Modelling of total dissolved solids in water supply systems using regression and supervised machine learning approaches

Anthony Ewusi¹ · Isaac Ahenkorah² · Derrick Aikins²

Received: 1 June 2020 / Accepted: 21 December 2020 / Published online: 14 January 2021
© The Author(s) 2021

Abstract

Monitoring of water quality through accurate predictions provides adequate information about water management. In the present study, three different modelling approaches: Gaussian process regression (GPR), backpropagation neural network (BPNN) and principal component regression (PCR) models were used to predict the total dissolved solids (TDS) as water quality indicator for the water quality management. The performance of each model was evaluated based on three different sets of inputs from groundwater (GW), surface water (SW) and drinking water (DW). The GPR, BPNN and PCR models used in this study gave an accurate prediction of the observed data (TDS) in GW, SW and DW, with the R^2 consistently greater than 0.850. The GPR model gave a better prediction of TDS concentration, with an average R^2 , MAE and RMSE of 0.987, 4.090 and 7.910, respectively. For the BPNN, an average R^2 , MAE and RMSE of 0.913, 9.720 and 19.137, respectively, were achieved, while the PCR gave an average R^2 , MAE and RMSE of 0.888, 11.327 and 25.032, respectively. The performance of each model was assessed using efficiency based indicators such as the Nash and Sutcliffe coefficient of efficiency (E_{NS}) and the index of agreement (d). The GPR, BPNN and PCR models, respectively, gave an E_{NS} of (0.967, 0.915, 0.874) and d of (0.992, 0.977, 0.965). It is understood from this study that advanced machine learning approaches (e.g. GPR and BPNN) are appropriate for the prediction of water quality indices and would be useful for future prediction and management of water quality parameters of various water supply systems in mining communities where artificial intelligence technology is yet to be fully explored.

Keywords Water quality · Total dissolved solids · Artificial neural network · Principal component regression

Introduction

Provision of safe and quality drinking water is a major concern in many developing countries due to rapid growth in urbanization and industrialization. An estimate of about 1.8 million people die every year, predominantly in developing countries, due to water-borne diseases and inadequate supply of quality water (Ishii and Sadowsky 2008; Corcoran

2010). To reduce the impact caused by water-related issues, water quality assessment based on water quality indices, drinking water standards and guidelines are used to evaluate the chemical, physical and biological constituents of water. Among other constituents, total dissolved solid (TDS) is one of the most vital constituents or parameters in assessing the overall suitability and quality of various water supply systems (Atta et al. 2018; Li et al. 2018; Pan et al. 2019). Therefore, accurate measurement and prediction of TDS may provide an indication of the salinity (total organic and inorganic dissolved substances) in various water resource systems.

Several models have been developed and applied for analysis and monitoring of water quality parameters (Ghosh et al. 2015; Sen et al. 2018; Adiat et al. 2020; Emami and Parsa 2020). Traditional (deterministic and stochastic) models, such as statistical approaches and visual modelling, have been commonly used in literature (Sun and Gui 2015; Tziritis and Lombardo 2017; Chen et al. 2018; Karami et al.

✉ Anthony Ewusi
aewusi@umat.edu.gh

Isaac Ahenkorah
isaac.ahenkorah@mymail.unisa.edu.au

Derrick Aikins
derrick.aikins@mymail.unisa.edu.au

¹ University of Mines and Technology, UMaT, Box 237, Tarkwa, Ghana

² University of South Australia, UniSA STEM, Adelaide, SA 5000, Australia

2018). Statistical-based water quality models, such as cluster analysis (CA), hierarchical cluster analysis (HCA) and principal component analysis (PCA), have been commonly used to classify and evaluate correlations between water constituents or parameters (Liu et al. 2011; Gu et al. 2016; Hamil et al. 2018; Lu and Ma 2020). However, data requirements for these approaches are enormous, difficult, time-consuming and expensive to obtain. Furthermore, many statistical models assume a linear relationship between response and prediction variables (parameters). Therefore, utilizing statistical approaches for nonlinear relations among variables is usually ineffective.

Multiple linear regression (MLR) and PCA models, despite their inefficiency on nonlinear relations among variables, have also been used in many hydrological studies possibly due to the easiness to use and interpret relationships between parameters (Adeloye 2009; Chenini and Khemiri 2009; Gholami et al. 2011; Viswanath et al. 2015; Lu and Ma 2020). For example, Viswanath et al. (2015) proposed a prediction model for TDS concentration with 10 other water quality parameters as a variable in watersheds by combining PCA with MLR. In their model, PCA was used to isolate less significant parameters, whereas MLR model was used to predict TDS in terms of other statistically significant parameters. However, the PCA prediction model used in their study utilized the entire dataset for model development with no validation. Alternatively, the principal component regression (PCR) technique, which combines PCA with MLR, has been developed and successfully applied in solid waste generation rate prediction (Azadi et al. 2016) and TDS prediction (Jacintha et al. 2017; Pan et al. 2019).

In addition to the classical statistical regression methods, supervised machine learning (SML) approaches such as artificial neural network (ANN), support vector machine (SVM) and adaptive neuro-fuzzy inference system (ANFIS) have been adopted in many hydrological studies (Suen and Eheart 2003; Asadollahfardi et al. 2012; Shamshirband et al. 2015; Alrashed et al. 2018; Yaseen et al. 2018; Sinshaw et al. 2019; Haghbin et al. 2020). These studies include forecasting nitrate concentration in rivers (Suen and Eheart 2003; Haghbin et al. 2020), modelling total phosphorus and total nitrogen in wetlands (Asadollahfardi et al. 2010), predicting TDS in rivers (Asadollahfardi et al. 2012), estimating the concentration of total nitrogen and total phosphorus in lakes (Sinshaw et al. 2019), analysing the thermal behaviour and performance of nano-suspensions in water supply systems (Shamshirband et al. 2015; Alrashed et al. 2018; Karimi-pour et al. 2019) and predicting water quality parameters including TDS, biochemical oxygen demand (BOD), and chemical oxygen demand (COD) using three new ensemble machine learning models (Asadollah et al. 2020; Sharafati et al. 2020).

