ISSN 1028-334X, Doklady Earth Sciences, 2023, Vol. 512, Part 1, pp. 907–914. © Pleiades Publishing, Ltd., 2023.
ISSN 1028-334X, Doklady Earth Sciences, 2023. © Pleiades Publishing, Ltd., 2023.

Forecasting Groundwater Quality Parameters using Machine Learning Models: a Case Study of Khemismiliana Plain, Algeria

A. Tachi[a],*, M. Metaiche[b], A. Messoul[a], H. Bouguerra[c], and S. E. Tachi[d]

Received April 19, 2023; revised April 22, 2023; accepted May 29, 2023

Abstract—Total Dissolved Solids, one of the most extensively used indicators for assessing groundwater quality, it useful to estimate salinity and hardness in water. The objective of the present study is to develop accurate and dependable machine learning models for forecasting the total dissolved solids, parameter; as well to evaluate and explain the relationship of total dissolved solids with the mineral salts. Four machine learning models Decision tree, Random forest, Adaboost and support victor regression SVR have been successfully employed for modeling the total dissolved solids using Electrical Conductivity (EC) and concentrations of major elements (Ca^{2+}, Mg^{2+}, Na^+, K^+, Cl^-, SO_4^{2-}, HCO_3^-, NO_3^-) of the groundwater aquifer in upper Cheliff plain (the northwestern of Algeria). One hundred ninety-one of observations collected from wells by the ANRH (national water resources agency, Algeria) for a period of 8 years between 2008 and 2016, were randomly divided into training and validation sets. The overall prediction performance results indicated that the models provided satisfactory estimation with priority to the support vector regression model, based on the four parameters including: EC, Na^+, SO_4^{2-}, and Cl^-, with the best support vector machine results of RMSE = 0.0328; NS = 0.9455. Feature selection method revealed that the correlation analysis results were reliable and could be utilized as a first step in selecting the optimum input data for forecasting groundwater quality parameters. Generally, the proposed models are useful in predicting groundwater quality parameters and may aid decision-makers in developing and managing groundwater plans.

Keywords: groundwater, total dissolved solid (TDS), support vector regression, Catboost, electrical conductivity, Algeria

DOI: 10.1134/S1028334X23600792

1. INTRODUCTION

Water is a crucial resource for supporting the survival and growth of all living organisms on Earth. It plays a vital role in maintaining ecosystem balance and providing habitats for a variety of plant and animal species. Among these resources, the groundwater is recognizing as one of the most valuable and essential natural sources of freshwater in the world for human consumption, agriculture and domestic purposes. The demand for groundwater is expected to increase in the future due to population growth and economic expansion [1]. Assessment of groundwater quality is an integral element of water management strategies, as it ensures optimal utilization of groundwater resources [2, 3].

Total dissolved solids (TDS) is a crucial parameter in determining the suitability of groundwater for drinking purposes, and is often used as a proxy for water quality assessment. Accurate estimation of TDS levels is crucial for effective freshwater resource management, as high TDS levels in groundwater can lead to salinization, rendering it unsuitable for irrigation and consumption [4] that can negatively affect soil quality and crop growth [5].Total dissolved solids (TDS) are primarily comprised of inorganic salts, such as nitrates (NO_3^-), sulfate (SO_4^{2-}), chloride (Cl^-), and bicarbonates (HCO_3^-) as anions, as well as cations such as magnesium (Mg^{2+}), sodium (Na^+), potassium (K^+), and calcium (Ca^{2+}). These salts originate from natural sources, including rocks and minerals, as well as anthropogenic activities such as agricultural and industrial practices. Apart from inorganic salts, TDS

[a] *Vegetal Chemistry—Water—Energy Laboratory (LCV2E), Department of Hydraulic, Faculty of Civil Engineering and Architecture, University of Hassiba Benbouali, Chlef, 0201 Algeria*

[b] *PM3E Laboratory, University of Bouira, Bouira, 10000 Algeria*

[c] *Water Resources and Sustainable Development Laboratory, Department of Geology, Faculty of Earth Sciences, Badji Mokhtar—Annaba University, P.O. Box 12, Annaba, 23000 Algeria*

[d] *Laboratoire de recherche des sciences de l'eau, Department of Hydraulic, National Polytechnic School, Algiers, 16200 Algeria*

**e-mail: tachiamina@gmail.com*

also includes dissolved organic materials, which contribute to the overall water quality. The World Health Organization (WHO) has published standards [6] for TDS measurement in water, which provide a benchmark for ensuring water quality. Although laboratory studies and computational methods are commonly employed to quantify TDS, traditional laboratory testing and manual calculations can be time-consuming, error-prone, and may lack the precision necessary for accurate calculations. Furthermore, such methods can often be cost-prohibitive [7, 8].

