

RESEARCH ARTICLE

Precision in Prediction: Groundwater Level Forecasting with Random Forest Regression in Coimbatore's Upper Bhavani River Basin Area

Ravanashree. M^{1*}, Arunadevi, K ¹, Raviraj, A², Balaji kannan², Sumathi, CS²

^{1*}Department of Soil and Water Conservation Engineering, Agricultural Engineering College and Research Institute, Tamil Nadu Agriculture University, Coimbatore, India

²Agricultural Engineering College and Research Institute, Tamil Nadu Agriculture University, Coimbatore, India

ABSTRACT:

This study presents a highly accurate method for predicting groundwater levels using Random Forest Regression (RFR) in Coimbatore, India's Upper Bhavani River Basin area. Daily groundwater level data from 1995 to 2021 were analysed along with relevant environmental factors. The model demonstrated exceptional presentation, with R² values of 0.9999 and 0.9994 for training and testing datasets.

Received: 20 Jan 2025

Revised: 19 Feb 2025

Accepted: 18 Mar 2025

Keywords: Groundwater level prediction, Random Forest Regression, Upper Bhavani River Basin, Machine learning.

INTRODUCTION

Accurate prediction of groundwater levels is essential for effective water resource management, especially in regions facing water scarcity or overexploitation of aquifers. The Upper Bhavani River Basin in Coimbatore, Tamil Nadu, is one area where sustainable groundwater management is vital for agriculture and urban water supply. Traditional methods often struggle to capture the complex interactions between various environmental factors and groundwater dynamics. Machine learning approaches have shown promise in addressing these challenges in recent years. This study explores the application of Random Forest Regression (RFR) in predicting groundwater levels with high precision in this critical watershed.

2. Study Area:

The Upper Bhavani River Basin is located in the western part of Coimbatore district, Tamil Nadu, India. The Upper Bhavani River Basin, an integral part of the

larger Bhavani River system and a significant tributary of the Cauvery River, is situated in the Western Ghats. The basin is predominantly agricultural, with Horticulture crops being the primary cultivation. Hydro geologically, the area features a description of the aquifer system, rock types, and groundwater availability. The Upper Bhavani River Basin is crucial for irrigation, domestic water supply, and industrial use in the region, making accurate groundwater level predictions essential for sustainable water resource management. The Upper Bhavani Basin faces challenges such as specific water-related issues in the area, e.g., seasonal water scarcity, overexploitation of groundwater, making accurate groundwater level prediction essential for sustainable water management.

Groundwater level forecasting has become increasingly important for sustainable water resource management, especially in regions facing water scarcity. Over the past decade, machine learning

Corresponding author mail: arunadeviswce@gmail.com



Copyright: © The Author(s), 2025. Published by Madras Agricultural Students' Union in Madras Agricultural Journal (MAJ). This is an Open Access article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited by the user.

techniques have gained prominence due to their ability to handle complex, non-linear relationships in hydrological systems. Random Forest Regression (RFR) has emerged as a powerful tool for groundwater level prediction. Rajaei et al. (2019) compared various machine learning techniques and found that RFR often outperforms other methods in terms of accuracy and robustness. They attributed this to RFR's ability to handle high-dimensional data and its resistance to overfitting. In a study focused on semi-arid regions, Sahoo et al. (2017) demonstrated the effectiveness of RFR in predicting groundwater levels under varying climatic conditions. They highlighted the importance of feature selection in improving model performance. To address data scarcity, Naghibi et al. (2020) proposed a hybrid approach combining RFR with other data-driven methods. Their results showed improved prediction accuracy, especially in areas with limited historical data. Raghavendra et al. (2014) utilized multiple linear regression (MLR) for predicting pest incidence in cotton crops. "Raghavendra et al. (2014) also used MLR for predicting pest incidence of cotton. "For the specific context of river basins, Chen et al. (2020) applied RFR to forecast groundwater levels in a complex river-aquifer system. They found that incorporating river stage data significantly enhanced the model's predictive power. In agricultural watersheds, similar to the Upper Bhavani River Basin, Nair and Kumar (2018) demonstrated the superiority of RFR over traditional time series models. They emphasized the importance of including land use and irrigation data as input features. Recent work by Prasad et al., (2022) has focused on integrating remote sensing data with RFR models. Their approach showed promise in improving long-term groundwater level forecasts, particularly in data-scarce regions.

