American Journal of Water Resources, 2025, Vol. 13, No. 3, 86-96
Available online at http://pubs.sciepub.com/ajwr/13/3/3
Published by Science and Education Publishing
DOI:10.12691/ajwr-13-3-3

Predictive Modelling of Groundwater Quality in the Nakanbé River Basin Using Machine Learning Techniques

Issoufou OUEDRAOGO[1,2,*], W. J. P. SANDWIDI[1,2], Fatoumata KABORE[3], Mahamadou KONARE[4], Cheick Abdramane OUATTARA[3]

[1]Mining Engineering Department, University Yembila-Abdoulaye-TOGUYENI, BP 54 Fada N' Gourma, Burkina Faso
[2]Geosciences and Environment Laboratory (LaGE). Joseph KI-ZERBO University, BP 7021, Ouagadougou, Burkina Faso
[3]Direction de la Qualité des Eaux/Ministère de l'Environnement de l'Eau et de l'Assainissement, Burkina Faso
[4]Laboratoire Sciences et Technologie (LaST), Université Thomas SANKARA, 12 BP 417 Ouagadougou 12, Burkina Faso
*Corresponding author: ouedraogo.issoufou03@gmail.com

Received July 18, 2025; Revised August 20, 2025; Accepted August 28, 2025

Abstract Continuous monitoring of groundwater quality is essential for protecting public health and the environment, particularly in vulnerable regions such as Burkina Faso's Nakanbé Basin, where groundwater serves as a primary source of potable water. This study aimed to develop and evaluate machine learning (ML) models to predict two key water quality parameters: Total Dissolved Solids (TDS) and Total Alkalinity (TA), using data provided by the General Directorate of Water Resources (DGRE). A total of 1,765 groundwater samples were analyzed, encompassing nineteen physicochemical parameters. Prior to modelling, multicollinearity analysis was conducted to ensure the reliability of the input variables. Three regression algorithms Random Forest Regression (RFR), Multiple Linear Regression (MLR), and Decision Tree Regression (DTR) were compared for their predictive performance. Among them, Random Forest demonstrated the highest accuracy, with the highest R^2 and lowest error metrics (MAE, RMSE) across both training and testing datasets for both TDS and TA. While MLR offered consistent and interpretable results, particularly for TA, DTR exhibited strong overfitting, with lower generalizability on test data. The results highlight the superiority of ensemble learning approaches, particularly RFR, in capturing complex, nonlinear relationships within groundwater quality datasets. ML application in this context provides a cost-effective and scalable alternative to conventional laboratory-based monitoring methods. It also enables the identification of influential water quality parameters, supports risk assessment of contamination, and contributes to evidence-based water resource management strategies. These findings demonstrate the potential of ML tools to enhance groundwater monitoring and advance sustainable water governance in arid and semi-arid regions.

Keywords: *Machine Learning, Regression, Groundwater Quality Parameters, Nakanbé basin, Burkina Faso*

Cite This Article: Issoufou OUEDRAOGO, W. J. P. SANDWIDI, Fatoumata KABORE, Mahamadou KONARE, and Cheick Abdramane OUATTARA, "Predictive Modelling of Groundwater Quality in the Nakanbé River Basin Using Machine Learning Techniques." *American Journal of Water Resources*, vol. 13, no. 3 (2025): 86-96. doi: 10.12691/ajwr-13-3-3.

1. Introduction

Groundwater is a vital resource supporting a wide range of human activities, including irrigation, drinking water supply and industrial operations. Ensuring its quality is essential for protecting public health, sustaining agricultural productivity and supporting economic development.

The Sahel region, located just south of the Sahara Desert and characterized by a semi-arid climate, confronts major water resource challenges due to fluctuating climatic conditions, expanding populations, and rising water demand. Groundwater, which plays a vital role in sustaining communities and ecosystems, is under growing pressure from both excessive extraction and pollution risks [1,2]. Contamination presents a worldwide risk with profound impacts on both human health and the stability of natural environments [3]. This degradation often results from human activities including industrial discharge and agricultural practices, which contribute to the leaching of organic matter, pesticides and nitrates into aquifers [4].

In Africa, these issues are particularly urgent. For example, in Burkina Faso Nakanbé Basin of about 44,000 km², groundwater quality concerns are compounded by limited monitoring infrastructures putting additional strain on public health systems and local economies. The region, overseen by the Nakanbé Water Agency, faces numerous challenges such as declining rainfall, water governance challenge, demographic pressures, pollution and geological constraints that limit groundwater availability. High evaporation rates also threaten reservoirs while land degradation and sedimentation adversely affect river systems. Addressing these complex issues requires sustainable management practices and adaptive strategies to cope with the impacts of climate change. Moreover, water quality remains a problem in the Nakanbé region, as shown by several previous studies [5,6,7,8]. There is almost no continuous monitoring of water quality, due to the cost of analyses and the timely availability of reagents.

The assessment of groundwater suitability for consumption and other uses often relies on analysing essential physicochemical parameters such as pH, electrical conductivity (EC), total hardness (TH), and concentrations of major ions including calcium (Ca^{2+}), magnesium (Mg^{2+}), sodium (Na^+), potassium (K^+), sulphate (SO_4^{2-}), chloride (Cl^-), bicarbonate (HCO_3^-), nitrate (NO_3^-), phosphate (PO_4^{3-}) and fluoride (F^-). These indicators have been widely used in groundwater quality studies [9,10,11,12]. Nonetheless, collecting and analysing this comprehensive set of parameters can be financially burdensome and time-consuming, especially in regions with limited resources [13].

Evaluating the quality of groundwater remains a key step in determining its suitability for various applications, including drinking, domestic use, irrigation, and industry. This process typically involves comparing water quality data against national or international standards [14].

Traditionally, methods such as hydrochemical analysis, multivariate statistical techniques and Geographic Information Systems (GIS) have been employed [15,16,17,18,19,20,21]. However, these conventional approaches often rely on manual data processing and basic statistics, limiting their effectiveness in analysing large, multidimensional datasets.

In recent years, Machine Learning (ML) has emerged as a promising tool for groundwater quality assessment, particularly in contexts of limited data availability [22,23,24,25,26,27]. ML models can process diverse range of input variables like hydrological, geological, chemical, biological and physical characteristics [28,29,30,31,32,33] to generate accurate and robust predictions.

The integration of ML with GIS further enhances spatial analysis and risk mapping of groundwater contamination, supporting informed policy-making and sustainable water management [34]. Despite its potential, the use of ML in groundwater studies is still challenging for high-quality data unavailability, infrastructure limitations and the need for interpretable models to support decision-making.

This research focuses on applying Machine Learning techniques to forecast groundwater quality within the Nakanbé Basin. The main goals include:

i Identification of the principal groundwater quality indicators that are specific to the Nakanbé basin and evaluation of their importance in determining water usability;
ii Development and evaluation of ML models that can deliver accurate and reliable predictions of groundwater quality;
iii Comparison of various ML algorithms performance to identify the most effective approach for the basin's specific environmental and data conditions.

