameters at the Idanak and Tang-e- Tekab stations, respectively, further confirming the excellent performance of the selected models.

Overall, the remarkable agreement between the observed and predicted time series, together with the slope values approaching unity in the fitted regression lines, highlights the high accuracy and robustness of the models across both stations. This visual consistency strongly corroborates the numerical evaluation results presented in the assessment tables.





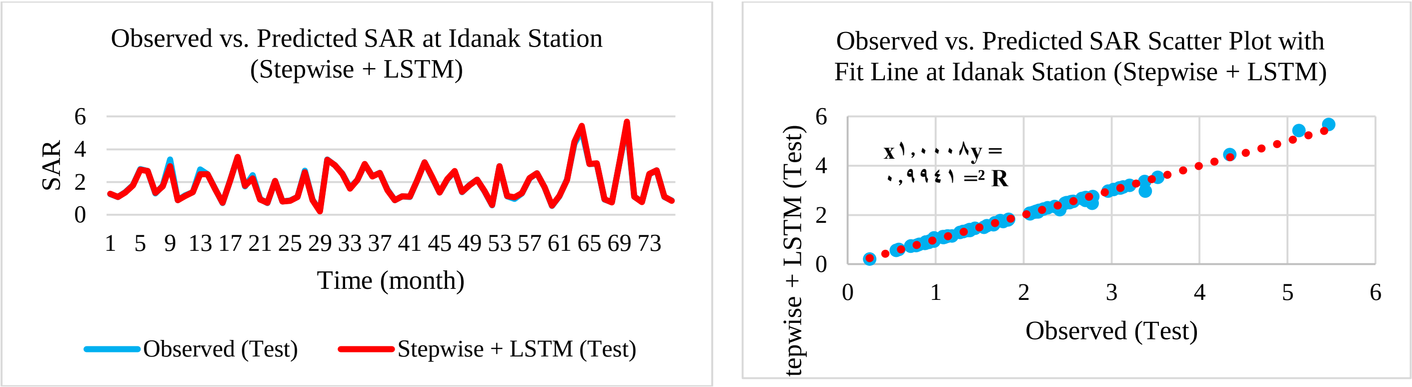


Figure 5: Observed vs. Predicted EC and SAR at Idanak Station (Stepwise Input Selection+LSTM)

