Project 6

Kaden Hart, Robiul Islam, Razin Issa
Github: https://github.com/Razin1996/CS5830_Project6
Presentation:
https://docs.google.com/presentation/d/1KnK3FsWpIZCrS_m8uH86fSRLpyo4ZHo4x-IXCW2JGSk/edit?usp=sharing

Introduction

Accurate and timely prediction of baseflow in river systems is critical for effective water resource management, particularly in regions that rely heavily on rivers for irrigation, water supply, and ecological health. Understanding how different environmental factors like evapotranspiration, precipitation, and irrigation pumping impact baseflow over time can inform stakeholders such as water managers and policymakers to make data-driven decisions regarding water allocation, flood control, and drought management. In this analysis, we aim to develop a predictive model that estimates monthly baseflow using lagged variables of key hydrological factors. By capturing the temporal dependencies of baseflow, our model will help anticipate changes in river flows, allowing for proactive management and better adaptation to varying weather conditions. This is especially important for regions experiencing climate variability, as well as for agricultural sectors that depend on consistent water availability. Our analysis not only provides insights into the dynamic interactions of hydrological factors but also helps improve water use efficiency and sustainable planning for river basins. It contributes to building resilient water management systems, ultimately supporting both human needs and ecosystem conservation.

Dataset

The dataset utilized for this project comprises monthly observations of baseflow across various river segments, along with related hydrological variables such as evapotranspiration, precipitation, and irrigation pumping. Each entry represents a unique river segment at a specific month, capturing both temporal and spatial dynamics essential for accurate baseflow prediction. Key features include the observation date, segment ID, spatial coordinates, and observed values of relevant variables. To prepare the dataset for analysis, several cleaning steps were undertaken. The date was reformatted to represent days since January 1, 1900, making it more interpretable. Missing values were identified and addressed to ensure data completeness. All numerical features were standardized to ensure consistent scaling, and segment IDs were treated as categorical variables. The dataset's comprehensive coverage of both hydrological and temporal variables makes it highly suitable for this predictive modeling project, enabling a detailed exploration of how different factors contribute to baseflow changes over time.

Analysis Technique

The primary analysis method used in this project was **linear regression**, chosen for its ability to model relationships between continuous variables, making it well-suited for predicting monthly baseflow from hydrological factors. This technique aligns with the project's goal of understanding both temporal and spatial influences on baseflow, as discussed in the introduction. Linear

regression is effective for analyzing temporal features and spatial attributes (e.g., segment-specific effects), both of which are central to understanding baseflow variability over time. Additionally, linear regression provides clear interpretability, allowing us to identify the relative importance of different predictors, which is crucial for water management decisions.

To enhance the model's performance and better capture the dynamics of baseflow, several **feature engineering techniques** were applied. Another technique involved **one-hot encoding** of the categorical features, 'Month' and 'Segment ID', transforming them into a numerical format that allowed the model to recognize the distinct patterns associated with different months and river segments. Additionally, two novel features, **Month_Correlation** and **Segment_Correlation**, were introduced to aggregate the influence of different months and segments on baseflow. These features were designed to summarize relationships with observed baseflow, providing a holistic representation of temporal and spatial patterns. Throughout the analysis, the model was refined iteratively by testing different combinations of features, a process that helped identify the most significant predictors and improved both model accuracy and efficiency. The linear regression model, in combination with these techniques, proved to be suitable for the dataset, effectively implementing methods that aligned with the project's purpose of understanding baseflow variation over time and space.

Results

The results of the analysis revealed varying levels of success in predicting baseflow, with significant improvements observed after optimizing the feature set. The initial model, which included all numeric features, achieved an **R^2 score of 0.075** and a **mean squared error (MSE) of 3,042**, indicating that it was not capturing the patterns effectively. This prompted an iterative approach to feature selection, where less significant variables like evapotranspiration and irrigation pumping were excluded, resulting in a simplified model that used only 'y', 'Precipitation', and 'x'. While this model had a similar R^2 score of 0.065, the exclusion of additional numeric features confirmed that they were not contributing meaningfully to baseflow prediction.

The introduction of **one-hot encoded Segment IDs** marked a turning point in model performance, as it led to a substantial increase in the R^2 score to **0.808** and a decrease in the MSE to **631**. This finding emphasized the importance of segment-specific characteristics in predicting baseflow. The inclusion of one-hot encoded months provided further improvements, raising the R^2 score to **0.815** and lowering the MSE to **607**, suggesting that seasonal effects, while not as strong as segment-specific effects, still play a role in baseflow variability. The correlation analyses, visualized through various plots, provided additional insights into the relationships between features and observed baseflow. For instance, the **correlation of months** with observed baseflow, shown in Figure 1, revealed that certain months like February and March had stronger positive correlations, while others, such as July and August, had negative correlations. Similarly, the **correlation of segment IDs** with observed baseflow, depicted in Figure 2, highlighted that segments like **Segment 56** and **Segment 98** had significant positive correlations, suggesting strong location-based effects. The **correlation matrix** for the entire dataset, presented in Figure 3, confirmed that precipitation had the highest correlation among numeric variables, reinforcing its importance in baseflow prediction. While the linear regression model provided a clear

understanding of baseflow prediction, alternative approaches could offer additional insights or improvements. For example, regularization techniques like **Ridge** or **Lasso** could help manage multicollinearity among features, potentially enhancing performance by penalizing less significant variables. More complex models like **decision trees** or **random forests** could capture non-linear interactions, while **time-series models** such as ARIMA or LSTM could offer better temporal modeling of baseflow, especially when longer dependencies exist in the data.

