Introduction

Water is the most important thing for the existence of life. To keep this vital resource available, learning how to plan ahead and predict the baseflow from previous years is key. Baseflow is the portion of a river's discharge sustained by groundwater. Understanding baseflow dynamics is crucial for maintaining stream flow during dry periods and informing water management. The dataset contains river segment observations with spatial and temporal attributes. The variables include evapotranspiration, precipitation, irrigation pumping, and observed baseflow.

By fine-tuning the data, our analysis offers hydrologists and water conservationists crucial information, empowering them to plan for times of drought and secure this crucial resource. With this enhanced ability to predict baseflow, we plan to help maintain ecological balance and support research in other areas that depend on accurate baseflow predictions, making our findings highly valuable.

Github: https://github.com/ablibranix/project6
Presentation: Presentation

Dataset

This dataset contains information on baseflow for various river segments over the course of a single month. The data includes details on the date, segment id, location of the gaging station, evapotranspiration, precipitation, irrigation pumping, and observed baseflow. Our objective is to employ linear regression to predict baseflow using the other variables provided. After analyzing the data, we found that no attribute was significantly more crucial than another. However, we found the 'Segment_id' column to be more vital than the 'x' and 'y' columns and used it as the primary reference for our analysis.

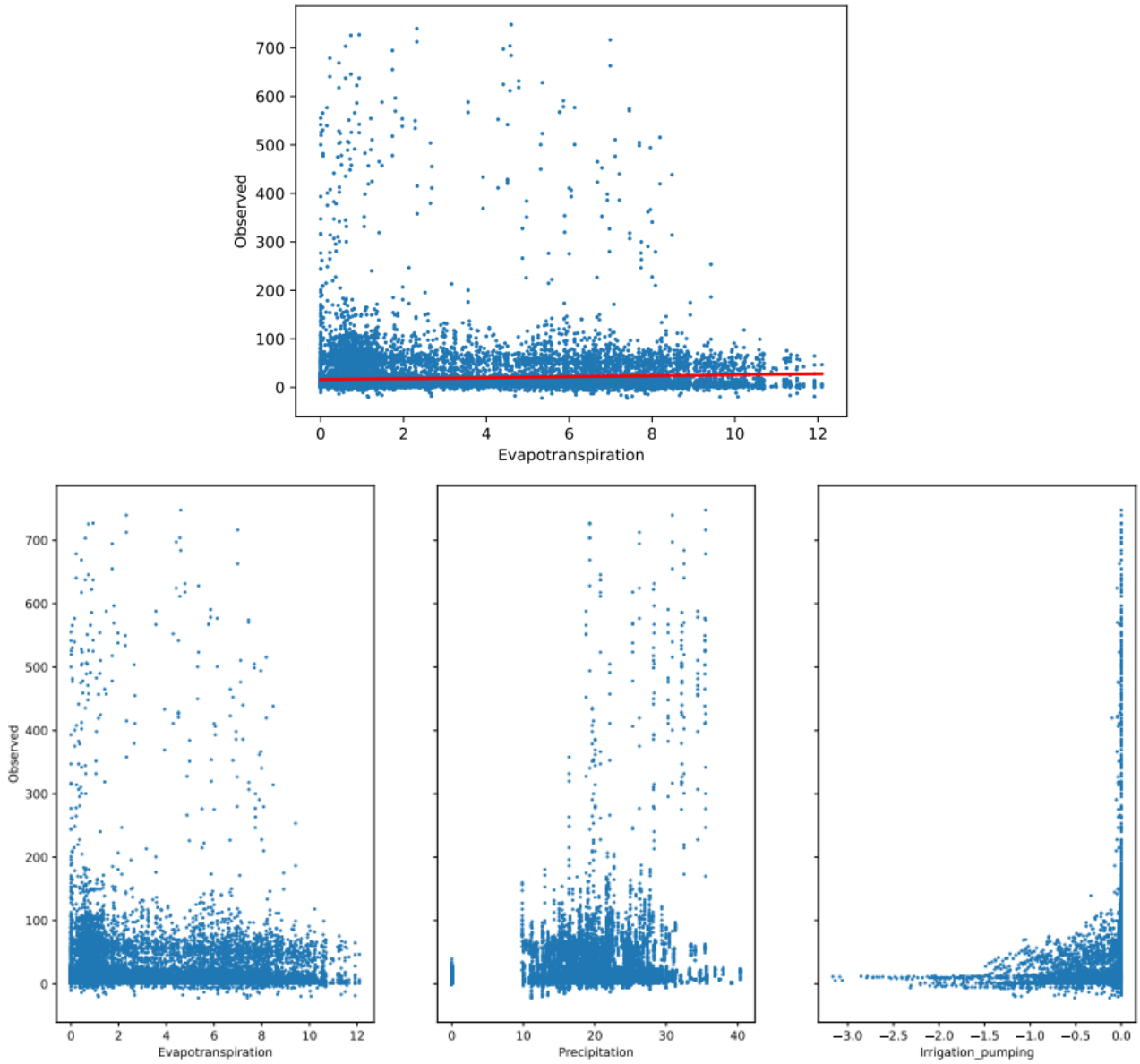
Analysis technique

Linear regression was used to mode the relationship between evapotranspiration, precipitation, irrigation pumping, and observed baseflow to predict baseflow. The technique quantified the relationship between these factors in the dataset, making it suitable for this analysis. Linear regression is a widely accepted tool for analyzing the relationship between variables and predicting outcomes from observed data.

