Indianapolis Ground Water Analysis

Projecting water levels using temperature and precipitation

Abstract

This study examines the relationship between historical groundwater levels, temperature, and precipitation data to assess the potential impact of climate change on groundwater availability in the urban Indianapolis region. Using machine learning techniques, particularly Random Forest regression, we model groundwater level fluctuations and investigate how different climate variables influence these changes. The study also explores the assignment of climate data to groundwater stations using unsupervised learning, specifically K-Means clustering, to optimize station selection for improved model accuracy. Our results highlight key trends and the predictive capabilities of machine learning in hydrological studies. However, the results left room for improvement as the model underperformed for a subset of the data.

Contents

1 Introduction

Water resources are increasingly impacted by climate variability, requiring robust analytical methods to predict and manage groundwater levels. This study integrates NOAA temperature and precipitation data from Indianapolis International Airport and Eagle Creek Airport with USGS groundwater level records from Marion 35 as well as Marion 39 to develop predictive models. Given the geographic proximity of Indianapolis Airport and Eagle Creek Airport to the USGS monitoring stations Marion 35 and 39 respectively, we assess the suitability of climate data

assignment using unsupervised learning to improve groundwater level prediction. This research contributes to the understanding of climate-groundwater interactions and advances data-driven decision-making in water resource management.