It is important to support agricultural and hydrological management studies by modeling the concentration of total dissolved solids (TDS) [9]. Numerous mathematical models have been widely used to predict and estimate TDS, with various approaches proposed in the literature. Among these, two popular methods include Sorensen's method [10], which estimates TDS based on the sum of ion concentrations in the water, and the method proposed by [11], which estimates TDS concentration using a linear regression equation derived from electrical conductivity. These methods are used to provide valuable insights into the quality of groundwater and its suitability for different uses. To ensure sustainable management of groundwater quality, which includes monitoring parameters such as TDS, EC, and major elements, it is crucial to utilize innovative and efficient evaluation methods. Robust and flexible models are essential for accurately predicting water quality and estimating future supplies, which can aid in addressing the challenges of groundwater planning and management [12, 13].

During the last two decades, the use of Artificial Intelligence (AI) techniques has increased in many fields, where have been explored algorithms for water quality modeling recently, which is an alternative and effective methods, these algorithms investigate hidden and complicated correlations between input data and output data in order to create models that effectively represent these correlations [12]. AI models have been widely used by researchers for predicting water quality variables; for instance, artificial neural network (ANN), multiple linear regressions (MLR), random forest regression (RFR), adaptive neuro-fuzzy inference system (ANFIS), support vector machines (SVM), and decision tree (DT). The literature review showed that these models have its strengths and weaknesses and its behavior depends on the different data and parameters in the selected study area. Moreover, these models have proven that is highly accurate in prediction; where these models include several advantages over traditional statistical and physical based models. Artificial intelligence (AI) models are known for their robust, non-linear, and flexible structures, which enable them to handle vast amounts of data at varying time scales. Furthermore, AI models are generally less sensitive to missing data than traditional approaches [12].

Extensive research and studies have demonstrated that machine learning (ML) models are highly accurate in forecasting and evaluating water quality, both for surface and groundwater. However, it is important to note that the efficiency of ML models is not solely dependent on prediction accuracy, but also on the nature and number of predictors used in the model [13].

The aim of this paper is to forecast the total dissolved solids in the upper Chellif aquifer using Random Forest, Decision tree, Support victor regression and Categorical boosting to obtain the optimum parameters to model TDS and by comparing the performances and the effectiveness of the different used machine learning algorithms.

2. MATERIALS

2.1. Study Area

The upper Cheliff alluvial plain is situated in the northwest region of Algeria and is an intra-mountainous depression that spans 500 km from east to west and 30 to 80 km wide. It is also known as the KhemisMiliana plain (wilaya of Ain Defla) and covers an area of approximately 21035 km^2. The plain is bounded to the north by the Zaccar mountains, to the south by the foothills of Ouarsenis, to the west by the Doin massif, and to the east by the threshold of Djendel (see Fig. 1). The area has a semi-arid Mediterranean-type climate with an average annual rainfall of approximately 380 mm, and the majority of the surface has a relatively low slope of less than 10%.

The Upper Cheliff alluvial plain comprises a mixed aquifer, consisting of Mio-Plio-Quaternary alluvial deposits such as sandstone with a thickness of 50 to 100 m at the level of Chellif wadi. The substratum of this aquifer horizon is composed of marls. Groundwater in this area is mainly derived from the alluvial aquifer of the upper Cheliff plain and is extracted through wells and drilling. The aquifer is recharged by seepage water from precipitation and runoff from wadis (such as Deurdeur, Chellif, Souffay, Boutane, etc.), as well as excess irrigation water.