2. Study area and Methodology

2.1. Description of the Study Area

Burkina Faso, a landlocked country in the Sahel region of West Africa, is mainly drained by four major watersheds: the Niger, Nakanbé, Mouhoun, and Comoé basins [35]. The Nakanbé basin, situated in the upper Volta basin, covers approximately 44,000 km² about 15% of the national territory extending centrally from north to south between latitudes 10°9′N and 14°1′N and longitudes 2°5′W and 0°1′E [36]. The climate is predominantly Sudano-Sahelian, characterized by a seasonal rainfall pattern lasting from June-July to September-October, with a strong precipitation gradient from the wetter south (about 1200 mm annually) to the drier north (around 400 mm) [37]. The basin includes three climatic zones of Sahelian (300 - 600 mm/year), Sub-Sahelian (600 - 900 mm/year), and North Sudanese (900 - 1200 mm/year) [38,39]. The Nakanbé river, a major tributary of the transboundary Volta River system flowing through Benin, Burkina Faso, Côte d'Ivoire, Ghana, Mali, and Togo, drains this densely populated basin, which holds over half of Burkina Faso reservoirs [36]. Topographically, the terrain is mostly flat, ranging in elevation from 259 to 526 meters, with a crystalline basement geology made of hard, impermeable magmatic, metamorphic, and sedimentary rocks that limits groundwater productivity [35]. Consequently, both surface and groundwater resources are highly sensitive to climate variability [35]. Several key reservoirs lie upstream of the Wayen gauging station, including Lake Bam (41 million m³), the Douroudam (100 million m³) and the Ziga dam (200 million m³), constructed in 1995 and 2000 respectively [40,41]. The Ziga dam is particularly important, supplying nearly 70% of Ouagadougou's drinking water [40,41], while the Douroudam mainly supports agriculture. Groundwater is vital for agriculture, livestock, and drinking water especially in arid and semi-arid parts of the basin where surface water is seasonal [42]. In the northern and eastern areas, groundwater depths range from 20 to 25 meters, with recharge heavily dependent on rainfall [38]. However, the basin faces degradation caused by climate change and human activities such as agriculture expansion, deforestation, and overgrazing, which threaten water resources [37]. Land use and land cover in the basin changed significantly between 1972 and 1992 but have remained relatively stable since then [43].

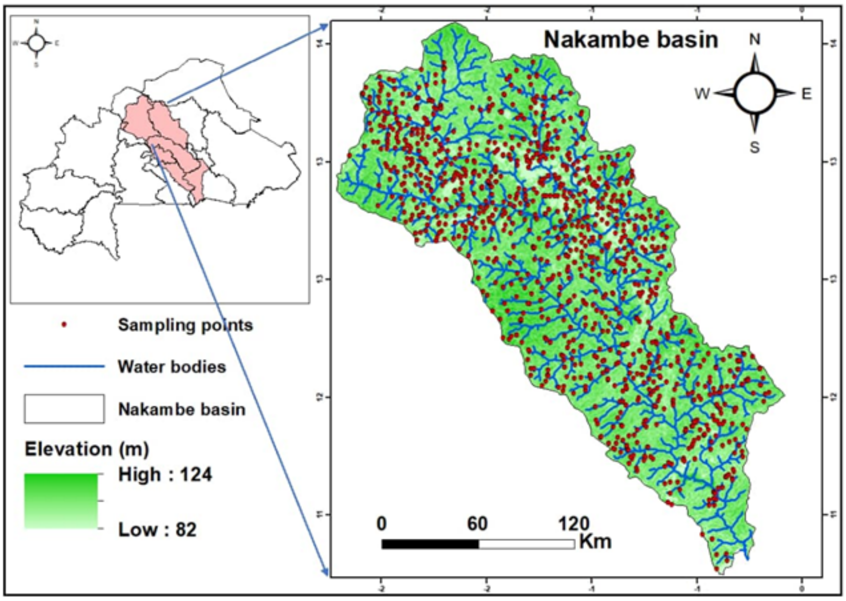


Figure 1. Location of Nakanbé river basin in Burkina Faso

2.2. Data Acquisition and Processing

A database of groundwater quality parameters has been built. The dataset included quality results from 1765 boreholes in the basin, collected in 2020 from the *Direction Générale des Ressources en Eau* (DGRE); Burkina Faso Water Resources' Directorate. The parameters used in the dataset included Temperature (Temp), pH, Electrical Conductivity (EC), Turbidity, Total Dissolved Solids (TDS), Total Alkalinity (TA), Calcium (Ca), Magnesium (Mg), Sodium (Na), Potassium (K), Ammonium (NH_4), Bicarbonates (HCO_3), Chloride (Cl), Sulphate (SO_4^{2-}), Nitrite (NO_2^-), Nitrate (NO_3^-), Phosphate (PO_4^{3-}). Some trace elements of Fluoride (F) and Iron (Fe) were also part of the dataset.

The dataset undergone thorough preprocessing to enhance its accuracy and consistency. This included steps such as data cleaning, organization, removal of duplicates, correction of errors, validation of records, and imputation of missing values. The cleaned and finalized dataset was then compiled using Excel 2016 and prepared for further analysis and modelling. The modelling process utilized statistical and programming environments, specifically R (version 4.2.1) and Python (accessed via Spyder), to implement three distinct ML algorithms.

For building predictive models of groundwater quality in the Nakanbé Basin, these software tools were applied to the processed data. Additionally, the relationships between variables were examined through correlation analysis conducted with the "corrplot" package in R, applying the Pearson correlation coefficient method.

Considering the importance of TA and TDS in groundwater quality assessment, several predictive models have been developed using other qualitative indicators as demonstrated by Mohebbi and Mohebbi [44] in the Golpayegan Plain in Iran, where clustering, principal component analysis (PCA), and factor analysis were applied for hydrochemical modelling; by Zarajabad et al. [45], who evaluated the performance of machine learning models including Decision Tree (DT), Random Forest (RF), and Multiple Linear Regression (MLR) to predict both TA and TDS in the Aran VaBidgol plain; and by Amini et al. [46], who proposed a hybrid approach combining wavelet analysis, trend-seasonal decomposition, and machine learning algorithms (including Random Forest and XGBoost) to enhance TDS prediction accuracy in the Karkheh River basin in Iran.

➢ **Total Alkalinity (TA)**

Total alkalinity is a key parameter for assessing groundwater quality as it represents the capacity of water to neutralize acids. Water TA is primarily caused by the presence of dissolved bicarbonates (HCO_3^-), carbonates (CO_3^{2-}), and hydroxides (OH^-) as well as other alkaline substances such as hydroxides of calcium, magnesium, and sodium. Alkalinity plays a crucial role in maintaining the pH balance of water and preventing drastic pH fluctuations that could harm aquatic life. High levels of alkalinity can buffer the water against acidification, whereas low alkalinity can make the water more susceptible to pH changes. Total alkalinity is often measured in milligrams per liter (mg/L) as $CaCO_3$. High alkalinity levels can indicate that the water has a strong buffering capacity, which is important for the suitability of water for drinking, agricultural, and industrial uses [47].

➢ **Total Dissolved Solids (TDS)**

TDS is an important parameter for assessing groundwater quality. The term refers to the inorganic salts and small amounts of organic matter dissolved in water. The principal constituents of TDS are typically calcium, magnesium, sodium and potassium cations, as well as carbonate, bicarbonate, chloride, sulphate and nitrate anions. High TDS levels generally indicate hard water and can lead to issues such as scale buildup in pipes, reduced efficiency of water filters and hot water heaters, and aesthetic problems like a bitter or salty taste [47,48].

2.3. Machine Learning (ML)Techniques

2.3.1. Models Description

Three ML models applied to the dataset were Decision Tree Regression (DTR), Random Forest Regression (RFR) and Multiple Linear Regression (MLR). Before the models were applied, data were first scaled and then split into 80% training and 20% testing datasets. The following describe the three ML models applied to the dataset.

