Article

Not peer-reviewed version

Groundwater Level Prediction with Machine Learning to Support Sustainable Irrigation in Water Scarcity Regions

Wanru Li *, Mekuanent Muluneh Finsa, Kathryn Blackmond Laskey, Paul Houser, Rupert Douglas-Bate

Posted Date: 18 September 2023

doi: 10.20944/preprints202309.1165.v1

Keywords: machine learning; groundwater table; ground water level; sustainable irrigation; drinking water; water-scarcity regions; AI; gradient boosting regression

Article

Groundwater Level Prediction with Machine Learning to Sup-port Sustainable Irrigation in Water Scarcity Regions

Wanru Li [1], *, Mekuanent Muluneh Finsa [2, 5], Kathryn B. Laskey [1], Paul Houser [3] and Rupert Douglas-Bate [4]

[1] Department of System Engineering and Operational Research, George Mason University, Fairfax, VA 22030, USA; klaskey@gmu.edu
[2] Institute of Hydrogeology, Engineering Geology and Applied Geophysics, Charles University, Czechia; finsa@natur.cuni.cz
[3] Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA 22030, USA; phouser@gmu.edu
[4] Global MapAid, United Kingdom; rupertdouglasbate@globalmapaid.org
[5] Water Resource Research Center, Arba Minch University, Arba Minch, Ethiopia; mekuanent.muluneh@amu.edu.et
* Correspondence: wli15@gmu.edu.

Abstract: In water scarcity regions, using data-driven approaches to predict groundwater level is challenging due to limited data availability. However, these regions have substantial water needs and require cost-effective groundwater utilization strategies. In this study, we use artificial intelligence to predict groundwater levels to provide guidance for drilling shallow boreholes for subsistence irrigation. The Bilate watershed, which is located in southern Ethiopia, was selected as the study area. This is typical of areas in Africa with high demand for water and limited availability of well data. Using a non-time-series database of 75 boreholes, machine learning models including multiple linear regression, multivariate adaptive regression spline, artificial neural networks, random forest regression, and gradient boosting regression (GBR) were constructed to predict the depth to the water table. 20 independent variables were considered in the models. GBR performed the best of the approaches with an average 0.77 R-squared value on testing data. Finally, a map of predicted water levels in the Bilate watershed was created based on the best model with water levels ranging from 1.6 to 245.9 meters. With the limited set of borehole data, the results show a clear signal that can provide guidance for borehole drilling decisions for sustainable irrigation with additional implications for drinking water.

Keywords: machine learning; groundwater table; ground water level; sustainable irrigation; drinking water; water-scarcity regions; AI; gradient boosting regression

1. Introduction

Ethiopia is one of the countries in East Africa that is threatened by water scarcity. In Ethiopia, rainfed agriculture and small farming families predominate. A recent study shows that around 95% of the agricultural areas in Ethiopia are rainfed areas [1]. A Food and Agriculture Organization (FAO) study [2] shows that small farming families make up 72% of the total population. 74% of Ethiopia's farmers come from small farming families and 67% of these live below the national poverty line. About 75% of farmland is devoted to cereals. Maize and wheat dominate, complemented by teff, barley, sorghum, and rice. Drought is a major stressor that reduces cereal yields. Moreover, climate change has induced significant and erratic deviations in rainfall patterns over the year and across the country, significantly reducing crop yields overall, and especially cereals [3]. To mitigate the impact of water stress and reduce hunger, utilizing groundwater by drilling wells for irrigation is a potential solution to address the problem of increasingly erratic rainfall.

Predictions of groundwater level, or the depth-to-water table, could support decisions on where to drill wells to extract groundwater. Artificial intelligence (AI) has been widely used to predict both the surface [4-10] and groundwater [11-21] levels globally. Regarding surface water level prediction,

Khan and Coulibaly [4] used a Support Vector Machine (SVM) to examine long-term water level in Lake Erie in North America based on mean monthly water level data from 1918 to 2001. The authors compared SVM with a multilayer perceptron (MLP) and with a conventional multiplicative seasonal autoregressive model. They found that the SVM outperformed the other two models with an overall RMSE less than 0.25 m. Liang et al. [5] applied SVM and a deep learning model based on a Long Short-Term Memory (LSTM) network to predict daily surface water levels in Dongting Lake in China. They found that the LSTM has better accuracy than the SVM model with less than 0.1 m RMSE. A river water level study performed by Chen and Qiao in 2021 [6] also confirms that LSTM has good performance in predicting surface water levels. Choi et al. [7] used four machine learning algorithms including artificial neural networks (ANN), decision tree, random forest (RF), and SVM based on the daily water level from 2009 to 2013 to predict water levels from 2013 to 2015 in Upo wetlands, South Korea. They found that random forest outperforms the other three algorithms with a 0.09 RMSE.

Regarding groundwater level prediction, in 2013, Sahoo and Jha [11] constructed seventeen site-specific AI models to predict groundwater levels in Japan. Compared to multiple linear regression (MLR) models, ANN-predicted groundwater levels have a better agreement with RMSE values ranging from 0.04 to 0.4 m for 17 sites. Sahoo et al. [12] developed a modeling framework using Multilayer Perceptron (MLP) network architecture to simulate groundwater level changes in two agricultural regions in the US. They found ANN performed better than the MLR and multivariate nonlinear regression model with RMSE less than 2 m for both the agricultural regions. Zhang et al. [13] developed a new model based on LSTM to predict groundwater levels, which outperforms feed-forward neural networks and double LSTM with a 0.14 m RMSE. In 2021, Liu et al. [14] applied SVM combined with the data assimilation (DA) technique for predicting changes in groundwater level. The researchers predicted the change in groundwater levels at 1 to 3-month time scales for 46 wells located in the northeast United States and found that both the SVM and SVM with DA can adequately predict groundwater levels with RMSE less than 4 meters. Hikouei et al. [15] demonstrated that Extreme Gradient Boosting exhibits superior performance over the RF model for predicting groundwater levels in the tropical peatlands of Indonesia. Many recent studies employed wavelet-transformed analysis [16-17] and hybrid AI techniques [18-19], which have a good performance in groundwater level prediction using time-series data.

Most of these studies that use machine learning algorithms employed a large amount of time series data to predict future water levels for lakes [4-5, 8], rivers [6, 9-10], wetlands [7], basins [11, 17, 19], regions [13, 16, 18, 20], aquifers [12], and watersheds [14, 21]. These models generally have good performance with low mean square error in predictions of water levels. These studies were conducted in regions with high data availability on both the time series of water level data and climate variables. However, in many regions of interest, a large amount of time series of water level data is unlikely to be available due to a high cost of data collection for local government or organizations. Many water scarcity areas have great need for groundwater development for sustainable irrigation although the lack of good data can make models perform poorly. To the best of our knowledge, there is no research on groundwater level prediction using machine learning based on non-time-series data in rainfed agricultural regions. It is much more difficult to predict water levels accurately in the absence of time series data, because previous water level is a strong predictor of current water level.

The objective of this study is to use AI to identify suitable drilling locations for sustainable irrigation for subsistence agriculture in water scarcity regions using sparse non-time-series data on existing wells. To achieve this objective, five machine learning models were constructed to predict groundwater levels including MLR, multivariate adaptive regression spline (MARS), ANN, random forest regression (RFR), and gradient boosting regression (GBR). The models were developed using data from 75 existing boreholes in the Bilate watershed in southern Ethiopia. The best-performing model was used to predict the groundwater level for hundreds of thousands of grid points covering the Bilate region. Finally, a map of predicted water levels was created to provide guidance for decision making on drilling locations for local individuals and organizations. Figure 1 shows the workflow of this study.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2023 doi:10.20944/preprints202309.1165.v1

The rest of the paper is organized as follows. Section 2 describes the study area, data source and the methodology used in this study; Section 3 shows the main results from each of the machine learning models; Section 4 provides a discussion of the results; Section 5 concludes the paper.

