



# Comparative Assessment of Machine Learning Models for Groundwater Quality Prediction Using Various Parameters

Majid Niazkar<sup>1,2</sup> · Reza Piraei<sup>3</sup> · Mohammad Reza Goodarzi<sup>4,5</sup> · Mohammad Javad Abedi<sup>6</sup>

Received: 26 June 2024 / Accepted: 29 January 2025  
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2025

## Abstract

This study aims to assess performances of eleven Machine Learning (ML) methods in predicting the Groundwater Quality Index (GWQI) for Yazd, an arid province in Iran. The ML models encompass Multiple Linear Regression (MLR), Support Vector Regression (SVR), K-Nearest Neighbors, Decision Tree Regression, Adaptive Boosting or AdaBoost, Random Forest Regression, Gradient Boosting Regression (GBR), XGBoost Regression (XGBR), Gaussian Process (GP), Artificial Neural Network (ANN), and Multi-Gene Genetic Programming (MGGP). After optimizing ML hyperparameters, ML-based estimation models were developed for three scenarios depending on which water quality parameters were used as input data: (1)  $K^+$  and pH; (2)  $K^+$ , pH,  $Na^+$ ,  $Ca^{2+}$ ,  $SO_4^{2-}$ ,  $HCO_3^-$  and  $Mg^{2+}$ ; and (3)  $K^+$ , pH,  $Na^+$ ,  $Ca^{2+}$ ,  $SO_4^{2-}$ ,  $HCO_3^-$ ,  $Mg^{2+}$ ,  $Cl^-$ , EC, TH, and TDS. For each scenario, ML-based models were assessed further by conducting (i) reliability analysis, (ii) ranking analysis, and (iii) confidence limits check. The results of the first scenario (with two input data) demonstrated the superiority of ANN, MGGP and GP, whereas ANN, MGGP and GBR were the most robust for the second scenario (with seven input data). Furthermore, the ranking analysis indicated that MLR, GP and ANN achieved the first highest ranks when eleven water quality parameters (third scenario) were used. The reliability analysis revealed that GP, MGGP, MLR, ANN, GBR, and XGBR achieved the highest reliability percentages across different scenarios, with ANN consistently ranking among the top models. Finally, the comprehensive comparative analysis of ML performances in this study reveals their potential for predicting GWQI.

## Highlights

- Groundwater quality was assessed using a dataset collected from wells of an arid region
- Eleven machine learning models were evaluated for estimating groundwater quality index
- Three scenarios were compared based on WHO permissible limits
- Reliability and ranking analyses were conducted for estimations of each ML model

**Keywords** Groundwater quality index · Water quality parameters · Machine learning · Multi-gene genetic programming · XGBoost

## 1 Introduction

Water is essential for human life and activities. Having access to safe, sufficient, and good-quality water is among the most prominent conditions for achieving sustainable development (Srebotnjak et al. 2012). Access to surface and groundwater resources and suitable quality water are decreasing due to various factors, such as population growth, agricultural expansion, industrialization, and urbanization worldwide (Tyagi et al. 2012). For instance, surface water and groundwater resources are highly vulnerable to various pollutants due to their rapid growth and development (UNEP 2016).

Physicochemical parameters in drinking water hold a significant importance, where their concentrations can directly or indirectly impact human health (Jonnalagadda and Mhere 2001). Natural factors including evaporation, watershed geography, and regional geology, as well as human factors, play roles in controlling chemical, physical, and biological compositions of water resources (Mishra et al. 2017). Consequently, assessing the quality status of water resources is necessary to implement suitable strategies for preventing degradation or enhancing its quality (Fernández et al. 2004). In this regard, analyzing various water physicochemical parameters, or water quality parameters, is difficult and inevitably cumbersome. In this regard, quality indices, e.g., Water Quality Index (WQI) or Groundwater Quality Index (GWQI), are methods that convert the qualitative attributes of water into a numerical value, allowing them to be utilized for water quality management, analysis, and monitoring over time and space (Carbajal-Hernández et al. 2013).

WQI for surface water or GWQI for groundwater is a technique proposed by Horton (1965) and has been used for evaluating water quality. It typically entails general water parameters, such as dissolved oxygen, acidity, hardness, dissolved solids, temperature, turbidity, nitrates, nitrites, and some major ions (Prusty and Farooq 2020; Tang et al. 2022). Some studies employ statistical techniques, utilizing weighted scores for each analyzed parameter to assess WQI (Jamshidzadeh 2020). It categorizes water quality status into easy-to-understand ranges on a scale from less than 50 to greater than 300, where higher values indicate lower water quality and vice versa (Prasad et al. 2019). Therefore, it aids in interpreting water quality through numerical values.

Although GWQI-based analysis approach is useful and applicable, it requires field survey to collect many water quality parameters from wells following laboratory analyses, which may be significantly costly and not practical in some cases, particularly in underdeveloped countries. Therefore, the need for alternative methods, e.g., artificial intelligence techniques, has emerged. According to the literature, various studies have already employed Machine Learning (ML) models to predict individual water quality parameters (Kheradpisheh et al. 2015; Ransom et al. 2017, 2022; Bedi et al. 2020; Stackelberg et al. 2021). Despite their proven utility, such application is limited to prediction of specific parameters, which may not delineate the overall status of water or groundwater quality. Therefore, a few studies have exploited ML models to estimate GWQI according to the literature review. To be more specific, some of them targeted at predicting Irrigation WQI (IWQI) (Trabelsi and Bel Hadj Ali 2022; Ibrahim et al. 2023; Hussein et al. 2024), whereas the present study emphasizes ML applications for estimating GWQI in the context of drinking water.

Most studies with the aim of drinking water have evaluated performances of Artificial Neural Network (ANN) and Support Vector Regression (SVR) (Sakizadeh 2016; Kulisz et al. 2021; Mohammed et al. 2023), while others used other ML models including Random Forest Regression (RFR) (Norouzi and Moghaddam 2020), Additive Regression

(AR), M5P tree model (M5P), Random Subspace (RSS) (Elbeltagi et al. 2022), Multiple Linear Regression (MLR), Locally Weighted Linear Regression (LWLR) (Kouadri et al. 2021), Deep Neural Network (DNN), Gradient Boosting Regression (GBR), eXtreme Gradient Boosting (XGBoost) Regression (XGBR) (Raheja et al. 2022), Naive Bayes, K-Nearest Neighbors (KNN) (Khiavi et al. 2023), Ensemble of Trees (ENT), Gaussian Process (GP), Regression Tree (RT) (Sajib et al. 2023), Ridge Regression (El-Rawy et al. 2024), Decision Tree Regression (DTR), Adaptive Neuro-Fuzzy Inference System (ANFIS) (Jibrin et al. 2024). These studies have proved the usefulness of ML models in estimating GWQI. Additionally, a few studies treated the estimation of GWQI as a classification problem (Agrawal et al. 2021; El-Magd et al. 2023; Sahour et al. 2023). In addition to GWQI, a few studies applied ML models to predict alternative indices, like Entropy-based WQI (EWQI) (Singha et al. 2021; Aju et al. 2024; Yang et al. 2024). Although ML models proved to be efficient in previous studies, they still require a sufficient amount of data for effective implementation. Due to practical challenges of measuring all groundwater quality parameters, it is crucial to estimate GWQI when not all water quality parameters are available. Hence, GWQI estimations rely not only on the measured data but also how the methodology operates.

In the domain of data-driven models, ML techniques have been applied to address estimation tasks for solving water resources problems, while estimation of water quality is among such problems. The critical issue is how ML-based models can predict GWQI in the absence of measurements of a few water quality parameter. Thus, further investigation is practically necessary as all water quality parameters may not be obtained by in situ measurements in some cases. In this regard, different scenarios of water quality parameters can be considered. For each scenario, ML models can be applied to the available parameters and consequently, performances of ML-based estimation models can be assessed for predicting GWQI. While several studies have employed feature importance or sensitivity analysis on their proposed GWQI estimation models (Kouadri et al. 2021; Raheja et al. 2022; El-Magd et al. 2023; Khiavi et al. 2023; Mohammed et al. 2023; Sahour et al. 2023; Sajib et al. 2023; El-Rawy et al. 2024), few have incorporated approaches, like Pearson correlation, to split their input parameters to conduct multiple scenarios with different input set of water quality parameters (Kulisz et al. 2021; Elbeltagi et al. 2022; Jibrin et al. 2024). However, variations of their results highlight a gap in the literature regarding the most effective splitting techniques.

Haggerty et al. (2023) conducted a review on applications of ML models to groundwater quality modeling. They suggested the need for further exploration of comparing performances of different ML models. Likewise, Torres-Martínez et al. (2024) emphasized that comparative evaluations of ML algorithms and the selection of appropriate evaluation metrics are critical for assessing model reliability, with 33% of studies focusing on a single algorithm, 18% comparing two algorithms, 41% comparing three to five methods, and only 8% comparing more than five. In essence, the availability of numerous ML algorithms poses a challenge when selecting an appropriate ML model for a given study. In addition, performances of different ML algorithms can vary significantly based on the data in question. Selecting a random ML model may compromise both the efficiency and the accuracy. Therefore, it is critical not only to validate the data but also to select the ML algorithms suited to the specific task (Rajeev et al. 2024). As a result, there is a clear need for comprehensive studies that assess the capabilities of ML models for GWQI estimations. Such line of investigations can contribute to more efficient and accurate water quality analysis and assessment.

This study gathered extensive field water quality parameters from Yazd Province, Iran, to develop various predictive models for forecasting GWQI. Three distinct scenarios were considered, each using different combinations of water quality parameters as input. Unlike previous studies that often utilized Pearson correlation for selecting input combinations, this study categorized parameters based on different ranges of standard values (weights), leading to a more consistent division method. This scenario-based ML modeling demonstrates how accurate GWQI can be estimated when a specific set of water quality parameters data is missing. For this purpose, eleven ML models, including MLR, SVR, KNN, DTR, Adaptive Boosting or AdaBoost (AB), RFR, GBR, XGBR, GP, ANN, and Multi-Gene Genetic Programming (MGGP), were evaluated for GWQI prediction. To the best of the authors' knowledge, comparing performances of these many ML models for predicting GWQI is novel, while this is the first time that MGGP has been employed for this application. The wide range of ML models offers a comprehensive evaluation of their potential. ML performances were assessed across the scenarios using six statistical metrics, and an overall ranking method was applied to identify the best-performing model across all scenarios. Furthermore, reliability and confidence limits analyses were conducted.

The remainder of this paper is structured as follows: Sect. 2 presents the Methods and Materials, which includes details on the case study, data for modeling, the groundwater quality index, model scenarios, and descriptions of the ML models. It also outlines the performance criteria, ranking approach, and reliability analysis. Section 3 provides the Results and Discussion, focusing on the optimization of ML hyperparameters and the results of the models for three different scenarios. It also includes a dedicated subsection for post-processing, where the ML models are ranked, and their reliability is analyzed. The final part of Sect. 3 is a discussion, which interprets the results and compares them with the findings obtained from other studies. Finally, Sect. 4 concludes the paper with a summary of key findings and implications for future research.

## 2 Materials and Methods

### 2.1 Case Study

The province of Yazd, located in the central plateau of Iran, covers an arid area of approximately 74,493 km<sup>2</sup>, accounting for 5.4% of Iran's total land area. Figure 1 illustrates the designated area and the locations of the sampling stations. As shown, the study area geographically lies between 29° 30' and 33° 20' North latitude and 52° 45' and 56° 40' East longitude. According to the Domartan climate classification, the province is divided into three climatic regions: arid, semi-arid, and Mediterranean. Except for the mountainous region of Shirkooh, which has a Mediterranean climate, the province becomes progressively drier when moving from the southwestern and western parts towards the northeastern and eastern areas.

The precipitation pattern of the province follows a Mediterranean type, with the maximum occurring during the winter season. The average annual precipitation ranges from 60 to 80 mm, while the average annual evaporation varies between 2,500 mm and 4,200 mm, which is significantly higher than the precipitation amount (IRIMO 2015).