➢ **Decision Tree Regression (DTR)**

Decision Tree Regression is a supervised learning method that splits data into branches to make predictions. Known for its simplicity and interpretability, DTR is also suitable for both classification and regression tasks [49]. In this study, the model was implemented using Python's Scikit-learn library. The dataset was scaled and divided into training and testing sets (80:20) before training. Decision Trees can produce accurate continuous predictions; however, they are often prone to overfitting, a limitation that can be mitigated through pruning techniques [50] or by employing ensemble methods such as Random Forests [51].

➢ **Random Forest Regression (RFR)**

Random Forest is a tree-based algorithm used for both classification and regression. It combines the results of multiple decision trees to improve accuracy and reduce overfitting [52]. In this study, the data was scaled and split (80% training, 20% testing), then the model was trained using Scikit-learn'sRandomForestRegressor. Despite being more complex and time-consuming, Random Forest effectively identified key features.

➢ **Multiple Linear Regression (MLR or LR)**

Linear regression is a fundamental predictive technique used to estimate the relationship between independent and dependent variables [53]. Simple linear regression involves one predictor, while multiple linear regression (MLR) uses two or more. In this study, the dataset was scaled and split (80% training, 20% testing) before applying MLR using Scikit-learn.

Given that regression models require independence among predictor variables, multi-collinearity statistics were used to analyse the matrix from Spyder python 3.12.

2.3.2. Evaluation performance of ML models

The methods discussed below were used to evaluate and compare the performance of all the models. A common and simple approach in evaluating models is to regress predicted values against observed values (or vice versa) and compare the slope and intercept parameters against the 1:1 line. The coefficient of correlation (R), root mean square error (RMSE), mean absolute error (MAE), and Mean Squared Error (MSE) are commonly used to assess the performance of ML models during training and testing. These statistical indicators can be defined by the following equations.

➢ **Coefficient of Correlation (R)**

The coefficient of correlation, often called Pearson's R, measures the strength and direction of the linear relationship between predicted values and observed values. It ranges from -1 to +1.

$$R = \frac{\sum_{i=1}^{n}\left(X_{obs,i} - \bar{X}_{obs}\right)\left(X_{model,i} - \bar{X}_{model}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_{obs,i} - \bar{X}_{obs}\right)^2} \cdot \sqrt{\sum_{i=1}^{n}\left(X_{model,i} - \bar{X}_{model}\right)^2}}$$

Where:

- Xobs,i: the observed value at index *i*
- Xmodel,i: the modeled/predicted value at index *i*
- Xobs: mean of observed values
- Xmodel: mean of modeled values
- n: number of observations

➢ **Mean Squared Error (MSE)**

The Mean Squared Error measures the average of the squares of the errors between predicted and observed values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(X_{obs,i} - X_{model,i}\right)^2$$

➢ **Root Mean Square Error (RMSE)**

The Root Mean Square Error is the square root of the MSE and indicates how much the predictions typically deviate from the observed values [54].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(\mathrm{X}_{\mathrm{obs,i}} - \mathrm{X}_{\mathrm{model,i}}\right)^2}{\mathrm{n}}}$$

➢ **Mean Absolute Error (MAE)**

According to Idri et al. [55], The Mean Absolute Error is the average of the absolute differences between predicted and observed values. It provides a direct interpretation of the prediction error.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\left(Actual\ value_i - Prediction\ value_i\right)\right| \Big/ n$$

Using these values, the efficiency of the three models in predicting the response variables was evaluated. A higher R^2 and a lower RMSE indicate a better model performance.

3. Results and Discussion

3.1. Groundwater Descriptive Statistics

Descriptive statistics provide a summary of the central tendency, dispersion, and shape of the distribution of the groundwater dataset. For each water quality parameter, the statistics summarized in Table 1 were calculated.

The descriptive statistics indicate considerable variability in groundwater quality parameters. Turbidity and electrical conductivity (EC) exhibit notably high standard deviations, indicating possible localized pollution sources. While most parameters comply with WHO standards, elements such as turbidity, iron (Fe), potassium (K), and fluoride (F) exceed recommended limits in some samples. The average pH values generally fall within acceptable ranges, but nitrate (NO_3^-) concentrations in certain cases approach or surpass safety thresholds, pointing potential human-induced contamination.

Table 1. Descriptive statistics for groundwater quality parameters along with WHO[1] standard

Parameter	Mean	Std_Dev	Variance	Min	Median	Max	WHO
Temp (°C)	31.215	1.58	2.496	7.24	31.2	38.8	-
pH (-)	6.879	0.479	0.229	3.26	6.93	10.93	6.5-8.5
EC (µS/cm)	385.953	241.482	58313.623	32	339	4150	1000
Turbidity	13.446	353.447	124925.113	0	0.89	14835	5
TDS (mg/l)	365.483	234.028	54769.195	26	315	4150	1000
TA (mg/l)	20.141	11.13	123.877	0	18.1	162.9	-
Ca (mg/l)	33.078	25.801	665.675	1.1	27.1	554.4	200
Mg (mg/l)	25.08	17.847	318.503	0	20.8	130	50
Na (mg/l)	12.536	13.589	184.666	0.1	7.9	106.8	200
K (mg/l)	2.694	3.098	9.597	0	1.9	70.8	12
Fe (mg/l)	0.311	0.727	0.529	0	0.04	8.9	0.3
NH_4 (mg/l)	0.087	0.191	0.037	0	0.03	2.34	1.5
HCO_3 (mg/l)	245.834	135.655	18402.225	17.5	220.8	1987.3	250
Cl (mg/l)	1.217	1.055	1.113	0.02	0.97	18.64	250
SO_2 (mg/l)	7.533	19.551	382.261	0	2	330	-
NO_2 (mg/l)	0.029	0.069	0.005	0	0.01	1.092	0.2
NO_3 (mg/l)	9.703	16.267	264.616	0	3.96	183.04	50
PO_4 (mg/l)	0.732	0.609	0.371	0.05	0.67	14.41	-
F (mg/l)	0.352	0.41	0.168	0	0.26	6	1.5

[1]WHO: World Health Organisation

Table 2. Correlation matrix analysis

	Temp	**pH**	**EC**	**Turbidity**	**TDS**	**TA**	**Ca**	**Mg**	**Na**	**K**	**Fe**	**NH_4**	**HCO_3**	**Cl**	**SO_2**	**NO_2**	**NO_3**	**PO_4**	**F**
Temp	1																		
pH	0.027	1																	
EC	0.007	0.403	1																
Turbidity	0.002	-0.023	-0.025	1															
TDS	-0.028	0.390	0.927	-0.022	1														
TA	-0.012	0.410	0.824	-0.016	0.827	1													
Ca	-0.038	0.235	0.693	-0.007	0.708	0.727	1												
Mg	0.009	0.318	0.653	-0.015	0.653	0.837	0.333	1											
Na	0.098	0.244	0.334	-0.018	0.320	0.372	0.132	0.225	1										
K	-0.093	0.102	0.228	-0.005	0.232	0.272	0.146	0.195	0.283	1									
Fe	-0.072	-0.085	-0.077	0.135	-0.059	-0.066	-0.042	-0.078	-0.082	0.081	1								
NH_4	-0.091	-0.055	0.114	0.080	0.128	0.106	0.168	0.027	-0.043	0.273	0.551	1							
HCO_3	-0.011	0.409	0.824	-0.016	0.827	0.999	0.728	0.837	0.372	0.272	-0.066	0.106	1						
Cl	0.034	0.035	0.436	-0.014	0.430	0.349	0.422	0.228	0.312	0.079	-0.045	0.076	0.350	1					
SO_2	0.007	0.075	0.310	-0.008	0.310	0.261	0.354	0.266	0.203	0.105	-0.043	0.056	0.261	0.140	1				
NO_2	-0.013	0.036	0.132	-0.009	0.135	0.112	0.114	0.088	0.069	0.064	0.162	0.072	0.112	0.104	0.021	1			
NO_3	0.052	-0.043	0.216	-0.014	0.239	0.139	0.203	0.140	0.290	0.053	-0.078	-0.066	0.140	0.489	0.028	0.232	1		
PO_4	-0.005	-0.136	-0.110	0.012	-0.116	-0.119	-0.077	-0.088	-0.093	-0.017	-0.035	-0.012	-0.120	-0.101	-0.023	-0.046	-0.108	1	
F	-0.019	0.029	0.088	-0.023	0.089	0.062	0.100	0.008	0.143	0.015	-0.063	-0.054	0.063	0.129	0.101	0.028	0.143	-0.057	1

