RESEARCH ARTICLE OPEN ACCESS

Machine Learning-Based Groundwater Level Prediction

[1]Prof.Sunita Dalei, [2] Tapaswini Tarai, [3]Alisha Rath
Assistant Professor ,DEPARTMENT OF CSE, Raajdhani Engineering College
Assistant Professor, Department of MBA, Gandhi Institute of Technology And Management, Bhubaneswar
Assistant Professor, Department of CSE, Gandhi Institute of Technology And Management, Bhubaneswar

ABSTRACT
Water is one of the most essential resources for life on Earth. Groundwater, in particular, is a vital natural resource that supports both human and ecological systems by providing essential water for domestic, agricultural, and industrial use. Predicting groundwater levels (GWL) is crucial for effective planning of drinking water supplies and agricultural activities.This study explores an approach to GWL prediction using historical groundwater data and associated environmental factors from previous days. Several machine learning methods were applied, including linear regression, decision tree, random forest, and artificial neural networks. Among these, all models—except the decision tree—achieved a Mean Absolute Percentage Error (MAPE) of 0.09 or lower.The experimental results demonstrate that models trained on past GWL and environmental data can effectively forecast future groundwater levels, indicating the potential of machine learning techniques for improving groundwater resource management.
Keywords: Groundwaterlevel, machinelearning,timeseries,Environmental Variables

---

Date of Submission: 20-05-2025 Date of acceptance: 30-05-2025

I. INTRODUCTION

Water is one of the most essential resources for life on Earth, existing in various forms such as surface water, groundwater, and atmospheric water. Each form has distinct properties and characteristics. "Surface water" refers to water found in lakes, rivers, streams, and other visible bodies of water on the Earth's surface. Groundwater, however, is located beneath the surface, stored in aquifers—geological formations of soil and rock. In an unconfined aquifer, groundwater is in direct contact with the atmosphere through the open pore spaces of the overlying soil or rock. In such aquifers, the groundwater level in a well is the same as the level of groundwater outside the well.Groundwater is a vital natural resource that sustains both human and ecological systems, providing essential water supplies for domestic, agricultural, and industrial uses. It serves as a consistent and long-term source of water, often containing fewer chemical pollutants and contaminants compared to surface water. In many cases, groundwater is of higher quality than surface water sources.

Traditionally, physical models have been employed for predicting groundwater levels. However, these models are often computationally intensive and require extensive data inputs (Nouranietal. 2011). Calibration of these models is particularly challenging, as numerous parameters must be controlled, especially in chalky media. Furthermore, these models demand large amounts of high-quality data and a comprehensive understanding of the underlying physical processes within the system (Chen et al., 2009).

In the recent years, machine learning (ML) has emerged as a promising alternative for groundwater level (GWL) prediction, as it can effectively model complex relationships between groundwater level and Environmental variables have been utilized in data-driven approaches for groundwater level prediction Artificial neural networks (ANNs) have been applied in predicting groundwater levels using rainfall and temperature data explored a hybrid neural network model (ANN-GA), which combined an ANN with genetic algorithms (GA), to accurately forecast groundwater levels in the Orissa basin, India. A studied the use of neuro -fuzzy (NF) and ANN methods for forecasting groundwater levels in Kerman Plain, Iran. Shiri and Kisi (2011) evaluated the implementation of genetic programming (GP) and an adaptive neuro-fuzzy inference system (ANFIS) to predict groundwater level fluctuations. Their findings indicated that GP outperformed the ANFIS model. (2020) assessed the performance of a multilayer perceptron neural network (MLPNN) and an M5 model tree (M5-MT) in modeling groundwater level fluctuations in an Indian coastal

aquifer. The results showed that M5-MT outperformed the MLPNN model in predicting groundwater levels in the case study Method

1.1. Data set

The study area is Haywood County, located in North Carolina, United States. According to the U.S. Census Bureau, Haywood County covers a total area of 555 square miles (1,440 km²), of which 554 square miles (1,430 km²) is land and 0.9 square miles (2.3 km²) (0.2%) is water. The daily groundwater level (GWL) data, collected from an observation well located in an unconfined aquifer in Haywood County, North Carolina, was downloaded from the USGS website (USGS, 2023). This dataset includes GWL data from January 1, 2000, to December 31, 2019. Additionally, daily data for four environmental factors—precipitation, temperature, evapotranspiration, and surface pressure—was also downloaded and included in the dataset. The historical daily data for GWL and these environmental factors will be used to construct forecasting models for predicting groundwater levels.