In relation to the prediction of TDS, some recent studies have utilized various SML techniques. Asadollahfardi et al. (2012) developed ANN model was to predict TDS in the Talkheh Rud River (Iran). In their study, the Elman network, which includes the multilayer perceptron (MLP) and recurrent neural network (RNN), was developed and applied. The results from their study indicate that the Elman network predicts the TDS very close to the observed values ($R=0.964$). Schuttrumpf (2018) developed a recurrent neural network (RNN)-based model for predicting and forecasting TDS of a river. They observed that the RNN model gave an accurate prediction of the observed parameters. Pan et al. (2019) compared the efficiency of dual-step MLR, hybrid PCR and BPNN models in predicting TDS in monitoring wells. According to their study, hybrid PCR and dual-step MLR models provide better prediction compared to the BPNN model. Banadkooki et al. (2020) employed the ANFIS, SVMs and ANN models for prediction of TDS of aquifers in Yazd plain (Iran). The results from their study showed that the hybrid ANFIS had a better improved accuracy over the ANN and SVM models by 1.4% and 3.8%, respectively. They also observed that the SVM model had the least Nash–Sutcliffe efficiency value among all the models.

Despite lots of studies on water quality analysis using deterministic and stochastic approaches, statistical regression methods and SML models, studies on using SML, PCR and machine learning techniques such as the Gaussian process regression (GPR) are limited, especially in developing African countries. While the world is geared towards the fourth industrial revolution where artificial intelligence (AI), machine learning, augmented reality, internet of things (IOT), etc., is playing a major role, Africa is yet to come to the moment of realization. More importantly, developing countries in Africa such as Ghana are yet to adopt, apply and test the efficiency of machine learning in the water supply system management. This study will therefore create awareness for practitioners to appreciate the robustness of the methods used in this study. Also, to the best of the author's knowledge, existing models for predicting water quality parameters are usually developed using a single water resource (rivers or monitoring wells or lakes), which may affect the long-term application of the models in different water supply systems. Hence, a comprehensive study on predictive models is required for monitoring and predicting water quality parameters, especially in mining communities such as Tarkwa, Ghana. Therefore, the objectives of this study are to (i) develop hybrid predictive models using GPR, BPNN and PCR by utilizing non-overlapping testing dataset from groundwater (GW), surface water (SW) and drinking water (DW); (ii) predict TDS concentration in SW, GW and DW using the proposed predictive models developed in this study; and (iii) evaluate and compare the performance of the

models using series of performance evaluation metrics and statistical indices.

Methodology

Study area

Hydrogeological setting

A total of about 386 data points (GW = 189, SW = 110 and DW = 87), obtained between February to March 2015, were used in this study. The datasets were taken from Tarkwa, a mining (mainly gold and manganese) community and the capital of Tarkwa-Nsuaem Municipal Assembly, Western Region, Ghana. The area was selected for this study due to high-level pollution of water supply systems from mining activities (Bhattacharya et al. 2012; Ewusi et al. 2017a, b; Baah-Ennumh and Adom-Asamoah 2019). The area is located between latitudes $4^{\circ} 0' 0''$ N and $5^{\circ} 40' 0''$ N and longitudes $1^{\circ} 45' 0''$ W and $2^{\circ} 1' 0''$ W. Figure 1 shows the location and geological setting of the study area using QGIS.

The domestic and commercial water supply systems in the area mainly consist of GW (boreholes and hand-dug wells) and SW (streams and rivers). The majority of these water supply systems (GW and SW) serve as a source of DW for nearby communities. The average well depth in the area is 35.4 m. Borehole yields range between $0.4 \text{ m}^3/\text{h}$ and $18 \text{ m}^3/\text{h}$ with an average of $2.4 \text{ m}^3/\text{h}$ (Bhattacharya et al. 2012). The Bonsa, Huni and Ankobra Rivers and their tributaries are the main sources of recharge for nearby streams and GW (Bhattacharya et al. 2012). The quality of the water supply systems in the area is highly affected by mine contaminants and mining-related activities, leakage from underground storage tanks, improper waste disposal and agrochemicals from agricultural fields (Ewusi et al. 2017b).

Data description

A total of 10 parameters, which include arsenic (As), cadmium (Cd), mercury (Hg), copper (Cu), cyanide (CN), total dissolved solids (TDS), total suspended solids (TSS), pH, turbidity and electric conductivity (EC), were obtained from GW, SW and DW systems in the study area. These

Fig. 1 Location map of the study area

parameters were carefully chosen based on their data availability, significance and concentrations with respect to the WHO guideline values. Table 1 presents a statistical summary of all the parameters used in this study. TDS was selected as the target parameter for all modelling and analyses as its concentration is affected by many of the studied parameters due to the high pollution of water supply systems from mining activities in the area. The remaining 9 parameters were used as the input parameters to build the prediction model for TDS.

In this study, three modelling approaches (GPR, BPNN and PCR) were used to predict the concentration of TDS in GW, SW and DW systems. To reduce modelling errors and avoid possible bias from the inputs, all the datasets from GW, SW and DW were combined and were used for training (70%) and testing (30%). The entire datasets were then used to perform model validation. This approach was adopted from previous studies (Konaté et al. 2015a; Ziggah et al. 2016) to understand the possible variability in the dataset and to determine the extent at which the developed model can be generalized should the size of the data increase in the study area. Moreover, it indicates the model predictive capability across the entire data extent in the study area. After the training, testing and validation stage, the model is then used for prediction using 3 different sets of inputs from GW, SW and DW. As such, each model was evaluated 3 times with different datasets. This approach allows a fair and systematic assessment of the methods and modelling precision by applying consistent training and testing inputs in multiple trials. This also helps to identify and reveal possible bias from the input data. All modelling and analysis were performed using MATLAB (ver. R²020a).