The water balance in the Yazd province is influenced by its arid and semi-arid climate, with an annual rainfall between 100 and 200 mm, primarily during spring and autumn. High summer temperatures lead to elevated evaporation rates, intensifying water stress

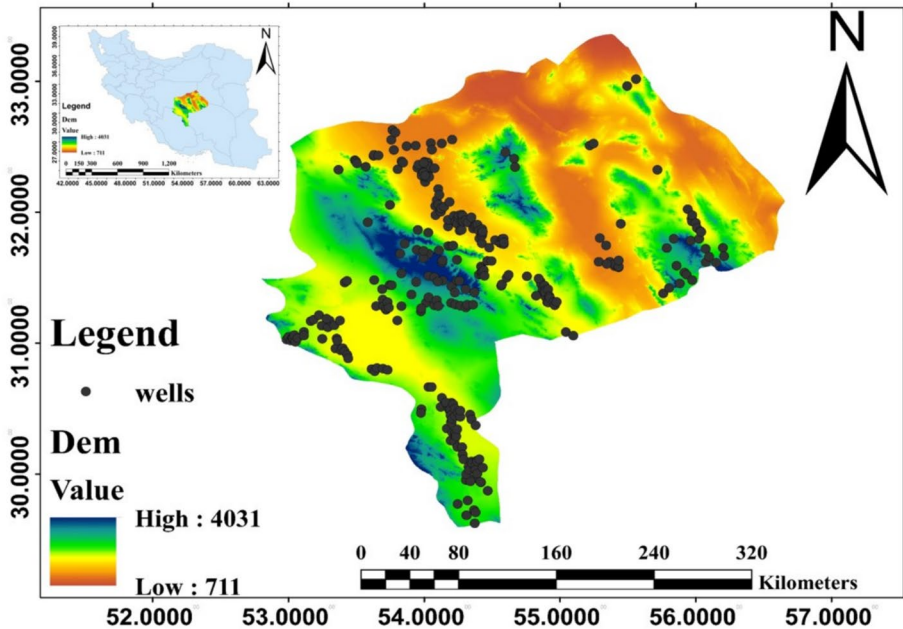


Fig. 1 Study area and locations of the sampling stations

(Eslamian et al. 2018). Groundwater, as the main water resource in the province, is extracted by wells across regions. Over-extraction of water has caused declining water levels and increased salinity (Salehi et al. 2014). Geologically, Yazd features layers of clastic rocks, clay, and sand, which affect groundwater permeability. Furthermore, the Yazd province contains several important aquifers, including karst, alluvial, and travertine aquifers. However, excessive groundwater use, combined with agricultural and industrial activities, has degraded water quality, with rising nitrate and salinity levels being major concerns (Farhadinejad et al. 2014).

## 2.2 Data for Modeling

The data used for modeling in this study includes eleven water quality parameters. They are Potassium ( $K^+$ ), Sodium ( $Na^+$ ), pH, Calcium ( $Ca^{2+}$ ), Sulfate ( $SO_4^{2-}$ ), Chloride ( $Cl^-$ ), Bicarbonate ( $HCO_3^-$ ), Electrical Conductivity (EC), Total Hardness (TH), Total Dissolved Solids (TDS), and Magnesium ( $Mg^{2+}$ ). All notations used in this study are presented in the Supplementary Material. The data was divided into two parts: (i) train data (240 measurements) and (ii) test data (80 measurements). Table 1 presents the minimum (min), maximum (max), average (mean), and standard deviation (Std.) of the water quality parameters (see Sect. 2.3) values for both train and test data. As shown, the minimum of each parameter of the train data is lower than the minimum of the corresponding parameter in the test data. Furthermore, the maximum of each parameter of the train data is larger than the maximum of the corresponding parameter of the test data.

A small part of the data (96 measurements), which belongs to a subregion of the Yazd province, was previously considered in another study (Goodarzi et al. 2023), while the rest

**Table 1** Statistical characteristics of the data used in this study

Data	Parameter (mg/L)	K <sup>+</sup>	pH	Na <sup>+</sup>	Ca <sup>2+</sup>	SO <sub>4</sub> <sup>2-</sup>	HCO <sub>3</sub>	Mg <sup>2+</sup>	Cl <sup>-</sup>	EC	TH	TDS
Train data	Min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Mean	0.096	0.560	0.114	0.083	0.168	0.113	0.104	0.079	0.104	0.092	0.104
	Max	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Std	0.093	0.148	0.138	0.111	0.158	0.109	0.126	0.112	0.122	0.114	0.118
Test data	Min	0.027	0.265	0.006	0.006	0.009	0.012	0.005	0.002	0.005	0.005	0.008
	Mean	0.107	0.553	0.182	0.108	0.237	0.108	0.142	0.125	0.156	0.123	0.159
	Max	0.253	0.810	0.590	0.387	0.690	0.321	0.453	0.413	0.453	0.383	0.457
	Std	0.066	0.113	0.171	0.096	0.185	0.069	0.125	0.123	0.137	0.105	0.138

(224 measurements) is new to the literature. In other words, this study worked on the data gathered from the whole Yazd province. Moreover, the previous study considered all water quality parameters to estimate GWQI (Goodarzi et al. 2023), whereas this study aimed to evaluate three scenarios with different set of water quality parameters as input. Finally, the data was further divided into three sets of scenarios, which will be explained later in Sect. 2.5.

## 2.3 Ground Water Quality Index

GWQI is a practical method employed for assessing groundwater quality. It is an important tool for informing policymakers about the condition of potable water as it synthesizes multiple water quality parameters into a single numerical value by identifying, weighting, and integrating them (Abba et al. 2024). GWQI assigns standard values to each water quality parameter, while the corresponding weights vary based on their relevance to water quality assessment and potential health implications. For example, TDS receives the highest weight due to its significant impact on water quality and associated health risks when exceeding permissible limits in drinking water. In contrast,  $K^+$  is among water quality parameters with the lowest weights, indicating its lesser significance and minimal health effects (Abidi et al. 2024). The process of computing GWQI for each data sample is described in the following steps.

### 2.3.1 Calculation of Water Quality Classification

The water quality is classified by Eq. (1) (Prasad et al. 2019):

$$q_n = 100 \left( \frac{V_n - V_i}{S_n - V_i} \right) \quad (1)$$

where  $q_n$  represents the water quality rating for the parameter  $n$ ,  $V_n$  is the observed value for parameter  $n$ ,  $S_n$  is the standard value for the parameter  $n$ , and  $V_i$  is an ideal value for the parameter  $n$ .

### 2.3.2 Calculation of Unit Weight

The unit weight of each water quality parameter ( $W_n$ ) corresponds inversely to the recommended standard value (Prasad et al. 2019), as shown in Eq. (2):

$$W_n = \frac{K}{S_n} \quad (2)$$

where  $K$  is the standard value for the parameter  $n$ , as given by Eq. (3):

$$K = \frac{1}{\sum \left( \frac{1}{S_n} \right)} \quad (3)$$

Finally, GWQI can be calculated by linearly incorporating the quality grade into the unit weight, as shown in Eq. (4):

$$\text{GWQI} = \frac{\sum q_n W_n}{\sum W_n} \quad (4)$$

Typically, the GWQI values are classified into five categories (Prasad et al. 2019): (i) Class A representing excellent quality ( $\text{GWQI} < 50$ ), (ii) Class B indicating good quality ( $51 \leq \text{GWQI} < 100$ ), (iii) Class C denoting poor water quality ( $101 \leq \text{GWQI} < 200$ ), (iv) Class D representing very poor water quality ( $201 \leq \text{GWQI} < 300$ ), and (v) Class E showing water unsuitable for drinking use ( $\text{GWQI} < 300$ ).

## 2.4 Model Scenarios

We developed three scenarios in which different sets of input water quality parameters were used as input variables to implement the ML models. These scenarios were defined based on the standard value  $S_n$  of each chemical parameter. To be more specific, permissible limits of  $S_n$  recommended by WHO were classified into three groups depending on each water quality parameter: (i) low ( $S_n < 100$ ), (ii) medium ( $100 < S_n < 400$ ), and (iii) high ( $400 < S_n$ ). Following this classification, the first scenario entails the two water quality parameters with low  $S_n$  (i.e., 8.5 for pH and 12 for  $\text{K}^+$ ) based on the World Health Organization (WHO) recommendation (WHO 2011). Additionally, the second scenario includes the water quality parameters with low to medium  $S_n$  (200 for  $\text{Na}^+$ , 200 for  $\text{Ca}^{2+}$ , 250 for  $\text{SO}_4^{2-}$ , 120 for  $\text{HCO}_3^-$ , and 150 for  $\text{Mg}^{2+}$ ) permissible limits, while the third scenario consists of all water quality parameters, including those with high  $S_n$  (600 for  $\text{Cl}^-$ , 1500 for EC, 500 for TH, and 1500 for TDS), as input. For better clarification, the input variables of the three scenarios are presented in Table 2, while the output for all scenarios is GWQI.

## 2.5 Machine Learning Models

This study employed 11 ML models, namely MLR, SVR, KNN, DTR, AB, RFR, GBR, XGBR, GP, ANN, and MGPP. The first 10 ML models were implemented in Python, while GPTIPS tool in MATLAB was used to exploit MGPP. To implement the MLR, SVR, KNN, DTR, AB, RFR, GBR and GP models, the scikit-learn library was used, while the keras and xgboost libraries were used for ANN and XGBR, respectively. A concise overview of each ML method is presented in the following.

### 2.5.1 Multiple Linear Regression

MLR is one of the most common and easy-to-implement regression algorithms, employed to ascertain a linear correlation between dependent and independent variables using the

**Table 2** Input variables in different scenarios

Scenario	Input variables										
1	$\text{K}^+$	pH	-	-	-	-	-	-	-	-	-
2	$\text{K}^+$	pH	$\text{Na}^+$	$\text{Ca}^{2+}$	$\text{SO}_4^{2-}$	$\text{HCO}_3^-$	$\text{Mg}^{2+}$	-	-	-	-
3	$\text{K}^+$	pH	$\text{Na}^+$	$\text{Ca}^{2+}$	$\text{SO}_4^{2-}$	$\text{HCO}_3^-$	$\text{Mg}^{2+}$	$\text{Cl}^-$	EC	TH	TDS



least squares method (Nathan et al. 2017). It can be used to handle a linear relationship among various water quality parameters for GWQI estimations:

$$GWQI = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_i x_i \quad (5)$$

where  $x_i$  is the value of the  $i^{\text{th}}$  water quality parameter,  $\alpha_0$  is the regression constant, and  $\alpha_i$  is the coefficient of the  $i^{\text{th}}$  water quality parameter.

## 2.5.2 Support Vector Regression

SVR is a versatile supervised ML algorithm for both linear and nonlinear regressions. It handles non-linear data by transforming it into a higher-dimensional space. SVR is a variant of Support Vector Machine (SVM) designed for predicting continuous numerical values. In contrast to conventional ANN, SVM offers the advantage of enhancing data network performance, which can be used to improve the accuracy of data transmission, reduce the latency of data transmission, and improve the security of data transmission (Mohammed et al. 2023). As a supervised classifier, SVM efficiently predicts by finding an optimal hyperplane that maximizes the margin between the hyperplane and input variables. SVM constructs a hyperplane for class label prediction, whereas SVR generates various functions based on the training data to predict numerical values. SVR relies on the computational framework of linear regression, where the inputs are transformed into a high-dimensional feature space using a non-linear kernel function (Elzain et al. 2022). The SVR efficiency depends on the kernel choice, hyperparameters, and the regularization parameter (Piraei et al. 2023a). For more details on the specific mathematics behind SVR, interested readers are referred to Elzain et al. (2022).

## 2.5.3 K-Nearest Neighbors

KNN is a widely used supervised ML algorithm that can be utilized in nonlinear-regression tasks. KNN, as a non-parametric model, does not assume that variables or residuals follow a normal distribution (Elzain et al. 2023). In essence, it involves arranging training data to make predictions for test data. KNN identifies the nearest data points in the training dataset, known as neighbors, to predict outcomes for a given test point. Furthermore, it selects a specified number of closest neighbors to a query point. It then calculates the average of the target values associated with the selected data points. Moreover, the influence of neighbors on the average is determined by assigning weights. The KNN model employs a distance function to quantify the similarity between the query point and the data points in the training set. Notably, three widely-recognized distance functions for continuous variables are (i) Euclidean, (ii) Manhattan, and (iii) Minkowski (Piraei et al. 2023a). Finally, the best value for K, i.e., the number of closest neighbors, is determined by cross validation process. For more details on the specific mathematics behind KNN, readers are encouraged to refer to Elzain et al. (2023).

