Gound Water Level Prediction Using Machine Learning Concepts

Valluri Keerthi Ram[1], Kamma Ajay Nageswara Rao[2], Abhay Gokavarapu[3]
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
keerthiramvalluri@gmail.com, kammaajay11@gmail.com, abhaygokavarapu@gmail.com

***Abstract*—Groundwater depletion was always a problem to sustainable development and water security. This study uses historical hydrological and meteorological data to forecast groundwater levels based on a number of traditional ML models, including Random Forest, Gradient Boosting and Support Vector Regression. Once again, Random Forest delivered the best forecasting accuracy of all the models tested. Overall, Random Forest provides a viable and affordable solution for forecasting groundwater levels. Rainfall prediction will be used in future studies to potentially improve model performance further.**

***Index Terms*—Groundwater prediction, Machine learning, Random Forest, Regression models, Water management**

I. Introduction

Groundwater is one of the most critical sources of freshwater for businesses, agriculture, and human beings. It is the main or unique water supply for many locations world wide (i.e., mostly and particularly in drought area zones). However, groundwater levels have increasingly dropped due to unsustainable extraction, uncontrolled use, and climate variability. This problem has been exacerbated by ignorance, bad planning, and lack of forecasting / prediction tools. Therefore, anticipating groundwater availability has become fundamental for water resource managers and policymakers.

Depletion of groundwater has far-reaching effects. Crop failure, decreased access to drinking water, higher pumping costs, land subsidence, and even groundwater contamination can result from falling water tables. Small-scale farmers and rural communities are especially at risk because they frequently rely solely on wells for domestic and irrigation needs. From a societal standpoint, water insecurity can exacerbate health problems, migration, and economic instability. In order to guarantee food security, water equity, and long-term environmental sustainability, this issue must be resolved.

Piezometers and observation wells are used for physical measurements in traditional groundwater monitoring. Despite their accuracy, these methods are labor-intensive, expensive to maintain, and have a limited spatial coverage. As a result, numerous statistical and empirical models have been created to examine trends in groundwater. These methods, however, frequently fall short in capturing intricate, nonlinear relationships between influencing factors like land use, temperature, humidity, and historical water table data.

This study suggests using machine learning algorithms to model and forecast groundwater levels using historical data in order to get around the drawbacks of traditional approaches. The emphasis is on supervised learning methods that can model nonlinear relationships between several variables, such as Random Forest Regression, Support Vector Regression, and Gradient Boosting. Historical datasets containing groundwater levels as well as pertinent hydrological and meteorological parameters are used to train these models. The ultimate objective is to create a reliable, affordable, and deployable prediction system that can assist in groundwater management decision-making.

This research focuses on machine learning methods that are better suited for my work environment, as they need less data, are easier to implement, and more interpretable, rather than relying on deep learning and other more complex black-box methods like many other recent studies do. With this approach, accuracy and transparency are both achieved, therefore, it can be more readily applied by water boards, NGOs, and local governments. This research contributes directly to the United Nations Sustainable Development Goals 6 (Clean Water and Sanitation) by aiding in the bona fide monitoring, forecasting, and management of water resources.

The best way to increase the ground water level prediction is to incorporate the rainfall prediction as a future extension of this work. Rainfall prediction can significantly improve the accuracy of groundwater forecasting as it is a major source for groundwater recharge, therefore, it is vital for the model to account surface and subsurface water interactions. This will enable the system to improve in real-time and help develop a sophisticated water prediction system that integrates environmental dynamics.

II. LITERATURE SURVEY

M.B. Raj et al. (2024) conducted a single comprehensive evaluation of machine learning methods for predicting groundwater levels. The authors focused strictly on the assessment of the supervised learning models' performance to accurately predict groundwater levels. They showcased the two-pronged approach on the importance of algorithm selection and apply-pre-processing of data as a means of improving performance.

In the authors' view, machine learning approaches, if implemented appropriately, can provide effective and adaptable solutions for managing water resources, especially in areas where conventional hydrological models are ineffective as a result of limited data [1].

T. Zeng et al. (2022) applied a hybrid model of Artificial Intelligence (AI) for groundwater level forecasting. The authors used AI techniques to capture the nonstationary, nonlinear variations of groundwater levels using Empirical Mode Decomposition (EMD) and Support Vector Regression (SVR) modelling methods. The authors showed that the application of physically based and data-driven approaches could enhance forecasting accuracy in areas with complex geological conditions, supporting risk management and landslide early warning systems using forecasting data to predict risk [2].