2.2. Used Data

One hundred ninety-one samples of ground water quality from upper Chellif aquifer have been considered in this research that was collected in 36 wells record by the national agency of hydraulic resources. Groundwater sampling was carried out during the period 2008 and 2016.These samples were analyzed for several physicochemical variables, that is include the measurements of the parameters that influencing on water quality: TDS, EC, Ca^{2+}, Mg^{2+}, Na^+, K^+, Cl^-, SO_4^{2-}, HCO_3^-, and NO_3^-. The data exhibits outlier values on nearly all parameters, which makes it crucial to conduct thorough analysis prior to the modeling pro-

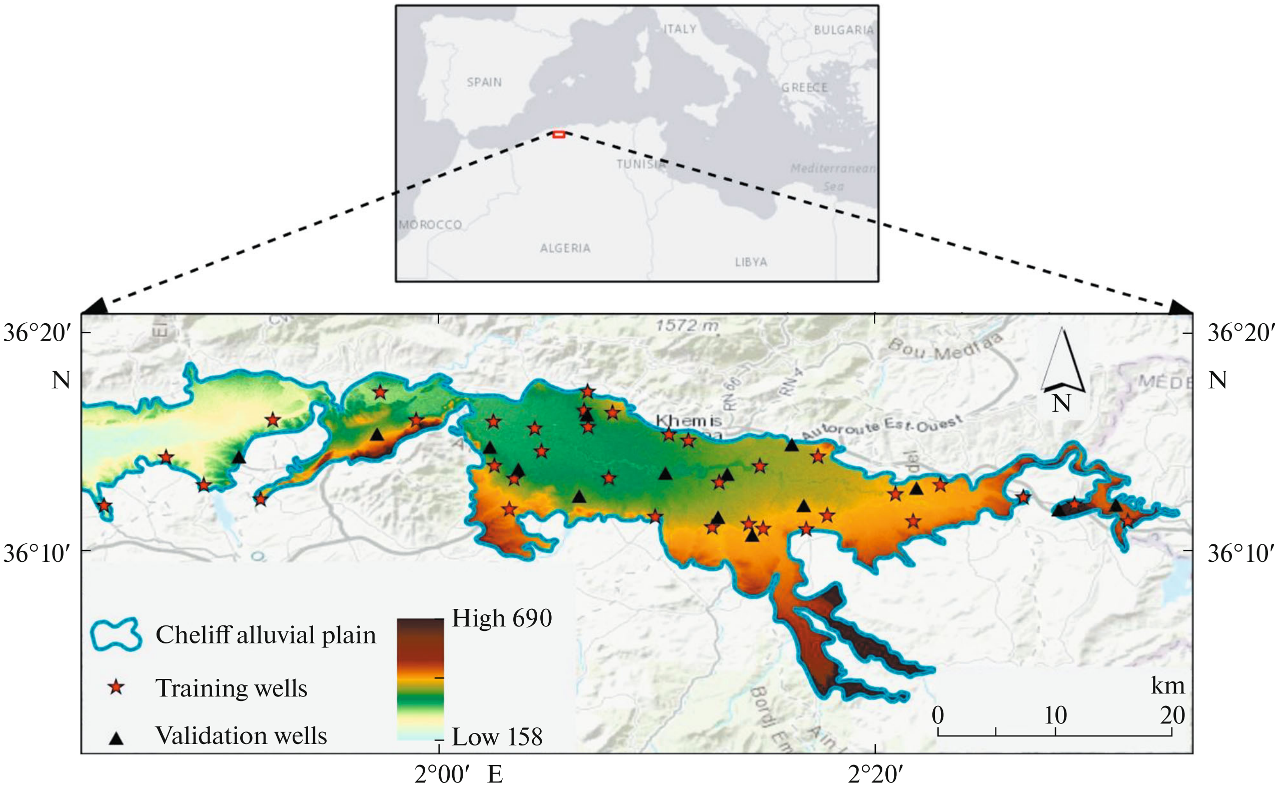

Fig. 1. Location map of the study area.

cess. In the development of machine learning models, exploring and cleaning the data are critical steps that ensure the accuracy and reliability of the resulting models.