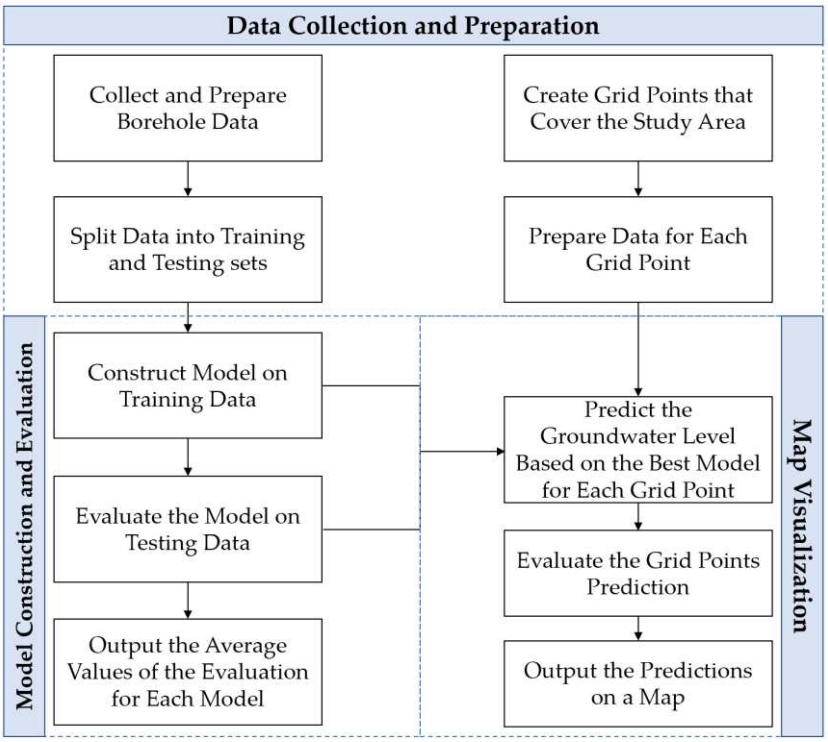

Figure 1. The workflow of this study

2. Materials and Methods

2.1. Study Area

The Bilate watershed is located in southeastern Ethiopia at latitude 6˚34' to 8˚6' N and longitude 37˚46' to 38˚18' E. The total area is 5276.25 square kilometers. Bilate is one of the largest watersheds in the Ethiopia Rift Valley Basins [22,23]. Elevation of the region ranges from 1194 m to 3216 m, as shown in Figure 2. The Bilate region was classified as three-season climate type, namely major rains from June to September, dry season from October to January and minor rains from February to May [24,25]. In Bilate, the minor rains typically start in March. However, for our analysis, we adhere to the general classification, commencing from February, as referenced in sources 24 and 25. The annual average precipitation within the Bilate watershed spans from 769 mm in lower regions to 1339 mm in the highlands. Meanwhile, the mean annual temperature fluctuates between 11 °C and 22 °C in Hossana, situated upstream, and ranges from 16 °C to 30 °C at Bilate Tena, the lower stream of the watershed [22].

The vegetation density in Bilate can be generally described as sparse to dense vegetation in the minor and major rainy seasons, and sparse vegetation in the dry season. Land in the region is mainly cultivated, grassland, plantation, shrub-land and wetland. The northern and western part of the study area is mountainous while the southern and eastern parts are lowlands. Distributed lakes in the study area indicate the presence of shallow groundwater. Pyroclastic and volcano-sedimentary rocks are dominant outcrops in the study area. The productivity of the aquifers varies from moderately productive to highly productive with average transmissivity range between 1.03×10^{-5} and 2.78×10^{-1} m/s^2 [26] .

2.2. Data Description

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2023 doi:10.20944/preprints202309.1165.v1

In this study, the dependent variable is the static water level. Field data on 75 boreholes were collected by Arbaminch Water Technology Institute (AWTI) in 2007 [27]. We fully acknowledge the limitations associated with our dataset, which was collected fifteen years ago. However, a change in static water level within the last fifteen years, even if it occurred, would have very little impact on our analysis. Any change would likely be within a few centimeters, which is small in relation to the accuracy of our analysis. Therefore, the existing historical dataset is useful to demonstrate the value of machine learning for predicting water levels.

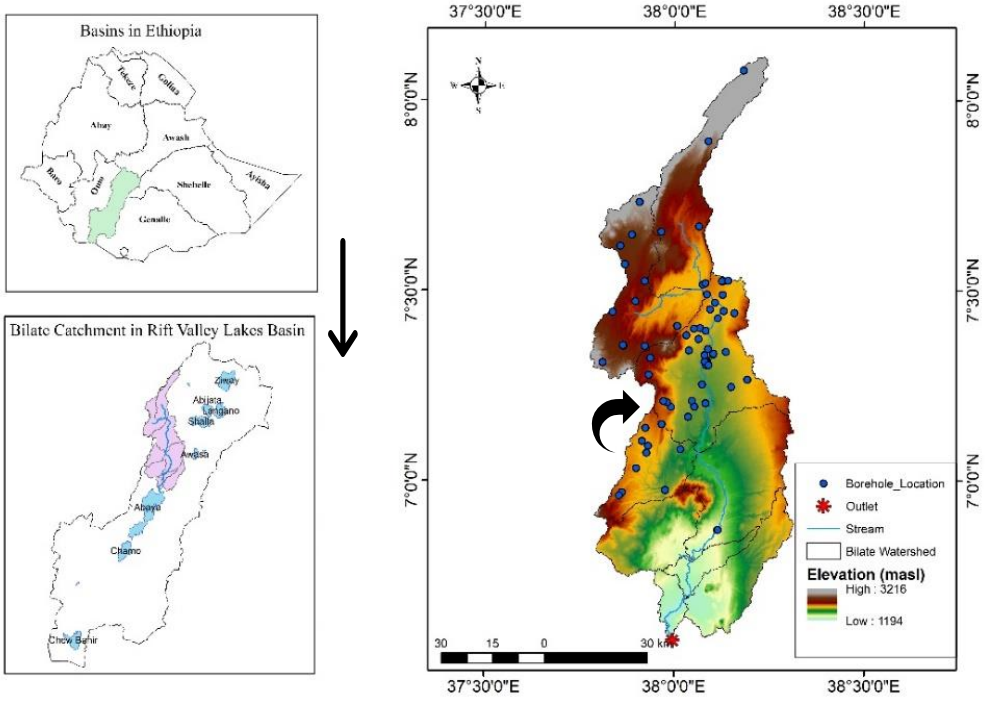

Figure 2. Bilate watershed (pink region on the lower left plot) belongs to a basin called Rift Valley (light green basin on the upper left plot). The map of Bilate watershed with boreholes and elevation [28] is shown on the right.

Initially, we explored a broader range of variables, identifying a subset that demonstrated greater influence on our predictions. This directed our attention towards incorporating this subset, resulting in the consideration of twenty independent variables for our modeling and analysis. These include elevation, soil type, meteorological variables (i.e. precipitation, specific humidity, wind speed, land surface temperature (LST) at day and night time) and vegetation (i.e. NDVI) for the three seasons in 2007. We have the coordinates for each of the 75 boreholes and have extracted the values for each variable from the raw dataset. Given that the resolution varies for different independent variables, borehole points that are close to each other might have the same extracted value, particularly for predictors with sparse resolution. Details on the sources of the dependent and independent variables are summarized in Table 1 and 2.

Nearly all the borehole points in the data set are located in areas categorized as cultivation or grassland. As mentioned, the purpose of this study is to predict groundwater levels to support decision-making on the optimal borehole drilling locations for agricultural irrigation. Based on the local conditions, it is unusual to place a borehole in forest, wetland or shrubland. For this reason, only observations in cultivated and grassland areas were included in the training and testing data sets, and the model does not predict groundwater levels for forest, wetland, and shrubland.

Machine learning models were built using a training dataset consisting of 63 randomly selected observations. Performance was then evaluated on a testing set made up of the remaining 12 observations. Each algorithm was subjected to fifteen individual experiments. For MLR and MARS, models were constructed based on fifteen diverse and randomly selected training sets. For ANN, RFR, and GBR, five different random data separations were used, and three experiments were carried

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2023 doi:10.20944/preprints202309.1165.v1

out for each training dataset. Performing a variety of experiments is important to verify the model's ability to effectively generalize to unseen data. This paper mainly presents and discusses models with median performance, and are based on consistent training and testing data separations across all algorithms. The average performance scores from the fifteen experiments are also reported, providing a comprehensive understanding of each model's effectiveness.

To predict groundwater levels for a larger area in the Bilate region, we generated a grid with resolution 100m * 100m that covered the Bilate region. The data for twenty independent variables for each grid point were prepared and processed in the same manner as for the original dataset of 75 boreholes. Because the resolution of the grid points is relatively high, grid points that fall within the same spatial resolution unit of a variable will have identical extracted values. The software tools used for data preparation and visualization were mainly QGIS 3.24.1 [29] and R 4.1.3 [30].

Table 1. Summary of data source.

Data	Unit	Source and Description	Type
Static water level	m	75 borehole points collected by AWTI in 2007	Numerical
Elevation	m	USGS Digital Elevation SRTM with 30 m resolution [31]	Numerical
Soil type	--	FAO Harmonized World Soil Database v 1.2 [32]. Four categories: chromic luvisols, eutric vertisols, humic nitisols, and vitric or mollic andosols	Categorical
Precipitation	mm/hour	NASA Global Precipitation Measurement with 0.1 degree spatial resolution [33]	Numerical
Specific humidity	Kg/kg	NASA FLDAS Noah Land Surface Model with 0.1 degree spatial resolution [34]	Numerical
Wind speed	m/s		
LST at daytime	°K	USGS MODIS Terra Land Surface Temperature with 1 km spatial resolution [35]	Numerical
LST at nighttime			
NDVI	--	USGS MODIS Terra Vegetation Indices 16-day at 250 m spatial resolution [36]	Numerical

Table 2. Description of dependent and independent variables.

