Journal of Geoscience and Environment Protection, 2024, 12, 218-241
https://www.scirp.org/journal/gep
ISSN Online: 2327-4344
ISSN Print: 2327-4336

Modeling of Total Dissolved Solids (TDS) and Sodium Absorption Ratio (SAR) in the Edwards-Trinity Plateau and Ogallala Aquifers in the Midland-Odessa Region Using Random Forest Regression and eXtreme Gradient Boosting

Azuka I. Udeh[1], Osayamen J. Imarhiagbe[1], Erepamo J. Omietimi[2*]

[1]Department of Geoscience, University of Texas Permian Basin, Odessa, Texas, USA
[2]Department of Geology, University of Pretoria, Pretoria, South Africa
Email: *erepamo.omietimi@tuks.co.za

How to cite this paper: Udeh, A. I., Imarhiagbe, O. J., & Omietimi, E. J. (2024). Modeling of Total Dissolved Solids (TDS) and Sodium Absorption Ratio (SAR) in the Edwards-Trinity Plateau and Ogallala Aquifers in the Midland-Odessa Region Using Random Forest Regression and eXtreme Gradient Boosting. *Journal of Geoscience and Environment Protection, 12,* 218-241.
https://doi.org/10.4236/gep.2024.125013

Received: March 25, 2024
Accepted: May 27, 2024
Published: May 30, 2024

Abstract

Efficient water quality monitoring and ensuring the safety of drinking water by government agencies in areas where the resource is constantly depleted due to anthropogenic or natural factors cannot be overemphasized. The above statement holds for West Texas, Midland, and Odessa Precisely. Two machine learning regression algorithms (Random Forest and XGBoost) were employed to develop models for the prediction of total dissolved solids (TDS) and sodium absorption ratio (SAR) for efficient water quality monitoring of two vital aquifers: Edward-Trinity (plateau), and Ogallala aquifers. These two aquifers have contributed immensely to providing water for different uses ranging from domestic, agricultural, industrial, etc. The data was obtained from the Texas Water Development Board (TWDB). The XGBoost and Random Forest models used in this study gave an accurate prediction of observed data (TDS and SAR) for both the Edward-Trinity (plateau) and Ogallala aquifers with the R^2 values consistently greater than 0.83. The Random Forest model gave a better prediction of TDS and SAR concentration with an average R, MAE, RMSE and MSE of 0.977, 0.015, 0.029 and 0.00, respectively. For the XGBoost, an average R, MAE, RMSE, and MSE of 0.953, 0.016, 0.037 and 0.00, respectively, were achieved. The overall performance of the models produced was impressive. From this study, we can clearly understand that Random Forest and XGBoost are appropriate for water quality prediction and monitoring in an area of high hydrocarbon activities like Midland and Odessa

and West Texas at large.

Keywords

Water Quality Prediction, Predictive Modeling, Aquifers, Machine Learning Regression, eXtreme Gradient Boosting

1. Introduction

The importance of water for life sustainability cannot be overestimated, especially in arid to semi-arid regions like West Texas. This area gets most of its water resources from groundwater. Most ground waters are replenished during or after precipitation, which is very low in west Texas on account of the high rate of evaporation due to the hot climate. The Pecos Valley Alluvium (PVA), Ogallala, and Edwards-Trinity Plateau are the three most notable newly discovered local aquifers in the state of Texas. In addition to being utilized for human consumption, all are employed in home, industrial, and agricultural settings. Although groundwater is a replenishable resource, dry regions such as West Texas offer less natural replenishment (George et al., 2011). Owing to the large-scale industrial activities currently going on in the Midland-Odessa area, especially concerning the oil and gas exploration and varied nature of groundwater recharge like agricultural runoff into streams, pesticides applied to crops, etc., measures need to be kept in place for consistent monitoring of the water quality parameters (TDS, SAR, Nitrate, Arsenic, etc.) in this region. Contaminants in groundwater might be biological, radioactive, organic, or inorganic. Inorganic contaminants such as arsenic, chromium, copper, lead, and nitrate are the most prevalent in water and pose the greatest threat to human health among all other types of contaminants (Sharma & Bhattacharya, 2017). To reduce the impact of water-borne illnesses, assessment of water quality indices, Environmental Protection Agency (EPA) recommended water quality standards and guidelines are used to ascertain the biological, chemical, and physical constituents of water. One of the very important water quality indices is TDS (Atta et al., 2018; Li et al., 2018; Pan et al., 2019) which is the amount of solid remnant of both organic and inorganic matter after a liter of water is evaporated and is measured in milligrams per liter or ppm. Mainly considered a secondary contaminant and really not harmful to humans, higher TDS levels in drinking water can give the water bad taste and, in some cases, bad odor, on the other hand, very low concentration of TDS can result in water having flat taste. The presence of TDS in water stems from the dissolution of inorganic salts and some organic matter from either natural or anthropogenic sources. These inorganic salts include calcium, sodium, magnesium. Elevated levels of dissolved solids may potentially have technical implications. Hard water, which forms deposits and films on fixtures, the insides of hot water pipes, and boilers, can be created by dissolved solids. Hard water reduces the amount of lather that soaps and detergents generate.

Because hard water contains more minerals than other types, water filters will eventually wear out sooner. According to Sulthonuddin et al. (2018), an extreme value of TDS indicates the existence of dissolved salts and minerals, including carbonates, nitrates, bicarbonates, and chlorides which can be harmful to plants when used for irrigation purposes, because it increases the soil salinity.

Among all the measures used to assess the water quality for irrigation, SAR is the most important one (Sposito & Mattigod, 1977). It is the ratio of sodium ion to calcium and magnesium ion.

$$\mathrm{SAR} = \frac{\left(\mathrm{Na}^{+}\right)}{\sqrt{\frac{1}{2}\left[\left(\mathrm{Mg}^{2+}\right)+\left(\mathrm{Ca}^{2+}\right)\right]}} \tag{1}$$

Problems with infiltration arise from high salinity levels in water, which alter soil permeability. This is because when exchangeable sodium is present in the soil, it replaces the calcium and magnesium that are adsorbed on the soil clays and causes the soil particles to disperse (i.e., if the predominant cations adsorbed on the soil exchange complex are calcium and magnesium, the soil tends to be easily cultivated and has a permeable & granular structure). Soil aggregates break down as a result of this dispersion. When the soil is dry, it is compacted and hardens, and its structure is affected by a decrease in the rate at which air and water penetrate it. High SAR in soils can decrease the rate of infiltration of water into the soil and also the specific conductivity (Blaine et al., 1993). The amount of sodium rises with an increase in SAR values, which raises sodium dangers and reduces the usefulness of water for irrigation (Michael, 2008). In general, the infiltration rate of soils with SAR between 0 and 15 will be larger than that of a soil SAR between 16 and 30.

Models have been developed and applied in the past using different machine-learning algorithms for the prediction of TDS, and SAR and other water quality parameters (Ghosh et al., 2015; Emami & Parsa 2020; Elsayed et al., 2021; Sepahvand et al., 2018; Mohd Zebaral Hoque et al., 2022). The works of Sun and Gui, 2015 and Chen et al. (2018) which are deterministic and stochastic-based approaches such as statistical approaches and visual modeling have drawbacks in the sense that they require a lot of data, are time-consuming and quite expensive rendering the traditional model building approaches mostly ineffective.

The above mentions were made to reveal the intention and importance of this study which includes, developing a model for the prediction of the total dissolved solids (TDS) in the Edward-Trinity aquifer, and Ogallala aquifer using Random Forest and XGBoost to help the authorities monitor the contamination and pollution level of these aquifers. The machine learning approach tends to ease the task of water quality prediction. By predicting changes in the water quality index (WQI) based on past data. AI-based WQI prediction system helps in the provision of timely and effective water pollution prevention and response systems (Mohd Zebaral Hoque et al., 2022). Artificial neural network (ANN) technique was used to develop a model for predicting the WQI of groundwater from

physicochemical data obtained from 19 wells near a shale gas extraction site (Kulisz et al., 2021). In another study (Othman et al., 2020), using the ANN algorithm, a model was created using water quality parameters like COD, BOD, DO, SS, OH, and Ammoniacal Nitrogen (AN), and a high correlation of 98.78% was obtained. Wang et al. 2021 developed a model using the random forest regression algorithm for the water quality distribution for the Taihu Lake basin in Zhejiang Province, China and adopted the Shapley Additive explanation (SHAP) method to interpret the underlying driving forces. In their work with twelve machine learning algorithms, Khoi et al. (2022) concluded that the XGBoost may be employed for WQI prediction with a high level of accuracy which will further improve water quality management and monitoring. Many studies in hydrology (groundwater, rivers, wells) have adopted the machine learning approach, these studies include nitrate concentration forecasts in rivers (Suen & Erheart, 2003), predicting TDS in rivers (Mohd Zebaral Hoque et al., 2022), and predicting water quality parameters including chemical oxygen demand (COD), biochemical oxygen demand ((BOD), and TDS (Asadollah et al., 2020). Despite the wealth of machine learning algorithms available in our world today, studies involving the prediction of water quality parameters in the Midland-Odessa region using the machine learning approach are still very limited. Here we use two aquifers to create machine learning models for the prediction of TDS and SAR. These models will be applied long-term in these aquifers for TDS and SAR prediction to help the authorities monitor the aquifers to ensure the safety of the beneficiaries.