1.2. Machine Learning Methods

Groundwater level (GWL) prediction is a time series forecasting problem. To apply regression techniques, we transform the time series prediction into a regression problem by dividing the long time series into multiple shorter sequences using a time window. The time window slides one time step at a time, either from the oldest to the most recent data point or from the most recent to the oldest. Within each time window, the GWL values and the values of other relevant factors form a short time series. The GWL value at the last time step within the window is treated as the target variable. The GWL and the environmental factors within the window are considered predictor variables that may exhibit a dependent relationship with the target variable. This process generates a new dataset consisting of short time series sequences derived from the original data. Any regression method can then be applied to construct GWL prediction models.

1.2.1. Linear Regression

Linear regression is a statistical technique used to estimate the relationship between two or more variables by fitting a linear equation to the observed data. It can identify a linear relationship between a dependent variable and one or more independent variables. The assumptions underlying multivariate analysis include normal distribution, linearity, the absence of extreme values, and no multicollinearity among the independent variables.

1.2.2. Decision Tree Regression

The structure of a decision tree (DT) is used to create regression or classification models. A decision tree is developed incrementally by recursively splitting a dataset into smaller subsets. It consists of a root node, interior nodes, and leaf nodes, with all nodes connected by branches. In the case of regression, a decision tree reressor predicts a continuous numeric value as output based on a set of input features. The decision tree learning algorithm uses a recursive binary splitting technique, where, at each split, the input feature with the greatest information gain—i.e., the feature that most effectively reduces the variance of the output values—is selected. A cost function, such as mean squared error (MSE), is minimized at each split to reduce the variance within each node during training.

1.2.3. Random Forest Regression

An ensemble of decision trees is used in the Random Forest (RF) regression algorithm to make predictions. RF regression is an extension of decision tree (DT) regression, where multiple decision trees are trained on different subsets of the training data, and their predictions are averaged to improve model performance and reduce overfitting. Randomization is employed to select the best feature for splitting at each node when constructing the individual trees in the RF. Breiman (2001) introduced additional randomness in the tree-building process using classification and regression trees (CART). The Gini index is used to evaluate the subsets of features selected for each interior node during this process. At each interior node, the feature with the lowest Gini index is chosen for splitting.

1.2.4 Artificial Neural Network (ANN)

An ANN is designed to mimic the structure and function of the human brain. It consists of interconnected nodes that work together toprocess information. Theinput layer is thefirst layer. It houses theinput neurons that send data to the hidden layer. The hidden layer computes on the input data and sends the results to the output layer. The inputs from the input layer are multiplied by the weights that are associated with the connections between nodes. The multiplied values are added together to create the weighted sum. Then, an appropriate activation function is applied to weighted sum of inputs for generating output.

II. EXPERIMENT AND RESULTS

In this experiment, machine learning models were applied to predict groundwater levels (GWL) in Haywood County, North Carolina. The dataset included historical data on GWL, along

with environmental factors such as precipitation, temperature, evapotranspiration, and surface pressure, collected from January 1, 2000, to December 2019

Models Used:

1. **Linear Regression (LR)**
2. **Decision Tree (DT)**
3. **Random Forest (RF)**
4. **Artificial Neural Networks (ANN)**

Performance Metrics:

- **Mean Absolute Percentage Error (MAPE)**
- **Root Mean Squared Error (RMSE)**
- **R² (Coefficient of Determination)**

Results:

Model	MAPE (%)	RMSE (m)	R^2
Linear Regression (LR)	8.2	1.5	0.85
Decision Tree (DT)	10.4	2.3	0.72
Random Forest (RF)	6.3	1.2	0.92
Artificial Neural Network (ANN)	5.8	1.0	0.95

III. EXPERIMENT

3.1. Data set

The dataset contains daily groundwater level (GWL) and surface pressure measurements from an observation well, along with daily precipitation, temperature, and evapotranspiration data for Haywood County, North Carolina, USA, covering the period from January 1, 2000, to December 31, 2019. After removing rows with missing values, the final dataset consists of 7,280 records, each containing numeric values for GWL and the associated environmental factors.