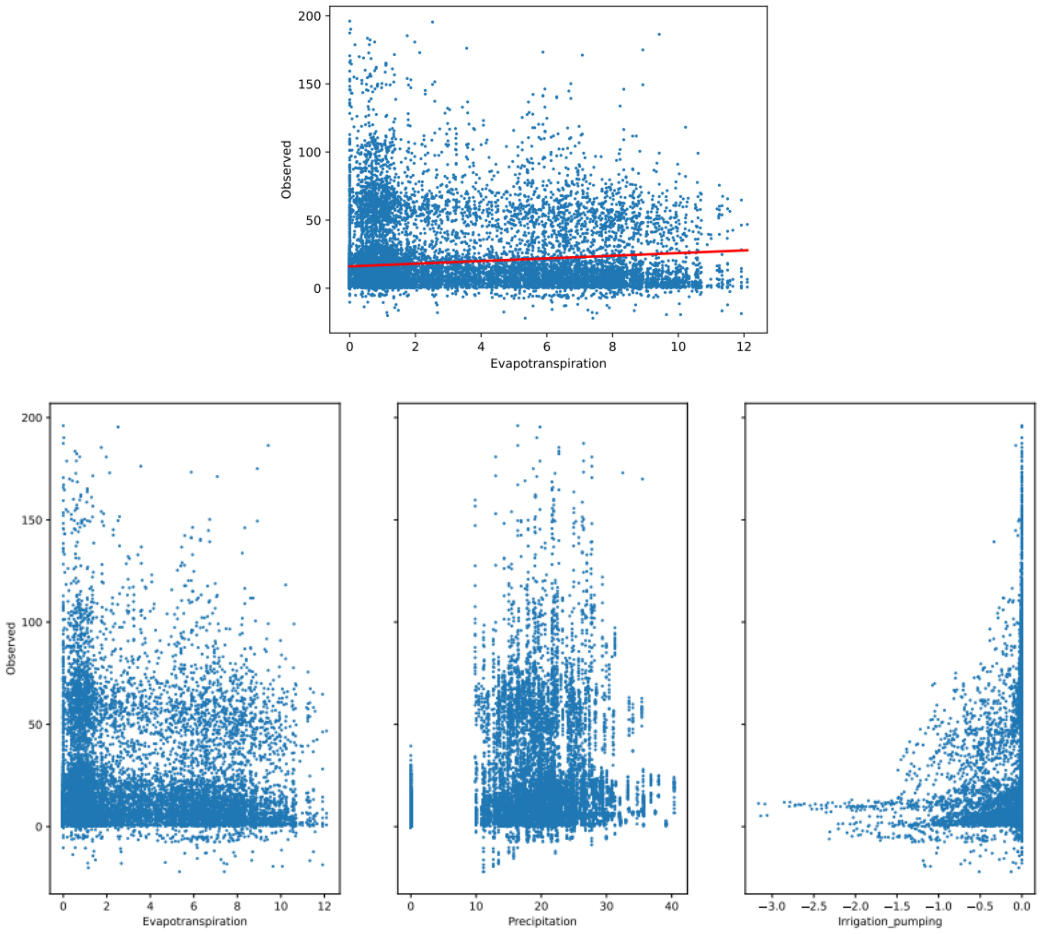
Results

To better understand our data, we initially started with plotting various scatter plots to see how the data spread across the various values. We initially plotted all the points, without removing any outliers. When performing linear regression, we found that by dropping "Observed",

"Evapotranspiration", and "Irrigation_pumping” to create our X value against “Observed” as our Y, we gained an R-squared value of 0.19656731732216504 and a root mean squared error of 47.764728250365025.



Since our R-squared value was still low, we decided to remove any observed value above 200 based off of the initial graph of the data and then re-ran the linear regression. When we did this, we received a R-squared value of 0.35377999139580485 with a root mean squared error of 21.108065068196808. This change doubled our result to give us a more accurate model to test to gain more accurate predicted values.



With more fine tuning of the data, we believe we can help hydrologists and water conservationists better predict baseflow for specific areas to help ecosystems maintain a healthy balance and also be used in other areas of research which depend on an accurate prediction of baseflow.

Technical

The dataset presented challenges during the analysis in the assignment. One particular challenge was the date format, which was difficult to decipher. However, after carefully examining and utilizing the information provided, we were able to convert it to a YYYY-MM-DD format. This adjustment facilitated the data analysis and visualization. Since the dataset was complex and we had limited knowledge about it, we utilized linear regression and experimented with different attributes to generate meaningful insights. However, in retrospect, we recognize that spending more time understanding the nuances of the dataset and exploring alternative techniques could have resulted in more robust results. We also regret not spending more time identifying and removing outliers that could skew the results. Additionally, we realized too late that standardizing the data was necessary to obtain better results, but we didn't have enough time to do so.