Nonlinear regression analyses

Gaussian process regression

The Gaussian process regression is a powerful nonlinear prediction tool, which can be used for both supervised and unsupervised learning frameworks. It is a nonparametric stochastic process that generalizes the Gaussian probability distribution. A Gaussian process sometimes is described as a distribution over functions ($P(f)$), where f is a function that projects input space (vector X) to feature space (vector r) and for any finite subset of X , the marginal distribution over that subset $P(f)$ has a Gaussian distribution. The f could be an infinite-dimensional quantity. As a result, the Gaussian process extends multivariate Gaussian distributions to infinite dimensionality (Rasmussen and Williams 2006). One of the advantages of a Gaussian process model is that its formulation is probabilistic. This is useful for probabilistic prediction and also enables the model parameters inference for kernel shape and noise-level control (Chu

Table 1 Statistical summary of parameters and guideline values

Parameters	Unit	WHO limit	Surface water (SW)				Groundwater (GW)				Drinking water (DW)			
			Min	Max	Mean	Std dev	Min	Max	Mean	Std dev	Min	Max	Mean	Std dev
pH	pH units	6.5–8.5	4.2000	10.4700	7.2845	1.0279	4.4400	10.2900	6.7714	1.0084	4.9000	8.1800	6.4054	0.7626
TDS	mg/L	1000	11.1200	415.0000	110.7063	77.1992	8.8100	534.0000	88.4952	74.6134	9.8900	274.0000	67.7446	53.9078
TSS	mg/L	–	1.0000	899.0000	74.0273	159.5645	1.0000	7650.0000	62.8936	560.4256	1.0000	35.0000	4.9310	7.6307
Turbidity	NTU	5	0.8300	1990.0000	108.5450	284.3563	0.7000	374.0000	16.4665	38.5612	0.0400	16.4000	2.3417	2.3583
EC	uS/cm	–	23.2000	895.0000	228.4673	156.3387	20.3000	684.0200	181.8928	138.1115	4.4500	469.0000	139.1661	99.8480
As	mg/L	0.01	0.0005	0.0290	0.0028	0.0058	0.0005	0.0350	0.0019	0.0047	0.0005	0.0031	0.0008	0.0005
Cd	mg/L	0.003	0.0001	0.0010	0.0001	0.0001	0.0001	0.0010	0.0001	0.0001	0.0001	0.0020	0.0001	0.0002
CN	mg/L	0.07	0.0010	0.0890	0.0049	0.0083	0.0010	0.1700	0.0074	0.0204	0.0010	0.0570	0.0053	0.0101
Cu	mg/L	2	0.0010	0.2870	0.0154	0.0413	0.0010	0.0920	0.0046	0.0066	0.0010	0.0050	0.0040	0.0016
Hg	mg/L	0.001	0.0001	0.0010	0.0001	0.0001	0.0001	0.0010	0.0001	0.0001	0.0001	0.0008	0.0001	0.0001

Mean concentrations exceeding guideline values are italicized

and Ghahramani 2005). Given a dataset $M = \{X, y\}$, where $X = x_1, \dots, x_n$ represent the matrix composed by input vectors, $y = y_1, \dots, y_n$ represents the output, x_i is a vector and y_i is a variable (Eq. 1). The relationship between the input and the output can be given as:

$$y_i = f(x_i) + \varepsilon_n \quad (1)$$

where $f(x)$ is the underlying regression function, and ε_n is the noise term.

Principal component regression

PCA is commonly used in hydrological studies to reduce the number of variables, extract useful information and to eliminate the noise from data (Konaté et al. 2015b). PCA extracts eigenvalues from the original dataset and forms new principal components (PC) that are linear combinations of the parameters (Pearson 1901). The resulting PCs are orthogonal to each other after varimax rotation (Abou Zakhem et al. 2017; Ravikumar and Somashekar 2017; Pan et al. 2019), which helps to avoid multicollinearity between model parameters. PCs with eigenvalues greater than unity (one) are considered significant (Abou Zakhem et al. 2017; Selvakumar et al. 2017; Pan et al. 2019), and each significant PC explains a portion of the total variance of the dataset. In this study, PCR models are developed by using PCs identified by PCA as independent variables in MLR. PCR is more advantageous than conventional MLR modelling since it retains more original predictor variables and minimizes multicollinearity between variables. For a given trial, PCs on TDS are first identified from the training dataset and MLR is carried out using the significant PCs (total variance > 95%) to obtain a TDS prediction model. The original MLR equation (Eq. 2) derived from the training dataset was used for testing and validation. The MLR interaction equation used in the current PCR model is expressed by:

$$\text{TDS} = 1 + (V_1 \times V_2) + (V_1 \times V_3) + (V_2 \times V_3) \quad (2)$$

where V_1 , V_2 and V_3 represent principal components of the independent variables derived from the PCA.

Artificial neural network model

An artificial neural network is a computational model that consists of highly interconnected elements (nodes or neurons) and is used to simulate the structure and/or functional aspects of biological neural networks. ANN applications can be categorized as classification or pattern recognition, clustering or prediction and modelling. The advantages of ANNs are the unrestricted number of inputs and outputs and the clearly defined number of hidden layers and hidden neurons. In the present study, the back-propagation training

algorithm was used to adjust connection weights and bias values training.

