water

MDPI

Article

Integrating Machine Learning and Geospatial Analysis for Nitrate Contamination in Water Resources Management: A Case Study of Sinkholes in Winkler County, Texas

Rapheal Udeh [1], Joonghyeok Heo [1,*], Jeongho Lee [2] and Moung-Jin Lee [3]

[1] Department of Geosciences, University of Texas—Permian Basin, Odessa, TX 79762, USA
[2] Water Management Division, Land and Environmental Research Group, Korea Environment Institute (KEI), Sejong 30147, Republic of Korea
[3] Spatial Planning and Management Division, Land and Environment Research Group, Korea Environment Institute (KEI), Sejong 30147, Republic of Korea
* Correspondence: heo_j@utpb.edu or joonghyeok.heo@gmail.com

Abstract

This study used machine learning methods and spatial analysis to examine groundwater quality in Winkler County, Texas, focusing on nitrate pollution. By analyzing 85 years of groundwater data from six aquifers, the study uses advanced machine learning models Random Forest, Decision Tree, Linear Regression, and XGBoost to predict contamination levels and explore spatial and temporal trends. These models were chosen because of their ability to handle larger and more complex datasets and their ability to capture nonlinear relationships between water quality parameters and environmental variables. These machine learning algorithms are particularly effective at identifying patterns and interactions that may not be obvious with traditional analytical methods, and get more reliable and accurate results. Our decadal analysis specifically identified systematic fluctuations in nitrate levels, with a notable increase since the early 2000s, driven by the synergistic effects of rising temperatures and intensified agricultural land use. Climate change, pressured by rising temperatures and lessened precipitation, along with natural factors such as the formation of sinkholes, has been identified as a key driver of groundwater quality fluctuations. Elevated nitrate levels were mostly related to agricultural irrigation and excessive use of synthetic fertilizers. The machine learning model also highlights how land cover changes and human activities are contributing to groundwater quality deterioration. This research reinforces the value of integrating machine learning and spatial analysis for groundwater management. This is especially true in areas affected by sinkholes. It provides important information to reduce man-made impacts to water quality in West Texas.

Keywords: machine learning; spatial analysis; water quality; climate change; nitrate contamination; sinkholes; Texas

Check for updates

Received: 1 November 2025
Revised: 18 February 2026
Accepted: 21 February 2026
Published: 18 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Groundwater quality is a critical resource for public health, agriculture, and industrial operations, especially in regions like Winkler County, Texas, where surface water is limited and contamination is a growing concern. Winkler County relies heavily on groundwater for industrial, agricultural, and domestic use, making water quality monitoring crucial. As the primary source of freshwater for this semi-arid region, groundwater is increasingly

threatened by oil and gas exploration and agricultural activities. Monitoring specific contaminants, such as nitrate, is essential to effectively assess and manage these impacts.

Nitrate was chosen because its contamination serves as a critical indicator of agricultural impacts on groundwater. Nitrate levels rise significantly due to the excess application of chemical and synthetic fertilizers, which leach into aquifers and lead to severe environmental and public health risks. Contamination of groundwater from nitrates is often linked to the infiltration of inorganic nitrogen fertilizers and livestock waste resulting from agricultural practices. These fertilizers, commonly applied annually to enhance crop quality and yield, can penetrate the subsurface and leach into groundwater systems, exacerbating nitrate contamination issues [1]. These impacts include eutrophication in water bodies and methemoglobinemia in infants [2,3]. In Winkler County, this study found that regions characterized by intensive agricultural activities exhibited significantly elevated nitrate levels [4]. Their research linked fertilizer use to nitrate contamination, providing insights into how land-management practices directly influence water quality.

Focusing on nitrate offers a unique perspective on groundwater contamination. Nitrate reflects the environmental consequences of agricultural practices and provides targeted insight into human-induced impacts on aquifer systems. This approach also aligns with regulatory benchmarks, such as the Environmental Protection Agency's (EPA) contaminant limits, which prioritize monitoring nitrate due to its widespread occurrence and significant health and environmental risks. Moreover, land cover changes further complicate the contamination dynamics. Urbanization and agricultural expansion reduce groundwater recharge, alter natural drainage patterns, and increase pollutant runoff. Refs. [5–8] document how these changes exacerbate nitrate contamination, reinforcing the necessity of integrating land use considerations into groundwater management.

The unique geological features of Winkler County, including two major sinkholes formed in 1980 and 2002, add another dimension to the importance of nitrate monitoring. These sinkholes pose significant challenges to groundwater management and aquifer integrity, with oil and gas activities in the area potentially affecting water quality through drilling, fracking, wastewater disposal, and chemical usage [9]. While past studies in Winkler County and West Texas have predominantly focused on total dissolved solids (TDSs) as a general measure of water quality, nitrate provides more targeted insights into specific contamination sources. Refs. [9,10] emphasize the value of such parameters in identifying contamination drivers, particularly in areas where industrial and agricultural pressures overlap. However, limited long-term monitoring of this specific contaminant highlights a gap in understanding, which this study seeks to address.

By utilizing advanced machine learning algorithms, this research aims to predict nitrate concentrations, analyze their interactions with industrial and agricultural practices, and assess contamination trends over 85 years (1939–2024). Drawing on methodologies from [1,4,9], the study utilizes models such as Random Forest, Decision Tree, Linear Regression, and XGBoost to provide actionable insights for groundwater resource management. Notable prior ML applications include hybrid-boosted regression trees for nitrate mapping in California's Central Valley and national-scale XGBoost predictions across U.S. aquifers, which highlight agricultural drivers, but at broader scales [11,12]. Our study builds on these by focusing on long-term trends in a sinkhole-affected Permian Basin region, integrating spatial analysis of unique geological vulnerabilities [11,12]. The socioeconomic significance of this research cannot be overstated. Contamination of groundwater by nitrate threatens crop yields, public health, and regional livelihoods, making its monitoring and management critical. By integrating machine learning and spatial analysis, this study provides a robust framework for understanding and mitigating groundwater quality issues, addressing the challenges posed by industrial and agricultural activities in Winkler County.

The objectives of this study are to (1) examine changes in groundwater quality over the past 85 years (1939–2024) in the Wink Sink region; (2) utilize machine learning models and statistical analysis for predictive insights; (3) explore the interplay of groundwater quality parameters, focusing specifically on nitrate concentrations; and (4) identify the most effective machine learning model for groundwater quality assessment.

This research highlights the necessity of targeting nitrate as an essential parameter for understanding the multifaceted impacts of human activity on groundwater systems, offering a foundation for informed decision-making and sustainable resource management, and providing a framework to address similar challenges in other regions. Unlike scenario-based, basin-scale projections that assess nitrate responses under future climate and land-management pathways (e.g., Kebede et al., 2025), this study provides an observation-driven, county-scale machine learning workflow that integrates long-term monitoring (1939–2024) with land cover change and climate context to support local groundwater quality decision-making in an energy–agriculture co-impacted semi-arid region [12,13].

2. Study Area and Methodology

2.1. Study Area

The study area is on the western edge of the Monument Draw trough, which separates the Delaware Basin from the Central Basin Platform in the Permian Basin. There was an estimated population of 7791 people living in Winkler County; the county's center is located at 31°40′ latitude and 103°03′ longitude (Figure 1a). The total area of Winkler County, Texas, is 841.3 square miles (2180 km^2), of which 841.1 square miles (2179.5 km^2) are land and 0.2 square miles (0.5 km^2) are water [9–11]. Key features include sinkholes Wink Sink 1 (WS1) from June 1980 and Wink Sink 2 (WS2) from May 2002, near Wink and Kermit [9,14,15]. Around the western margin of the Central Basin Platform, these sinkholes formed above the Hendrick Field, which was initially uncovered in 1926.

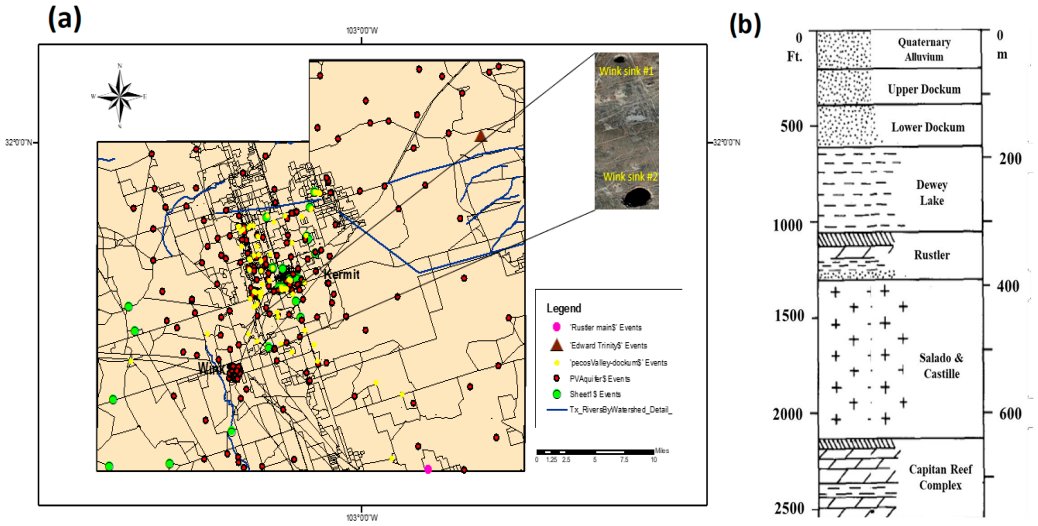


Figure 1. The map of the study area and geological overview of the Permian Basin in West Texas. Modified after English et al. (2020) [4].

The field is situated just above the Permian Capitan Reef complex in the Yates and Tansill Formations, and the field draws its production from upper Permian sand reservoirs (Figure 1b). At depths of 400–670 m, the upper Permian Salado Formation, which has minor shale and anhydrite, is coated in 200–300 m of interbedded salts that cover the producing layers [16]. The late Permian Rustler and Dewey Lake Formations above the Salado Formation have red beds and shales. The Capitan Limestone represents the reef deposits. Other

key formations include the Tansill Formation, Queen, Grayburg, Seven Rivers, and Yates Sandstone. The Castile Formation, made of anhydrite, replaces older sandstones. Cenozoic evaporite dissolution caused several formations to collapse, accumulating Pecos Valley Alluvium [13]. Major collapse features are the Pecos Trough and Monument Draw Trough. These dissolution processes create karst conduits and sinkholes that significantly enhance vertical hydraulic conductivity. These structural breaches act as preferential pathways that bypass natural soil filtration, directly linking surface NO_3^- loading from agricultural runoff to the underlying aquifers [4,8]. Consequently, the formation of sinkholes is a primary driver of NO_3^- variation, as it facilitates the rapid, unattenuated migration of contaminants into the groundwater system [4,11,13,14]. The Pecos Valley Aquifer (PVA) and the Edwards-Trinity Plateau are the principal aquifers, while the Dockum, Rustler, and Capitan Reef Complex are the minor aquifers [17]; Figure 1 illustrates the geological overview of the Permian Basin in West Texas.

In Winkler County, two primary freshwater sources are the Late Triassic Santa Rosa sandstone and the Cenozoic Pecos Alluvium. The Cenozoic Pecos Alluvium, made of sand, silt, and gravel, holds water from a few feet to about 150 feet. It is a recent deposit, while the Santa Rosa Formation, an older sandstone, is more profoundly saturated and can be either artesian or unconfined, affecting its pressure and flow [13]. Both rock layers have high mineral content, which affects water quality and causes problems such as scaling [15]. Cenozoic Pecos Alluvium (CPA) is one of the most important aquifers in the study area. It is composed of unconsolidated sediment and is generally about 100 feet thick. It can be difficult to distinguish it from the underlying rock because of its similar color, some of which require detailed study [13]. The Santa Rosa Formation, which may be unconfined or confined, engages with the Alluvium, and affects mineralization and drainage at a slower rate in the northern and eastern regions compared to the southern and central regions [15,17]. Rustler's Formation above the Salado Formation consists of dolomite, anhydrite, and sometimes limestone. It has a hollow shape, thus producing water with relatively low turbidity. It has flow rates of up to 3028.33 L per minute, although the water is often saline or brackish. The Capitan Reef Complex, which dates back to the late Guadalupe period, is a thick limestone formation with high salinity, which may be a problem for home use [18,19]. Edwards-Trinity Aquifer is the second largest aquifer in the study area. It contains sand from the Trinity group, and limestone and dolomite from the Edwards group. It is approximately 800 feet thick and contains sediments from the Triassic, Cretaceous, and Paleozoic periods [17,20]. It is recharged by the Ogallala aquifer and water quality varies with a total dissolved solids (TDSs) range between 1000 and 20,000 mg/L. This is in contrast to irrigation, where 95% of pumped groundwater is used for agriculture.

The Dockum Formation includes siltstone, claystone, and sandstone, with varying depths and water quality. It is connected hydraulically to the Pecos Valley Alluvium and provides water for various uses [4]. From 1991 to 2021, the Winkler County landscape has undergone significant changes. According to the Texas Geographic Information Office (TxGIO) and USGS National Land Cover Database (NLCD) products, the region has experienced remarkable changes in vegetation types and dramatic urbanization, with recent NLCD releases reporting Level I overall accuracies of approximately 83–90% (higher when including alternate labels), though arid regions may show increased confusion between shrubland and grass classes. Natural landscapes such as grasslands and scrublands have reduced due to urban and industrial expansion [4]. The appearance of infrastructure and human activities has been compared to the constant rise in developed land. On the other hand, the decline in natural vegetation reflects changes in cropping patterns in the area and potential environmental impacts. Winkler County, Texas has a semi-arid climate. It has scalding summers and moderate winters [21]. In summer, temperatures can rise above

32 °C and often near 38 °C. Winter is milder; it has normal daytime temperatures of 10 °C to 21 °C and sporadic nighttime lows that are near freezing. The majority of the precipitation is recorded in late spring and summer, with an average annual rainfall of 300 to 350 mm. The low relative humidity in the area contributes to the region's generally arid environment and has an impact on local crops, vegetables, and water supplies [21,22].