F. Feng et al. (2024) conducted a study regarding the effectiveness of traditional machine learning models and deep learning architectures in predicting groundwater levels. It featured models including LSTMs and CNNs, in addition to traditional machine learning heuristics such as Decision Trees, Random Forests, and Support Vector Machines. It was noted by the authors that although traditional models were cheaper to compute, deep learning approaches, through the use of LSTMs, performed significantly better in time-series data with long-range dependencies in comparison to the simpler algorithms. Such research exemplifies the increasing influence of deep learning algorithms in hydrology [3].

V. Kumar et al. (2023) investigated how various machine learning models can be utilized to forecast rainfall in urban municipalities. While the primary purpose was not rainfall forecasting but groundwater modeling, the research results can still be utilized because of the highly correlated nature of rainfall trends and groundwater recharge. They evaluated model performance for models including Gradient Boost, Random Forest, and Deep Neural Networks and provided a performance baseline across the urban settings. One caveat is that the respective spatial and temporal characterizations/histories of the sites greatly affect model performance accuracy in these environmental forecasting applications [4].

N. Mungale et al. (2024) did an extensive comparison study between conventional machine learning models and deep learning models for dynamic rainfall forecasting. They used real-world environmental data and utilized models such as XGBoost, ANN, and LSTM networks. They found that the deep learning models, specifically LSTM, produced better accuracy in forecasting over the long term, while conventional machine learning models were better equipped to provide short-term forecasts to their computational efficiency and other overheads. Thus, the findings of this research will be useful for forecasting groundwater level, particularly in instances that rainfall is the dominant contributor for recharge [5].

T. V. Tran et al. (2025) exemplified a new way to reconstruct and predict groundwater time series with machine learning in the Allertal area of Germany. They took on the issue of missing data, which is an ongoing issue for groundwater datasets, using imputation procedures and regression-based machine learning models, including Random Forest and Gradient Boosting. The authors demonstrated to some extent the validity of these models to identify seasonal and interannual variation in groundwater levels, and provided useful information about how machine learning can enhance approaches to water resource planning and resilience in these temperate climate zones [6].

A. Jari et al., (2023) described the application of individual and ensemble machine learning models to map groundwater potential zones in the arid area of Tan-Tan in Morocco. The authors utilized input parameters such as geology, slope, rainfall, soil type, and land use, and applied models such as Random Forest, Logistic Regression, and ensemble methods to produce groundwater potential maps. The research showed the utility of ensemble models by higher accuracy and generalization. The implications of the findings are critical for groundwater exploration and sustainable utilization in arid and water-scarce regions [7].

A. Ali et al. (2024) developed an advanced machine learning model that proposed using Transformer models for forecasting groundwater levels in the Thames Basin, London. The authors investigated more standard models (i.e., GRU, LSTM) as well as more recent deep architectures and illustrated that Transformers improved long-horizon prediction ability much better than other models. The authors emphasized the importance of high reliability in the temporal observations and outside features such as weather and land use in ultimately improving model performance. This work represents one of the first example of demonstrating Transformers can be used in hydrogeology for time-series forecasting, and thus opens up additional avenues for research [8].

G. Tuysuzoglu et al. (2023) have recommended the use of an ensemble machine learning model to improve rainfall prediction, using the K-Star algorithm. The study aims to improve prediction of rainfall patterns by introducing K-Star, a lazy learning algorithm that integrates instance-based learning with distance functions based on entropy in an ensemble configuration. The models were evaluated against a range of individual learners and superior models were detected, especially where the meteorological data was noisy and irregular. The authors commented on the strength of ensemble methods to promote generalization and robustness, particularly with chaotic weather systems. Overall, this research makes a rich contribution to model selection for hydrological variables that may indirectly impact elevated levels of groundwater since rainfall is often a critical driver of recharge [9].

P. N. Triveni et al. (2023) offered a detailed review of rainfall forecasting methods, highlighting the application of Multiple Linear Regression (MLR) and Artificial Neural Networks (ANN). They integrated the findings from several studies and weighed the advantages and shortcomings of MLR (a linear statistical technique) and ANN (a non-linear adaptive model) in describing spatiotemporal patterns of rainfall. The review highlighted that as MLR is simple to interpret and beneficial for linear relationships, ANN is better at discovering sophisticated, non-linear relationships inherent in meteorological data. The authors concluded hybrid models fusing

statistical and machine learning techniques hold great promise. Although the paper is a review, it presents the basics important for researchers of groundwater level forecasting, considering rainfall plays a core role in hydrological cycles [10].

III. METHODOLOGY

This project entailed examining district-wise ground water level data from the Atal Jal Disclosed Ground Water Level dataset (2015–2022) and implementing machine learning techniques to classify and analyze the trends in water levels. The methodology was organized in a step-by-step manner encompassing data preprocessing, statistical analysis, distance metric analysis, model training, and performance testing.