2. Methodology

2.1. Study Area

The study area extends across two counties in west Texas, Midland, and Odessa Counties (**Figure 1**). The climate in these regions is primarily semi-arid. Snowfall is rare in Midland and Odessa; rainfall is the primary source of precipitation (Kimmel et al., 2016). The two cities can be categorized as semi-arid environments by the United Nations Environmental Program since the annual average precipitation was almost equal to 13.78 inches (UNEP, 2019). Because of the arid environment and lack of water resources, it is essential to exploit groundwater aquifers. Consequently, because water recharge rates in semi-arid areas are slow, groundwater contamination has a substantial negative impact on public health (Heo et al., 2015). Anthropogenic activities like hydraulic fracturing and agriculture also play a crucial role in introducing contaminants to the aquifers, especially the unconfined aquifers. In the Midland and Odessa region, municipal wells pull water from the Pecos Valley aquifer (PVA), Ogallala aquifer, and the Edwards-Trinity plateau aquifer. Our interest in this study is the Ogallala aquifer which is the largest aquifer by area extent in the United States and the Edwards-Trinity aquifer (plateau). A transition boundary is not clearly defined between the Ogallala and the Edwards-Trinity aquifers in most parts of West Texas.

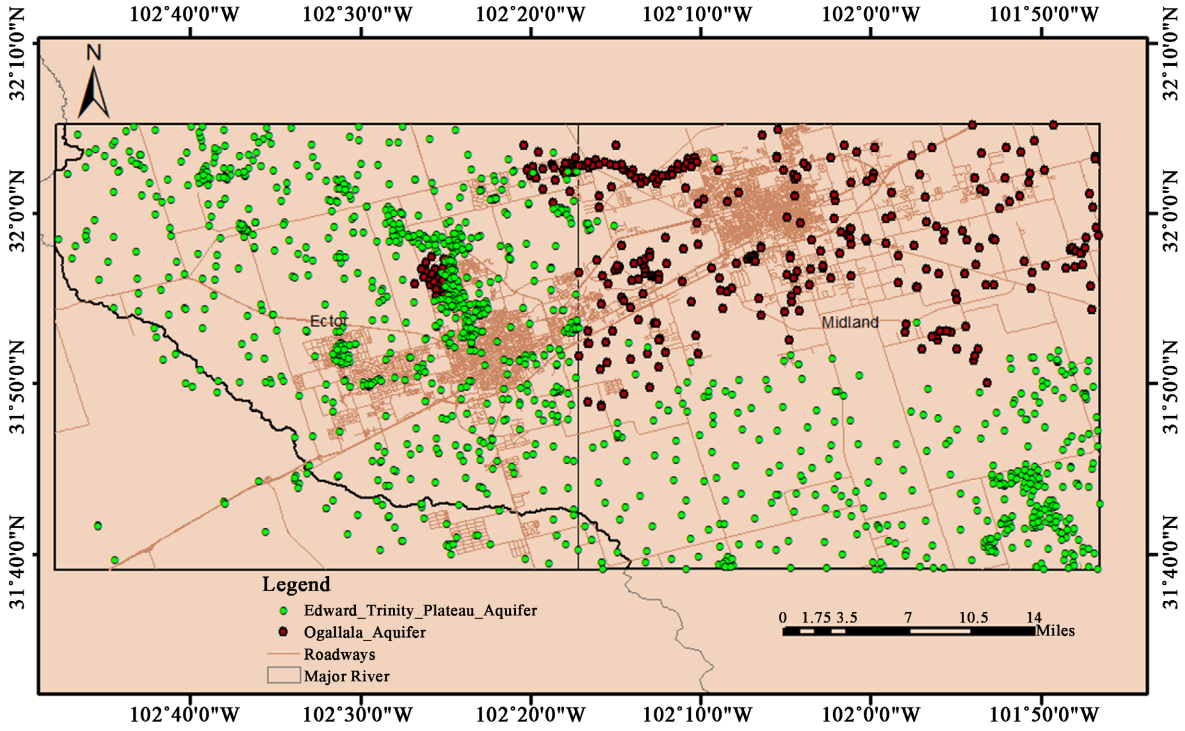

Figure 1. Map of the study area showing wells sunk into aquifers.

2.1.1. The Edwards-Trinity (Plateau) Aquifer

In the Pecos County region of western Texas, the Edwards-Trinity aquifer provides an essential groundwater resource for industrial, public supply, and agricultural applications (Barker & Ardis, 1992). The Edwards-Trinity aquifer is the most significant source of water for the Edwards plateau and covers approximately 23,000 square miles in southwest Texas (Blandford & Blazer, 2004). It is a major aquifer that extends to much of the counties in the southwestern part of Texas. The water-bearing units are composed predominantly of limestones and dolomites of the Edward Group and sands of the Trinity Group. While the maximum thickness of the freshwater saturated layer in the Edward-Trinity aquifer is about 433 feet, the total aquifer thickness is greater than 800 feet (George et al., 2011). The Edward-Trinity plateau was deposited during the early to middle Cretaceous. The total dissolved solids (TDS) range from 1100 - 3000 milligrams per liter, making it fall within the range of fresh water to slightly saline. The Edward-Trinity Plateau Aquifer is made up of two aquifers in close proximity, namely the Edwards and the Trinity aquifers. In some parts of West Texas, these two aquifers commingle due to faultings, as seen from **Figure 2**. In some other parts they are separated by confining layers.

2.1.2. Ogallala Aquifer

The Ogallala aquifer spans about eight states in the United States ranging from Texas to South Dakota and it is regarded as the largest aquifer by area extent in the United States. The aquifer comprises mostly sand, gravel, clay, and silt with a maximum thickness of 800 feet (George et al., 2011). It was deposited during the

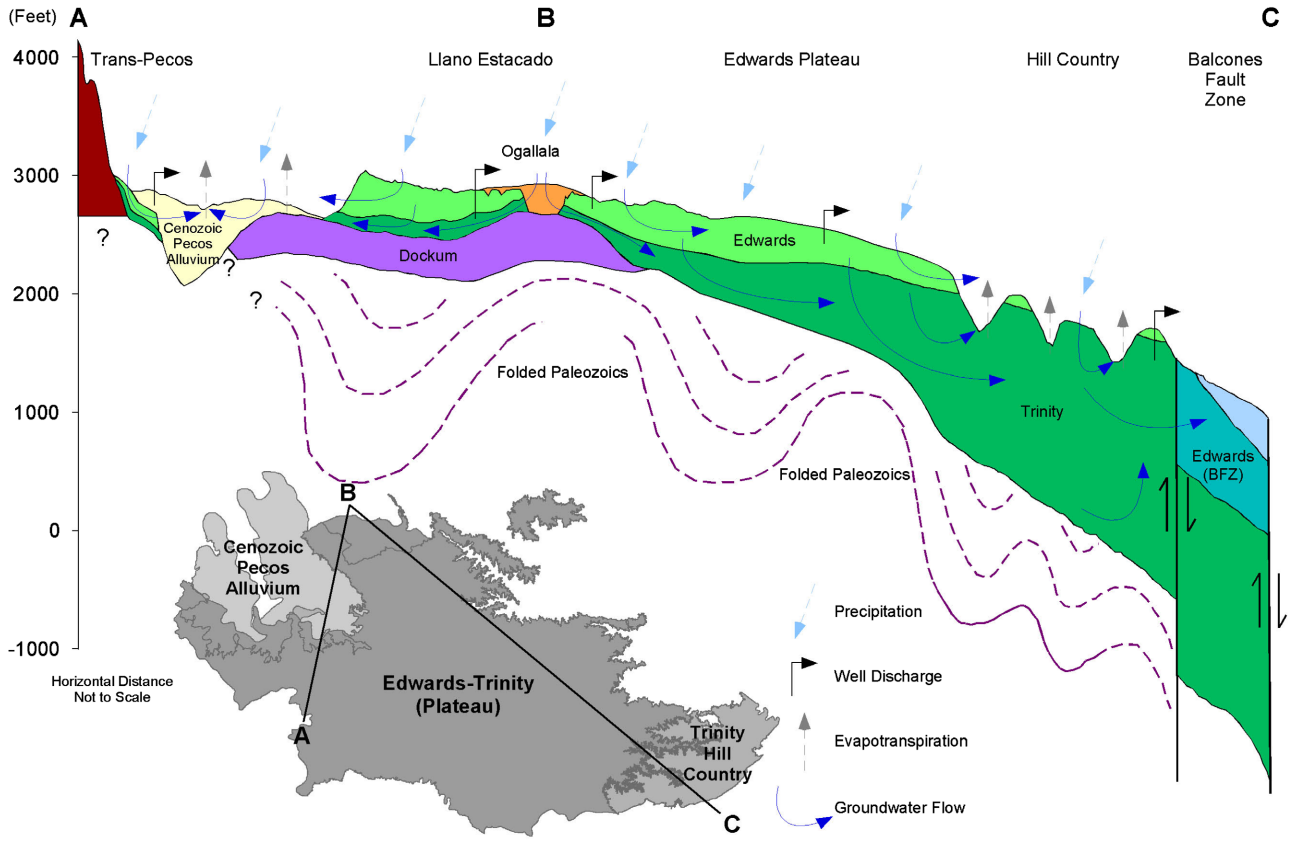

Figure 2. Conceptual model of the Edwards-Trinity (Plateau) and Pecos Valley aquifers and Hill Country part of the Trinity Aquifer (modified from Anaya and Jones, 2004, 2009). BFZ = Balcones Fault Zone.

late Miocene to early Pliocene. This aquifer provides significantly more water than other aquifers in Texas and is used mostly for irrigation.

