Water Conservation Science and Engineering (2025) 10:63
https://doi.org/10.1007/s41101-025-00390-z

RESEARCH

TDS Prediction with Wavelet Analysis and Trend-Seasonal Decomposition and Machine Learning Algorithms, Case Study: Karkheh River, Iran

Hossein Amini[1] · Farshid Fakheri[2] · Jaffer Yousuf Dar[3] · Reza Shakeri[4] · Hossein Nejati[4] · Man Yue Lam[1] · Banafsheh Zahraie[4] · Reza Ahmadian[1]

Received: 14 January 2025 / Revised: 31 May 2025 / Accepted: 17 June 2025

Abstract

Water quality assessment is definitely important, as the available water resources are highly stressed by population growth, climate change due to anthropogenic activities, and a significant change in consumption patterns. This study aims an innovative framework to predict the total dissolved solids (TDS) with more accuracy in the rivers, case study: "Karkheh River", Iran, with the integration of signal analysis with machine learning algorithms. First, continuous wavelet transform (CWT) was applied to decompose the time series of water quality variables (e.g., Ca, HCO_3, SO_4, and Cl) into their trends, seasonality, and residuals, extracting features that capture temporal dynamics. These features served as input for non-linear machine learning models (XGBoost, Random Forest, Decision Tree) in differenct scenarios to compare which way of adding new feature would improve the model performance in terms of the TDS predictions. Adding new features characterized by only TDS signal analysis improved the TDS predictions and was compared with adding all variables signal characterization and compared with only using raw data to predict TDS level. Using a 50-year dataset from three different hydrometric stations, the models could achieve over 95% accuracy, and XGBoost outperformed others in terms of taking the advantage of the new extracted features from signals. The results indicates that signal-driven features significantly contribute to ccurately TDS prediction by 30% improvement in RMSE, and it can offer a scalable approach for real-time water quality monitoring in semi-arid river systems, which leads to a better early warning system for designing future mitigation strategies.

Keywords Water quality · Machine learning · TDS · Signal analysis

✉ Hossein Amini
AminiH@cardiff.ac.uk

Farshid Fakheri
farshid73@aut.ac.ir

Jaffer Yousuf Dar
Yousuf.Dar@igb-berlin.de

Reza Shakeri
Shakeri.R@ut.ac.ir

Hossein Nejati
Hossein.nejati@ut.ac.ir

Man Yue Lam
LamM7@cardiff.ac.uk

Banafsheh Zahraie
bzahraie@ut.ac.ir

Reza Ahmadian
AhmadianR@cardiff.ac.uk

[1] Hydro-Environmental Research Centre, School of Engineering, Cardiff University, Cardiff, UK

[2] Department of Civil and Environmental Engineering, Amirkabir University of Technology, Tehran, Iran

[3] Department of Experimental Limnology, Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany

[4] School of Civil Engineering, College of Engineering, University of Tehran, Tehran, Iran

Published online: 25 June 2025

Introduction

Water is a fundamental and highly limited resource, essential for the sustenance of life and the health of ecosystems. Water quality is a pivotal priority for public health, ecosystem sustainability, and economic development, and it faces unprecedented challenges from population growth, climate change, and anthropogenic activities [23, 36, 60]. Given the strong and direct correlation between water quality and public health, governments and international organizations have made it a top priority [16]. Globally, the abovementioned challenges in freshwater resources are mainly from agricultural runoff, industrial discharges, and urban wastewater degrading surface and groundwater quality [13, 41]. According to the World Health Organization (WHO), poor water quality accounts for a significant portion of global diseases and mortality rates, especially in developing regions where access to safe drinking water is limited (WHO, 2011). Water quality is influenced by a variety of factors, including natural processes and human activities. Pollutants from agricultural runoff, industrial discharges, and urban wastewater are some of the main contributors to water contamination which are affecting both surface and groundwater sources. In semi-arid regions, such as the Middle East and North Africa, these challenges are more important and necessary to be considered due to the water scarcity, rising temperatures, and continuous and long droughts, which increases the necessity for robust water quality monitoring and management [37]. Effective monitoring of key indicators, such as Total Dissolved Solids (TDS), is essential to ensure water safety and inform sustainable resource management [47, 52].

Among the various indicators that are used to assess the water quality, Total Dissolved Solids (TDS) and Electrical Conductivity (EC) are among the most widely recognized. TDS represents the total concentration of dissolved substances in water, including minerals, salts, and organic matter, which can affect the taste, hardness, and health impact of water [2]. EC is similar to TDS,in EC, we measure the water's ability to conduct electrical current, which is directly related to the concentration of dissolved ions in water [52]. Elevated levels of TDS and EC can indicate the presence of pollutants that may be harmful to human health, necessitating rigorous monitoring and management [30]. In West Asian countries like Iran, an increase in TDS levels, often caused by natural processes and human activities (anthropogenic activities), can have risks to public health [30]. The Karkheh River, Iran's third-longest river, exemplifies these challenges, with its water quality affected by rapid urbanization, industrial activities, and climate variability [5]. Traditional water quality assessment methods, such as chemical analysis and biological monitoring, are expensive and human resource demanding, and sometimes inefficient to capture completely the temporal dynamics of the non-stationary time series of the variables. To elaborate more, sometimes, in traditional methods, the seasonal fluctuations or abrupt pollution events cannot be tracked or captured [32]. In a sense, there is an essential need for advanced methodologies that can accurately predict TDS levels and support real-time monitoring in complex hydrological systems.

In recent years, the integration of signal processing techniques, such as Fourier Analysis, Wavelet Transform, and Continuous Wavelet Transform (CWT), alongside the advancement and improvements in Artificial Intelligence (AI) and Machine Learning (ML), has opened new avenues for the precise and efficient monitoring of water quality [32, 40, 54]. Wavelet analysis, particularly Continuous Wavelet Transform (CWT), has been used to analyze non-stationary water quality signals, revealing temporal patterns and anomalies [24, 58]. Similarly, ML models, such as Artificial Neural Networks and Random Forests, have shown high accuracy in predicting water quality parameters [27, 59]. However, few studies have integrated wavelet-based trend-seasonal decomposition with ML to predict TDS in semi-arid river systems. For instance, Dong et al. [19] combined signal decomposition with deep learning for dissolved oxygen prediction, but their approach did not address TDS or leverage ensemble ML models. This gap highlights the need for a framework that combines the time–frequency localization of CWT with advanced ML to enhance TDS prediction accuracy and interpretability. The novelty of this work lies in showing which way of doing signal analysis and adding features might improve the model capability in TDS prediction. In a sense, it is important to be computationally efficient, as this signal analysis might increase the computational costs of the process,however, the accuracy does not improve significantly. Thus, in this study, we try to show which factor can improve the TDS prediction accuracy.

This study tries to propose a novel and new method to predict TDS levels in the Karkheh River, Iran, by integrating CWT for trend-seasonal decomposition and using the extracted features from signals in the ensemble ML models, including XGBoost, Random Forest, and Decision Tree. Using a 50-year dataset from three hydrometric stations, we decompose water quality signals (e.g., Ca, SO_4) into trends, seasonality, and residuals, extracting features that capture non-stationary dynamics. Given this feature extraction, three different scenrios of dataset construction are defined including considering only original data, considering all variables to be analyzed through signal analisys and add them to the original dataset, and finally considering only features from TDS signal to be added to the original dataset. We hypothesize that adding more features would improve the TDS prediction. These extracted features, along with the original data, serve as inputs for ML models to achieve high prediction accuracy. The study aims to address the limitations of prior approaches by leveraging signal-derived features and ensemble ML, offering a scalable solution for real-time water quality monitoring in semi-arid regions. By applying this framework to the Karkheh River, we demonstrate its potential to inform water resource management and safeguard public health.

Materials and Methods

Study Area

Hydro-environmental research often faces the challenge of inconsistent and limited data. To overcome this, the Karkheh River was chosen for its extensive 50-year data record, which makes it ideal for studying the impacts of climate change, urban development, and human activities on water quality [5]. The Karkheh River, Iran's third longest river at approximately 950 km, originates in the Zagros Mountains in western Iran and flows into the Persian Gulf through the Hurolazim wetland, which is located at the Iran-Iraq border. This river has an average flow rate of 177.8 m^3/s and a catchment area of about 51,912.3 km^2 [35, 44]. The region experiences cold winters, hot summers, 360 mm of annual precipitation, and 3200 mm of annual evaporation. The average temperature is 19 °C with 37% relative humidity from the early 1990s to around 2020 [42], and elevations range from 3000 to 500 m [50]. The Karkheh River is crucial for irrigation, drinking water, and industrial uses in the regions it traverses. The study area's location and the river's path within Khuzestan province in the southwest of Iran are shown in Fig. 1 [43].

Data Collection and Preparation

Data on both qualitative and quantitative variables including TDS, EC, pH, cations (Sodium (Na^{+}), Magnesium ($Mg2^{+}$), Calcium ($Ca2^{+}$)), anions (Chloride (Cl^{-}), Sulfate ($SO_4 2^{-}$), Bicarbonates (HCO_3^{-})), and Discharge (Flow rate, Q) were collected from three hydrometric stations on the Karkheh River, provided by the Iran Water Resources Management Company. These data, sampled monthly from