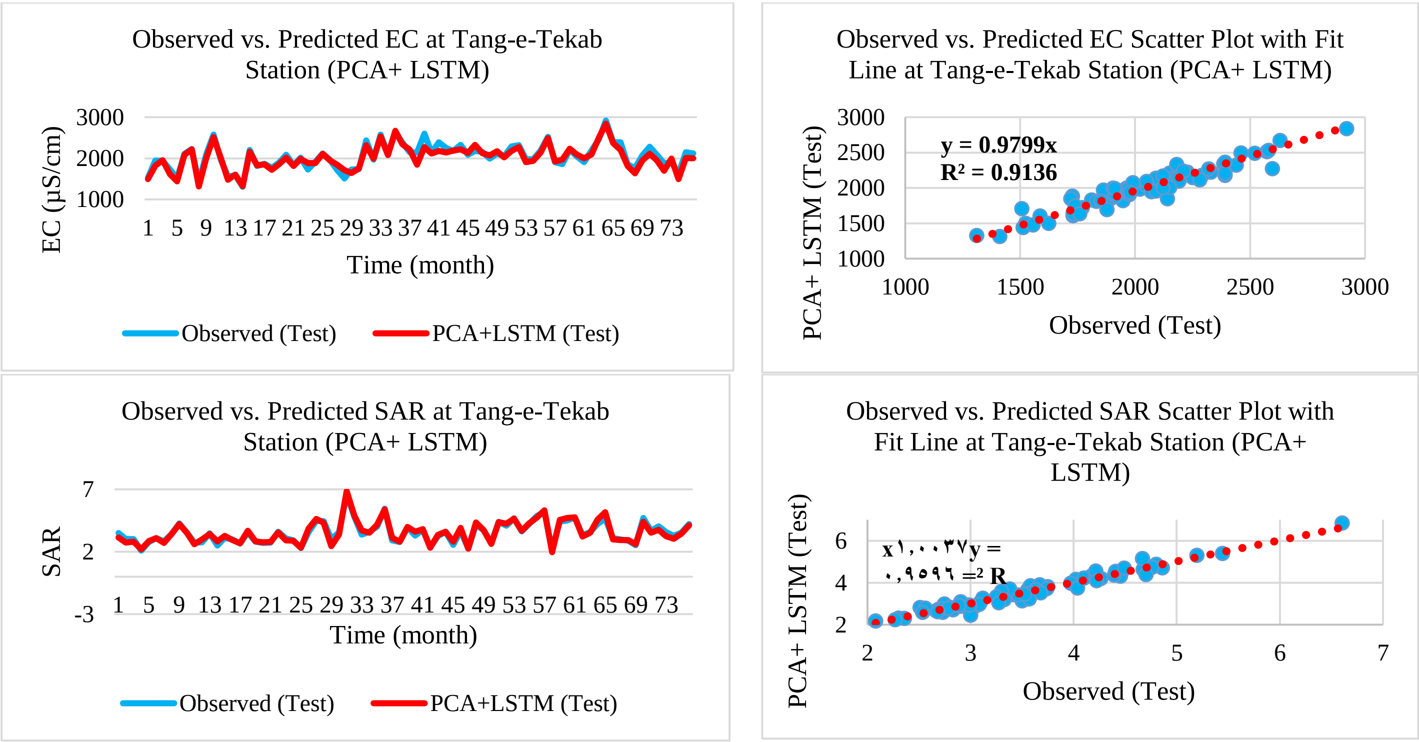


Figure 6: Observed vs. Predicted EC and SAR at Tang-e- Tekab Station (PCA Input Selection+LSTM)

Conclusion

The comparative analysis reveals that the LSTM model consistently outperforms the RF model across most scenarios, particularly in SAR prediction, owing to its superior ability to capture complex nonlinear temporal patterns. Dimensionality reduction techniques play a critical role: while PCA enhances model performance by effectively reducing multicollinearity and compressing inputs especially beneficial under complex hydrogeochemical conditions at the Tang-e- Tekab station Stepwise Regression preserves variable interpretability and excels in more stable, less disturbed environments like the Idanak station. The geographical and hydrological context significantly influences model efficacy; upstream Idanak's relatively straightforward conditions favor interpretable input selection paired with LSTM, whereas downstream Tang-e- Tekab's complex dynamics require PCA's dimensionality reduction combined with LSTM's advanced pattern recognition. Multi-criteria TOPSIS evaluation confirms the consistent superiority of LSTM-based models, and visual comparisons of observed versus predicted data further validate their high accuracy and robustness. These findings underscore the importance of tailoring model architecture and input preprocessing strategies to site-specific conditions for optimal water quality parameter prediction, highlighting the practical implications of the research.

Also, the findings of this study hold significant practical value for supporting water resource management decisions, particularly in regulating the quality of water released from the Maroon Dam and managing irrigation practices in downstream agricultural lands. At the Tang-e- Tekab station located directly downstream of the dam accurate prediction of salinity-related parameters such as EC and SAR using the LSTM model can serve as an advanced early warning system to prevent salinity stress in agricultural water. For instance, if the model forecasts a sharp increase in SAR or EC in the coming months, reservoir release volumes and timing can be strategically adjusted to dilute salinity levels and mitigate potential risks to crops. similarly, at the upstream Idanak station, where hydrochemical conditions remain relatively stable, accurate forecasts of EC and SAR can serve as reliable benchmarks for detecting both natural variations and anthropogenic impacts on water quality. Such predictive insights are particularly beneficial for farmers along the Maroon River, especially during sensitive agricultural periods such as June to September, when salt-sensitive crops like rice are cultivated. Early knowledge of future water quality enables more informed decisions regarding crop selection, irrigation timing, and planting strategies.

These findings underscore that model selection and dimensionality reduction strategies should not follow a one-size-fits-all approach; instead, they must be tailored to the specific hydrological and geo-environmental context of each monitoring site. Ultimately, integrating these predictive models into regional decision-support systems could significantly enhance the intelligence and adaptability of water resource management.

Reference

[1.] Kaushal SS, Likens GE, Pace ML, Reimer JE, Maas CM, Galella JG, Utz RM, Duan S, Kryger JR, Yaculak AM, Boger WL. Freshwater salinization syndrome: from emerging global problem to managing risks. Biogeochemistry. 2021 Jun;154:255-92.