3.2. Statistical Analysis and Modelling of Groundwater Quality Parameters

3.2.1. Correlation Analysis

Correlation analysis was conducted to identify significant relationships among the water quality parameters. The Pearson correlation coefficient was used to quantify the linear relationships between pairs of parameters [56]. In this study, nineteen (19) physicochemical parameters were analysed. The Pearson correlation matrix of the water samples is presented in Table 2, showing the correlation coefficients (R) between the parameters.

Several strong and moderate correlations were observed, revealing underlying geochemical interactions. Electrical Conductivity (EC) shows a very strong positive correlation with Total Dissolved Solids (TDS) (R = 0.93), as well as strong correlations with Total Alkalinity (TA) (R = 0.82) and Bicarbonate (HCO_3^-) (R = 0.82), indicating that EC is primarily influenced by dissolved salts and buffering ions in the water. Furthermore, TA and HCO_3^- are almost perfectly correlated (R = 0.999), confirming their shared chemical basis. TA also shows strong correlations with Magnesium (Mg^{2+}) (R = 0.84) and Calcium (Ca^{2+}) (R = 0.73), suggesting a common origin through natural mineral dissolution, such as carbonate weathering.

In contrast, Turbidity, Iron (Fe), and Phosphate (PO_4^{3-})

exhibit weak or negative correlations with most other parameters. For example, PO_4^{3-} shows slightly negative correlations with EC (R = -0.11), TDS (R = -0.12), and TA (R = -0.12), while Fe shows negligible or negative correlations across the matrix.

These results emphasize the influence of both natural hydrogeochemical processes and site-specific anthropogenic factors in determining groundwater quality in the study area [57,58,59].

To comprehensively interpret the dynamics of groundwater quality and the implications of correlation patterns among chemical constituents, it is essential to examine key hydrochemical indicators particularly Total Dissolved Solids (TDS) and alkalinity (TA). These parameters serve as fundamental proxies for understanding both natural mineralization processes and potential contamination sources within aquifer systems.

TDS reflect the cumulative concentration of dissolved inorganic and organic substances in water and are widely recognized as a primary indicator of the degree of mineralization. According to Sako and Kafando [57], variations in TDS provide critical insights into the intensity of water-rock interactions, especially in fractured crystalline basement aquifers typical of the Central Plateau of Burkina Faso. Elevated TDS levels are commonly associated with the dissolution of lithological components, such as carbonates and silicates, which contribute to the natural geochemical evolution of groundwater. However, such increases may also signal anthropogenic influences, including the infiltration of agricultural runoff, seepage from latrines, or leaching of domestic waste, which introduce additional ions into the groundwater system.

Similarly, alkalinity, dominated by bicarbonate ions (HCO_3^-), plays a vital role in characterizing the buffering capacity of groundwater and is considered a direct product of carbonate minerals weathering like calcite ($CaCO_3$) and dolomite ($MgCO_3$). As highlighted by Sako and Ouangaré [58], alkalinity acts as a robust geochemical tracer of carbonate dissolution, particularly within semi-confined or weathered basement aquifers. It provides valuable evidence for differentiating between endogenous geogenic processes and exogenous chemical inputs, thereby aiding in the identification of localized contamination sources.

Together, TDS and TA form a coherent interpretative framework for groundwater quality assessment, enabling the distinction between natural geochemical baselines and human-induced alterations in hydrochemical regimes.

3.2.2. Multi-collinearity assessment of parameters

We conducted a multicollinearity analysis to assess the linear relationships among nineteen hydrochemical variables. This was done using the Variance Inflation Factor (VIF) and Tolerance statistics, as presented in the table below. According to established thresholds, multicollinearity is considered problematic when the VIF exceeds 10.0 or the Tolerance falls below 0.1 [60]. In our analysis, all Tolerance values were greater than 0.1 and all VIF values were below 10.0, except for Ca, Mg, and HCO_3^-, which exhibited high multicollinearity. Therefore, most selected variables were deemed appropriate for inclusion in the modelling process.

Table 3. Multi-collinearity values for several explanatory factors

Variable	VIF	Tolerance
Temp	1.037	0.964291
pH	1.354	0.738513
EC	3.731	0.268051
Turbidity	1.021	0.979285
Na	4.436	0.225417
K	1.283	0.779159
Fe	1.577	0.634123
NH_4	1.678	0.595975
Cl	1.762	0.567679
SO_2	2.794	0.357872
NO_2	1.113	0.898110
NO_3	2.309	0.433058
PO_4	1.054	0.948641
F	1.055	0.947445
Ca	23.343	0.042839
Mg	30.585	0.032696
HCO_3	69.070	0.014478

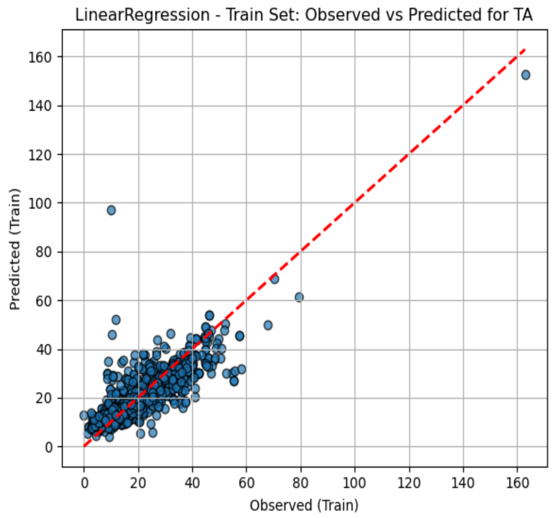
3.2.3. Predictive Modelling of Groundwater Quality Parameters