## 2.5.4 Tree-Based Models

Tree-based models are a class of ML methods that leverage hierarchical structures to make predictions. They utilize decision trees as their fundamental building blocks, enabling them to capture complex relationships within a dataset. This study utilized five distinct tree-based models, namely DTR, AB, RFR, GBR, XGBR, and MGGP, where each one exploits

decision trees in a distinct way not only to enhance the prediction accuracy but also tackle overfitting. They are briefly presented in the following.

DTR is a tree-structured model that employs regression trees to perform nonlinear regression tasks. DTR demonstrated significant potential for predicting the sensitivity of groundwater contamination, even when working with limited data or complex nonlinear relationship within the dataset (Jibrin et al. 2024). It relies on a hierarchical structure whose internal nodes represent tests on observed values, branches denote test outcomes, and leaf nodes offer output predictions. DTR starts at the root node and recursively splits data subsets based on observation values until it reaches a leaf node. Then, predictions are derived from the average of instances within each leaf (Piraei et al. 2023a). For more details on the specific mathematics behind DTR, readers are encouraged to refer to Jibrin et al. (2024).

RFR is a popular robust, flexible, and easy-to-use ensemble algorithm for nonlinear regression that offers a strong balance between accuracy and stability (Elzain et al. 2024). It employs multiple decision trees to make predictions. Each tree is constructed using bootstrapped data, which involves resampling the training dataset to create distinct samples. The samples are then used to build individual decision trees, ensuring diverse patterns are captured. The final prediction of RF is an average of outcomes obtained by all trees, contributing to reduce overfitting as compared to standalone decision trees. For more details on the specific mathematics behind RFR, readers are encouraged to refer to Di Nunno et al. (2023).

AB is an ensemble technique that combines weak learners to create a strong learner. Unlike RFR, AB constructs trees with a single node and two leaves, often referred to as stumps. The stumps basically serve as weak learners. To be more precise, AB iteratively trains a series of weak learners, assigning a higher emphasis on misclassified data points from preceding learners. Furthermore, the sequence of constructing weak learners matters because errors from previous iterations impact subsequent ones. The final prediction of AB is a weighted summation of the weak learners predictions (Piraei et al. 2023b).

GBR is another ensemble technique that employs gradient descent to optimize the loss function, resulting in accurate and flexible estimations. Generally, it starts with a single leaf as an initial estimation, which is typically the average of continuous data. Subsequent trees are trained to correct errors from preceding ones. The construction of trees continues until a predefined threshold is reached. Finally, the GBR prediction is a weighted summation of tree predictions (Piraei et al. 2023b).

XGBR, as an advancement of GBR, incorporates features like regularization and tree pruning to mitigate overfitting. It is renowned for its efficiency in handling large datasets, parallel processing, and continuous algorithm enhancements (Piraei et al. 2023b). Based on the gradient boosting framework, XGBR incrementally builds an ensemble of trees. Each new tree focuses on addressing the errors made by previous trees by minimizing a unique objective function. It comprises some components designed to reduce model complexity and refine residuals through an iterative process. Similar to GBR, the final prediction made by XGBR is a weighted summation of tree predictions. Additional significant innovations of XGBR involve the use of approximate greedy algorithms for efficient tree construction, the ability to customize objective functions for diverse learning tasks, and the estimation of feature importance, which enhances model interpretability (El-Rawy et al. 2024). For more details on the boosting methods (i.e., AB, GBR, and XGBR), readers are encouraged to refer to Wade and Glynn (2020).

Finally, MGGP is a modified version of genetic programming and has been used to solve some problems in water resources (Niazkar 2023). The latter is an ML method with a tree-based structure, whose nodes are classified into three groups: (1) root, (2) functions, and (3)

terminals. In MGGP, each individual can be one or more than one tree (or gene), whereas in the traditional genetic programming, each tree can only be one tree. Both variants exploit the genetic algorithm as their search engine to optimize the fitness function, which is minimizing the difference between estimated and measured values. The flexible architecture of MGGP enables it to capture nonlinear relationships between independent and dependent variables (Niazkar et al. 2023). For more information regarding the MGGP algorithm readers are encouraged to refer to Niazkar et al. (2023).

### 2.5.5 Gaussian Process

GP is a stochastic model that involves utilizing random variables, assuming the underlying data follows a Gaussian distribution. Unlike parametric models that rely on predetermined structures, GP is a non-parametric method and constructs its model framework directly from the observed data (Suphawan and Chaisee 2021). This flexibility allows GP to adapt to various data patterns. Furthermore, GP is characterized by its mean and covariance functions, which depend on the input vector. In GP, the common choice for the covariance function is the Radial Basis Function (RBF) kernel, which maps data to a higher-dimensional space. By evaluating the joint distribution of training and testing data, GP can make predictions for unseen input data (Piraei et al. 2023b). For further details on GP and its applications, refer to the relevant literature (Roushangar et al. 2023).

### 2.5.6 Artificial Neural Network

ANN mimics the biological nervous system by processing data in parallel, like the human brain. Among various versions of ANN, multilayer perceptron neural networks are common for environmental problems (Mohammed et al. 2023). It is a type of feed-forward neural network consisted of an input layer (containing input variables), one or more hidden layer(s), and an output layer (output variable that is GWQI in this study). Initial operation involves linking inputs to hidden layers and refining weights using regularization. Specifically, the hidden layers process the data in the input layer by applying a weighted linear summation, followed by a non-linear activation function. Connections between neurons in adjacent layers are linked to trainable parameters, called weights, with each neuron typically having a bias term that aids in transformation. This process continues iteratively through the neurons from the input layer to the output layer, adjusting the weights in conjunction with biases until an optimal result is achieved (Granata et al. 2024).

## 2.6 Performance Criteria and Ranking Approach

To evaluate the performances of the ML techniques in predicting GWQI under different scenarios, six metrics were employed. These metrics are: (i) Root Mean Square Error (RMSE); (ii) Mean Absolute Error (MAE); (iii) Nash–Sutcliffe Efficiency (NSE); (iv) Coefficient of Determination ( $R^2$ ); (v) Maximum Absolute Relative Error (MXARE); and (vi) Mean Absolute Relative Error (MARE). The following equations are the mathematical formulas of these criteria (Goodarzi et al. 2023):

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (O_i - P_i)^2}{n}} \quad (6)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |O_i - P_i| \tag{7}$$

$$NSE = 1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n \left(O_i - \frac{\sum_{i=1}^n O_i}{n}\right)^2} \tag{8}$$

$$R^2 = \left\{ \frac{\sum_{i=1}^n \left[ \left(O_i - \frac{\sum_{i=1}^n O_i}{n}\right) \left(P_i - \frac{\sum_{i=1}^n P_i}{n}\right) \right]}{\sqrt{\sum_{i=1}^n \left(O_i - \frac{\sum_{i=1}^n O_i}{n}\right)^2 \sum_{i=1}^n \left(P_i - \frac{\sum_{i=1}^n P_i}{n}\right)^2}} \right\}^2 \tag{9}$$

$$MXARE = \max \left( \left| \frac{O_i - P_i}{O_i} \right| \right) \text{ for } i = 1, \dots, n \tag{10}$$

$$MARE = \frac{1}{n} \sum_{i=1}^n \left| \frac{O_i - P_i}{O_i} \right| \tag{11}$$

where  $n$  represents the number of data points,  $O_i$  is the  $i^{\text{th}}$  observed GWQI, and  $P_i$  indicates the  $i^{\text{th}}$  predicted GWQI.

The definitions associated with each metric imply that achieving more accurate estimates of GWQIs corresponds to reduced RMSE, MAE, MARE, and MXARE values, and elevated  $R^2$  and NSE values.

After conducting a comprehensive analysis of metrics across all methods, a ranking strategy was employed to facilitate a more robust comparison of the ML methods based on their performances across all metrics. In this approach, each of the 12 metric results (comprising 6 metrics for each of the two datasets) was utilized to delineate performances of the ML models, which were subsequently arranged from the highest performance to the lowest one. Thus, 11 ML models were assigned rankings ranging from 1 (indicating superior performance) to 11 for each of the 12 metric results. Subsequently, the summation of the ranks for all metrics within each of the training and testing datasets was computed independently for each method and organized in an ascending order. Given that a lower total rank denotes a better performance, the models were then re-ranked based on the summations, with the lowest total rank receiving the first position and the highest one receiving the final position. The outcome in a condensed representation of the rankings, presenting them as two columns: one for the ranks pertaining to the training data and the other one for the ranks associated with the testing data. These ranks were once again subjected to the summation for both training and testing datasets. Then, the total summation was sorted in an ascending order and re-ranking. Finally, the ultimate ranking for each ML model was determined.

### 2.7 Reliability Analysis

Reliability analysis evaluates the consistency of a predictive model in relation to a pre-defined threshold. This assessment involved computing the relative error ( $RE$ ), as shown

in Eq. (12), for each ML model. It was then constrained with a predefined threshold set at 20% based on the literature (Piraei et al. 2023b). The count of cases where the relative error meets or falls lower than the designated threshold was divided by the total number of cases. The resulting percentage serves as an index of reliability, which assesses the overall coherence of an estimation model.

$$RE = \frac{|P_i - O_i|}{O_i} \quad (12)$$

### 3 Results

#### 3.1 Optimizing ML Hyperparameters

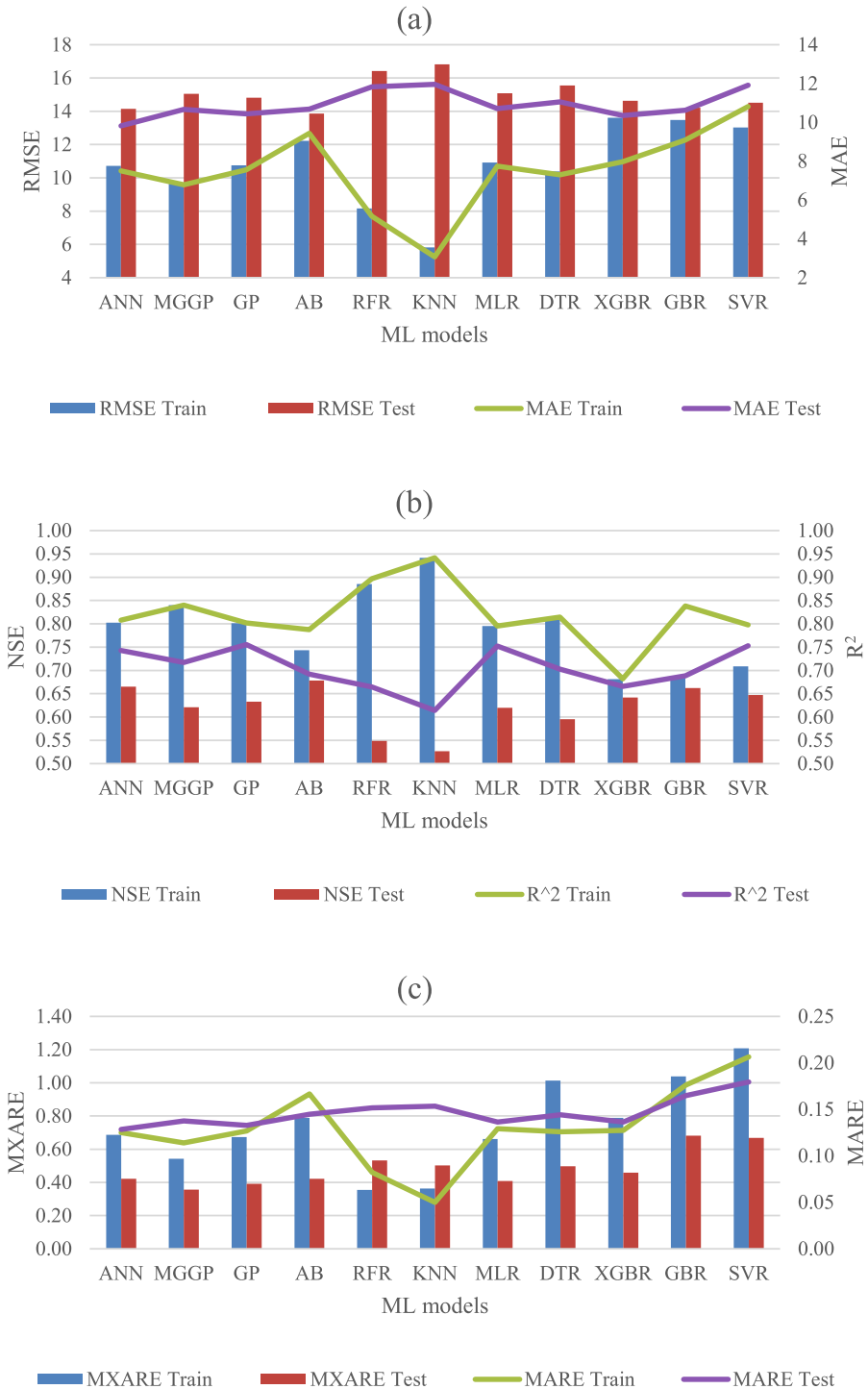
To enhance the performance of ML models, it is imperative to tune their hyperparameters in alignment with the characteristics of the training data. The optimization process is crucial for ensuring that the ML models exhibit an optimal performance when applied to the testing data. The process of optimizing hyperparameters was conducted for each scenario using the grid search method, and the results are summarized in Supplementary Material file for interested readers. The performance results of ML models for each scenario considering optimized hyperparameters are presented separately in following sections.