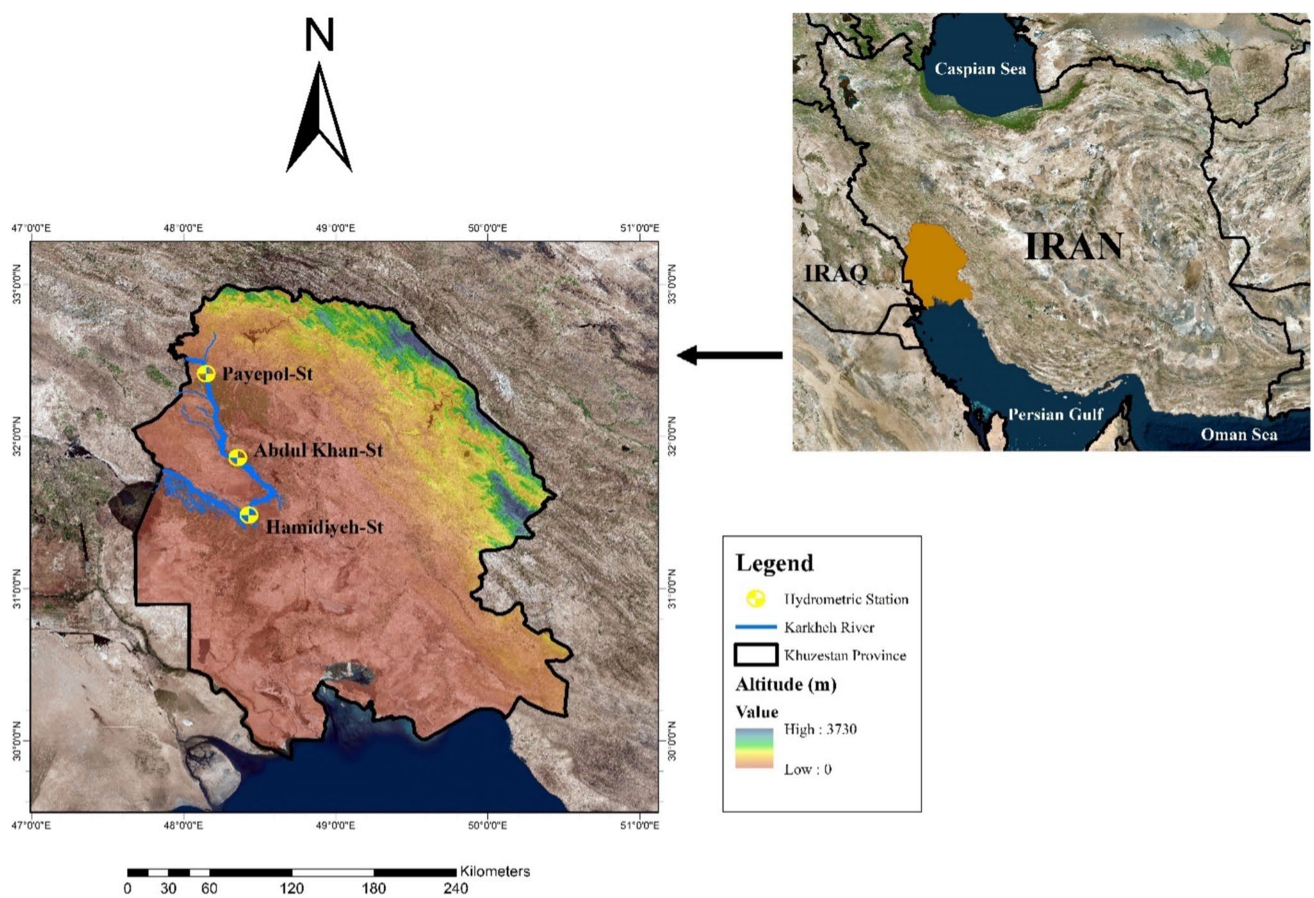

Fig. 1 The path of the Karkheh River in Khuzestan province, southwestern Iran

1968 to 2018 (50 years), come from stations strategically positioned along the river:

- Payepol station: 48° 08′ E, 32° 24′ N
- Abdul Khan station: 48° 21′ E, 31° 51′ N
- Hamidiyeh station: 48° 25′ E, 31° 29′ N

To address data gaps, interpolation techniques using linear regression were applied to the nearest periods of measured values. Additionally, ML statistical methods were used to assess whether the data followed a normal distribution. Table 1 compares the quality variables of the Karkheh River to the permissible values according to WHO guidelines. But before employing the data, it is essential to check the statistical condition of the measured data by statistical tests.

Table 1 Comparative analysis of monthly water quality variables of the Karkheh River against WHO drinking water standards

Value/variable	Average of three stations				Max acceptable value (WHO)	Max allowed value (WHO)
	Min	**Max**	**Mean**	**Standard deviation**		
TDS (mg/l)	344.3	1601	857.2	231.5	600	1000
EC (µs/cm)	613	2446	1316.4	360.1	500	1500
Na (mg/l)	1.2	14.2	5.6	2.5	200	
Mg (mg/l)	1	5	2.7	0.74	50	150
Ca (mg/l)	2.5	12.5	5	1.3	75	200
pH (µs/cm)	5.3	8.5	7.9	0.26	6.5–8.5	
Cl (mg/l)	0.95	14.7	5.5	2.5	200	600
SO_4 (mg/l)	1.7	11.7	5.1	1.7	200	400
HCO_3 (mg/l)	1.1	4	2.7	0.43	150	

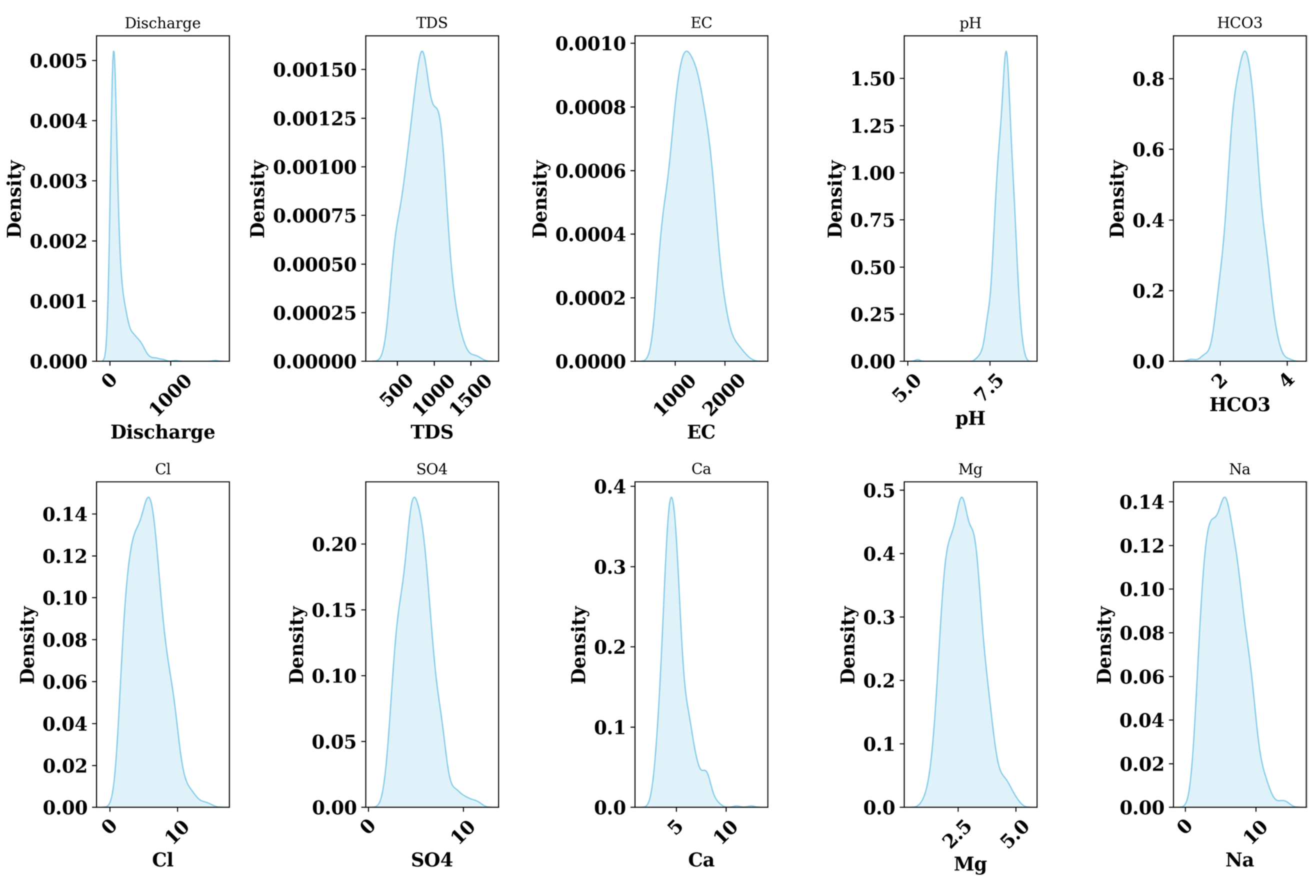

Fig. 2 The density distribution of water quality variables

The data was checked through Kolmogorov–Smirnov normality test, and the variables show normal distribution (TDS, HCO_3, Cl, SO_4, Mg, Na) and discharge; pH and Ca do not show normal distribution (Fig. 2). However, not all data variables are completely normally distributed (Table 2), but the machine learning model selection can be done in a way that this problem gets resolved. In the sense, there are machine learning models that are scale-invariant, and they do not require normal distribution.

In this study, the selected models are those that are not sensitive to this problem. It is also noteworthy that for this study, the data gets normalized in order to avoid data imbalance, but at the end of the modeling, to obtain an estimation of real TDS prediction values versus actual values, the predicted data was rescaled back to normal value ranges.

Wavelet Transform Analysis

In environmental hydrology and water quality investigation, the temporal dynamics of water quality indicators are evaluated using signal analysis to provide useful information for well-informed decision-making. Complex patterns, such as long-term trends and periodic components with low frequencies (about the sampling frequency rate), are commonly observed in the data stream. Understanding the underlying hydrological processes and creating efficient water management plans require a thorough characterization of these patterns [7, 21, 32]. There are methods to find more details from a signal, such as Fourier Transform and Wavelet Transform. As an example, a mathematical method is available, called Fourier Transformation (FT), and is used to break down a signal or function into a linear mixture of sinusoidal functions with different frequencies. Synthesizing a complex waveform from simpler, harmonic components is comparable to this procedure. These periodic functions are expressed as the sum of sine and cosine functions via the Fourier series, a specific application of the FT [53]. [21, 33, 36]).

Table 2 Kolmogorov–Smirnov normality test for the variables in this study

Variables	K-S p-value
Discharge	0
TDS	0.51
EC	0.30
pH	0.01
HCO_3	0.97
Cl	0.17
SO_4	0.40
Ca	0
Mg	0.18
Na	0.09

$$g(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos\left(2\pi n f_0 t\right) + b_n \sin(2\pi n f_0 t) \quad (1)$$

where a_0 is the constant, and this function decomposes the signal to an endless infinite number of cosines and sines. While FT is a powerful tool for signal analysis, it is limited in capturing abrupt temporal changes due to its global frequency representation, and this method has not been used in this study. The sum of sines and cosines is not well localized in terms of time and space [38, 53]. Therefore, to accurately describe a signal, there must be another class of function that is well localized in the time–frequency domain.

Transformation is a sort of mapping that transforms the input function, which includes the place and time, to the output function. The first reason would be to facilitate the modeling process [7, 53]. The main aim of the transformation is to understand and highlight details in the data that conventional exploratory data analysis may miss. The aforementioned detail would be frequency, scale, etc., that can describe the phenomena more clearly. In other words, the main aim of wavelet transform is the extraction of the features (Fig. 3).

In the context of wavelet transform, similarly, there is also a decomposition procedure that takes place similar to what is done in FT but with a different concept. In wavelet analysis, two types of basis functions are typically used: the scaling function (often referred to as the "father wavelet") and the wavelet function (the "mother wavelet").

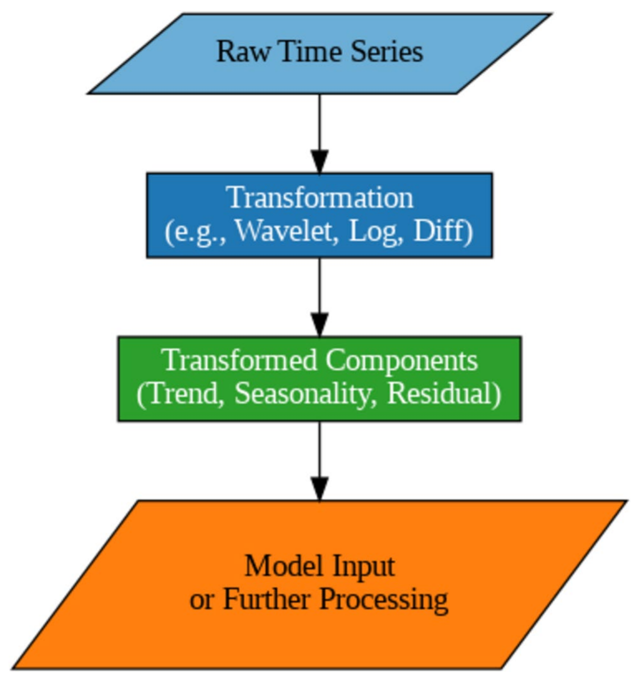

Fig. 3 A general view of using a transformation for the input data to have an output extracted features that are used as the new input for the ML model

$$x(t) \approx \sum_k S_{j,k}\phi_{j,k}(t) + \sum_{j=J}^{1} \sum_k d_{j,k}\psi_{j,k}(t) \quad (2)$$

where the first part is the father wave, and it is good to capture the approximation of the smoothness of the signal; it also captures low frequencies, while the second part is the mother wave of decomposition, and this part is for understanding the detail of the signals [11, 36],Kumar & Foufoula-Georgiou, 1997; [55].

