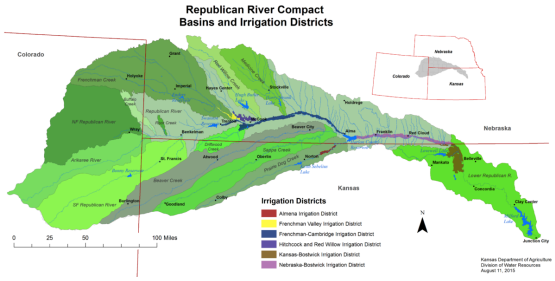
Project 6 Report:
Republican Valley River Base Flow Prediction

CS5830
Karl Poulson and Blake Barber
Slides: Google Slides
Github: Repository

Introduction

The Republican River Compact regulates how water is allocated among Nebraska, Kansas, and Colorado. The Republican River region plays an important role in supplying water to farmers across each of these three states. Disputes and tensions over water rights has been a common theme amongst farmers in this area. This project aims to forecast the baseflow, or the flow between precipitation events, in the basin, which will provide essential data for policymakers to be informed and allow for fair and effective irrigation regulations.



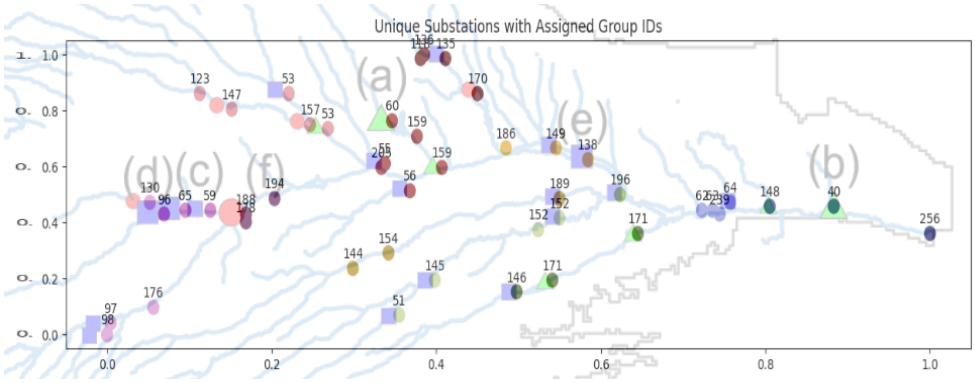
Dataset

The hydrological dataset included 3 prominent features we found compelling for analysis: precipitation, irrigation pumping, and evapotranspiration (regular evaporation and transpiration from plants). The actual baseflow is in the column called 'Observed Flow,' and represents what the actual baseflow was at the time of the recording. Other features include the recording substation id, and x and y coordinates of where the substation was located, which is particularly useful for identifying where the majority of the precipitation occurs. Our goal was to use the three prominent features to predict the baseflow. The dataset is suitable for this project as it contains over 15,000 data points from across the region. It includes data points from as far back as 1940 through the 2000's. To prepare the

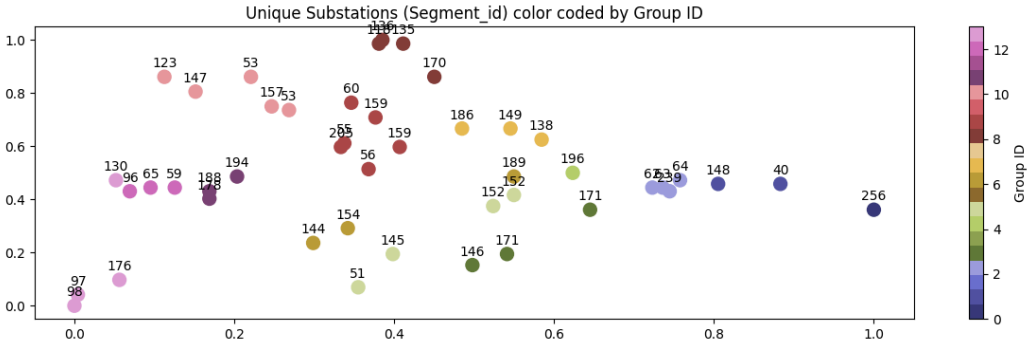
data, we dropped duplicate entries and standardized the data to ensure the predictions had the highest likelihood of success. We also chose to remove outliers beyond 3 standard deviations to ensure they did not mess with our algorithm.

```
Date,Segment_id,x,y,Evapotranspiration,Precipitation,Irrigation_pumping,Observed
717945,154,1138990,14497920,2.8,34.66,-0.011344,14.31557377
717976,154,1138990,14497920,1.89,34.66,-0.039628,16.0803279
718006,154,1138990,14497920,1.38,34.66,-0.011344,18.4803279
```