Variable	Description	Variable	Description	Variable	Description
Y	Static water level	X1	Elevation	X2	Soil type
X3	Precipitation Oct to Jan (monthly ave)	X4	Precipitation Feb to May (monthly ave)	X5	Precipitation Jun to Sep (monthly ave)
X6	Specific humidity Oct to Jan (daily ave)	X7	Specific humidity Feb to May (daily ave)	X8	Specific humidity Jun to Sep (daily ave)
X9	Wind speed Oct to Jan	X10	Wind speed Feb to May	X11	Wind speed Jun to Sep

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2023 doi:10.20944/preprints202309.1165.v1

	(daily ave)		(daily ave)		(daily ave)
X12	LST daytime Oct to Jan (daily ave)	X13	LST daytime Feb to May (daily ave)	X14	LST daytime Jun to Sep (daily ave)
X15	LST nighttime Oct to Jan (daily ave)	X16	LST nighttime Oct to Jan (daily ave)	X17	LST nighttime Oct to Jan (daily ave)
X18	NDVI Oct to Jan (16-day ave)	X19	NDVI Feb to May (16-day ave)	X20	NDVI Jun to Sep (16-day ave)

2.3. Resampling Methods

2.3.1. Leave-One-Out Cross-Validation

Cross-validation (CV) is a resampling technique to check the generalization ability of a model on a limited sample. The k-fold CV method randomly divides the observations into k groups/folds of approximately equal size [37]. The first fold is used as the validation set. The remaining folds are used to fit a model. This process repeats k times and a different fold is used as the validation set each time. The CV estimate is finally computed as the average of the mean square error on each validation set. Leave-one-out CV (LOOCV) is a special case of cross-validation, where k is the number of samples [37]. In LOOCV, the model is fit based on the entire dataset with one observation excluded. Next, the fitted model is used to predict the one observation that was left out. Then, the process is repeated n times, leaving out a different observation each time, where n is the number of observations in the data set. Since the dataset is small, we performed LOOCV across the machine learning models in this study.

2.3.2. Bootstrapping

Bootstrapping is a resampling method that randomly samples values with replacement. It is mainly employed in the construction of RFR models. In the context of RFR, each decision tree within the forest is trained on a distinct dataset, generated by bootstrapping the original training set. This ensures each dataset is of the same size as the original, but composed of a subset of the original data, with some samples likely repeated. This process introduces randomness into the model-building phase, which aids in preventing overfitting and improves model robustness by ensuring that each individual tree within the forest learns from a slightly different sample of the data.

2.4. Machine Learning Algorithms

Machine learning is a branch of AI focused on building systems that learn from data. It has been widely applied in science and engineering. Application areas include as hydrogeology [11-19], cyber security [38-39], transportation [40-41], and aerospace engineering [42-44]. By identifying patterns in data, it allows the computer to learn a decision rule from data without explicit programming. Over time, with more data, machine learning models can make increasingly accurate predictions or decisions.

2.4.1. Multiple Linear Regression

A multiple linear regression model [45] is built to identify factors that affect groundwater level and to estimate the groundwater level. The general form of a linear regression model is defined as

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j, \quad (1)$$

where β_0 is the intercept; $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients; $X_1, X_2, \ldots, X_p$ are the predictors.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2023 doi:10.20944/preprints202309.1165.v1

2.4.2. Multivariate Adaptive Regression Spline

MARS is a nonparametric regression method that creates a piece-wise linear model so that the relationship between the predictor and the response can be different for different ranges of the predictors [46]. The model begins with a simple constant model and iteratively adds basis functions that best reduce the sum of squared residuals. The basis functions are created using hinge functions that form piecewise linear relationships. A hinge function takes a single variable and a constant as input, returning the difference between the variable and the constant when the variable is greater than the constant, and zero otherwise. For a cut point a, a pair of hinge functions are defined as

$$\begin{aligned} h(x-a) &= \{x-a \text{ if } x > a, 0 \text{ otherwise}\} \\ h(a-x) &= \{a-x \text{ if } x < a, 0 \text{ otherwise}\}, \end{aligned} \tag{2}$$

where x is a given variable. Each time a pair of basis functions is added, the model chooses the variable and the constant that minimize the residual sum of squares (RSS). In order to control overfitting, MARS then performs a backward deletion process. It removes basis functions that contribute the least to the model's predictive power based on generalized cross-validation (GCV) score or Akaike Information Criterion (AIC) score. The backward pass prunes the model by eliminating the terms one by one until the best subset is found for the model. The one with the lowest GCV or AIC score is usually chosen as the final model.

2.4.3. Artificial Neural Networks

ANN is a machine learning model inspired by the neural structure of the brain. An ANN model transforms inputs into outputs through a series of hidden layers to an output layer. For a regression model with a single hidden layer, the form of the ANN model [47] is defined as

$$\begin{aligned} p(y \mid x, \theta) &= N(y \mid w^T z(x), \sigma^2) \\ z(x) &= g(Vx) = [g(v_1^T x), \ldots, g(v_H^T x)], \end{aligned} \tag{3}$$

Where g is a non-linear activation function; z(x) is the hidden layer, which is a deterministic function of the input; H is the number of hidden units, V is the weight matrix from the inputs to the hidden nodes and w is the weight vector from the hidden nodes to the output. Figure 3 shows a simple neural network with only one hidden layer.

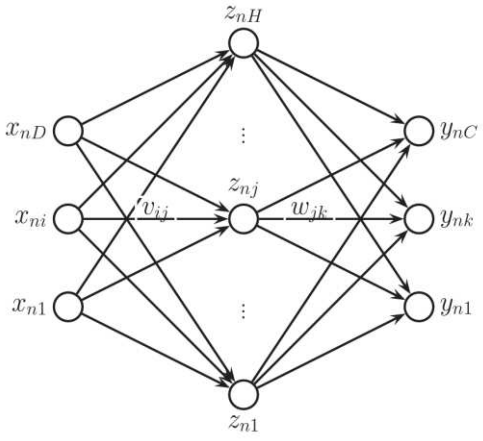

Figure 3. A simple Neural Network with one hidden layer [47].

2.4.4. Random Forest Regression

The random forest regression algorithm, proposed by Breiman in 2001, is a supervised learning algorithm that produces a forecast based on many decision trees [48]. Each tree is built based on a bootstrap sample of the training data. At each decision node of each tree, the best split is chosen among the number of randomly selected predictors. The splitting process terminates at a leaf node

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2023 doi:10.20944/preprints202309.1165.v1

when a termination criterion is met. To make a prediction at a point, the tree is traversed to find the leaf node corresponding to the independent variables corresponding to the point, and the dependent variables for the data points at the leaf nodes are averaged [49]. This procedure, creating decision trees based on different bootstrap samples and then averaging the predictions from all the trees, is called bootstrap aggregation, or bagging.

The bagging technique in the RFR model tends to reduce the variance of predictions, but a bias still exists. Specifically, since the prediction is the average of the output from all leaf nodes, observations with small values tend to be overestimated and those with large values tend to be underestimated. This tendency to bias has been identified in previous studies [50, 51]. Our results showed a bias in the initial RFR model. To correct the bias, a post-processing bias-correcting transformation to the RFR predictions could be made [51]. For the linear transformation, a linear regression model was fitted to find the intercept (β_0) and slope (β_1) of the transformed prediction:

$$f(\hat{y}) = \beta_0 + \beta_1 * \hat{y} \quad (4)$$

where $\hat{y}$ is the predicted response value by the initial random forest model on training data; $\beta 0$ is the coefficient for the intercept, and $\beta 1$ is the coefficient for $\hat{y}$. The objective is to find the parameters that minimize the mean square error:

$$\min_{\beta_0,\beta_1} 1/n * \sum(f(\hat{y}_i) - y_i)^2 \quad (5)$$

2.4.5. Gradient Boosting Regression

The gradient boosting regression (GBR) is an ensemble learning algorithm that combines multiple weak prediction models, typically decision trees, to create a strong predictive model. Initially, the ensemble is empty. The model starts with an initial prediction, typically the mean of the response variable. The difference between the observed values and the initial prediction, known as the residuals, is calculated. These residuals represent the errors that the model needs to correct. Then, a decision tree is trained to predict the residuals. The tree is constructed to minimize a specific loss function using gradient descent optimization. The loss function we used in this study is defined as

$$L(y_i, f(x_i)) = \sum_{i=1}^{n} \frac{1}{2}(y_i - f(x_i))^2 \quad (6)$$

where y_i is the i^{th} observed value; $f(x_i)$ is the predicted response value; n is the number of observations. Next, The predictions from the decision tree are combined with the current ensemble's predictions to obtain an updated prediction. This updated prediction is added to the ensemble. Then, the residuals are recalculated using the updated predictions. The new residuals represent the errors that were not captured by the current ensemble. The process continues for a specified number of iterations or until a certain stopping criterion is met. The final prediction is obtained by summing the predictions from the entire ensemble. By iteratively correcting the errors of the previous models, gradient boosting regression is able to learn complex relationships and improve predictive accuracy.