In this study; the data pre-processing was conducted following two steps; (i) the reliability of the data was analyzed and consisted based on the verification of the ionic balances. This verification was preceded by an analysis of the major ions involved in the evaluation of the ion balances (IB). The ionic balance represents the difference between the sum of the anions and the sum of the cations, in other words, the anion-cation balance which naturally equals zero, that is to say the anions and the cations are equal (1). An ion balance of less than 5% was obtained; this shows that the results of the analysis are reliable. (ii) Data normalization from 0.05 to 0. To avoid the impact of variable dimensionality on model performance and increase model generalization; and on the other side, inaccuracy caused by a difference in measurement units between input and output parameters (2).

$$\text{IonBalance: IB}(\%) = 100\times\frac{\left|\sum \text{cations}\left(\frac{meq}{l}\right)-\sum \text{anions}\left(\frac{meq}{l}\right)\right|}{\sum \text{cations}\left(\frac{meq}{l}\right)+\sum \text{anions}\left(\frac{meq}{l}\right)}, \quad (1)$$

where

$$x_{\text{norm}} = \frac{(0.95-0.05)(x_i - x_{\min})}{(x_{\max}-x_{\min})} + 0.05, \quad (2)$$

where x_{norm} is the normalized data set, x_i is the original data set, $x_{\min}$ and $x_{\max}$ are respectively the minimum and the maximum values of the original dataset. Table 1 presents the statistical data of the training set 70% (36 wells) and the testing set 30% (15 wells).

3. METHODOLOGY

3.1. Decision Tree Model (DT)

Decision trees are supervised machine learning algorithms that are widely utilized for classification and regression applications to predict a response to data. The decision tree is one of the most important data mining methods. Thus, the Training data is used by this algorithm to create a set of decision rules. Thereafter the decision tree represented as a tree structure with a root node, branches, and leaf nodes. The decision tree is built by determining which characteristic or feature has the greatest information gain at each level of the algorithm [14].

3.2. Random Forest Model (RF)

Random Forest method is a supervised machine learning algorithm based on decision tree, introduced

Table 1. Statistical parameters of the training and validation datasets

Variables	Training dataset					Validation dataset				
	Min	Max	Mean	SD	CV	Min	Max	Mean	SD	CV
TDS (mg/L)	452.00	6571.00	1975.91	914.12	0.46	609.00	5686.00	1973.39	962.87	0.48
EC (μS/cm)	610.00	6800.00	2943.10	1257.23	0.43	993.00	8100.00	2941.77	1330.96	0.45
Ca^{2+} (mg/L)	62.96	489.80	246.96	98.02	0.40	117.00	470.00	247.47	82.67	0.33
Mg^{2+} (mg/L)	3.00	224.00	84.69	46.19	0.54	2.45	349.09	91.44	63.74	0.69
Na^+ (mg/L)	12.00	800.00	195.01	138.35	0.71	13.10	540.00	179.14	120.46	0.67
K^+ (mg/L)	0.10	48.00	6.07	7.39	1.21	0.20	34.20	5.98	7.26	1.20
Cl^- (mg/L)	60.00	1390.00	518.22	313.46	0.60	98.00	1690.00	529.80	315.59	0.59
SO_4^{2-} (mg/L)	40.00	2185.00	320.31	230.23	0.72	67.00	1540.00	330.13	247.38	0.74
HCO_3^- (mg/L)	146.40	762.50	334.86	101.09	0.30	148.84	457.50	303.21	71.30	0.23
NO_3^- (mg/L)	0.00	270.00	47.57	35.34	0.74	0.00	175.00	59.36	38.31	0.64

by [15], which can be applied for classification and regression learning [13]. This model is enhanced and extensively used in numerous studies and provided good performance. RF has emerged as a serious competitor with the boosting approach [16]. To generate predictions, the data is first partitioned into multiple sub-datasets using bootstrapping, a method employed by the random forest model. Decision trees are then constructed for each sub-dataset. These sub-decision trees are then combined using the random forest model to generate final predictions, which reflect the overall performance of the random forest model [17].

3.3. Categorical Boosting (CATBOOST)

The categorical boosting (catboost) is a new gradient boosting based decision tree (GBDT) algorithm proposed by [18]. It makes use of a sophisticated ensemble learning approach based on the gradient descent framework. During model training, a set of decision trees (DTs) are built sequentially to generate each DT, which learns from the previous tree and affects the next tree to improve model performance, resulting in a robust learner. This algorithm is different from the next traditional of gradient boosting trees (GBTs) algorithms; thus, it has two notable features are: ordered boosting and efficient handling of categorical features [18].