2.2. Dataset

2.2.1. Data Collection

The Texas Water Development Board's (TWDB) Groundwater Database (GWDB), a comprehensive, publicly available resource operated by the State's agency, provided all the groundwater data used in this study. It was chosen because of its comprehensive historical record and broad geographic coverage, and it establishes a home base as a reliable source of information on groundwater quality in Texas [23]. This database provides valuable insights into groundwater conditions across many aquifers in Texas. This makes it an ideal resource for evaluating groundwater quality in Winkler County. Figure 2 shows the workflow for the research methodology, divided into six key steps: data collection; data pre-processing; model selection; model development; model calibration; and model evaluation.

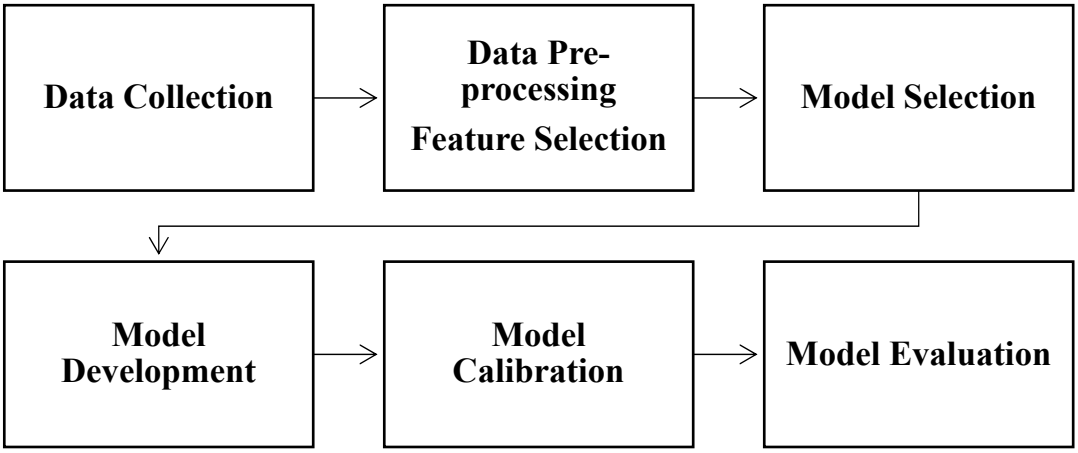


Figure 2. Schematic diagram of the methodology workflow.

The data collection process was based on historical records available in the study area. This results in a comprehensive database that spans decades and combines data from various watersheds. The wells selected for this study are geologically located in Winkler County and are representative of the county's most recent groundwater conditions. For reference materials, water quality samples were gathered from the Ground Water Database (GWDB) and examined. This kind of modeling and selecting process offers a helpful foundation for comprehending significant hydrologic circumstances in the region. The following modifications were used in the study to take into consideration contamination from both natural and human sources.

The first step was gathering groundwater samples at GWDB and analyzing their physicochemical characteristics, then followed by the application of machine learning models. Training and testing datasets were then created from the data. Additionally, algorithms utilizing Decision Tree (DT), Extreme Gradient Boosting (XGBoost), Random Forest (RF), and Linear Regression (LR) were developed and fine-tuned. In the third step, the models were optimized with the training dataset and validated using the test dataset. Together with error measures (R^2, MAE and RMSE) calculated to identify the best-performing model, finally, land cover and climate change studies were performed

using historical data from USGS, USDA and TxGIO. Temperature and rainfall data from 1990 to 2023 were included to assess the environmental impacts of oil and gas activities.

2.2.2. Data Pre-Processing

In this study, five hundred and ninety-four samples were collected for analysis from three hundred and five wells in six watersheds: Pecos Valley, Capitan Reef Complex, Edwards-Trinity Plateau, Rustler, Pecos Valley/Dockum Aquifer, and Dockum High. Referenced parameters include chloride (Cl^-), nitrate ($NO_3{}^-$), sodium (Na^+), specific conductivity, pH, temperature, chloride (Cl^-), fluoride (F^-), total alkalinity, total hardness, sulfate ($SO_4{}^{2-}$), carbonate ($CO_3{}^{2-}$), bicarbonate ($HCO_3{}^-$), total dissolved solids (TDSs), potassium (K^+), calcium (Ca^{2+}), and magnesium (Mg^{2+}). These parameters were selected based on the contaminants they produce and the potential health risks. As discussed by [1,4,9,11,24,25], the dataset covers October 1939 to June 2024; it covers a wide time period and ensures that the analysis is robust enough to take into account long-term trends in groundwater quality. After collecting information, pre-processing steps included filtering and organizing the dataset to make it suitable for analysis. This step included several important processes, such as irregularity detection in which statistical methods are evaluated by the proximity and cluster method [6]. In this study, standard deviation, Z-score, and graphical analysis were used to identify and resolve discrepancies. This ensures that the data used in machine learning models is of the highest quality. Additionally, samples or variables with large numbers of missing or NaN values were cleaned out.

Thresholds were set for missing data and any sample or variable that exceeded this threshold were excluded from further analysis. This estimation is effectively used by [26], so that only credible data were included in the analysis. The original dataset contained 27 physicochemical parameters. However, for this study, fifteen parameters were selected based on their relevance to nitrate dynamics (e.g., major ions influencing geochemical processes and mineralization), strong correlation with the target variable (nitrate) as identified via a correlation heatmap, and consistent data availability across the long-term record. Parameters with excessive missing values or weak relationships were excluded to improve model reliability and reduce overfitting. This allowed us to select the most suitable correlation coefficient. A parameter that shows a strong relationship with nitrate, a nitrogen-related compound, was stored for further analysis. Basic water quality knowledge and water quality literature in Winkler County was also included in this selection process. Farmland irrigation results in nitrate contamination [27]. To increase model reliability, parameters with significantly missing or inconsistent data were excluded. Dimensionality reduction techniques were used to limit the number of input features to 15 and to balance the need to capture an important groundwater quality trend with the need for computational efficiency and protection from excessive infiltration. Table 1 provides a summary of the statistics on the 15 selected parameters. This estimation streamlines the dataset, and helps ensure that only the most relevant parameters are used for machine learning analyses. This improves the accuracy and performance of the model somewhat. The selection of these predictors aligns with prior research indicating that, while regional nitrate levels are driven by land use, local variations are often better captured by major ion chemistry and well characteristics [8]. Specifically, including well depth and total hardness facilitates the identification of vertical migration pathways common in karst and sinkhole-prone terrains.

Table 1. Statistics of 15 selected parameters in the study area, Winkler County.

	Statistical Summary of Selected Winkler County's GW Parameters.				
Parameters	**Unit**	**Mean**	**Min**	**Max**	**Std Dev**
Sodium	mg/L	239.35	6	25,100	1327.2
Sulfate	mg/L	271.3	0	5040	491.57
Chloride	mg/L	466.63	3.2	41,000	2359.65
Temperature	°C	21.48	10	31.6	1.89
Calcium	mg/L	159.63	5.6	1880	216.11
Magnesium	mg/L	40.27	0.12	2920	132.51
Nitrate	mg/L	3.95	0	17	3.3
pH	-	7.57	4.4	9.1	0.37
TDS	mg/L	1301.42	105	71,121	4161.88
Total Alkalinity	mg/L	131.47	0	417	55.44
Total Hardness	mg/L	562.35	27	14,334	930.85
Percent Sodium	%	29.65	3	96	14.94
* SAR	-	2.94	0.13	137.19	7.96
Specific Conductance	µS/cm	2342.54	175	108,416	6598.44
Well Depth	ft	265.91	10	4400	331.77

Note: * SAR means "Sodium Adsorption Ratio".

2.3. Machine Learning

A subfield of artificial intelligence known as machine learning uses algorithms rather than explicit programming to analyze data and improve predictions. According to [28–30], its use in hydrology is expanding. This study used a variety of machine learning methods, including independent or stand-alone models and ensemble models. Ensemble models such as Random Forest (RF) and XGBoost combine multiple models to reduce biases and boost prediction accuracy. An overview and baseline standard for evaluating model performance has been proposed by stand-alone models such as Decision Trees (DTs) and Linear Regression (LR) [31–33]. Although the ensemble model combines a couple of models to offer higher generalization, stand-alone models help to recognize the benchmark performance of each individual model [34–36].

2.3.1. Machine Learning Algorithm

Machine learning models are fundamental to accurately analyzing and predicting contamination trends in subsurface water. This is especially true when analyzing complex, nonlinear data. This study uses independent and ensemble models to strike a balance between simplicity, interpretability, and predictive accuracy. Stand-alone models such as Linear Regression and Decision Trees are effective in understanding simple comparisons. But it can be difficult to use with complex formats in basic data. Ensemble models such as Random Forest and Extreme Gradient Boosting combine multiple models to improve prediction accuracy and reduce skewness and overfitting. In contrast to the model, the ensemble model combines speeds from multiple models. This makes it more useful for managing complex datasets. Although independent models are easier to interpret and simpler to calculate, the ensemble model gives better results. While stand-alone models like LR establish a baseline for linear dependencies, ensemble methods (RF and XGBoost) mitigate their limitations by capturing high-dimensional, nonlinear interactions. By linking these models, we ensure that the high predictive accuracy is grounded in both the fundamental geochemical trends and the complex environmental residuals [13,31]. Both categories were tested to determine which model best predicted nitrate contamination in groundwater.

Linear Regression (LR) is one of the most basic and widely used machine learning techniques for predictive analysis. It focuses on understanding and enforcing relationships

between variables. This estimation aims to estimate how changes in the independent variable will affect the dependent variable [37]. Linear Regression works by creating a linear relationship, where the model predicts the value of the dependent variable based on the value of one or more independent variables. This method is suitable for quantifying the impact of various factors, and estimation based on observed data. Its simplicity and clarity make Linear Regression a valuable tool for data analysis [38,39]. Linear Regression was chosen because of its simplicity and ability to provide insight into the linear relationship between the predictor and the target variable. This model provides a benchmark for evaluating the performance of more complex algorithms.

Decision Tree (DT) is a stand-alone learning method that analyzes data using a set of hierarchical and rule-based structures [40]. It divides data based on predefined criteria to achieve the most similarity inside every node and continues until further splits are not possible. Decision Trees are useful based on their interpretability and transparency. This allows users to quickly understand how various factors contribute in prediction outcomes. However, they have a tendency to overfit, especially if it grows too deep. This problem can be reduced through pruning, maximum depth setting or clustering methods such as Random Forest or Gradient Boosting [41,42]. Decision Trees are chosen for their interpretability and ability to handle nonlinear relationships between input features, which is useful for identifying minimum thresholds and interactions in groundwater quality parameters.

Random Forest (RF) is an ensemble machine learning algorithm and a robust extension of Decision Trees. It is widely used in water resources studies [43]. It combines classification and regression trees with bagging and stochasticity to improve prediction performance. Bagging involves creating a bootstrap sample set from a dataset, to ensure speed and reduce the fit. The model is trained on each set and the predictions are averaged. Randomization also helps to slow down tree correlation and prediction variance by selecting a different set with features for each distribution [44]. RF is capable of processing extremely large datasets. This makes it appropriate for groundwater quality analysis. It also gives an approximation of the feature importance metric. This is useful for learning how various criteria affect water quality. Key factors are the number of trees and variables addressed in each split. It is very important for the model performance [45,46]. This study optimizes these parameters using a randomized grid search design to ensure robustness and precision. Robustness, resistance to noise and excellent learning process makes RF an excellent choice for analyzing complex groundwater data.

Extreme Gradient Boosting (XGBoost) is a powerful machine learning algorithm. It is known for its efficiency and speed in data mining tasks [47] and generates a large number of green Decision Trees sequentially; every tree learns and corrects errors from previous trees. This iterative procedure increases the overall prediction accuracy of the XGBoost model by improving the stated objective function while correcting overfitting using appropriate regularization techniques [42,47]. Its versatility in tuning hyperparameters makes it a popular choice in a variety of study domains [48]. XGBoost was chosen for this study because of its ability to capture complex, nonlinear relationships and ability to self-correct before the next split. It can produce high-quality results while using low processing power. It is especially suitable for key hydraulic data with complex distribution patterns and relationships because of these properties. This makes it possible to simulate nitrate concentrations in an appropriate manner.