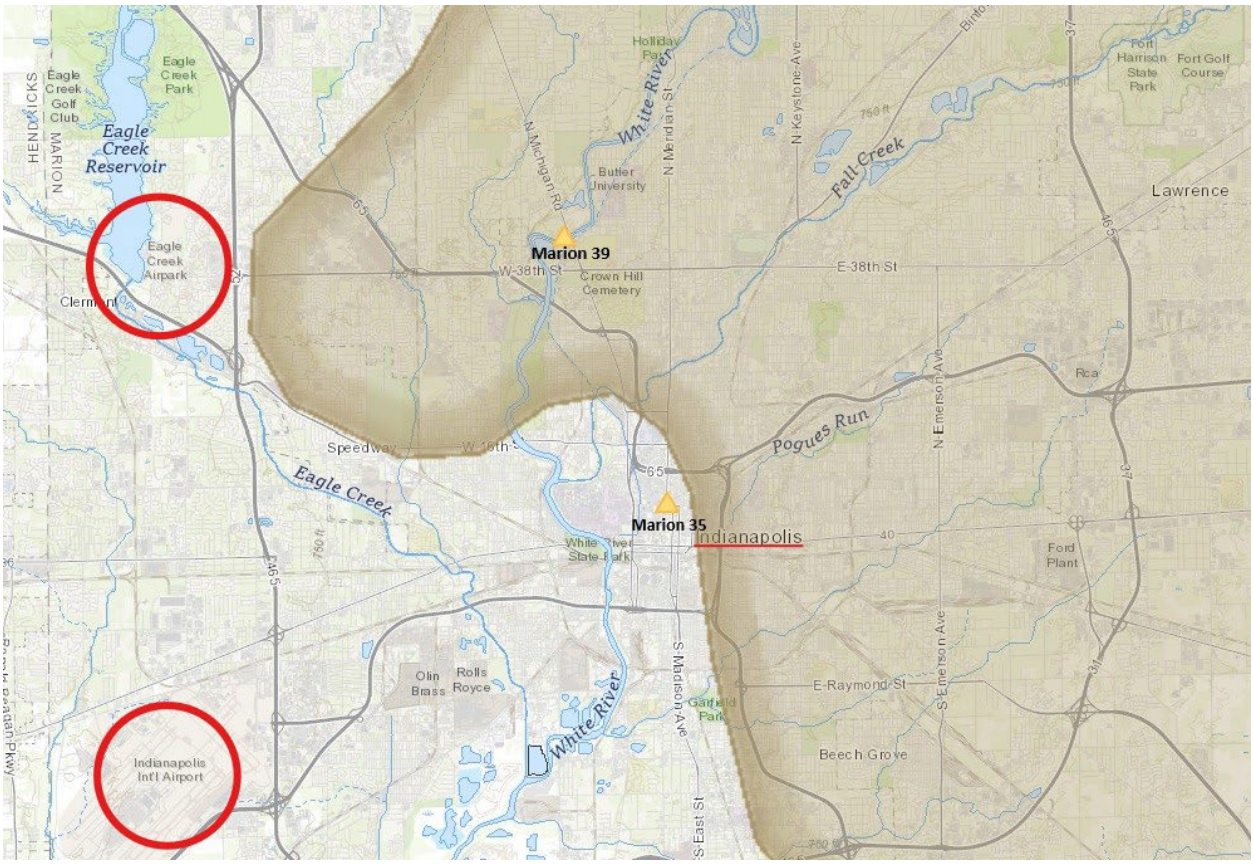

Figure 1: Selected sites with Aquifer Overlay in Indiana

2 Methodology

2.1 Data Collection and Cleaning

Climate Data (temperature, precipitation) was collected from NOAA for both Airports, while groundwater level data was collected from USGS for the Marion stations. All datasets were obtained as 'csv' files with daily data covering the period 2015-2024. Cleaning involved handling missing values, removing redundant information and ensuring consistency across datasets.