The water demand used in fracking, especially in the Midland area, is mostly met by a collection of aquifers that filled up over millions of geologic years. Care and precautions should be put in place to monitor the chemicals and harmful substances that may find their way into the aquifers. This task is made even more difficult because the Texas Water Code exempts oil and gas producers from reporting exactly how many water wells are used for fracking. Estimates from the U.S. Geological Survey indicate that freshwater usage for hydraulic fracturing in the Texas Permian Basin surged by 2400% from 2010 to 2019, reaching a total of 72 billion gallons. This quantity is almost 1.5 times the volume of water used by the City of Austin during the same year. In 2019, the entire basin, which includes sections of New Mexico, yielded 1.4 billion barrels of oil. However, according to a report by the Texas consortium, Permian wells generated 3.93 billion barrels (about 165 billion gallons) of wastewater from fracking operations (Baddour, 2022). The majority of the water is transported through underground pipes and disposed away, while a portion is recycled for use in fracking. The planners have considered using it for crop irrigation rather than treating it, similar to how California handles wastewater that has lower salinity levels and does not contain fracking fluids. However, achieving this goal of cleaning the highly

polluted Permian water currently remains a high-tech goal for the time being.

2.2. Dataset

The data used for this study was accessed from the website of the Texas Water Development Board (TWDB). The data are freely available for anyone to access thus, the models produced from this study can easily be replicated. The historical dataset of water quality parameters from two aquifers Edward Trinity Plateau and Ogallala aquifers was taken. A total of 10 parameters, which include Calcium (Ca), Sodium (Na), Magnesium (Mg), Sulphate (SO_4^{2-}), Chloride (Cl^-), Total Hardness, Specific Conductance, (SC), Total Dissolved Solids (TDS), Sodium Absorption Ratio (SAR), and Percent Sodium, (PC), were extracted from the data. These parameters were chosen based on their correlation coefficients with the two dependent water quality parameters (SAR and TDS) which are also the target parameters we are trying to develop prediction models for. **Table 1** shows a statistical summary of all parameters used for this study.

Two machine learning algorithms Random Forest (RF), and XGBoost were used to predict TDS concentration and the sodium absorption ratio (SAR) in the Edward-Trinity (plateau) and Ogallala aquifers in the Midland and Odessa regions. The whole data for the Edward-trinity (plateau) aquifer from both counties were combined to get a model that can be applied to the whole region. The same approach was used for the Ogallala aquifer. This approach reduces modeling errors and possible introduction of bias into the model. The data were split into training (80%) and testing (20%) before using the algorithms for model prediction.

2.2.1. Random Forest Regression (RF)

A supervised learning algorithm and a bagging technique called random forest regression employ an ensemble learning approach for machine learning regression. In random forests, the trees grow in parallel; therefore there is no interaction between them as they grow (Shaikh & Barbé, 2021).

An assortment of tree predictors is called a random forest. $h(\mathbf{x}; \theta_k)$, $k = 1, \ldots, K$, where θ_k and x are independent, identically distributed (iid) random vectors, and x is the observed input (covariate) vector of length p with associated random vector X. As previously said, we primarily address regression settings where we have a numerical outcome, Y, but we also touch on some issues related to categorization (categorical outcome) concerns. Assumed to be independently selected from the joint distribution of (X, Y), the observed (training) data consists of n $(p + 1)$-tuples $(x_1, y_1), \ldots, (x_n, y_n)$ (Segal, 2004).

2.2.2. EXtreme Gradient Boosting for Regression (XGBoost)

XGBoost is a technique for group learning. Scalability is the key to this algorithm's power. It allows for economical memory utilization and rapid learning via distributed and parallel computing (Kiangala & Wang, 2021). In terms of mathematics, XGBoost is an ensemble learning technique that generates a strong

prediction by aggregating the predictions of several weak models. Decision trees, which are trained using gradient boosting, are the weak models in XGBoost (Dong et al., 2023). In other words, the algorithm fits a decision tree to the residuals of the previous iteration at each iteration. In many ML hackathons, XGBoost is the primary algorithm of choice. Its regular accuracy and time-saving benefits show just how beneficial it is. The capacity of XGBoost to accept missing values well is one of its main advantages for water prediction (Patel et al., 2023). This means that real-world water quality data may be handled by the algorithm without requiring a lot of pre-processing.

2.3. Models' Performance Measurement

The performance of each predictive model is evaluated and compared in this section.

2.3.1. Linear Correlation Coefficient

The capacity of a model to accurately predict the observed (actual) data is quantified by the linear correlation coefficient (R). R values typically range from −1.0 to 1.0. When there is no difference between the expected and observed, there is a total positive correlation (a value of 1.0), and vice versa. Another term for it is a value that represents the direction and strength of the linear relationship between two variables, x and y. Finding the covariance ratio between the two variables and multiplying their standard deviations by one another are the steps involved in computing this value.

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

2.3.2. Coefficient of Determination (R^2)

The R^2 measures how well the model's predictions explain the variance in the observed data. A higher R^2 value suggests that the model has better prediction accuracy. A number between 0 and 1 known as the coefficient of determination indicates how well a statistical model predicts a result. Even in cases where the correlation is negative, the coefficient of determination is always positive.

$$R^2 = \left(\frac{\sum_{i=1}^{n} \left(t_i - \underline{t}\right)\left(y_i - \underline{y}\right)}{\sqrt{\sum_{i=1}^{n} \left(t_i - \underline{t}\right)^2 \sum_{i=1}^{n} \left(y_i - \underline{y}\right)^2}} \right)^2 \tag{3}$$

2.3.3. Root-Mean-Squared Error (RMSE)

Root-mean-squared error, or RMSE for short, is the square root of the mean square error. The average distance between an observed data point and the measured model line, or the standard deviation of the prediction errors, is the definition of the root mean square error (RMSE). The RMSE is given by the following equation. By quantifying the degree of dispersion of these residuals, the RMSE offers information on how closely the observed data matches the predicted val-

ues. Because there are fewer mistakes in the model, the RMSE drops as the data points go closer to the regression line. A model that has less error produces more accurate predictions.

$$\text{RMSE} = \sqrt{\frac{\sum(y'-y)^2}{n}} \tag{4}$$

2.3.4. Mean Absolute Error (MAE)

The mean absolute error (MAE), which is the arithmetic of the absolute errors, is the statistical indicator of a model's predictive power. The MAE is often used in quantitative prediction models because it indicates the relative overall fit or goodness of fit. One of the most popular loss functions for regression problems, MAE helps users transform learning problems into optimization problems. It also offers regression problems with a straightforward, quantifiable error measurement.

$$\text{MAE} = \frac{\sum_{i=1}^{n}|y-y'|}{n} \tag{5}$$

2.3.5. Cross Validation (K-Fold Method)

Validation is determining whether the numerical results quantifying proposed relationships between variables are appropriate for characterizing the data. Therefore, what we need is a procedure that leaves enough data for both validation and training of the model. That is precisely what K Fold cross-validation achieves. The data in K Fold cross-validation is separated into k subgroups. The holdout approach is now performed k times, with each repetition using one of the k subsets as the test or validation set and the remaining $k-1$ subsets combined to create a training set. To determine our model's overall effectiveness, we use the average of the error estimation over the k trials (see Table 4). Each data point appears exactly once in a validation set and k times in a training set. The majority of the data is also utilized in the validation set. This considerably minimizes variance as well as bias because the majority of the data is used for fitting. This technique is made more effective by switching up the training and test sets. $k = 5$ or 10 is typically favored based on empirical evidence and general guidelines, however, it can take any value. In this study, we used $k = 10$.

3. Results and Discussion

Table 1 presents a summary of the concentration of the eight water quality parameters used in this study. The mean concentration of TDS was by far higher than the United States Environmental Protection Agency (EPA) recommended values, which was placed at 1000 mg/L. Certain parameters had maximum values that were significantly higher than the recommended limits, even while the mean concentrations of other parameters, such as SAR, were lower than the recommended value. Accurate prediction models for ongoing TDS and SAR

Table 1. Statistical Analysis of all Ten Parameters from Ogallala and Edwards-Trinity Plateau aquifers.

		OGALLALA AQUIFER				EDWARD-TRINITY PLATEAU AQUIFER			
Parameters	**Unit**	**Min**	**Max**	**Mean**	**Std dev**	**Min**	**Max**	**Mean**	**Std dev**
TDS	mg/L	1597.360	9296.000	1597.396	1527.060	274.000	10,313.000	1021.798	863.729
SAR	mg/L	3.550	22.530	3.557	3.082	0.26	61.330	2.452	2.742
Calcium	mg/L	182.030	767.000	4.610	142.790	32	800.000	147.195	93.402
Sodium	mg/L	263.100	2250.000	263.105	340.177	9.75	3610.000	141.739	210.136
Magnesium	mg/L	71.7900	502.000	71.798	66.031	1	328.000	33.280	29.481
Chloride	mg/L	448.260	5098.000	448.268	757.543	6.63	4633.000	173.905	377.830
Sulfate	mg/L	411.980	2883.000	411.982	478.914	10	3490.000	318.473	375.455
% Sodium		36.330	75.000	36.336	9.607	6	92.000	31.425	10.234
Specific Conductance	µS/cm	2917.090	19040.000	2917.096	3018.124	439	19,890.000	1748.964	1503.457
Total Hardness	mg/L	750.900	3694.000	750.905	592.743	125	3345.000	506.293	335.702

monitoring in the research area are necessary given the salinity issue with the water supply systems in the area to minimize the time and expense associated with employing traditional methods. The correlation between TDS, SAR and other input parameters was evaluated. The results show that there is a high correlation between Specific conductance and TDS with R = 0.94 and 0.94 in the Edwards-Trinity plateau and Ogallala aquifers respectively and also a correlation between Specific conductance and SAR with R = 0.76 and 0.90 in the Edward-Trinity plateau and Ogallala aquifers respectively as presented in Table 3. The performance of each of the eight models was evaluated, and the result summarized in Table 2.