3.4. Support Victor Regression Model (SVR)

Support Victor Machine (SVM) is considered as a supervised machine learning algorithm that was developed earlier by [19], which could successfully have applied to solve classification, regression and pattern recognition problems [19]. This model demonstrates the correlation between input and output, with the purpose of minimizing the variance between observed and predicted data [20]. Support Vector Regression (SVR) is a newer and more refined data-driven model that seeks to minimize structural risk, thereby reducing high-limit errors compared to other machine learning methods that only focus on local training errors.

3.5. Model Development

The study aimed to evaluate the influence of different input data combinations on the total dissolved solids (TDS) using feature importance, a method that ranks input variables based on their significance in output uncertainty. This method is crucial in the development of predictive models. Sets of input combinations were selected based on the highest contribution of each parameter to TDS, determined by the order of importance using the random forest technique (see Table 2). The resulting input combinations at different steps are summarized in the table (see Table 3).

Table 2. Feature importance order of the used parameters

Input variables	EC	Na^+	Cl^-	SO_4^{2-}	Mg^{2+}	CO_3H	K^+	NO_3^-	Ca^{2+}
Value (%)	70.79	12.67	07.53	04.80	02.18	0.60	0.58	0.48	0.37
Rank	1	2	3	4	5	6	7	8	9

Table 3. Various input combinations used in the modeling

ML	Different input combinations
ML_{01}	EC
ML_{02}	EC and Na^+
ML_{03}	EC, Na^+, and Cl^-
ML_{04}	EC, Na^+, Cl^-, and SO_4^{2-}
ML_{05}	EC, Na^+, Cl^-, SO_4^{2-}, and Mg^+

3.6. Model Performance Evaluation

To evaluate and compare the models performance of the validation set. Two statistical criteria were used: Nash-Sutcliffe efficiency (NS) and The Root Mean Square Error (RMSE). These criteria were defined as follows:

The NS coefficient (NS) may be used to evaluate machine learning model's predicting capacity, where the higher NS value indicates that models are more accurate [21]. NS can accept values between $-\infty$ and 1. A score of one show full agreement, whereas a value of zero indicates that the model explains none of the initial variation

$$NS = 1 - \frac{\sum\left(TDS_{i_{pred}} - TDS_{i_{obs}}\right)^2}{\sum\left(TDS_{i_{obs}} - \overline{TDS}_{i_{obs}}\right)^2}. \quad (4)$$

The RMSE values represent the square root of the variance of the residual errors between the observed and predicted values. Hence; the accuracy of the model estimation is greater when the RMSE values are lower.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(TDS_{i_{pred}} - TDS_{i_{obs}}\right)^2}, \quad (5)$$

where $TDS_{i_{obs}}$ and $TDS_{i_{pred}}$ are the observed and predicted TDS; $\overline{TDS_{i_{obs}}}$ is the mean of observed TDS; N is the total number of observations.

4. RESULTS AND DISCUSSION

The TDS comprise the total dissolved inorganic salts and some amounts of organic substances that are dissolved in water, where it is influenced by many parameters. In the present study TDS is the target parameter, which is considered in the present work as one of the most important parameters for overall assessment of groundwater quality. In this paper Decision Tree, Random Forest, Support Victor Machine and Cat Boost machine learning models were used for forecasting of the TDS. Thus, researchers adopted broadly on use the ML models in various fields, where can be useful to predict future scenarios of nature and conservation processes. The Table 4 illustrates the results obtained by ML models to predict TDS. Each model was trained on the training dataset after selected the most influential input variable and was evaluated with the testing dataset; and the results were compared by means of NS and RMSE statistics (see Table 4).

According the Table 4, in the first application, five input combinations were evaluated using DT, Catboost, RF and SVR models to predict TDS in groundwater quality of the upper Chellif plain.

The performance results of the selected combinations varied between 0.65 and 0.94 in terms of Nash-Sutcliff, with superiority indicated using the support vector regress or depending on EC, Na^+, Cl^- & SO_4^{2-} variables with NS 0.94 and RMSE of 0.0328. Followed by RF model were using the same input combination with NS 0.93 and RMSE of 0.0359. The results using input combinations 3 depending on EC, Na^+, & Cl^- showed very good values with small inferiority to the latter results with NS 0.92, 0.92, and 0.91 and RMSE of 0.0374, 0.0387, and 0.0402 for the models RF, Catboost, and SVR, respectively. In regards to SVM model, it was achieving best performance, which had excellent accuracy with average NS (0.9455 to 0.9006) and RMSE (0.0328 to 0.0442).