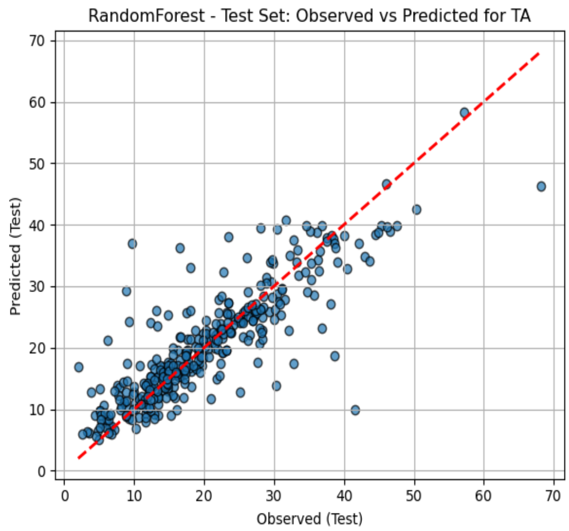
Modelling Total Alkalinity (TA) and Total Dissolved Solids (TDS) as key groundwater quality parameters are essential due to their critical role in assessing water chemistry and suitability for various uses. TA primarily reflects the buffering capacity of water, mostly controlled by bicarbonate ions, which regulate the pH and acid-base balance, providing important information on geochemical processes in the aquifer system. Meanwhile, TDS represents the total concentration of dissolved minerals and salts, influencing water taste, hardness, and overall usability for drinking, irrigation, or industrial purposes [61]. Both parameters are highly sensitive to natural processes such as rock weathering and water-rock interactions, as well as anthropogenic influences including agricultural runoff and wastewater discharge. Consequently, modelling TA and TDS help predict spatial and temporal variations in groundwater quality, enabling better management of water resources [62]. Additionally, these parameters are frequently measured in groundwater monitoring networks, making them practical for developing reliable predictive models. The complementary nature of TA and TDS with TA indicating buffering capacity and TDS measuring total solute load offers a comprehensive understanding of groundwater chemistry, which is crucial for sustainable exploitation and protection of water resources [61,62].

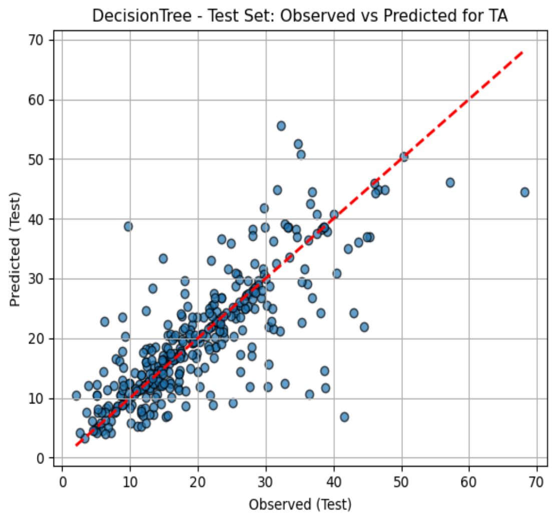
In this section, are presented the results of the predictive models developed for key groundwater quality parameters in the Nakanbé basin. The three multivariate regressions of MLR, DTR and RFR were used to predict these parameters. The dataset was divided into 80% for training and 20% for testing.

The performance of these models was evaluated using several statistical metrics including MAE, MSE, RMSE and R^2 for both training and testing phases as shown in Table 3. These types of regression models and performance indicators have also been widely used in previous studies for the prediction of groundwater quality variables. For instance, Zarajabad et al. [45] applied similar statistical approaches in their analysis. They found that RF outperformed the other models and MLR model was found with strong performance particularly when predicting TH.

The scatter plot in Figure 2 and Figure 3 shows respectively the observed versus predicted TA and TDS values using the three models.







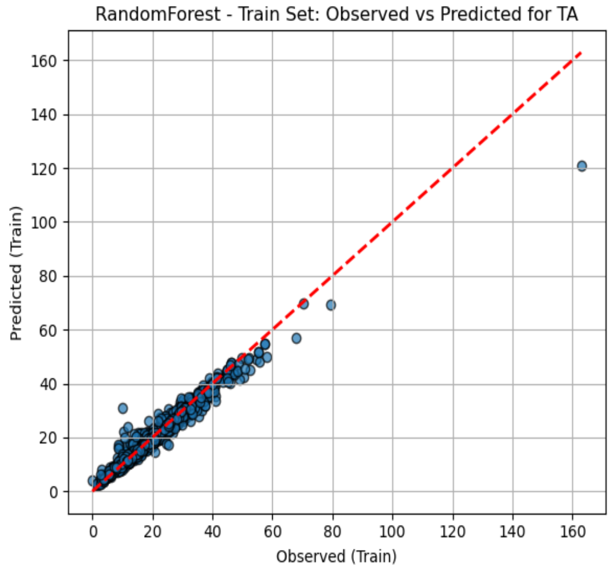
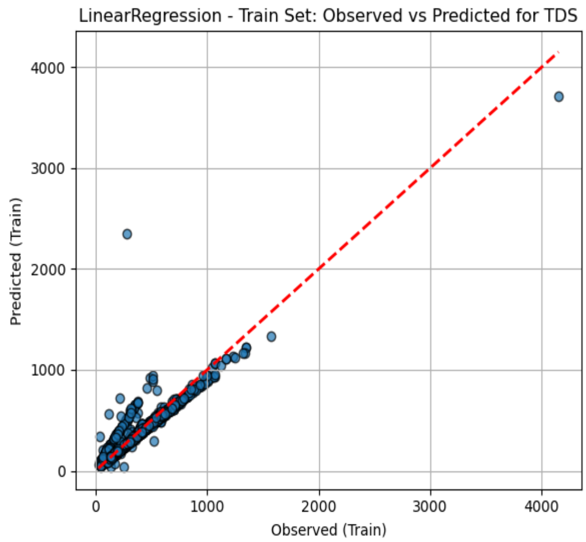
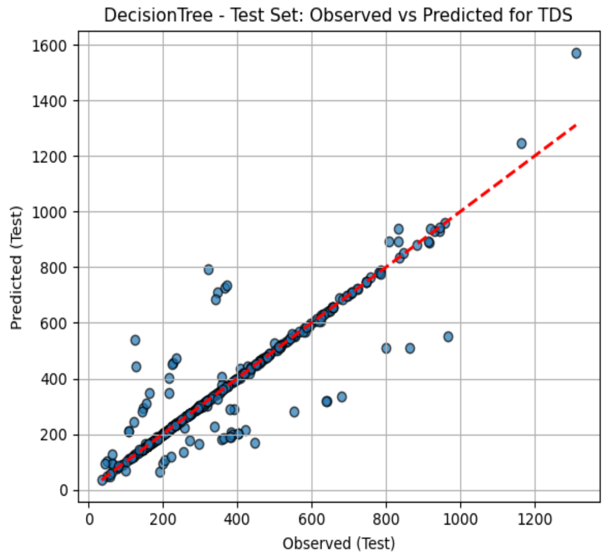


Figure 2. The performance of predictive models of the training and testing TA data