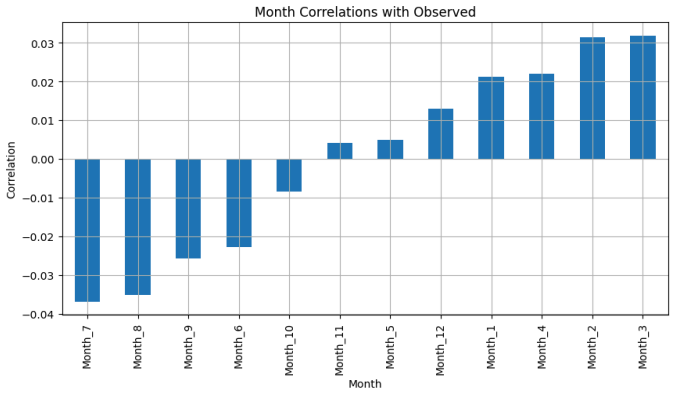

Figure 1: Correlation of months with observed baseflow.

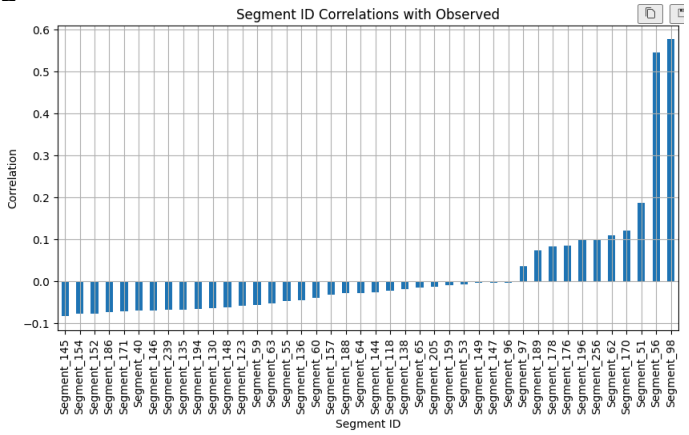

Figure 2: Correlation of segment IDs with observed baseflow.

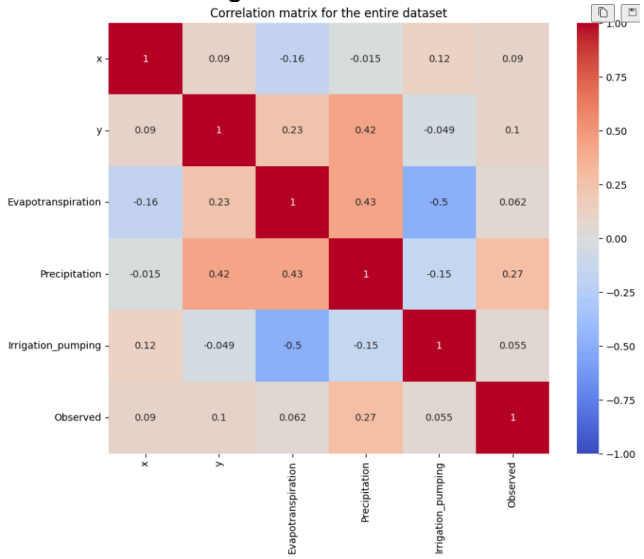

Figure 3: Correlation matrix for the entire dataset.

Project 6

Technical

The technical aspects of the analysis involved a comprehensive data preparation process, feature engineering, and iterative modeling. Initially, the dataset was loaded into a **pandas dataframe** from a CSV file. The **date column** was in a non-standard format, representing days since January 1, 0000, so it was adjusted by subtracting 693,963 to convert it to days since January 1, 1900. This column was then converted to actual dates using **pandas.to_datetime()**, making it more interpretable and suitable for analysis. A new 'Month' column was created by extracting the month from the 'Date' column, allowing for seasonal analysis. Missing values were addressed by removing rows with incomplete data, ensuring a clean dataset for model training.

Feature engineering was a crucial part of the analysis. The categorical features, 'Month' and 'Segment ID', were **one-hot encoded** to allow the model to recognize distinct temporal and spatial effects. Additionally, the creation of **Month_Correlation** and **Segment_Correlation** features provided a more comprehensive representation of temporal and segment-specific impacts on baseflow. These features were calculated by multiplying one-hot encoded values with their respective correlation values derived from the correlation matrix, effectively summarizing the relationships with observed baseflow.

The modeling process involved an iterative approach to selecting the optimal set of features. The initial model, which included all numeric features, performed poorly with an R^2 score of 0.075. Subsequent models focused on refining the feature set by testing various combinations. Adding one-hot encoded segment IDs led to a significant improvement, raising the R^2 score to 0.808. Further inclusion of one-hot encoded months provided a minor boost to the R^2, bringing it to 0.815. The evaluation metrics primarily used were **mean squared error (MSE)** and **R^2 score**, with the final model achieving an MSE of 607 and explaining approximately 81.5% of the variance in baseflow.

Conclusion

The analysis effectively demonstrates the temporal and segment-specific dependencies in baseflow prediction. While certain factors like precipitation showed some correlation with observed baseflow, the results highlight that river segment characteristics play a more dominant role in baseflow variability. The use of one-hot encoding for segment IDs and months significantly improved the model's predictive performance, making it a more viable tool for water resource management. The linear regression model, although simple, provides actionable insights and can be further enhanced by exploring more complex algorithms or incorporating additional hydrological variables. This analysis serves as a foundational step towards developing a reliable predictive tool for water managers and policymakers, ultimately supporting informed decision-making in river basin management.