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 RFR(Random Forest Regression), GB(Gradient Boosting) and SVR(Support Vector Regression). 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.

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 SVR(Support Vector Regression modelling) and EMR(Empirical Mode Decomposition 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 ML models and DL architectures in predicting groundwater levels. It featured models including LSTMs and CNNs, in addition to traditional ML 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

Q1.)To use regression with one feature, you're going to first split your data into training and test sets, select one feature (independent variable) from the training data set, and with the target variable (dependent variable) train your LinearRegression model with using one feature, then predict on your training set.

Q2.)To assess the model that has been trained, calculate MSE, RMSE, MAPE, and R^2 for both the training and testing predictions, and then compare the resulting metrics to understand the accuracy of your model and how it may be overfitting or underfitting..

Q3.)Re-run the regression procedure from problem A1, but this time use multiple features (or all the features) and again calculate metrics as you did in problem A2 to compare the performance of the model that only used one feature with the multiple feature model.

Q4.)In order to conduct k-means clustering, we simply remove the target variable from the dataset, fit a KMeans model with n-clusters=2 on the training data, and retrieve the cluster labels plus the cluster centers.

Q5.)For your clustering in A4, calculate the three evaluation metrics: Silhouette Score (describes cluster cohesion and separation), Calinski–Harabasz Score (describes the dispersion between clusters), and Davies–Bouldin Index (described as the average cluster similarity).

Q6.)Carry out k-means clustering for a number of values k, (e.g., from 2 - 10), then compute the Silhouette Score, Calinski–Harabasz Score, and Davies–Bouldin Index for each k. Once you have these metrics computed for each k, you can plot these metrics against k and visually determine the most appropriate number of clusters.

Q7.)Also use the elbow method: for k = 2 - 20, plot k against the model's inertia (i.e., the sum of squared distances within clusters) and observe where inertia reduces appreciably slower — that is, this latest k value would represent the essentially optimal value of k.

IV. RESULTS

The regression model performed excellently in the training set with only one attribute, with R^2 = 0.95, and a near-zero MSE and RMSE, and very low MAPE (0.03 percent). But in the test set, the performance dropped significantly, with R^2 = 0.00, higher errors, and a huge test MAPE (in the order of trillions), indicating instability in the model possibly due to division by zero, or zero/near zerotarget values. This is good evidence of overfitting and poor generalization from the train to test data.(Table 1)

The computed MSE, RMSE, MAPE, and R^2 for both train and test all confirm the same conclusion: that the model is memorizing the training data, rather than learning from it and generalizing, ultimately leading to poor performance on test.(Table 2)

TABLE I
REGRESSION COEFFICIENTS (Q1)

Feature	Coefficient
JAN_R/F_2018	0.0234
FEB_R/F_2018	0.4063
MAR_R/F_2018	0.4750
APR_R/F_2018	0.0138
MAY_R/F_2018	-0.0385
JUN_R/F_2018	0.0013
JUL_R/F_2018	-0.0006
AUG_R/F_2018	-0.0014
SEP_R/F_2018	0.0010
OCT_R/F_2018	-0.0682
NOV_R/F_2018	0.2186
DEC_R/F_2018	-0.0307
JAN_R/F_2019	0.0110
FEB_R/F_2019	-0.0355
MAR_R/F_2019	-0.0707
APR_R/F_2019	0.0131
MAY_R/F_2019	0.0034
JUN_R/F_2019	-0.0021
JUL_R/F_2019	-0.0009
AUG_R/F_2019	-0.0003
SEP_R/F_2019	0.0007
OCT_R/F_2019	0.0010
NOV_R/F_2019	-0.0142
DEC_R/F_2019	-0.0288
JAN_R/F_2020	0.0395
Intercept	0.3629
Training MSE	0.00
Training R^2	0.95

TABLE II
TRAIN AND TEST SET METRICS (Q2)

Metric	Train Set	Test Set
MSE	0.00	2.16
RMSE	0.00	1.47
MAPE	0.03%	1328010457387.68%
R^2	1.00	0.00

TABLE III
REGRESSION METRICS (ALL FEATURES)

Dataset	MSE	RMSE	MAPE (%)	R^2
Train	48.20	6.94	572774389609.56	0.71
Test	330.20	18.17	7494498313180.48	-1.78

TABLE IV
K-MEANS CLUSTERING METRICS (K=2)

Metric	Score
Silhouette Score	0.2844
Calinski-Harabasz Score	9.5640
Davies-Bouldin Index	1.1120

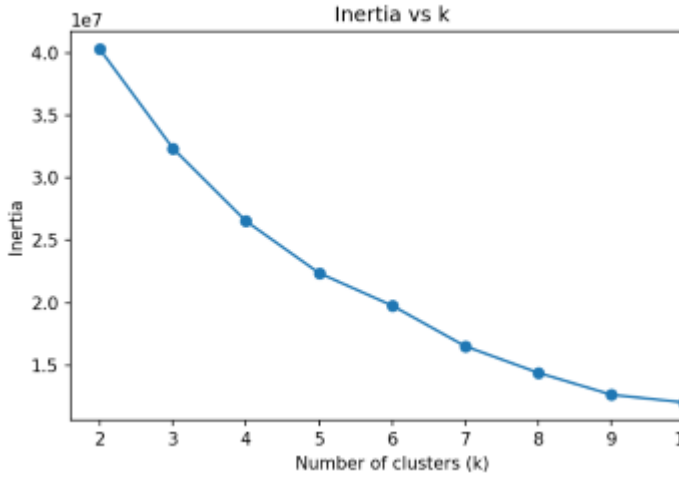

Fig. 1. Evaluation of K-means clustering performance for different values of k using Inertia.