#### 3.2 ML Model Performances for the First Scenario

Statistical metrics serve as a valuable means for enhancing the comparative evaluation of the accuracy of ML models. The findings of the metrics are summarized in Fig. 2. The x-axis represents the ML models, while the y-axis illustrates the results of metrics. In Fig. 2, the MAE and RMSE results are depicted as bar charts and lines for each dataset. All ML models exhibited satisfactory and comparable performance levels. Notably, with respect to RMSE, the ANN model demonstrated the most impressive performance on the testing data, yielding an RMSE value of 14.15, while the KNN model showed the least favorable performance (RMSE=16.82). According to Fig. 2, RMSE achieved by ML are quite close, with a mere 2.67-unit discrepancy between the best and worst RMSE performances.

Based on Fig. 2, the MAE results followed a similar pattern, with the ANN model achieving the lowest MAE (=9.82), and the KNN model obtaining the highest MAE (=11.96). According to NSE, AB emerged as the top performer with a value of 0.68, whereas KNN displayed the weakest performance (NSE=0.53). Moreover, regarding  $R^2$ , the GP model outperformed others with an  $R^2$  value of 0.76, whereas KNN reached the weakest performance with an  $R^2$  of 0.61.

The MARE results presented in Fig. 2 are closely clustered, with both ANN and GP yielding the best MARE results (0.13), and SVR delivering slightly weaker outcomes (MARE=0.18). Regarding MXARE, MGGP achieved the best value of 0.36, while GBR resulted in the least favorable performance (MXARE=0.68). Based on the narrow margins between the metric results and the utilization of various criteria to evaluate each ML model, employing a ranking scheme is advisable to draw more conclusive findings.



**Fig. 2** Statistical metric results for the first scenario: (a) RMSE and MAE, (b) NSE and R<sup>2</sup>, (c) MXARE and MARE

### 3.3 ML Model Performances for the Second Scenario

Figure 3 presents the statistical findings for the second scenario. Compared to the first scenario, all ML models exhibited significantly improved performance in the second scenario. Nonetheless, there were notable variations in model performances, with certain models outperforming others to a considerable extent. According to Fig. 3, with respect to RMSE, the MGGP model demonstrated the most impressive performance on the testing data, yielding an RMSE value of 0.22, while the DTR model achieved the least favorable performance (RMSE = 8.51). The MAE results followed a similar pattern as the ANN and MGGP models obtained the lowest MAE of 0.12 and 0.13, respectively, whereas the DTR model reached the highest MAE (= 6.35).

According to Fig. 3, ANN, MGGP, MLR, and GP achieved an exceptional NSE of 1, whereas DTR displayed the weakest performance with NSE equal to 0.88. Moreover, the results of  $R^2$  were also the same. To be more specific, the ANN, MGGP, MLR, and GP models outperformed others with an  $R^2$  value of 1, while DTR obtained the weakest performance with an  $R^2$  of 0.88. Furthermore, the MARE results were closely clustered, with the ANN, MGGP, MLR, and GP models yielding the best results (almost 0), and AB delivering slightly weaker outcomes (0.08). Regarding MXARE, the MGGP, MLR and GP models achieved the best score at 0.01, whereas SVR exhibited the least favorable performance (MXARE = 0.39).

### 3.4 ML Model Performances for the Third Scenario

Figure 4 presents the metric results for the third scenario. Overall, the ML models exhibited an improved performance in both second and third scenarios, as opposed to the first scenario. Nevertheless, there were slight disparities in performances of ML models when comparing the results of the second scenario with the third one, where most of ML models displayed a slight improvement in their performances, while others showed a slight weaker performance. Notably, MLR showed an exceptional performance with RMSE, MAE, MARE and MXARE values close to 0, and  $R^2$  and NSE values close to 1. In addition, GP showed the best results regarding the MARE, MXARE,  $R^2$  and NSE. However, in terms of RMSE and MAE, it showed a slight error of 0.03 and 0.02, respectively. While other ML models indicated a commendable and satisfactory performance, the SVR, AB and DTR models exhibited much weaker performances compared to others. Consequently, the SVR model demonstrated the poorest performance according to most of the performance criteria (RMSE = 7.07, MAE = 5.78, NSE = 0.92, MARE = 0.09, and MXARE = 0.34). DTR also showed the weakest performance regarding  $R^2$  with a value of 0.92. Like previous scenarios, employing a ranking scheme is advisable to draw more conclusive findings.

### 3.5 Postprocessing of ML Results

To enhance the assessment of diverse ML models across different scenarios, this study employed various postprocessing analyses. In this regard, Table 3 presents the outcomes of the ranking analysis for different scenarios. As shown, for the first scenario, while the ANN model initially secured the fourth position in terms of its performance on the training data, its exceptional performance on the testing data elevated it to the top

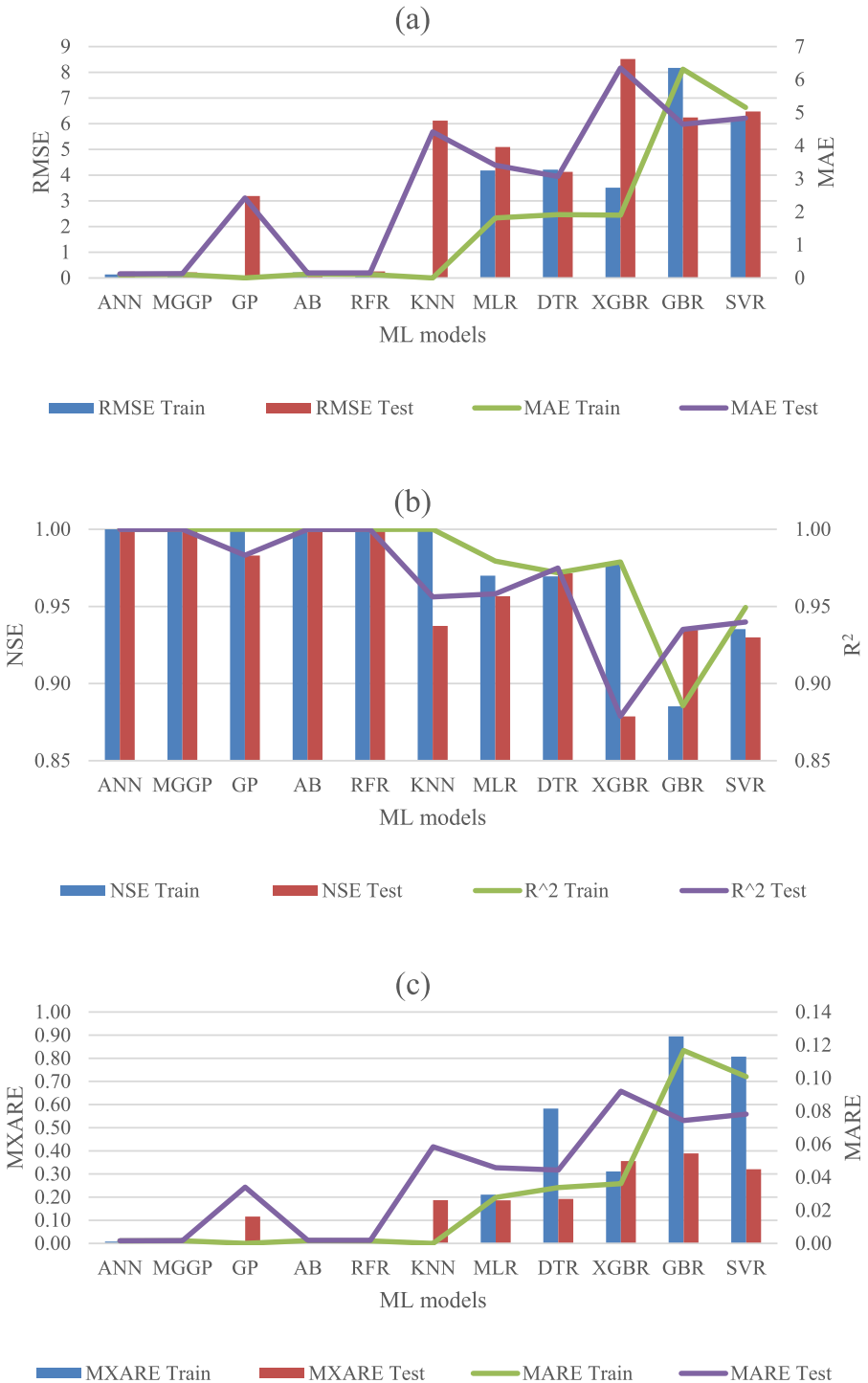


Fig. 3 Statistical metric results for the second scenario: (a) RMSE and MAE, (b) NSE and R<sup>2</sup>, (c) MXARE and MARE