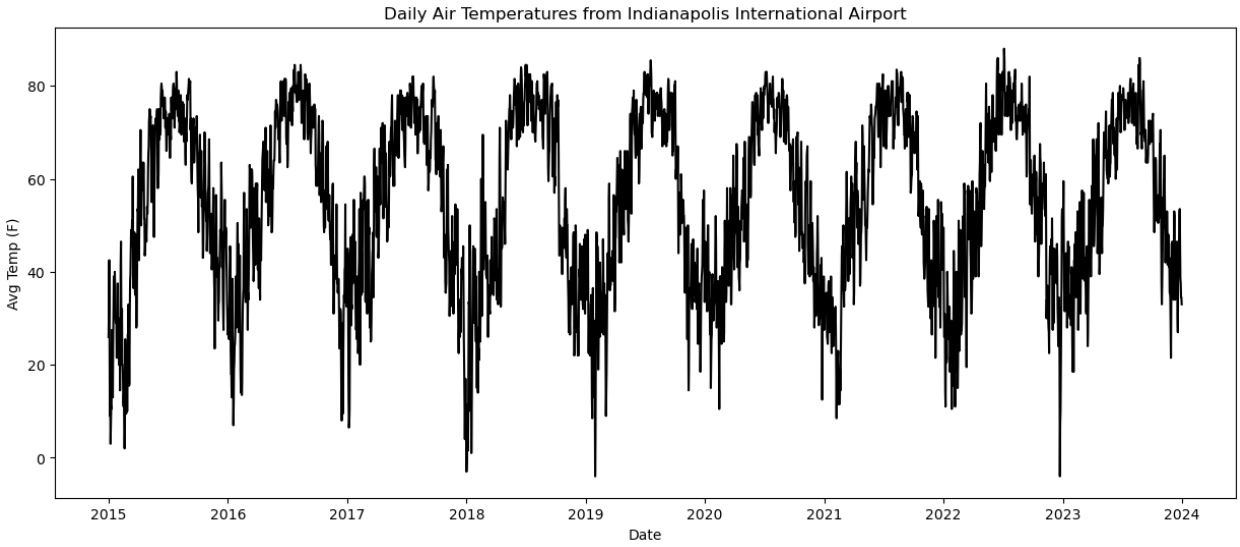

Fig 2.1.1: Historic Air Temperature data for Indianapolis International Airport

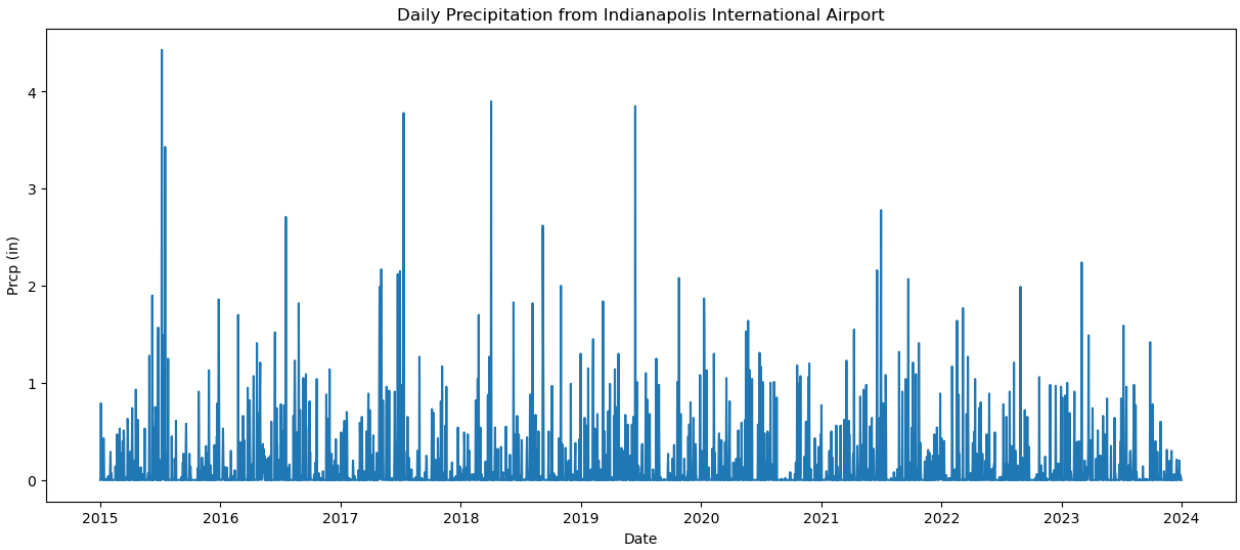

Fig 2.1.2: Historic Precipitation data for Indianapolis International Airport

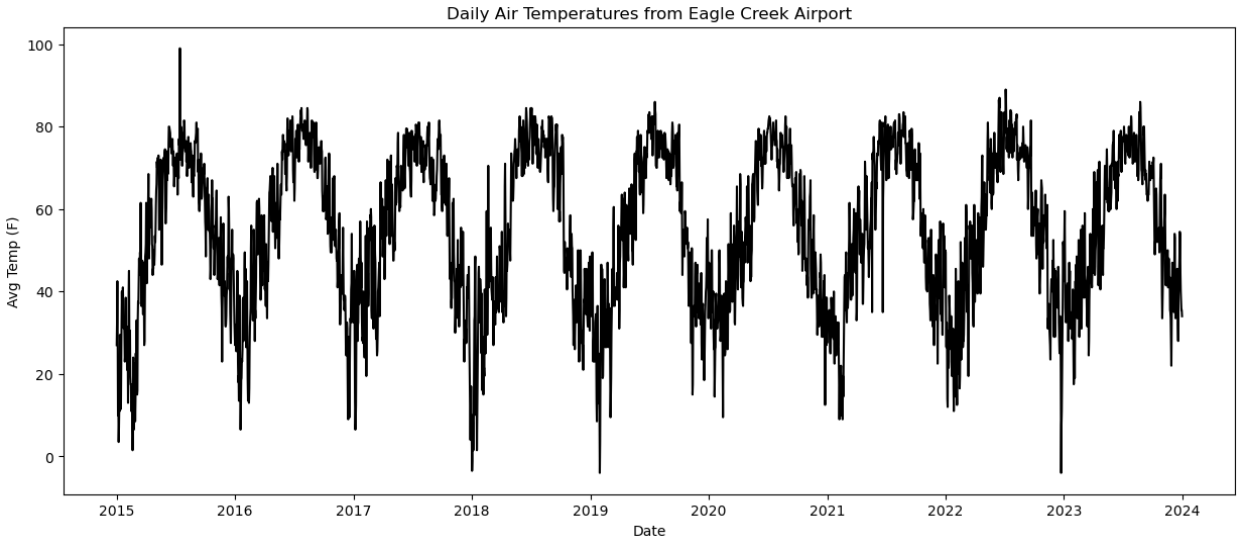

Fig 2.1.3: Historic Air Temperature data for Eagle Creek Airport

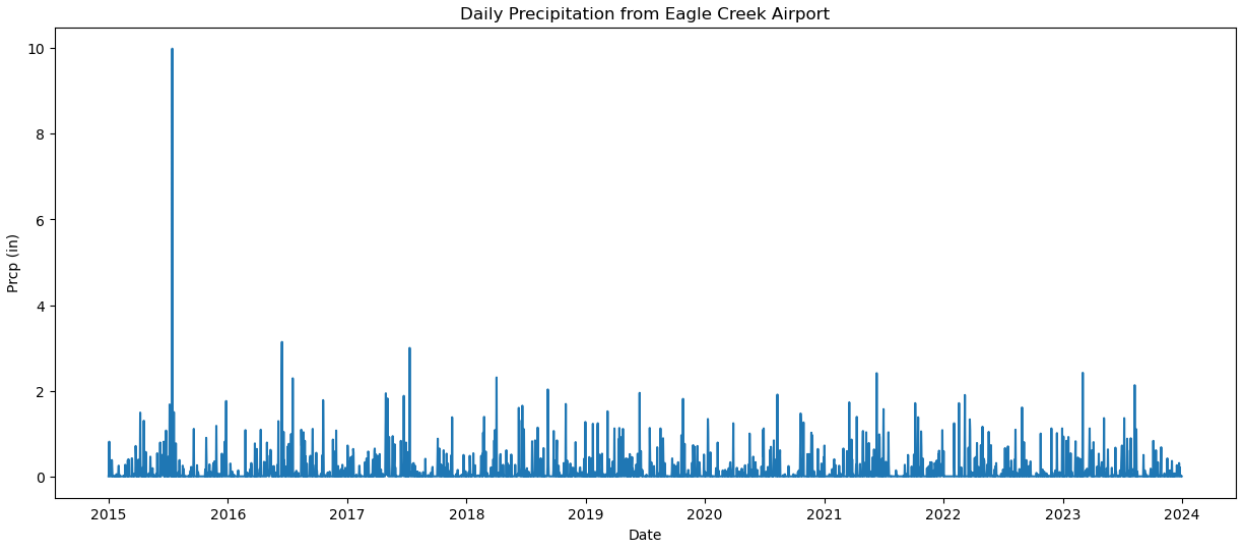

Fig 2.1.4: Historic Precipitation data for Eagle Creek Airport