3.1 Data Collection:

Daily groundwater level data, measured meters below ground level (mbgl), were collected from 1995 to 2021. The data was sourced from the Water Resources Information System (WRIS). In addition to groundwater levels, the following environmental variables were included in the analysis. Rainfall data was sourced from the Indian Meteorological Department (IMD), and additional climatic parameters such as soil moisture, relative humidity, minimum temperature, and maximum temperature were collected from NASA POWER. The basin is predominantly agricultural, with Coconut being the primary cultivation.

Selection of parameters

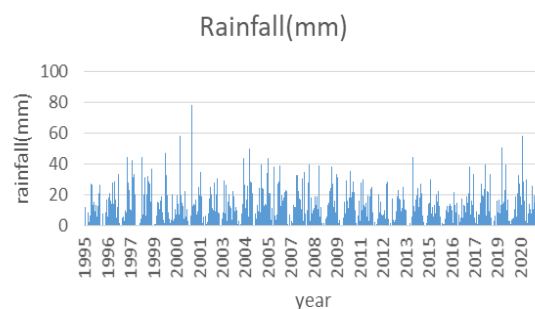


Fig 1 Rainfall over the year

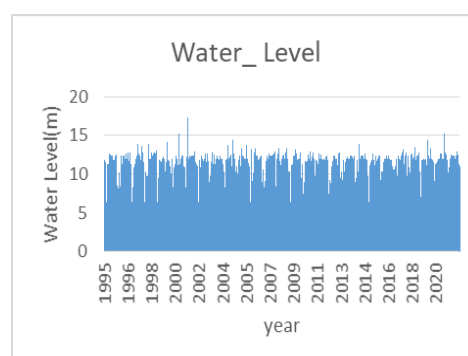


Fig 2 Water level (m) over year

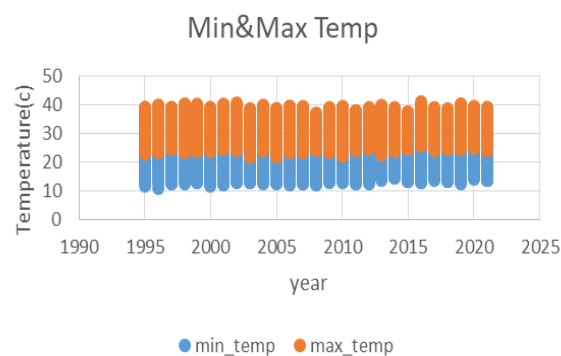


Fig 3 Minimum and Maximum Temperatures over the year

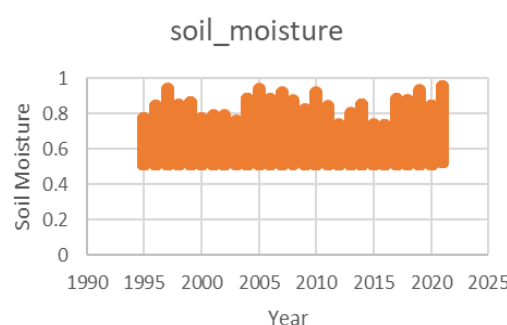


Fig 4 Soil Moisture over the year

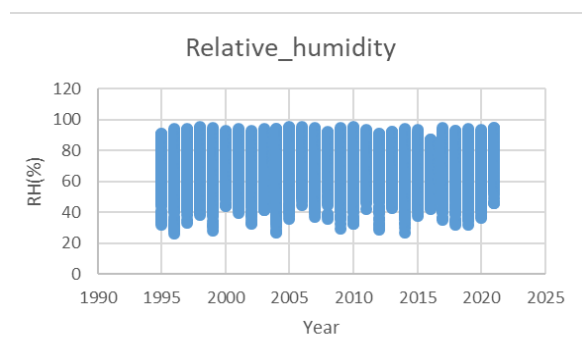


Fig 5 Relative Humidity over the year

Correlation coefficients, particularly the Pearson correlation coefficient (r), are crucial in assessing linear relationships between variables. The formula $r = \frac{\sum[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{[\sum(x_i - \bar{x})^2][\sum(y_i - \bar{y})^2]}}$ calculates this coefficient, where x_i and y_i represent individual data points, and $\bar{x}$ and $\bar{y}$ denote the means of the x and y variables respectively. The process involves calculating means, computing differences from these means, determining products of these differences, summing these products and squares of differences, and dividing the sum of products by the square root of the product of summed squares. The resulting r -value ranges from -1 to 1, with -1 indicating a perfect

negative linear relationship, 1 signifying a perfect positive linear relationship, and 0 suggesting no linear relationship. Values between 0 and 1 (or 0 and -1) indicate varying degrees of positive (or negative) linear relationships. This method forms the foundation for creating correlation heatmaps, visually representing relationships among multiple variables.

