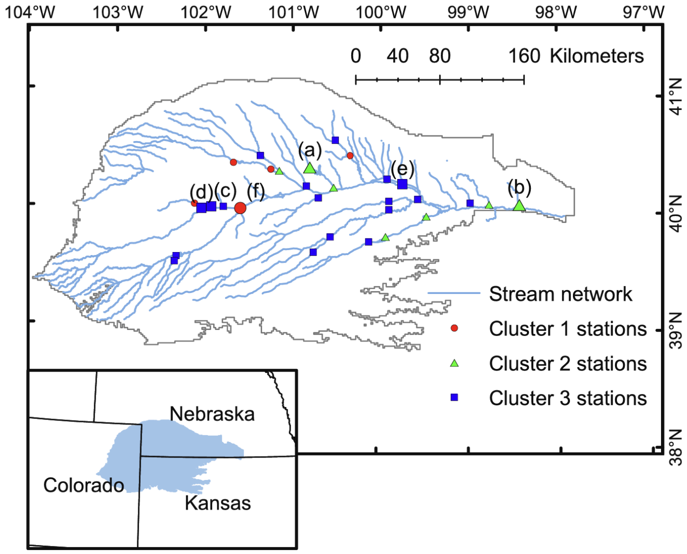
Introduction:

This project analyzes river baseflow data to understand how seasonal, spatial, and environmental factors influence variations in baseflow. Using a comprehensive dataset of rivers, monthly observations, and environmental variables such as ‘precipitation’ and ‘evapotranspiration’, we applied predictive modeling techniques to estimate baseflow. Scatter plots of actual versus predicted values helped evaluate model accuracy. Also, residual plots examined potential biases against environmental factors and spatial coordinates. Coefficient analyses reveal the significance of specific months and river segments in influencing baseflow, which highlights seasonal and spatial patterns. These visualizations provide insights into the model’s performance that offer a detailed assessment of baseflow dynamics in response to the mentioned environmental factors.The Slides and GitHub are available here.

Dataset:

The dataset utilized in this project focuses on river baseflow and includes a blend of temporal, geographical, and environmental features essential for predicting baseflow dynamics. Key features include the date, which has been converted into year, month, day, and season to capture temporal patterns, and the segment ID, which represents different river segments treated as categorical variables. The dataset also contains geographical coordinates (x and y) to locate each observation, along with environmental variables like evapotranspiration, precipitation, and irrigation pumping that influence baseflow. Observed baseflow serves as the target variable for prediction. To enhance the dataset’s utility, additional columns were created to analyze seasonal and yearly trends, and One-Hot encoding was applied to categorical variables, preparing the data for a multiple linear regression model. With 15,591 samples and no missing values, this clean dataset provides a comprehensive analysis of baseflow variations across temporal and spatial dimensions.



Analysis Technique:

Our project employs multiple analysis techniques to understand and predict river baseflow behavior effectively. The primary method is multiple linear regression, chosen for its ability to model relationships between baseflow and various independent variables, including evapotranspiration, precipitation, and irrigation pumping. The dataset required extensive preprocessing, including temporal feature extraction (e.g., year, month, season) and One-Hot encoding of categorical variables like segment ID and season, making it suitable for regression analysis. Exploratory Data Analysis methods, like scatter plots, residual plots, and correlation matrices were applied to uncover patterns, trends, and relationships among variables. Model performance was evaluated using R^2 and Adjusted R^2 metrics. Advanced techniques, including Ridge and Lasso regression were also explored to refine model coefficients and reduce overfitting, but were unsuccessful in improving the model. Finally, these combined approaches provided valuable insights into the environmental and spatial factors influencing baseflow dynamics.

Results:

The multiple linear regression model demonstrated strong predictive ability for river baseflow (distribution shown if Fig. 1), achieving a final R-squared value of 0.889 with 58 explanatory variables. In the initial model, incorporating segment IDs and monthly variables was essential, as omitting these resulted in a significant drop in model performance. A Box-Cox transformation of the dependent variable (Y) with λ = 0.25 helped address non-constant variance observed in the residuals, improving R-squared values considerably. The scatter plot of actual vs. predicted values (Fig. 2 & 3) shows the model's accuracy, particularly after transformation, as most points align closely along the diagonal, with notable deviations at higher values.