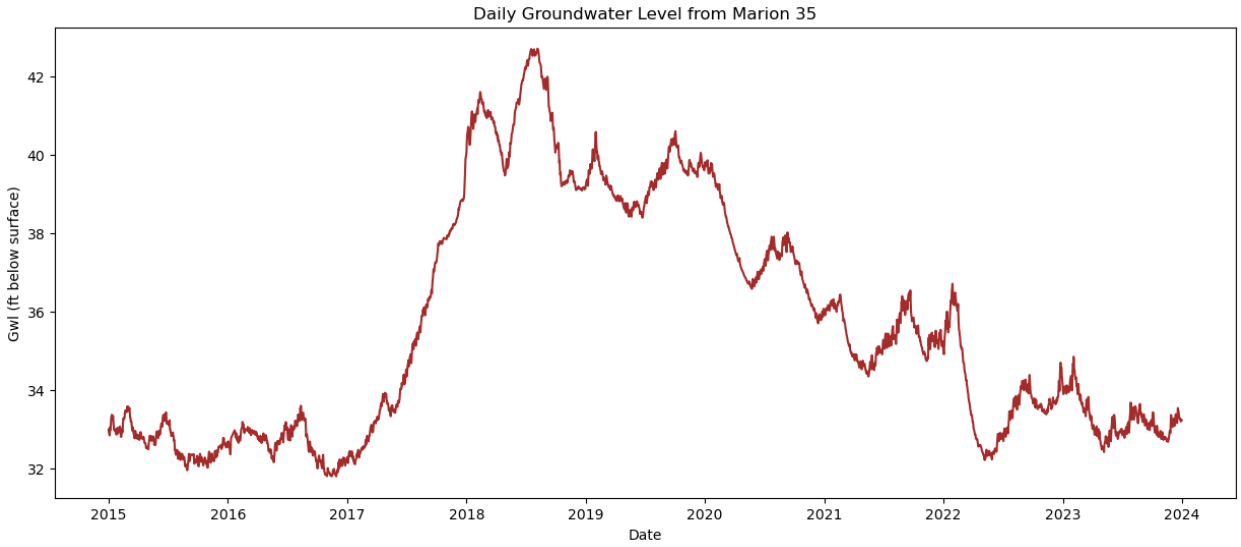

Fig 2.1.5: Historic Groundwater Level from Marion 35

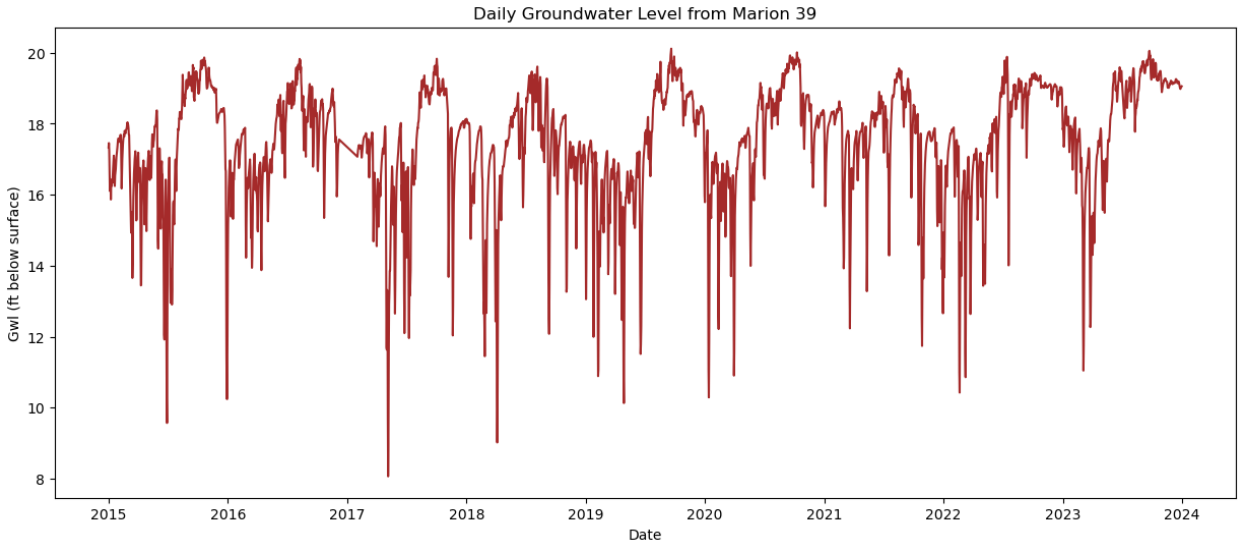

Fig 2.1.6: Historic Groundwater Level from Marion 39

2.2 Feature Engineering

Rolling averages (seven-day) and lagged values (one-day) were computed for temperature and precipitation to capture climate patterns. Similarly, sine and cosine values were calculated for each month and day value to capture temporality.

2.3 Model Selection and Training

A Random Forest regression model was chosen for groundwater level prediction, leveraging engineered features. The model was trained on data from 2015-2023 and validated on the year 2024, with performance evaluated using Mean Squared Error and R-squared Score.

For the Marion 35 Station over the evaluation period, the model Mean Squared Error was 3.32 and the R-squared Score was -1.44.