2.5. Evaluation Metrics

The evaluation metrics presented in this paper include mutual information (MI), root mean square error (RMSE), median absolute error (MAE), and R-squared.

MI measures the degree of relatedness between the datasets [52]. A higher MI value shows that the dependent variable (y) has higher relatedness to the corresponding independent variable (x).

$$MI(Y;\ X) = \sum_{x\in X}\sum_{y\in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right), \quad (7)$$

where p(x, y) represents the joint probability function of x and y; p(x) and p(y) are the marginal probability functions of x and y, respectively.

RMSE describes how far the predictions deviate from the actual values. A small RMSE represents a good performance of the model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2023 doi:10.20944/preprints202309.1165.v1

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (8)$$

MAE measures the median of absolute errors between the predicted and observed values. Similar to RMSE, it is also used to describe how well the data fit the model.

$$\text{MAE} = \text{Median}(|\hat{y}_i - y_i|) \quad (9)$$

R-squared represents how much variation for a dependent variable is explained by an independent variable.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad (10)$$

where y denotes the observed values; $\hat{y}$ denotes the predicted values; $\bar{y}$ denotes the mean of the observed values.

3. Results

In this section, we detail the mutual information analysis and outcomes from all the implemented methods. We primarily focus on results from one training and testing data partition used consistently across all methods. This training and testing split was chosen because it yielded close to median performance across all models. For the ANN, RFR, and GBR methods, three experiments were conducted for each data partition, hence models with median performance were selected. Primary results associated with all algorithms include the residuals versus predicted values outlined in Table 7, and the observed versus predicted values for both training and testing data detailed in Table 8. Furthermore, Tables 5 and 6 summarize the performance metrics of the models based on the single experiment that has a median performance and average performance score based on the fifteen experiments, respectively.

3.1. Mutual Information Analysis

A zero MI value indicates no mutual dependence or relatedness between the specific independent variable and the static water level. A larger MI value represents a stronger relatedness. From Figure 4, variables precipitation Oct to Jan (X3), precipitation Jun to Aug (X5), and specific humidity Jun to Sep (X8) have relatively stronger relatedness to static water level. On the other hand, Wind speed Oct to Jan (X9), Wind speed Jun to Sep (X11), NDVI Feb to May (X19) showed weaker relatedness. The soil type (X2) is removed from the mutual information analysis, because it is a categorical variable.

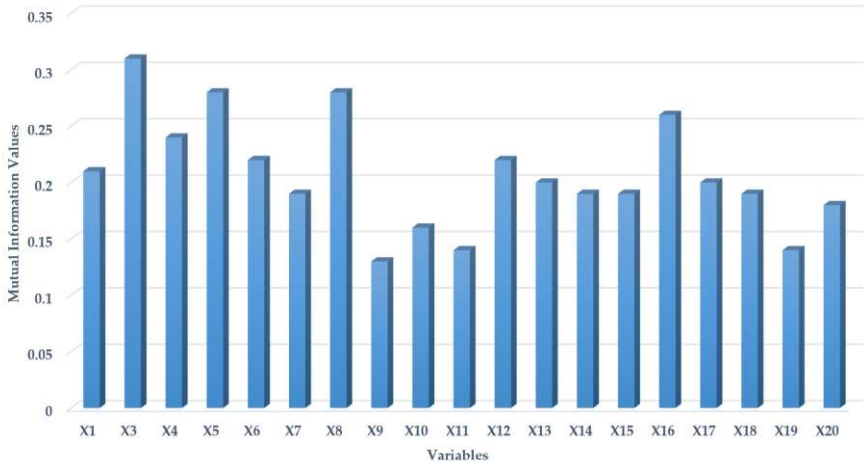

Figure 4. Mutual information between independent variables and the dependent variable.

3.2. Multiple Linear Regression

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2023 doi:10.20944/preprints202309.1165.v1

Before building a MLR, highly correlated predictors were removed. If a pair of predictors have a correlation equal to or greater than 0.85, the R function findCorrelation() randomly picks one predictor to remove. Table 3 shows the predictors remaining after removal along with their coefficients. We found the factors including the euric vertisols soil type (X2) and NDVI from Jun to Sep (X20) have a significant relationship with the static water level at 0.05 significance level.

To examine the normality assumption of a linear regression model, we created a Quantile-Quantile (Q-Q) plot and a residual plot. From the Q-Q plot shown in Figure 5, we see that the points are approximately distributed along the line with light upper and lower tails. No obvious pattern was found on the residual plot (Figure 7a). We also performed a Shapiro-Wilk normality test on residuals with a null hypothesis – the residual data are normally distributed. The p-value is approximately 0.51; therefore, the null hypothesis of normality is not rejected at the 0.05 significance level. The model performance results are shown in Table 4 and 5 and will be discussed in section 4.2.

Table 3. Summary of MLR

Variables	Coefficients	Standard Error	P value
Intercept	354.04	1963	0.85
X2 Eutric Vertisols	-97.05	40.22	0.02**
X2 Humic Nitisols	-13.83	27.07	0.61
X2 Vitric & Mollic Andosols	-19.95	33.77	0.56
X3	-9.25	5.93	0.12
X4	-8.96	11.28	0.43
X5	2.03	1.58	0.20
X10	42.15	57.00	0.46
X13	3.30	6.26	0.60
X15	-5.53	7.23	0.45
X19	65.85	162.0	0.69
X20	249.16	91.97	0.009**

** 0.05 significance level

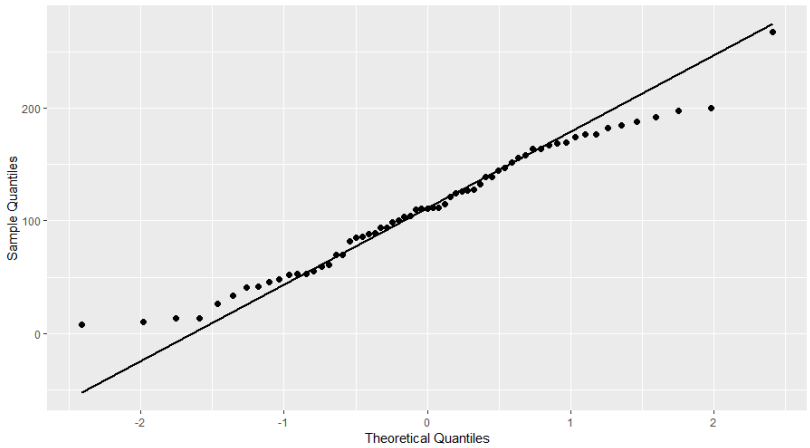

Figure 5. Quantile-Quantile (Q-Q) plot.

3.3. Multivariate Adaptive Regression Spline

Similar to MLR, we built a MARS model using the training data. The hyperparameters tuned for the MARS model included the interaction complexity (degree) and the number of terms to retain in the final model (nprune). We evaluated the degree from 1 to 3 and the nprune from 2 to 20. The final setting for the degree is 3 and nprune is 8.

The MARS model was built with a backward elimination feature selection process. The backward pass prunes the model by eliminating the terms one by one until the best subset is found

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2023 doi:10.20944/preprints202309.1165.v1

for the model. The Generalized cross-validation (GCV) criterion was used to evaluate each subset and was also considered as the variable importance measure. From Figure 6, the first three most important variables include LST daytime from Feb to May (X13), precipitation from Feb to May (X4), and wind speed from Oct to Jan (X9). Soil type (X2) eutric vertisols has an importance value of zero indicating that this predictor did not contribute to the predictive power of the model and was never used in any of the MARS basis functions in the pruned final model.

Even though MARS, being a nonparametric technique, does not assume linearity and homoscedasticity, the residual plot can still provide valuable diagnostic information about how well the model fits the data. Figure 7 shows that the residuals are randomly scattered around zero, indicating that the variance of the error is approximately constant across all levels of the independent variables, which is a desirable property.

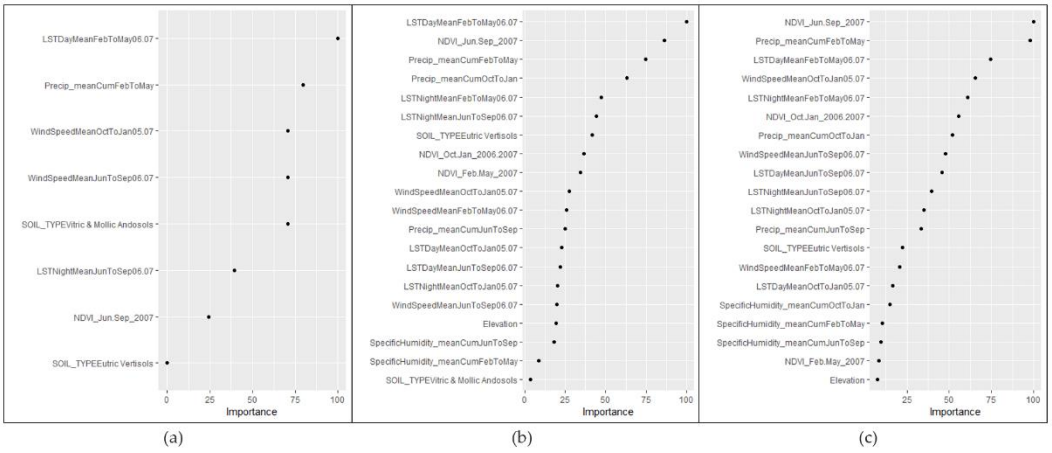

Figure 6. Variable importance plots: (a) MARS; (b) RFR; (c) GBR.