The heat map reveals several important relationships among environmental variables. Rainfall shows positive correlations with both soil moisture and water level, indicating that increased precipitation leads to higher soil moisture content and elevated water levels. Conversely, relative humidity exhibits negative correlations with minimum and maximum temperatures, suggesting higher temperatures are associated with lower humidity levels. A strong positive correlation exists between minimum and maximum temperatures, which is expected given their related nature. Additionally, water level and soil moisture demonstrate a positive correlation, implying that higher water levels coincide with increased soil moisture. These interrelationships provide valuable insights into the complex dynamics of environmental factors, highlighting how changes in one variable can influence others within the ecosystem.

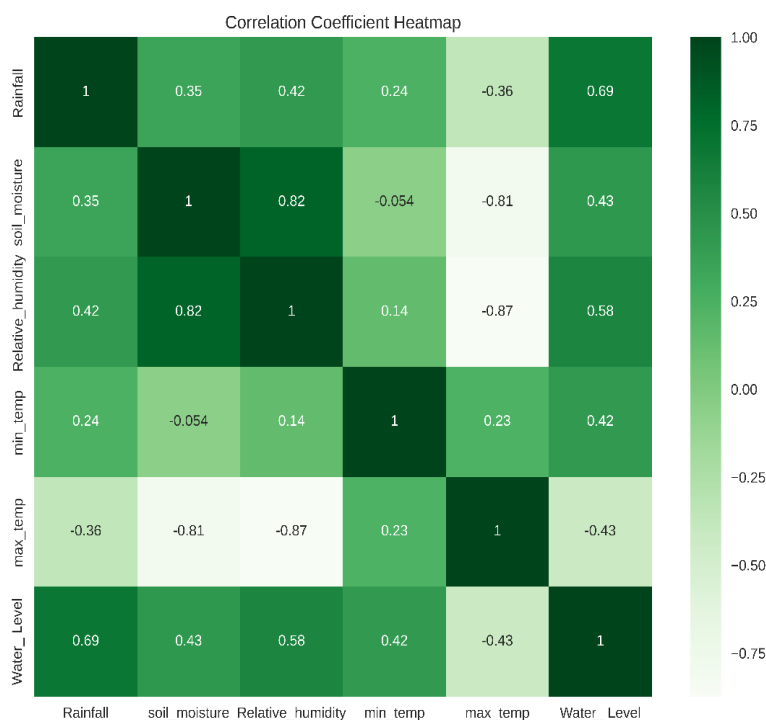


Fig. 6. Correlation heat map

3.2 Data Pre-processing:

To confirm data quality, rows with missing values (NA) were removed using the pandas function `df.dropna(inplace=True)`. The dataset was then split into training (70%) and testing (30%) sets using the `train_test_split` function from scikit-learn, with a random state 42 for reproducibility. Feature scaling was performed using Standard Scaler to normalize the input variables, which is crucial for many machine learning algorithms to perform optimally.

3.3 Model Development:

A Random Forest Regression model was developed using the pre-processed data. The model was trained on the scaled training dataset and evaluated on both the training and testing sets.

$$\hat{y} = 1/B * \sum_{b=1 \text{ to } B} fb(x)$$

Where:

$\hat{y}$ = the predicted output

B = the number of trees in the forest

$fb(x)$ = the prediction of the bth tree

x = the input features

$$\hat{y}(x) = 1/B * \sum_{b=1 \text{ to } B} \sum_{m=1 \text{ to } M} cmb * I(x \in Rmb)$$

Where:

M = the number of regions (leaf nodes) in each tree

cmb = the predicted value in region m of tree b

Rmb = the mth region in the bth tree

$I(\bullet)$ = an indicator function that equals 1 if the condition is true, 0 otherwise

This formula shows that each tree partitions the feature space into regions, and the prediction for a given input is determined by which region it falls into for each tree. The final prediction is the average of these individual tree predictions.