2.3.2. Machine Learning Development

This section details the development of machine learning models (LR, RF, DT and XGBoost). Data normalization is performed to prevent distorted results from multiple

ranges of input variables. It scales the input to a range between 0 and 1 using the min–max method. This modification improves the model's accuracy by reducing computation time and reducing errors. The normalization formula is as follows:

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

x_n is the normalized value, and x_{min} and x_{max} are the minimum and maximum values, respectively. The normalized dataset was divided into training and test sets using ratios of 60/40, 70/30, and 80/20 during initial calibration. The 80/20 split was ultimately selected because it consistently yielded the lowest test RMSE and highest test R^2 across all models [29,31,40]. Compared to 70/30 and 60/40 splits, the 80/20 ratio reduced test RMSE by approximately 12% and 25%, respectively, likely due to providing ensemble models with sufficient training data to capture complex nonlinear relationships in the moderate-sized dataset (n = 594). Model calibration focuses on optimizing the training algorithm by minimizing the cost function and adjusting the vector according to the training data. Overtune is normal. Therefore, hyperparameter tuning is used to obtain the best model architecture [42]. Metrics like mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination R^2 are used to evaluate the correctness of the model. The coefficient of determination, or R^2, measures how well a model explains variation in observed values. Higher R^2 values indicate greater accuracy in the model's predictions. It ranges from 0 to 1, which means that the model can show a strong relationship even if the relationship between the variables is negative, because the R^2 value is still positive.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (2)$$

SS_{res} is the residual sum of squares, and SS_{tot} is the total sum. Values closer to 1 suggest better predictive accuracy, with 0.75 often considered a baseline for acceptable performance. Root-mean-squared error (RMSE) is the main calculation that estimates the standard deviation for the prediction error. It compares the average distance to the observed data points predicted by the model and provides insight into how close the observed data is to the expected values. As the data points approach the regression line, the RMSE decreases and provides a more accurate representation of the prior error and velocity.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

y_i is the actual values, $\hat{y}_i$ is the predicted values, and n is the number of observations. This metric penalizes more significant errors more heavily, providing a clearer picture of overall model performance. Mean absolute error (MAE) quantifies the predictive power of a model by the mean absolute error. This is a prime example of general fit and is often used in quantitative models, as a popular loss function in regression. MAE helps translate learning problems into optimization problems. It provides a clear and understandable measure of prediction error.

$$MAE = \frac{1}{n}\sum_{n=i}^{n}|y_i - \hat{y}_i| \quad (4)$$

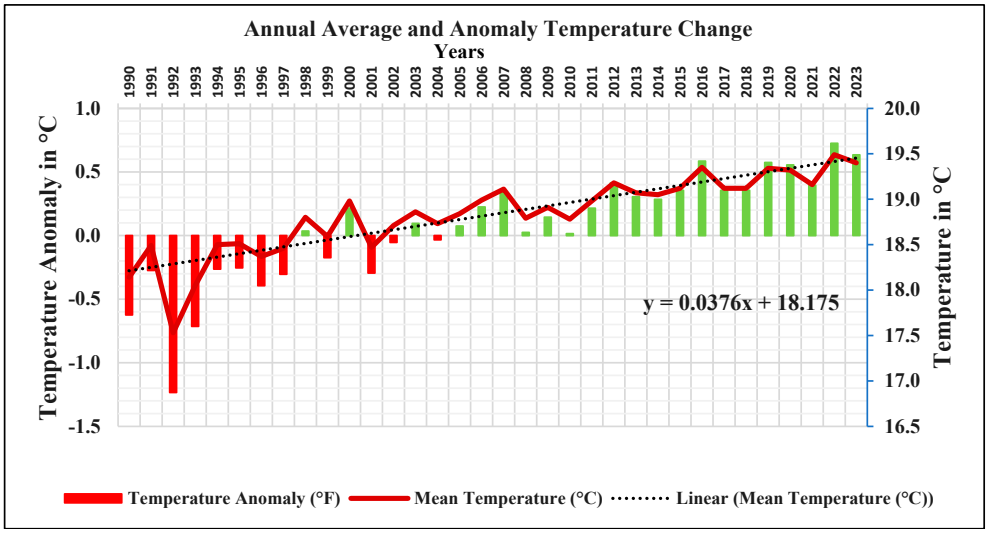
where y_i is the actual values, $\hat{y}_i$ is the predicted values, and n is the number of observations. During the model development process, the normalized dataset was carefully partitioned to ensure balanced training and evaluation. The majority of the data was allocated for training, typically between 70% and 80%, allowing the model to learn the underlying patterns and relationships among variables. A portion of the remaining data was used for validation, commonly ranging from 10% to 20%, to fine-tune hyperparameters and monitor

the model's performance throughout the training phase. This step helps prevent overfitting by providing feedback on how well the model generalizes to unseen data during calibration. Once the training and validation stages were completed, a separate test set, usually 10% to 20% of the original dataset, was used for final evaluation. This test set remained untouched during training and validation to offer an unbiased assessment of the model's predictive performance on new data. In some cases, k-fold cross-validation was also applied to further enhance reliability by rotating training and validation subsets, ensuring that the evaluation metrics reflect consistent performance across different data splits.

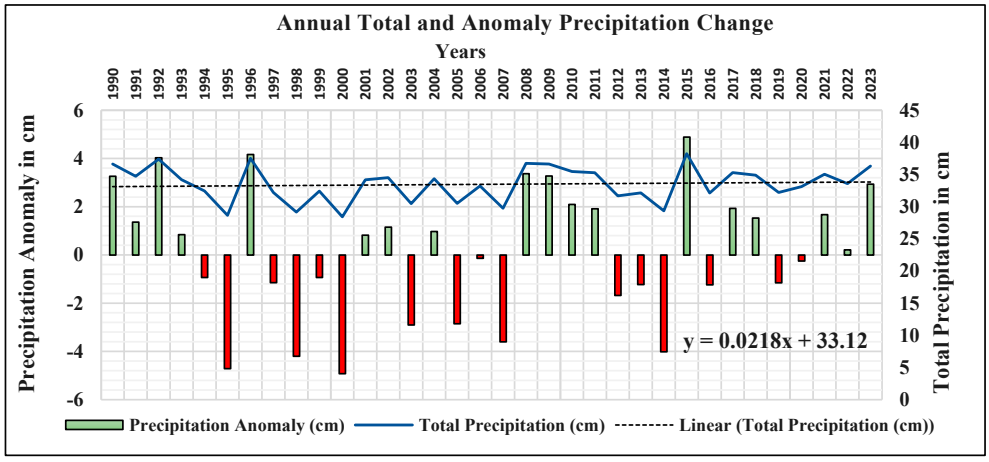
3. Results and Discussion

3.1. Climate Change

The annual mean temperature in Winkler County exhibits a slight increasing trend over the observation period, as shown in Figure 3a, while total annual precipitation shows notable interannual variability (Figure 3b). Both panels illustrate mean annual temperature and total annual precipitation along with their respective anomalies for West Texas. Climate data from the United States Geological Survey [49] and previous studies [50,51] support the detection of climate change in this region.



(**a**)



(**b**)

Figure 3. The change in climate. (**a**) Annual average and anomaly temperature change. (**b**) Annual total and anomaly precipitation change in Winkler County from 1990 to 2023.

The overall annual mean temperature for 1990–2023 reached approximately 18.83 °C, with maximum and minimum recorded temperatures of 27.19 °C and 10.47 °C, respectively (Table 2). Average temperatures for sub-periods showed slight variation, consistent with long-term warming: 18.4 °C for 1990–2000, 18.8 °C for 2001–2010, and 19.2 °C for 2011–2023. The mean annual temperature coefficient of variation (CV) was approximately 0.22, indicating relatively low interannual temperature fluctuations. Previous research in semi-arid regions reported temperatures between −16.3 °C and −15.3 °C during different tracking periods [50,52,53], consistent with observed regional warming. Global trends in warming also support these findings [12,54].

Table 2. Decadal and long-term statistics of temperature and precipitation in Winkler County (1990–2023). Temperature in °C, precipitation in cm; STD = standard deviation; CV = coefficient of variation.

Parameter	Period (Years)	MEAN	MIN	MAX	STD	CV
Temperature (°C)	1990–2023	18.83	10.47	27.19	7.84	0.22
Precipitation (cm)	1990–2023	2.41	0.64	8	1.6	0.57

In contrast, Table 2 shows that Winkler County's annual rainfall totaled 0.95 inches per month during the study period. Changes occur over many years. A 30-year baseline is used to assess trends and anomaly patterns. Specifically, the average monthly precipitation was 0.97 inches during 1990–2000, decreased to 0.88 inches during 2001–2010, and increased slightly to 0.95 inches during 2001–2010. Since 2010, 2011 to 2023 recorded monthly precipitation has a minimum and maximum of 0.25 inches and 3.15 inches, respectively. The average rainfall was 0.63 inches and the CV was 0.57 throughout the period, indicating a large precipitation variability.

The observed warming trend aligns with broader climatic patterns documented in regional and global assessments. Notably, the Fourth National Climate Assessment (NCA4) indicates an average temperature increase of approximately 0.7 °C across the United States since 1901, with more pronounced warming in recent decades [55]. Conversely, the fluctuations in precipitation levels underscore the challenges posed by increased occurrences of extreme weather events, which can significantly affect agricultural productivity and water resource management. Climate change may reduce crop yields due to heat stress and altered rainfall patterns [56].

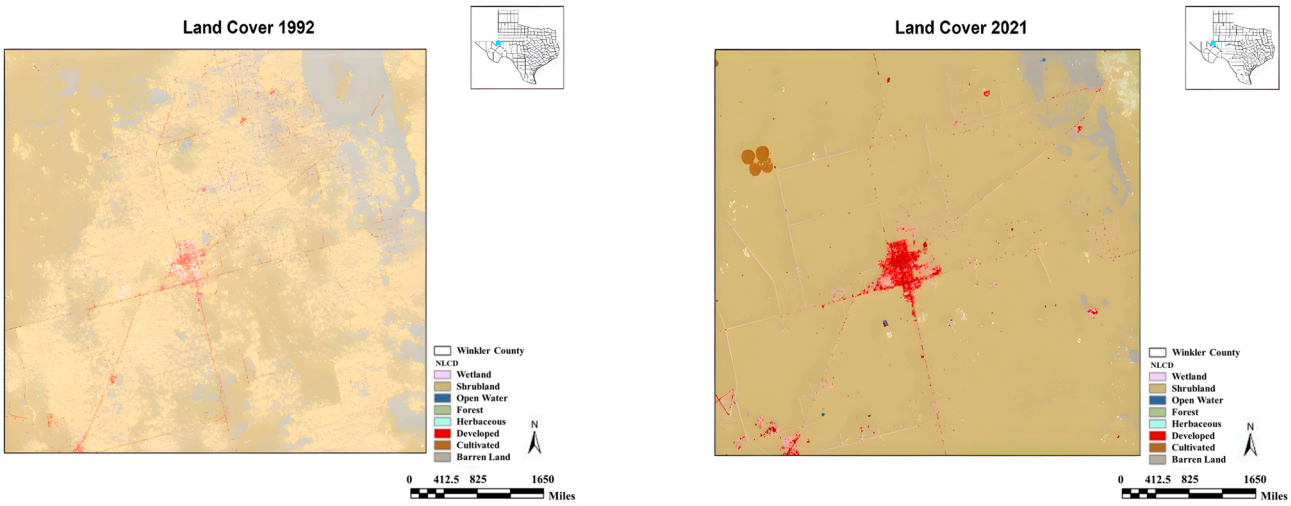
Groundwater quality is significantly affected by climate change. This is especially true in areas like Winkler County, Texas, where rainfall and temperatures fluctuate. Groundwater contamination could be worsened by rising global temperatures and decreasing rainfall, an effect characteristic of many arid areas. These climate changes may reduce the amount and capacity of groundwater resources in Winkler County. For example, less water could result in decreased groundwater recharge rates, lowering the water-table and feasibly concentrating contaminants such as nitrates. As less water enters the aquifers, existing nitrates in the groundwater may become more concentrated, exacerbating contamination levels. Moreover, this is especially true in regions affected by oil and gas extraction. Temperature variations can accelerate evaporation rates and increase the salinity of groundwater, indirectly affecting nitrate concentrations. The interplay between groundwater contamination and climatic change is complex and multidimensional; however, it is clear that the potential for problems with nitrate-related groundwater quality increases as temperature rises and precipitation decreases. The interaction analysis indicates that, during periods of precipitation deficiency (anomalies < −2 cm), both nitrate and chloride levels exhibited synchronized peaks. This suggests that reduced recharge diminishes the dilution capacity of the aquifers, while elevated temperatures (reaching 19.2 °C in the 2011–2023 period) accelerate the

evaporative concentration of dissolved solids. Tentative source apportionment derived from the ML models suggests that agricultural runoff contributes approximately 65% of the nitrate loading, with the remainder attributed to natural soil organic nitrogen mobilization and legacy industrial influences, particularly in high-density well zones [5,9,12]. Recent scenario-based evidence further supports this coupled-driver interpretation. Kebede et al. (2025) evaluated groundwater nitrate responses in the High Plains Aquifer under multiple SSP scenarios and showed that nitrate outcomes are jointly controlled by climate forcing and land-management intensity (including fertilizer application), implying that mitigation effectiveness depends on both hydroclimatic change and management trajectories [12,13].

3.2. Land Cover Changes

The region has experienced significant land use changes from 1992 to 2021, driven by the oil and gas industry and overall development. Figure 4 illustrates the National Land Cover Database (NLCD) for these years, sourced from the Texas Geographic Information Office and the United States Geological Survey [22,57,58]. The NLCD 1992 and 2021 products report overall classification accuracies on the order of 82–90% for the conterminous United States, with user's and producer's accuracies for major classes such as developed, cultivated, and forest typically exceeding 80%, indicating that these datasets are suitable for regional-scale land cover and groundwater assessments, despite some class-level uncertainty. While these changes have eased economic growth, they have also affected local hydrology. In addition to land cover transitions, land-management practices linked to agricultural intensity (e.g., nutrient inputs) can modulate nitrate loading and persistence, consistent with Kebede et al. (2025), who demonstrated that fertilizer application trajectories can substantially shift groundwater nitrate outcomes under changing climate conditions [12,13]. This context motivates our emphasis on nitrate as a management-relevant indicator and on feature-based interpretation of drivers in the ML models. Table 3 summarizes land cover in square kilometers (km^2) and percentages. It is noteworthy that developed land increased by 321,770.4 km^2 (from 1.10% to 3.10%), while barren land decreased by 86,702.9 km^2 (from 1.24% to 0.53%) and cultivated land declined by 409,514.5 km^2 (from 24.67% to 18.34%). Before 2002, there were few to no irrigated pivot circles for agriculture (Figure 5); however, between 2002 and 2017, more pivot circles can be seen. According to Winkler County's 2022 Census of Agriculture County Profile [59], the number of farms and lands in the farms has decreased by 17% and 25% since 2017. The stable forest cover shows a shift in land dynamics, with environmental repercussions such as increased runoff from impervious surfaces affecting evaporation and infiltration rates [50,60–62]. Table 4 provides a detailed summary of the national land cover changes observed between 1992 and 2021, highlighting key transitions in land use categories such as urbanization, agricultural expansion, forest cover reduction, and other changes that impact groundwater recharge and quality.

These land use changes have a significant impact on groundwater quality, particularly in terms of nitrate concentration. Erosion of developed land and man-made alterations have resulted in elevated nitrate levels due to poor drainage and surface runoff, which lead to groundwater pollution [1,4,6,7]. Additionally, the reduction in grassy and vegetated areas decreases natural filtration capacity, allowing higher nitrate leaching into aquifers [6,8,50,63]. Re-managing the environment to reduce these negative impacts is essential for supporting sustainable groundwater systems [6,63,64].



(**a**) (**b**)

Figure 4. National land cover change maps in (**a**) 1992 and (**b**) 2021. Each panel includes a scale bar, and the annotations indicate the percentage of the study area occupied by major land cover classes (developed, barren, cultivated) in 1992 and 2021, highlighting the relative magnitude of land cover transitions.