[2.] Al-Mukhtar M, Al-Yaseen F. Modeling Water Quality Parameters Using Data-Driven Models, a Case Study Abu-Ziriq Marsh in South of Iraq. Hydrology. 2019; 6(1):24. https://doi.org/10.3390/hydrology6010024

[3.] Adib A, Farajpanah H, Mahmoudian Shoushtari M, Ahmadeanfar I. Estimation of Water Quality Parameters in the Sepidrood River by ANFIS, GEP and LS-SVM Models. Journal of Water and Wastewater. 2020 Nov 21;31(5):1-0.

[4.] Adib A, Farajpanah H, Shoushtari MM, Lotfirad M, Saeedpanah I, Sasani H. Selection of the best machine learning method for estimation of concentration of different water quality parameters. Sustainable Water Resources Management. 2022 Dec;8(6):172.

[5.] Victoriano JM, Santos ML, Vinluan AA, Carpio JT. Predicting pollution level using random forest: a case study of Marilao River in Bulacan Province, Philippines. arXiv preprint arXiv:2202.06066. 2022 Feb 12.

[6.] Bui DT, Khosravi K, Tiefenbacher J, Nguyen H, Kazakis N. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Science of the Total Environment. 2020 Jun 15;721:137612.

[7.] Wu X, Zhang Q, Wen F, Qi Y. A water quality prediction model based on multi-task deep learning: a case study of the Yellow River, China. Water. 2022 Oct 27;14(21):3408.

[8.] Ubah JI, Orakwe LC, Ogbu KN, Awu JI, Ahaneku IE, Chukwuma EC. Forecasting water quality parameters using artificial neural network for irrigation purposes. Scientific Reports. 2021 Dec 24;11(1):24438.

[9.] Nouraki A, Alavi M, Golabi M, Albaji M. Prediction of water quality parameters using machine learning models: A case study of the Karun River, Iran. Environmental Science and Pollution Research. 2021 Oct;28(40):57060-72.

[10.] Trach R, Trach Y, Kiersnowska A, Markiewicz A, Lendo-Siwicka M, Rusakov K. A study of assessment and prediction of water quality index using fuzzy logic and ANN models. Sustainability. 2022 May 7;14(9):5656.

[11.] Adib A, Farajpanah H, Shoushtari MM, Lotfirad M, Saeedpanah I, Sasani H. Selection of the best machine learning method for estimation of concentration of different water quality parameters. Sustainable Water Resources Management. 2022 Dec;8(6):172.

[12.] Ibrahim A, Ismail A, Juahir H, Iliyasu AB, Wailare BT, Mukhtar M, Aminu H. Water quality modelling using principal component analysis and artificial neural network. Marine Pollution Bulletin. 2023 Feb 1;187:114493.

[13.] Adjovu GE, Stephen H, Ahmad S. A machine learning approach for the estimation of total dissolved solids concentration in lake mead using electrical conductivity and temperature. Water. 2023 Jul 2;15(13):2439.

[14.] Pourhosseini FA, Ebrahimi K, Omid MH. Prediction of total dissolved solids, based on optimization of new hybrid SVM models. Engineering Applications of Artificial Intelligence. 2023 Nov 1;126:106780.

[15.] Pyo J, Pachepsky Y, Kim S, Abbas A, Kim M, Kwon YS, Ligaray M, Cho KH. Long short-term memory models of water quality in inland water environments. Water research X. 2023 Dec 1;21:100207.

[16.] Karbasi M, Ali M, Bateni SM, Jun C, Jamei M, Farooque AA, Yaseen ZM. Multi-step ahead forecasting of electrical conductivity in rivers by using a hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) model enhanced by Boruta-XGBoost feature selection algorithm. Scientific reports. 2024 Jul 1;14(1):15051.

[17.] Jaafer NS, Al-Mukhtar M. Prediction of Water Quality Parameters of Tigris River in Baghdad City by Using Artificial Intelligence Methods. Ecological Engineering & Environmental Technology. 2024;25.

[18.] Ismail R, Rawashdeh A, Al-Mattarneh H, Hatamleh R, Dua’a BT, Jaradat A. Artificial intelligence for application in water engineering: The use of ANN to determine water quality index in rivers. Civil Engineering Journal. 2024 Jul 1;10(7):2261-74.