4. RESULTS AND DISCUSSION

4.1 Model Performance

The RFR model demonstrated exceptional performance in predicting groundwater levels.

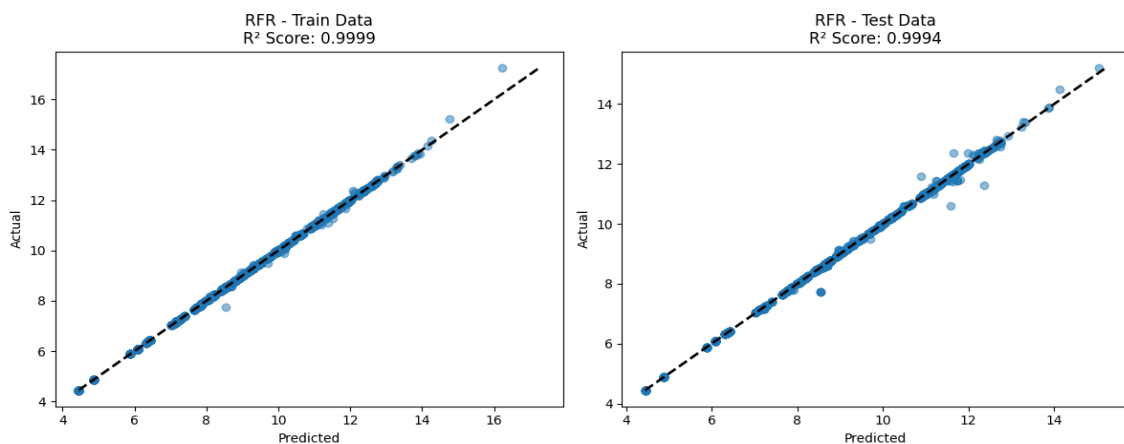


Fig. 7 Prediction of water level for train data

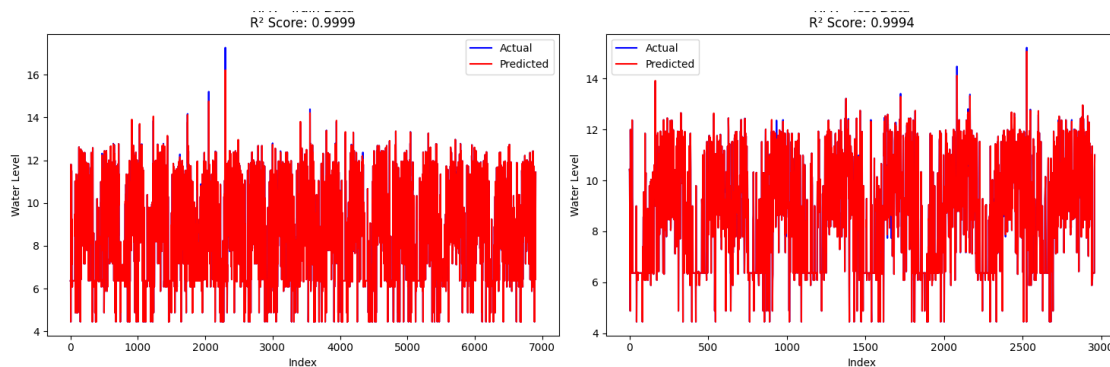


Fig. 8 Prediction of water level for test data



Table 1: Model Performance

Metric	Training	Testing
R ²	0.9999	0.9994
MSE	0.0004	0.0025
RMSE	0.0196	0.0499
MAE	0.002	0.0062

The extremely high R² values (0.9999 for training and 0.9994 for testing) indicate that the model explains nearly all the variance in both datasets, suggesting an exceptionally high level of accuracy. The very low Mean Squared Error (MSE) values (0.0004 for training and 0.0025 for testing) reflect minimal average squared differences between predicted and actual values, demonstrating excellent model precision. The Root Mean Squared Error (RMSE) values (0.0196 for training and 0.0499 for testing) further confirm the model's high accuracy. The slight increase in RMSE for the testing data is expected and indicates good generalization. The low Mean Absolute Error (MAE) values (0.002 for training and 0.0062 for testing) show that the average magnitude of prediction errors is minimal.