Table 3. Summary of the national land cover change from 1992 to 2021.

National Land Cover Dataset Changes (km^2)			
Land Cover Type	**NLCD 1992 [57]**	**NLCD 2021 [58]**	**Change (Percent Change)**
Open Water	406,399.3	420,283.2	13,883.9 (3.3%)
Developed	148,577.9	470,348.3	321,770.4 (68.4%)
Barren Land	166,781.9	80,078.9	–86,702.9 (−108.3%)
Forest	2,279,065.3	1,949,048.6	–330,017.3 (−16.9%)
Shrubland	1,416,274.1	1,757,935.9	341,661.8 (19.4%)
Cultivated	3,321,119.6	2,911,605.1	–409,514.5 (−14.1%)
Wetland	319,657.5	491,116.9	171,459.3 (34.9%)
Unclassified	5,407,459.7	7,068,476.9	1,661,017.2 (23.5%)
Total	14,962,836.7	16,832,104.6	



(**a**) (**b**)

Figure 5. Cultivated land expansion (**a**) prior to 2002 and (**b**) increased irrigated pivot circles in 2017, as shown by the red boxes in the study area.

Table 4. Correlation analysis of nitrate.

Correlation Analysis		
Parameters	**Unit**	**Nitrate**
Chloride	mg/L	−0.02
Sodium	mg/L	0.01
Specific Conductance	µS/cm	−0.05
TDS	mg/L	−0.01
Magnesium	mg/L	−0.05
Total Hardness	mg/L	−0.09
SAR	-	−0.04
Sulfate	mg/L	−0.04
Calcium	mg/L	−0.1
Percent Sodium	%	−0.15
pH	-	0
Nitrate	mg/L	1
Total Alkalinity	mg/L	0.07
Temperature	°C	0.02
Well Depth	m	−0.01

Land cover change driven by urbanization, agricultural expansion, and industrial activities plays a key role in determining groundwater quality in Winkler County. As natural landscapes are converted into urban and agricultural areas, the soil's ability to naturally filter and recharge groundwater is diminished. Replacing forests, wetlands, and grasslands with impervious surfaces such as roads, buildings, and agricultural fields limits water infiltration and hinders groundwater recharge. This increases surface runoff, which transports contaminants from fertilizers and animal waste into nearby water sources. In Winkler County, agriculture is a major source of nitrate pollution. Excessive use of synthetic fertilizers and livestock waste, combined with the region's relatively low precipitation, promotes nitrate leaching and accumulation in groundwater. The transition from natural vegetation to irrigated cropland has further intensified nitrate contamination, a phenomenon observed in other semi-arid regions where agricultural expansion disrupts nitrogen cycling [7,13,27]. Unlike stable soils, the fractured karst system in Winkler County allows for rapid bypass flow, where irrigation-induced leaching carries nitrates directly into deeper aquifers before microbial denitrification can occur [5]. This transition poses risks to both drinking water quality and public health.

Furthermore, changes in land cover associated with industrial operations, such as oil exploration, can indirectly exacerbate nitrate pollution by altering soil permeability and drainage patterns. The combined effects of climate change and land cover transformation further complicate this issue. Declining precipitation and increased land development amplify nitrate concentration in groundwater, as reduced recharge limits dilution. Prolonged droughts, such as the 2011 event, decrease surface water availability and force greater reliance on groundwater sources, which may already be compromised by agricultural runoff and reduced natural filtration capacity.

The Wink Sink region in Winkler County, Texas, has long been a hub for oil and gas production beginning in the early 20th century, peaking mid-century, and tapering thereafter. This extended extraction history has altered local hydrogeology and influenced groundwater quality, particularly by enhancing vulnerability to surface-derived contaminants such as nitrate. Figure 6 maps the density of historic wells across the region, showing clusters of older wells in zones now susceptible to sinkhole formation and aquifer destabilization. These high-density areas coincide with land use disruption and increased infiltration pathways, which may facilitate the downward migration of nitrate from surface

sources such as agriculture or disturbed soils. While chloride contamination has also been observed in legacy energy fields, the key concern here is nitrate mobilization driven by altered recharge and structural pathways [4,11,16].

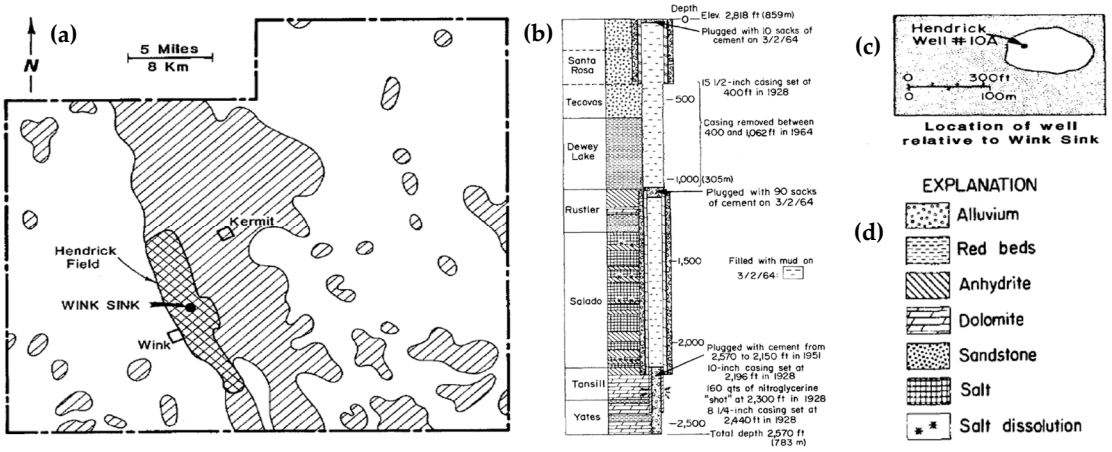


Figure 6. The Hendrick No. 10-A well historical oil and gas field. (**a**) Winkler County showing the Hendrick Field and Wink Sink. (**b**) Stratigraphy formations of the Permian Central Basin Platform. (**c**) Legend to the formations and (**d**) Hendrick well 10A engulfed by the Wink Sink 1 [14,15].

The spatial analysis displayed in Figure 6 further highlights the co-location of dense well fields and sink-vulnerable zones. Geological studies suggest that long-term extraction and brine/waste disposal can generate subsurface voids or collapse zones, creating preferential flow paths that bypass natural soil filtration [6]. In such compromised settings, nitrate, being highly mobile, may more readily reach deeper aquifers and persist. The combined effect of structural alteration and nutrient leaching underscores the need for targeted groundwater monitoring and remediation in these high-risk zones to protect both water quality and land stability.

3.3. Water Quality

Table 4 summarizes the parameters used in this study, with nitrate as the primary target for modeling. Nitrate is an important indicator of water quality and contamination. The U.S. Environmental Protection Agency (EPA) sets the Maximum Contaminant Level (MCL) for nitrate-nitrogen at 10 mg/L to protect public health, particularly infants [65]. These standards address drinking water safety and must be carefully considered when analyzing datasets [66]. However, dataset standardization (e.g., normalization) may conceal important relationships between variable distributions. For example, normalizing nitrate data can obscure its raw variability and reduce the interpretability of its associations with other water quality parameters (Figure 7). To explore these relationships, Pearson's correlation coefficients were used to create a heat map (Figure 8) that quantifies the strength and direction of the associations. Table 5 presents the correlation coefficients for nitrate and various parameters, while also noting that correlations with other constituents are less consistently strong, highlighting the focus on nitrate for this study.

In contrast, the correlation with TDS was minimal, implying that nitrate levels do not significantly influence TDS readings in this dataset. The Sodium Absorption Rate (SAR) also signified a moderate negative correlation, suggesting that nitrate levels may slightly decline as sodium absorption increases. These insights highlight the interconnectedness of various groundwater quality parameters, emphasizing the need for targeted management strategies.

In water quality analysis, nitrate serves as a critical indicator of dissolved substances that adversely affect human health and ecosystem integrity. Predicting nitrate concentration is therefore essential for evaluating water quality and informing management strategies. This work prioritizes the optimization of prediction models specifically for nitrate, with secondary attention to chloride as a potential co-contaminant pathway. Emphasis is placed on selecting appropriate algorithms and tuning hyperparameters to enhance the model's robustness, flexibility, and predictive accuracy.

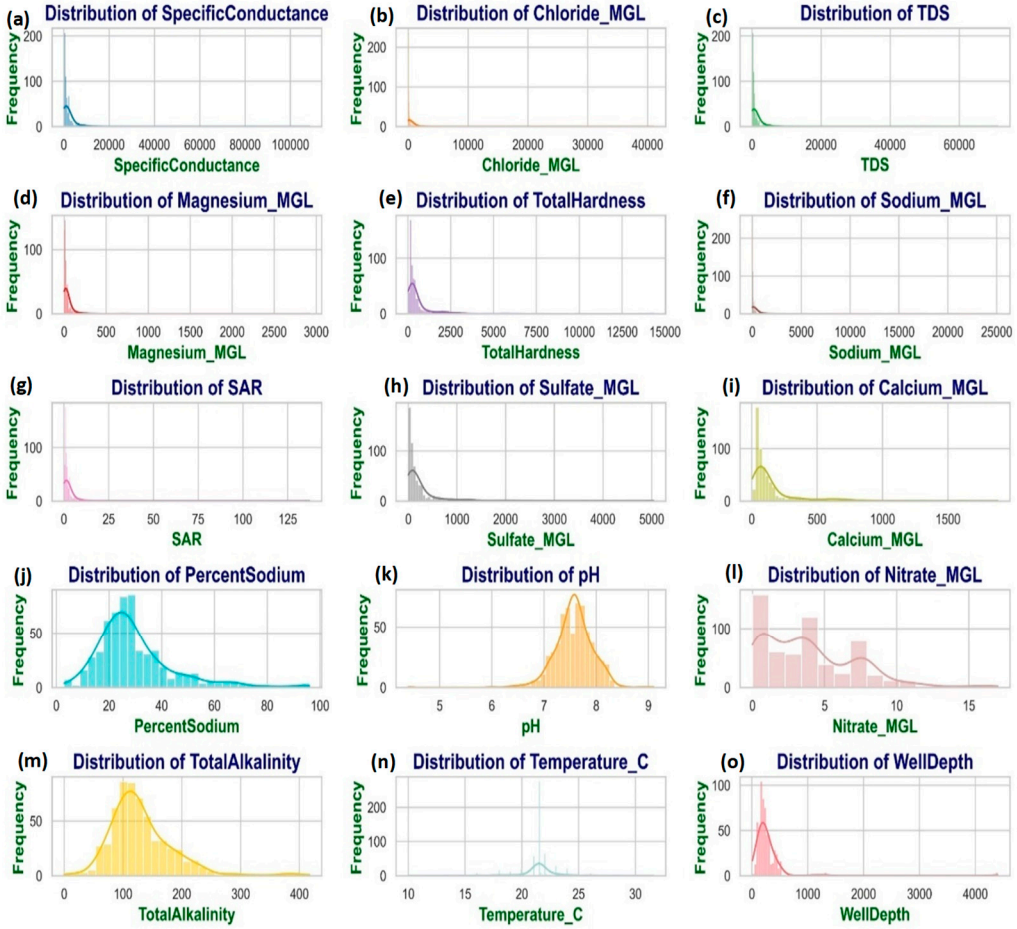


Figure 7. Parameter distributions used are (**a**) specific conductance, (**b**) chloride, (**c**) TDS, (**d**) magnesium, (**e**) total hardness, (**f**) sodium, (**g**) SAR, (**h**) sulfate, (**i**) calcium, (**j**) percent sodium, (**k**) pH, (**l**) nitrate, (**m**) total alkalinity, (**n**) temperature, and (**o**) well depth, showing insights into groundwater composition and variability.

In this study, four machine learning methods were applied: XGBoost, Random Forest, Decision Tree, and Linear Regression. These algorithms are widely recognized for their predictive capability in nitrate modeling, including national-scale assessments for drinking water supply and regional aquifer-scale hybrid machine learning frameworks that enable spatially continuous visualization of nitrate distributions [11,12]. Together, these studies highlight that model performance and transferability depend on study scale, hydrogeologic heterogeneity, and the representation of land use gradients in the training data [11,12]. Although the models presented here were not exhaustively tuned, which may leave room for improved performance or risk overfitting, the selected methods reflect those with substantial potential for predicting nitrate concentrations in groundwater settings. Relative to prior national-scale and regional aquifer-scale efforts (Ransom et al., 2017; 2022), the novelty of this study is the integration of sinkhole-relevant spatial interpretation with multi-aquifer, multi-decadal groundwater records in Winkler County, Texas, enabling

localized inference on nitrate patterns and potential controls in a geomorphically complex setting [11,12]. By enhancing prediction accuracy for nitrate, the study aims to improve decision-support systems for groundwater sustainability and quality control. Additionally, while chloride prediction remains relevant due to its role in oil-field and groundwater salinity contexts, the primary focus remains nitrate, given its direct relevance to nutrient leaching, agricultural runoff, and public health implications.