Feed-forward network

A feed-forward network with one hidden layer was selected, in which the input data ($x_1, x_2, \dots, x_n$) are included in the first layer, and the network progressively processes those data throughout subsequent layers to produce the results ($y_1, y_2, \dots, y_k$) in the output layer. The input neurons are linked to those in the intermediate layer by w_{ji} weights (weight connecting the i th neuron in the input layer and the j th neuron in the hidden layer), and the neurons in the intermediate layer are linked to those in the output layer by w_{kj} weights (weight connecting the k th neuron in the output layer and the j th neuron in the hidden layer). The ANNs, based on the nonlinear activation functions, map the relationship between the inputs and the output. Thus, the explicit correlation for the output values is expressed in Eq. 3.

$$y_k = f_0 \left(\sum_{j=1}^s w_{kj} \cdot f_h \left(\sum_{i=1}^{s'} w_{ji} x_i + b_j \right) + b_k \right) \quad (3)$$

where f_h = activation function of the nodes in the hidden layer; f_0 = activation function of the nodes in the output layer; s and s' = number of nodes in the input and hidden layers, respectively; b_j = bias for the j th hidden neuron; and b_k = bias for the k th output neuron.

The five training algorithms commonly used in ANNs are Levenberg–Marquardt, gradient descent, gradient descent with momentum, gradient descent momentum and adaptive learning rate, and gradient descent with adaptive learning rate. As the Levenberg–Marquardt (LM) algorithm is assumed to be one of the fastest methods for training ANNs, it was chosen in this study.

Back-propagation neural network

The number of input and output nodes in the BPNN is determined by the nature of the actual input and output variables. The number of hidden nodes, however, depends on the complexity of the mathematical nature of the problem and is determined by the modeller, often by trial and error. Each hidden and output node processes its input by multiplying each of its input values by a weight, summing the product and then passing the sum through a nonlinear transfer function (e.g. sigmoid function) to produce a result (Eq. 4). It can be expressed as:

$$Y = f \left(\sum WX - \theta \right) \quad (4)$$

where X = input or hidden node value; Y = output value of the hidden or output node; $f()$ = transfer function; W = weights

connecting the input to hidden, or hidden to output, nodes; and θ =bias or threshold for each node.

The input, hidden and output layer nodes are interconnected by adjustable connection weights to recognize different patterns of information. A decision about the number of hidden layers and the number of hidden nodes is an important aspect of a neural network design process because it significantly affects the final output. For many practical problems, one hidden layer is sufficient to provide the required accuracy (Khalil et al. 2011; Wu et al. 2015; Azadi et al. 2016; Sinshaw et al. 2019). The current BPNN model developed in this study uses the same input and target variables as other methods. A BPNN structure of 9-10-1, representing 9 input parameters, 10 nodes in the hidden layer, and 1 output variable (TDS), was adopted (Fig. 2).

Performance measurement of models

The performance of each model is evaluated and compared using the methods discussed in this section.

Linear correlation coefficient

The linear correlation coefficient (R) is a measure of how well a particular model can accurately predict the observed (actual) data. The values of R usually range from -1.0 to 1.0. A value of 1.0 indicates a perfect positive correlation between the observed and the predicted and vice versa. The value of R is calculated using Eq. 5 given below:

$$R = \frac{n \sum y \cdot y' - (\sum y)(\sum y')}{\sqrt{[n(\sum y^2) - (\sum y)^2][n(\sum y'^2) - (\sum y')^2]}} \quad (5)$$

where y =observed value; y' =predicted value; and n =number of data samples.

Coefficient of determination (R^2)

The R^2 measures how much the variance in the observed values is explained by the model prediction. The higher the R^2 value, the better the model prediction accuracy.

Fig. 2 Proposed structure of a feed-forward back-propagation neural network

Table 2 Correlation coefficients of statistically input parameters on TDS

	pH	TSS	Turbidity	EC	As	Cd	CN	Cu-D	Hg
TDS	0.498	0.255	0.302	0.934	0.595	0.133	0.139	0.460	0.122

Root-mean-squared error

Root-mean-squared error (RMSE) is the square root of the mean square error. The RMSE is thus the average distance of an observed data point from the model line measured or the standard deviation of the prediction errors (Eq. 6). The RMSE is given by the following equation:

$$\text{RMSE} = \sqrt{\frac{\sum (y' - y)^2}{n}} \quad (6)$$

Mean absolute error

The mean absolute error (MAE) is an arithmetic of the absolute errors and statistically measures the predictive accuracy of a model. The MAE is commonly used in quantitative predictive models because it indicates the relative overall fit (i.e. the goodness of fit). The MAE is given by Eq. 7 below:

$$\text{MAE} = \frac{\sum_{i=1}^n |y - y'|}{n} \quad (7)$$

Table 3 Performance indices in training, testing and validation stages

Model	R^2	MAE	RMSE
GPR: training	0.997	2.255	5.276
GPR: testing	0.946	11.982	18.887
GPR: validation	0.978	5.790	11.684
BPNN: training	0.953	8.403	15.649
BPNN: testing	0.839	14.718	30.185
BPNN: validation	0.912	10.689	21.686
PCR: training	0.896	9.139	23.099
PCR: testing	0.796	19.068	33.849
PCR: validation	0.874	11.225	25.611

Nash and Sutcliffe coefficient of efficiency (E_{NS})

The Nash–Sutcliffe efficiency (E_{NS}) is a normalized statistic that determines the relative magnitude of the residual variance compared to the measured data variance (Nash and Sutcliffe 1970). E_{NS} indicates how well the plot of observed versus predicted data fits the 1:1 line. $0 < E_{NS} < 1$ indicate

Fig. 3 Prediction accuracy of GPR model for: **a** training; **b** testing; and **c** validation

Fig. 4 Performance of GPR model: **a** predicted TDS for each sampling points in GW; **b** predicted vs observed TDS in GW; **c** predicted TDS for each sampling points in SW; **d** predicted vs observed TDS in

SW; **e** predicted TDS for each sampling points in DW; and **f** predicted vs observed TDS in DW

a perfect match of the model to the observed data. $E_{NS} < 0$ indicates an unsatisfactory performance of the model. The value of E_{NS} is calculated using Eq. 8 given below:

$$E_{NS} = 1 - \frac{\sum_{i=1}^n (y - y')^2}{\sum_{i=1}^n (y - y_m)^2} \quad (8)$$

where y_m = mean of the observed value.

