Project 6
Carson Stoker
Seth Beckett

Github: https://github.com/sethbeckett/cs6830_project6
Slides: https://docs.google.com/presentation/d/1t89DZOXjXb-TUyC4NCXj6sRW8sWse1L_RkEFZh_LeQQ/edit?usp=sharing

# Groundwater Model in the Republican River Basin

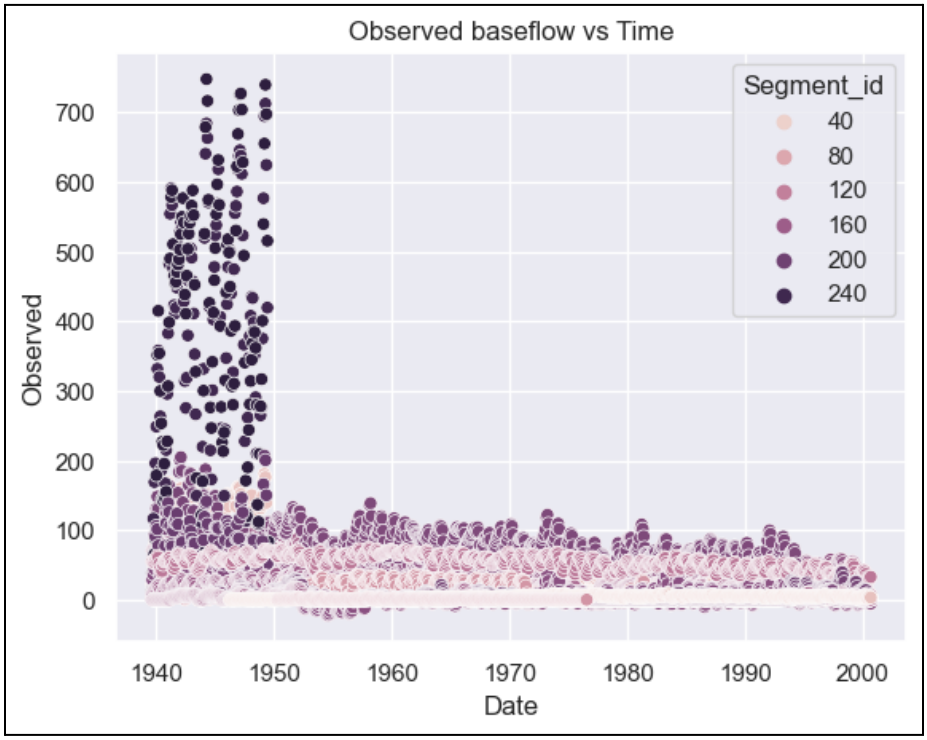
## Introduction

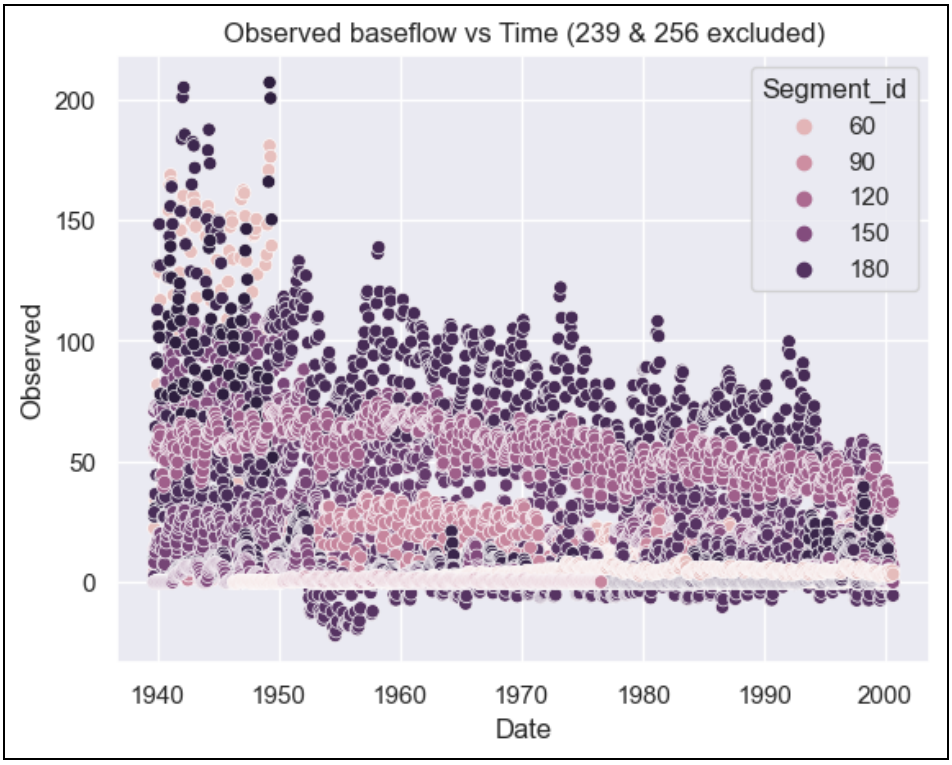
In this project, we perform analysis on groundwater baseflow data from the Republic River Basin in Kansas, Nebraska, and Colorado. The data comes from the Republican River Compact Administration. The purpose of this analysis is to predict the baseflow value of a river or stream based on other metrics such as precipitation and evapotranspiration. This will be done using linear regression models. This is a critical topic due to the importance of water supply in the Plains states, where the primary industry is agriculture. Increased pumping for irrigation from the groundwater has led to a decrease in the stream levels in the river basin. Understanding groundwater models and being able to predict how water is moving between the groundwater and the streams is an important step in managing water supply and resolving conflicts over water rights. This is especially important in drought years, when the input to a stream from runoff is low. Thus, this project will inform the government, water management agencies, and farmers on baseflow levels, assisting them in managing water in the Great Plains.