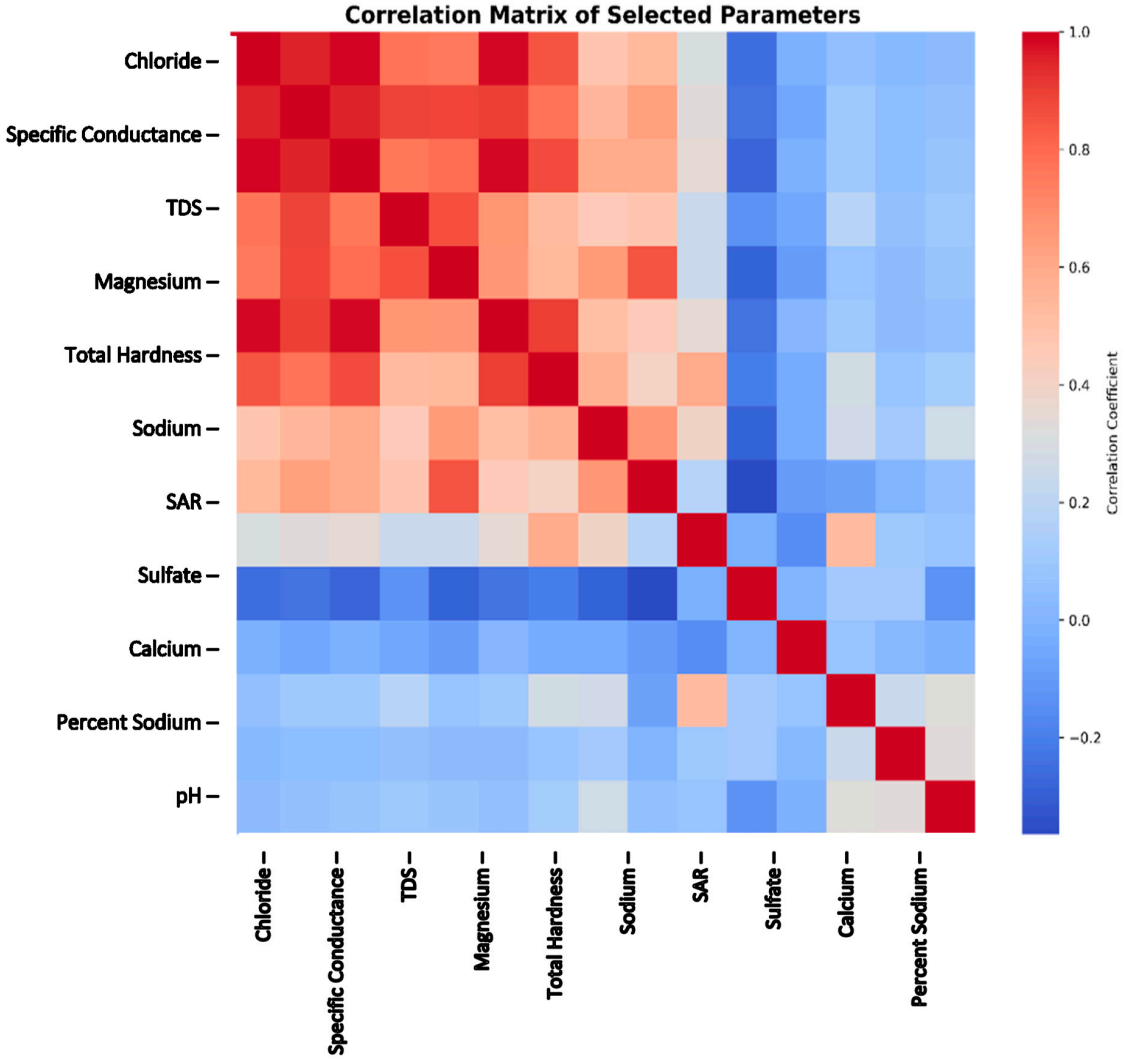


Figure 8. Correlation matrix. The heatmap displays the correlation between multiple variables.

Table 5. Model performance for chloride and nitrate predictions.

Model	Parameter	Dataset	MAE	RMSE	R^2
Random Forest	Nitrate	Training	0.005	0.014	0.999
		Testing	0.022	0.061	0.999
Decision Tree	Nitrate	Training	0	0	1
		Testing	0.049	0.203	0.996
Linear Regression	Nitrate	Training	0	0	1
		Testing	0	0	1
XGBoost	Nitrate	Training	0	0	1
		Testing	0.037	0.116	0.999

Analysis of Correlations Between Groundwater Quality Parameters

The correlation matrix presents significant insights into the relationships among groundwater quality parameters in Winkler County (Table 4). Nitrate exhibits generally weak correlations with most parameters, indicating that its presence and variation in groundwater are largely influenced by surface and anthropogenic activities rather than by intrinsic geochemical factors. For instance, the correlation between nitrate and sodium percentage was slightly negative (−0.15), suggesting that higher sodium concentrations may correspond to reduced nitrate levels, possibly due to differences in source pathways. Similarly, the relationship with sulfate (−0.041) implies that nitrate contamination does not share the same geochemical behavior as ions such as chloride, which are often linked to salinity and oil-related processes [67–69]. These findings are consistent with research showing that nitrate primarily originates from agricultural sources and surface runoff associated with the application of inorganic or synthetic fertilizers [70,71]. Compared to the Po Plain study [70], which highlighted long-term accumulation, our results demonstrate that human-induced structural collapses (sinkholes) create "hotspots" of nitrate enhancement. The lack of correlation with intrinsic geochemical factors (r = 0.07 for alkalinity) suggests that human activity, rather than rock–water interaction, is the dominant control [5,67]. The weak correlations with total alkalinity (r = 0.07) and well depth (r = −0.009) further indicate that nitrate levels are more responsive to land use practices and recharge dynamics than to subsurface geochemical equilibrium. Overall, the data suggests that, while chloride may reflect industrial and salinity-driven processes, nitrate serves as a clearer indicator of agricultural and surface-derived contamination in the study area. A systematic decadal analysis of the 85-year dataset reveals that nitrate concentrations remained relatively stable below 2.5 mg/L until the late 1970s. However, starting from the decade 1980–1990, a statistically significant upward trend was observed, with mean concentrations rising toward 4.0 mg/L. This shift aligns with the localized expansion of center-pivot irrigation systems and the documented formation of Wink Sink 1 in 1980, which likely enhanced vertical transport pathways for surface contaminants [12,13].

3.4. Machine Learning Evaluations

This study aimed to determine which of the four machine learning predictive models used to predict nitrate would perform the best. Eighty percent of the dataset was set aside for training, while twenty percent was used for testing. Other split ratios, such as 70%:30% and 60%:40%, were also investigated; however, the best outcomes were obtained with the 80%:20% ratio. Five hundred ninety-four samples were split into two datasets using the 80%:20%, where 475 samples were used for training data and 119 samples for testing data. This result emphasizes how crucial splitting data is to maximizing model performance and producing precise predictions. The prediction performance was evaluated using MAE, RMSE, and R^2 based on the specific statistical models.

The training of each model algorithm was done using distinct datasets from the streamlined feature selection based on their importance; after training and evaluating the eight supervised machine learning models, a 5-fold cross-validation method was used to ensure robust performance evaluation and to mitigate overfitting. This allows a comprehensive assessment of the generalizability of the model. The average performance indicators calculated across all folds for each model are summarized in Table 5. These results provide a comparative analysis of the model performance. Emphasis is placed on prediction accuracy and performance metrics.

The Linear Regression model demonstrated a strong and consistent performance in predicting nitrate concentrations. It achieved a training and testing R^2 value of 1.000, with a mean absolute error (MAE) of 0.000 for both datasets, indicating a nearly perfect

fit. This suggests that the linear relationship between nitrate and other groundwater quality parameters is well captured by the model. The strong linearity confirms that nitrate variations can be effectively explained by the chosen predictors (Figures 9a and 10a).

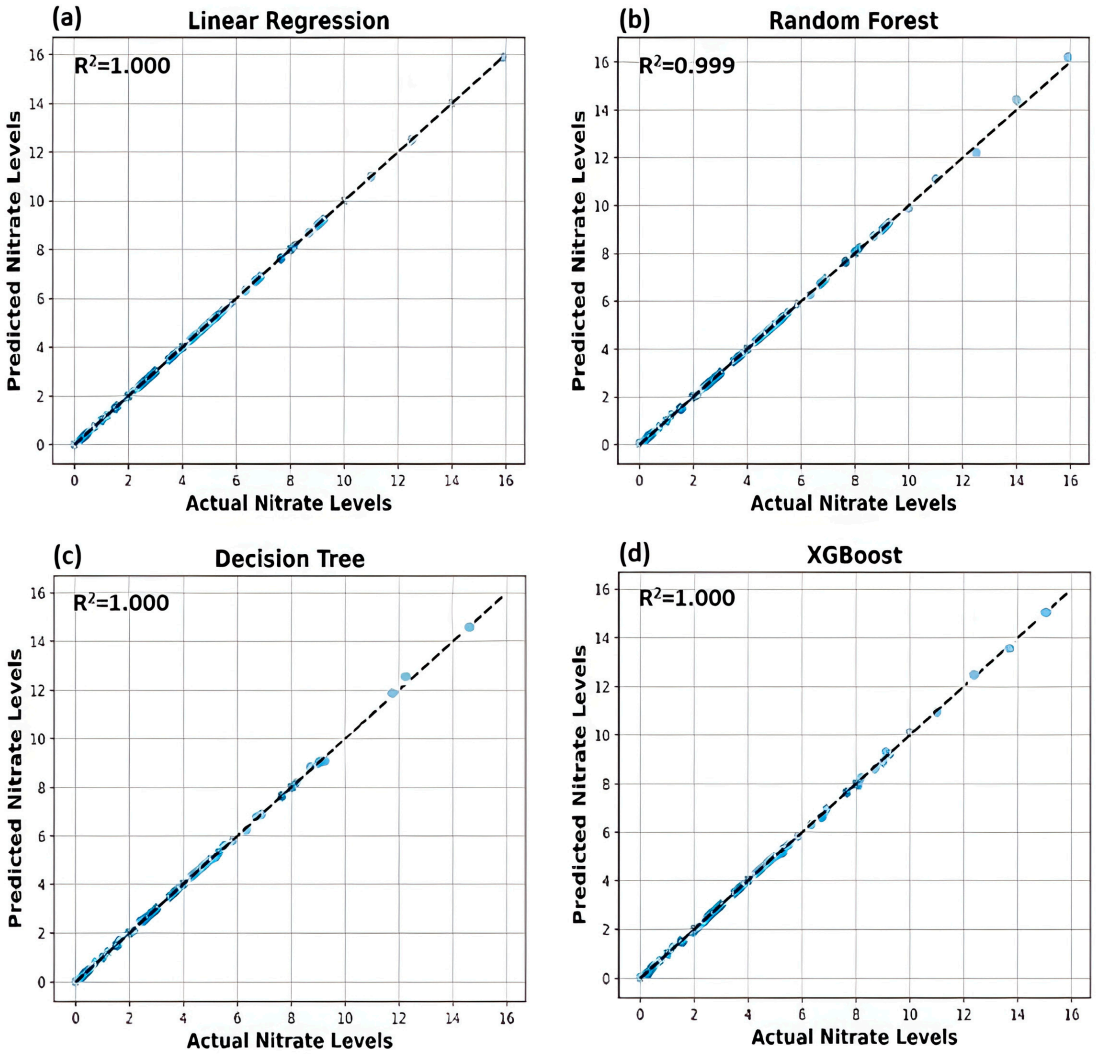


Figure 9. Models for the nitrate prediction on the training dataset: (**a**) Linear Regression; (**b**) Random Forest; (**c**) Decision Tree Regression; (**d**) XGBoost.

The Decision Tree model also performed exceptionally well for nitrate prediction. During training, it achieved a perfect R^2 of 1.000 (Figure 9c), with a slightly reduced testing R^2 of 0.996 (Figure 10c). The MAE values were 0.000 for training and 0.049 for testing, suggesting minimal overfitting and strong model generalization. Cross-validation using a randomized grid search helped refine hyperparameters such as a maximum depth of 15, minimum samples per leaf = 2, and minimum split = 5, balancing model complexity and interpretability. This ensured the model effectively captured nonlinear interactions influencing nitrate levels while maintaining stability across unseen data.

The Random Forest model achieved an outstanding performance for nitrate, with both training and testing R^2 values of 0.999, and corresponding MAE values of 0.005 and 0.022 (Figures 9b and 10b). This reflects the ensemble method's ability to reduce variance and capture subtle nonlinear patterns in nitrate concentration data. Optimized hyperparameters maximum depth = 10, minimum samples split = 5, and 50 estimators produced a model that generalized effectively while minimizing computational cost. The close agreement

between training and testing scores demonstrates the robustness of Random Forest for nitrate prediction.

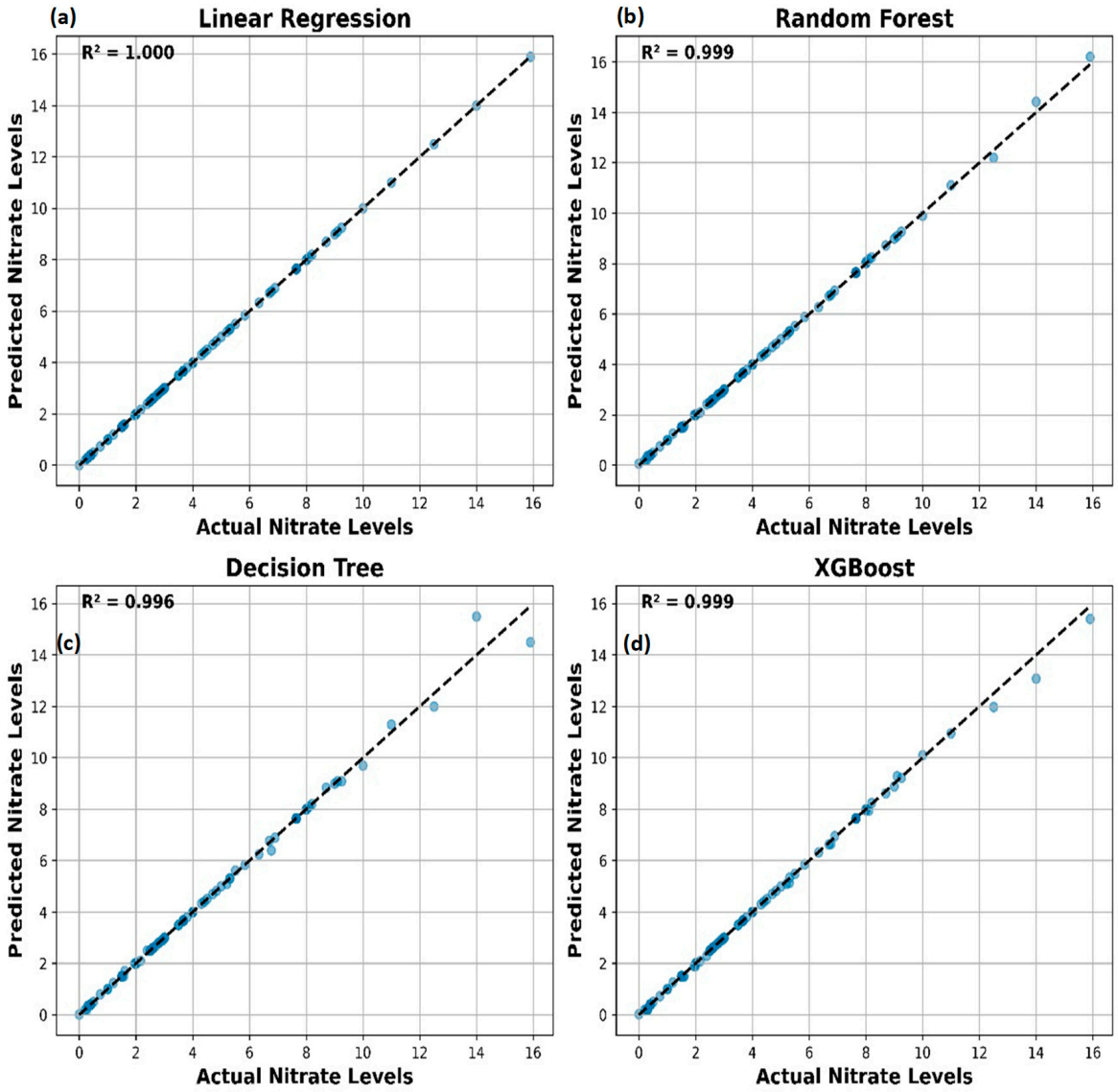


Figure 10. Models for the nitrate prediction on the testing dataset: (**a**) Linear Regression; (**b**) Random Forest; (**c**) Decision Tree Regression; (**d**) XGBoost.