The observed and modeled TDS variations in the validation phase are presented in figures bellow (see Fig. 2) for the best combinations in our ML models. As it could be notice, there is very goodness of fit

Table 4. Performance indices of the four machine learning models in the validation phase

Input combination	Machine learning models							
	DT		Catboost		RF		SVR	
	NS	RMSE	NS	RMSE	NS	RMSE	NS	RMSE
ML_{01}	0.6580	0.0869	0.6679	0.0809	0.7596	0.0688	0.9006	0.0442
ML_{02}	0.7960	0.0634	0.8327	0.0574	0.8076	0.0616	0.9017	0.0440
ML_{03}	0.8770	0.0492	0.9178	0.0402	0.9291	0.0374	0.9240	0.0387
ML_{04}	0.8677	0.0510	0.8914	0.0463	0.9346	0.0359	0.9455	0.0328
ML_{05}	0.8691	0.0508	0.9141	0.0411	0.9252	0.0384	0.9390	0.0347

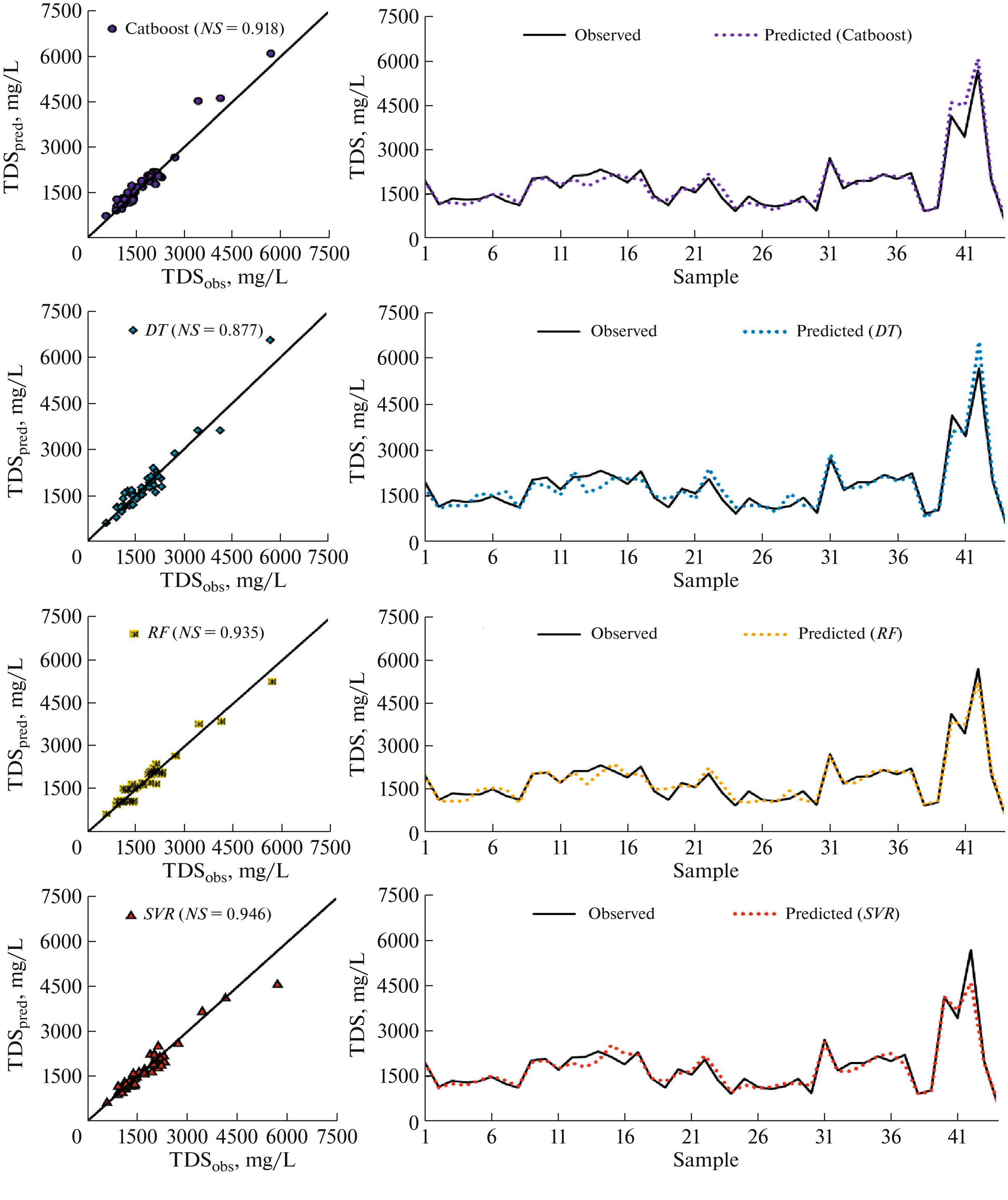

Fig. 2. Observed and predicted TDS of DT, Catboost, RF, and SVR algorithms of the best input combination.

between the observed and predicted TDS, and it is obviously observed that there is a linear relationship of the observed and predicted values presented in the scatterplots.

4.1. BRGM Analysis Reports of TDS

On the other hand; Based on BRGM analyzes report of TDS [22]; where it divided and classified the main elements according to their importance through their concentration in the water into four basic partials groups; and from it, we got the following results:

The TDS1 is the sum of the concentrations (mg/L) of the main major elements: HCO_3^-, NO_3^-, SO_4^{2-}, and Cl^-.

TDS2 is the sum of the major "secondary" elements: Ca^{2+}, Na^+, Mg^{2+}, K^+, and SiO_2. These elements are less frequently measured than those included in the calculation of TDS 1. They are also expressed in mg/L.

TDS3 is the sum of F, Fe, and Mn. These three elements are expressed in mg/L in the database

The TDS4 corresponds to the sum of trace elements expressed in pg/L in the Observatory and the most frequently measured: Cu, Pb, T.v, As, Cd, and Al.

The results of this study were compared to the BRGM report using the best performing combination (ML N°04), which combines the elements Na^+, SO_4^{2-}, and Cl^-. These elements are known to have high concentrations in TDS, with Na falling under TDS partial 2, and SO_4^{2-} and Cl^- falling under TDS partial 1. Therefore, this comparison supports the results obtained in our study.

CONCLUSION