3.1. Performance Evaluation of Models

The recognition of the best-performing regression model for TDS and SAR prediction from the two regression models (Random Forest and XGBoost) used in this study has been the focus of this study. Eight different models were generated, four each for the two algorithms. Each of the two algorithms was used to generate models for TDS and SAR separately for Edward-Trinity (Plateau) and Ogallala aquifers.

3.1.1. XGBoost Regression Model

XGBoost is a state-of-the-art algorithm and a decision tree enhancement approach appreciated for its skill in managing sparsity. To maximize its performance, a variety of hyperparameters were set up for the XGBoost regression model. To validate the model's performance, the resulting process involves training the XGBoost regression model on the training subset and then using it on the test subset. The coefficient of determination (R^2) for both training and testing for all four models is never below 0.80 in both aquifers. The model for

Table 2. The Performance of different machine Learning algorithms for estimation of TDS and SAR.

Performance evaluation of the algorithms							
Aquifers	Algorithm	Parameter		R^2	MAE	RMSE	MSE
Edward-Trinity Plateau Aquifer	Random Forest	TDS	Training	0.980	0.010	0.020	0.000
			Testing	0.930	0.040	0.070	0.100
		SAR	Training	1.000	0.000	0.010	0.000
			Testing	1.000	0.000	0.010	0.000
	XGBoost	TDS	Training	1.000	0.000	0.000	0.000
			Testing	0.920	0.040	0.060	0.000
		SAR	Training	1.000	0.000	0.000	0.000
			Testing	1.000	0.010	0.017	0.000
Ogallala Aquifer	Random Forest	TDS	Training	0.990	0.010	0.000	0.000
			Testing	0.930	0.040	0.089	0.100
		SAR	Training	1.000	0.000	0.017	0.000
			Testing	0.990	0.020	0.020	0.000
	XGBoost	TDS	Training	1.000	0.000	0.000	0.000
			Testing	0.870	0.040	0.105	0.011
		SAR	Training	1.000	0.000	0.000	0.000
			Testing	0.840	0.040	0.121	0.014

Table 3. Correlation coefficients of statistically input parameters on TDS.

		Calcium	Sodium	Magnesium	Chloride	Sulfate	Total Hardness	Specific Conductance	% Sodium
Edward-Trinity plateau	TDS	0.76	0.92	0.79	0.83	0.66	0.82	0.94	0.47
	SAR	0.34	0.94	0.44	0.81	0.35	0.4	0.76	0.68
Ogallala	TDS	0.92	0.93	0.82	0.91	0.66	0.93	0.94	0.59
	SAR	0.74	0.97	0.58	0.92	0.35	0.71	0.9	0.82

SAR in the Edward-Trinity (plateau) aquifer gave one of the best models with the coefficient of determination for both training and testing as 1.00 and 1.00 respectively (Table 1). The MAE, RMSE, and MSE for both the training and testing datasets in all four models produced were all below 0.5 (Table 1). The XGBoost model gave very good predictive models (Figure 5, Figure 6, Figure 9, Figure 10). It also gave very low errors of the models. An XGBoost model with the above evaluation metrics is a model that seems to perform exceptionally well on the given data. The RMSE of all four XGBoost models suggests that the model's predictions are spot on with the actual values. The models recorded very low RMSE values (Table 1). An MSE of 0.00 means no errors, as the squared errors are all zero. One would consider these numbers too good or a sign of overfitting,

hence it was crucial to evaluate the model on a separate test set to ensure generalization.

3.1.2. Random Forest Regression Model

The technique aggregates the predictions of several decision trees by averaging them. Additionally, it performs well when there are more observations than variables (Biau & Scornet, 2016). The RFR algorithm performed well in both training and testing for all four models (SAR and TDS) produced from both the Edward-Trinity (plateau) and Ogallala aquifers with coefficient of determination values always higher than 0.90 for training and testing datasets. With a closer look at **Figure 7(c)**, one can easily deduce that there is almost a perfect match between the predicted and observed values. The models developed for SAR in **Figure 4** and **Figure 8** for the Edward-Trinity and Ogallala aquifers using Random Forest also performed well. This agrees with Meshram et al., 2020 that the RFR model could provide successful modeling. Error analysis results were used to obtain the performance of the RFR prediction. If the difference between the models and the real data becomes smaller, the prediction is more accurate (see **Figure 3**). With an MAE mostly lower than 0.05, RMSE mostly lower than 0.08, and MSE mostly lower than 0.02 for both the training and testing datasets in both aquifers (**Table 2**), it can safely be concluded that the RFR performed well. Based on these data, the random forest model appears to be operating better.

4. Implications of Model Findings for Local Water Policy and Management Strategies

The results of this study provide understanding of groundwater quality in the study area, with implications for local water policies and management strategies. This research introduces a modeling approach that could be helpful for managing and predicting water quality parameters in the future in West Texas. The findings of the study suggest that machine learning methods such as XGBoost and Random Forest are effective in predicting water quality indicators. Furthermore,

Table 4. Training and testing results for cross validation (Coefficient of determination).

			Cross Validation (K-Fold Method)	
Aquifers	**Algorithm**	**Parameter**	**Training**	**Testing**
Edward-Trinity Plateau Aquifer	Random Forest	TDS	0.9120	0.900
		SAR	0.9000	0.939
	XGBoost	TDS	0.9080	0.818
		SAR	0.9890	0.943
Ogallala Aquifer	Random Forest	TDS	0.8790	0.813
		SAR	0.9800	0.870
	XGBoost	TDS	0.9410	0.734
		SAR	0.9720	0.894

the models derived from this research could serve as the foundation for improved decision-making processes aimed at supporting the maintenance and improvement of water supply management systems, particularly in mining regions.

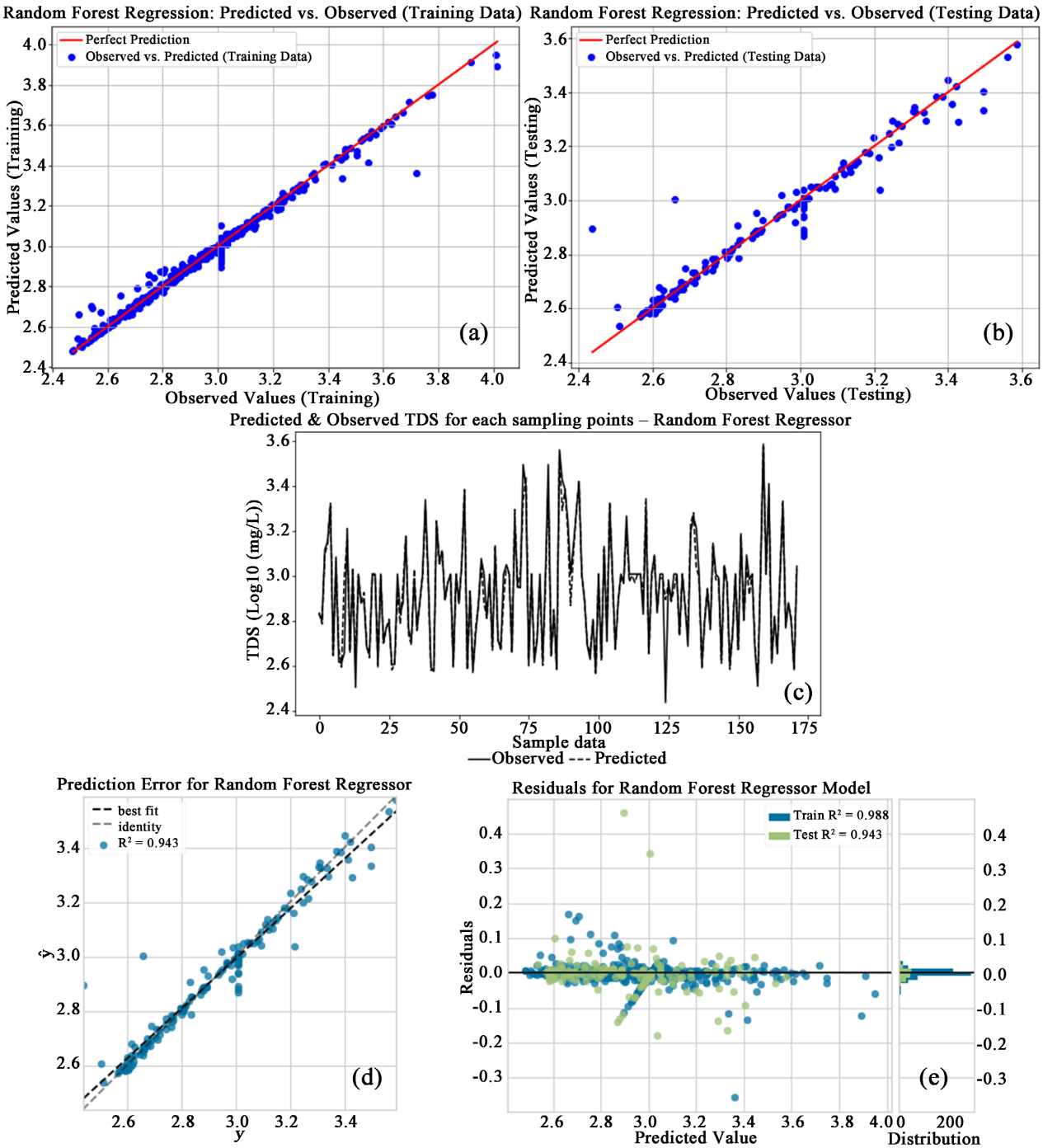

Figure 3. Performance of Random Forest algorithm in Edward-Trinity (plateau) aquifer for TDS in (a). Training (b). Testing (c). Predicted versus Observed (d). Prediction error (e). Residuals.