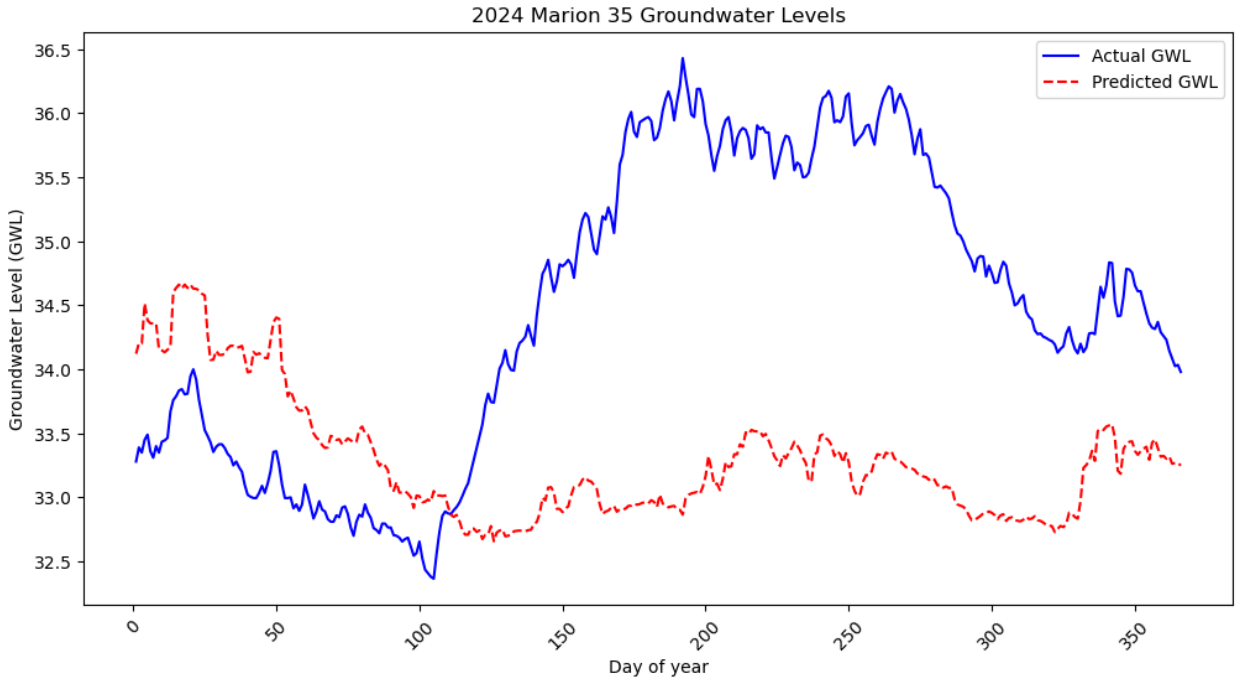

Fig 2.3.1: Evaluating Predicted Groundwater Level for Marion 35 in 2024

While the Marion 39 model Mean Squared Error was 0.87 and the R-squared Score was 0.67 over the evaluation period.

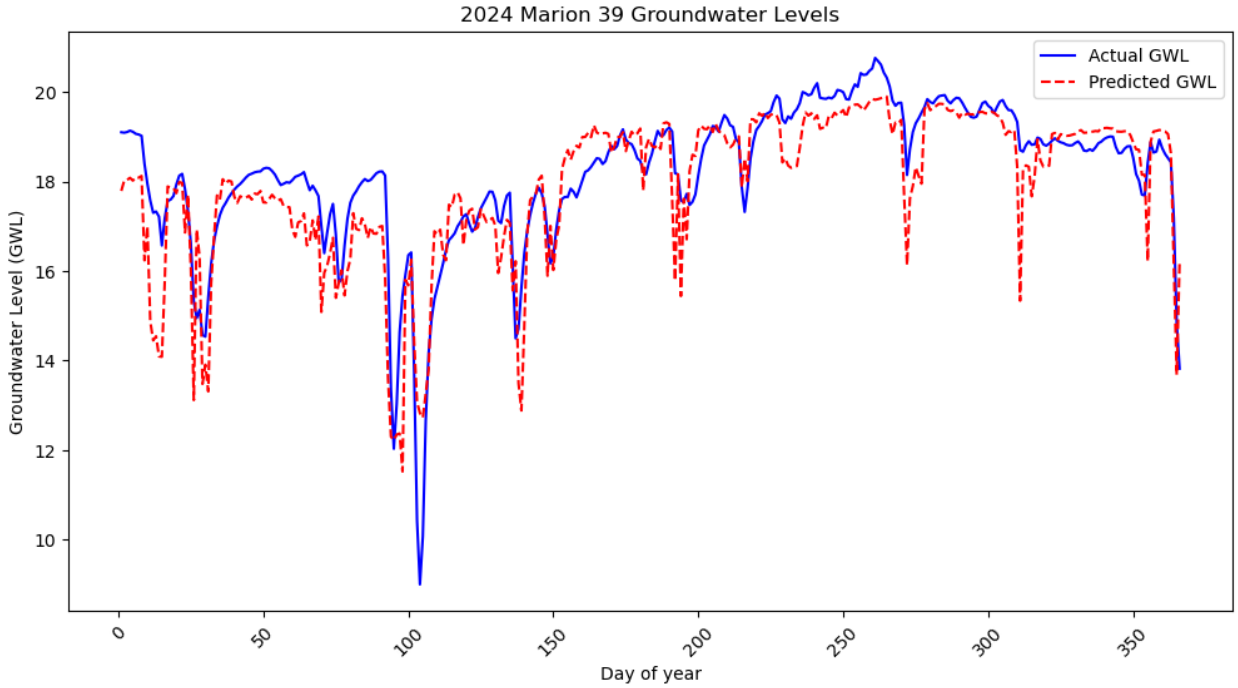

Fig 2.3.2: Evaluating Predicted Groundwater Level for Marion 39 in 2024

2.4 Unsupervised Learning for Climate Data Assignment

K-Means clustering was applied to temperature and precipitation variables to determine the most suitable climate station assignment per day for each groundwater monitoring station. This method tests the previously assumed pairing method of geological proximity and offers similar performance compared to it, reaffirming the validity of our previous assumptions.