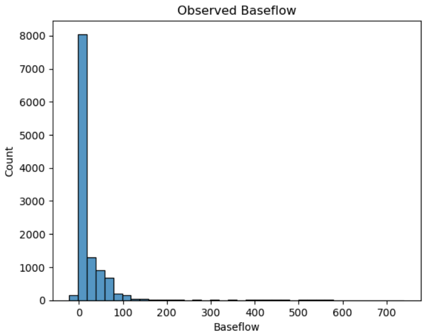


Fig. 1

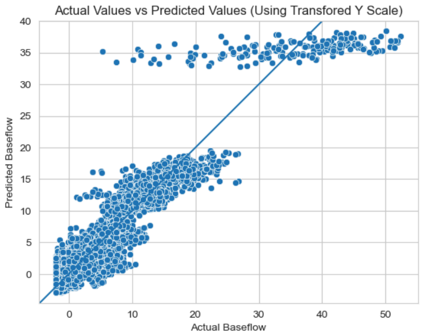


Fig. 2

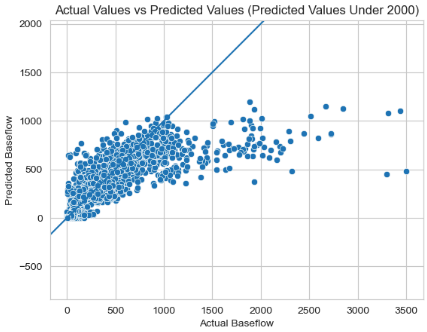


Fig. 3

The analysis revealed pronounced seasonal patterns, with elevated baseflows observed in spring, likely due to snowmelt contributions, and reduced flows during summer months, correlating with increased evapotranspiration demands (Fig. 5). This seasonal variation highlights the influence of climatic factors on river baseflow. Further breakdown by individual river segments underscores these patterns, as segments #256 and #239 consistently exhibit higher baseflows throughout the year, possibly due to their unique hydrological or geographical characteristics. Additionally, the relationship between mean evapotranspiration and mean baseflow across different seasons (Fig.4) provides further insight into temporal variability, confirming that evapotranspiration is a key factor driving baseflow reductions in warmer months. These findings emphasize the interconnected effects of seasonal shifts and specific river segments on baseflow dynamics.

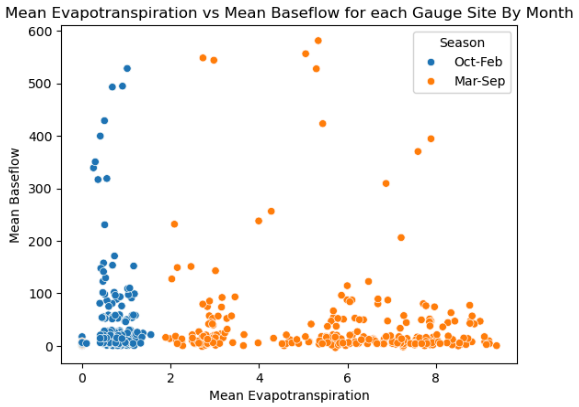


Fig. 4

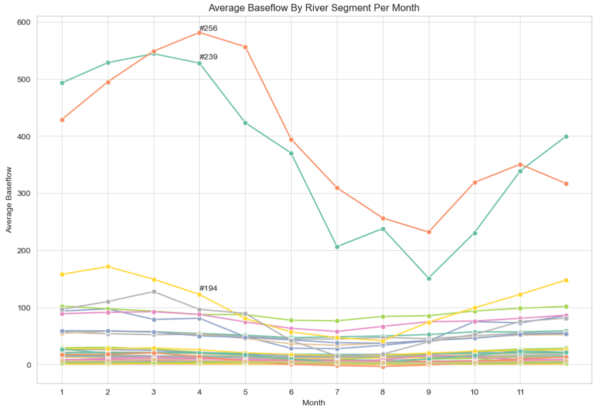


Fig. 5

Spatially, certain river segments, such as #256 and #239, exhibited significantly higher average baseflows. This could be attributed to favorable conditions for groundwater recharge or reduced water extraction pressures in these areas. The spatial variability in baseflow values is visually represented in a geographic distribution plot (Fig. 6), which highlights regions where baseflows are consistently and notably higher across the dataset.

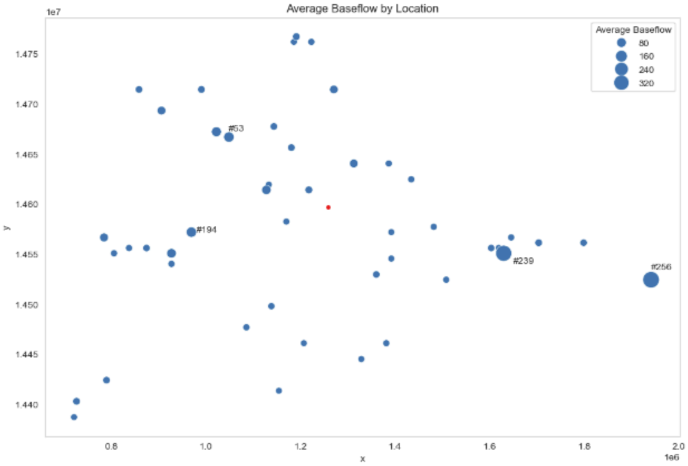


Fig. 6

Analysis of model coefficients underscored the importance of seasonal and spatial factors in predicting baseflow. Spring months and high-flow segments had the highest positive contributions, as displayed in the coefficient plot by month (Fig. 7). The segment-specific coefficient analysis (Fig. 8) identifies segments #147 and #157 as having substantial influences, reinforcing the geographic impact on model predictions.