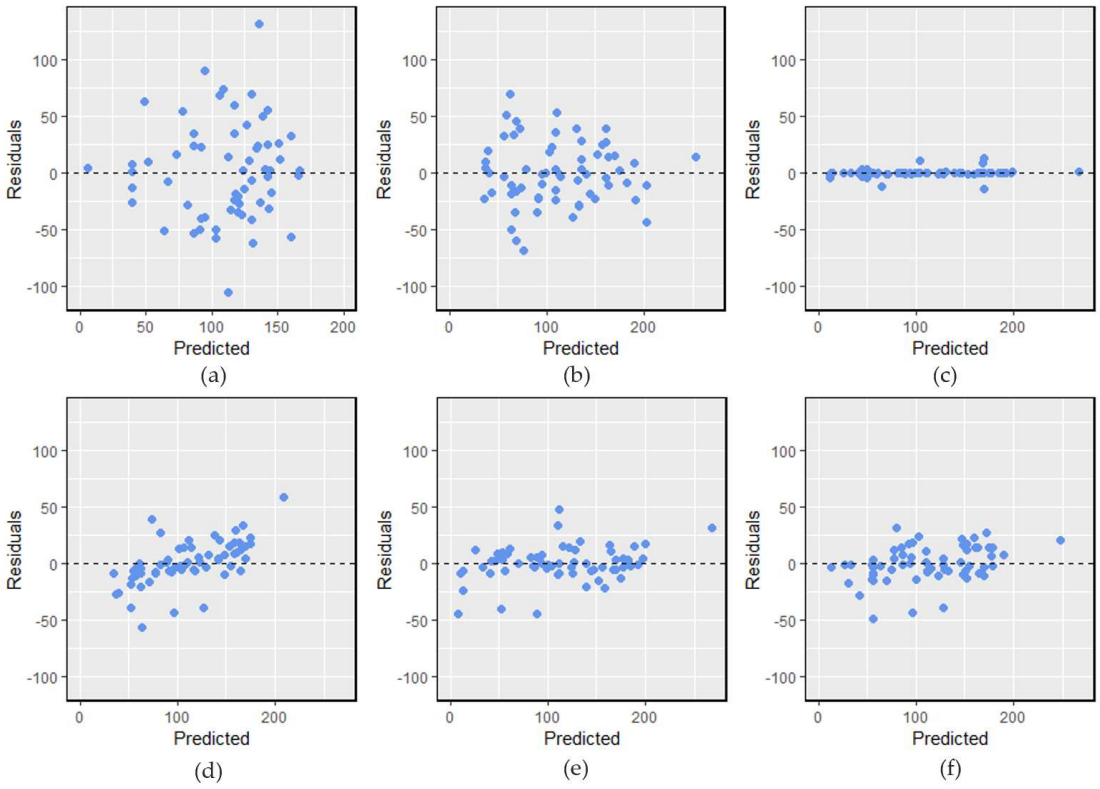

Figure 7. Plot of residuals (m) versus predicted values (m) for (a) MLR; (b) MARS; (c) ANN; (d) Original RFR; (e) RFR with linear transformation; (f) GBR.

 doi:10.20944/preprints202309.1165.v1

3.4. Artificial Neural Networks

The training of an ANN model is executed using the nnet and caret packages. This nnet package is designed to support a single hidden layer sandwiched between the input and output layers. In the preprocessing stage, the model is set to center and scale predictors, which is a common strategy to normalize variable scales.

Suited to regression tasks, the output layer of the neural network uses a linear activation function. The model incorporates key hyperparameters, such as weight decay for regularization, and the size, which determines the number of nodes in the hidden layer.

We set the range for 'size' from 1 to 10, and evaluated the 'weight decay' at 0, 0.01, and 0.1. After the tuning and training process, the model identifies optimal hyperparameters: a decay value of 0.1 and a size of 10. A residual plot is generated to check the assumption of homoscedasticity. As indicated in Figure 7, the residuals are distributed randomly, suggesting that the assumption of the constant variance of errors is reasonable.

3.5. Random Forest Regression

An RFR model was trained using randomForest package in R based on the training dataset with applying LOOCV as the resampling method. The hyperparameters for the RFR model include the number of randomly selected predictors at each split (mtry), the node size, and the number of trees (ntree). These hyperparameters were collectively tuned within a for loop. For mtry, a grid range of 1 to 10 was set and fine-tuned using the 'caret' package, while the node size was assessed at 5, 6, and 7, and the ntree parameter was evaluated between 50 and 200. Following this comprehensive grid search, the optimal settings were found to be an mtry value of 8, a node size of 5, and a ntree value of 60.

To assess the importance of the predictors, a variable importance plot was generated, revealing LST at daytime from Feb to May (X13), NDVI from Jun to Sep (X20), and precipitation from Feb to May (X4) to be of higher importance compared to the other variables, as depicted in Figure 6.

In response to the inherent bias of the RFR model, as detailed in Section 2.4.4, we employed a post-processing step to correct for this bias. This involved applying a linear transformation to the RFR predictions. Specifically, we fitted a linear regression model with the predicted water level from the initial RFR model as the independent variable and the actual water level as the dependent variable. The parameters of the transformation were estimated on the training sample. The estimated coefficients for the intercept ($\beta 0$) and predicted water level variable ($\beta 1$) are -29.65 and 1.3, respectively. Comparing the residual plot for the original RFR model and the post-processed model, we see that the bias of the original RFR model has been mitigated.

3.6. Gradient Boosting Regression

GBR was constructed using gbm package in R. The hyperparameters such as the minimum number of observations in a node required for a split (n.minobsinnode), the boosting model's complexity (interaction.depth), number of iterations, and the learning rate were optimized via grid search. We assessed n.minobsinnode at values of 5 and 10, interaction.depth within a range of 1 to 3, iterations at 50, 70, 100, and 120, and learning rates at 0.1 and 0.01. The final optimal hyperparameters, which led to the best model performance, were 5, 2, 100, and 0.1, respectively.

To evaluate the importance of the predictors, a variable importance plot was generated, revealing NDVI from Jun to Sep (X20), precipitation from Feb to May (x4), and LST at daytime from Feb to May (X13) to be of higher importance compared to the other variables, as depicted in Figure 6.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2023 doi:10.20944/preprints202309.1165.v1

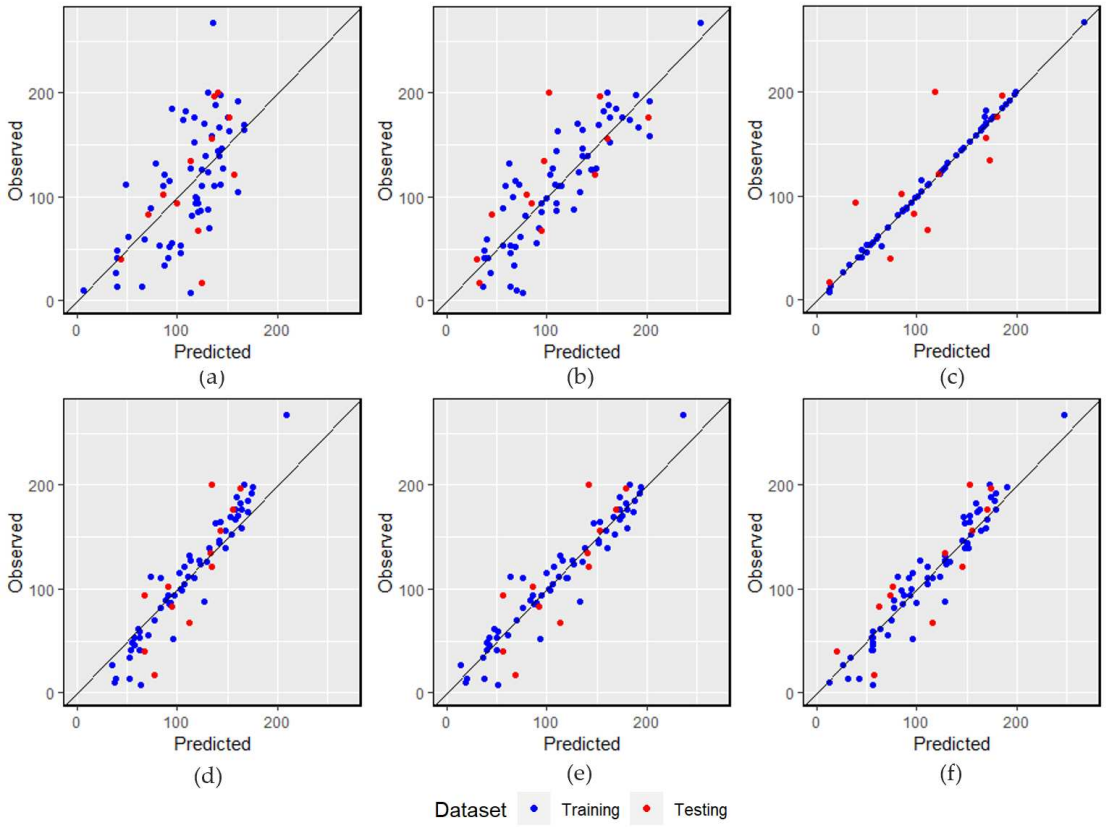

Figure 8. Observed (m) versus predicted (m) plots for (a) MLR; (b) MARS; (c) ANN; (d) Original RFR; (e) RFR with linear transformation; (f) GBR.