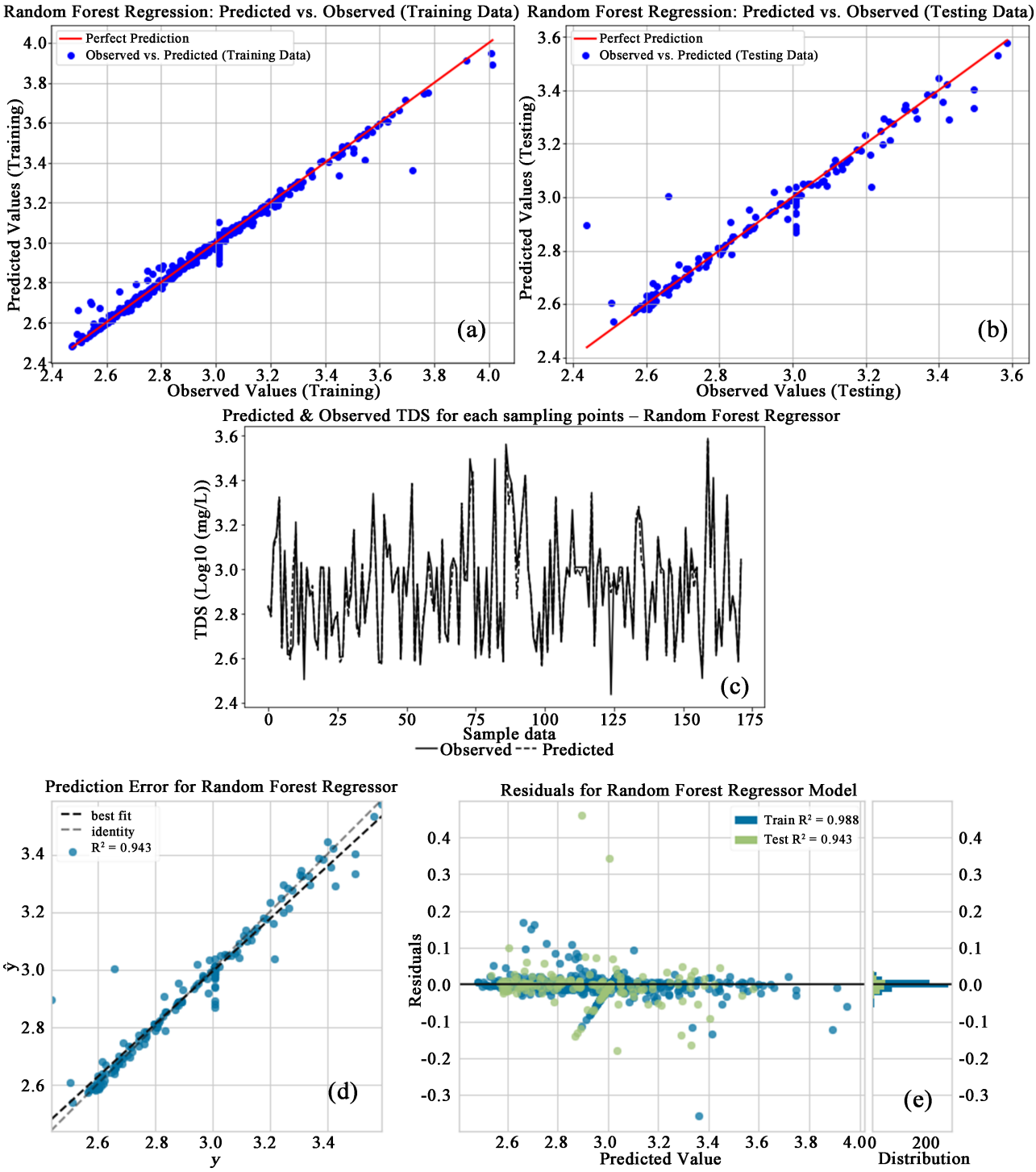

Figure 4. Performance of Random Forest algorithm in Edward-Trinity (plateau) aquifer for SAR in (a). Training (b). Testing (c). Predicted versus Observed (d). Prediction error (e). Residuals.

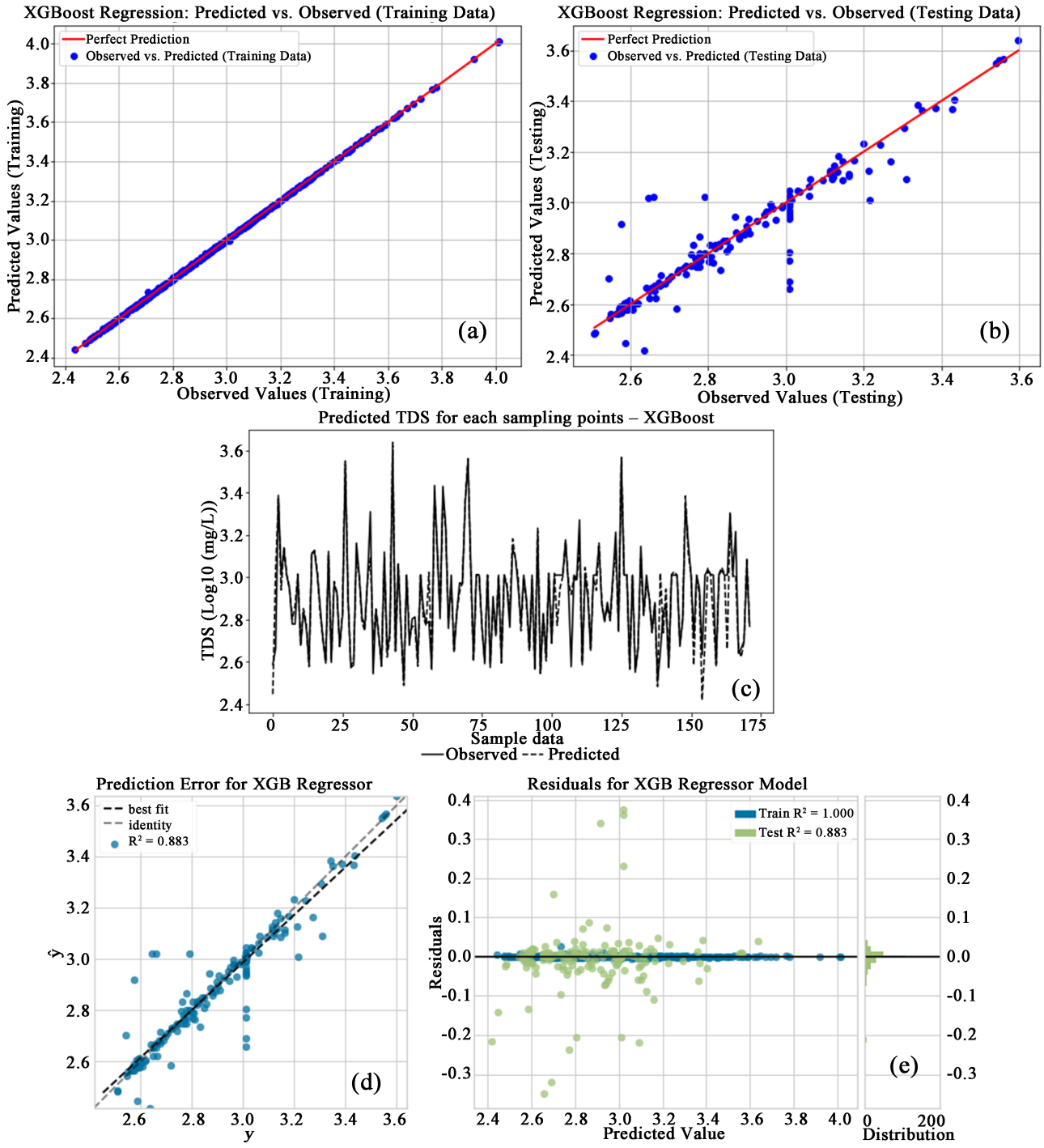

Figure 5. Performance of XGBoost algorithm in Edward-Trinity (plateau) aquifer for TDS in (a). Training (b). Testing (c). Predicted versus Observed (d). Prediction error (e). Residuals.