Depending on the decomposition level, the detail and information obtained by the model increase significantly (Fig. 4).

In this context, several terms should be described: scaling and shifting. Scaling refers to the process of stretching or shrinking the signal in time, to show how much a signal is scaled in time. This factor is inversely proportional to frequency. Also, the act of shifting the wavelet function along the time or space axis is referred to in wavelet transforms. This is carried out in order to analyze various signal components that are done in both Time and Frequency [36, 46, 56] (Fig. 5).

In general, there are two different wavelet transforms: (a) Continuous Wavelet Transform (CWT) and (b) Discrete Wavelet Transform (DWT). From a discretization standpoint, the decomposition scale in CWT is finer, which provides more information (depending on the problem's nature), while in DWT, the scale parameter is always discretized to the power of 2.

In this work, CWT is used to analyze both periodic patterns and the trend of the indicators describing the water quality. Since wavelet analysis can capture both temporal and frequency information, CWT was chosen over Fourier transform for the analysis of non-stationary signals with localized fluctuations. Wavelet transform permits multi-resolution analysis, allowing the discovery of transitory patterns and changes in signal over time, in contrast to the Fourier transform, which only offers global frequency content. The outcome of CWT (extracted features from the time series) as input, along with the raw data, is the matrix of complex coefficients, including rows defined as the frequency and columns as time points. For example, for each variable's signal (e.g., Mg), CWT computes the coefficient matrix; the mean power, representing the average energy of the signal across all time and scales; the max power, indicating the strongest localized oscillation observed; and, finally, the dominant scale, which corresponds to the scale (or frequency range) where the signal shows the highest overall energy. There have been many prior research studies investigating the signal analysis in water quality using CWT [12, 29, 36, 45], but in this study, we try to find an optimum way of adding engineered features which might improve the TDS prediction.

In order to create the final dataset, first, CWT was performed for all variables, and then three different datasets were created: (i) only raw measured original dataset, (ii) the original dataset along with engineered features of "all" time series, and (iii) the original dataset along with the engineered features of only "TDS" time series. The results were then compared with different metrics for different models. The main difference between the second and third datasets is that in the second dataset, signal analysis results of all variables were added to the original datasets and used as

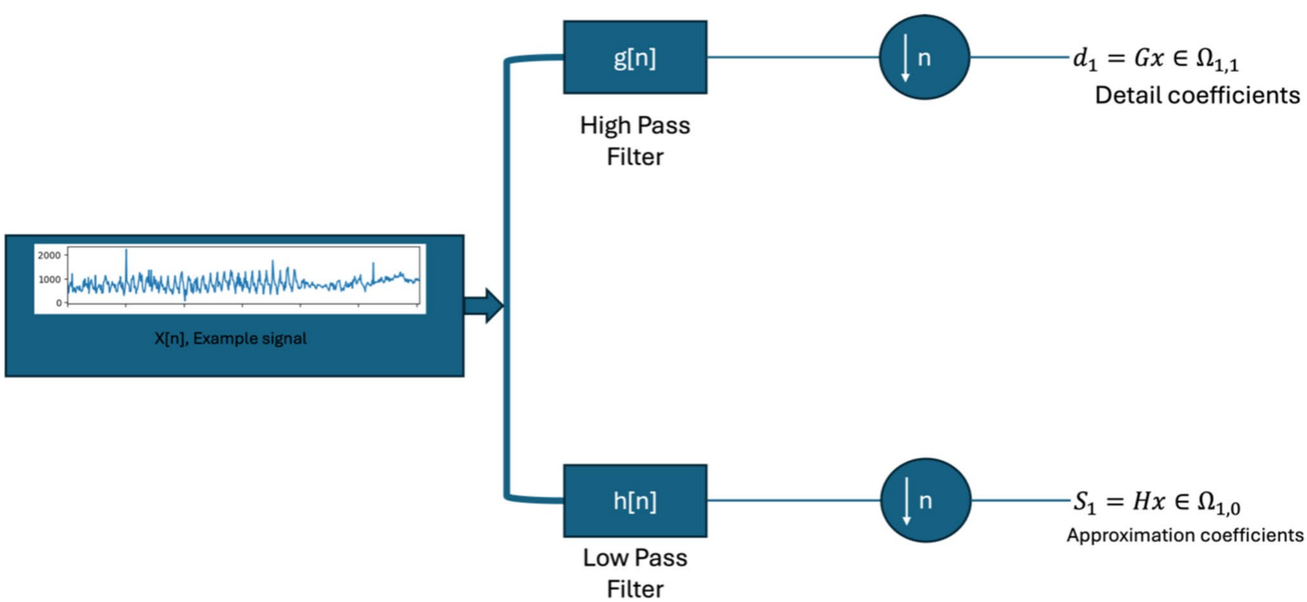

Fig. 4 A systematic and schematic illustration of the way that a signal is checked using both high-pass and low-pass filters to extract features from the signal

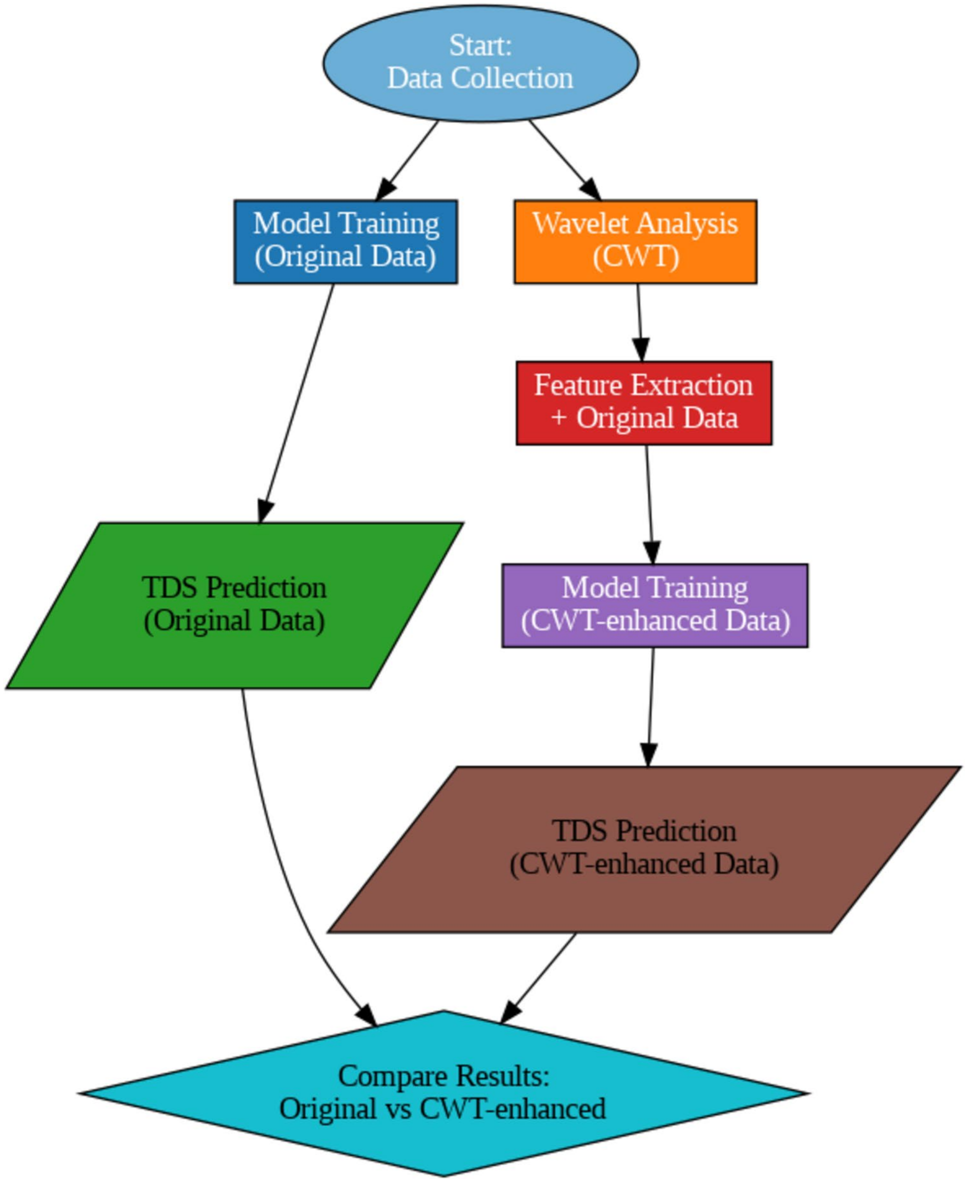

Fig. 5 Flowchart of the signal analysis

the input, while in the third dataset, only the "TDS" signal analysis results were added to the original dataset.

Machine Learning Model

In this paper, several ML algorithms have been employed to predict the TDS level using the extracted features from signal analysis along with the original datasets, including eXtreme Gradient Boosting (hereinafter XGBoost), Random Forest, and Decision Tree. XGBoost is distinct from other decision-based models due to its optimization and parallelizability. XGBoost is an efficient implementation of the gradient boosting technique. It is possible to utilize the model directly for predictive regression modeling [1], Bentéjac et al., 2019; [8, 10]. Random Forest is an ensemble learning model that incorporates multiple Decision Trees (base models) to generate a single, often more robust and accurate prediction. To arrive at a single outcome, Random Forest aggregates the predictions of these individual Decision Trees using methods like averaging or majority voting. In contrast, a single Decision Tree is a standalone supervised learning algorithm that makes predictions based on a hierarchical set of decisions derived from the training data [15, 22, 26, 28, 51].