The process began with data preparation. The data was loaded into pandas and preprocessed by dropping rows containing missing or non-numeric data. Ground water levels in pre-monsoon seasons over the years were key features, along with spatial features like latitude and longitude. In order to facilitate binary classification, a new label column was introduced by applying a threshold to the 2022 pre-monsoon ground water level values. The samples whose levels were 25 meters or more below ground level were categorized as class 1 (deep), and the remaining ones as class 0 (shallow).

Then, to know how statistically distant two areas are, data for two particular districts—Pune and Bengaluru Rural—were selected. For both districts, mean ground water level (also known as the class centroid) and the spread (standard deviation) were determined. These values served to understand the intra-class variability. The Euclidean distance between the centroids of the two classes was then calculated to measure the inter-class distance. Histogram for the two districts was also plotted to graphically represent the comparison between their distributions of ground water levels.

Next, we performed a feature-level density analysis. One feature—specifically, the pre-monsoon ground water level from 2015—was chosen first to determine its distribution across the dataset. A histogram was made with ten bins using numpy and matplotlib. From this histogram, we calculated both the mean and mean of the distribution found. This analysis was used to determine how concentrated or dispersed a single feature was across all records.

To view the behavior of different distance measures, two random feature vectors were selected from the data. The two feature vectors were compared by computing the Minkowski distance between the two as r varied from 1 to 10. This included standard cases like Manhattan distance (r=1) and Euclidean distance (r=2). We plotted this to see how the distance changed as r varied, which informs our understanding of how different measures affect neighborhood based models like k-NN.

After the exploratory analyses were completed split the dataset up into a training and test (hold-out) subsets, using a 70-30 split. This split was accomplished with the use of the train-test-split() function of scikit-learn library. The features used were the pre-monsoon water levels from 2015 to 2022 and the labels were the binary classifications based on the levels from 2022.

After the training data was prepared the k-Nearest Neighbors (k-NN) classifier was trained with k=3 as a standard starting point to balance bias and variance. The training was carried out using scikit-learn's KNeighborsClassifier and then saved by the joblib library for future use in making predictions.

The model's performance was evaluated on the test set using the .score() method, which returns accuracy. In addition, the classifier was used to get predictions from the individual test vectors with the .predict() method. The predictions were then compared to the actual labels to see if the classification was correct.

To further investigate the effect of different neighborhood sizes, the number of nearest neighbors, k, was changed from 1 to 11 and accuracy was examined for each model. A line graph was established relating accuracy to the different neighbor values of k, which allowed us to determine the best value for our dataset.

Finally, a thorough performance analysis was performed using confusion matrices and classification reports. The confusion that was provided included true positives, true negatives, false positives and false negatives, for both the training and test set. These were calculated in order to find the other metrics, precision, recall, and F1-score, from the classification-report() method from scikit-learn. The model displayed consistency in accuracy when trained on the training and test data which demonstrated that appropriate balance exists between underfitting and overfitting. Overall review of the development process suggested that the model achieved a normal fit, with general applicability to real data

IV. RESULTS

This histogram displays the frequency distribution of groundwater depths during the pre-monsoon season of 2015. The data has been binned into ranges (buckets), highlighting a highly right-skewed distribution, with most groundwater levels falling between 0 and 25 meters below ground level. The highest frequency is observed in the 0–25 meter range, indicating shallow groundwater in most regions. A long tail extends beyond 100 meters, representing deeper groundwater levels in a few locations(as shown in Fig 1).

Accuracy comparison between NN (k=1) and k-NN classifiers for k = 1 to 11. The Nearest Neighbor (NN) model (k=1) achieves an accuracy of approximately 96.6 percent, while k-NN with k=3 improves performance to around 97.2 percent. Overall, the accuracy varies slightly across k values, peaking near k=4 and k=5. Accuracy improves from k=1 (NN) to k=3 (k-NN), with peak performance around k=4–5 before slightly declining.(as shown in Fig 2).

Based on our observations the train and test accuracy we got are:

Train Accuracy: 98.89 percent

Test Accuracy: 96.83 percent

Train Confusion matrix:

[1101 9]

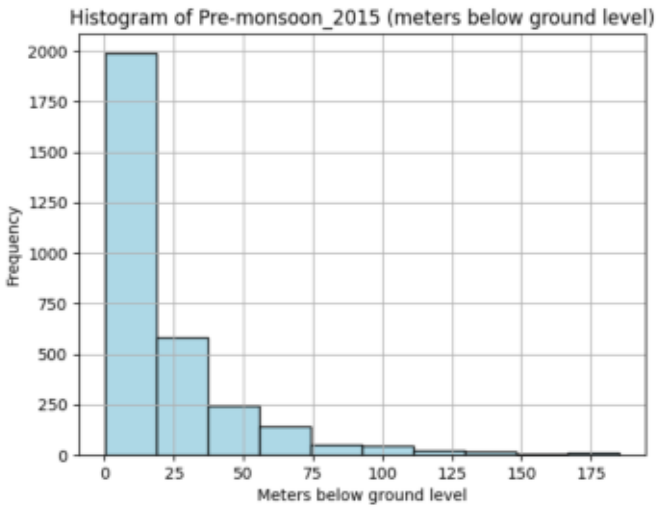

Fig. 1. Histogram of Groundwater Depth – Pre-monsoon 2015 (in meters below ground level).