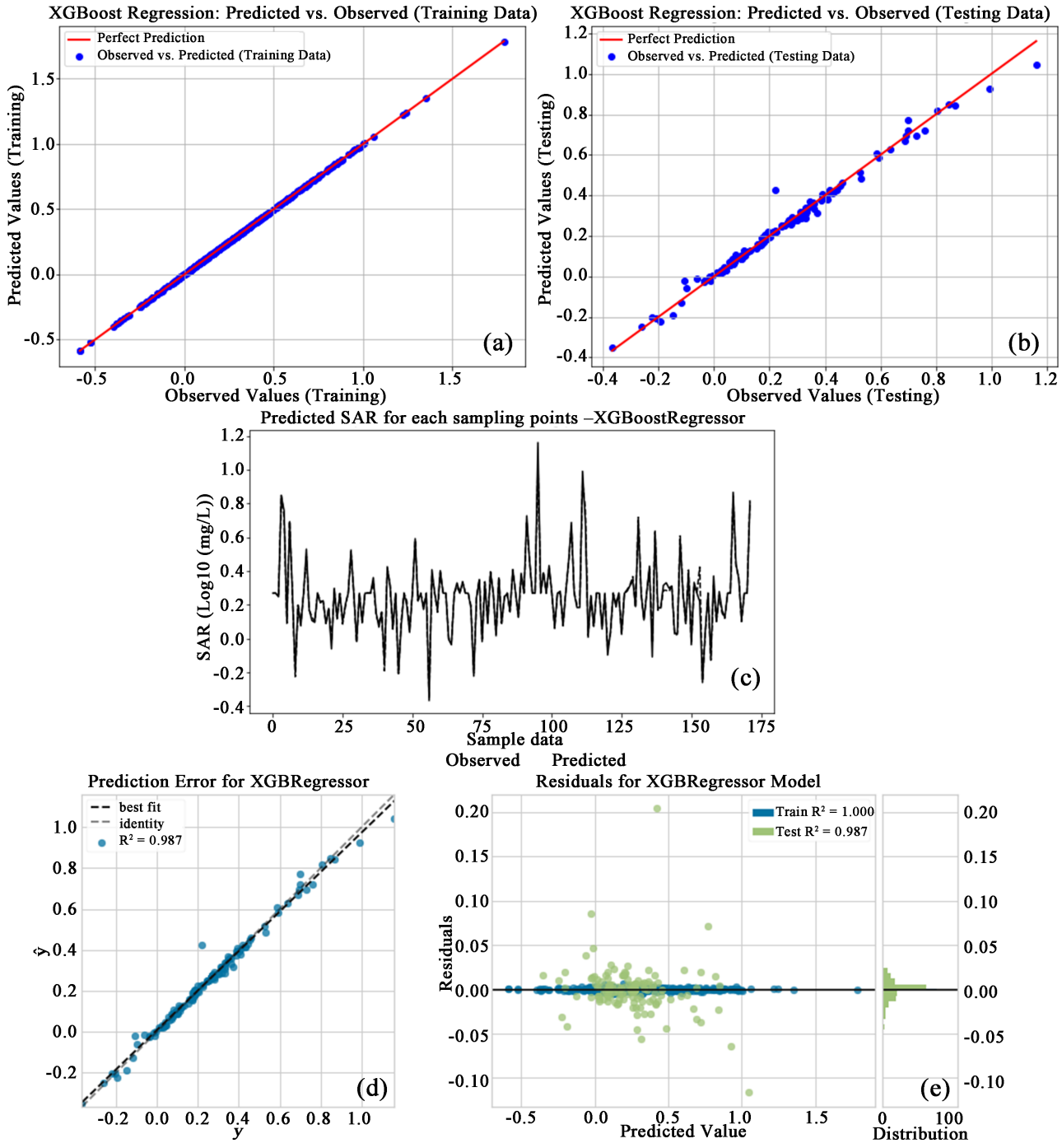

Figure 6. Performance of XGBoost algorithm in Edward-Trinity (plateau) aquifer for SAR in (a). Training (b). Testing (c). Predicted versus Observed (d). Prediction error (e). Residuals.

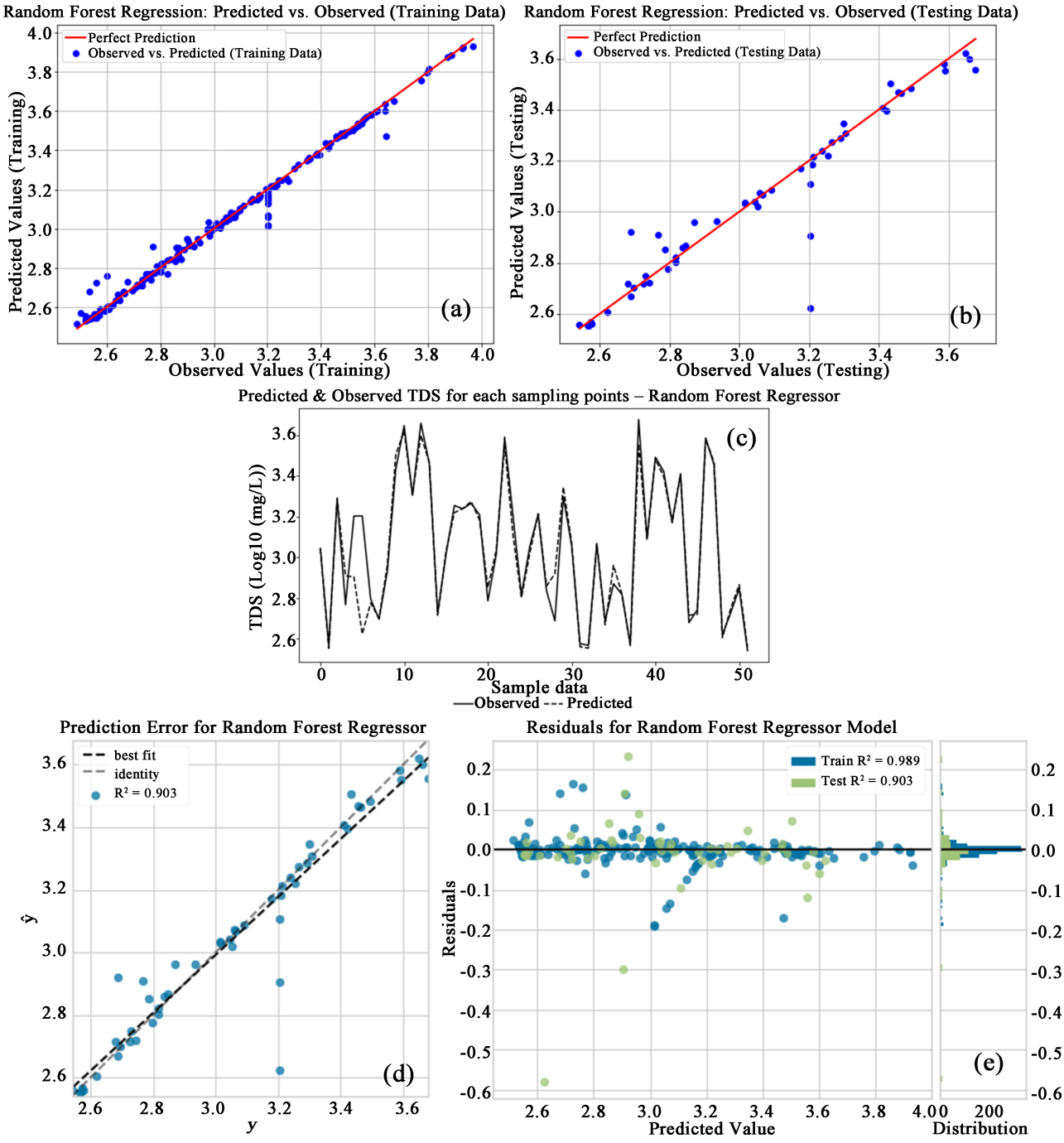

Figure 7. Performance of Random Forest algorithm in Ogallala aquifer for TDS in (a). Training (b). Testing (c). Predicted versus Observed (d). Prediction error (e). Residuals.

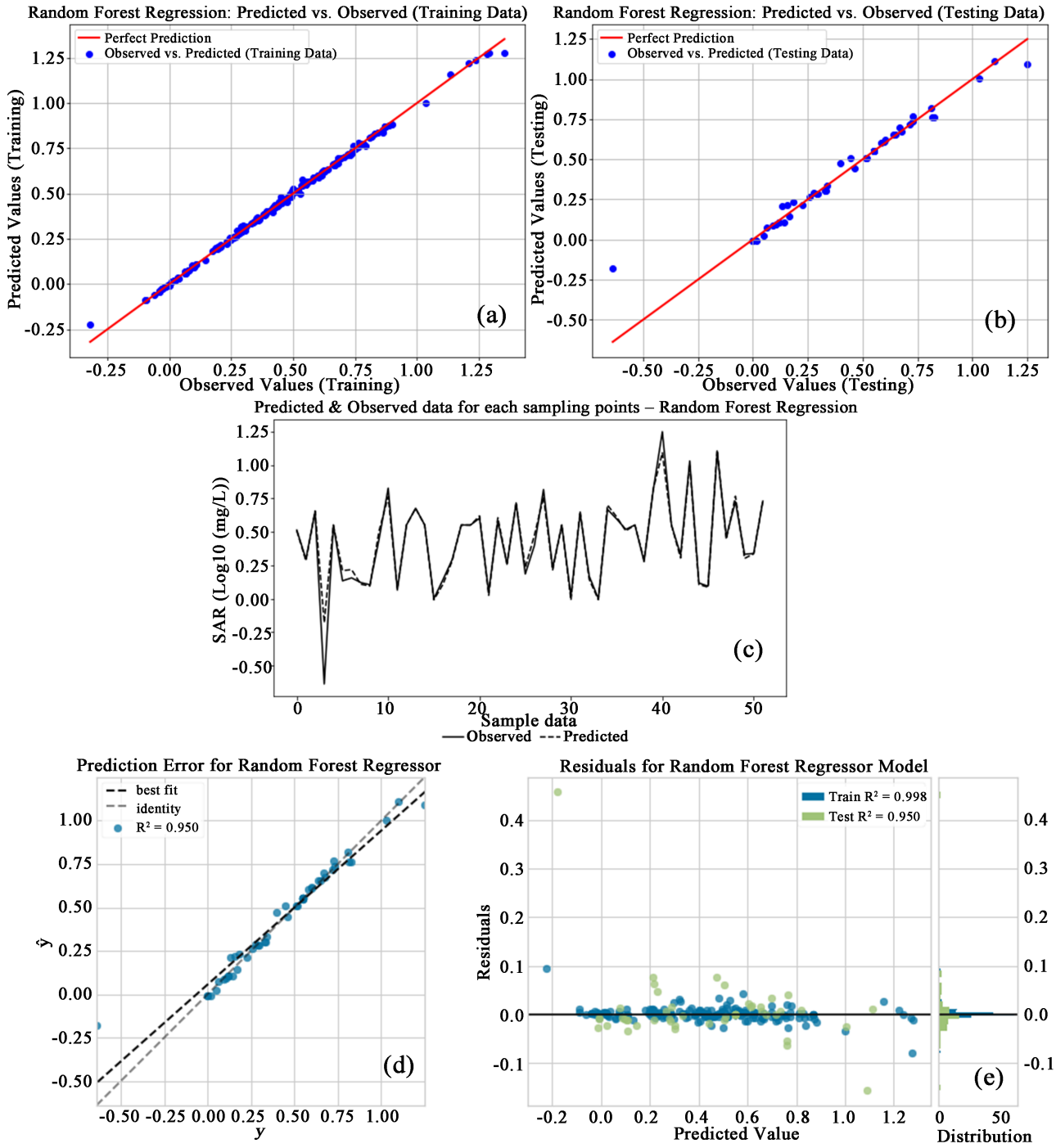

Figure 8. Performance of Random Forest algorithm in Ogallala aquifer for SAR in (a). Training (b). Testing (c). Predicted versus Observed (d). Prediction error (e). Residuals.