Accurately forecasting water quality parameters is crucial for effective groundwater management, enabling the monitoring of pollution levels, population needs, and human activities. The present study aimed to investigate whether machine learning (ML) models, including Catboost, Random Forest (RF), Decision Tree (DT), and Support Vector Regression (SVR), are effective tools for forecasting the salinity parameter (TDS). This study also examined the sensitivity of these models to the output variable using hydro chemical and physicochemical groundwater parameters as inputs. Therefore, 190 groundwater samples of Upper Cheliffalluvial plain in northwestern Algeria were used for training and testing for applied the Catboost, RF, DT and SVR models. In consideration of the correlation coefficient of TDS to physicochemical parameters, the model's inputs were chosen based on the priority by using feature selection of the Random Forest algorithm of all parameters, and five sets of variable combinations were constructed as inputs to build and evaluate the models using these variables (Na^+, HCO_3^-, Ca^{2+}, Mg^{2+}, and SO_4^{2-}).

In overall, the inter-comparison of the results indicated that the SVR model had the best performance in forecasting TDS values among all the models with the best result was obtained by application of input combination ML4, followed by RF and Catboost models of the input combination 4 and 3. Although, the differences between the ML models were not high significant, but the SVR model was ultimately chosen as the finest model according to better conditions during the validation performance. In general, the performance results indicate that the SVR, RF, Catboost and DT, machine learning-based models are acceptable for groundwater evaluation and that can be considered as effective tools for predicting TDS values.

FUNDING

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

COMPETING INTERESTS

I declare that the authors have no competing interests as defined by Springer, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

AUTHOR'S CONTRIBUTIONS

PhD Student Tachi Amina contributed in design, analysis, interpretation of data, programming and modelling, writing manuscript; *Dr Hab Metaiche Mehdi* contributed in design, analysis, interpretation of data; *Dr Messoul Abd Errahmen* contributed in design; *Dr Bouguerra Hamza* contributed in design, figures preparation, proof reading manuscript; *Dr Hab Tachi Salah Edine* contributed in design, programming and modeling, writing and proof reading manuscript.