As can be seen in Fig. 6, the process of reaching a prediction is done via asking questions and answering. In order to check whether or not the model has performed the process of training and testing properly, there are a number of metrics that quantitatively evaluate this subject. The metrics that have been used to check the performance are as follows:

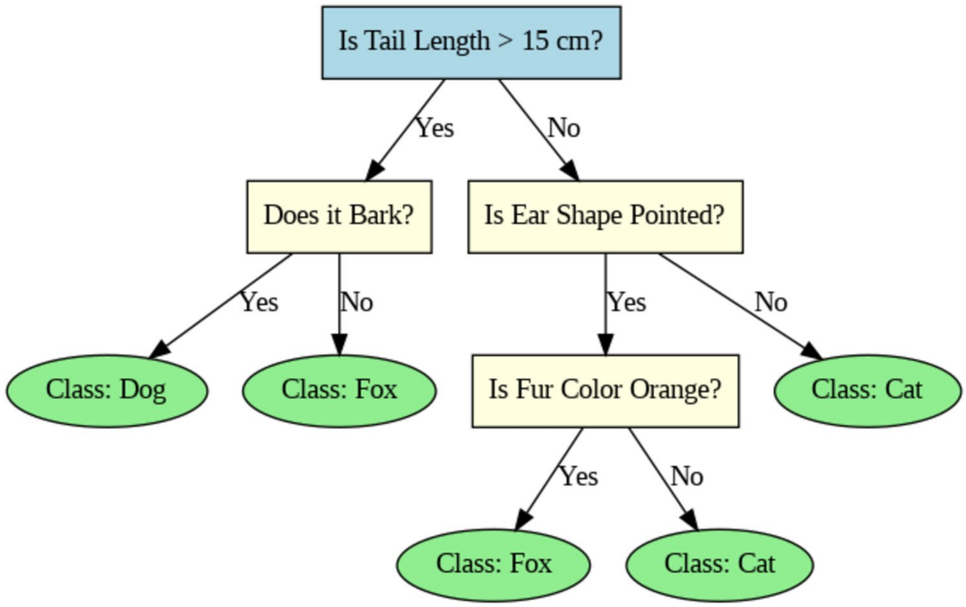

Fig. 6 A schematic view of tree-based ML algorithms for a classification problem to achieve the best prediction through questioning and answering

$$MSE = 1/n \sum (y_i - \hat{y}_i)^2 \tag{3}$$

$$R^2 = 1 - \left(\frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}\right) \tag{4}$$

where y_i is the actual value in the dataset, $\hat{y}_i$ is the predicted data using trained model, and $\bar{y}$ is the mean of the actual data. R^2 measures proportion of variance in the dependent variable explained by the model, including all ensemble models, e.g., XGBoost model (Mitchell, 1999; Miura et al., 2023; Uddin et al., 2022), and *MSE* quantifies the prediction error as the average of squared differences between observed and predicted values. R^2 reports the fact that the model could have captured the complexities of the dataset, and it can be generalized to the other datasets, while *MSE* reports that the existing model for this current dataset has predicted with an accuracy between the actual data and predicted data.

Model architecture plays a crucial role in machine learning performance. The hyperparameters employed in this study are listed in Table 3. It is essential to compare which machine learning takes more advantage of adding more features (extracted by CWT) to predict the target; thus these models have been chosen to have a quantitative comparisons between different models.

In this study, the ratio between train and test is 70/30%. To elaborate more, the train set is 70% of the whole dataset in every scenario, and 30% is considered for the test. This is an essential step that should be done for the prevention of overfitting.

Results and Discussions

The study's findings show that combining machine learning (ML) with continuous wavelet transform (CWT) is an effective way to forecast total dissolved solids (TDS) in Iran's Karkheh River. Power spectra analysis is used to measure frequency-domain dynamics, signal decomposition is used to uncover temporal patterns, and TDS prediction performance is examined across several dataset circumstances in

Table 3 Model characteristics and architecture

Random forest				
Criterion	**Max depth**	**Splitter**	**ccp_alpha**	**min_samples_split**
Squared_error	none	Best	0.0	2
XGBoost				
Learning rate	**Max depth**	**Objective**	**Sampling method**	
0.3	6	reg:linear	Uniform	
Decision tree				
Criterion	**Max depth**	**Splitter**	**ccp_alpha**	**min_samples_split**
Squared_error	none	Best	0.0	2

this section. In order to meet the demand for precise, scalable water quality monitoring in semi-arid areas, this study combines CWT with ensemble machine learning models to forecast TDS in the Karkheh River. The succeeding part of this section tries to delve into deeper lessons learned from the integration of signal decomposition and machine learning models, evaluates the model's performance against earlier research, and considers the management implications for water quality.

Signal Decomposition and Feature Extraction and Power Spectra Analysis

This section presents the result of Continuous Wavelet Transform (CWT) that was applied to decompose the water quality time series to analyze the seasonality, trend, and residual for the variable. The results regarding some of the variables are presented in the Fig. 7, and the rest of the variables' details are presented in the supplementary document. As can be seen in Fig. 7, EC and TDS generally show an upward trend. They have approximately similar seasonality patterns which have been repeated in time. The CWT result for pH is different as it has downtrend behavior in time, and also, its seasonality pattern differs from those of the other variables.

In terms of HCO_3, there are fluctuations in the trend, as seen also in the signal itself; however, the seasonality does not seem similar to the abovementioned variables. This cyclical behavior is due to the nature of the variable and the climatic condition, but the change in the trend can be a result of climate change and direct/indirect anthropogenic activities (e.g., water sewage and increasing environmental pollution). Since this research deals with time-dependent and time-independent variables, it is essential to check the stationary condition of the variables within the dataset.

Figure 8 illustrates non-stationary parameters that have been separated to check how they are when they get detrended and deseasonalized to compare the signal and do the analysis. In CWT analysis, the preprocessing processes of detrending and deseasonalizing guarantee that the attention is on the pertinent, localized dynamics of the data, unhindered by regular periodicities or large-scale trends. This clarifies the results by exposing the actual time–frequency features of the data. Among different parameters, EC, TDS, SO_4, and Ca are not in steady state condition. Then the power spectrum analysis has been conducted for

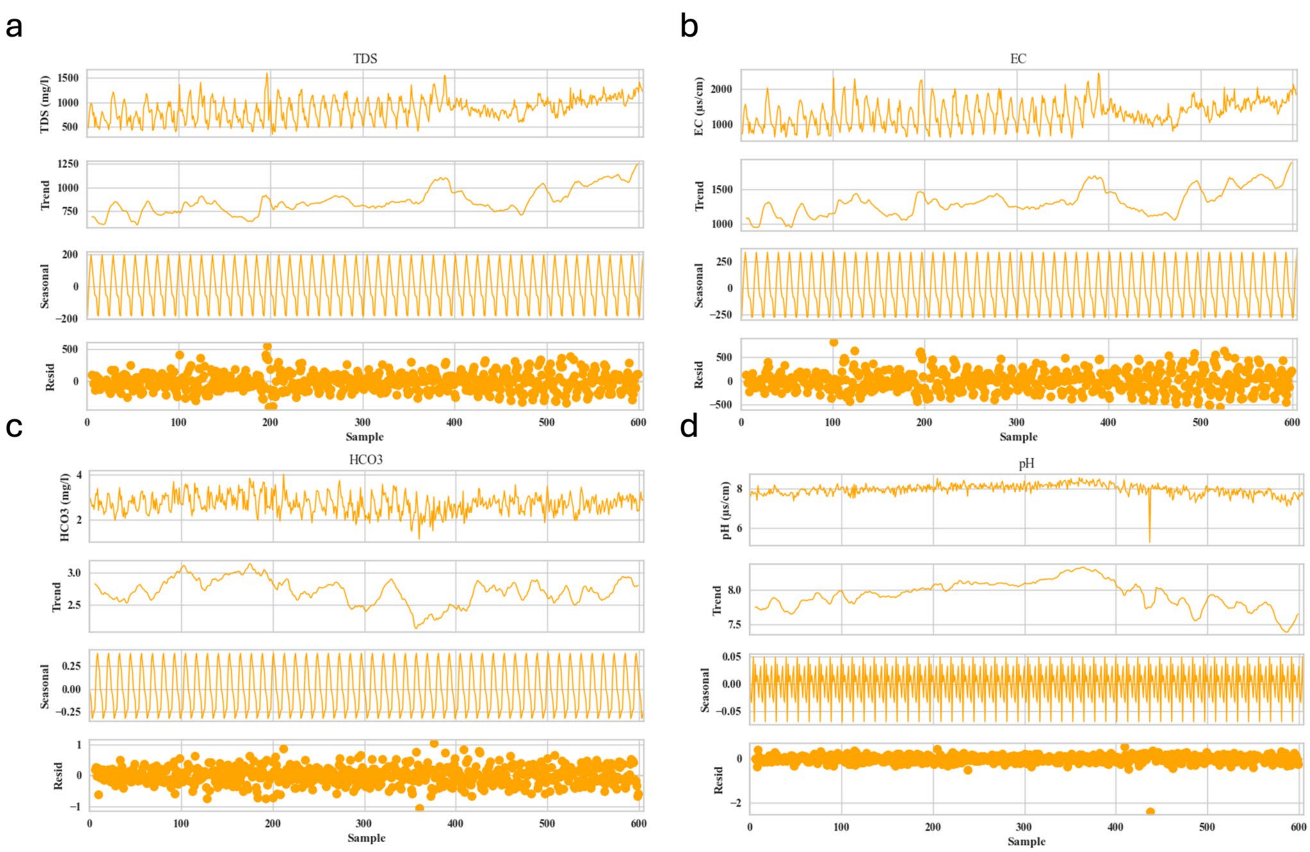

Fig. 7 Different variables signal decomposed into their trends, seasonality, and residuals for different variables **a** TDS, **b** EC, **c** HCO_3, and **d** pH. The plots start with the "signal" and are followed by its "trend, seasonality, and residuals"

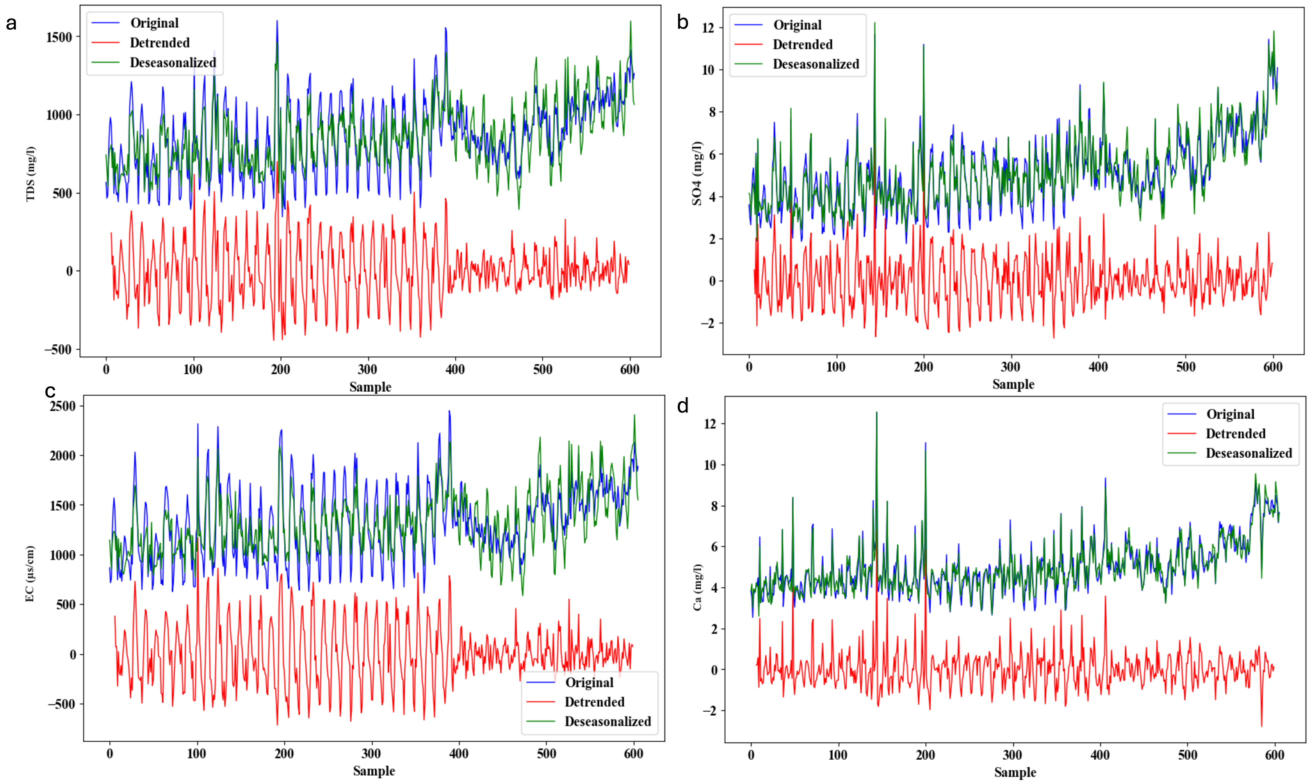

Fig. 8 Non-stationary variables **a** TDS, **b** SO_4, **c** EC, and **d** Ca that have been detrended and deseasonalized to have a clearer insight into the signals to check the power spectra as well