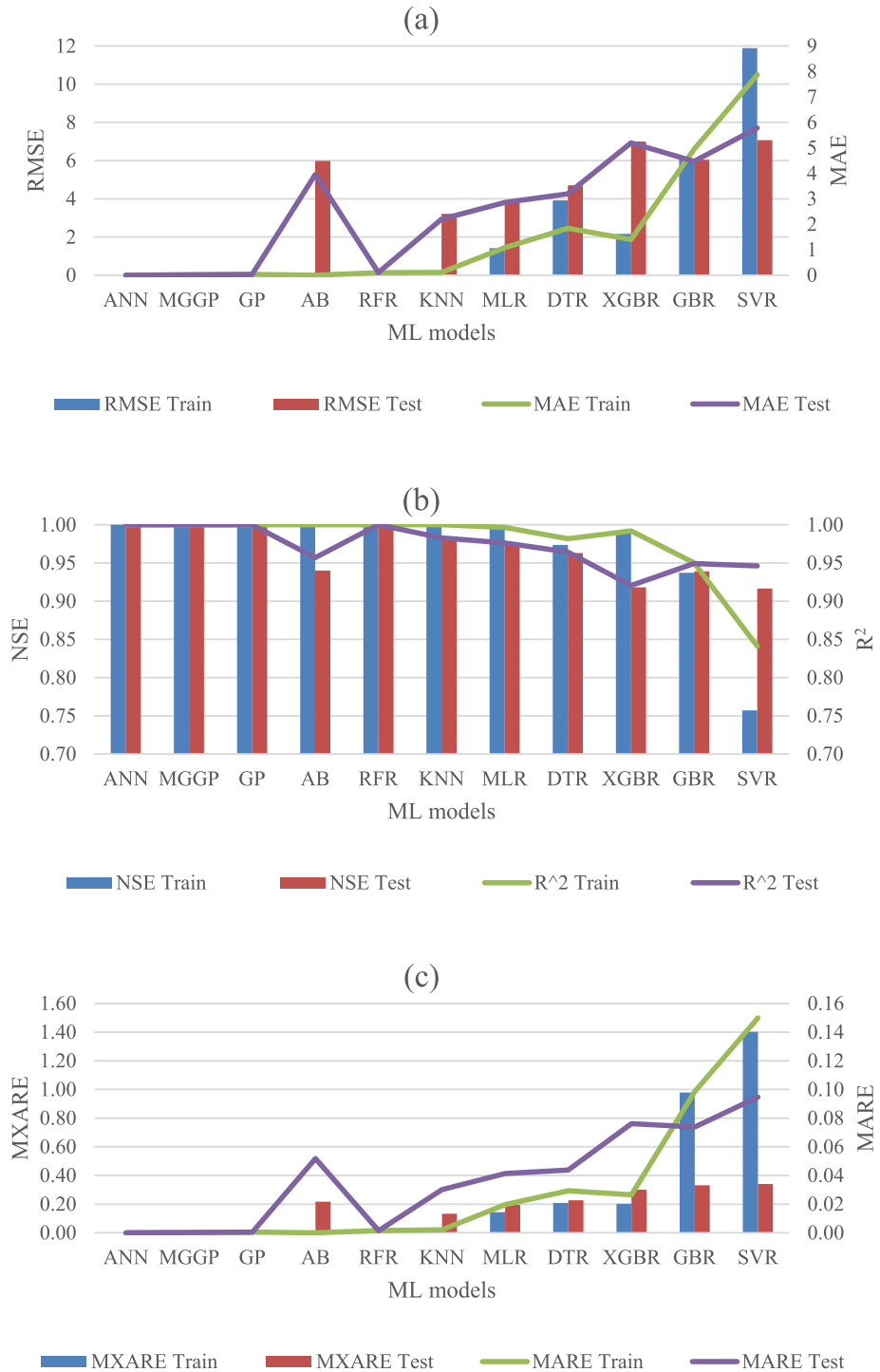


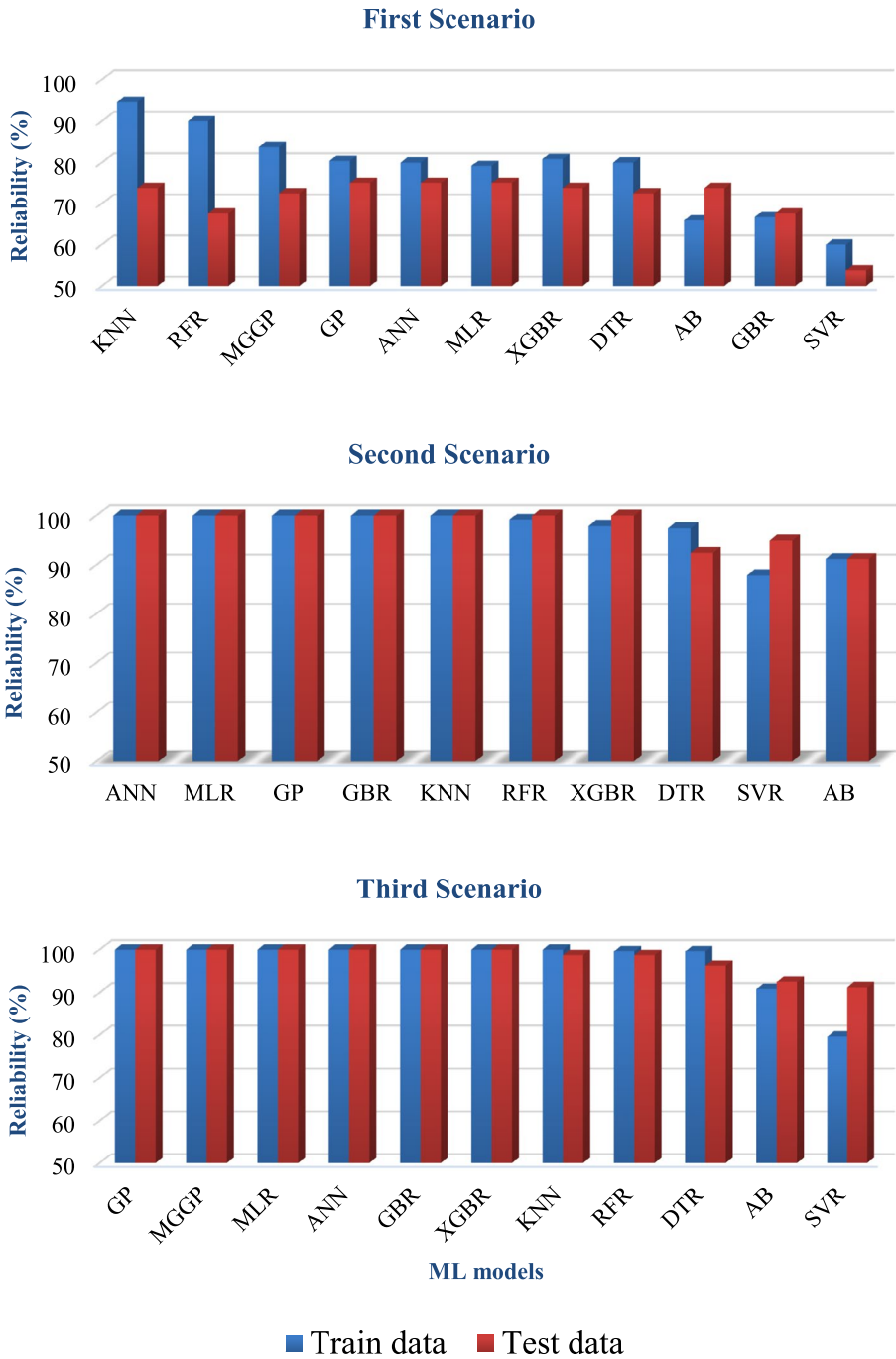
Fig. 4 Statistical metric results for the third scenario: (a) RMSE and MAE, (b) NSE and R<sup>2</sup>, (c) MXARE and MARE

**Table 3** Results of ranking analysis of different ML models for three scenarios considered in this study

ML models	First scenario			Second scenario			Third scenario		
	Train	Test	Total	Train	Test	Total	Train	Test	Total
ANN	4	1	1	3	2	1	4	3	3
MGGP	3	4	2	6	1	2	6	4	5
GP	6	2	3	4	4	4	3	2	2
MLR	7	6	7	5	3	4	2	1	1
KNN	1	11	5	1	8	6	1	8	4
GBR	8	7	10	2	5	2	5	5	5
RFR	2	10	5	7	7	7	9	7	8
XGBR	10	4	9	9	6	8	7	6	7
AB	8	3	4	10	10	10	10	9	10
DTR	4	9	7	8	11	9	8	10	9
SVR	11	8	11	11	9	10	11	11	11

ranking. Following ANN, the MGGP, GP and AB models obtained subsequent ranks. Furthermore, the KNN model generally demonstrated a strong fit with the training data, while its comparatively weaker performance for the testing data put it in the fifth rank, similarly to the RFR model. Likewise, the MLR and DTR models were jointly ranked seventh. Finally, the XGBR, GBR and SVR models exhibited the lowest performances. Regarding the second scenario, like the first scenario, ANN outperformed other ML models. While the MGGP model initially secured the sixth position for the training data, it moved up to the second ranking position because of its exceptional performance for the testing data. Following MGGP, the GBR, MLR, and GP models achieved subsequent ranks, respectively. Additionally, the KNN model demonstrated an adequate performance for the training data, while its considerably weaker performance for the testing data (the eighth rank for the testing dataset) led it to being ranked as the sixth model overall. Moreover, the RFR, XGBR, and DTR models were placed at the subsequent ranks, respectively. Lastly, the SVR and AB models exhibited the lowest performances and jointly placed as the last model for the second scenario. Regarding the third scenario, the MLR model outperformed other ML models, following by the GP and ANN models obtaining the second and third ranks, respectively. Like other scenarios, the KNN model showed a weaker performance for the testing dataset (the eighth rank), which positioned it as the fourth robust model. In addition, the MGGP and GBR models achieved satisfactory performances and jointly were placed fifth. Furthermore, the XGBR, RFR, DTR and AB models were placed at the subsequent ranks, respectively. Finally, the SVR model was placed last due to its lowest performance.

Figure 5 depicts the results of the reliability analysis, with the x-axis representing the reliability percentage and the y-axis indicating each ML model. As shown, for the first scenario, KNN achieved the highest reliability percentage for both training data (94.58%) and testing data (73.75%). Furthermore, the reliability percentages obtained by RFR (90% for training data and 67.5% for testing data) and MGGP (83.75% for training data and 72.5% for testing data) were the second and third highest values, respectively. Subsequently, the ANN, MLR, XGBR and DTR models exhibited acceptable reliability scores ranging from 70 to 80%. In contrast, both AB and GBR yielded a lower training reliability of 66%, much



**Fig. 5** Reliability analysis results for the three scenarios considered in this study. Scenario 1 includes water quality parameters with low  $S_n$ , Scenario 2 includes parameters with low to medium  $S_n$ , and Scenario 3 considers all parameters, including those with high  $S_n$

lower than the reliability score achieved by other ML models. Finally, SVR demonstrated the least reliability percentage among all ML models as it reached a training reliability of 60% and a testing reliability of 53.75%.

Regarding the second scenario, all ML models exploited in this study are more reliable compared to their corresponding reliability scores obtained in the first scenario. Furthermore, ANN, MLR, GP, GBR, and KNN demonstrated a remarkable performance, achieving 100% reliability scores for both training and testing datasets. Additionally, while RFR and XGBR yielded slightly lower reliability percentages for the training data (99.17% and 97.92%, respectively), they reached 100% reliability for the testing data. In contrast, DTR, SVR and AB emerged as the least reliable models for the second scenario. Specifically, SVR displayed the lowest training reliability (i.e., 87.92%), while AB reached the lowest reliability percentage for the testing dataset (i.e., 91.25%).

Regarding the third scenario, the results of the reliability analysis indicate that all ML models overall exhibited a higher reliability compared to other scenarios. Particularly GP, MGGP, MLR, ANN, GBR and XGBR showed remarkable performances by attaining the highest reliability scores of 100% for both training and testing datasets. Although KNN displayed slightly diminished reliability for the testing data (98.75%), it maintained a 100% reliability score for the training data. Furthermore, the RFR and DTR models resulted in adequate reliability percentages for both training and testing datasets. On the other hand, SVR and AB emerged as the least reliable models for the third scenario, with AB achieving the lowest training and testing reliability scores (i.e., 79.58% and 91.25%, respectively).

Figure 6 compares the confidence limits of various ML models. The y-axis displays the GWQI values, while the vertical lines represent the ML models, each line featuring (i) the 95% confidence interval, and (ii) the average value predicted by ML models. The initial vertical line, referred to as the “benchmark”, illustrates the confidence limits for the observed GWQI values. For the first scenario, upon analyzing the confidence limits depicted in Fig. 6, it becomes evident that none of the ML models in the first scenario closely align with the observed limits. To be more precise, most ML models exhibit lower confidence limits compared to that of the benchmark, signifying an underestimation of most GWQI. Conversely, the GBR model has higher confidence limits than that of the benchmark values, indicating a tendency to overestimate GWQI. In the case of the SVR model, while its average limit closely approximates that of the benchmark, the maximum and minimum values of its confidence limit deviate from those limits of the benchmark. This suggests that the SVR model tends not only to overestimate small GWQI but also to underestimate large GWQI.