Analysis Techniques

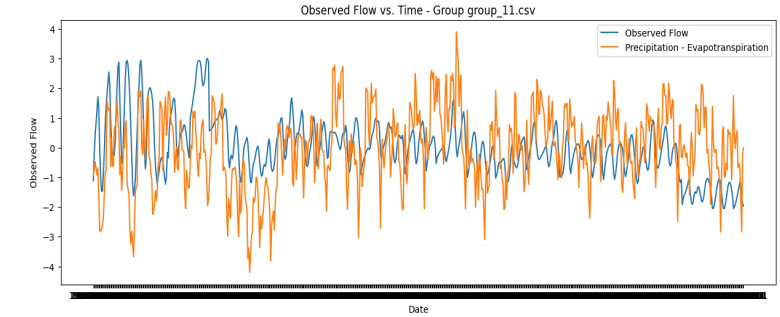


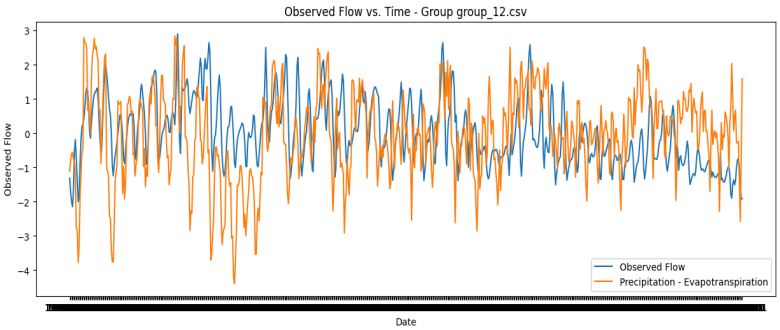
The observation stations are not uniformly placed along the river, and therefore the geographical placement between each other along the river path was an important metric to understand. We manually selected substations that were grouped near each other and located on the same tributaries and waterways. We then aggregated each of the selected features by the group id and took the average. This helped to prevent duplicate data points interfering with the linear regression algorithm, as stations near each other may provide too similar data.

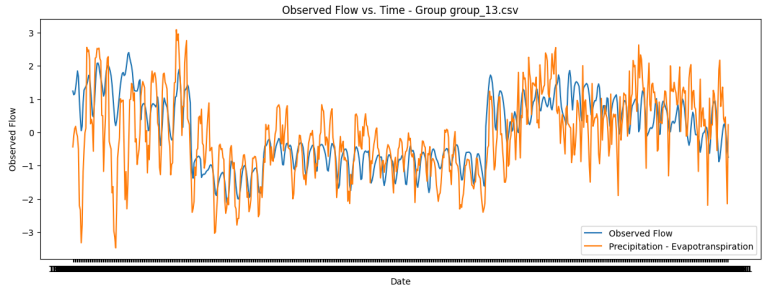


Results

Initial data exploration proved to be unsuccessful until we tested including the date column as a feature. For each group, we plotted the observed value vs the date. We found that some groups exhibit interesting cyclic patterns in river flow.