Table 4. Model performance evaluation based on one experiment that has a median performance

Dataset	Model	RMSE (m)	MAE (m)	R Squared
Training	MLR	43.56	28.54	0.40
	MARS	27.85	18.67	0.76
	ANN	3.55	0.27	0.99
	Original RFR	19.51	9.05	0.88
	RFR with linear transformation	15.55	6.41	0.92
	GBR	15.66	9.28	0.92
Testing	MLR	45.46	23.62	0.37
	MARS	38.32	24.94	0.55
	ANN	35.48	15.43	0.61
	Original RFR	33.66	23.79	0.65
	RFR with linear transformation	30.34	16.97	0.72
	GBR	27.86	21.84	0.76

Table 5. Average model performance score based on multiple experiments did for each model

Dataset	Model	RMSE (m)	MAE (m)	R Squared
Training	MLR	42.49	29.92	0.43
	MARS	40.56	28.46	0.47
	ANN	7.58	2.10	0.96
	RFR	19.92	12.83	0.88
	GBR	12.74	7.58	0.95
	MLR	46.81	26.14	0.31

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2023 doi:10.20944/preprints202309.1165.v1

Testing	MARS	49.63	34.39	0.23
	ANN	36.45	23.74	0.49
	RFR	29.46	18.77	0.68
	GBR	24.55	18.92	0.77

4. Discussion

4.1. Important Variables Analysis

In our study, we assessed variable importance using methods such as MI, significance of predictors through MLR, and GCV-based importance scoring for MARS, RFR, and GBR models. While ANNs can be powerful predictors, they are not typically preferred for identifying important variables due to their 'black box' nature, which complicates the interpretation of individual feature contributions. Therefore, we did not assess variable importance for the ANN model.

The results of the variable importance analysis for each method were presented in section 3. Collectively, the findings suggest that three variables, namely LST at daytime from February to May (X13), NDVI from June to September (X20), and precipitation from February to May (X4), consistently hold importance across different models. However, the significance of other variables fluctuates depending on the specific model used. This suggests a potential correlation among the predictor variables, and the variables identified the most important might differ from one model to another.

Some inconsistencies are observed between the analysis and the importance values from the machine learning models. This may stem from their differing methodological approaches to examining variable importance. Mutual information is a non-parametric method that quantifies the degree of dependence or relatedness between two variables. It can capture both linear and non-linear relationships, but it doesn't account for interactions between variables. Therefore, a variable might be rated as less important in mutual information analysis but may prove critical in a model where interactions between variables are considered. Machine learning models like MARS, RFR, and GBR not only consider individual variables' contributions but also consider the interactions and combinations of variables. Consequently, these models can identify variables as important even if their individual relationships with the outcome variable are not strong, provided their combined effect with other variables is substantial. This highlights the importance of applying diverse methods when investigating variable importance. Each method brings its strengths and can offer unique insights. Mutual information can identify variables that have a strong individual effect, whereas machine learning models can capture complex interaction effects.

4.2. Model Performance Evaluation and Comparison

As previously highlighted in section 2.2, numerous experiments were carried out for each machine learning algorithm. We selected the model with median performance for this paper, and its results are detailed in Table 4 To provide a more comprehensive understanding of each model's performance, we calculated the average performance score from the fifteen experiments, which is presented in Table 5.

Examining Table 4, we find that GBR outperforms the other models with a high R-squared value of 0.76. While the performance of the RFR with linear transformation is slightly inferior to the GBR it still outperforms the remaining models. In contrast, MLR appears insufficient for modeling our data effectively. The MARS model exhibits some predictive capability, though it doesn't fully capture the data's complexity. With an R-square value of 0.99 on the training data and 0.61 on the testing data, the ANN model shows clear evidence of overfitting to the training data. This occurs despite the model having a single hidden layer and parameters that have been optimally tuned to avoid overfitting. The overfitting may stem from the small and potentially less diverse dataset, causing the model to memorize training data rather than learning to generalize from it. Future work might resolve this by enriching the dataset and/or simplifying the model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2023 doi:10.20944/preprints202309.1165.v1

Table 5 presents the average performance score based on multiple experiments conducted for each model. Consistent with the findings from the single experiment (Table 4), GBR has the best performance among the models. The RFR model's performance is slightly lower than that of GBR. MARS, on the other hand, exhibits the weakest average performance among the five models. Collectively, these results suggest that GBR and RFR are most suitable for predicting depth to water table for our data set.

4.3. Grid Points Prediction Evaluation Based on the Best Model

As detailed in Section 2.2, the grid points were generated to comprehensively represent the study area. We utilized the primary GBR model presented in this paper (Table 4) to generate predictions for these grid points. Given the map's discretized nature, we evaluated prediction performance solely based on these grid points. This involved comparing the observed water levels from the existing 75 boreholes with the predicted water levels for their closest respective grid points (Figure 9). Table 7 illustrates the evaluation of water level predictions. It's worth noting that the predicted water levels are slightly worse on both the training and testing data in comparison to the performance reported in Section 4.2. This is to be expected: is naturally more accurate to predict based on predictor variables at the well location rather than based on predictor variables some distance away. However, this analysis shows that our model can be useful even if, as may be the case in practice, predictor values are available only at grid points and not at a specific well location. Thus, despite the slight drop in prediction accuracy, model results can still be valuable to support decisions about drilling locations in the region.

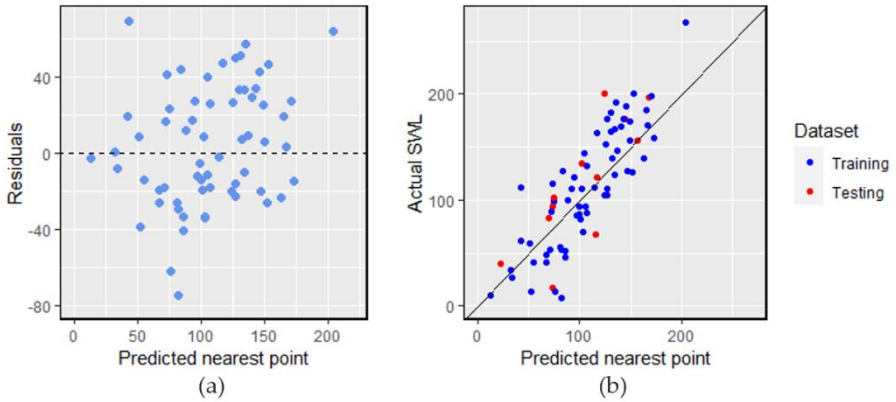

Figure 9. (a) Residual plot for the predicted water level for the nearest grid point; (b)Actual static water level versus predicted water level for the nearest grid point.