For the Marion 35 Station over the evaluation period, the enhanced model Mean Squared Error was 3.32 and the R-squared Score was -1.45.

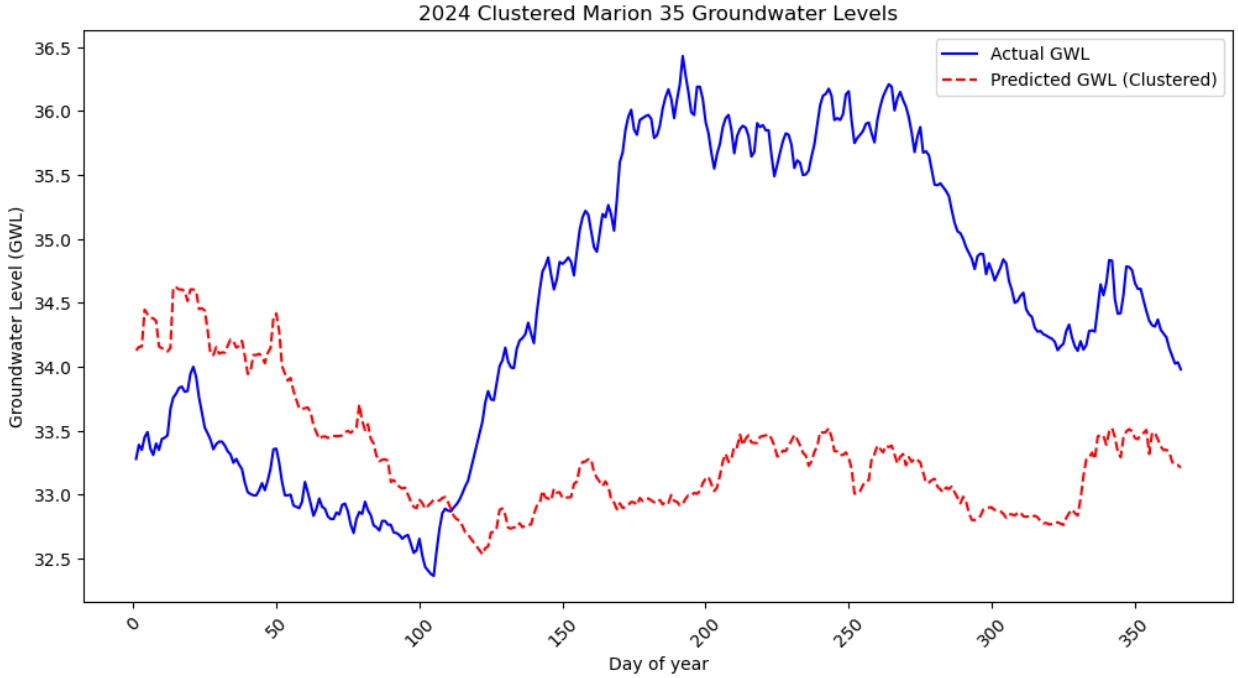

Fig 2.4.1: Evaluating Clustering Effect on Predicted Groundwater Level for Marion 35 in 2024

While the Marion 39 model Mean Squared Error was 0.88 and the R-squared Score was 0.64 over the evaluation period.