## Dataset

The dataset used in this project comes from the Republican River Compact Administration, an organization dedicated to overseeing the Republican River Compact, a congressional law passed in the 1940's to manage water in the region. The Republican River has its headwaters in eastern Colorado and flows through Nebraska and Kansas until it joins the Smoky Hill River to form the Kansas River. The dataset contains baseflow (net flow from groundwater to stream) measurements at 46 different gaging stations across the river basin between the years 1939 and 2000. Some of the attributes the dataset contains include the date (number of days since Jan 1, 1900), the segment ID of the river segment and the x and y coordinates corresponding to the location of the gaging station. It also contains measurements of precipitation, evapotranspiration, and irrigation pumping in the area adjacent to the river segment, and of course the observed baseflow at the station. The measurements were taken once a month; however it should be noted that not every gaging station was operational for the entire duration of the dataset. This can lead to perceived relationships that don't actually exist. For example, before the year 1950 there were many baseflow observations above about 150. After about 1950, these measurements cease. It might seem that there was a decrease in water levels around this time. However, looking more closely at the data reveals that some river segments (likely downstream ones where the river is larger and there is greater baseflow) stopped being measured around this time, leading to a perceived drop in baseflow.

Little data cleaning was needed for this project. Mainly, the original Date attribute was said to be the number of days since Jan 1, 0000. We subtracted 693963 from this value to obtain the number of days since Jan 1, 1900 and added that as a new column. We then changed the date to be the actual date of the measurement in year-month-day format. Additionally, we created a new column titled x,y that combined the x and y coordinates of the gaging station in order to identify the number of unique measurement sites. We also decided to remove observations for the river segments 239 and 256, since they would skew the model, and were only measured between 1939-1949. Below we can see the comparison of baseflow over time when segments 239 and 256 were included versus when they were excluded.






**Analysis Techniques**

We employ multiple linear regression analysis in this project. Linear regression is a technique that uses features or predictor values to predict a quantitative response. Multiple linear regression uses multiple features to predict the value. It works by finding the line which minimizes the sum of squared errors (or the distance between observed and predicted points). This is a suitable technique as it allows us to create a line of best fit and predict the quantitative baseflow values based on other measured metrics at the site at the time. Doing so enables us to fulfill the purpose of the project and inform interested parties about the predicted state of groundwater and stream levels in the Republican River Basin. We also employ hypothesis testing, confidence intervals, and R-squared values to assess how well our model fits the data and how confident we are in the results. These are suitable techniques as they allow us to detail to the interested parties how relevant, significant, and reliable our results are.

**Results**

After comparison of multiple models created using multiple linear regression, we found the model which best explains the data and can hopefully be used to predict future baseflow best. As seen in the table below, our final model's predicting variables include evapotranspiration, precipitation, and irrigation pumping. In addition to these quantitative variables, the station (unique combination of x y coordinates) was used as a categorical variable. We found that all of these variables were statistically significant, and their 95% confidence intervals did not include 0. This model had an adjusted R^2 value of 0.853. This means that the model explains the data well.

| Feature | Coefficient |
|---|---|
| Constant (y-int) | 88.26 |
| Evapotranspiration | -0.71 |
| Precipitation | 0.24 |
| Irrigation_pumping | 5.90 |

To interpret these coefficients, we would say that a unit increase in evapotranspiration leads to a unit increase in baseflow. This makes sense since water that is evaporating and transpiring enters the atmosphere and cannot enter the stream. So more evapotranspiration, would mean a lower baseflow. For precipitation, a unit increase in precipitation leads to a 0.24 unit increase in baseflow. This makes sense since more rain means more water available in the system to enter the stream from the groundwater. For Irrigation pumping, a 5.90 coefficient means that for every unit increase in water pumped for irrigation, there is a 5.9 unit increase in baseflow. A big thing to be aware of is that irrigation pumping is described as a negative value. It ranged from about -3 to 0. Thus, -3 means more pumping and 0 would mean no pumping is going on. This is why the coefficient is positive. As irrigation goes up (becomes more positive) the magnitude actually goes down meaning less pumping is going on, which is why it is related to an increase in baseflow. This is opposite of evapotranspiration, which is measured as a positive value and thus had a negative coefficient even though irrigation pumping and evapotranspiration have the same relationship with baseflow.

Below is a table showing the P-values and confidence intervals for the quantitative variables. We can see that all of the p-values are significant and that none of the confidence intervals contain 0. This means we are quite confident that our results are not due to chance and that the true coefficient for each feature lies somewhere within its confidence interval. We will not list the coefficients for the dummied-out x,y variables since there are 40+ of them. However, they all had significant p-values. Most of them had a negative coefficient, but the coefficients ranged from -89 to 16.

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 88.2571 | 0.946 | 93.269 | 0.000 | 86.402 | 90.112 |
| Evapotranspiration | -0.7141 | 0.034 | -20.706 | 0.000 | -0.782 | -0.646 |
| Precipitation | 0.2350 | 0.018 | 12.830 | 0.000 | 0.199 | 0.271 |
| Irrigation_pumping | 5.8987 | 0.358 | 16.473 | 0.000 | 5.197 | 6.601 |

Baseflow is the net groundwater discharge in streams, and is an especially key component of water supply during droughts. By using our model, water management of the Republican River Basin will be able to effectively predict how much baseflow will be available for use at different times and in different places. This will allow for an equitable splitting of water rights and a responsible management of the streamflow.