Index of agreement (d)

Index of agreement (d) represents the ratio of the mean square error and the potential error (Willmott 1982). It is a standardized measure of the degree of model prediction error which varies between 0 and 1. A d value of 1 indicates

Fig. 5 Prediction accuracy of BPNN model for: **a** training; **b** testing; and **c** validation

a perfect match of the model to the observed data, and 0 indicates unsatisfactory performance of the model. The d is given by Eq. 9 below:

$$d = 1 - \frac{\sum_{i=1}^n (y - y')^2}{\sum_{i=1}^n (y_x - y'_x)^2} \quad (9)$$

where $y_x = |y - y'_m|$, $y'_x = |y' - y'_m|$, y'_m = mean of the predicted value.

Results and discussion

Table 1 presents a summary of the concentration of parameters from GW, SW and DW used in this study. The mean concentration of turbidity was considerably higher than the guideline value in GW, SW and DW. This is possibly due to a high amount of effluents released as a result of numerous mining activities in the area. Although the mean concentrations of other parameters including TDS were lower than the guideline value, some parameters had a considerably high

maximum value above the guideline values. Considering the salinity problem associated with the water supply systems in the study area, accurate predictive models for constant monitoring of TDS in the study area are required to reduce the time and cost involved using conventional methods. The correlation between TDS and other input parameters was evaluated. The results show that there is a high correlation between EC and TDS with $R=0.934$ as presented in italic in Table 2. The performance of each model was also evaluated, and the results are summarized in Table 3.

Performance evaluation of models

Gaussian process regression model

For selecting the optimum covariance function for the proposed GPR model, the following covariance functions were tried and tested: (1) the squared exponential covariance function, (2) the exponential covariance function, (3) the rational quadratic covariance function, and (4) the Matérn 3/2 covariance function. The GPR model with the exponential

Fig. 6 Performance of BPNN model: **a** predicted TDS for each sampling points in GW; **b** predicted vs observed TDS in GW; **c** predicted TDS for each sampling points in SW; **d** predicted vs observed TDS in

SW; **e** predicted TDS for each sampling points in DW; and **f** predicted vs observed TDS in DW

covariance function had the highest R value with the least MSE and was then selected as the optimum covariance function. The GPR model was used to predict the concentration of TDS in this study. Compared to other models, the GPR model gave a satisfactory R^2 , MAE, RMSE, E_{NS} and d. The results of the GPR model during training, testing and validation stages are presented in Figs. 3a–c, respectively. The

performance of the GPR model was evaluated using 3 sets of inputs from GW, SW and DW. As shown in Figs. 4a, b, the TDS in GW was accurately predicted using the GPR model with R^2 of 0.980. The model also gave a good prediction of TDS in SW ($R^2=0.982$) and DW ($R^2=0.999$) as shown in Figs. 4c–f, respectively. Overall, the GPR model gave an accurate prediction for DW with R^2 , MAE, RMSE, E_{NS} and

Table 4 Summary of PCs loadings for input parameters

Parameters	PC1	PC2	PC3
pH	0.0001378506	0.0023913694	0.0036821250
TSS	0.9914666691	0.1279330402	0.0250382271
Turbidity	0.0941815455	0.8357669478	0.5409468015
EC	0.0901313566	0.5339672468	0.8406759696
As	0.0000010062	0.0000149702	0.0000130833
Cd	0.0000000068	0.0000000199	0.0000001504
CN	0.0000005769	0.0000064031	0.0000156789
Cu	0.0000055473	0.0000810995	0.0000267496
Hg	0.0000000004	0.0000000340	0.0000001219

d of 0.999, 0.006, 0.013, 0.976 and 0.994, respectively, as shown in Table 5.

Back-propagation neural network model

The BPNN developed consists of three layers, i.e. input, hidden, and output layers. In accordance with existing literature (Hornik et al. 1989; Arthur et al. 2020), one hidden layer was

used in this study due to its capability to universally approximate any complex problem. The network, the hyperbolic tangent sigmoid and linear transfer functions were utilized in the hidden and output layers due to the nonlinearity of the input datasets. Training of the BPNN was done using the Levenberg–Marquardt algorithm (Moré 1978). The optimum BPNN obtained for this study had 9 inputs, one neuron in the hidden layer and one output, with the structure [9-1-1]. Figure 5a–c presents the results of the BPNN model during training, testing and validation stages, respectively. The performance of the BPNN model was further evaluated using 3 different sets of inputs from GW, SW and DW, and the results are presented in Figs. 6a–f. It is worth noting that the TDS concentrations in GW, SW and DW were accurately predicted using the BPNN model with R^2 of 0.945 (Fig. 6a, b), 0.936 (Fig. 6c, d) and 0.859 (Figs. 6e, f), respectively. Contrary to the GPR model, the BPNN model gave an accurate prediction for GW with R^2 , MAE, RMSE, E_{NS} and d of 0.945, 9.100, 16.672, 0.950 and 0.987, respectively, as shown in Table 5.