Fig. 2. Evaluation of K-means clustering performance for different values of k using Silhouette Score.

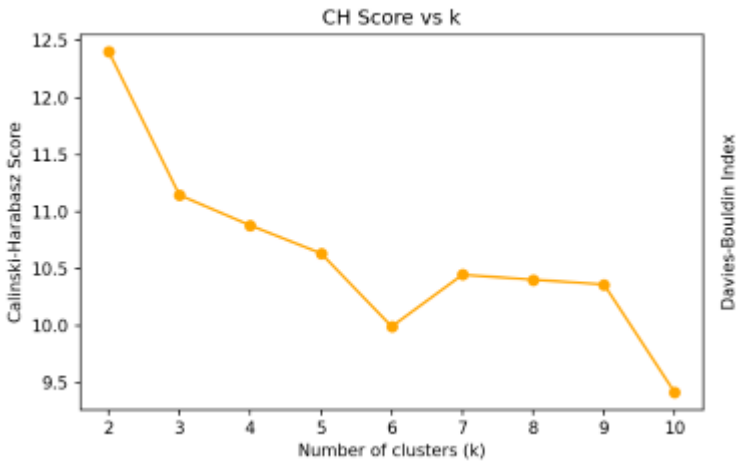

Fig. 3. Evaluation of K-means clustering performance for different values of k using Calinski–Harabasz Score.

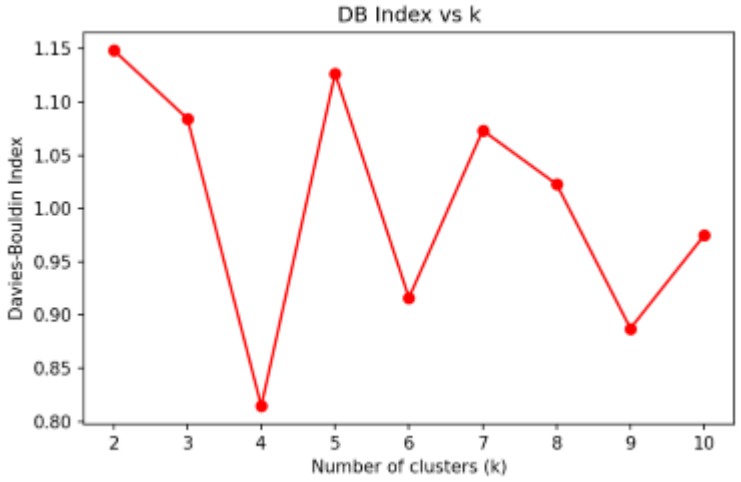

Fig. 4. Evaluation of K-means clustering performance for different values of k using Davies–Bouldin Index.

With all attributes, the training set R^2 improved to 0.71, but test R^2 dropped sharply to -1.78, along with substantial increases in error values, which indicated serious overfitting. The MAPE values were again extremely high due to the target values being equal to or near-zero, and like before, would not be reliable without some form of adjustment.(Table 3)

Clustering exhibits weak separability when k=2. The Silhouette Score (0.284) and Calinski-Harabaz Score (9.56) were low representing weak cluster definition. The Davies-Bouldin Index (1.11) can be considered moderate, although not desired.(Table 4)

When varying k from 2 to 10, the Silhouette Score was maximized at k=4, suggesting greater separation at k=4. The CH Score was highest at k=2 but began to experience decline after that. The Davies–Bouldin Index was lowest at k=4 and k=9, which indicates more compact clusters. All-in-all, k=4 seems to provide a good mix of separation and compactness.(Fig 1,2,3,4)

The Elbow Method plot also shows a marked inflection point at approximately k=4, reinforcing the Silhouette Score's suggested best-fit model. Therefore, k=4 is the best fit model with respect to cluster count, since it appears to balance both performance metrics and the rate of inertia reduction.(Fig 5)

performance on the training set but failed to generalize to the test set, with a drastic drop in R^2 and large errors. The multiple-feature regression improved training fit but performed even worse on the test data, confirming the overfitting issue. The excessively large MAPE values indicate that the dataset contains zero or near-zero target values, making MAPE unreliable in its raw form.

From the clustering tasks (Q4–Q7), k=2 clustering produced weak separation and low-quality clusters according to the Silhouette and Calinski–Harabasz Scores. When exploring multiple k values, k=4 consistently showed better separation

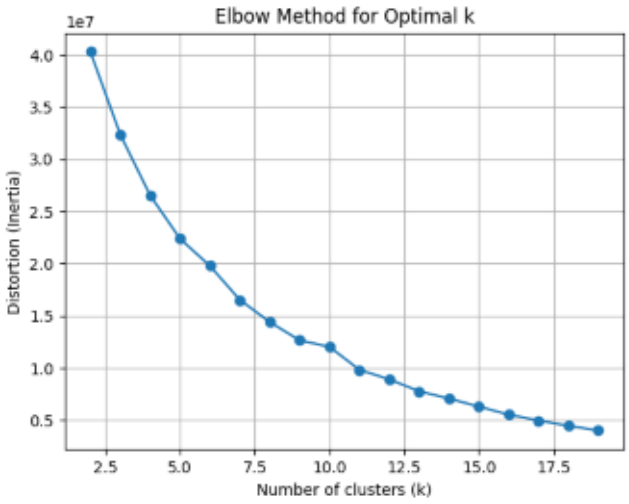

Fig. 5. Elbow method plot showing distortion (inertia) against the number of clusters k to determine the optimal cluster count.

(highest Silhouette, low Davies–Bouldin Index) and a clear bend in the Elbow Method plot, making it the optimal cluster choice for this dataset.

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].