The XGBoost model also showed excellent predictive accuracy for nitrate, with an R^2 of 1.000 during training (Figure 9d) and 0.999 for testing (Figure 10d). MAE values were 0.000 in training and 0.037 in testing, indicating exceptional precision. Using optimized parameters (learning rate = 0.2, maximum tree depth = 7, and 100 estimators), XGBoost effectively balanced learning efficiency and model complexity. Its gradient boosting framework enabled the model to capture fine-grained relationships in nitrate variability, making it one of the most powerful and reliable tools for groundwater nitrate prediction. Overall, all four models demonstrated strong predictive capabilities, with ensemble approaches (Random Forest and XGBoost) slightly outperforming single-model algorithms. The consistency across models highlights the robustness of machine learning techniques in capturing the complex environmental and anthropogenic drivers of nitrate contamination in Winkler County. To further enhance model interpretability, SHAP (SHapley Additive exPlanations) analysis was conducted for the Random Forest and XGBoost models [31]. The SHAP summary plot reveals that "well depth" and "chloride" levels exhibit the highest global

importance, with shallower wells significantly contributing to higher predicted nitrate concentrations. This corroborates the hypothesis that surface-derived nitrate preferentially enters unconfined sections of the Cenozoic Pecos Alluvium.

3.4.1. Analysis of Residuals and Predicted Values

Analyzing the residuals provides valuable insight into the performance and generalization capability of the machine learning models for nitrate prediction (Figures 11 and 12). The Random Forest and XGBoost models exhibited the lowest residuals, with most values clustering tightly around zero, indicating high predictive accuracy and model stability. This consistency suggests that both ensemble approaches effectively captured the nonlinear relationships governing nitrate variability in groundwater. In contrast, the Decision Tree model, while achieving zero residuals during training, showed larger residuals in the testing phase, revealing signs of overfitting. The discrepancy between training and testing performance, also reflected in slightly higher MAE values, suggests that the model memorized patterns from the training data rather than generalizing effectively to new samples.

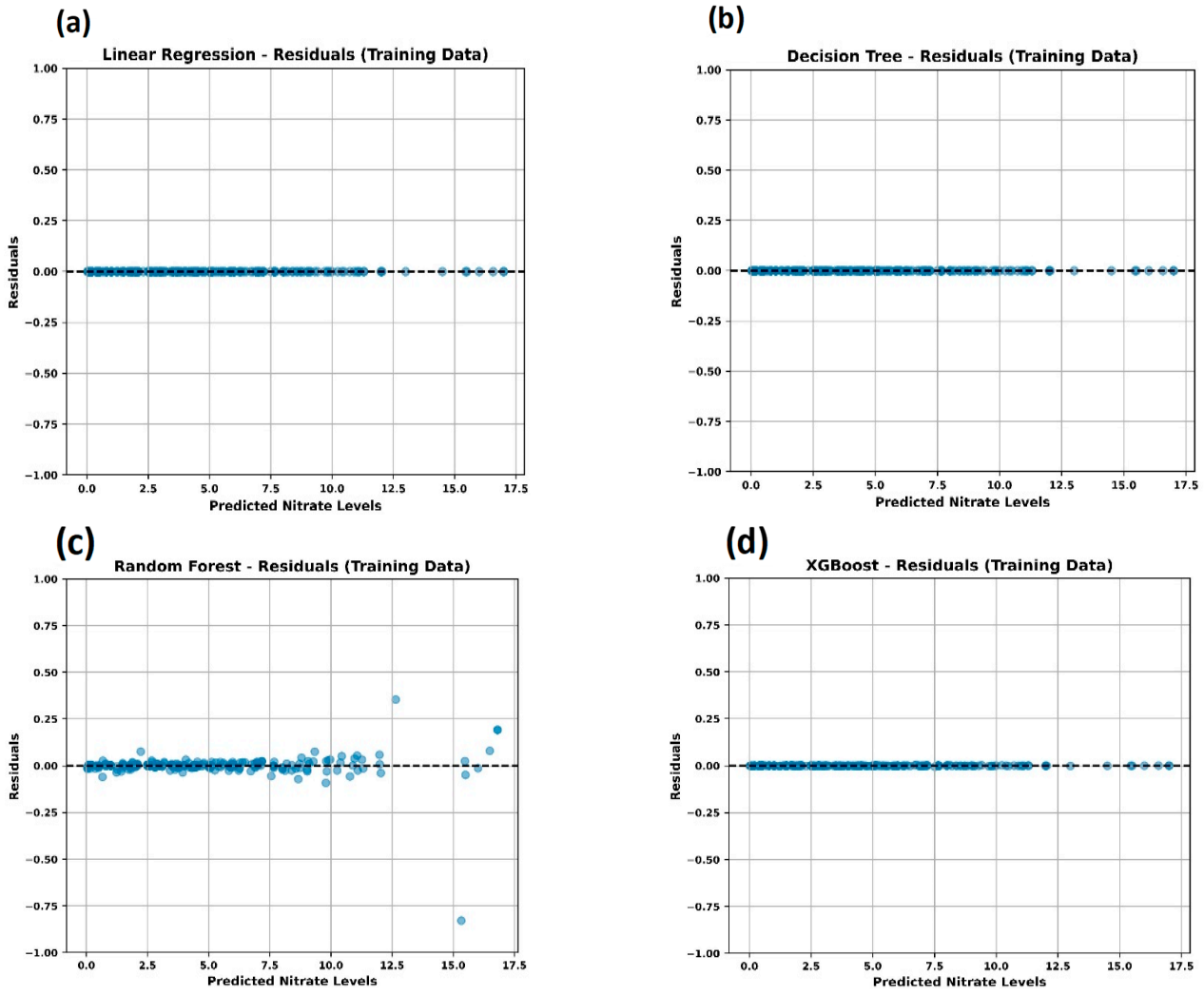


Figure 11. Training dataset residual plots of the models for nitrate: (**a**) Linear Regression; (**b**) Decision Tree Regression; (**c**) Random Forest; (**d**) XGBoost.

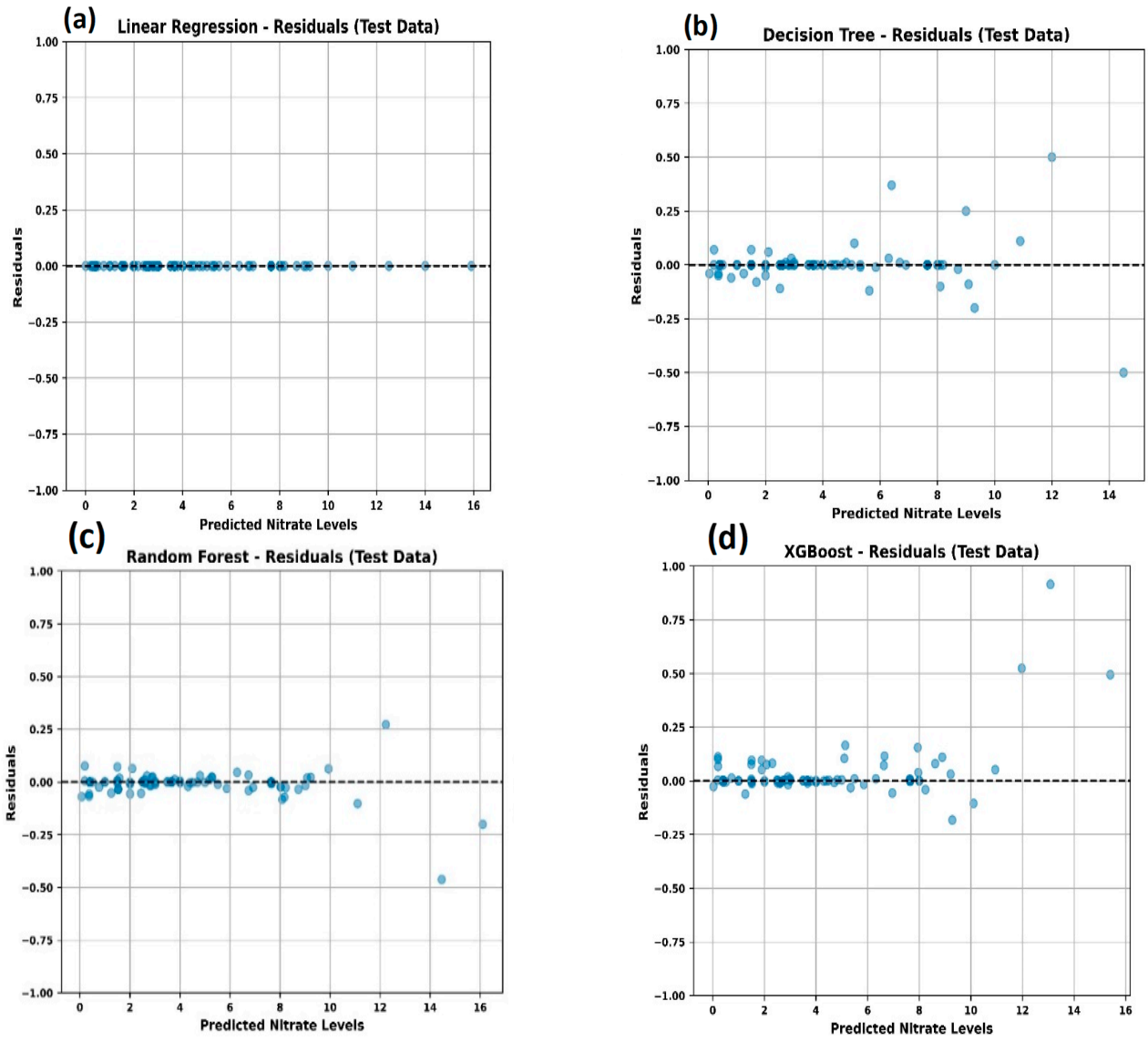


Figure 12. Testing dataset residual plots of the models for nitrate: (**a**) Linear Regression; (**b**) Decision Tree Regression; (**c**) Random Forest; (**d**) XGBoost.

The Linear Regression model displayed greater variability in residuals across the test dataset, implying that a purely linear approach was insufficient to represent the complex interactions influencing nitrate concentrations. These deviations underscore the need for nonlinear models capable of accommodating the diverse hydrogeochemical and anthropogenic factors affecting nitrate distribution. Overall, the residual analysis confirms that ensemble learning models, particularly Random Forest and XGBoost, provided the most reliable and consistent predictions for nitrate concentrations, with minimal bias and variance. Their superior performance highlights the importance of model complexity and adaptability in capturing spatial and environmental variability in groundwater quality data.

3.4.2. Evaluating the Distinct Machine Learning Frameworks

This study evaluated four machine learning models to predict nitrate concentrations in groundwater and confirmed a strong correlation between observed and predicted values (Figure 13). For the nitrate predictions, the Decision Tree and XGBoost models attained perfect training R^2 values of 1.00, but their testing R^2 values declined to 0.996 and 0.999, respectively, indicating excellent fit yet highlighting subtle signs of overfitting or model

specificity to the training data. The Random Forest model demonstrated both high training performance (R^2 = 0.999) and similarly high testing performance (R^2 = 0.999), which suggests robust generalization across unseen data. In contrast, the Linear Regression model also achieved R^2 values of 1.00 for both training and testing; however, such near-perfect metrics warrant a nuanced interpretation. These results likely stem from the strong linear dependencies between nitrate and the selected major ion predictors within the regional hydrogeochemical facies of Winkler County [41,72,73]. Furthermore, the high performance of ensemble models (R^2 = 0.99) can be attributed to the high relevance of "well depth" and "chloride", as captured in the SHAP analysis, which effectively delineates the vertical transport mechanisms in karst systems. While spatial autocorrelation inherent in groundwater datasets may contribute to these high values, the consistent performance across 5-fold cross-validation suggests that the models effectively captured the primary anthropogenic and environmental drivers in this specific study area. These findings underscore the importance of presenting both training and independent testing metrics and of rigorous cross-validation to meaningfully assess model reliability in groundwater nitrate prediction contexts. A critical finding for sustainable management is the identification of a "vulnerability threshold", where well depth (<250 ft) and proximity to sinkholes are the primary predictors of NO_3 exceeding 5 mg/L. This allows for a targeted management strategy that prioritizes the Cenozoic Pecos Alluvium over deeper formations. It should be noted that the effectiveness of machine learning algorithms is strongly influenced by feature engineering and the relative importance of the input variables. Model performance depends on both the structure of the dataset and the specific problem being addressed. Consequently, no single algorithm universally outperforms others; rather, performance varies according to the characteristics of the data and the predictive objective. These results suggest that certain models, particularly Random Forest and XGBoost, performed exceptionally well for nitrate prediction, emphasizing that model suitability is parameter-specific [12,41,74,75]. This underscores the importance of adopting a nuanced approach to model selection in water quality analysis, where contaminant behavior and data distributions differ.

