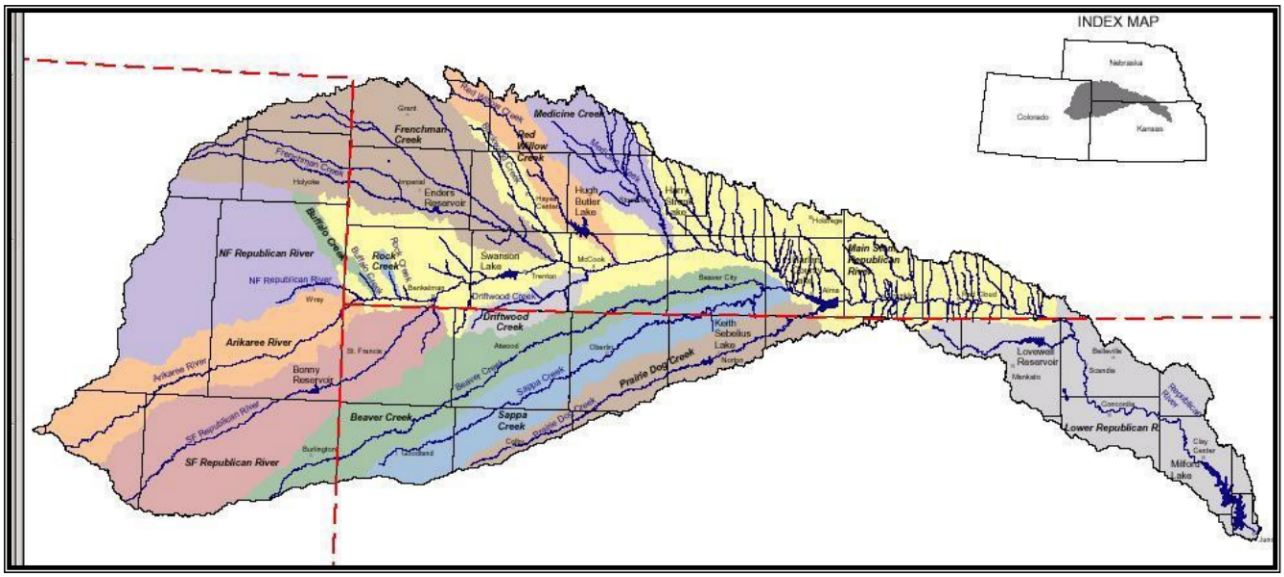
Introduction

Long gone the days where man brutally tames nature must be; using these data can ensure a long and prosperous future for great plains generations to come. In our analysis, we show that to some extent, base flow for each section of river can be predicted by other variables. This will let local and state leaders both be informed on the matter, as well as be given the knowledge on how to mitigate this crisis. The baseflow of a river is how much water is consistently available in between precipitation events.

Presentation and GitHub



Dataset

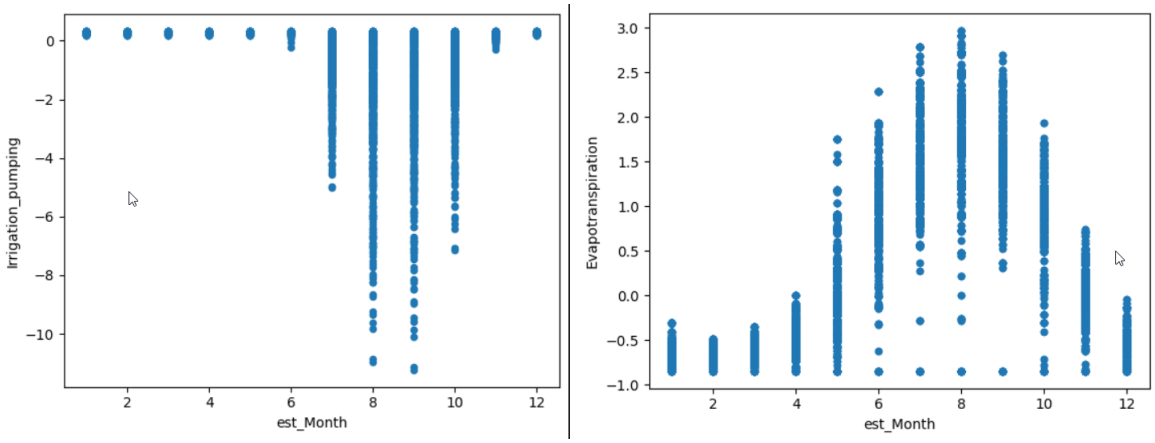
In our dataset, there are almost 16,000 samples, taken monthly from dozens of sites across the Republic River Basin. In each sample, monthly perception was measured, as well as local irrigation usage from that section. Furthermore, in the data was included an estimation for the amount of water which was lost from base flow due to evaporation or transpiration on the surrounding shores. Most importantly, baseflow was measured for each sample. The data are provided by a tri state coalition of Colorado, Nebraska and Kansas governments and is reputable.

Analysis Technique

Linear regression is an efficient manner to relate variables that are correlated one with another. There is some nuisance in ensuring that correlation doesn't get carried away into causation, and that the data are explained well by linear functions. This would entail that our explanatory variables don't cause a change in themselves. As an example from our data, more irrigation doesn't positively feedback into itself. Linear regression is simple and can be graphically presented simply enough for any layperson (including government officials) to understand.

Results

Our initial analysis discovered that, when grouping by month, irrigation pumping and evapotranspiration increased in scale dramatically in the summer months. This helped us to know that we needed to plot the data grouped by month when searching for a strong linear regression line.



We found that when we fit the data based on the month we estimated the current row to be in, that we tended to get better R^2 values. Before we factored in the months, we averaged around .2 for our R^2. After we fit the data according to month, our best R^2 ended up being around .34. The features that we used were evapotranspiration, precipitation, and irrigation pumping, and a number we calculated for distance away from the source. We standardized each of them and found that the results we got improved by a margin of .03 for R^2.

These are the ideal factors that we found for predicting the baseflow of a particular section of the river. 0.34 isn't a fantastic R^2 value for showing how good the data is spread against our predictors, but it is an improvement upon the naive guessing that was done initially when we simply plotted all the data with no regard to segment or month.

Technical

"A fundamental assumption of least squares regression is that model residuals can be described by a noise term corresponding to measurement error and that the noise term is uncorrelated and Gaussian distributed. This assumption is often violated when the groundwater model has significant input and structural errors. As a result, simulations made with the calibrated model could be biased and the resulting predictive uncertainty intervals may be unreliable" (Honti et al., 2013).

This is one of the first remarks made by the researchers in the included study. Such a preface is needed for this technical section as well.

We prepared the data by removing the sections of the river which had no useful data. This entailed those sections for which evapotranspiration, precipitation and irrigation were all zero for the entire course of the study, which eliminated 7 sections of the river from our study. Furthermore, we had to first justify, then implement the modification from the date of the sample

to the calendar month in which the sample was taken. Additionally, since it wasn't included, a distance to the head of the Republic River was calculated and used in the analysis.

Linear regression is a model which predicts a quantitative outcome based on other variables, with the assumption (or rather hope) that there is some cause linking them. A linear line is fit to the explanatory variables, minimizing the variance in the distance from the line. The standard euclidean norm was used in our project. It is equivalent to projecting the explanatory variables onto the first principle component of the data.

```
All together Mean R^2 scores: 0.341453740485771 Variance of R^2 scores: 0.09000837511757476
```