both stationary and non-stationary variables. Power spectra analysis using CWT reveals the frequency-domain dynamics of water quality variables, with and without detrending, to isolate localized periodicities. Figures 9 and 10 show the wavelet power spectra for some of the key variables before and after detrending, respectively. In both plots—with and without the detrended process—the TDS level shows some significance during different periods. As an example, strong seasonal and annual cycles occurred during the 1980 s and early 2000s. In the normal (without detrending) analysis, after 2000, long-term trends are dominant, while in the detrended plot, the TDS level shows shorter periodicities (e.g., 10–20 months). These findings are reflected in red/yellow patches observed during different periods. “Ca” also shows a strong dominance during the mid-80s and weaker signals later (after the 2000 s). Both “Ca” and “TDS” illustrate periodic behavior on the same time scale, which means changes in “Ca” concentration influence the “TDS” level. In the long term, both “TDS” and “Ca” could stem from shared environmental factors, such as an increase or decrease in Ca, causing changes in “TDS” level over time (which is due to the human activities and climate change).

This behavior (similarity between variables, e.g., Ca, Cl, SO_4, HCO_3 on the TDS level) can be seen in other plots. SO_4 shows a dominant long-term trend at larger periods. This again confirms that TDS level is influenced by SO_4 as both Ca and SO_4 are contributors to the total dissolved solids concentration. Similar to Ca, there is a temporal alignment in the power spectra of SO_4 in the mid-80 s. These shared periodic patterns suggest that factors such as rainfall, runoff, and other factors like industrial pollution impact the TDS level simultaneously. Detrended spectra provide clearer insight into how the periodicities in SO_4 and Ca align with the TDS level.

Different periodicities and temporal dynamics throughout the time series are revealed by the variables’ wavelet power spectra. At lower frequencies (longer periods), the majority of variables show substantial power concentrations, especially around annual to multi-annual scales. This suggests that seasonal and possibly interannual climatic or hydrological cycles are influencing these parameters. A substantial seasonal component that is congruent with anticipated hydrological cycles in river systems is indicated by variables like discharge, TDS, and SO_4, which display notable bands of power linked with 12- to 24-month intervals.

TDS Prediction Performance

As it was mentioned in this study, we investigated the impact of incorporating the features driven by Continuous Wavelet

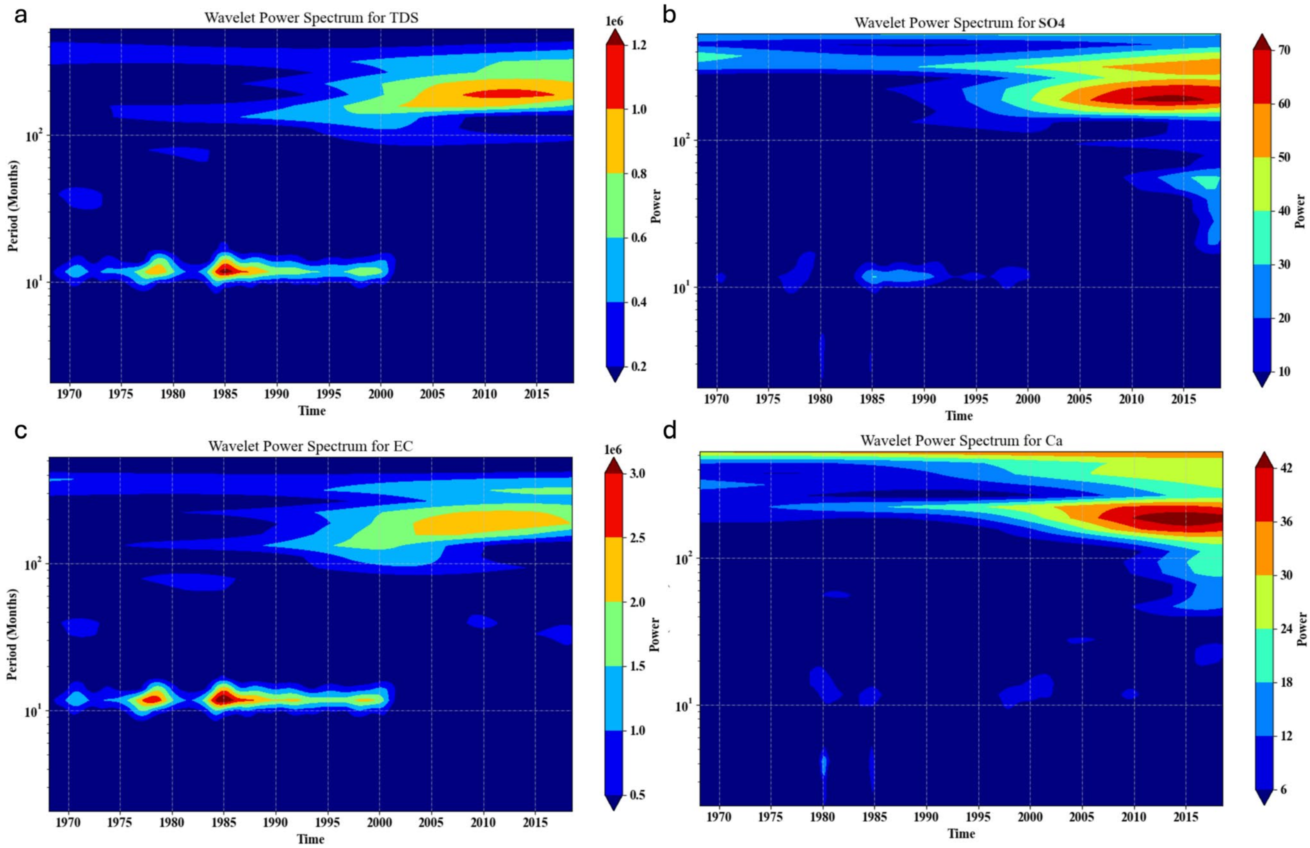

Fig. 9 Power spectra of the variables **a** TDS, **b** SO_4, **c** EC, and **d** Ca

Transform (CWT) on the predictive performance of machine learning models for TDS estimation. The analysis compared three main scenarios: (i) using only raw measured data; (ii) combining raw data with the "all" signal-base characteristics, the features that are extracted from CWT analysis; and (iii) combining raw data with only "TDS" signal-base characteristics, the features that are extracted from CWT analysis. The models showed different performance**;** however**,** "XGBoost" outperformed the other algorithms. Table 4 shows the model performance for each algorithm using the metrics.

As indicated in Table 4, XGBoost model showed more advantages of adding features to the original dataset in predicting the TDS level. It trained on measured data only, and it could achieve an R^2 of 0.93, indicating that the model could explain around 93% of the variance in the TDS values. When the data was augmented using CWT-driven features only from "TDS", the R^2 score improved to 0.97, reflecting a slight yet meaningful increase in the model's ability to capture the target values' variability. Additionally, integrating CWT-driven feature to the dataset (measured data + TDS signal CWT-driven features) resulted in approximately 30% reduction in Root Mean Squared Error (RMSE). This reduction in RMSE and improvement in R^2 reports that TDS predictions were close to the actual values. In water quality problems, this RMSE is significant, since even small errors in the prediction (in early warning systems) can influence the decisions related to water treatment. This result shows only that adding TDS's signal characteristics would improve the performance of the model and increase the accuracy of TDS prediction which is definitely crucial for early warning systems and water treatment designations that is done by policy makers. Considering all variables' time series via signal analysis would improve the prediction; however, this enhancement is more computationally expensive compared to only TDS signal characterization.

The use of CWT for variables, because of its capacity to break down the signal (in this case TDS) into time–frequency components, may uncover patterns like periodic oscillations or transient changes that may not be fully represented by raw sensor data. The observed increases in RMSE (by about 30% reduction) and R^2 (from 0.93 to 0.97) show that these characteristics successfully captured extra temporal dynamics, boosting the predictive potential of the model [4, 17, 45, 48]. The other algorithms including Random Forest, and Decision Tree, could have improved due to the existence of engineered features from TDS singal analysis but not as much as XGBoost. Also, when all variables' time