Comparing the result of first and second scenario, it is obvious that the confidence limits of the ML models in the second scenario indicate a much closer proximity to those of the observed values. Most of ML models (ANN, MGGP, GBR, MLR, and GP) demonstrated confidence limits that replicated the benchmark. Additionally, while RFR, XGBR, DTR, and SVR exhibited a slight deviation from the benchmark, their confidence limits remained in a close alignment with it. Notably, the AB model displayed slightly larger confidence limits than that of the benchmark, indicating an overestimation of certain cases of GWQI. Conversely, KNN achieved lower confidence limits and the most significant deviations from the benchmark, suggesting a tendency to underestimate GWQI.

Regarding the third scenario, the results are almost the same, where most of ML models demonstrated confidence limits close to the confidence limits of the benchmark.

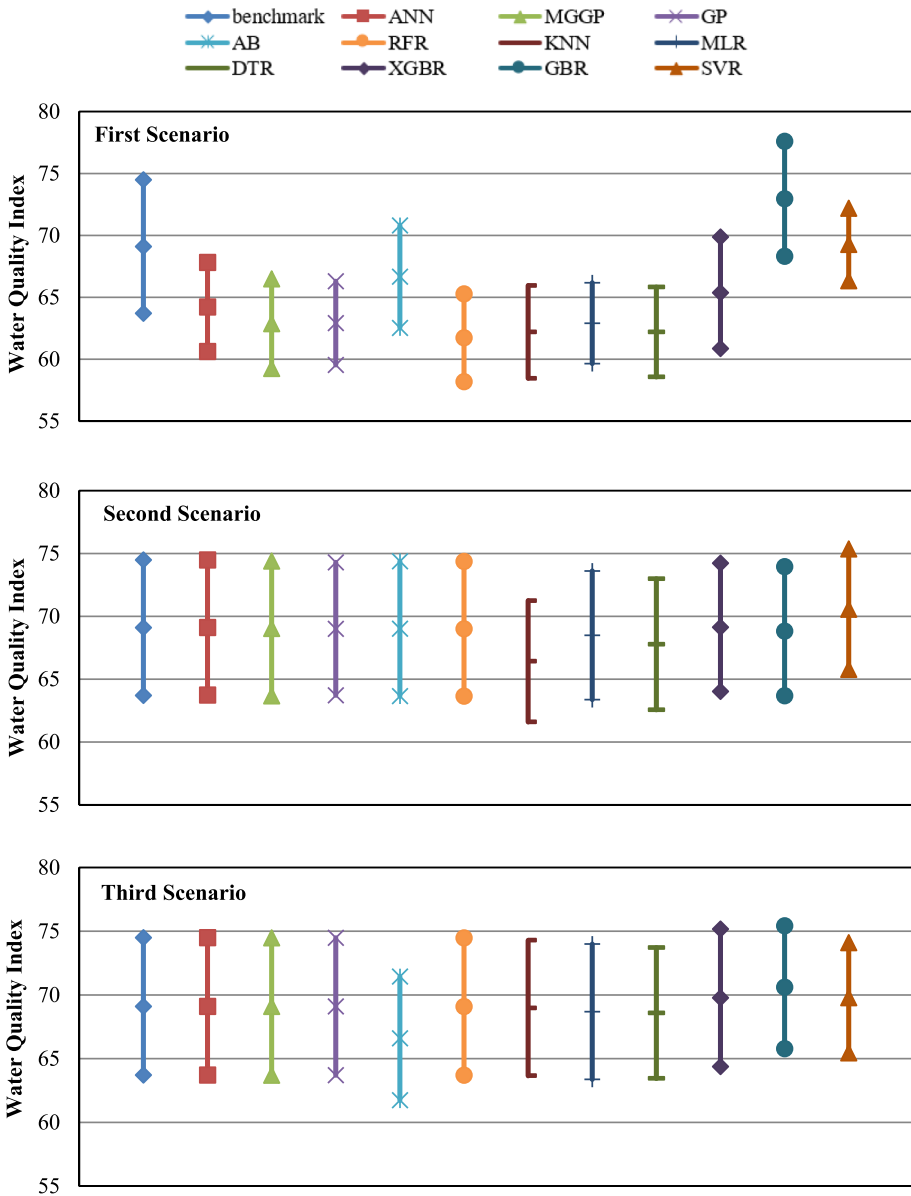
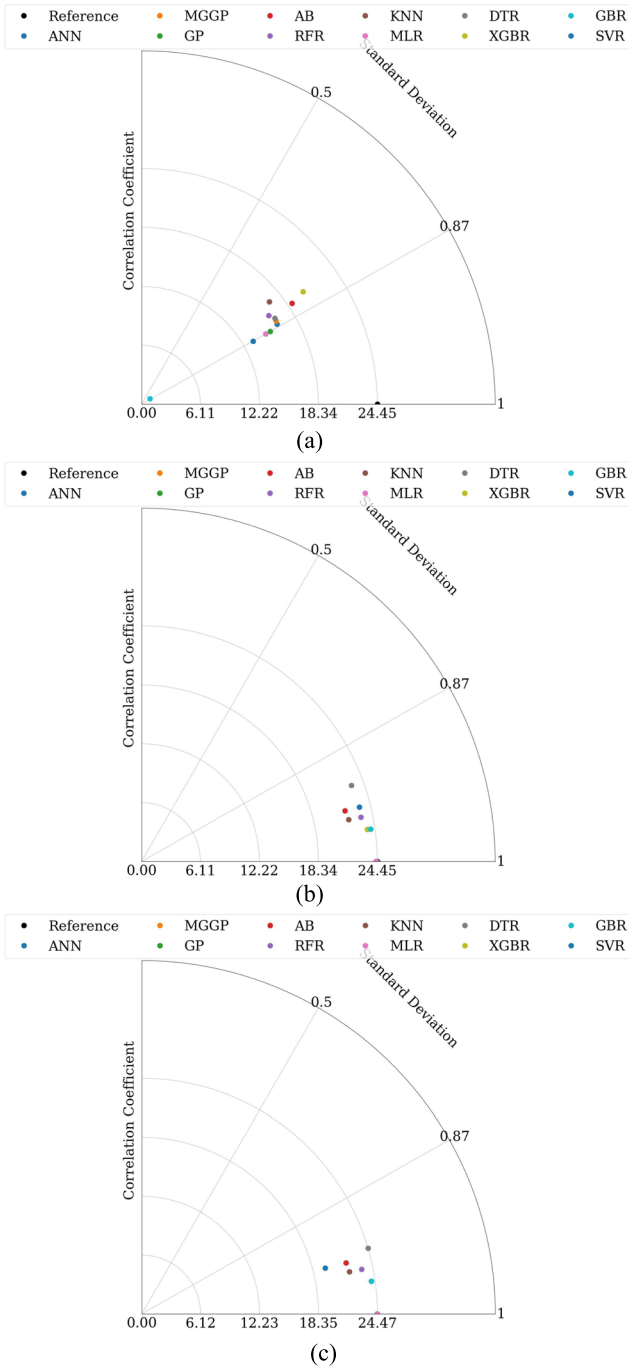


Fig. 6 Confidence limits of the observed and estimated GWQI obtained by the ML models for three scenarios considered in this study

Additionally, KNN and AB showed performances like they did in the second scenario. Finally, the SVR model in the third scenario tends to overestimate smaller GWQI and underestimate larger GWQI, similar to its performance in the first scenario.



**Fig. 7** Taylor diagrams for different ML models for the testing data of: (a) first scenario; (b) second scenario; and (c) third scenario

Figure 7 depicts the Taylor diagram for the testing data of each scenario. As shown, the performances of ML models in the Taylor diagram are in line with the results of previous analyses.

## 4 Discussion

This study provides a comprehensive evaluation of eleven ML models applied to the prediction of GWQI under three different input scenarios, providing valuable insights into how different models handle varying data availability. The findings reveal notable distinctions in model performance across different scenarios, driven by both the availability of water quality parameters and the characteristics of the models themselves. These results address key gaps in the existing literature on ML applications for GWQI estimation, particularly in terms of scenario-based analysis and the exploration of underutilized models.

In the first scenario, where fewer input variables were used (low  $S_n$  values), ANN outperformed the other models. ANN strong performance with limited input data can be attributed to its ability to model complex relationships, even with fewer features making it ideal for modelling aquatic ecosystems (Hadjisolomou et al. 2021). Following ANN, the MGGP, GP, and AB also performed well, demonstrating their ability to handle limited input variables. However, the extreme relative errors (i.e., MXARE) ranged from 0.36 to 0.68, with models like XGBR, GBR and SVR exhibiting the weakest performance.

In the second scenario, which incorporated a broader set of input features ( $S_n$  values from low to medium), all ML models demonstrated improved performance. The range of MXARE narrowed significantly, from 0.01 to 0.39, indicating that the broader input data enhanced the accuracy of all models. Once again, ANN ranked highest, followed by MGGP, GBR, MLR and GP. Models like SVR and AB showed the weakest performance, and RFR, XGBR and DTR failed again to rank among the top performers. The high performance of most models in this scenario suggests that the combination of inputs with low to medium  $S_n$  values is sufficient for accurate GWQI prediction. Notably, there was no major leap in performance when transitioning to the third scenario, reinforcing the notion that adding more features beyond a certain threshold may not significantly improve model accuracy. Same outcome was also achieved in various studies that used feature importance analysis on their proposed ML models (Kulisz et al. 2021; Elbeltagi et al. 2022; El-Rawy et al. 2024; Jibrin et al. 2024).

In the third scenario, where all input features ( $S_n$  values from low to high) were used, MLR achieved the best performance. This is likely because the GWQI calculation is based on a linear equation, and MLR excels at capturing linear relationships between inputs and outputs. The proficiency of MLR or other simple linear algorithms such as Ridge Regression was also stated in previous studies in the literature (Kouadri et al. 2021; El-Rawy et al. 2024). However, MLR required all available input features to achieve optimal performance, making it less versatile for scenarios with fewer parameters. This limitation was also noted by Kouadri et al. (2021), who demonstrated that while MLR excels as the best estimator when all parameters are available, its performance significantly declines when only TDS and TH are accessible. GP and ANN followed MLR in performance, securing second and third ranks, respectively. Models like XGBR, RFR, DTR, AB and SVR ranked at the bottom. The MXARE in this scenario ranged from 0.01 to 0.34, showing slight improvements for some models and minor declines for others, reinforcing the conclusion that additional features did not lead to significant gains.

Regarding the reliability of the ML models, in the first scenario all models exhibited testing reliability percentages higher than 50%, ranging from 53 to 75%. KNN achieved the highest training reliability (94.58%), while its testing reliability was 73.75%. Notably, despite ANN being the top performer across most scenarios, it had lower reliability in training (80%), but the testing reliability was the highest (75%), indicating effective training with close reliability percentages across datasets. In the second scenario, reliability improved significantly across all models, with all achieving testing reliabilities above 90%. Five models, i.e., ANN, MLR, GP, GBR, and KNN, achieved perfect reliability (100%) in both the training and testing datasets, reflecting their ability to consistently predict GWQI with the broader input dataset. This trend continued in the third scenario, where the reliability remained high, with most models exceeding 90% reliability. Models such as ANN, MLR, GP, GBR, MGGP, and XGBR achieved perfect reliability in both datasets. However, SVR exhibited the lowest training reliability 79.58% indicating it remained less reliable than the other models. The confidence limit analysis provides additional insight into the robustness of the model predictions. In the first scenario, most models displayed confidence limits that deviated from the benchmark values, with models such as GBR overestimating GWQI and SVR tending to overestimate smaller values and underestimate larger values. In the second scenario, the confidence limits of most models, including ANN, MGGP, GP, and MLR, aligned closely with the benchmark, indicating a strong fit between predicted and observed values. The third scenario exhibited similar results, where confidence limits remained tightly aligned with the benchmark for top-performing models like MLR, GP, and ANN, further confirming model reliability and accuracy. Although numerous studies in the literature have explored the application of ML models in GWQI prediction, very few have employed post-processing analyses, such as reliability assessments. Raheja et al. (2022) is one of the few that conducted a reliability analysis, demonstrating high reliability in their GBR, XGBR, and particularly their DNN model. Future studies must consider conducting post-processing analyses, such as reliability analysis, since they are crucial for further validating the results of ML models, ensuring the robustness and dependability of the predictions in real-world applications.

