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.)As part of preprocessing, continuous rainfall and percentage departure values were transformed into categorical values through the use of a custom binning function, with equal-width and equal-frequency binning methods applied; default parameters were used when a parameter was not supplied.

Q2.)Entropy for each attribute was calculated using the formula , with the probabilities derived from the frequency distribution of the binned values.

Q3.)The Gini index for each attribute was calculated using the formula, which measures impurity by generating probability for each category.

Q4.)To establish the root node of the decision tree, information gain was calculated for each feature and the feature with the highest gain was chosen as the root.

Q5.)Continuous attributes were binned either by using the equal-width method or the equal-frequency method parameter depending on the choice of parameter, and all attributes were categorical before any splits could take place.

Q6.)Armed with a decision tree based on our dataset, we greedily generated splits based on the rules of maximum information gain to create the tree. We then created a graph structure to visualize it, which displayed the composite relationships between the original root node and the leaf nodes.

Q7.)Two numeric features were selected to build a decision tree classifier. The decision boundary was then plotted in a 2D vector space and the decision boundaries illustrate how the decision tree has separated the feature space into different classes.

IV. RESULTS

TABLE I
COLUMN ENTROPY VALUES

Column	Entropy
JAN_R/F_2018	0.195909
JAN_%DEP_2018	0.195909
FEB_R/F_2018	0.583584
FEB_%DEP_2018	0.583584
MAR_R/F_2018	0.714653
⋮	⋮
OCT_%DEP_2022	1.529399
NOV_R/F_2022	0.583584
NOV_%DEP_2022	0.583584
DEC_R/F_2022	-0.000000
DEC_%DEP_2022	-0.000000

Entropy months, like OCT-DEP-2022 (1.529) indicates our data is more centrally dispersed across values range, suggesting greater variability and unpredictability from the districts. While low entropy months, like JAN-R/F-2018 (0.196) indicates months with rainfall values more densely grouped into fewer ranges; indicating more uniform conditions across

TABLE II
COLUMN GINI INDEX VALUES

Column	Gini Index
JAN_R/F_2018	0.058770
JAN_%DEP_2018	0.058770
FEB_R/F_2018	0.170799
FEB_%DEP_2018	0.170799
MAR_R/F_2018	0.222222
⋮	⋮
OCT_%DEP_2022	0.582185
NOV_R/F_2022	0.170799
NOV_%DEP_2022	0.170799
DEC_R/F_2022	0.000000
DEC_%DEP_2022	0.000000

TABLE III
BINNING RESULTS FOR JAN_R/F_2018

Index	Original Data	EWB (4 bins)	EFB (4 bins)
0	0.7	0	0
1	0.0	0	0
2	0.0	0	0
3	0.0	0	0
4	0.8	0	0
5	0.0	0	0
6	3.4	3	1
7	0.2	0	0
8	0.0	0	0
9	0.0	0	0

Fig. 1. District network with class labels

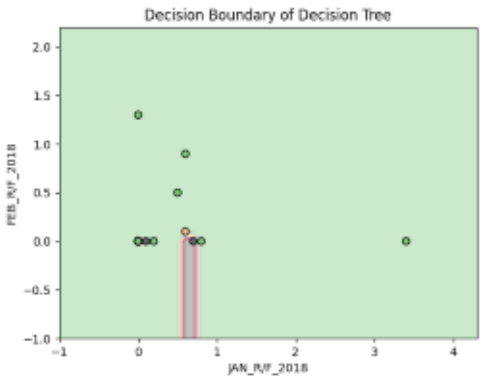

Fig. 2. Decision tree decision boundary

districts. An entropy value of 0, as is the case for DEC-R/F-2022, mirrors that all districts experienced rainfall in the same range suggesting no variability was present at all.(Table 1)

The values of the Gini Index measure the impurity or diversity of rainfall and percentage departure distributions among the districts. Low Gini values, like JAN-R/F-2018 (0.059), show that most of the districts were in the same rainfall range that depicted low diversity in values. Moderate values, like MAR-R/F-2018 (0.222), showed more of an even spread of rainfall within multiple bins. High Gini values, like OCT-DEP-2022 (0.582), showed more diverse distributions of values or equal representation among the bins. Gini Index values of 0, such as for DEC-R/F-2022 and DEC-DEP-2022, show no variability or that each of the districts had values in the same bin. (Table 2)

For the dataset, the most rainfall values in JAN-R/F-2018 fall very near to zero. The nearly-zero values assigned to the first bin using equal-width binning put almost all the values in the first bin, which demonstrates the skewness of the data. Even using equal-frequency binning, many values fell into the same bin due to so many zeros, which reduces our capacity to identify even small differences. Essentially, equal-width binning highlights the outliers, whereas equal-frequency binning distributes the data proportionately but risks masking any differences that may or may not exist when the regular values are so necessarily repeated.(Table 3)

The decision tree indicates District as the root node because it has the highest information gain for predicting January 2018 rainfall. Most of the districts fell in the lower rainfall bin (0), which suggested a very low amount of rainfall. The districts of Hanumangarh and some additional districts designated as bin (3) had a higher rainfall compared to the rest. This result shows that rainfall in January 2018 was based mainly on District, since most of the districts were dry and some had higher levels of rainfall.5Q.)

The decision tree selects ”District” as the root node, or the topmost attribute to check, meaning in terms of information gain on the target attribute (JAN-R/F-2018 after binning), the district attribute has the most information gain. In the decision tree visualization, showing the splits of the independent attributes using yes/no (i.e. case 0 and case 1) almost all districts yield the same class (class 0) with the exception of Hanumangarh which is categorized as class 3. This would appear to indicate that the rainfall behavior in January 2018 across the majority of districts is similar, with little variability except for some anomalies (like Hanumangarh) and thus the tree is a fairly simple one; it mainly splits on the district attribute but since most outcomes share one class as the clear majority class indicates that there is not much variability or diversity in rainfall for that month across the state of Rajasthan.(Fig 1)

The decision boundary indicates that, for most regions, the label is one dominant class. Thus, for January then February, the amount of rainfall is usually very low. The tree makes narrow decisions for seldom higher rainfalls. The data is extremely imbalanced, and the model is mainly predicting the majority class.(Fig 2)

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].