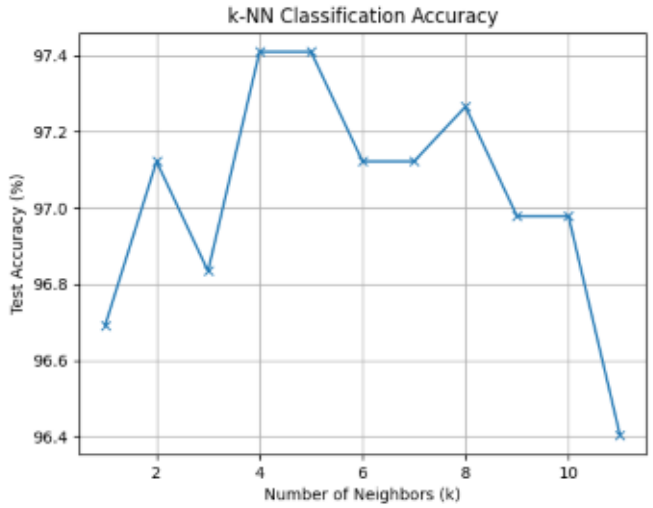

Fig. 2. Accuracy comaparison between NN and K-NN..

[9 500]
Test Confusion matrix:
[465 14]
[8 205]

TABLE I
TRAIN CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
0	0.99	0.99	0.99	1110
1	0.98	0.98	0.98	509
Accuracy			0.99	1619
Macro Avg	0.99	0.99	0.99	1619
Weighted Avg	0.99	0.99	0.99	1619

TABLE II
TEST CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
0	0.98	0.97	0.98	479
1	0.94	0.96	0.95	216
Accuracy			0.97	695
Macro Avg	0.96	0.97	0.96	695
Weighted Avg	0.97	0.97	0.97	695

REFERENCES

[1] M. B. Raj, C. Priya (2024). GROUND WATER LEVEL PREDICTION USING MACHINE LEARNING. International Research Journal of Modernization in Engineering Technology and Science.

[2] T. Zeng, K. Yin, H. Jiang, X. Liu, Z. Guo and D. Peduto (2022). Groundwater level prediction based on a combined intelligence method for the Sifangbei landslide in the Three Gorges Reservoir Area. Scientific Reports, 12(1). .

[3] F. Feng, H. Ghorbani and A. E. Radwan (2024). Predicting groundwater level using traditional and deep machine learning algorithms. Frontiers in Environmental Science, 12. h

[4] V. Kumar, N. Kedam, K. V. Sharma,K. M. Khedher and A. E. Alluqmani (2023). A comparison of machine learning models for predicting rainfall in urban metropolitan cities. Sustainability, 15(18), 13724.

[5] N. Mungale and J. Shinde (2024, January 12). Rainfall Forecasting: A Comparative Analysis of Deep Learning and Machine Learning Models with Application to Environmental Data.

[6] T. V. Tran, A. Peche, R. Kringel, k. Brömme and S. Altfelder (2025). Machine Learning-Based Reconstruction and Prediction of groundwater time Series in the Allertal, Germany. Water, 17(3), 433.

[7] A. Jari, E. M. Bachaoui, S. Hajaj, A. Khaddari, Y. Khandouch,, A. E. Harti, A. Jellouli and M. Namous (2023). Investigating machine learning and ensemble learning models in groundwater potential mapping in arid region: case study from Tan-Tan water-scarce region, Morocco. Frontiers in Water, 5.

[8] A. Ali, A. Ahmed and M. Abbod (2024, November 28). Groundwater level predictions in the Thames Basin, London over extended horizons using Transformers and advanced machine learning models.

[9] G. Tuysuzoglu, K. U. Birant and D. Birant (2023). Rainfall prediction using an ensemble machine learning model based on K-Stars. Sustainability, 15(7), 5889.

[10] P. N. Triveni, Dr. G. JawaherlalNehru, Dr. R. Santhoshkumar and S. BavanKumar (2023). A REVIEW ON RAINFALL PREDICTION USING MACHINE LEARNING ALGORITHMS: MLR AND ARTIFICIAL NEURAL NETWORK. In ResMilitaris: Vol. vol.13 (Issue n°4) [Journal-article].