Overall, ANN proved to be the most reliable and consistent performer across all scenarios, making it the best choice for GWQI prediction. Multiple studies have also highlighted the strong predictive capabilities of this model (Sakizadeh 2016; Kulisz et al. 2021; Sajib et al. 2023). It was closely followed by MGGP and GP, which also performed exceptionally well. Jibrin et al. (2024) also demonstrated the outperformance of their GP model compared to the ANFIS and DTR models in their study. Although ANN demonstrated superior performance, its high computational time could limit its applicability compared to alternatives like GP. Furthermore, while MLR achieved the highest performance in the third scenario, its results in other scenarios were less robust, placing it in the fourth position overall. KNN tendency to overfit to the training data limited its performance, but it still ranked fifth due to commendable testing results. However, this result highlights the need for careful validation when using KNN to avoid overfitting. This was also the case in various studies (Khiavi et al. 2023; Sahour et al. 2023; El-Rawy et al. 2024). Beside MGGP, other tree-based models like GBR, RFR, XGBR, AB and DTR, did not perform as well in this study, likely due to their preference for nonlinear regression tasks. Notably, studies that relied solely on such tree-based models or variants like the M5P model (Norouzi and Moghaddam 2020; Elbeltagi et al. 2022), might have benefitted from incorporating a broader range of ML models for GWQI prediction. Lastly, SVR ranked lowest across all scenarios, demonstrating its inability in capturing the underlying relationships inherent in GWQI estimation as effectively as other models. While some studies in the literature



support this finding (Kouadri et al. 2021; Elbeltagi et al. 2022; Sajib et al. 2023), the effectiveness of ML models is significantly influenced by the data at hand and the appropriate tuning of hyperparameters. This can lead to scenarios where models like SVR or RFR outperform ANN in certain studies (Mohammed et al. 2023; Sahour et al. 2023).

The results of this study emphasize the importance of selecting appropriate ML models for GWQI prediction. While all models demonstrated commendable performance, the scenario-based approach highlights how the proper choice of models and input parameters can lead to better predictive accuracy and reliability. As noted in the literature review in the introduction, most studies have employed Pearson correlation analysis prior to modeling, while only a few have integrated it directly into their modeling processes. This method relies on the available data, and one distinguishing aspect of this study, compared to previous research, is the utilization of a more stable choice of scenario design. These findings contribute to the broader literature on water quality modeling, suggesting that a more comprehensive selection of ML models could improve the accuracy of groundwater quality assessments in future studies. Furthermore, future studies could benefit from the splitting method applied in this research, which divides the data into three scenarios based on  $S_n$  ranges. This approach may lead to a more thorough investigation of the results, ultimately contributing to more reliable outcomes. Additionally, while this study demonstrated the effectiveness of the ML models in predicting GWQI, indicating a reduced necessity for more complex methods such as deep learning or hybrid models for this specific task, future research can explore these models to demonstrate that their use, even if they perform better, may not be essential. Lastly, one limitation of this study, similar to others in the literature, is the restricted scope of the study area, which is typically confined to a single location. Future research could incorporate data from diverse regions across the globe to improve the evaluation of ML models for this task.

## 5 Conclusions

Analysis of groundwater quality is essential as it serves as one of the most important water resources, particularly for drinking purpose in arid and semi-arid regions. For such analysis, availability of groundwater quality observations is in need. Since measuring all groundwater quality parameters may not be feasible in some regions, estimating GWQI based on available water quality parameters is inevitable. Thus, the main challenges of GWQI predictions include data availability, assessment of data-driven estimation models, and quantifying how much precision can be obtained when not all water quality parameters are known. This study took a step forward to address these challenges for a case study in an arid region, where groundwater has been used for drinking and agricultural purposes. In this regard, eleven ML methods were exploited to estimate GWQI considering three scenarios with different sets of input data. In addition to comprehensive assessment of various ML models in predicting GWQI, a special aspect of this study lies in its unique scenario-based approach, which highlights how varying data availability can impact on model performances. This approach not only enhances the understanding of the strength and weakness of each model but also offers a more practical framework (compared to traditional methods such as Pearson correlation) for future GWQI assessments. Comparing performances of different ML models indicated that ANN, MGGP, GP, GBR and MLR achieved the most robust estimations for GWQI across different scenarios. Among these models, ANN was among the first three ML models for all scenarios, while MGGP and

GP were among the first three ML models for two scenarios. This study also addresses critical limitations in existing studies by incorporating reliability analyses and confidence limit assessments, which are often overlooked. The analyses facilitate choosing more reliable estimation models that can be effectively utilized in real-world applications to predict groundwater quality, thereby aiding decision-makers in water resources management. The reliability analysis showed that GP, MGGP, MLR, ANN, GBR and XGBR reach the highest reliability percentages for different scenarios. Specifically, ANN constantly ranked among the top ML models for reliability. Furthermore, checking confidence limits of ML-based estimation models revealed that most models forecasted GWQI predictions closely aligned with the observed values, especially for the second and third scenarios. Particularly, GWQI estimations carried out by ANN, MGGP and GP demonstrate remarkable consistency with benchmark values. Moreover, KNN and AB tended to underestimate and overestimate GWQI, respectively. These findings highlight the effectiveness of ANN, MGGP and GP in providing robust and reliable GWQI predictions. Indeed, further studies on groundwater quality data from various regions are required to delineate performances of data-driven models for predicting GWQI, whereas this study along with previous ones are limited to a specific study area. While hybrid models or deep learning techniques could potentially enhance prediction accuracy, this study indicates that simpler models can achieve reliable results without the complexities of implementation. In this regard, future studies could further explore the trade of between complexity and accuracy of ML-based models for estimating GWQI. Finally, the robust performance of ML models, like ANN and MGGP, offers a reliable foundation for future research aimed at improving groundwater management strategies.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s40710-025-00751-9>.

**Authors Contributions** M. N. and M. R. G. contributed to the study conception and design. Material preparation was performed by M. N., R. P., and M. R. G. Data collection was performed by M. R. G. Analysis was performed by M. N. and R. P. The first draft of the manuscript was written by M. N. and R. P. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

**Funding** The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

**Data Availability** The data used in this study can be supplied from the corresponding author upon request.

## Declarations

**Ethical Approval** Not applicable.

**Consent to Participate** Not applicable.

**Consent to Publish** Not applicable.

**Competing Interest** The authors declare no competing interests.

## References

Abba S, Yassin MA, Shah SMH, Egbueri JC, Elzain HE, Agbasi JC, Saini G, Usaman J, Khan NA, Aljundi IH (2024) Trace element pollution tracking in the complex multi-aquifer groundwater system of

- Al-Hassa oasis (Saudi Arabia) using spatial, chemometric and index-based techniques. *Environ Res* 249:118320. <https://doi.org/10.1016/j.envres.2024.118320>
- Abidi JH, Elzain HE, Sabarathinam C, Selmane T, Selvam S, Farhat B, Mammou AB, Senapathi V (2024) Evaluation of groundwater quality indices using multi-criteria decision-making techniques and a fuzzy logic model in an irrigated area. *Groundw Sustain Dev* 25:101122. <https://doi.org/10.1016/j.gsd.2024.101122>
- Agrawal P, Sinha A, Kumar S, Agarwal A, Banerjee A, Villuri VGK, Annavarapu CSR, Dwivedi R, Dera VVR, Sinha J (2021) Exploring artificial intelligence techniques for groundwater quality assessment. *Water* 13(9):1172. <https://doi.org/10.3390/w13091172>
- Aju C, Achu A, Mohammed MP, Raicy M, Gopinath G, Reghunath R (2024) Groundwater quality prediction and risk assessment in Kerala, India: a machine-learning approach. *J Environ Manage* 370:122616. <https://doi.org/10.1016/j.jenvman.2024.122616>
- Bedi S, Samal A, Ray C, Snow D (2020) Comparative evaluation of machine learning models for groundwater quality assessment. *Environ Monit Assess* 192:1–23. <https://doi.org/10.1007/s10661-020-08695-3>
- Carbajal-Hernández JJ, Sánchez-Fernández LP, Villa-Vargas LA, Carrasco-Ochoa JA, Martínez-Trinidad JF (2013) Water quality assessment in shrimp culture using an analytical hierarchical process. *Ecol Ind* 29:148–158. <https://doi.org/10.1016/j.ecolind.2012.12.017>
- Di Nunno F, Zhu S, Ptak M, Sojka M, Granata F (2023) A stacked machine learning model for multi-step ahead prediction of lake surface water temperature. *Sci Total Environ* 890:164323. <https://doi.org/10.1016/j.scitotenv.2023.164323>
- Elbeltagi A, Pande CB, Kouadri S, Islam ARMT (2022) Applications of various data-driven models for the prediction of groundwater quality index in the Akot basin, Maharashtra, India. *Environ Sci Pollut Res* 1–15. <https://doi.org/10.1007/s11356-021-17064-7>
- El-Magd SAA, Ismael IS, El-Sabri MAS, Abdo MS, Farhat HI (2023) Integrated machine learning-based model and WQI for groundwater quality assessment: ML, geospatial, and hydro-index approaches. *Environ Sci Pollut Res Int* 30(18):53862. <https://doi.org/10.1007/s11356-023-25938-1>
- El-Rawy M, Wahba M, Fathi H, Alshehri F, Abdalla F, El Attar RM (2024) Assessment of groundwater quality in arid regions utilizing principal component analysis, GIS, and machine learning techniques. *Mar Pollut Bull* 205:116645. <https://doi.org/10.1016/j.marpolbul.2024.116645>
- Elzain HE, Chung SY, Senapathi V, Sekar S, Lee SY, Roy PD, Hassan A, Sabarathinam C (2022) Comparative study of machine learning models for evaluating groundwater vulnerability to nitrate contamination. *Ecotoxicol Environ Saf* 229:113061. <https://doi.org/10.1016/j.ecoenv.2021.113061>
- Elzain HE, Chung SY, Venkatraman S, Selvam S, Ahemd HA, Seo YK, Bhuyan MS, Yassin MA (2023) Novel machine learning algorithms to predict the groundwater vulnerability index to nitrate pollution at two levels of modeling. *Chemosphere* 314:137671. <https://doi.org/10.1016/j.chemosphere.2022.137671>
- Elzain HE, Abdalla O, Ahmed HA, Kacimov A, Al-Maktoumi A, Al-Higgi K, Abdallah M, Yassin MA, Senapathi V (2024) An innovative approach for predicting groundwater TDS using optimized ensemble machine learning algorithms at two levels of modeling strategy. *J Environ Manage* 351:119896. <https://doi.org/10.1016/j.jenvman.2023.119896>
- Eslamian S, Ostad-Ali-Askari K, Dehghan S, Singh VP, Dalezios NR, Ghane M (2018) Comparison of climatic features of wet and dry areas. *Int J Emerg Eng Res Technol* 6(5):12–22. <https://doi.org/10.2139/ssrn.4825518>
- Farhadinejad T, Khakzad A, Jafari M, Shoaee Z, Khosrotehrani K, Nobari R, Shahrokh V (2014) The study of environmental effects of chemical fertilizers and domestic sewage on water quality of Taft region, Central Iran. *Arab J Geosci* 7:221–229. <https://doi.org/10.1007/s12517-012-0717-0>
- Fernández N, Ramírez A, Solano F (2004) Physico-chemical water quality indices—a comparative review. *Bistua: Rev Fac Cienc Básicas* 2(1):19–30
- Goodarzi MR, Niknam ARR, Barzkar A, Niazkari M, Zare Mehrjerdi Y, Abedi MJ, Heydari Pour M (2023) Water quality index estimations using machine learning algorithms: a case study of Yazd-Ardakan Plain. *Iran Water* 15(10):1876. <https://doi.org/10.3390/w15101876>
- Granata F, Zhu S, Di Nunno F (2024) Dissolved oxygen forecasting in the Mississippi River: advanced ensemble machine learning models. *Environ Sci: Adv*. <https://doi.org/10.1039/D4VA00119B>
- Hadjisolomou E, Stefanidis K, Herodotou H, Michaelides M, Papatheodorou G, Papastergiadou E (2021) Modelling freshwater eutrophication with limited limnological data using artificial neural networks. *Water* 13(11):1590. <https://doi.org/10.3390/w13111590>
- Haggerty R, Sun J, Yu H, Li Y (2023) Application of machine learning in groundwater quality modeling—A comprehensive review. *Water Res* 233:119745. <https://doi.org/10.1016/j.watres.2023.119745>
- Horton RK (1965) An index number system for rating water quality. *J Water Pollut Control Fed* 37:300–306