4.2 Model Interpretation

The slight increase in error metrics from training to testing data is standard and indicates that the model generalizes exceptionally well with negligible over-fitting. The consistently low error values across all metrics suggest that the RFR model has effectively captured the underlying patterns in the data.

4.3 Limitations and Future Work

While the model shows remarkable performance, future studies could explore the following areas:

- Investigating the relative importance of each input variable in predicting groundwater levels
- Comparing the RFR model with other machine learning algorithms or traditional hydrological models
- Extending the study to include longer-term predictions or different geographical regions

5. Conclusion

This study demonstrates the high potential of Random Forest Regression in predicting groundwater levels with exceptional accuracy. The model's performance, as evidenced by near-perfect R² values and very low error metrics, suggests it can be a valuable tool for water resource managers and policymakers.

This approach can contribute to more effective and sustainable groundwater management strategies by accurately predicting groundwater levels.

Acknowledgments:

The authors sincerely thank TNAU for providing groundwater level data. We are also grateful to Agricultural Engineering College and Research Institute, Coimbatore for their assistance and access to computational resources. Special thanks to ArunaDevi K for valuable guidance, and to the anonymous reviewers for their constructive feedback.

REFERENCES

- Rahmati, O., Pourghasemi, H. R., & Melesse, A. M. (2016). Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: A case study at Mehran Region, Iran. *Catena*, 137, 360-372. <http://dx.doi.org/10.1016/j.catena.2015.10.010>
- Naghibi, S. A., Pourghasemi, H. R., & Dixon, B. (2016). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environmental Monitoring and Assessment*, 188(1), 44. <https://link.springer.com/article/10.1007/s10661-015-5049-6>
- Ranganai, E., & Kubanga, K. P. (2015). A comparative study of hydrological drought forecasting techniques in the Upper Limpopo River Basin of Botswana. *Journal of Hydrology: Regional Studies*, 4, 153-171. <http://dx.doi.org/10.5194/hess-19-1695-2015>
- Prasad, R., Deo, R. C., Li, Y., & Maraseni, T. (2018). Input selection and performance optimization of ANN-based streamflow forecasts in the drought-prone Murray Darling Basin region using IIS and MODWT algorithm. *Atmospheric Research*, 197, 42-63. <http://dx.doi.org/10.1016/j.atmosres.2017.06.014>
- Chen, W., Li, X., Wang, Y., Chen, G., & Liu, S. (2014). Forested landslide detection using LiDAR data and the random forest algorithm: A case study of the Three Gorges, China. *Remote Sensing of Environment*, 152, 291-301. <http://dx.doi.org/10.1016/j.rse.2014.07.004>
- Niu, W. J., Feng, Z. K., Cheng, C. T., & Wu, X. Y. (2018). A parallel multi-objective particle swarm

- optimization for cascade hydropower reservoir operation in southwest China. *Applied Soft Computing*, 70, 562-575. <https://doi.org/10.3390/w11102040>
- Sahoo, S., Russo, T. A., Elliott, J., & Foster, I. (2017). Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. *Water Resources Research*, 53(5), 3878-3895. <https://doi.org/10.1002/2016WR019933>
- Shiri, J., Kisi, O., Yoon, H., Lee, K. K., & Nazemi, A. H. (2013). Predicting groundwater level fluctuations with meteorological effect implications—A comparative study among soft computing techniques. *Computers & Geosciences*, 56, 32-44. <http://dx.doi.org/10.1016/j.cageo.2013.01.007>
- Baudron, P., Alonso-Sarria, F., García-Aróstegui, J. L., Cánovas-García, F., Martínez-Vicente, D., & Moreno-Brotóns, J. (2013). Identifying the origin of groundwater samples in a multi-layer aquifer system with Random Forest classification. *Journal of Hydrology*, 499, 303-315. <http://dx.doi.org/10.1016/j.jhydrol.2013.07.009>
- Kisi, O., Shiri, J., & Tombul, M. (2013). Modeling rainfall-runoff process using soft computing techniques. *Computers & Geosciences*, 51, 108-117. <http://dx.doi.org/10.1016/j.cageo.2012.07.001>
- Aswathi, V. S., & Duraisamy, M. R. (2018). Comparison of prediction accuracy of multiple linear regression, ARIMA and ARIMAX model for pest incidence of cotton with weather factors. <http://dx.doi.org/10.29321/MAJ.2018.000151>