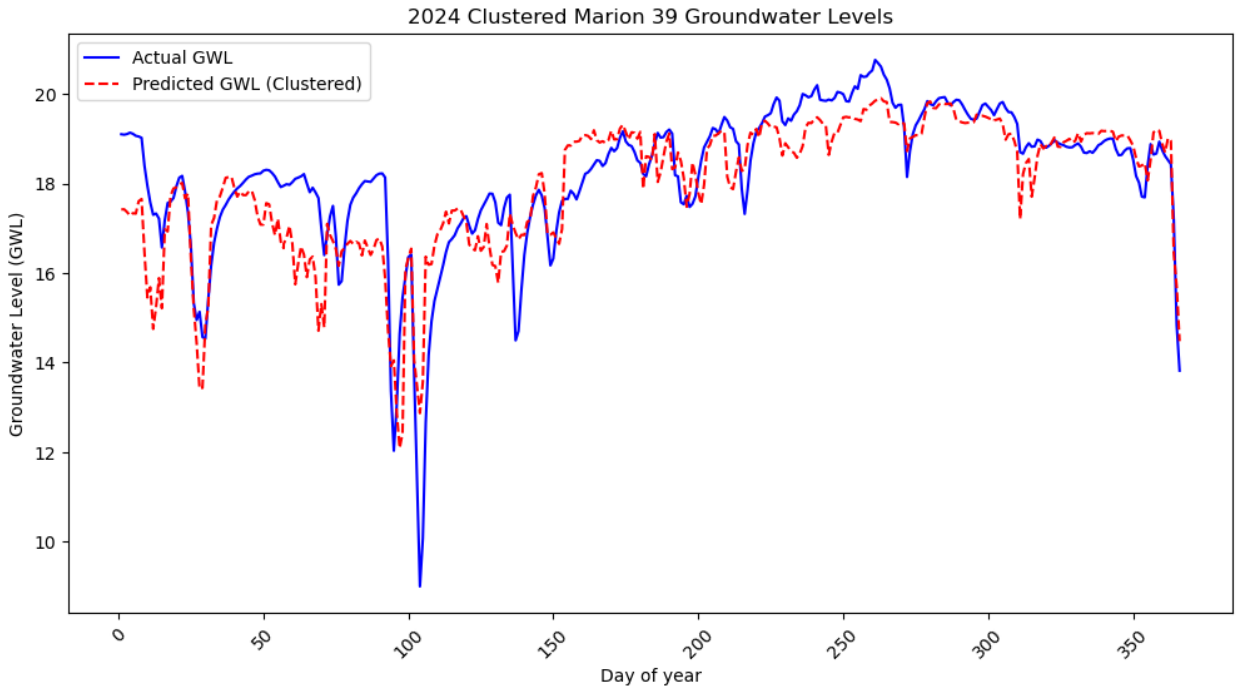

Fig 2.4.2: Evaluating Clustering Effect on Predicted Groundwater Level for Marion 39 in 2024

3 Results

Preliminary results indicate that clustering-based climate data assignment and geographical data assignment offer similar performance. The Random Forest model captures outlier measurements for Marion 39 effectively but underperforms in predicting the groundwater trends in the second half of 2024 for Marion 35, indicating a potential unexpected increase in consumption.

4 Limitations and Future Works

- The performance of the model could be improved by further scaling and normalizing features, as well as experimenting with different feature engineering techniques.
- Missing or estimated data could have affected model performance, introducing biases or inconsistencies that need to be addressed in future iterations.
- The lack of direct data measuring water consumption in urban regions limits the model's ability to fully capture groundwater depletion drivers.
- K-Means clustering did not provide a significant improvement in model accuracy, suggesting that other unsupervised learning techniques or alternative assignment methods should be explored.
- Future work could incorporate additional environmental factors such as soil moisture levels, vegetation indices, and human activity indicators to enhance predictive capabilities.

5 Conclusion

Machine learning proves to be a valuable tool in groundwater prediction using existing and publicly available data. However, limitations in data availability and modeling techniques highlight areas for further improvement. The results indicate that while the model performs well in certain cases, it struggles with unexpected variations in groundwater trends, particularly in urban regions. Further refinements in feature engineering, data scaling, and alternative unsupervised learning methods could enhance predictive performance and provide deeper insight into groundwater level fluctuations. Future research should also focus on integrating additional climate and urban consumption metrics to develop more comprehensive groundwater models.

6 Code Availability

All used code and data for the project is available here.

7 References

1. "NOAA" Accessed February 23, 2025. https://www.ncdc.noaa.gov/cdo-web/search

2. "USGS" Accessed February 23, 2025. https://dashboard.waterdata.usgs.gov/app/nwd/en/

3. Garret Flowers, Shivam Pandey "Colorado Ground Water Analysis" Projecting water levels based on RCP8.5 scenarios (December 2023)

4. Lin, Yi, and Kerstin Wiegand. "Low R2 in Ecology: Bitter, or B-Side?" *Ecological Indicators*, vol. 153, 1 Sept. 2023, pp. 110406–110406, https://doi.org/10.1016/j.ecolind.2023.110406.