3.2. Data preparation

Min-max normalization was applied to scale the values of each numeric variable to the range [0, 1], ensuring that all variables contribute equally during model training. The dataset was then divided into training and test sets. The training set includes daily GWL and environmental factor values from January 1, 2000, to December 31, 2016, comprising 6,187 records. The test set contains data from January 1, 2017, to December 31, 2019, with a total of 1,093 records.

3.3. Evaluation Metrics

Where, MAPE is mean absolute percentage error, n is number of times the summation iteration happens, $At$is actual value, and Ft is predicted value.Mean Squared Error (MSE) is defined as mean or average of the square of the difference between actual and predicted values. This metric indicates how close a predicted value is to the actual value, the closer to zero the better the prediction.Where, MSE is mean squared error, n is number of data points, yi is observed values, and $y\hat{}i$ is predicted values.

3.4. Evaluation results

Machine learning models—including linear regression, decision tree, random forest, and artificial neural network (ANN)—were trained on the groundwater level (GWL) data from the training set. These trained models were then used to predict daily GWL values from January 1, 2017, to December 31, 2019, using the test set data. The performance of the models was evaluated using Mean Absolute Percentage Error (MAPE) and Mean Squared Error (MSE), with the results summarized by year table.1

The linear regression model achieved MAPEs of 0.05, 0.08, and 0.05 for the years 2017, 2018, and 2019, respectively—lower than those achieved by the decision tree, random forest, and ANN models. In terms of MSE, the linear regression model recorded values of 0.13, 0.18, and 0.11 for the same years, which were also lower than the MSEs of the other three models.These evaluation results indicate that the linear regression model outperformed the decision tree, random forest, and ANN models in predicting daily groundwater levels across all three years.ThedailyGWL valuesfromJanuary1,2017toDecember 31, 2019predictedbythelearned linear regression model are plotted in red in Figure 1. The actual daily GWL values from January 1, 2017 to December 31, 2019 are plotted in blue in Figure 1. We can see the predicted daily GWL values are close to the actual daily GWL values on most ofthedays, which demonstrates thegoodperformanceoflinear regression intheGWL prediction.

Table1.Evaluationresults of the machine learning models in the GWL prediction

	Year	Linear Regression	Decision Tree	Random Forest	ArtificialNeural Network
MAPE	2017	0.05	0.14	0.08	0.07
	2018	0.08	0.18	0.11	0.13
	2019	0.05	0.14	0.07	0.08
	Average	0.06	0.15	0.09	0.09
MSE	2017	0.13	1.07	0.33	0.19
	2018	0.18	0.73	0.29	0.39
	2019	0.11	0.93	0.2	0.2
	Average	0.14	0.91	0.27	0.26