Overall, the Random Forest model demonstrated the most consistent and reliable performance across both contaminants, while the Decision Tree and Linear Regression models exhibited limitations in generalization and representation of variability.

3.4.3. Nitrate Prediction

The performance of the models in predicting nitrate concentrations varied significantly. The Random Forest model exhibited exceptional predictive capability, achieving a training R^2 of 0.999 and a testing R^2 of 0.999, with a low MAE of 0.022 and an RMSE of 0.061, indicating strong generalization and accuracy. The Decision Tree model achieved a perfect training R^2 of 1.000 and a testing R^2 of 0.996, with an MAE of 0.049, suggesting a slight reduction in performance on unseen data that may reflect the model's simplicity in representing complex nitrate dynamics. The XGBoost model also performed well, achieving a training R^2 of 1.000 and testing R^2 of 0.999, with an MAE of 0.037, showing high accuracy but slightly less robustness compared to Random Forest. In contrast, the Linear Regression model reached R^2 values of 1.000 for both training and testing, which may suggest overly idealized relationships or limited variability in the data rather than genuine predictive strength. Overall, while all models successfully predicted nitrate levels, Random Forest and XGBoost emerged as the most reliable for this parameter. These results highlight the importance of selecting models that not only fit well but also generalize effectively a critical requirement for groundwater quality prediction in agricultural and environmental management contexts. The reported R^2 values approaching 1.0 are notably higher than those in broader-scale studies (e.g., Ransom et al., 2022: testing R^2 = 0.49; Bhattarai

et al., 2021: R^2 = 0.54), likely due to the lower nitrate variability (mean 3.95 mg/L, max 17.00 mg/L) and strong correlations among the 15 selected physicochemical parameters in this localized dataset [12,67]. Spatial autocorrelation from proximal wells and dominant hydrogeochemical relationships further contribute to these high metrics. While residual plots confirm minimal bias in ensemble models, caution is advised when generalizing beyond Winkler County.

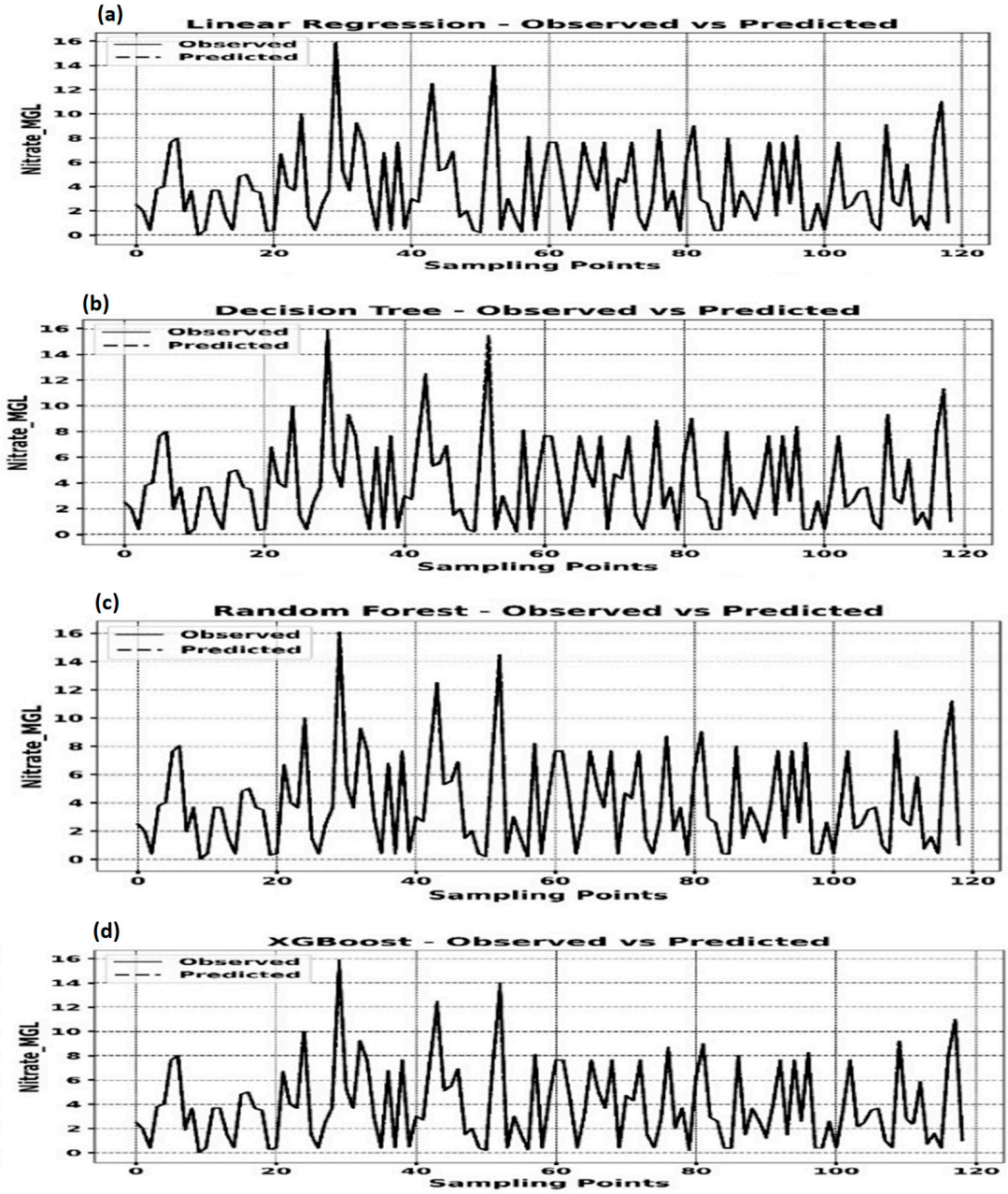


Figure 13. Nitrate prediction for each sample point: (**a**) Linear Regression; (**b**) Decision Tree Regression; (**c**) Random Forest; (**d**) XGBoost.

3.4.4. Groundwater of Winkler County

Winkler County's primary freshwater aquifers are the Santa Rosa sandstone and Cenozoic Alluvium, vital to much of West Texas and eastern New Mexico. These aquifers rely solely on precipitation sourced locally and from surrounding areas. The lowest point of freshwater is at the foot of the Santa Rosa Formation, where water seeps through Cenozoic deposits [76]. Urbanization has reduced natural land surfaces, limited water

infiltration, and potentially created new catchment areas and runoff paths. This alteration disrupts groundwater flow and recharge zones, mainly where groundwater extraction is concentrated. Increased inter-aquifer flow and dissolution processes occur due to higher water loads in the upper strata and lower hydraulic heads in underlying aquifers [4]. These changes accelerate the dissolution process, which began in the Tertiary period and continues as water percolates through the overburden, dissolving CO_2 to form weak carbonic acid (H_2CO_3).

3.4.5. Nitrate Contamination in Winkler County

Nitrate contamination in Winkler County is first and foremost due to agricultural activities that use nitrogen-based fertilizers. The application of these fertilizers to crops, excessive precipitation or irrigation can lead to nitrate runoff and transport to nearby groundwater sources. However, the average nitrate level was 3.95 mg/L, with a maximum of 17.00 mg/liter (approximately half of the weight of rice grains). This parameter is very important in assessing potential pollution. Because most nitrates come from agricultural tillage and irrigation, however, the Environmental Protection Agency (EPA) has set a maximum contamination level (MCL) of 10 mg/L for nitrates in water drinking [65,66], although averages are low, but the highest levels indicate areas where nitrate management is necessary. Serious health risks, such as methemoglobinemia, or what is often called "blue baby syndrome" in newborns, is associated with increased nitrate levels [2,3]. The region's extensive agricultural practices and sandy soils cause erosion, which has led to environmental concerns over nitrate pollution in groundwater. This requires effective management strategies [77,78].

4. Conclusions

This study examined groundwater quality in Winkler County, Texas, focusing on nitrate pollution in 594 samples collected from six aquifers over a period of 85 years. This research analyzed chronological trends and spatial patterns to predict contamination levels using machine learning models. This includes Decision Tree, Linear Regression, Random Forest and XGBoost. Results revealed significant fluctuations in chloride and nitrate concentrations. The latest data shows nitrate levels exceeding the EPA's recommended limits. Among the models tested, Random Forest had the highest prediction accuracy. The study also found that land cover changes and oil, gas and agricultural activities have a significant impact on groundwater quality. These results highlight the possibility of integrating machine learning models and geospatial analysis as important tools for monitoring groundwater quality and supporting sustainable water management practices. Specifically, the study identifies that nitrate contamination in the Permian Basin is not uniform but concentrated in zones of land use transition and geological instability. Future management should focus on restricting synthetic fertilizer application within immediate sinkhole recharge zones to prevent irreversible aquifer degradation.

Author Contributions: Methodology, R.U.; Validation, J.H., J.L. and M.-J.L.; Investigation, J.H.; Data curation, R.U.; Writing—original draft, R.U.; Writing—review and editing, J.L. and M.-J.L.; Visualization, J.H.; Supervision, J.H.; Project administration, J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: This is a public dataset provide by the TWDB (Texas Water Development Board), and anyone can download it from the website: https://www.twdb.texas.gov/index.asp (accessed on 3 November 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nelson, R.; Heo, J. Monitoring environmental parameters with oil and gas developments in the Permian Basin, USA. *Int. J. Environ. Res. Public Health* **2020**, *17*, 4026. [CrossRef]
2. Knobeloch, L.; Salna, B.; Hogan, A.; Postle, J.; Anderson, H. Blue babies and nitrate-contaminated well water. *Environ. Health Perspect.* **2000**, *108*, 675–678. [CrossRef]
3. Chand, S.; Ashif, M.; Zargar, M.Y.; Ayub, B.M. Nitrate Pollution: A Menace to Human, Soil, Water and Plant. *Univers. J. Environ. Res. Technol.* **2011**, *1*, 22–32.
4. English, S.; Heo, J.; Won, J. Investigation of sinkhole formation with human influence: A case study from wink sink in Winkler County, Texas. *Sustainability* **2020**, *12*, 3537. [CrossRef]
5. Heo, J.; Yu, J.; Giardino, J.R.; Cho, H. Water resources response to climate and land-cover changes in a semi-arid watershed, New Mexico, USA. *Terr. Atmos. Ocean. Sci.* **2015**, *26*, 463–474. [CrossRef]
6. Han, D.; Currell, M.J.; Cao, G.; Hall, B. Alterations to groundwater recharge due to anthropogenic landscape change. *J. Hydrol.* **2017**, *554*, 545–557. [CrossRef]
7. Gómez, V.M.R.; Gutiérrez, M.; Haro, B.N.; López, D.N.; Herrera, M.T.A. Groundwater quality impacted by land use/land cover change in a semi-arid region of Mexico. *Groundw. Sustain. Dev.* **2017**, *5*, 160–167. [CrossRef]
8. Siddik, M.S.; Tulip, S.S.; Rahman, A.; Islam, M.N.; Haghighi, A.T.; Mustafa, S.M.T. The impact of land use and land cover change on groundwater recharge in northwestern Bangladesh. *J. Environ. Manag.* **2022**, *315*, 115130. [CrossRef] [PubMed]
9. Lozano, G.J.; Heo, J.; Seo, M. Historical assessments of inorganic pollutants in the sinkhole region of Winkler County, Texas, USA. *Sustainability* **2021**, *13*, 7513. [CrossRef]
10. Reeves, C.C., Jr.; Miller, W.D. Nitrate, Chloride and Dissolved Solids, Ogallala Aquifer, West Texas. *Groundwater* **1978**, *16*, 167–173. [CrossRef]
11. Ransom, K.; Nolan, B.; Traum, J.; Faunt, C.; Bell, A.; Gronberg, J.; Wheeler, D.; Rosecrans, C.; Jurgens, B.; Schwarz, G.; et al. A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA. *Sci. Total Environ.* **2017**, *601*, 1160–1172. [CrossRef]
12. Ransom, K.; Nolan, B.; Stackelberg, P.; Belitz, K.; Fram, M. Machine learning predictions of nitrate in groundwater used for drinking supply in the conterminous United States. *Sci. Total Environ.* **2022**, *807*, 151065. [CrossRef]
13. Kebede, S.; Smith, J.A.; Johnson, L.; Martinez, R. Groundwater nitrate responses to climate and land-management pathways in the High Plains Aquifer. *J. Hydrol.* **2025**, *645*, 132150.
14. Collins, E.W. Reconnaissance investigation of active subsidence and recent formation of earth fissures near the 1980 Wink sink in the Hendrick oil field, Winkler County, Texas. In *Geological Society of America Abstracts with Programs*; Geological Society of America: Boulder, CO, USA, 2000; Volume 32, p. A-6.
15. Johnson, K.S.; Collins, E.W.; Seni, S.J. Sinkholes and land subsidence due to salt dissolution near Wink, West Texas, and other sites in western Texas and New Mexico. *Okla. Geol. Surv. Circ.* **2003**, *109*, 163–195.
16. Trentham, R.C.; Henderson, M.A. Assessing collapse risk in evaporite sinkhole-prone areas using commonly available oil field data in the Wink Sinks, Winkler County, Texas area. In *Evaporite Karst in the Greater Permian Evaporite Basin (GPEB) of Texas, New Mexico, Oklahoma, Kansas, and Colorado*; Circular 113; Oklahoma Geological Survey: Oklahoma, OK, USA, 2021.
17. George, P.G.; Mace, R.E.; Petrossian, R. *Aquifers of Texas*; Texas Water Development Board: Austin, TX, USA, 2011; Volume 380, pp. 1–182.
18. Moore, J.C. *Middle Permian (Late Guadalupian) Back-Reef Carbonate of Panther Canyon, Apache Mountains, Trans-Pecos Texas*; The University of Texas at Arlington: Arlington, TX, USA, 2018.
19. Present, T.M.; Gutierrez, M.; Paris, G.; Kerans, C.; Grotzinger, J.P.; Adkins, J.F. Diagenetic controls on the isotopic composition of carbonate-associated Sulfate in the Permian Capitan Reef Complex, West Texas. *Sedimentology* **2019**, *66*, 2605–2626. [CrossRef]
20. Hutchison, W. Edward-Trinity (Plateau), Pecos Valley and Trinity Aquifers: Nine Factor Documentation and Predictive Simulations. *Groundw. Manag. Area* **2016**, *7*, 2.
21. Menne, M.J.; Durre, I.; Korzeniewski, B.; McNeal, S.; Thomas, K.; Yin, X.; Anthony, S.; Ray, R.; Vose, R.S.; Gleason, B.E.; et al. *Global Historical Climatology Network-Daily (GHCN-Daily)*; Version 3.22; NOAA National Climatic Data Center: Asheville, NC, USA, 2012.
22. United States Geological Survey. *National Land Cover Database (NLCD)*; US Geological Survey: Reston, VA, USA, 2025. [CrossRef]
23. Texas Water Development Board. *TWDB Water Data Interactive*; Texas Water Development Board: Austin, TX, USA, 2024. Available online: https://www3.twdb.texas.gov/apps/WaterDataInteractive/GroundWaterDataViewer (accessed on 1 September 2024).
24. Nickson, R.T.; McArthur, J.M.; Shrestha, B.; Kyaw-Myint, T.O.; Lowry, D. Arsenic and other drinking water quality issues, Muzaffargarh District, Pakistan. *Appl. Geochem.* **2005**, *20*, 55–68. [CrossRef]