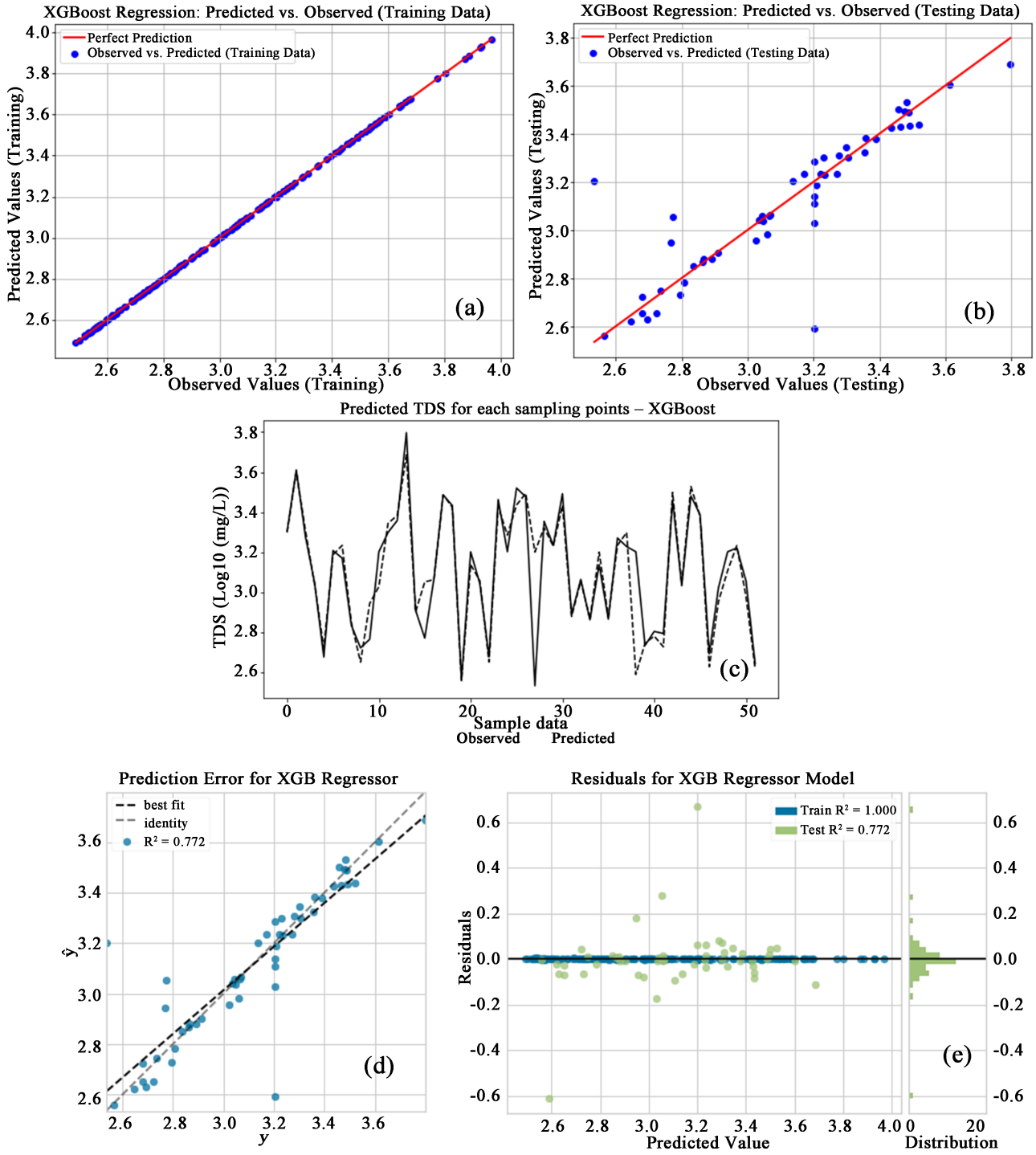

Figure 9. Performance of XGBoost algorithm in Ogallala aquifer for TDS in (a). Training (b). Testing (c). Predicted versus Observed (d). Prediction error (e). Residuals.

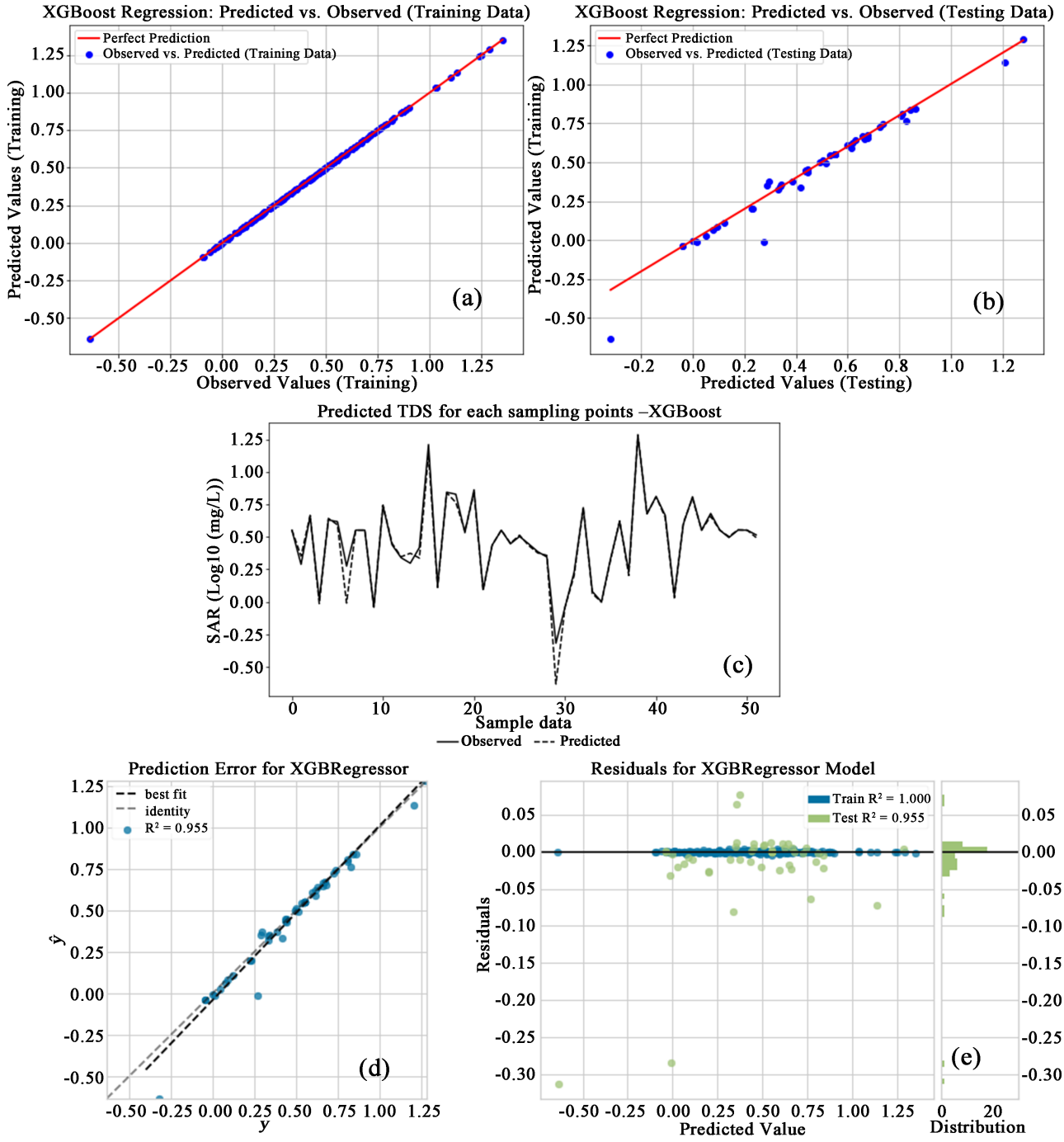

Figure 10. Performance of XGBoost algorithm in Ogallala aquifer for TDS in (a). Training (b). Testing (c). Predicted versus Observed (d). Prediction error (e). Residuals.

5. Conclusion

This work describes the performances of two regression-based modeling approaches and offers evidence for a suitable approach to estimate TDS and SAR. These methods included XGBoost and Random Forest Regression. When it comes to predicting water quality, machine learning has been shown to generate a simpler and more accurate model than physically and statistically based-data-

driven models. Accurate water quality prediction in West Texas is essential. Continuously monitoring and predicting water quality in the aquifers that generate the majority of water used for domestic and industrial activities can help authorities enact stricter laws to enable the long-term sustainability of these aquifers. Because of its distributed computing architecture and sophisticated parallel processing capabilities, XGBoost was the fastest method with a very good, reported performance. Using both TDS and SAR as dependent variables, Random Forest also generated excellent models that may be used to forecast the water quality in the Midland-Odessa area. Thus, models for predicting groundwater water quality can be created using machine learning techniques. These models can be used to gain a better understanding of the quality of the water and how inadequate monitoring of anthropogenic activities would eventually affect the availability of water. Furthermore, by offering precise and timely predictions of aquifer water quality, it can help make decisions by revealing the best approaches for forecasting water levels.