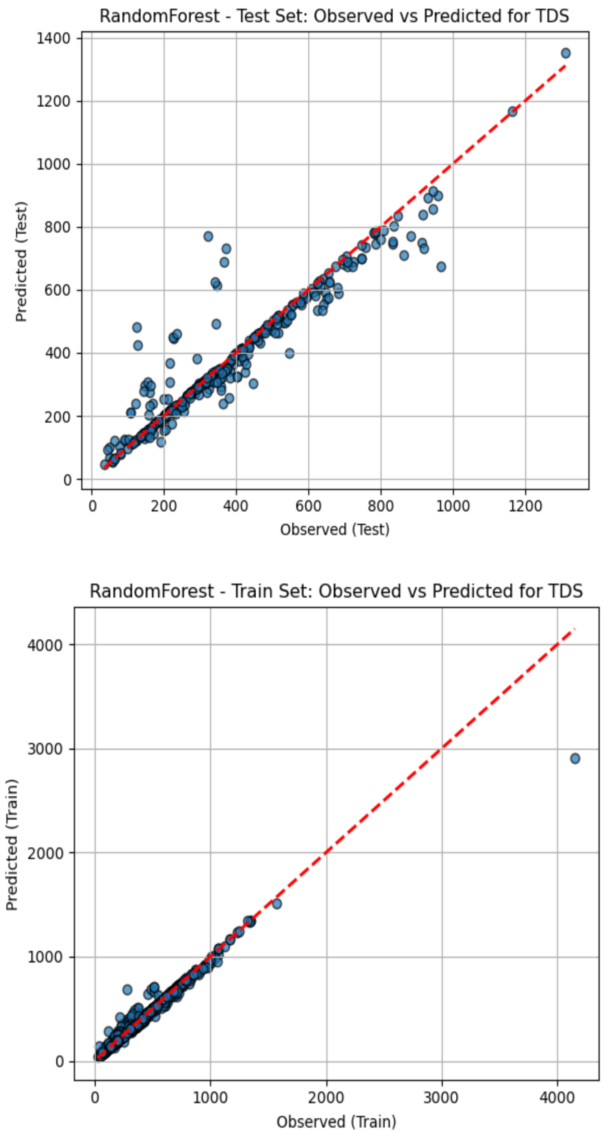


Figure 3. The performance of predictive models of the training and testing TDS data

3.2.4. Model Performance Evaluation and Comparison Analysis for TDS and TA

In this study, three models' algorithms of LR, DTR and RFR were used to predict TDS and TA concentrations. The predictive performance of each model was evaluated using the four-standard metrics of MAE, MSE, RMSE and R^2. The summarized results of the training and testing phases are in Table 4.

Among the evaluated models, RFR consistently exhibited superior predictive performance, robustness, and generalization capacity for TDS. On the testing dataset, RFR achieved the highest R^2 of 0.901 and the lowest RMSE of 68.14, indicating high-quality predictions with minimal deviation. These findings align with previous studies that highlighted the effectiveness and reliability of ensemble learning methods, particularly RFR, in modelling complex water quality parameters [45,63].

In contrast, DTR model, while attaining a perfect fit on the training data (R^2 = 1.000, RMSE = 0), demonstrated a substantial drop in performance on the test set (R^2 = 0.832, RMSE = 88.58), suggesting overfitting a limitation frequently reported in the literature for single-tree models [64]. The LR model also performed reasonably well (R^2 = 0.893, RMSE = 70.76), though with slightly high error metrics compared to RFR, confirming its adequacy for linear relationships but limited capacity to capture complex nonlinear interactions.

Table 4. The Performance of different models for estimation of TDS and TA

Target	Model	Set	MAE	MSE	RMSE	R^2
TDS	MLR	Training	41.749	8085.727	89.920	0.857
TDS	MLR	Testing	41.045	5007.464	70.763	0.892
TDS	DTR	Training	0.000	0.000	0.000	1.000
TDS	DTR	Testing	32.604	7845.736	88.576	0.832
TDS	RFR	Training	12.949	1852.016	43.035	0.967
TDS	RFR	Testing	31.544	4642.874	68.138	0.900
TA	MLR	Training	4.093	38.099	6.172	0.701
TA	MLR	Testing	3.956	32.257	5.679	0.700
TA	DTR	Training	0.000	0.000	0.000	1.000
TA	DTR	Testing	4.143	45.093	6.715	0.581
TA	RFR	Training	1.251	5.023	2.241	0.960
TA	RFR	Testing	3.410	28.442	5.333	0.736

Regarding TA, all models yielded acceptable results with differing predictive capacities. RFR once again outperformed the other models, delivering the best generalization on the test set (R^2 = 0.736, RMSE = 5.33), followed by LR (R^2 = 0.701, RMSE = 5.68). DTR displayed relatively weaker generalization (R^2 = 0.582, RMSE = 6.72), despite fitting the training data perfectly (R^2 = 1.000, RMSE ≈ 0). These outcomes highlight RFR's ability to model complex interactions even in smaller datasets, where linear and tree-based models may face limitations.

Overall, this comparative evaluation demonstrates that RFR is the most reliable and consistent model for predicting both TDS and TA, particularly in terms of minimizing error and enhancing predictive accuracy across both training and testing datasets. LR remains a valuable baseline model, especially for TA, while DTR should be applied cautiously due to its overfit tendency.

These results reinforce the suitability of ML techniques particularly ensemble models for water quality prediction tasks. From a sustainable development perspective, the deployment of such models offers a cost-effective and data-driven decision support tool for groundwater quality monitoring, supporting evidence-based water management, public health protection and long-term environmental planning in vulnerable regions such as the Nakanbé Basin in Burkina Faso.

4. Conclusion

Groundwater is a vital resource for drinking water, agriculture, ecosystem services and industries, particularly in regions like Burkina Faso Nakanbé river Basin where surface water availability is limited and highly variable. Ensuring the sustainable management of groundwater quality is therefore essential for public health, food security, environmental protection and long-term socio-economic development.

This study demonstrates the potential of machine learning techniques as efficient, scalable alternatives to traditional groundwater quality assessment methods. By evaluating and comparing three models Multiple Linear Regression, Decision Tree Regression and Random Forest Regression for the prediction of Total Dissolved Solids and Total Alkalinity, the analysis revealed that Random Forest Regression consistently outperformed the others, offering the best balance of accuracy, robustness and generalization across both training and testing datasets.

The strong predictive performance of Random Forest Regression underscores its utility not only in estimating key water quality parameters but also in identifying complex and nonlinear interactions between variables. This makes it a powerful decision-support tool for proactive groundwater management. In contrast, while Decision Tree Regression demonstrated excellent training performance, signs of overfitting were also exhibited. Multiple Linear Regression showed solid results, particularly for Total Alkalinity, but limited in capturing more complex relationships.

Going forward, refining machine learning models and integrating them with geospatial, hydrological and socio-environmental data could enhance the precision and applicability of predictive tools in groundwater management. Such integration can support a range of mitigation strategies, including early warning systems for groundwater degradation, targeted land-use regulation, and evidence-based policy design to reduce pollution risks and stabilize vulnerable zones.

Ultimately, the study reinforces the value of machine learning in advancing groundwater quality monitoring in data-scarce environments. The adoption of these models in the Nakanbé Basin study and beyond, can lead to more informed timely and cost-effective decision-making; therefore, contributing to globally achieve sustainable development goals in water resource management across arid and semi-arid regions.

Declaration of Competing Interest

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

ACKNOWLEDGEMENTS

The authors thank the Direction Générale des Ressources en Eau (DGRE) in Burkina Faso for the water samples' analysis and the anonymous reviewers for their helpful comments improving the paper.

References

[1] Xu, Y., and Usher, B. (2006). Groundwater pollution in Africa, Taylor & Francis/Balkema, The Netherlands. 353pp.

[2] MacDonald, A., M., R. Taylor, G. and H. Bonsor, C.,2013. (eds.) Groundwater in Africa - is there sufficient water to support the intensification of agriculture from "Land Grabs"." Hand book of land and water grabs in Africa. pp 376-383.

[3] Li, P., Karunanidhi, D., Subramani, T., & Srinivasamoorthy, K. (2021). Sources and consequences of groundwater contamination. Archives of environmental contamination and toxicology, 80(1), 1-10.