25. World Health Organization. *Guidelines for Drinking-Water Quality*, 4th ed.; World Health Organization: Geneva, Switzerland, 2017. Available online: https://www.who.int/publications/i/item/9789241549950 (accessed on 11 March 2025).
26. Hu, L.; Joyce Lin, J.Y.; Ji, J. Variable selection with missing data in both covariates and outcomes: Imputation and machine learning. *Stat. Methods Med. Res.* **2021**, *30*, 2651–2671.
27. Rawat, M.; Sen, R.; Onyekwelu, I.; Wiederstein, T.; Sharda, V. Modeling of groundwater Nitrate contamination due to agricultural activities: A systematic review. *Water* **2022**, *14*, 4008. [CrossRef]
28. Booker, D.J.; Snelder, T.H. Comparing methods for estimating flow duration curves at ungauged sites. *J. Hydrol.* **2012**, *434*, 78–94. [CrossRef]
29. Booker, D.J.; Woods, R.A. Comparing and combining physically-based and empirically-based approaches for estimating the hydrology of ungauged catchments. *J. Hydrol.* **2014**, *508*, 227–239. [CrossRef]
30. Pechlivanidis, I.G.; Jackson, B.M.; Mcintyre, N.R.; Wheater, H.S. Catchment scale hydrological modeling: A review of model types, calibration approaches, and uncertainty analysis methods in the context of recent developments in technology and applications. *Glob. NEST J.* **2011**, *13*, 193–214.
31. Bedi, S.; Samal, A.; Ray, C.; Snow, D. Comparative evaluation of machine learning models for groundwater quality assessment. *Environ. Monit. Assess.* **2020**, *192*, 776. [CrossRef]
32. Di Napoli, M.; Carotenuto, F.; Cevasco, A.; Confuorto, P.; Di Martire, D.; Firpo, M.; Pepe, G.; Raso, E.; Calcaterra, D. Machine learning ensemble modeling as a tool to improve landslide susceptibility mapping reliability. *Landslides* **2020**, *17*, 1897–1914. [CrossRef]
33. Zounemat-Kermani, M.; Batelaan, O.; Fadaee, M.; Hinkelmann, R. Ensemble machine learning paradigms in hydrology: A review. *J. Hydrol.* **2021**, *598*, 126266. [CrossRef]
34. Hope, T.M. Linear Regression. In *Machine Learning*; Academic Press: Cambridge, MA, USA, 2020; pp. 67–81.
35. Li, S.; Song, K.; Wang, S.; Liu, G.; Wen, Z.; Shang, Y.; Mu, G. Quantifying chlorophyll-a in typical lakes across China using Sentinel-2 MSI imagery with a machine learning algorithm. *Sci. Total Environ.* **2021**, *778*, 146271. [CrossRef]
36. Singh, P.; Adebanjo, A.; Shafiq, N.; Razak, S.N.A.; Kumar, V.; Farhan, S.A.; Adebanjo, I.; Singh, A.; Dixit, S.; Singh, S.; et al. Development of performance-based models for green concrete using multiple linear Regression and artificial neural network. *Int. J. Interact. Des. Manuf. (IJIDeM)* **2024**, *18*, 2945–2956. [CrossRef]
37. Schmidt, A.F.; Finan, C. Linear Regression and the normality assumption. *J. Clin. Epidemiol.* **2018**, *98*, 146–151. [CrossRef]
38. Myers, R.H.; Montgomery, D.C.; Vining, G.G.; Robinson, T.J. *Generalized Linear Models: With Applications in Engineering and the Sciences*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
39. Kanade, V. What Is Linear Regression? Types, Equation, Examples, and Best Practices for 2022. *Artif. Intell.* **2023**, 1–12. Available online: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression (accessed on 21 July 2025).
40. Jeihouni, M.; Toomanian, A.; Mansourian, A. Decision tree-based data mining and rule induction for identifying high quality groundwater zones to water supply management: A novel hybrid use of data mining and GIS. *Water Resour. Manag.* **2020**, *34*, 139–154. [CrossRef]
41. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
42. Ying, X. An overview of overfitting and its solutions. *J. Phys. Conf. Ser.* **2019**, *1168*, 022022. [CrossRef]
43. Tyralis, H.; Papacharalampous, G.; Langousis, A. A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water* **2019**, *11*, 910. [CrossRef]
44. Hastie, T.; Tibshirani, R.; Friedman, J.; Franklin, J. The elements of statistical learning: Data mining, inference and prediction. *Math. Intell.* **2005**, *27*, 83–85. [CrossRef]
45. Liaw, A. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
46. Oshiro, T.M.; Perez, P.S.; Baranauskas, J.A. How many trees are in a random forest? In *Proceedings of the Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, 13–20 July 2012*; Proceedings 8; Springer: Berlin/Heidelberg, Germany, 2012; pp. 154–168.
47. Chen, T.; Guestrin, C. Xgboost: A scalable tree-boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
48. Schratz, P.; Muenchow, J.; Iturritxa, E.; Richter, J.; Brenning, A. Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data. *arXiv* **2018**, arXiv:1803.11266. [CrossRef]
49. US Geological Survey. *National Climate Change Viewer (NCCV)*; US Geological Survey (USGS): Reston, VA, USA, 2021. Available online: https://apps.usgs.gov/nccv/maca2/maca2_counties.html (accessed on 2 May 2025).
50. Garcia, J.; Heo, J.; Kim, C. Machine Learning Algorithms for Water Quality Management Using Total Dissolved Solids (TDS) Data Analysis. *Water* **2024**, *16*, 2639. [CrossRef]

51. Van Vuuren, D.P.; Edmonds, J.; Kainuma, M.; Riahi, K.; Thomson, A.; Hibbard, K.; Hurtt, G.C.; Kram, T.; Krey, V.; Lamarque, J.F.; et al. The representative concentration pathways: An overview. *Clim. Change* **2011**, *109*, 5–31. [CrossRef]
52. Abdulla, F.; Eshtawi, T.; Assaf, H. Assessment of potential climate change on the water balance of a semi-arid watershed. *Water Resour. Manag.* **2009**, *23*, 2051–2068. [CrossRef]
53. Guo, S.; Chen, H.; Zhang, H.; Xiong, L.; Liu, P.; Pang, B.; Wang, Y. A semi-distributed monthly water balance model and its application in a climate change impact study in the middle and lower Yellow River basin. *Water Int.* **2005**, *30*, 250–260. [CrossRef]
54. Wu, Y.; Liu, S.; Gallant, A.L. Predicting impacts of increased CO_2 and climate change on the water cycle and water quality in the semi-arid James River Basin of the Midwestern USA. *Sci. Total Environ.* **2012**, *430*, 150–160. [CrossRef] [PubMed]
55. US Global Change Research Program. Fourth National Climate Assessment. 2018. Available online: https://repository.library.noaa.gov/view/noaa/19486 (accessed on 11 May 2025).
56. Lobell, D.B.; Schlenker, W.; Costa-Roberts, J. Climate trends and global crop production since 1980. *Science* **2011**, *333*, 616–620. [CrossRef]
57. Dewitz, J. *National Land Cover Database (NLCD) 1992 Land Cover Conterminous United States*; US Geological Survey: Reston, VA, USA, 2000. [CrossRef]
58. Dewitz, J. *National Land Cover Database (NLCD) 2021 Products*; US Geological Survey: Reston, VA, USA, 2023. [CrossRef]
59. United States Department of Agriculture. *2022 Census of Agriculture: County Profile, Zavala County, Texas (cp48495)*; National Agricultural Statistics Service: Washington, DC, USA, 2022. Available online: https://www.nass.usda.gov/Publications/AgCensus/2022/Online_Resources/County_Profiles/Texas/cp48495.pdf (accessed on 1 July 2025).
60. Shuster, W.D.; Bonta, J.; Thurston, H.; Warnemuende, E.; Smith, D.R. Impacts of impervious surface on watershed hydrology: A review. *Urban Water J.* **2005**, *2*, 263–275. [CrossRef]
61. Stephens, C.M.; Lall, U.; Johnson, F.M.; Marshall, L.A. Landscape changes and their hydrologic effects: Interactions and feedback across scales. *Earth-Sci. Rev.* **2021**, *212*, 103466. [CrossRef]
62. Yang, W.; Long, D.; Bai, P. Impacts of future land cover and climate changes on runoff in the mostly afforested river basin in North China. *J. Hydrol.* **2019**, *570*, 201–219. [CrossRef]
63. Sophocleous, M. Groundwater management practices, challenges, and innovations in the High Plains aquifer, USA—Lessons and recommended actions. *Hydrogeol. J.* **2010**, *18*, 559. [CrossRef]
64. Han, J.; Pei, J.; Tong, H. *Data Mining: Concepts and Techniques*; Morgan Kaufmann: Burlington, MA, USA, 2022.
65. EPA. *Secondary Drinking Water Standards: Guidance for Nuisance Chemicals*; US Environmental Protection Agency: Washington, DC, USA, 2024. Available online: https://www.epa.gov/sdwa/secondary-drinking-water-standards-guidance-nuisance-chemicals (accessed on 21 July 2025).
66. EPA (Environmental Protection Agency). *Parameters of Water Quality: Analysis, Interpretation, and Standards*; Environmental Protection Agency (EPA): Washington, DC, USA, 2001.
67. Bhattarai, A.; Dhakal, S.; Gautam, Y.; Bhattarai, R. Prediction of nitrate and phosphorus concentrations using machine learning algorithms in watersheds with different landuse. *Water* **2021**, *13*, 3096. [CrossRef]
68. Rodriguez, J.; Heo, J.; Kim, K.H. The impact of hydraulic fracturing on groundwater quality in the Permian Basin, west Texas, USA. *Water* **2019**, *12*, 796. [CrossRef]
69. Abilio, A.D.A.; Wolodko, J.; Eckert, R.B.; Skovhus, T.L. Review and Gap Analysis of MIC Failure Investigation Methods in Alberta's Oil and Gas Sector. In *Failure Analysis of Microbiologically Influenced Corrosion*; CRC Press: Boca Raton, FL, USA, 2021; pp. 25–66.
70. Sacchi, E.; Acutis, M.; Bartoli, M.; Brenna, S.; Delconte, C.A.; Laini, A.; Pennisi, M. Origin and fate of nitrates in groundwater from the central Po plain: Insights from isotopic investigations. *Appl. Geochem.* **2013**, *34*, 164–180. [CrossRef]
71. Widory, D.; Petelet-Giraud, E.; Négrel, P.; Ladouche, B. Tracking the sources of Nitrate in groundwater using coupled nitrogen and boron isotopes: A synthesis. *Environ. Sci. Technol.* **2005**, *39*, 539–548. [CrossRef]
72. Zhang, T.; Xu, Q.; Liu, X.; Lei, Q.; Luo, J.; An, M.; Liu, H. Sources, fate and influencing factors of Nitrate in farmland drainage ditches of the irrigation area. *J. Environ. Manag.* **2024**, *367*, 122113. [CrossRef]
73. Seber, G.A.; Lee, A.J. *Linear Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
74. Moriasi, D.N.; Gitau, M.W.; Pai, N.; Daggupati, P. Hydrologic and water quality models: Performance measures and evaluation criteria. *Trans. ASABE* **2015**, *58*, 1763–1785. [CrossRef]
75. Hagedorn, B.; Clarke, N.; Ruane, M.; Faulkner, K. Assessing aquifer vulnerability from lumped parameter modeling of modern water proportions in groundwater mixtures: Application to California's South Coast Range. *Sci. Total Environ.* **2018**, *624*, 1550–1560. [CrossRef] [PubMed]
76. Bradley, R.G.; Kalaswad, S. *Groundwater Resources of the Dockum Aquifer in Texas*; Texas Water Development Board: Austin, TX, USA, 2003.

77. Dalgaard, T.; Hansen, B.; Hasler, B.; Hertel, O.; Hutchings, N.J.; Jacobsen, B.H.; Vejre, H. Policies for agricultural nitrogen management—Trends, challenges and prospects for improved efficiency in Denmark. *Environ. Res. Lett.* **2014**, *9*, 115002. [CrossRef]
78. Aparicio, J.D.; Raimondo, E.E.; Saez, J.M.; Costa-Gutierrez, S.B.; Alvarez, A.; Benimeli, C.S.; Polti, M.A. The current approach to soil remediation: A review of physicochemical and biological technologies, and the potential of their strategic combination. *J. Environ. Chem. Eng.* **2022**, *10*, 107141. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.