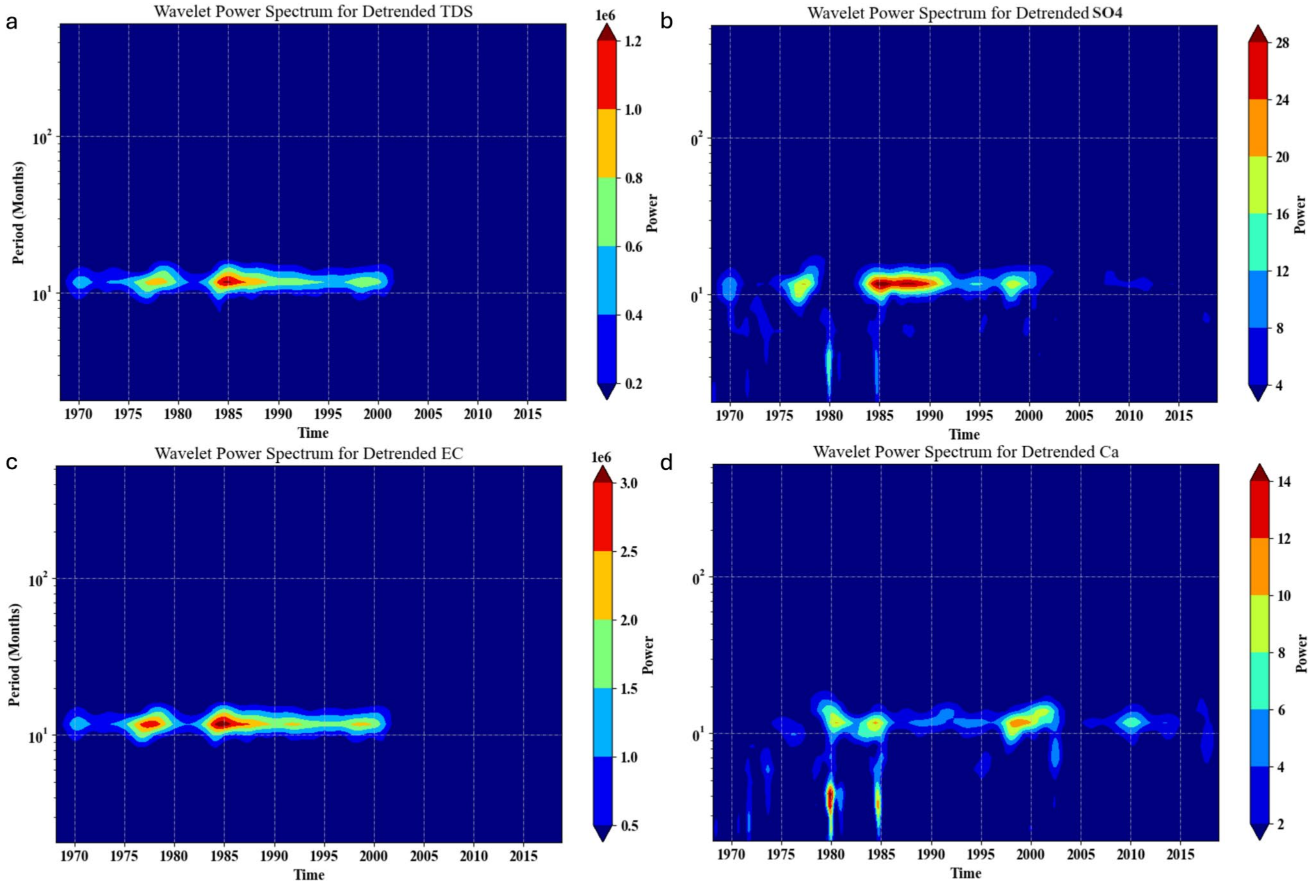

Fig. 10 Power spectra of detrended the variables **a** TDS, **b** SO_4, **c** EC, and **d** Ca

Table 4 Model performance in different input data condition

Model/metrics	R^2	RMSE
Only original measured data		
Random forest	0.94	57.03
XGBoost	0.93	61.96
Decision tree	0.89	81.21
Combination of engineered features of “All” variable + original data		
Random forest	0.94	55.09
XGBoost	**0.95**	**52.01**
Decision tree	0.88	85.1
Combination of engineered features of only for “TDS” variable + original data		
Random forest	0.95	51.63
XGBoost	**0.97**	**48.01**
Decision tree	0.93	62.10

The bold values show the significance of the values and shows the better prediction

series characteristics were added to the original dataset, the Decision Tree shows an increase in RMSE and a decrease in R2 which means underperformance in terms of TDS prediction. This shows that having more features in some machine learning models does not necessarily lead to an improvement in prediction accuracy [9].

Given these metrics information, the results regarding the model are illustrated in Figs. 11, 12, and 13. In Fig. 11 (scatter plot of actual vs. predicted TDS levels), the datapoints are scattered around the best fit line. As it is seen in Fig. 11, when “TDS” signal characteristics get added to the raw dataset and used for the prediction, the prediction results improved compared to the time that only raw data is considered or the time the all variables are analyzed through signal processing and then added to the dataset.

Figure 12 (time series comparison over 2008–2018) further shows this improvement, showing that the predicted TDS values from the combined dataset (orange dashed line) closely track the actual TDS measurements (black line), particularly during peak events (e.g., around 2010, 2012, and 2018). In contrast, the model trained on measured data alone (blue dashed line) exhibits larger deviations during these periods, underscoring the value of CWT features in capturing temporal fluctuations.

According to our comparison research, these results have important practical implications for real-world water

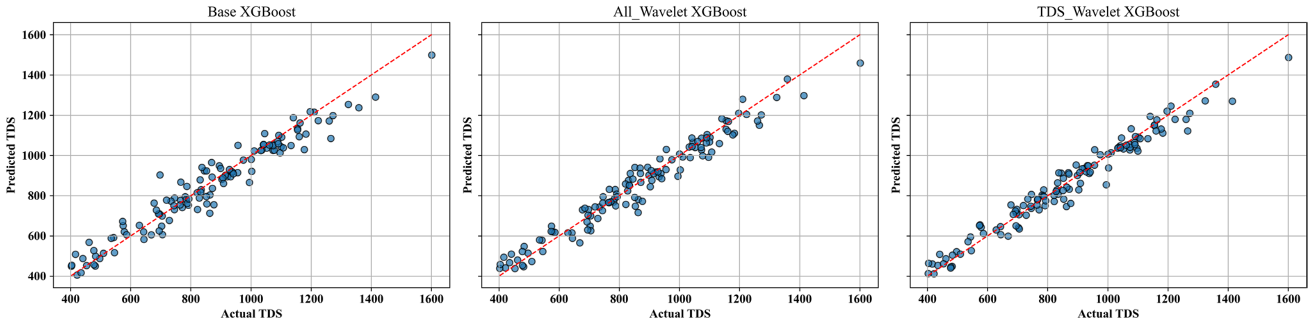

Fig. 11 Scatter plot predicted data vs actual data, when only measured data was used for TDS prediction (base XGBoost), when features of all variables are extracted via CWT (All_Wavelet XGBoost), and, ultimately, only "TDS" signal features are added to the original dataset (TDS_Wavelet XGBoost)

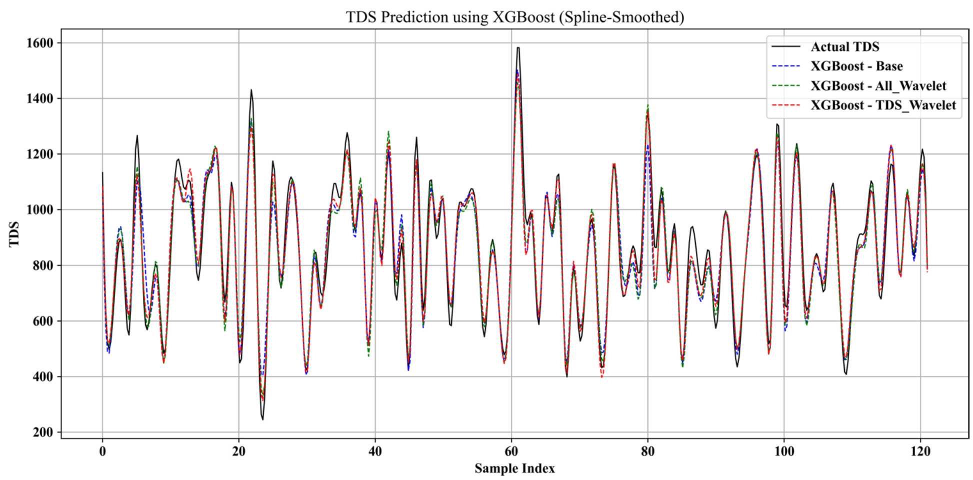

Fig. 12 Comparison between the predicted and actual time series of TDS level

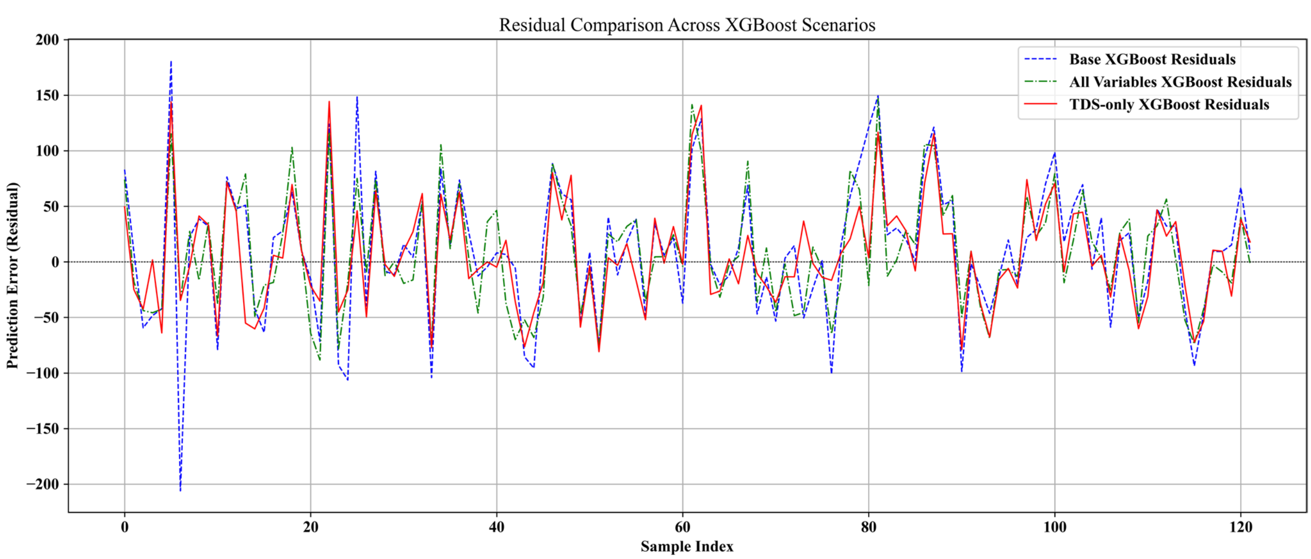

Fig. 13 Comparison between the predicted and actual time series of TDS level

monitoring and real-time water treatment. The accuracy of TDS predictions is improved by the 30% RMSE reduction, which is essential for real-world water quality monitoring, such as ensuring the safety of drinking water or evaluating environmental health. The model's applicability for treatment and regulatory decisions is further supported by the enhanced R^2 (0.97), which shows a greater connection between expected and actual TDS values.

Because of its capacity to simulate non-linear interactions and its resilience to noisy, high-dimensional water quality data, Random Forest showed better performance as it is a good option for integrating the complexity of CWT-derived features. Notwithstanding these improvements, there may be room for more optimizations, given the slight increase in R^2 (0.04) and the 30% RMSE decrease.

Figure 13 shows the residual error of the prediction versus actual data when a different scenario is considered for the model training and testing. As it is seen when only "TDS" characteristics are added to the raw dataset, the results are much closer to the real values, which in turn cause the residual to become closer to zero. Comparing it with the original data, it is seen that TDS prediction using the original data has a higher range of errors with respect to the zero-crossing line.

Water Quality Monitoring Implications, Limitations and Future Works

This proposed framework suggests a new way for water quality modeling in riverine system for the TDS evaluation. In semi-arid environments like west Asian countries like Iran, that is suffering from water scarcity, climate change, urbanization and industrialization have put the water quality condition in a highly crucial risk. This new TDS prediction shows better performance in terms of the prediction (Figs. 11 and 12). The CWT analysis along with Machine learning implementations (Figs. 7, 8, and 9) shows the importance of using the hidden features within a time series that can add value to the data set and improve the ML modeling. Aside from TDS, other hazards to aquatic habitats include algal blooms and other stresses brought on by pollutants, which can have a negative effect on ecosystems and human health. It is crucial to anticipate changes in climate, aquatic ecosystem disruption, oxygen depletion, and water usability early on since these elements could pose a health risk to the general public.