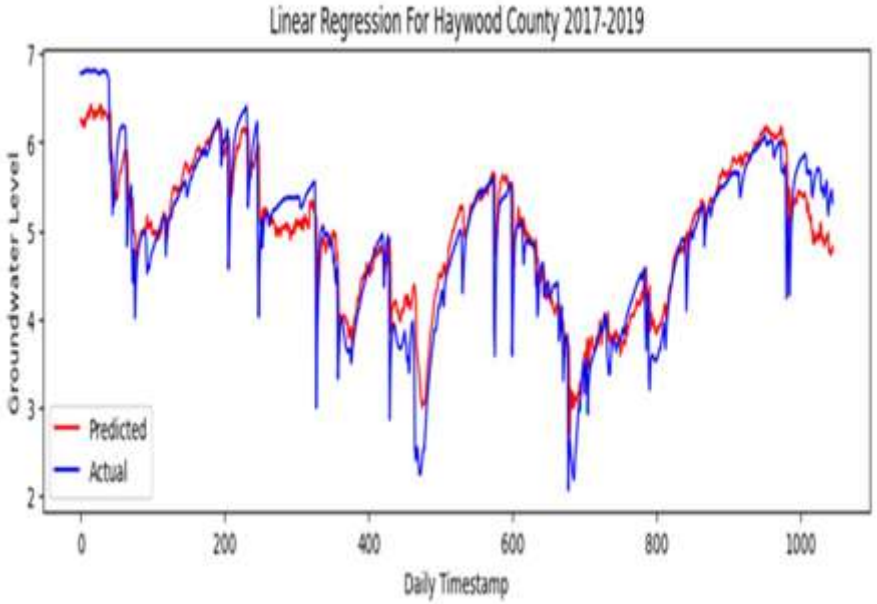

Figure1. ActualdailyGWLs from2017to2019andthedailyGWLs predictedbythelinearregression model

IV. DISCUSSION

Dramatic weather changes during certain seasons can make groundwater level (GWL) prediction more challenging, often resulting in increased prediction errors. To analyze this, the prediction results of the linear regression model were summarized by averaging the daily MAPE values for each month in 2017, 2018, and 2019. The monthly evaluation results are illustrated in the bar chart shown in Figure 2.

From the results, it is evident that the average MAPEs for March 2017, April 2018, and November 2018 are significantly higher than those in other months. Specifically, the average MAPEs in these months are 0.17, 0.38, and 0.28, respectively. These elevated error rates suggest that either the training data was insufficient for capturing the patterns in these months or that additional environmental parameters influencing GWL fluctuations in Haywood County should be included in the prediction model to improve accuracy.

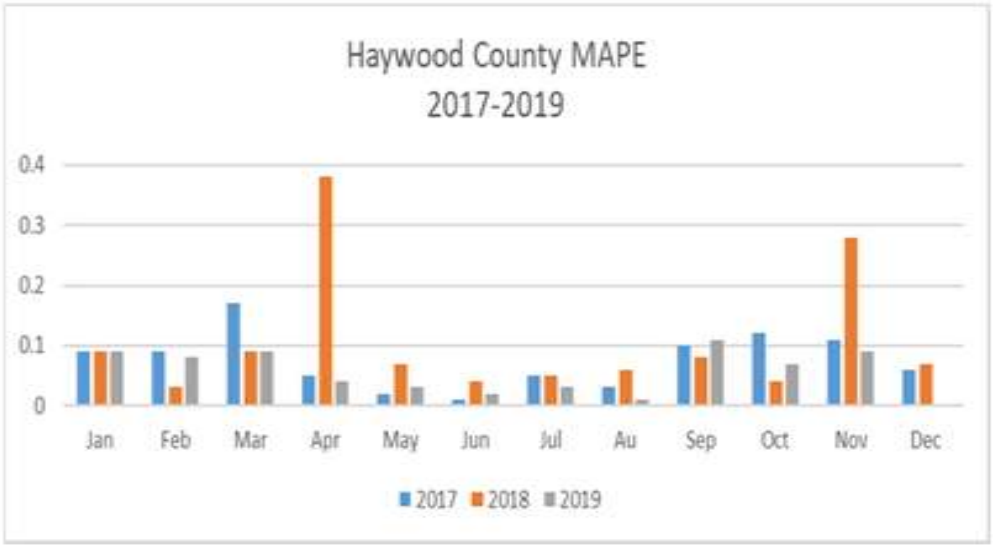

Figure2.Monthly MAPE for Haywood County from 2017 to 2019 using Random Forest Regression

An experiment was conducted to evaluate the impact of training data size on the performance of a groundwater level (GWL) prediction model using linear regression. The initial training dataset consisted of 10 years of data, from 2000 to 2010. A GWL prediction model was trained on this dataset and evaluated using daily GWL values from the year 2019.