Fig. 7 Prediction accuracy of PCR model for: **a** training; **b** testing; and **c** validation

Fig. 8 Performance of PCR model: **a** predicted TDS for each sampling points in GW; **b** predicted vs observed TDS in GW; **c** predicted TDS for each sampling points in SW; **d** predicted vs observed TDS in

SW; **e** predicted TDS for each sampling points in DW; and **f** predicted vs observed TDS in DW

Principal component regression

PCR model, which contains a hybrid PCA and MLR models, is constructed in this study to minimize the multicollinearity of the variables. Table 4 presents a summary of all three PCs factor scores for the input parameters with the highest scores in *italics*. Figure 7a–c presents the results of the

PCR model during training, testing and validation stages, respectively. New inputs GW, SW and DW were used to evaluate the performance of the PCR model, and the results are presented in Fig. 8a–f. The TDS concentrations in GW, SW and DW were accurately predicted using the PCR model with R^2 of 0.875, 0.870 and 0.919 as shown in Fig. 8a–f, respectively. Similar to the GPR model, the PCA model gave

Table 5 Comparison of models' performance for GW, SW and DW

Model	R^2	MAE	RMSE	E_{NS}	d
GPR: GW	0.980	4.857	8.983	0.959	0.990
GPR: SW	0.982	7.407	14.734	0.965	0.991
GPR: DW	<i>0.999</i>	<i>0.006</i>	<i>0.013</i>	0.976	0.994
BPNN: GW	<i>0.945</i>	<i>9.100</i>	<i>16.672</i>	0.950	<i>0.987</i>
BPNN: SW	0.936	9.946	19.489	0.936	0.983
BPNN: DW	0.859	10.114	21.251	0.858	0.962
PCR: GW	0.875	11.389	26.351	0.875	0.966
PCR: SW	0.870	13.432	27.668	0.876	0.966
PCR: DW	<i>0.919</i>	<i>9.159</i>	<i>21.076</i>	0.870	0.964

an accurate prediction for DW with R^2 , MAE, RMSE, E_{NS} and d of 0.919, 9.159, 21.076, 0.870 and 0.964, respectively, as shown in Table 5.

Comparison of model performance

The predictive techniques proposed in this study were evaluated by using the performance methods (R^2 , MAE and RMSE) discussed previously. In general, indices during the training stage may not provide a good reference for accurate model evaluation (Bagheri et al. 2017). Therefore, the performance of each model (GPR, BPNN and PCR) was evaluated using new inputs parameters from GW, SW and DW as shown in Table 5. The best model performance (i.e. highest R^2 value, or lowest MAE and RMSE values) for each model is in italics.

The MAE and RMSE for training, testing and validation are shown in Fig. 9a, b, respectively. It is worth noting that the GPR model showed the best performance with the least error during training, testing and validation (Fig. 9). It was found that all models adequately predict the observed data

Fig. 9 Performance of models at training, testing and validation stage: **a** MAE; and **b** RMSE **Fig. 10** Performance of TDS prediction in GW, SW and DW using three different models: **a** MAE; and **b** RMSE

Table 6 Comparison of model prediction accuracy with previous works

Reference	Sample source and size	Technique	Input parameters	Best model: performance measure (R^2)
Asadollahfardi et al. (2012)	Rivers (189)	Time series prediction of TDS using ANN, MLP and RNN.	TDS and discharge.	MLP: 0.964
Maedeh et al. (2013)	Monitoring wells (71)	TDS prediction using BPNN.	EC, SO_4 , Cl, HCO_3 , TH, SAR, K, Na, Mg, Ca.	BPNN: 0.960
Nasr and Zahran (2014)	GW samples (78)	TDS prediction using BPNN.	pH	BPNN: 0.900
Pan et al. (2019)	Monitoring wells (27)	TDS prediction using three different modelling approaches: MLR, hybrid PCR and BPNN.	AS, Ca, Mg, Mn, K, Na, U, HCO_3 , Cl, SO_4 TH, pH and EC.	MLR: 0.981
This study	A total of 386 data points (GW = 189, SW = 110 and DW = 87)	TDS prediction using three different modelling approaches: GPR, BPNN and PCR.	As, Cd, Hg, Cu, CN, TSS, pH, Turbidity and EC.	GPR: 0.987

Sodium absorption ratio (SAR), total hardness (TH), calcium (Ca), magnesium (Mg), manganese (Mn), potassium (K), sodium (Na), uranium (U), bicarbonate (HCO_3), chloride (Cl), Sulphate (SO_4)

in GW, SW and DW, with the R^2 consistently greater than 0.85. The average R^2 values of GPR model ($R^2_{\text{avg}} = 0.987$) and BPNN ($R^2_{\text{avg}} = 0.913$) are higher than the PCR ($R^2_{\text{avg}} = 0.888$). In general, accurate predictions were made with input data from DW based on the GPR ($R^2 = 0.999$) and PCR ($R^2 = 0.919$) models; however, the BPNN gave a good prediction in GW ($R^2 = 0.945$). The performance of each model was again assessed using efficiency based indicators such as E_{NS} and d. The GPR, BPNN and PCR models, respectively, gave an E_{NS} of (0.967, 0.915, 0.874) and d of (0.992, 0.977, 0.965). Overall, the GPR model gave a better prediction with the highest average R^2 , E_{NS} and d, and the lowest average error (MAE and RMSE) values as shown in Fig. 10.

Table 6 summarizes the primary previous works in TDS prediction. It is worth noting that R^2 varied from 0.900 to 0.987, which indicates good performances in the overall proposed models for predicting TDS. By comparing with other models proposed in the literature, this current study achieved the best predictive performance model with R^2 of 0.987. Unlike previous works, the model used in this study ensured good generalization capability.

Conclusions

The overall performance of various models for predicting the concentration of TDS in GW, SW and DW in the Tarkwa mining area was evaluated. Different water parameters are implemented to develop TDS models, and the performance of the models is evaluated by various statistical indices. The major findings are:

- Although the mean concentrations of all the parameters used in this study were lower than guideline value, except turbidity, TDS was chosen as the target parameter considering the salinity problem associated with the water supply systems in the study area.
- The GPR, BPNN and PCR models developed in this study gave an accurate prediction of the observed data (TDS) in GW, SW and DW, with the R^2 consistently greater than 0.850.
- The GPR model gave a better prediction of TDS concentration, with an average R^2 , MAE and RMSE of 0.987, 4.090 and 7.910, respectively. The performance of each model was assessed using efficiency-based indicators such as the Nash and Sutcliffe coefficient of efficiency (E_{NS}) and index of agreement (d). The GPR, BPNN and PCR models, respectively, gave an E_{NS} of (0.967, 0.915, 0.874) and d of (0.992, 0.977, 0.965).
- Compared with other models proposed in previous works, the proposed model in this study gave the best performance (highest R^2 value of 0.987) with a superior generalization capability because of the use of datasets from different water supply systems.