[4] Su, F., Wu, J. & He, S. (2019). Set pair analysis-Markov chain model for groundwater quality assessment and prediction: a case study of Xi'an city, China. Human and Ecological Risk Assessment.

[5] SDAGE - Schéma Directeur d'Aménagement et de Gestion de l'Eau de l'Agence de l'Eau du Nakanbé (2015). Tome I. Etat des lieux. Version validée par le Comité de Bassin Novembre 2015 https:// eaunakanbe.bf/ wp-content/uploads/2019/06/Etat-des-lieux-actualis%C3%A9-des-RE-du-Nakanb%C3%A9-_Tome-1-SDAGE.pdf.

[6] Millogo, D., Bazié, M.M., Koussoubé, Y., Zombré, P.N. and Da, E.C.D. (2018). Assessment of Agricultural and Mining Pollutions of Waterbodies within the Nakanbé Basin (Burkina Faso): The Case of the Goinré, Ziga and Bagré Reservoirs. Journal of Water Resource and Protection, 10, 41-58.

[7] Silga, R. P., Ouedraogo, I., Kabore, I., Sirima, D., Mano, K., Bance, V.& Gneme, A. (2022). Évaluation de la qualité des eaux de surface basée sur les paramètres physico-chimiques des eaux et les macroinvertébrés: cas du réservoir de Loumbila. Sciences Naturelles et Appliquées, 41(2 (1)), 57-73.

[8] Somé, K., Dembélé, Y., Somé, L. and Millogo-Rasolodimby, J. (2008). Pollution agricole des eaux dans le bassin du Nakanbé : Cas des réservoirs de Loumbila et de Mogtédo au Burkina Faso. Sud Sciences et Technologies. Semestriel, 16, 14-22.

[9] Vasanthavigar, M., Srinivasamoorthy, K., Rajiv Gantha, R., Vijayaraghavan, K., & Sarma, V. S. (2010). Characterization and quality assessment of groundwater with special emphasis on irrigation utility: Thirumanimuttar sub-basin, Tamil Nadu, India. Arab. Geosci J.

[10] Zahedi, S. (2017). Modification of expected conflicts between Drinking Water Quality Index and Irrigation Water Quality Index in water quality ranking of shared extraction wells using Multi Criteria Decision Making techniques. Ecological Indicators.

[11] Kumari, M. & Rai, S. C. 2020 Hydrogeochemical evaluation of groundwater quality for drinking and irrigation purposes using water quality index in Semi Arid Region of India. Journal of the Geological Society of India 95(2), 159-168.

[12] Maghrebi, M., Noori, R., Partani, S., Araghi, A., Barati, R., Farnoush, H. & Torabi Haghighi, A. 2021 Iran's groundwater hydrochemistry. Earth and Space Science 8(8), 1-18.

[13] Kumar, V. S., Amarender, B., Dhakate, R., Sankaran, S. & Raj Kumar, K. 2016 Assessment of groundwater quality for drinking and irrigation use in shallow hard rock aquifer of Pudunagaram, Palakkad District Kerala. Applied Water Science 6(2), 149-167.

[14] El-Rawy, M., Batelaan, O., Alshehri, F., Almadani, S., Ahmed, M.S., Elbeltagi, A. (2023). An Integrated GIS and Machine-Learning Technique for Groundwater Quality Assessment and Prediction in Southern Saudi Arabia. Water 15(13): 2448.

[15] Amiri, V, Nakagawa K (2021). Using a linear discriminant analysis (LDA)-based nomenclature system and self-organizing maps (SOM) for spatiotemporal assessment of groundwater quality in a coastal aquifer. J Hydrol 603: 127082.

[16] Elnazer, A.A, Salman, S.A, Mohamed, Y.M., Staford, J., Davies, P., El Nazer, H.A. (2023). Siwa oasis groundwater quality: factors controlling spatial and temporal changes. Environ Monit Assess 195(1): 61.

[17] Kaur, T., Bhardwaj R, Arora S (2017). Assessment of groundwater quality for drinking and irrigation purposes using hydrochemical studies in Malwa region, southwestern part of Punjab, India. Appl Water Sci 7: 3301-3316.

[18] Sargazi, S., Mokhtari, M., Ehrampoush, M.H., Almodaresi, S.A., Sargazi, H., Sarhadi, M. (2021). The application of geographical information system (GIS) approach for assessment of groundwater quality of Zahedan city, Sistan and Baluchestan Province, Iran. Groundw Sustainable Dev 12: 100509.

[19] Thakur D, Sharma A, Goel P, Thakur A, Raturi M (2023) Groundwater quality assessment in the alluvial region of upper yamuna basin, India. Groundw Sustainable Dev 22: 100969.

[20] Wang, X., Xiao, C., Yang, W., Liang, X., Zhang, L., Zhang, J. (2023). Analysis of the quality, source identifcation and apportionment of the groundwater in a typical arid and semi-arid region. J Hydrol 625: 130169.

[21] Kabore, F. (2024). Caractérisation de l'impact des seuils d'épandage sur la recharge des eaux souterraines dans un contexte de bas-fonds semi-aride : application au sous bassin de

Nariaré (Burkina Faso). Universite de Liège (Belgium). ProQuest Dissertations & Theses, 2024. 31855857.

[22] Sakizadeh, M. (2016). Artifcial intelligence for the prediction of water quality index in groundwater systems. Model Earth Syst Environ 2:1-9.

[23] Chaplot, B. (2021) Prediction of rainfall time series using soft computing techniques. Environ Monit Assess 193(11): 721.

[24] Latif, S.D., Hazrin NAB, Koo CH, Ng JL, Chaplot B, Huang YF, Ahmed, A.N., (2023). Assessing rainfall prediction models: exploring the advantages of machine learning and remote sensing approaches. Alexandria Eng J 82: 16-25.

[25] Fuladipanah, M., Shahhosseini A, Rathnayake N, Azamathulla HM, Rathnayake U, Meddage DPP, Tota-Maharaj K (2024). In-depth simulation of rainfall-runof relationships using machine learning methods. Water Pract Technol 19(6): 2442-2459.

[26] Kisi, O., Azamathulla, H.M., Cevat, F., Kulls, C., Kuhdaragh, M., Fuladipanah, M. (2024). Enhancing river fow predictions: comparative analysis of machine learning approaches in modeling stage-discharge relationship. Results Eng 22: 102017.

[27] Niazkar, M., Zakwan, M., Goodarzi, M.R., Hazi, M.A. (2024). Assessment of climate change impact on water resources using machine learning algorithms. J Water Clim Change 15(6): jwc2024002.

[28] Sahu, M., Mahapatra, S.S., Sahu, H.B., Patel, R.K. (2011). Prediction of water quality index using neuro fuzzy inference system. Water Qual Exposure Health 3: 175-191.

[29] Seyam, M., Mogheir, Y., (2011). Application of artifcial neural networks model as analytical tool for groundwater salinity. J Environ Prot 2(01): 56.

[30] Gholami, V., Khaleghi, M.R., Sebghati, M. (2017). A method of groundwater quality assessment based on fuzzy network-CANFIS and geographic information system (GIS). Appl Water Sci 7(7): 3633- 3647.

[31] Gholami, V., Sebghati M, Yousef Z., (2016). Integration of artifcial neural network and geographic information system applications in quality groundwater simulating 182-173: 43.

[32] Nathan, N.S, Saravanane, R., Sundararajan, T. (2017). Application of ANN and MLR models on groundwater quality using CWQI at Lawspet, Puducherry in India. J Geoscience Environ Prot 5(03): 99.