Table 7. Nearest grid point performance evaluation

Model	Data	RMSE (m)	MAE (m)	R Squared
GBR	Training	31.45	26.39	0.69
	Testing	36.61	30.02	0.60

4.4. Final Map of the Predicted Water Level

Utilizing QGIS, we generated a 100m*100m resolution map based on the grid points (Figure 10). The predicted groundwater levels of the grid points were broken into 20-meter categories and color-coded the figure for the purpose of display. The regular block pattern in some areas of the map is due to the low resolution of some of the predictor variables. The map suggests that areas of shallow groundwater (under 30 m) are predominantly located in the northern and southern regions, with some patches in the central area. Conversely, the eastern region mostly features deeper groundwater. This high-resolution map can serve as a practical guide for borehole drilling for sustainable irrigation, particularly beneficial for stakeholders such as drilling companies, government entities, and local farmers.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2023 doi:10.20944/preprints202309.1165.v1

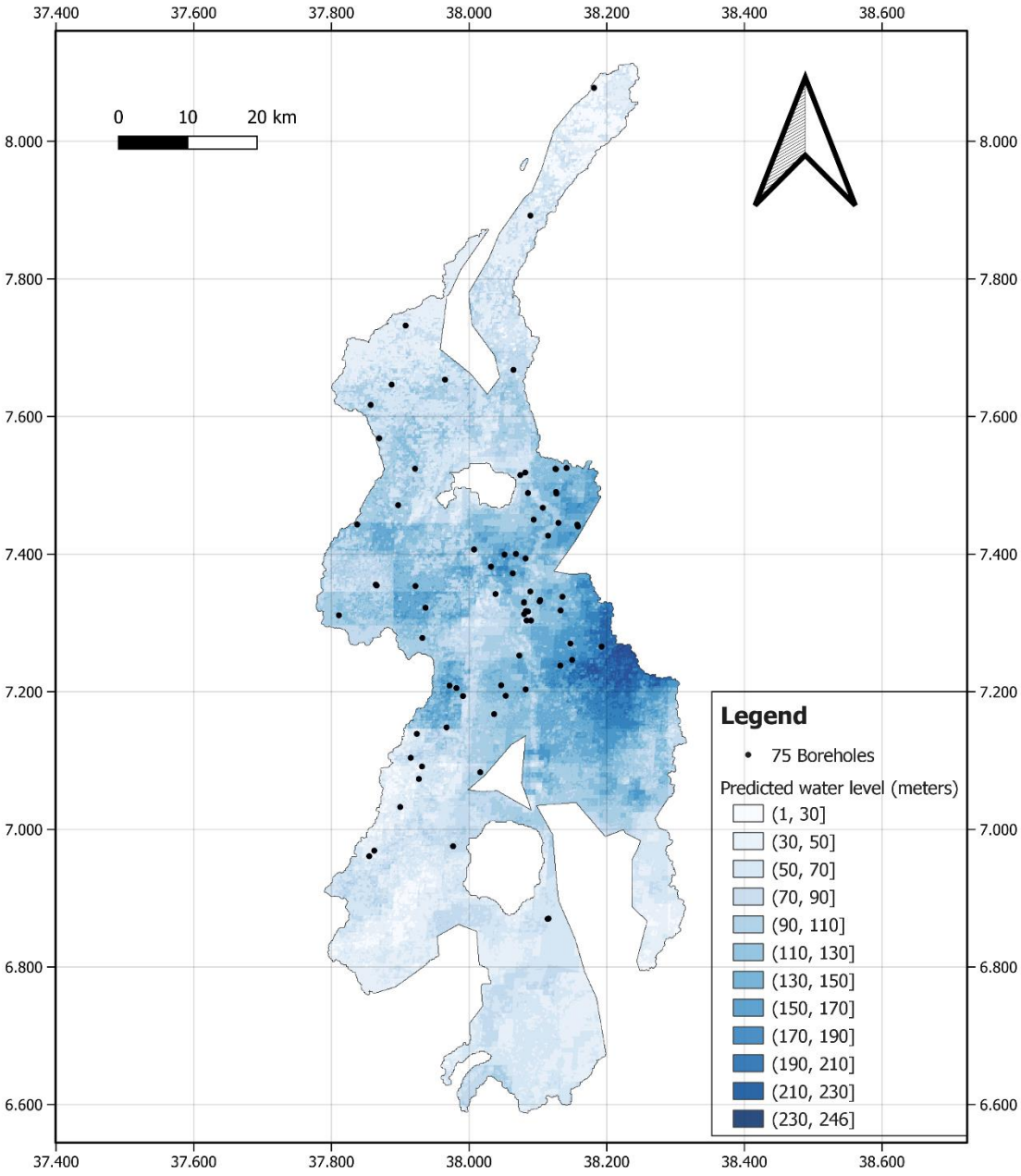

Figure 10. Predicted groundwater level for Bilate Ethiopia.

5. Conclusions

To conclude, this research emphasizes the application of AI in pinpointing viable drilling sites for sustainable irrigation and even drinking water, in water-deficient areas. Prior studies have typically utilized time-series data for groundwater level prediction. However, the challenge in water scarcity regions lies in the lack of data due to formidable data collection constraints. These regions, marked by higher water demand, necessitate effective strategies for groundwater exploitation.

Addressing this gap, we have utilized the available non-time-series data to devise five machine learning models for groundwater level prediction. Of these, Gradient Boosting Regression consistently demonstrated superior performance, with an average R-squared value of 0.77 across numerous experiments. The highest-performing model was subsequently employed to predict groundwater levels across the entire Bilate region. This process resulted in the development of a high-resolution map, anticipated to guide local communities and organizations in pinpointing the most suitable locations for sustainable irrigation drilling.

Investigating variable importance revealed that Land Surface Temperature during daytime from February to May, NDVI from June to September, and precipitation from February to May consistently demonstrated significance across models. We captured inconsistencies between the variable importance from the MI method and machine learning methods. The results from both should be considered complementary rather than contradictory. Using a combination of methods allows for a more robust and comprehensive understanding of variable importance, leading to a more reliable model. In case of substantial discrepancies, deeper investigations can be conducted to reconcile the findings.

There are some limitations of this study. Firstly, a potential concern is the relatively small dataset of 75 boreholes, which are not evenly distributed throughout the Bilate watershed. This may present a limitation for making region-wide predictions. Secondly, the predictor variables we considered were limited to ones that could be readily computed from publicly available data. Thirdly, the ANN model showed a tendency to overfit the training data, indicating the need for more extensive hyperparameter tuning and model simplification. Lastly, our data, having been collected in 2007, may be somewhat dated. Efforts are underway to acquire more recent data to verify prediction accuracy. Future research in this region should aim to improve the predictive power of groundwater levels by considering additional predictor variables (e.g., distance to water and elevation above permanent streams), forecast groundwater recharge, and analyze the impacts of climate change. This will provide comprehensive guidance for decision-making related to borehole drilling.

6. Patents

The findings in this paper are being incorporated into a system called WellMapr© and designed to support drilling decisions for shallow groundwater and drinking water wells in Ethiopia.

Author Contributions: All authors contributed to the conceptualization, design of the study, and reading and revising the manuscript. Methodology, W.L. and K.L.; software, W.L.; validation, W.L. and K.L.; formal analysis, W.L.; investigation, W.L. and K.L.; resources, M.F. and R.D.; data collection, M.F.; data curation, W.L.; writing—original draft preparation, W.L.; writing—review and editing, K.L., R.D., M.F., P.H., and W.L.; visualization, W.L. and M.F.; supervision, M.F., P.H., R.D., and K.L.; project administration, K.L. and R.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by a graduate research fellowship to W.L. from George Mason University's Center for Resilient and Sustainable Communities.

Acknowledgments: This work represents a collaboration among George Mason University, Arba Minch University and Global Map Aid with support by the Czech Geological Survey. We would like to express our gratitude to Dr. Jiří Bruthans for the valuable comments and feedback on the manuscript.

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

References

1. Chandrasekharan, K. M.; Subasinghe, C.; Haileslassie, A. *Mapping irrigated and rainfed agriculture in Ethiopia (2015-2016) using remote sensing methods* (Vol. 196). 2021, International Water Management Institute (IWMI).
2. FAO. Small Family Farms Country Factsheet Thiopia - food and agriculture. Available online: https://www.fao.org/3/i8911en/I8911EN.pdf (assessed on August 9, 2022)
3. Haileslassie, A. On-Farm Smallholder Irrigation Performance in Ethiopia: From Water Use Efficiency to Equity and Sustainability. 2016, ISBN 978-92-9146-468-5.
4. Khan, M.S.; Coulibaly, P. Application of Support Vector Machine in Lake Water Level Prediction. *J. Hydrol. Eng.* **2006**, *11*, 199–205, doi:10.1061/(ASCE)1084-0699(2006)11:3(199).
5. Liang, C.; Li, H.; Lei, M.; Du, Q. Dongting Lake Water Level Forecast and Its Relationship with the Three Gorges Dam Based on a Long Short-Term Memory Network. *Water* **2018**, *10*, 1389, doi:10.3390/w10101389.
6. Chen, S.; Qiao, Y. Short-Term Forecast of Yangtze River Water Level Based on Long Short-Term Memory Neural Network. *IOP Conf. Ser.: Earth Environ. Sci.* **2021**, *831*, 012051, doi:10.1088/1755-1315/831/1/012051.
7. Choi, C.; Kim, J.; Han, H.; Han, D.; Kim, H. S. Development of water level prediction models using machine learning in wetlands: A case study of Upo wetland in South Korea. *Water* **2019**, *12*(1), 93.
8. Wang, Q.; Wang, S. Machine learning-based water level prediction in Lake Erie. *Water* **2020**, *12*(10), 2654.