In general, this research work provides an integrated analytical and modelling methods that would be useful for future prediction and management of water quality parameters in various water supply systems. The results obtained from this study suggest that advanced NLR techniques and machine learning approaches are appropriate for the prediction of water quality indices. Moreover, the models obtained from this study could form a basis for a more effective decision-making process which will help in maintaining

and improving the management of water supply systems, especially in mining communities.

Although the models used in this study have some predictive capabilities to some degree, the data used for validation and testing were very limited to the study area. It is therefore recommended that future studies should validate these models with large datasets. It is also recommended that future researches on predicting water quality parameters in developing African countries should examine novel models such as extreme learning machines (ELM), hybrid and ensemble models.

Data availability statement Data generated or analysed during the study are available from the corresponding author by request.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abou Zakhem B, Al-Charideh A, Kattaa B (2017) Using principal component analysis in the investigation of groundwater hydrochemistry of Upper Jezireh Basin. *Syria Hydrol Sci J* 62:2266–2279
- Adeloye AJ (2009) Multiple linear regression and artificial neural networks models for generalized reservoir storage–yield–reliability function for reservoir planning. *J Hydrol Eng* 14:731–738
- Adiat K, Ajayi O, Akinlalu A, Tijani I (2020) Prediction of groundwater level in basement complex terrain using artificial neural network: a case of Ijebu-Jesa, southwestern Nigeria. *Appl Water Sci* 10:8
- Alrashed AA, Gharibdousti MS, Goodarzi M, de Oliveira LR, Safaei MR, Bandarra Filho EP (2018) Effects on thermophysical properties of carbon based nanofluids: experimental data, modelling using regression, ANFIS and ANN. *Int J Heat Mass Transfer* 125:920–932
- Arthur CK, Temeng VA, Ziggah YY (2020) Novel approach to predicting blast-induced ground vibration using Gaussian process regression. *Eng Comput* 36:29–42
- Asadollah SBHS, Sharafati A, Motta D, Yaseen ZMJJoECE (2020) River water quality index prediction and uncertainty analysis: a comparative study of machine learning models, pp 104599
- Asadollahfardi G, Khodadadi A, Gharayloo R (2010) The assessment of effective factors on Anzali wetland pollution using artificial neural networks *Asian. J Water Environ Pollut* 7:23–30
- Asadollahfardi G, Taklily A, Ghanbari A (2012) Application of artificial neural network to predict TDS in Talkheh Rud River. *J Irrig Drainag Eng* 138:363–370
- Atta HSAF, Amer AWM, Atta SAF (2018) Hydro-chemical study of groundwater and its suitability for different purposes at Manfalut District. *Assuit Govern Water Sci* 32:1–15
- Azadi S, Amiri H, Rakhshandehroo GR (2016) Evaluating the ability of artificial neural network and PCA-M5P models in predicting leachate COD load in landfills. *Waste Manag (Oxford)* 55:220–230
- Baah-Ennumh TY, Adom-Asamoah G (2019) Land use challenges in mining communities-the case of Tarkwa-Nsuaem municipality *Environ. Ecol Res* 7:139–152
- Banadkooki FB, Ehteram M, Panahi F, Sammen SS, Othman FB, Ahmed E-S (2020) Estimation of total dissolved solids (TDS) using new hybrid machine learning models. *J Hydrol* 1:124989
- Bhattacharya P et al (2012) Hydrogeochemical study on the contamination of water resources in a part of Tarkwa mining area. *Western Ghana J African Earth Sci* 66:72–84
- Chen T, Zhang H, Sun C, Li H, Gao Y (2018) Multivariate statistical approaches to identify the major factors governing groundwater quality. *Appl Water Sci* 8:215
- Chenini I, Khemiri S (2009) Evaluation of ground water quality using multiple linear regression and structural equation modeling. *Int J Environ Sci Technol* 6:509–519
- Chu W, Ghahramani Z (2005) Gaussian processes for ordinal regression. *J Mach Learn Res* 6:1019–1041
- Corcoran E (2010) Sick water?: the central role of wastewater management in sustainable development: a rapid response assessment. *UNEP/Earthprint*
- Emami S, Parsa J (2020) Comparative evaluation of imperialist competitive algorithm and artificial neural networks for estimation of reservoirs storage capacity. *Appl Water Sci* 10:1–13
- Ewusi A, Ahenkorah I, Kuma J (2017a) Groundwater vulnerability assessment of the Tarkwa mining area using SINTACS approach and GIS Ghana. *Min J* 17:18–30
- Ewusi A, Apeani B, Ahenkorah I, Nartey R (2017b) Mining and metal pollution: assessment of water quality in the Tarkwa mining area Ghana. *Min J* 17:17–31
- Gholami R, Kamkar-Rouhani A, Ardejani FD, Maleki S (2011) Prediction of toxic metals concentration using artificial intelligence techniques. *Appl Water Sci* 1:125–134
- Ghosh A, Das P, Sinha K (2015) Modeling of biosorption of Cu (II) by alkali-modified spent tea leaves using response surface methodology (RSM) and artificial neural network (ANN). *Appl Water Sci* 5:191–199
- Gu Q et al (2016) Assessment of reservoir water quality using multivariate statistical techniques: a case study of Qiandao Lake. *China Sustain* 8:243
- Hagbini M, Sharafati A, Dixon B, Kumar VJ, AoCMiE (2020) Application of soft computing models for simulating nitrate contamination in groundwater: comprehensive review. *Assessm Fut Oppor* 1:1–23
- Hamil S, Arab S, Chaffai A, Baha M, Arab A (2018) Assessment of surface water quality using multivariate statistical analysis techniques: a case study from Ghrib dam. *Algeria Arab J Geosci* 11:754
- Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. *Neural Networks* 2:359–366