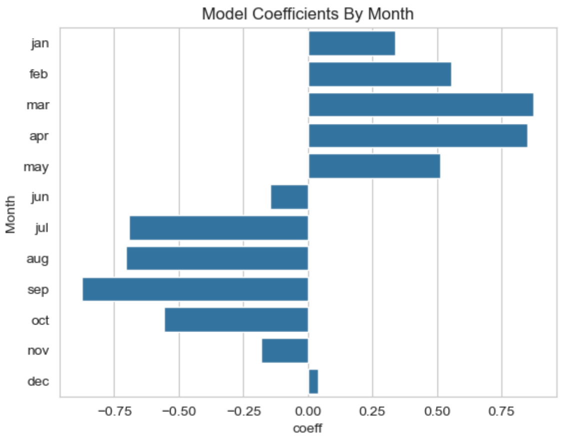


Fig. 7

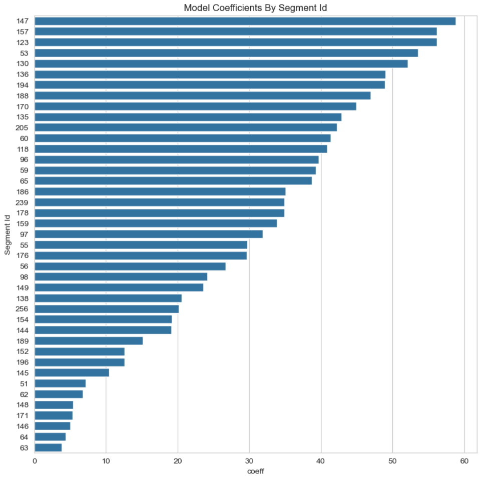


Fig. 8

The residuals were examined against key variables such as precipitation and evapotranspiration to assess potential patterns in prediction errors. The residuals show a generally random distribution around zero, which suggests that the model effectively captures the main trends of the data without significant systematic bias. However, there is slight clustering in some ranges of residuals, which indicates areas where the model's accuracy could benefit from further refinement. This clustering may hint at relationships or interactions that are not fully captured by the current model structure. Addressing these through additional transformations or by integrating interaction terms between predictors could potentially enhance model performance, especially for extreme or anomalous values within the data. This analysis highlights opportunities for further optimization in capturing nuanced relationships in baseflow predictions.

Removing non-significant variables, such as Irrigation_pumping and Segment_id_40, helped streamline the model without compromising accuracy. Though removing spatial coordinates (x & y) impaired the model, season and location variables had the most considerable impact on baseflow predictions, suggesting future efforts could focus on refining these variables. A recommendation for further improvement is to cluster river segments by geographic proximity to reduce variable complexity.

Model	R²	Adjusted R²	Key Features and Modifications
Initial Model	0.816	0.815	Included all months (one-hot encoded), summer variable, all river segments, evapotranspiration, precipitation, x, y, year, irrigation pumping.
Final Model	0.889	0.889	Dropped non-significant variables ('Summer', 'Irrigation_pumping', 'Segment_id_40'), transformed Y with Box-Cox (λ = 0.25), retained essential variables.

Technical:

The dataset required some preparation to enable linear regression analysis. First, we separated the date field into day, month, and year columns and explored seasonal and geographic patterns, which eventually led us to One-Hot encode the months and river segments.

Linear regression was appropriate for analysis after we had transformed all the explanatory variables into numerical variables and we had a continuous response variable of baseflow. Inspecting the residuals and scatterplots of our explanatory variables with our response variable indicated that many variables did not have a linear relationship with our response variable. However, using a box cox transformation (with lambda = 0.25) on our response variable did significantly improve the distribution of residuals in the model with most explanatory variables. Various transformations were attempted on Evapotranspiration, Precipitation, and Irrigation Pumping to try and improve the linearity of the data, but none of these transformations improved the model or the residuals of the model in any meaningful way. Multicollinearity is also likely an issue among our predictors, and with more time care should be taken to address this.

We first split the data into a training and testing data set after performing some basic data cleaning. Our analysis began with a thorough exploration of variable distributions and their relationships with baseflow, including examining geographic and temporal patterns. After encoding categorical variables, we built an initial model using all features. By reviewing residuals, predicted values, and p-values, we identified insignificant predictors and subsequently dropped them. Transforming the Y variable improved model accuracy, and all p-values became significant upon re-running the model.

To further refine the model, we tested Lasso and Ridge regression for dimensionality reduction; however, their R^2 values were considerably lower than the OLS model, so we chose not to pursue refining these models. Finally, we evaluated the R^2 value on the hold-out test set to check for potential overfitting. With more time, we would address the multicollinearity problems, focus on more dimensionality reduction, and explore more possible including of higher order terms in the model.