[19.] Satish N, Anmala J, Rajitha K, Varma MR. A stacking ANN ensemble model of ML models for stream water quality prediction of Godavari River Basin, India. Ecological Informatics. 2024 May 1;80:102500.

[20.] Adnan RM, Ewees AA, Wang M, Kisi O, Heddam S, Parmar KS, Zounemat-Kermani M. Enhancing BOD5 forecasting accuracy with the ANN-Enhanced Runge Kutta model. Journal of Environmental Chemical Engineering. 2025 Apr 1;13(2):115430.

[21.] Khosravi K, Farooque AA, Karbasi M, Ali M, Heddam S, Faghfouri A, Abolfathi S. Enhanced water quality prediction model using advanced hybridized resampling alternating tree-based and deep learning algorithms. Environmental Science and Pollution Research. 2025 Feb 24:1-20.

[22.] Abushandi E. Water Quality Assessment and Forecasting Along the Liffey and Andarax Rivers by Artificial Neural Network Techniques Toward Sustainable Water Resources Management. Water. 2025 Feb 6;17(3):453.

[23.] Al-Khuzaie MM, Abdul Maulud KN, Wan Mohtar WH, Yaseen ZM. Modelling Euphrates river water quality index based on field measured data in Al-Diwaniyah City, Iraq. Scientific Reports. 2025 Jan 2;15(1):51.

[24.] Isık H, Akkan T. Water quality assessment with artificial neural network models: Performance comparison between SMN, MLP and PS-ANN methodologies. Arabian Journal for Science and Engineering. 2025 Jan;50(1):369-87.

[25.] Ahmadpour A, Mirhashemi SH, Panahi M. Comparative evaluation of classical and SARIMA-BL time series hybrid models in predicting monthly qualitative parameters of Maroon river. Applied Water Science. 2023 Mar;13(3):71.

[26.] Sayahi F, Divband Hafshejani L, Tishehzan P, Abdolabadi H. The combination of dimensionality reduction methods and machine learning algorithms in the optimization of Maroon River water quality prediction. Iranian Journal of Soil and Water Research. 2024 Nov 21;55(9):1601-15.

[27.] Farajpanah H, Adib A, Lotfirad M, Esmaeili-Gisavandani H, Riyahi MM, Zaerpour A. A novel application of waveform matching algorithm for improving monthly runoff forecasting using wavelet–ML models. Journal of Hydroinformatics. 2024 Jul 1;26(7):1771-89.

[28.] Breiman L. Random forests. Machine learning. 2001 Oct;45:5-32.

[29.] Salehinejad H, Sankar S, Barfett J, Colak E, Valaee S. Recent advances in recurrent neural networks. arXiv preprint arXiv:1801.01078. 2017 Dec 29.

[30.] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997 Nov 15;9(8):1735-80.

[31.] Chen H, Yang J, Fu X, Zheng Q, Song X, Fu Z, Wang J, Liang Y, Yin H, Liu Z, Jiang J. Water quality prediction based on LSTM and attention mechanism: A case study of the Burnett River, Australia. Sustainability. 2022 Oct 14;14(20):13231.

[32.] Farajpanah H, Lotfirad M, Adib A, Esmaeili-Gisavandani H, Kisi Ö, Riyahi MM, Salehpoor J. Ranking of hybrid wavelet-AI models by TOPSIS method for estimation of daily flow discharge. Water Supply. 2020 Dec 1;20(8):3156-71.

[33.] Wright MN, Ziegler A, König IR. Do little interactions get lost in dark random forests?. BMC bioinformatics. 2016 Dec;17:1-0.

[34.] Lai G, Chang WC, Yang Y, Liu H. Modeling long-and short-term temporal patterns with deep neural networks. InThe 41st international ACM SIGIR conference on research & development in information retrieval 2018 Jun 27 (pp. 95-104).

[35.] Kratzert F, Klotz D, Shalev G, Klambauer G, Hochreiter S, Nearing G. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrology and Earth System Sciences. 2019 Dec 17;23(12):5089-110.

مع تحليل المكونات RFو LSTM التنبؤ بجودة مياه النهر باستخدام التعلم الآلي باستخدام نماذج .الرئيسية والانحدار التدريجي لتقليل الأبعاد: دراسة حالة لحوض نهر مارون

التنبؤ بجودة مياه النهر باستخدام التعلم الآلي باستخدام نماذج LSTM وRF مع تحليل المكونات الرئيسية والانحدار التدريجي لتقليل الأبعاد: دراسة حالة لحوض نهر مارون.