The accuracy and robustness of predictive models are paramount for effective water quality policymaking. Poorly designed methodologies or models with low predictive power can lead to misinformed policies, resulting in inadequate treatment measures or resource misallocation. The CWT-ML approach, by contrast, provides a reliable foundation for policy design by extracting meaningful features from complex datasets, as demonstrated in the improved TDS predictions. This accuracy is crucial for crafting policies that address immediate water quality concerns while anticipating future risks. For example, early warning systems built on this framework (Fig. 14) can guide the designation of mitigation strategies, such as adjusting treatment processes or regulating industrial discharges, to prevent water quality degradation. Designing an early warning system based on the integration of CWT and ML models could level up the mitigation strategy designation appropriately. Limitations in this study were ensuring the data gaps that might cause minor biases. Another limitation would be the computational costs that might occur for real-time implementation. To this end, future studies should be focused on model optimization and reduction in computational costs. This study suggests the incorporation of "Causal Inference" modeling within designing early warning systems for mitigation strategy, to not only understand the features' correlation but also to understand the causal relationships between variables. Since this study was aiming to understand the effect of the signal analysis on the TDS prediction and not looking to understand the importance of the features, it is suggested to conduct feature importance analysis to realize the variable's contribution in the model prediction. Further analysis can be done using "SHapply Additive exPlainations or SHAP" and "LIME".

Conclusion

The available freshwater bodies are under severe stress which makes the studies regarding the water quality significantly important. Among different methods of water quality investigation/prediction, data-based solutions are making an important contribution to level up the quality of the studies as the availability of the observation is increasing and the necessity to process the actual data observed in the environment has been essential. In this study, we tried to use not only machine learning models to predict the water quality (in this case, Total Dissolved Solids or TDS), but also, in order to employ the ML models, we tried to see if we can find more information from the time series through signal analysis that can lead to a better TDS prediction or not. The observed data was treated as signals that could have potentially features that need extraction. Then, according to the comparison approach, three scenarios were defined: (i) using only original data to predict the TDS level, (ii) adding features extracted by CWT for the all variables to the original dataset and TDS prediction, and finally, (iii) only considering "TDS" signal characteristics and adding them to the original data. Different signals had different features including seasonality, trends and residuals. For this study, XGBoost, Random Forest, and Decision Tree were employed. Among different machine learning algorithms,

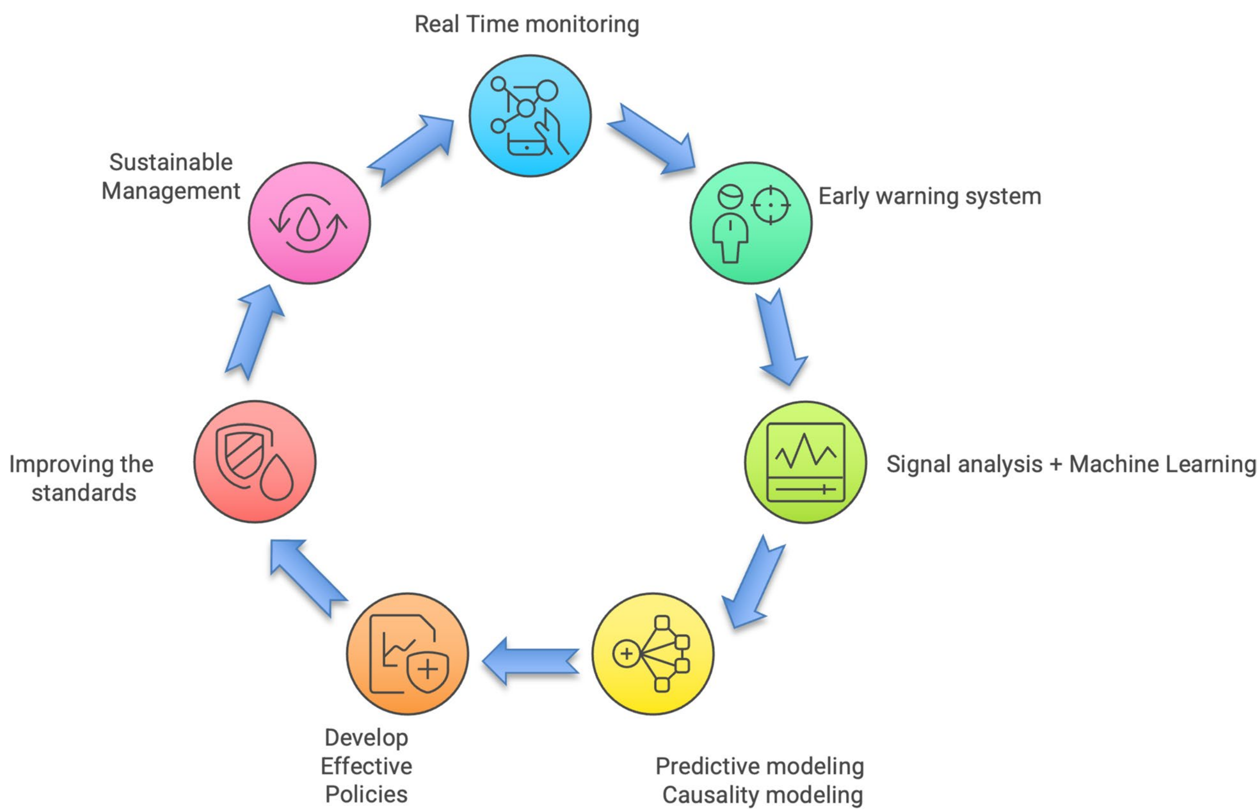

Fig. 14 Strategy designing and designation process for water quality assessment

XGBoost showed a better performance with an R^2 score of 97% and an RMSE of around 48 when the TDS signal characteristics were added to the measured data and used for TDS prediction with significantly lower RMSE, while using only measured data had an RMSE of around 62. The accuracy of the prediction which was over 95% showed that the signal analysis had the potential to be employed along with the original data as it can gain more information about the fluctuations of the variables within the duration of the measurement. Although, the results indicated an improvement in the TDS level prediction by considering the signal analysis and adding the time series characterization to the original dataset, not all variables can make a significant contribution to the prediction accuracy improvement. Considering the signal analysis for only the target can significantly improve the TDS prediction while this reduction in signal analysis is relatively cheaper than analyzing all signals. This study offers a scalable and novel way of TDS prediction by integrating signal analysis and machine learning in predicting the TDS level.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s41101-025-00390-z.

Acknowledgements The authors appreciate the Iran Water Resources Management Company (a subset of the Ministry of Energy) in supplying the part of the required data for the real-case studies described in this paper.

Author Contribution H.A. conceptualized, investigated, analyzed, Visualized, wrote and reviewed the main text. F.F. did the data curation, visualized, wrote and reviewed the main text. J.Y.D supervised. R.Sh. wrote and reviewed the main text. H.N investigated and reviewed the main text. M.Y.L wrote the revision text. B.Z supervised and reviewed the revision text. R.A supervised and reviewed the revision text. All authors reviewed the manuscript.

Data Availability The data is available upon request.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Amini H, Monegaglia F, Shakeri R, Tubino M, Zolezzi G (2024) Meanders on the move: can AI-based solutions predict where they will be located? Water 16(17):2460. https://doi.org/10.3390/w16172460
2. Alfitrah SP, Kirana KH, Fitriansyah RA, Shafaria M (2024) Preliminary study of pH, EC, TDS, and heavy metals analysis of Oxbow's sediments in Bandung region. Phi Jurnal Pendidikan Fisika dan Terapan 10(1):98–105. https://doi.org/10.22373/p-jpft.v10i1.22714
3. Arab M (2024) Investigating water policies and managing hydro politic challenges in border areas with an analytical view of Iran's climate. Power Syst Technol 48(2):583–610
4. Asadollahfardi G, Taklify A, Ghanbari A (2012) Application of artificial neural network to predict TDS in Talkheh Rud River. J Irrig Drain Eng 138(4):363–370. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000402
5. Ashraf Vaghefi S, Mousavi SJ, Abbaspour KC, Srinivasan R, Yang H (2014) Analyses of the impact of climate change on water resources components, drought, and wheat yield in semi-arid regions: Karkheh River Basin in Iran. Hydrological processes 28(4):2018–2032. https://doi.org/10.1002/hyp.9747
6. Bentéjac C, Csörgő A, Martínez-Muñoz G (2020) A comparative analysis of gradient boosting algorithms. Artif Intell Rev 54:1937–1967. https://doi.org/10.1007/s10462-020-09896-5
7. Coraggio E, Han D, Gronow C, Tryfonas T (2022) Water quality sampling frequency analysis of surface freshwater: a case study on Bristol Floating Harbour. Frontiers in Sustainable Cities 3:791595. https://doi.org/10.3389/frsc.2021.791595
8. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. Proc 22nd ACM SIGKDD Int Conf Knowledge Discov Data Mining - KDD '16 1(1):785–794. https://doi.org/10.1145/2939672.2939785
9. Chen J, Ooi LQR, Tan TWK, Zhang S, Li J, Asplund CL, Eickhoff SB, Bzdok D, Holmes AJ, Yeo BTT (2023) Relationship between prediction accuracy and feature importance reliability: an empirical and theoretical study. Neuroimage 274:120115. https://doi.org/10.1016/j.neuroimage.2023.120115
10. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. https://doi.org/10.1145/2939672.2939785
11. Chin CH, Abdullah S, Ariffin AK, Singh SSK, Arifin A (2024) A review of the wavelet transform for durability and structural health monitoring in automotive applications. Alex Eng J 99:204–216. https://doi.org/10.1016/j.aej.2024.04.069
12. Dökmen F, Aslan Z (2013) Evaluation of the parameters of water quality with wavelet techniques. Water Resour Manage 27(14):4977–4988. https://doi.org/10.1007/s11269-013-0454-5
13. Eid MH, Eissa M, Mohamed EA, Ramadan HS, Tamás M, Kovács A, Szűcs P (2024) New approach into human health risk assessment associated with heavy metals in surface water and groundwater using Monte Carlo method. Sci Rep 14(1):1008. https://doi.org/10.1038/s41598-023-50000-y
14. Ejaz U, Khan SM, Jehangir S, Ahmad Z, Abdullah A, Iqbal M, Svenning JC (2024) Monitoring the Industrial waste polluted stream-Integrated analytics and machine learning for water quality index assessment. J Clean Prod 450:141877. https://doi.org/10.1016/j.jclepro.2024.141877
15. Everaert G, Bennetsen E, Goethals PLM (2016) An applicability index for reliable and applicable decision trees in water quality modelling. Eco Inform 32:1–6. https://doi.org/10.1016/j.ecoinf.2015.12.004
16. European union (2006) Revised bathing water directive
17. Farzana SZ, Paudyal DR, Chadalavada S, Alam MJ (2024) Temporal dynamics and predictive modelling of streamflow and water quality using advanced statistical and ensemble machine learning techniques. Water 16(15):2107. https://doi.org/10.3390/w16152107
18. Dohan K, Whitfield PH (1997) Identification and characterization of water quality transients using wavelet analysis. I. Wavelet analysis methodology. Water Sci Technol 36(5):325–335. https://doi.org/10.1016/S0273-1223(97)00490-3
19. Dong J, Wang Z, Wu J, Huang J, Zhang C (2023) A water quality prediction model based on signal decomposition and ensemble deep learning techniques. Water Sci Technol 88(10):2611–2632. https://doi.org/10.2166/wst.2023.357
20. Edition F (2011) Guidelines for drinking-water quality. WHO Chronicle 38(4):104–108
21. Hamidi M, Sohrabi MR, Tehrani MS, Nik SM (2024) Continuous wavelet transform and integration of discrete wavelet transform with principal component analysis and fuzzy inference system for the simultaneous determination of ethinyl estradiol and drospirenone in combined oral contraceptives. Spectrochim Acta Part A Mol Biomol Spectrosc 320:124541. https://doi.org/10.1016/j.saa.2024.124541
22. Ha HT, Nguyen DTB, Stoeckel T (2024) What is the best predictor of word difficulty? A case of data mining using random forest. Language testing. https://doi.org/10.1177/02655322241263628
23. Hammoumi D, Al-Aizari HS, Alaraidh IA, Okla MK, Assal ME, Al-Aizari AR, Moshab MS, Chakiri S, Bejjaji Z (2024) Seasonal variations and assessment of surface water quality using water quality index (WQI) and principal component analysis (PCA): a case study. Sustainability (Switzerland) 16(13):5644. https://doi.org/10.3390/su16135644
24. Hatvani IG, Clement A, Korponai J, Kern Z, Kovács J (2017) Periodic signals of climatic variables and water quality in a river–eutrophic pond–wetland cascade ecosystem tracked by wavelet coherence analysis. Ecol Ind 83:21–31. https://doi.org/10.1016/j.ecolind.2017.07.018
25. He L, Huang GH, Zeng GM, Lu HW (2008) Wavelet-based multiresolution analysis for data cleaning and its application to water quality management systems. Expert Syst Appl 35(3):1301–1310. https://doi.org/10.1016/j.eswa.2007.08.009
26. Ha HT, Nguyen DTB, Stoeckel T (2024) What is the best predictor of word difficulty? A case of data mining using random forest. Lang Test 41(4):828–844. https://doi.org/10.1177/02655322241263628
27. Islam MS, Yin H, Rahman M (2024) Long-term trend prediction of surface water quality of two main river basins of China using machine learning method. Procedia Comput Sci 236:257–264. https://doi.org/10.1016/j.procs.2024.05.029
28. Jaafari A, Zenner EK, Pham BT (2018) Wildfire spatial pattern analysis in the Zagros Mountains, Iran: A comparative study of decision tree based classifiers. Eco Inform 43:200–211. https://doi.org/10.1016/j.ecoinf.2017.12.006
29. Jamei M, Ahmadianfar I, Chu X, Yaseen ZM (2020) Prediction of surface water total dissolved solids using hybridized wavelet-multigene genetic programming: new approach. J Hydrol 589:125335. https://doi.org/10.1016/j.jhydrol.2020.125335
30. Jamshidzadeh Z, Latif SD, Ehteram M, Sheikh Khozani Z, Ahmed AN, Sherif M, El-Shafie A (2024) An advanced hybrid deep learning model for predicting total dissolved solids and electrical conductivity (EC) in coastal aquifers. Environ Sci Eur 36(1):20. https://doi.org/10.1186/s12302-024-00850-8
31. Jannat JN, Islam ARMT, Mia MY, Pal SC, Biswas T, Jion MMMF, Senapathi V (2024) Using unsupervised machine learning models to drive groundwater chemistry and associated health risks in Indo-Bangla Sundarban region. Chemosphere 351:141217. https://doi.org/10.1016/j.chemosphere.2024.141217
32. Jayaraman P, Nagarajan KK, Partheeban P, Krishnamurthy V (2024) Critical review on water quality analysis using IoT