- Ishii S, Sadowsky MJ (2008) *Escherichia coli* in the environment: implications for water quality and human health. *Microb Environ* 23:101–108
- Jacintha TGA, Rawat KS, Mishra A, Singh SK (2017) Hydrogeochemical characterization of groundwater of peninsular Indian region using multivariate statistical techniques. *Appl Water Sci* 7:3001–3013
- Karami S, Madani H, Katibeh H, Marj AF (2018) Assessment and modeling of the groundwater hydrogeochemical quality parameters via geostatistical approaches. *Appl Water Sci* 8:23
- Karimipour A, Bagherzadeh SA, Taghipour A, Abdollahi A, Safaei MR (2019) A novel nonlinear regression model of SVR as a substitute for ANN to predict conductivity of MWCNT-CuO/water hybrid nanofluid based on empirical data. *Physica A* 521:89–97
- Khalil B, Ouarda T, St-Hilaire A (2011) Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis. *J Hydrol* 405:277–287
- Konaté AA, Pan H, Khan N, Ziggah YY (2015a) Prediction of porosity in crystalline rocks using artificial neural networks: an example from the Chinese continental scientific drilling main hole. *Stud Geophys Geod* 59:113–136
- Konaté AA, Pan H, Ma H, Cao X, Ziggah YY, Oloo M, Khan N (2015b) Application of dimensionality reduction technique to improve geophysical log data classification performance in crystalline rocks. *J Petrol Sci Eng* 133:633–645
- Li Z et al (2018) Groundwater quality and associated hydrogeochemical processes in Northwest Namibia. *J Geochem Explor* 186:202–214
- Liu W-C, Yu H-L, Chung C-E (2011) Assessment of water quality in a subtropical alpine lake using multivariate statistical techniques and geostatistical mapping: a case study. *Int J Environ Res Public Health* 8:1126–1140
- Lu H, Ma X (2020) Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere* 249:126169
- Maedeh P, Mehrdadi N, Bidhendi G, Abyaneh HZ (2013) Application of artificial neural network to predict total dissolved solids variations in groundwater of Tehran plain. *Iran Int J Environ Sustain* 2:10–20
- Moré JJ (1978) The Levenberg-Marquardt algorithm: implementation and theory. In: *Numerical analysis*. Springer, pp 105–116
- Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I—a discussion of principles. *J Hydrol* 10:282–290
- Nasr M, Zahran HF (2014) Using of pH as a tool to predict salinity of groundwater for irrigation purpose using artificial neural network. *Egypt J Aquat Res* 40:111–115
- Pan C, Ng KTW, Fallah B, Richter A (2019) Evaluation of the bias and precision of regression techniques and machine learning approaches in total dissolved solids modeling of an urban aquifer. *Environ Sci Pollut Res* 26:1821–1833
- Pearson K (1901) On lines and planes of closest fit to systems of points in space. *Lond Edinb Dubl Philos Mag J Sci* 2:559–572
- Rasmussen CE, Williams CK (2006) *Gaussian processes for machine learning*. MIT Press, Cambridge, MA
- Ravikumar P, Somashekar R (2017) Principal component analysis and hydrochemical facies characterization to evaluate groundwater quality in Varahi river basin, Karnataka state, India. *Appl Water Sci* 7:745–755
- Schuttrumpf H (2018) Prediction and forecasting of total dissolved solids (TDS) by recurrent neural networks. *J Adv Res Dyn Control Syst* 10:1
- Selvakumar S, Chandrasekar N, Kumar G (2017) Hydrogeochemical characteristics and groundwater contamination in the rapid urban development areas of Coimbatore, India. *Water Resour Ind* 17:26–33
- Sen S, Nandi S, Dutta S (2018) Application of RSM and ANN for optimization and modeling of biosorption of chromium (VI) using cyanobacterial biomass. *Appl Water Sci* 8:148
- Shamshirband S et al (2015) Performance investigation of micro-and nano-sized particle erosion in a 90 elbow using an ANFIS model. *Powder Technol* 284:336–343
- Sharafati A, Asadollah SBHS, Hosseinzadeh MJPS, Protection E (2020) The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty
- Sinshaw TA, Surbeck CQ, Yasarer H, Najjar Y (2019) Artificial neural network for prediction of total nitrogen and phosphorus in US Lakes. *J Environ Eng* 145:04019032
- Suen J-P, Eheart JW (2003) Evaluation of neural networks for modeling nitrate concentrations in rivers. *J Water Resour Plan Manag* 129:505–510
- Sun L, Gui H (2015) Hydro-chemical evolution of groundwater and mixing between aquifers: a statistical approach based on major ions. *Appl Water Sci* 5:97–104
- Tziritis E, Lombardo L (2017) Estimation of intrinsic aquifer vulnerability with index-overlay and statistical methods: the case of eastern Kopaïda, central Greece. *Appl Water Sci* 7:2215–2229
- Viswanath NC, Kumar P, Ammad K (2015) Statistical analysis of quality of water in various water shed for Kozhikode City, Kerala, India. *Aquatic* 4:1078–1085
- Willmott CJ (1982) Some comments on the evaluation of model performance. *Bull Am Meteor Soc* 63:1309–1313
- Wu M-L, Wang Y-S, Gu J-D (2015) Assessment for water quality by artificial neural network in Daya Bay, South China Sea. *Ecotoxicology* 24:1632–1642
- Yaseen ZM, Ehteram M, Sharafati A, Shahid S, Al-Ansari N, El-Shafie AJW (2018) The integration of nature-inspired algorithms with least square support vector regression models: application to modeling river dissolved oxygen concentration. *Water* 10:1124
- Ziggah YY, Youjian H, Yu X, Basommi LP (2016) Capability of artificial neural network for forward conversion of geodetic coordinates (ϕ , λ , h) to cartesian coordinates (X, Y, Z). *Math Geosci* 48:687–721

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