- Hussein EE, Derdour A, Zerouali B, Almaliki A, Wong YJ, Ballesta-de los Santos M, Minh Ngoc P, Hashim MA, Elbeltagi A (2024) Groundwater quality assessment and irrigation water quality index prediction using machine learning algorithms. *Water* 16(2):264. <https://doi.org/10.3390/w16020264>
- Ibrahim H, Yaseen ZM, Scholz M, Ali M, Gad M, Elsayed S, Khadr M, Hussein H, Ibrahim HH, Eid MH (2023) Evaluation and prediction of groundwater quality for irrigation using an integrated water quality indices, machine learning models and GIS approaches: a representative case study. *Water* 15(4):694. <https://doi.org/10.3390/w15040694>
- IRIMO (2015) Iran meteorological organization (in Persian). <http://www.irimo.ir>. Accessed 25 Feb 2024
- Jamshidzadeh Z (2020) An integrated approach of hydrogeochemistry, statistical analysis, and drinking water quality index for groundwater assessment. *Environ Process* 7(3):781–804. <https://doi.org/10.1007/s40710-020-00450-7>
- Jibrin AM, Al-Suwaiyan M, Aldrees A, Dan'azumi S, Usman J, Abba SI, Yassin MA, Scholz M, Sammen SS (2024) Machine learning predictive insight of water pollution and groundwater quality in the Eastern Province of Saudi Arabia. *Sci Rep* 14(1):20031. <https://doi.org/10.1038/s41598-024-70610-4>
- Jonnalagadda S, Mhere G (2001) Water quality of the Odzi River in the eastern highlands of Zimbabwe. *Water Res* 35(10):2371–2376. [https://doi.org/10.1016/S0043-1354\(00\)00533-9](https://doi.org/10.1016/S0043-1354(00)00533-9)
- Kheradpisheh Z, Talebi A, Rafati L, Ghaneian MT, Ehrampoush MH (2015) Groundwater quality assessment using artificial neural network: A case study of Bahabad plain, Yazd Iran. *Desert* 20(1):65–71. <https://doi.org/10.22059/jdesert.2015.54084>
- Khiavi AN, Tavooosi M, Kuriqi A (2023) Conjunct application of machine learning and game theory in groundwater quality mapping. *Environ Earth Sci* 82(17):395. <https://doi.org/10.1007/s12665-023-11059-y>
- Kouadri S, Elbeltagi A, Islam ARMT, Kateb S (2021) Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). *Appl Water Sci* 11(12):190. <https://doi.org/10.1007/s13201-021-01528-9>
- Kulisz M, Kujawska J, Przysucha B, Cel W (2021) Forecasting water quality index in groundwater using artificial neural network. *Energies* 14(18):5875. <https://doi.org/10.3390/en14185875>
- Mishra BK, Regmi RK, Masago Y, Fukushi K, Kumar P, Saraswat C (2017) Assessment of Bagmati river pollution in Kathmandu Valley: scenario-based modeling and analysis for sustainable urban development. *Sustain Water Qual Ecol* 9:67–77. <https://doi.org/10.1016/j.swaqe.2017.06.001>
- Mohammed MAA, Khleel NAA, Szabó NP, Szűcs P (2023) Modeling of groundwater quality index by using artificial intelligence algorithms in northern Khartoum State, Sudan. *Model Earth Syst Environ* 9(2):2501–2516. <https://doi.org/10.1007/s40808-022-01638-6>
- Nathan NS, Saravanane R, Sundararajan T (2017) Application of ANN and MLR models on groundwater quality using CWQI at Lawspet, Puducherry in India. *J Geosci Environ Prot* 5(03):99. <https://doi.org/10.4236/gep.2017.53008>
- Niazkar M, Goodarzi MR, Fatehifar A, Abedi MJ (2023) Machine learning-based downscaling: application of multi-gene genetic programming for downscaling daily temperature at Dogonbadan, Iran, under CMIP6 scenarios. *Theoret Appl Climatol* 151(1–2):153–168. <https://doi.org/10.1007/s00704-022-04274-3>
- Niazkar M (2023) Multigene genetic programming and its various applications. *Handbook of Hydroinformatics*, pp 321–332. <https://doi.org/10.1016/B978-0-12-821285-1.00019-1>
- Norouzi H, Moghaddam AA (2020) Groundwater quality assessment using random forest method based on groundwater quality indices (case study: Miandoab plain aquifer, NW of Iran). *Arab J Geosci* 13(18):912. <https://doi.org/10.1007/s12517-020-05904-8>
- Piraei R, Niazkar M, Afzali SH (2023a) Assessment of data-driven models for estimating total sediment discharge. *Earth Sci Inf* 16(3):2795–2812. <https://doi.org/10.1007/s12145-023-01069-6>
- Piraei R, Niazkar M, Afzali SH, Menapace A (2023b) Application of machine learning models to bridge afflux estimation. *Water* 15(12):2187. <https://doi.org/10.3390/w15122187>
- Prasad M, Sunitha V, Reddy YS, Suvarna B, Reddy BM, Reddy MR (2019) Data on water quality index development for groundwater quality assessment from Obulavaripalli Mandal, YSR district, A.P India. *Data Brief* 24:103846. <https://doi.org/10.1016/j.dib.2019.103846>
- Prusty P, Farooq SH (2020) Application of water quality index and multivariate statistical analysis for assessing coastal water quality. *Environ Process* 7:805–825. <https://doi.org/10.1007/s40710-020-00453-4>
- Raheja H, Goel A, Pal M (2022) Prediction of groundwater quality indices using machine learning algorithms. *Water Pract Technol* 17(1):336–351. <https://doi.org/10.2166/wpt.2021.120>
- Rajeev A, Shah R, Shah P, Shah M, Nanavaty R (2024) The potential of big data and machine learning for ground water quality assessment and prediction. *Arch Comput Methods Eng* 1–15. <https://doi.org/10.1007/s11831-024-10156-w>
- Ransom KM, Nolan BT, Traum JA, Faunt CC, Bell AM, Gronberg JAM, Wheeler DC, Rosecrans CZ, Jurgens B, Schwarz GE (2017) A hybrid machine learning model to predict and visualize nitrate

- concentration throughout the Central Valley aquifer, California, USA. *Sci Total Environ* 601:1160–1172. <https://doi.org/10.1016/j.scitotenv.2017.05.192>
- Ransom KM, Nolan BT, Stackelberg P, Belitz K, Fram MS (2022) Machine learning predictions of nitrate in groundwater used for drinking supply in the conterminous United States. *Sci Total Environ* 807:151065. <https://doi.org/10.1016/j.scitotenv.2021.151065>
- Roushangar K, Davoudi S, Shahnazi S (2023) The potential of novel hybrid SBO-based long short-term memory network for prediction of dissolved oxygen concentration in successive points of the Savannah River, USA. *Environ Sci Pollut Res* 30(16):46960–46978. <https://doi.org/10.1007/s11356-023-25539-y>
- Sahour S, Khanbeyki M, Gholami V, Sahour H, Kahvazade I, Karimi H (2023) Evaluation of machine learning algorithms for groundwater quality modeling. *Environ Sci Pollut Res* 30(16):46004–46021. <https://doi.org/10.1007/s11356-023-25596-3>
- Sajib AM, Diganta MTM, Rahman A, Dabrowski T, Olbert AI, Uddin MG (2023) Developing a novel tool for assessing the groundwater incorporating water quality index and machine learning approach. *Groundw Sustain Dev* 23:101049. <https://doi.org/10.1016/j.gsd.2023.101049>
- Sakizadeh M (2016) Artificial intelligence for the prediction of water quality index in groundwater systems. *Model Earth Syst Environ* 2:1–9. <https://doi.org/10.1007/s40808-015-0063-9>
- Salehi M, Mousavi M, Kashkooli AB (2014) Assessment and classification of environmental problems based on sustainable development indexes (Case Study: Cities of Yazd Province). *Assessment* 4(2):110–119
- Singha S, Pasupuleti S, Singha SS, Singh R, Kumar S (2021) Prediction of groundwater quality using efficient machine learning technique. *Chemosphere* 276:130265. <https://doi.org/10.1016/j.chemosphere.2021.130265>
- Srebotnjak T, Carr G, de Sherbinin A, Rickwood C (2012) A global water quality index and hot-deck imputation of missing data. *Ecol Ind* 17:108–119. <https://doi.org/10.1016/j.ecolind.2011.04.023>
- Stackelberg PE, Belitz K, Brown CJ, Erickson ML, Elliott SM, Kauffman LJ, Ransom KM, Reddy JE (2021) Machine learning predictions of pH in the glacial aquifer system, Northern USA. *Groundwater* 59(3):352–368. <https://doi.org/10.1111/gwat.13063>
- Suphawan K, Chaisee K (2021) Gaussian process regression for predicting water quality index: a case study on Ping River basin, Thailand. *AIMS Environ Sci* 8(3). <https://doi.org/10.3934/environsci.2021018>
- Tang M, Zeng H, Wang K (2022) Bayesian water quality evaluation model based on generalized triangular fuzzy number and its application. *Environ Process* 9(1):6. <https://doi.org/10.1007/s40710-022-00562-2>
- Torres-Martínez JA, Mahlknecht J, Kumar M, Loge FJ, Kaown D (2024) Advancing groundwater quality predictions: machine learning challenges and solutions. *Sci Total Environ* 174973. <https://doi.org/10.1016/j.scitotenv.2024.174973>
- Trabelsi F, Bel Hadj Ali S (2022) Exploring machine learning models in predicting irrigation groundwater quality indices for effective decision making in Medjerda River Basin, Tunisia. *Sustainability* 14(4):2341. <https://doi.org/10.3390/su14042341>
- Tyagi PK, Shruti SV, Ahuja A (2012) Synthesis of metal nanoparticles: a biological prospective for analysis. *Int J Pharm Innov* 2(4):48–60
- UNEP (2016) A snapshot of the world's water quality: towards a global assessment. United Nations Environment Programme, Nairobi, Kenya, p 162
- Wade C, Glynn K (2020) Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python. Packt Publishing Ltd, p 310
- WHO (2011) Guidelines for drinking-water quality. World Health Organization, 4th edition, Geneva, p 564
- Yang S, Luo D, Tan J, Li S, Song X, Xiong R, Wang J, Ma C, Xiong H (2024) Spatial mapping and prediction of groundwater quality using ensemble learning models and shapley additive explanations with spatial uncertainty analysis. *Water* 16(17):2375. <https://doi.org/10.3390/w16172375>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## Authors and Affiliations

**Majid Niazkar<sup>1,2</sup> · Reza Piraei<sup>3</sup> · Mohammad Reza Goodarzi<sup>4,5</sup>  · Mohammad Javad Abedi<sup>6</sup>**

✉ Majid Niazkar  
majid.niazkar@cmcc.it; majid.niazkar@unive.it

Reza Piraei  
piraeir@gmail.com

Mohammad Reza Goodarzi  
Goodarzimr@um.ac.ir; Goodarzimr@yazd.ac.ir

Mohammad Javad Abedi  
mgabedi13@yahoo.com

<sup>1</sup> Euro-Mediterranean Center On Climate Change, Porta Dell’Innovazione Building - 2Nd Floor, Via Della Libertà, 12 - 30175 Venice, VE, Italy

<sup>2</sup> Ca ‘Foscari University of Venice, Venice, Italy

<sup>3</sup> Department of Civil Engineering, Shiraz University, Shiraz, Iran

<sup>4</sup> Department of Civil Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

<sup>5</sup> Department of Civil Engineering, Yazd University, Yazd, Iran

<sup>6</sup> Department of Civil Engineering, Water Resources Management Engineering, Yazd University, Yazd, Iran