احمد مجيد محمد دلويى[١] فلاح حمه فرج علي[٢]
سالم نعمت عزيز[٣] هيوا فرج پناه[٤]

[١] جامعة كرميان التقنية، المعهد كلار التقني، قسم المساحة، السليمانية، إقليم كردستان، العراق.
[٢] جامعة السليمانية التقنية، المعهد دربنديخان التقني، قسم المساحة، السليمانية، إقليم كردستان، العراق.
[٣] جامعة السليمانية التقنية ، المعهد التقني بكرجو، قسم الزراعة المحمية، . السليمانية، إقليم كردستان، العراق.
[٤] جامعة الشهيد شمران ، كلية الهندسة المدنية والعمارة، ، الأهواز، ايران .

الخلاصة

تعد مراقبة جودة مياه الأنهار والتنبؤ بها أمرًا حيويًا لإدارة المياه الحضرية والزراعة والاستدامة البيئية، وخاصة في المناطق الحارة والجافة. يمكن أن يؤدي وجود السدود على طول مجرى النهر إلى تغيير كبير في جودة المياه من خلال التأثير على أنظمة التدفق وتراكم الملح، مما قد يؤدي إلى زيادة الملوحة وغيرها من المشاكل ذات الصلة. ومع ذلك، يمكن إدارة هذه التغييرات بفعالية من خلال عملية النمذجة الدقيقة والتنبؤ.

تهدف هذه الدراسة إلى تقييم أداء نموذجين من نماذج تعلم الآلة (*Machine learning models*)وهما الغابة العشوائية (*RF*) والذاكرة طويلة المدى قصيرة المدى (*LSTM*) في التنبؤ بالتوصيل الكهربائي (*EC*) ونسبة امتصاص الصوديوم (*SAR*) في حوض نهر مارون في إيران. تم استخدام تحليل المكونات الرئيسية (*PCA*) والانحدار التدريجي لتقليل أبعاد المدخلات وتعزيز كفاءة النموذج. تشير النتائج إلى أن نموذج *LSTM* يتفوق باستمرار على *RF* في كل من محطة *Idank* (أعلى سد مارون-*upstream*) ومحطة *Tang-e-Tekab* (أسفل سدمارون*downstream-*) لكلا المعاملين، وخاصة في التنبؤ بـ *SAR*. في محطة *Idnak*، حقق نموذج *LSTM*، مقترنا بالانحدار التدريجي، أعلى دقة في تنبؤات *EC*، حيث بلغ R^2 *0.96*، و*RMSE 61.56*، و*KGE 0.96* في مجموعة بيانات الاختبار. أما بالنسبة لـ *SAR* في المحطة نفسها، فقد أظهر *LSTM* أداءً استثنائيًا، محققًا R^2 *0.99*، و*RMSE 0.08*، و*KGE 0.99*. أما في محطة *Tang-e-Tekab* ، فقد حقق

LSTM مع تحليل المكونات الرئيسية (*PCA*) أدق تنبؤات *EC* ($R^2 = 0.96$، و$RMSE = 76.60$، و$KGE = 0.97$). وبالمثل تم الحصول على أفضل تنبؤات *SAR* في هذه المحطة باستخدام *LSTM* مع تحليل المكونات الرئيسية (*PCA*) ($R^2 = 0.96$، و$RMSE = 0.18$، و$KGE = 0.95$). تؤكد هذه النتائج فعالية دمج شبكات *LSTM* مع تقنيات اختيار المدخلات المصممة خصيصًا لظروف كل موقع، مما يُبرز إمكانية تطبيقها في أنظمة دعم قرارات موارد المياه. بشكل عام، تظهر هذه الدراسة أنه على الرغم من تأثير عمليات السدود على جودة المياه، إلا أنه يمكن إدارة هذه التأثيرات بنجاح من خلال النمذجة التنبؤية المتقدمة لتسهيل الإدارة المستدامة لموارد المياه.

الكلمات المفتاحية: نهر المارون في ايران ، جودة المياه ، التوصيل الكهربائي ، نسبة امتصاص الصوديوم ، الغابة العشوائية ، الذاكرة طويلة المدى قصيرة المدى.