9. Assem, H.; Ghariba, S.; Makrai, G.; Johnston, P.; Gill, L.; Pilla, F. Urban Water Flow and Water Level Prediction Based on Deep Learning. In *Machine Learning and Knowledge Discovery in Databases*: *European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22,* **2017**, *Proceedings, Part III 10* (pp. 317-329). Springer International Publishing.
10. Kim, D.; Han, H.; Wang, W.; Kim, H. S. Improvement of Deep Learning Models for River Water Level Prediction Using Complex Network Method. *Water* **2022**, *14*(3), 466.
11. Sahoo, S.; Jha, M.K. Groundwater-Level Prediction Using Multiple Linear Regression and Artificial Neural Network Techniques: A Comparative Assessment. *Hydrogeol J* **2013**, *21*, 1865–1887, doi:10.1007/s10040-013-1029-5.
12. Sahoo, S.; Russo, T.A.; Elliott, J.; Foster, I. Machine Learning Algorithms for Modeling Groundwater Level Changes in Agricultural Regions of the U.S. *Water Resources Research* **2017**, *53*, 3878–3895, doi:10.1002/2016WR019933.
13. Zhang, J.; Zhu, Y.; Zhang, X.; Ye, M.; Yang, J. Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. *Journal of hydrology* **2018**, *561*, 918-929.
14. Liu, D.; Mishra, A. K.; Yu, Z.; Lü, H.; Li, Y. Support vector machine and data assimilation framework for Groundwater Level Forecasting using GRACE satellite data. *Journal of Hydrology* **2021**, *603*, 126929.
15. Hikouei, I. S.; Eshleman, K. N.; Saharjo, B. H.; Graham; L. L.; Applegate, G.; Cochrane, M. A. Using machine learning algorithms to predict groundwater levels in Indonesian tropical peatlands. *Science of the Total Environment*, **2023**, *857*, 159701.
16. Rahman, A. S.; Hosono, T.; Quilty, J. M.; Das, J.; Basak, A. Multiscale groundwater level forecasting: Coupling new machine learning approaches with wavelet transforms. *Advances in Water Resources* **2020**, *141*, 103595.
17. Wen, X.; Feng, Q.; Deo, R. C.; Wu, M.; Si, J. Wavelet analysis–artificial neural network conjunction models for multi-scale monthly groundwater level predicting in an arid inland river basin, northwestern China. *Hydrology Research* **2017**, *48*(6), 1710-1729.
18. Bahmani, R.; Ouarda, T. B. Groundwater level modeling with hybrid artificial intelligence techniques. *Journal of Hydrology* **2021**, *595*, 125659.
19. Liu, W.; Yu, H.; Yang, L.; Yin, Z.; Zhu, M.; Wen, X. Deep Learning-Based Predictive Framework for Groundwater Level Forecast in Arid Irrigated Areas. *Water* **2021**, *13*(18), 2558.
20. Wu, Z.; Lu, C.; Sun, Q.; Lu, W.; He, X.; Qin, T.; Yan, L.; Wu, C. Predicting Groundwater Level Based on Machine Learning: A Case Study of the Hebei Plain. *Water* **2023**, *15*(4), 823.
21. Kochhar, A.; Singh, H.; Sahoo, S.; Litoria, P. K.; Pateriya, B. Prediction and forecast of pre-monsoon and post-monsoon groundwater level: using deep learning and statistical modelling. *Modeling Earth Systems and Environment* **2022**, *8*(2), 2317-2329.
22. Orke, Y.A.; Li, M. H. Hydroclimatic Variability in the Bilate Watershed, Ethiopia. *Climate* **2021**, *9*, 98, doi:10.3390/cli9060098.
23. Tekle, A. Assessment of Climate Change Impact on Water Availability of Bilate Watershed, Ethiopian Rift Valley Basin. *In Proceedings of the AFRICON 2015*; September **2015**; pp. 1–5.
24. Tsegay Wolde-Georgis; Aweke, D.; Hagos, Y. The Case of Ethiopia Reducing the Impacts of Environmental Emergencies through Early Warning and Preparedness: The Case of the 1997–98 El Niño. *National Meteorological Service Agency (NMSA): Addis Ababa, Ethiopia,* 2000, 1-73..
25. Legese, W.; Koricha, D.; Ture, K. Characteristics of Seasonal Rainfall and Its Distribution Over Bale Highland, Southeastern Ethiopia. *J Earth Sci Clim Change* **2018**, *09*, doi:10.4172/2157-7617.1000443.
26. Czech Geological Survey and Geological Survey of Ethiopia. Explanatory notes to the thematic geoscientific maps of Ethiopia at a scale of 1 : 50,000. 2018. Available online: http://www.geology.cz/etiopie-2018/outputs/dila/explanatory-notes-0638-c2-dila.pdf
27. Muluneh, M. Web-Based Decision Support Systems for Managing Water Resources of Abaya Chamo Basin Project; 2018.
28. Alaska Satelite Facility. Available online: https://asf.alaska.edu/ (Assessed on August 1, 2022)
29. QGIS Development Team. QGIS Geographic Information System. *Open Source Geospatial Foundation Project*, 2022. http://qgis.osgeo.org
30. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, 2022, Vienna, Austria. URL https://www.R-project.org/.
31. U.S. Geological Survey. USGS EROS Archive - Digital Elevation - Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global. Available online: https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-shuttle-radar-topography-mission-srtm-1#overview (assessed on August 1, 2022).
32. Food and Agriculture Organization of the United Nations. *Harmonized world soil database*. Available online: https://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/harmonized-world-soil-database-v12/en/ (accessed on November 16, 2022)

33. Huffman, G.J.; Stocker, E.F.; Bolvin, D.T.; Nelkin, E.J.; Tan, J. GPM IMERG Final Precipitation L3 1 month 0.1 degree x 0.1 degree V06. Goddard Earth Sciences Data and Information Services Center (GES DISC). Greenbelt, MD, 2019. Available online: https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGM_06/summary (accessed on July 31, 2022)
34. Amy McNally NASA/GSFC/HSL. FLDAS Noah Land Surface Model L4 Global Monthly 0.1 x 0.1 degree (MERRA-2 and CHIRPS). Goddard Earth Sciences Data and Information Services Center (GES DISC). Greenbelt, MD, USA, 2018. Available online: https://disc.gsfc.nasa.gov/datasets/FLDAS_NOAH01_C_GL_M_001/summary (accessed on July 31, 2022)
35. Wan, Z.; Hook, S.; Hulley, G. MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V006. *NASA EOSDIS Land Processes DAAC* 2015. Available online: https://doi.org/10.5067/MODIS/MOD11A1.006 (accessed on July 31, 2022)
36. Didan, K. MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006. *NASA EOSDIS Land Processes DAAC* 2015. Available online: https://doi.org/10.5067/MODIS/MOD13Q1.00636. (accessed on August 1, 2022)
37. Hastie, T.; Tibshirani, R.; Friedman, J. H.; Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*, New York: springer, 2009. Vol. 2, pp. 1-758.
38. Greitzer, F. L.; Li, W.; Laskey, K. B.; Lee, J.; Purl, J. Experimental investigation of technical and human factors related to phishing susceptibility. *ACM Transactions on Social Computing*, 2021, *4*(2), 1-48.
39. Tang, L.; Mahmoud, Q. H. A survey of machine learning-based solutions for phishing website detection. *Machine Learning and Knowledge Extraction*, 2021, *3*(3), 672-694.
40. Zhou, W. Condition State-Based Decision Making in Evolving Systems: Applications in Asset Management and Delivery, Doctoral dissertation, George Mason University, Fairfax, VA, 2023.
41. Zantalis, F.; Koulouras, G.; Karabetsos, S.; Kandris, D. A review of machine learning and IoT in smart transportation. *Future Internet*, 2019, *11*(4), 94.
42. Harvey, A.; Laskey, K.; Chang, K. C. Machine learning applications for sensor tasking with non-linear filtering. *Sensors*, 2022, *22*(6), 2229.
43. Fan, Z. Models and Algorithms for Data-Driven Scheduling, Doctoral dissertation, George Mason University, Fairfax, VA 2023.
44. Fan, Z.; Chang, K. C.; Raz, A. K.; Harvey, A.; Chen, G. Sensor Tasking for Space Situation Awareness: Combining Reinforcement Learning and Causality. In *2023 IEEE Aerospace Conference*, 2023, pp. 1-9.
45. Freedman, D. A. *Statistical models: theory and practice*. Cambridge university press, 2009.
46. Friedman, J. H. Multivariate adaptive regression splines. *The annals of statistics* **1991**, *19*(1), 1-67.
47. Murphy, K. P. Machine learning: a probabilistic perspective. MIT press, 2012.
48. Breiman, L. Random forests. *Machine learning* **2001**, *45*(1), 5-32.
49. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R news* **2022**, *2*(3), 18-22.
50. Zhang, G.; Lu, Y. Bias-corrected random forests in regression. *Journal of Applied Statistics* **2012**, *39*(1), 151-160.
51. Malhotra, S.; Karanicolas, J. A Numerical Transform of Random Forest Regressors corrects Systematically-Biased Predictions. 2020, arXiv preprint arXiv:2003.07445.
52. Ross, B.C. Mutual information between discrete and continuous data sets. *PLoS ONE* **2014**, *9*, e87357.