REFERENCES

1. A. E. Ercin and A. Y. Hoekstra, Environ. Int. **64**, 71–82 (2014). https://doi.org/10.1016/j.envint.2013.11.019
2. F. B. Banadkooki, M. Ehteram, A. N. Ahmed, C. M. Fai, H. A. Afan, W. M. Ridwam, and A. El-Shafie, Sustainability **11** (23), 6681 (2019). https://doi.org/10.3390/su11236681
3. F. B. Banadkooki, M. Ehteram, F. Panahi, S. S. Sammen, F. B. Othman, and E. S. Ahmed, J. Hydrol. **587**, 124989 (2020). https://doi.org/10.1016/j.jhydrol.2020.124989
4. F. B. Banadkooki, M. Ehteram, A. N. Ahmed, F. Y. Teo, C. M. Fai, H. A. Afan, and A. El-Shafie, Nat. Resour. Res. **29**, 3233–3252 (2020). https://doi.org/10.1007/s11053-020-09634-2
5. S. Zaman ZadGhavidel and M. Montaseri, Stochastic Environ. Res. Risk Assess., No. 8, 2101–2118 (2014). https://doi.org/10.1007/s00477-014-0899-y
6. *WHO*, *Guidelines for Drinking-Water Quality* (World Health Organization, 2004), Vol. 1.
7. T. M. Tung and Z. M. Yaseen, J. Hydrol. **585**, 124670 (2020). https://doi.org/10.1016/j.jhydrol.2020.124670

8. M. Jamei, I. Ahmadianfar, X. Chu, and Z. M. Yaseen, J. Hydrol. **589**, 125335 (2020). https://doi.org/10.1016/j.jhydrol.2020.125335

9. O. Kisi, N. Akbari, M. Sanatipour, A. Hashemi, K. Teimourzadeh, and J. Shiri, J. Environ. Inf. **22** (2), 92–101 (2013). https://doi.org/10.3808/jei.201300248

10. D. L. Sorensen, M. M. McCarthy, E. J. Middlebrooks, and D. B. Porcella, *Suspended and Dissolved Solids Effects on Freshwater Biota: a Review* (Environmental Research Lab., Office of Research and Development U.S. Environmental Protection Agency, Corvallis, OR, 1977).

11. J. D. Hem, *Study and Interpretation of the Chemical Characteristics of Natural Water* (Department of the Interior, US Geol. Surv., 1985), Vol. 2254.

12. A. M. Melesse, K. Khosravi, J. P. Tiefenbacher, S. Heddam, S. Kim, A. Mosavi, and B. T. Pham, Water **12** (10), 2951 (2020). https://doi.org/10.3390/w12102951

13. A. El Bilali, A. Taleb, and Y. Brouziyne, Agric. Water Manag. **245**, 106625 (2020). https://doi.org/10.1016/j.agwat.2020.106625

14. T. Thomas, A. P. Vijayaraghavan, and S. Emmanuel, *Machine Learning Approaches in Cyber Security Analytics* (Springer, Singapore, 2020), pp. 37–200. https://doi.org/10.1007/978-981-15-1706-8

15. L. Breiman, Mach. Learn. **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

16. G. Biau, J. Mach. Learn. Res. **13** (1), 1063–1095 (2012).

17. J. Wang, R. Zuo, and Y. Xiong, Nat. Resour. Res. **29**, 189–202 (2020). https://doi.org/10.1007/s11053-019-09510-8

18. L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, Adv. Neural Inf. Process. Syst. **31** (2018). https://arxiv.org/abs/1810.11363v1.

19. V. N. Vapnik, *The Nature of Statistical Learning Theory* (Springer, New York, 1995). https://doi.org/10.1007/978-1-4757-2440-0

20. S. Samantaray, O. Tripathy, A. Sahoo, and D. K. Ghose, in *Proc. 3rd Int. Conf. on Smart Computing and Informatics* (Springer Singapore, 2020), Vol. 1, pp. 767–774. https://doi.org/10.1007/978-981-13-9282-5_74

21. S. Maroufpoor, A. Fakheri-Fard, and J. Shiri, J. Hydraul. Eng. **25** (2), 232–238 (2019). https://doi.org/10.1080/09715010.2017.1408036

22. *BRGM Elaboration de Tests de Validation des Donnees Analytiques de l'Observation National de la Qualite des Eaux Souterraines (ONQES). Rap. BRGM.* (1998), R 40009.