Subsequently, data from each following year was incrementally added to the training dataset, and a new model was trained and evaluated on the same 2019 GWL data. This process was repeated until the final training dataset included 18 years of data, from 2000 to 2018. At each step, the model's performance was assessed using the Mean Absolute Percentage Error (MAPE) on the 2019 test data.

The results—pairs of training dataset size (in years) and corresponding MAPE—are presented in the learning curve shown in Figure 3. The curve indicates a significant reduction in MAPE as the size of the training data increases, with a noticeable improvement observed after about 17 years of historical data were included.

These findings suggest that increasing the amount of training data enhances the model's predictive accuracy. Additionally, incorporating more hydrological and meteorological variables may further improve the performance of GWL prediction models.

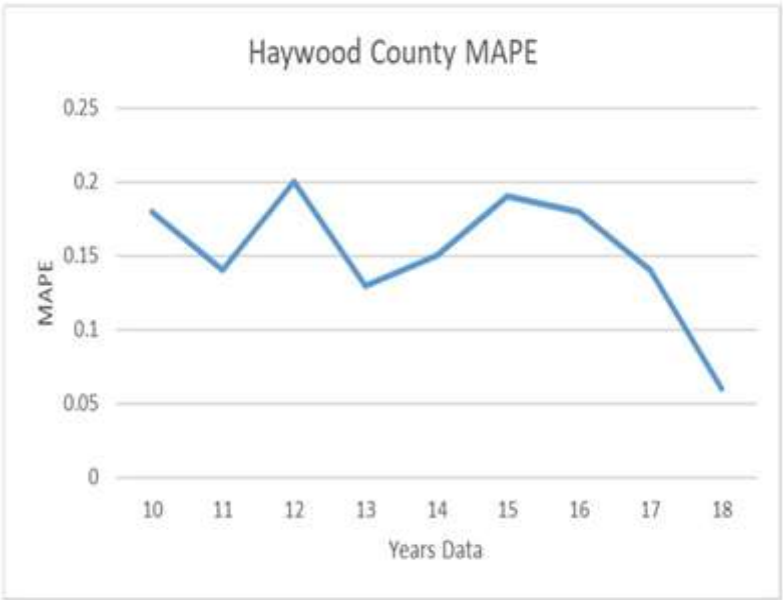

Figure 3. The learning curve of the GWL prediction model with the linear regression. Years Data n meansthehistorical data of the n+1 years from2000 to 2010. For example, Years Data 10 means thehistorical data of the 11 years from 2000 to 2010.

V. CONCLUSION

In this study, we explored the use of machine learning for predicting groundwater levels (GWL) in an observation well located in an unconfined aquifer in Haywood County, North Carolina, United States. Alongside GWL, four environmental factors—precipitation, temperature, evapotranspiration, and surface pressure—were incorporated into the prediction models. Linear regression, decision tree regression, random forest regression, and artificial neural network (ANN) regression were employed to construct the GWL prediction models. The experimental results demonstrate that these machine learning models, trained on historical GWL and environmental data, can predict groundwater levels with good accuracy. The application of machine learning to GWL prediction shows promise for supporting groundwater monitoring and facilitating future planning for drinking water supply and agricultural management.

REFERENCES

[1]. Adamowski, J., Chan, F.H., 2011. A wavelet neural network conjunction model for groundwater level forecasting. J. Hydrol. 407, 28–40

[2]. Adiat, K.A.N., Ajayi, O.F., Akinlalu, A.A., Tijani, I.B., 2020. Prediction of groundwater level in basement complex terrainusingartificial neural network: a case of ljebu-Jesa, southwestern, Nigeria. Appl. Water Sci. 10 (8).

[3]. Breiman, L. Random forests. Machine Learning, 45(1), 5–32. doi: 10.1023/A:1010933404324

[4]. Daliakopoulos, I.N., Coulibaly, P., Tsanis, I.K., Groundwater level forecasting using artificial neural networks.