This research work is significant, because it offers integrated modeling and analytical techniques that could be helpful for managing and predicting water quality parameters in the future, particularly in West Texas. The study's findings indicate that machine learning techniques like XGBoost and Random Forest are suitable for predicting water quality indices. Additionally, the models developed from this research may serve as the foundation for a better decision-making process that would support the upkeep and enhancement of water supply system management, particularly in mining areas.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Asadollah, H. S. B., Sharafati, A., Motta, D., & Yaseen, Z. (2020). River Water Quality Index Prediction and Uncertainty Analysis: A Comparative Study of Machine Learning Models. *Journal of Environmental Chemical Engineering, 9,* Article ID: 104599. https://doi.org/10.1016/j.jece.2020.104599

Atta, U. R., Khairullah, K., Wahab, K., Aurangzeb, K., & Saqia, B. (2018). Unsupervised Machine Learning Based Documents Clustering in Urdu. *EAI Endorsed Transactions on Scalable Information Systems, 5,* e5. https://doi.org/10.4108/eai.19-12-2018.156081

Baddour, D. (2022). *To Ease Looming West Texas Water Shortage, Oil Companies Have Begun Recycling Fracking Wastewater*. Inside the Texas Tribune. https://www.texastribune.org/2022/12/19/texas-permian-basin-fracking-oil-wastewater-recycling/

Barker, R. A., & Ardis, A. F. (1992). *Hydrogeologic Framework of the Edward-Trinity Aquifer System, West Central Texas*. U.S Geological Survey Professional Paper 1421-B.

Biau, G., & Scornet, E. (2016). Rejoinder on a Random Forest Guided Tour. *Test, 25,* 264-268. https://doi.org/10.1007/s11749-016-0488-0

Blaine, H., Grattan, S. R., & Fulton, A. (1993). *Agricultural Salinity and Drainage: A*

Handbook for Water Managers. University of California.

Blandford, T. N., & Blazer, D. J. (2004). *Hydrologic Relationships and Numerical Simulations of the Exchange of Water between the Southern Ogallala and Edwards-Trinity Aquifers in Southwest Texas* (pp. 115-132). Aquifers of the Edwards Plateau: Texas Water Development Board Report 360.

Chen, T., Zhang, H., Sun, C., Li, H., & Gao, Y. (2018). Multivariate Statistical Approaches to Identify the Major Factors Governing Groundwater Quality. *Applied Water Science, 8,* Article No. 215. https://doi.org/10.1007/s13201-018-0837-0

Dong, J., Zeng, W., Wu, L., Huang, J., Gaiser, T., & Srivastava, A. K. (2023). Enhancing Short-Term Forecasting of Daily Precipitation Using Numerical Weather Prediction Bias Correcting with XGBoost in Different Regions of China. *Engineering Applications of Artificial Intelligence, 117,* Article ID: 105579. https://doi.org/10.1016/j.engappai.2022.105579

Elsayed, S., Ibrahim, H., Hussein, H., Elsherbiny, O., Elmetwalli, A. H., & Moghanm, F. S. (2021). Assessment of Water Quality in Lake Qaroun Using Ground-Based Remote Sensing Data and Artificial Neural Networks. *Water, 13,* Article No. 3094. https://doi.org/10.3390/w13213094

Emami, S., & Parsa, J. (2020). Comparative Evaluation of Imperialist Competitive Algorithm and Artificial Neural Networks for Estimation of Reservoirs Storage Capacity. *Applied Water Science, 10,* Article No. 177. https://doi.org/10.1007/s13201-020-01259-3

George, P., Mace, R., & Petrossian, R. (2011). *Aquifers of Texas*. Texas Water Development Board, Austin.

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). *Applied Water Science, 5,* 191-199. https://doi.org/10.1007/s13201-014-0180-z

Heo, J., Yu, J., Giardino, J. R., & Cho, H. (2015). Water Resources Response to Climate and Land-Cover Changes in a Semi-Arid Watershed, New Mexico, USA. *Terrestrial, Atmospheric and Oceanic Sciences, 26,* 463-474. https://doi.org/10.3319/TAO.2015.03.24.01(Hy)

Kiangala, S. K., & Wang, Z. (2021). An Effective Adaptive Customization Framework for Small Manufacturing Plants Using Extreme Gradient Boosting-XGBoost and Random Forest Ensemble Learning Algorithms in an Industry 4.0 Environment. *Machine Learning with Applications, 4,* Article ID: 100024. https://doi.org/10.1016/j.mlwa.2021.100024

Kimmel, T. M., Nielsen-Gammon, J., Rose, B., & Mogil, H. M. (2016). The Weather and Climate of Texas: A Big State with Big Extremes. *Weatherwise, 69,* 25-33. https://doi.org/10.1080/00431672.2016.1206446

Kulisz, M., Kujawska, J., Przysucha, B., & Cel, W. (2021). Forecasting Water Quality Index in Groundwater Using Artificial Neural Network. *Energies, 14,* Article No. 5875. https://doi.org/10.3390/en14185875

Li, J., Liu, H., & Paul Chen, J. (2018). Microplastics in Freshwater Systems: A Review on Occurrence, Environmental Effects, and Methods for Microplastics Detection. *Water Research, 137,* 362-374. https://doi.org/10.1016/j.watres.2017.12.056

Meshram, S. G., Safari, M. J. S., Khosravi, K., & Meshram, C. (2020). Iterative Classifier Optimizer-Based Pace Regression and Random Forest Hybrid Models for Suspended. *Environmental Science and Pollution Research International, 28,* 11637-11649.

https://doi.org/10.1007/s11356-020-11335-5

Michael, A. M. (2008). *Water Wells & Pumps*. Tata McGraw-Hill Education.

Mohd Zebaral Hoque, J., Ab Aziz, N. A., Alelyani, S., Mohana, M., & Hosain, M. (2022). Improving Water Quality Index Prediction Using Regression Learning Models. *International Journal of Environmental Research and Public Health, 19,* Article No. 13702. https://doi.org/10.3390/ijerph192013702

Nguyen Khoi, D., Nguyen, Q., Do, L., Thi Thao Nhi, P., & Thuy, N. T. (2022). Using Machine Learning Models for Predicting the Water Quality Index in the La Buong River, Vietnam. *Water, 14,* Article No. 1552. https://doi.org/10.3390/w14101552

Othman, A. H. A., Kassim, S., Rosman, R. B., & Redzuan, N. H. B. (2020). Prediction Accuracy Improvement for Bitcoin Market Prices Based on Symmetric Volatility Information Using Artificial Neural Network Approach. *Journal of Revenue and Pricing Management, 19,* 314-330. https://doi.org/10.1057/s41272-020-00229-3

Pan, C., Ng, K. T. W., 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. *Environmental Science and Pollution Research, 26,* 1821-1833. https://doi.org/10.1007/s11356-018-3751-y

Patel, A., Arora, G. S., Roknsharifi, M., Kaur, P., & Javed, H. (2023). Artificial Intelligence in the Detection of Barrett's Esophagus: A Systematic Review. *Cureu, 15,* e47755. https://doi.org/10.7759/cureus.47755

Segal, M. R. (2004). *Machine Learning Benchmarks and Random Forest Regression*. Center for Bioinformatics and Molecular Biostatistics, University of California, San Francisco. https://escholarship.org/uc/item/35x3v9t4

Sepahvand, A., Singh, B., Sihag, P., Samani, A. N., Ahmadi, H., & Nia, S. F. (2021) Assessment of the Various Soft Computing Techniques to Predict Sodium Absorption Ratio (SAR). *ISH Journal of Hydraulic Engineering, 27,* 124-135. https://doi.org/10.1080/09715010.2019.1595185

Shaikh, M. A. H., & Barbé, K. (2021). Study of Random Forest to Identify Wiener-Hammerstein System. *IEEE Transactions on Instrumentation and Measurement, 70,* 1-12. https://doi.org/10.1109/TIM.2020.3018840

Sharma, S., & Bhattacharya, A. (2017). Drinking Water Contamination and Treatment Techniques. *Applied Water Science, 7,* 1043-1067. https://doi.org/10.1007/s13201-016-0455-7

Sposito, G., & Mattigod, S. V. (1977). On the Chemical Foundation of the Sodium Adsorption Ratio. *Soil Science Society of America Journal, 41,* 323-329. https://doi.org/10.2136/sssaj1977.03615995004100020030x

Suen, J.-P., & Eheart, J. W. (2003). Evaluation of Neural Networks for Modeling Nitrate Concentrations in Rivers. *Journal of Water Resources Planning and Management, 129,* 505-510. https://doi.org/10.1061/(ASCE)0733-9496(2003)129:6(505)

Sulthonuddin, I., Harton, D. M., & Utomo, S. W. (2018). Water Quality Assessment of Cimanuk River in West Java Using Pollution Index. *E3S Web of Conferences, 68,* Article No. 04009. https://doi.org/10.1051/e3sconf/20186804009

Sun, L., & Gui, H. (2015). Hydro-Chemical Evolution of Groundwater and Mixing Between Aquifers: A Statistical Approach Based on Major Ions. *Applied Water Science, 5,* 97-104. https://doi.org/10.1007/s13201-014-0169-7

UNEP (2019). Emissions Gap Report. https://www.unep.org/resources/emissions-gap-report-2019

Wang, F. E. et al. (2021). Spatial Heterogeneity Modeling of Water Quality Based on Random Forest Regression and Model Interpretation. *Environmental Research, 202,* Article ID: 111660. https://www.sciencedirect.com/science/article/abs/pii/S0013935121009543?via%3Dihub