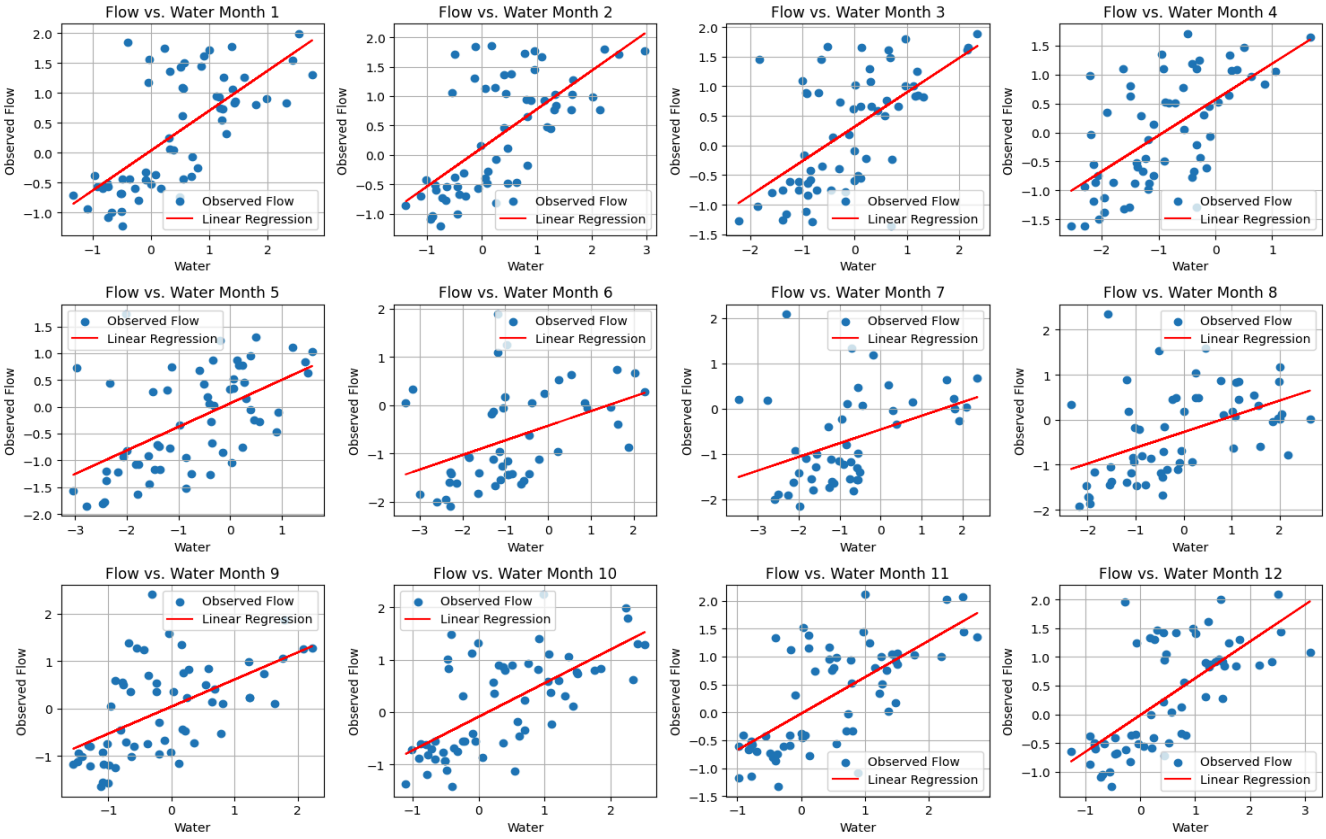




Groups 11, 12, and 13 are located at the upper portion of the flow along similar tributaries. These groups exhibited matching patterns between the cyclic nature of precipitation, evaporation and the observed river flow.

Which led us to our next question, how do we measure the linear correlation between two cyclic graphs? We measured the average water, which included both precipitation and evapotranspiration, as well as the observed river flow for each month across several years. We then plotted these values.

The graphs below depict the relationship between total water and the observed flow for each month for group 13. Each subplot represents a specific month, such as January, February, March, April, etc., with each point showing the corresponding total water and observed flow values. The analysis revealed a moderate level of correlation between the features, with an average linear correlation coefficient (r-squared) value of 0.353. This indicates a noticeable but not strong linear relationship between average water and observed river flow.



Technical

Data Preparation:

The code starts by reading in the CSV file containing the river basin data, merges it with geographic group IDs, and standardizes the coordinates to enable the creation of a scatter plot to help visualize the unique substations. The substations are then color-coded by group ID for additional visual clarity. Missing dates were filled in by forward filling the previous date values. We considered outliers to be values which were beyond the range of 3 standard deviations. The outliers were then removed to aid in the task of linear regression.

Analysis:

After conducting the initial and subsequent data explorations, we encountered challenges in identifying any significant correlation between our initial chosen features and the observed flow. It was only after we considered time factors and grouping the data by months that we began to notice some potential correlation between the features and the observed flow value. While linear regression may not have been the optimal solution for this dataset, it motivated us to thoroughly analyze the data and understand the underlying patterns more deeply.

Analysis Process:

During our analysis process, our primary goal was to understand the impact of irrigation on observed flow, but unfortunately, we were not successful in achieving this goal. However, we did discover a meaningful finding: there exists a linear relationship between total precipitation and evaporation within specific subgroups of our data. In conclusion, given the chance to do this project over again, we would look for alternative machine learning methods for predicting the baseflow of the river.