and machine learning models. Int J Info Manag Data Insights 4(1):100210. https://doi.org/10.1016/j.jjimei.2023.100210
33. Jiang J, Zheng Y, Pang T, Wang B, Chachan R, Tian Y (2020) A comprehensive study on spectral analysis and anomaly detection of river water quality dynamics with high time resolution measurements. J Hydrol 589:125175. https://doi.org/10.1016/j.jhydrol.2020.125175
34. Jie-xing Z, Jing W, Peng-fei W, Yang H, Bin LIU, Jing L (2012) Wavelet analysis of water quality changes in Dianchi Lake during the past 7a. Procedia Earth Planet Sci 5:280–288. https://doi.org/10.1016/j.proeps.2012.01.048
35. Kamali S, Saedi F, Asghari K (2023) Prediction of flood in Karkheh Basin using data-driven methods. ISPRS Ann Photogram, Remote Sens Spatial Info Sci 10:349–354. https://doi.org/10.5194/isprs-annals-X-4-W1-2022-349-2023
36. Kang S, Lin H (2007) Wavelet analysis of hydrological and water quality signals in an agricultural watershed. J Hydrol 338(1–2):1–14. https://doi.org/10.1016/j.jhydrol.2007.01.047
37. Karimi M, Tabiee M, Karami S, Karimi V, Karamidehkordi E (2024) Climate change and water scarcity impact on sustainability in semi-arid areas: Lessons from the South of Iran. Groundw Sustain Dev 24:101075. https://doi.org/10.1016/j.gsd.2023.101075
38. Kumar P, Foufoula-Georgiou E (1997) Wavelet analysis for geophysical applications. Rev Geophys 35(4):385–412. https://doi.org/10.1029/97RG00427
39. Liu B, Li T (2024) A machine-learning-based framework for retrieving water quality parameters in urban rivers using UAV hyperspectral images. Remote sensing 16(5):905. https://doi.org/10.3390/rs16050905
40. Liu H, Chen Z, Wang Z, Xu M, Wang Y, Geng J, Yin F (2024) Water quality system informatics: an emerging inter-discipline of environmental engineering. Engineering 43:115–124. https://doi.org/10.1016/j.eng.2024.03.018
41. Mawari G, Kumar N, Sarkar S, Frank AL, Daga MK, Singh MM, Singh I (2022) Human health risk assessment due to heavy metals in ground and surface water and association of diseases with drinking water sources: a study from Maharashtra, India. Environ Health Insights 16:11786302221146020. https://doi.org/10.1177/11786302221146020
42. Merrikhpour H, Badamfirooz J, Moridi A, Mousazadeh R, Zarandian A, Shiranzaei S (2024) Water and soil resource assessment using GIS-based modeling of climate variability (WaterWorld Policy Support System) in Karkheh hydrological unit-southwestern Iran. https://doi.org/10.21203/rs.3.rs-3852779/v1
43. Ministry of Energy of Iran (MOE), 2023
44. Morovati R, Kisi O (2024) Utilizing hybrid machine learning techniques and gridded precipitation data for advanced discharge simulation in under-monitored river basins. Hydrology 11(4):48. https://doi.org/10.3390/hydrology11040048
45. Montaseri M, Zaman ZadGhavidel S, Sanikhani H (2018) Water quality variations in different climates of Iran: toward modeling total dissolved solid using soft computing techniques. Stochastic Environ Res Risk Assess 32(8):2253–2273. https://doi.org/10.1007/s00477-018-1554-9
46. Nguyen TD, Nguyen PD (2024) Improvements in the wavelet transform and its variations: concepts and applications in diagnosing gearbox in non-stationary conditions. Appl Sci 14(11):4642. https://doi.org/10.3390/app14114642
47. Noori A, Mohajeri SH, Delnavaz M, Rezazadeh R (2023) A spatiotemporal monitoring model of TSM and TDS in arid region lakes utilizing Sentinel-2 imagery. J Arid Environ 216:105024. https://doi.org/10.1016/j.jaridenv.2023.105024
48. Roushangar K, Davoudi S, Shahnazi S (2024) Temporal prediction of dissolved oxygen based on CEEMDAN and multi-strategy LSTM hybrid model. Environmental Earth Sciences 83(6):158. https://doi.org/10.1007/s12665-024-11453-0
49. Parmar KS, Bhardwaj R (2013) Wavelet and statistical analysis of river water quality parameters. Appl Math Comput 219(20):10172–10182. https://doi.org/10.1016/j.amc.2013.03.109
50. Saki Malehi E, Ashrafi SM, Ghafouri HR (2022) Performance evaluation of using Shannon's entropy crossing time to monitor drought: a case study of the Karkheh. Hydrol Sci J 67(13):1971–1987. https://doi.org/10.1080/02626667.2022.2116981
51. Shakeri R, Amini H, Fakheri F, Ketabchi H (2023) Assessment of drought conditions and prediction by machine learning algorithms using standardized precipitation index and standardized water-level index (case study: Yazd province, Iran). Environ Sci Pollut Res 30(45):101744–101760. https://doi.org/10.1007/s11356-023-29522-5
52. Shakeri R, Amini H, Fakheri F, Lam MY, Zahraie B (2025) Comparative analysis of correlation and causality inference in water quality problems with emphasis on TDS Karkheh River in Iran. Sci Rep 15(1):2798. https://doi.org/10.1038/s41598-025-85908-0
53. Torrence C, Compo GP (1998) A practical guide to wavelet analysis. Bull Am Meteor Soc 79(1):61–78. https://doi.org/10.1175/1520-0477(1998)079%3c0061:APGTWA%3e2.0.CO;2
54. Wei KL, Chen M, Wen ZY, Xie YK (2014) Research on signal processing for water quality monitoring based on continuous spectral analysis. Spectroscopy and spectral analysis 34(12):3368–3373. https://doi.org/10.3964/j.issn.1000-0593(2014)12-3368-06
55. Wen N, Dai K, Deng J, Liu C, Liang R, Yu B, Feng W (2024) Periodic displacement accurate extraction of reservoir active slopes through InSAR observation and independent component analysis-based wavelet transform. Int J Appl Earth Obs Geoinf 130:103919. https://doi.org/10.1016/j.jag.2024.103919
56. Wu Y, Yan W, Liu Z, Zhang Q, Zhou L, Song J, Yuchi M (2024) Full waveform inversion using frequency shift envelope-based global correlation norm for ultrasound computed tomography. Phys Med Biol 69(10):105024. https://doi.org/10.1088/1361-6560/ad3c8f
57. Xu R, Hu S, Wan H, Xie Y, Cai Y, Wen J (2024) A unified deep learning framework for water quality prediction based on time-frequency feature extraction and data feature enhancement. J Environ Manage 351:119894. https://doi.org/10.1016/j.jenvman.2023.119894
58. Zhang X, Chen Q, Recknagel F, Li R (2014) Wavelet analysis of time-lags in the response of cyanobacteria growth to water quality conditions in Lake Taihu, China. Eco Inform 22:52–57. https://doi.org/10.1016/j.ecoinf.2014.05.001
59. Zhi W, Appling AP, Golden HE, Podgorski J, Li L (2024) Deep learning for water quality. Nature water 2(3):228–241. https://doi.org/10.1038/s44221-024-00202-z
60. Zhang B, Song X, Zhang Y, Han D, Tang C, Yu Y, Ma Y (2012) Hydrochemical characteristics and water quality assessment of surface water and groundwater in Songnen plain. Northeast China Water research 46(8):2737–2748. https://doi.org/10.1016/j.watres.2012.02.033