[33] Asadollah, S.B.H.S., Jodar-Abellan, A., Pardo, M.Á. (2024). Optimizing machine learning for agricultural productivity: a novel approach with RScv and remote sensing data over Europe. Agric Syst 218: 103955.

[34] Judeh, T., Almasri, M.N., Shadeed, S.M. et al. (2022). Use of GIS, Statistics and Machine Learning for Groundwater Quality Management: Application to Nitrate Contamination. Water Resour 49, 503-514 (2022).

[35] Gbohoui, Y. P., Paturel, J. E., Tazen, F., Mounirou, L. A., Yonaba, R., Karambiri, H., & Yacouba, H. (2021). Impacts of climate and environmental changes on water resources: A multi-scale study based on Nakanbé nested watersheds in West African Sahel. Journal of Hydrology: Regional Studies, 35, 100828.

[36] Ibrahim, B., (2012). Caractérisation des saisons de pluies au burkina faso dans un contexte de changement climatique et évaluation des impacts hydrologiques sur le bassin du nakanbé. PhD Thesis. 2iE et Université Pierre et Marie Curie, Ouagadougou et Paris VI.

[37] West, Thor, C., Benecky, S., Karlsson, C., Reiss, B., & Moody, A. J. (2020). Bottom-up perspectives on the re-greening of the Sahel: An evaluation of the spatial relationship between soil and water conservation (SWC) and tree-cover in Burkina Faso. *Land* 2020, *9*(6), 208.

[38] Ibrahim, B., Karambiri, H., & Polcher, J. (2015). Hydrological impacts of the changes in simulated rainfall fields on Nakanbe Basin in Burkina Faso. Climate, 3(3), 442-458.

[39] Karambiri, H.; García Galiano, S.; Giraldo, J.; Yacouba, H.; Ibrahim, B.; Barbier, B.; Polcher, J. (2011). Assessing the impact of climate variability and climate change on runoff in West Africa: The case of Senegal) and Nakambe River basins. Atmos. Sci. Lett. 2011, 12, 109-115.

[40] World Bank. (2001). Project Appraisal Document on a Proposed Credit in the Amount of SDR 55.5 Million (US$70 Million Equivalent) to Burkina Faso for a Ziga Water Supply Project. Report No: 21884-BUR. Retrieved from https://ocuments1.worldbank.org/curated/en/419491468743737560/pdf/multi0page.pdf.

[41] IW/LEARN. (2014). Visit at Ziga dam (Burkina Faso) https://fdmt.iwlearn.org/news/visit-at-ziga-dam-burkina-faso (Access 19 June 2025).

[42] Ciampi, L., Plumpton, H. J., Osbahr, H., Cornforth, R. J., & Petty, C. (2022). Building resilience through improving groundwater management for sustainable agricultural intensification in African Sahel. CABI Agriculture and Bioscience, 3(1), 63.

[43] Diello, P. (2007). Interrelations Climat - Homme - Environnement Dans Le Sahel Burkinabé : Impacts Sur Les Etats De Surface Et La Modélisation Hydrologique. Ph.D. Thesis. Montpellier 2-IRD/ 2iE, France/Burkina Faso.

[44] Mohebbi, T., G., & Mohebbi, T., A. (2020). Statistical approaches and hydrochemical modeling of groundwater in the Golpayegan Plain aquifer, Iran. *Modeling Earth Systems and Environment*, *6*(4), 2391-2404.

[45] Zarajabad, A. M., Hadi, M., Nodehi, R. N., Moradi, M., Ghalhari, M. R., Zeraatkar, A., & Mahvi, A. H. (2024). Providing predictive models for quality parameters of groundwater resources in arid areas of central Iran: A case study of kashan plain. Heliyon, 10(11).

[46] Amini, H., Fakheri, F., Dar, J. Y., Shakeri, R., Nejati, H., Lam, M. Y., ... & Ahmadian, R. (2025). TDS prediction with wavelet analysis and trend-seasonal decomposition and machine learning algorithms, Case Study: Karkheh River, Iran. *Water Conservation Science and Engineering*, *10*(2), 63.

[47] Jacintha, T. G. A., Rawat, K. S., Mishra, A., & Singh, S. K. (2017). Hydrogeochemical characterization of groundwater of peninsular Indian region using multivariate statistical techniques. Applied Water Science, 7, 3001-3013.

[48] WHO, (1993). Guidelines for Drinking Water Quality Recommendations. 2nd Edition, Vol. 1, World Health Organization, Geneva, Switzerland.

[49] Nasteski, V. (2017). An overview of the supervised machine learning methods. Horizons. b, 4(51-62), 56.

[50] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. Wadsworth.

[51] Breiman, L. Random Forests. Machine Learning 45, 5-32 (2001).

[52] IBM Cloud Education. (2025). What is Random Forest? IBM. https://www.ibm.com/cloud/learn/random-forest (Accessed 16 March 2025).

[53] Kanade, V. (2023). What Is Linear Regression? Types, Equation, Examples, and Best Practices for 2022. https:// www.spiceworks.com/ tech/artificial-intelligence/articles/what-is-linear-regression (Accessed online 27 Juanury 2025).

[54] Minku, L. L., & Yao, X. (2013). Ensembles and locality: Insight on improving software effort estimation. Elsevier, Information and Software Technology, 55(8), 1512-1528.

[55] Idri, A., azzahra Amazal, F., & Abran, A. (2015). Analogy-based software development effort estimation: A systematic mapping and review. Information and Software Technology, 58, 206-230.

[56] Nelyub, V. A., Tynchenko, V. S., Gantimurov, A. P., Degtyareva, K. V., & Kukartseva, O. I. (2023). Correlation Analysis and Predictive Factors for Building a Mathematical Model. In Proceedings of the Computational Methods in Systems and Software (pp. 14-25). Cham: Springer International Publishing.

[57] Sako, A., & Kafando, S. (2021). Hydrogeochemical and spatial assessment of groundwater quality from basement aquifers in the Central Plateau Region of Burkina Faso, West Africa. Environmental Earth Sciences, 80(9), 358.

[58] Sako, A., Ouangaré, C.A.C. (2023). Hydrogeochemical characterization and natural background level determination of selected inorganic substances in groundwater from a semi-confined aquifer in Midwestern Burkina Faso, West Africa. Environ Monit Assess 195, 519 (2023).

[59] Millogo, C., Sako, A., Sagnon, B. and Nakolendousse, S., (2024). Hydrochemical and Spatial Assessment of Groundwater Quality from Fractured Basement Aquifers in the South-Central of Burkina Faso, West Africa.

[60] Kutner, M. H. (2005). Applied linear statistical models. https://thuvienso.thanglong.edu.vn/handle/TLU/12233.

[61] Appelo, C. A. J., & Postma, D. (2005). Geochemistry, Groundwater and Pollution (2nd ed.). CRC Press.

[62] Hem, J. D. (1985). Study and Interpretation of the Chemical Characteristics of Natural Water (3rd ed.). U.S. Geological Survey Water-Supply Paper 2254.

[63] El Bilali, A., Taleb, A. (2020). Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment, Journal of the Saudi Society of Agricultural Sciences 19 (7), 439-451.

[64] Calvert, M.B. (2020). Predicting Concentrations of Selected Ions and Total Hardness in Groundwater Using Artificial Neural Networks and Multiple Linear Regression Models, Duke University.

© The Author(s) 2025. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).