CS 5830-Project 6

Estimating Baseflow

Megh KC | Supriyo Sadhya

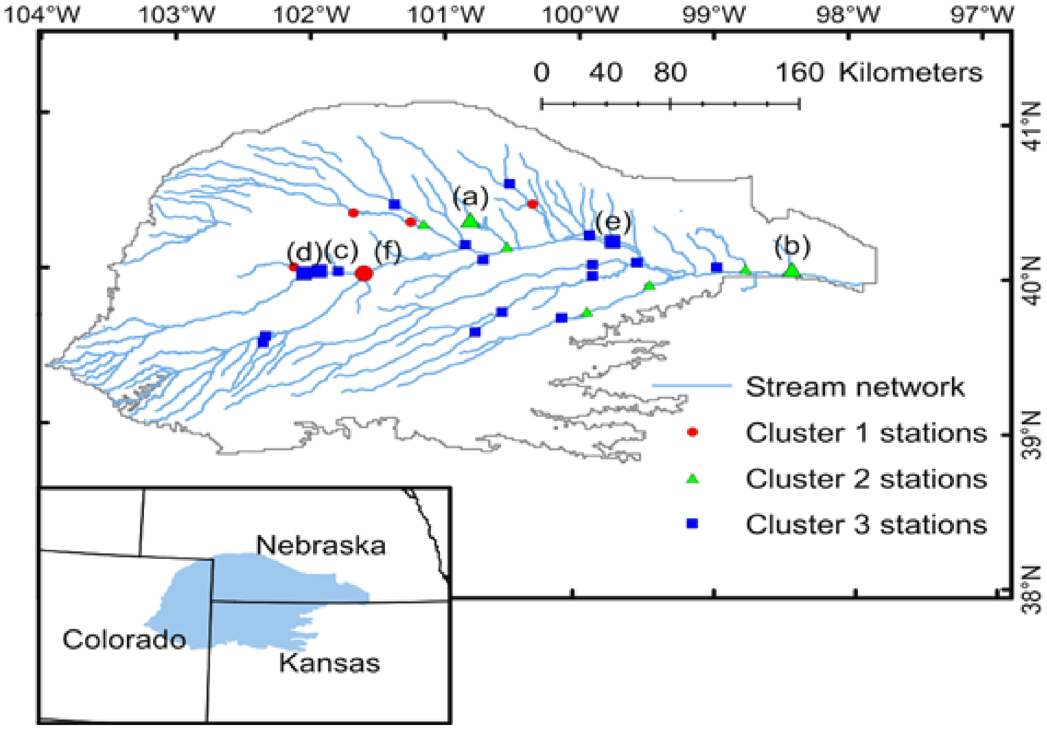
Introduction

Baseflow prediction for a given stream network dataset by using a linear regression model is the purpose of this project. The dataset contains information from over 60 years (1939-2000) in 3 states of the USA. The dependent variable is observed baseflow where independent variables are evapotranspiration, precipitation, and irrigation_pumping. We used a multiple linear regression model and predicted the base flow. The adjusted r-square value showed the significance of using multiple attributes.
Our analysis could be used by the hydrologic or water resource planning organizations to understand how the hydrologic, climatic & groundwater discharge affect the river flow. Also, planning officials, data analysts, and water engineers can be benefitted.

Presentation link: here
GitHub link: here

Dataset

The hydrologic dataset comprises 7 independent variables: evapotranspiration, precipitation, irrigation_pumping, (x,y) spatial location of the gaging station, date and segment_id over the 60-year period. The study area was 3 states of the USA namely Colorado, Nebraska & Kansas. To better understand the study area, an image was extracted from the referenced paper. We predicted the base flow on the river/stream segment using a linear regression model. The date field was corrected using the datetime function in python. The average number of days & months in a year is used to convert the given date field into respective years and months.

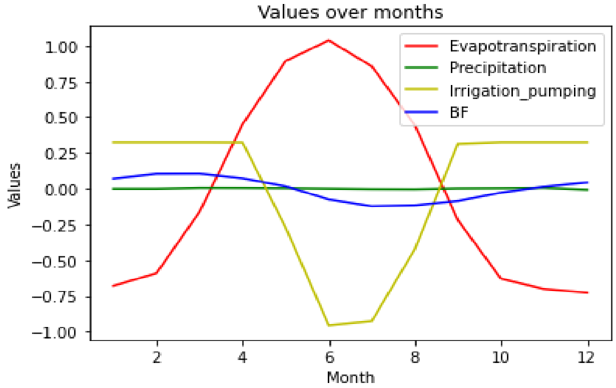


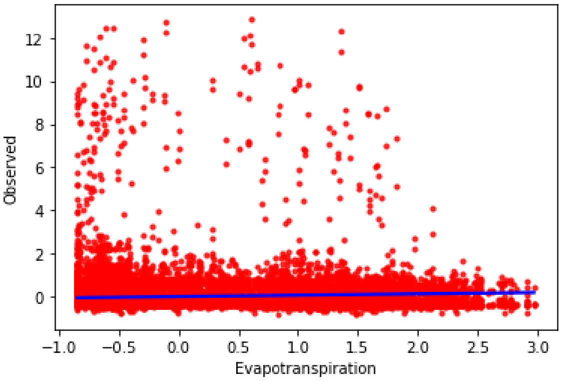
Analysis Techniques

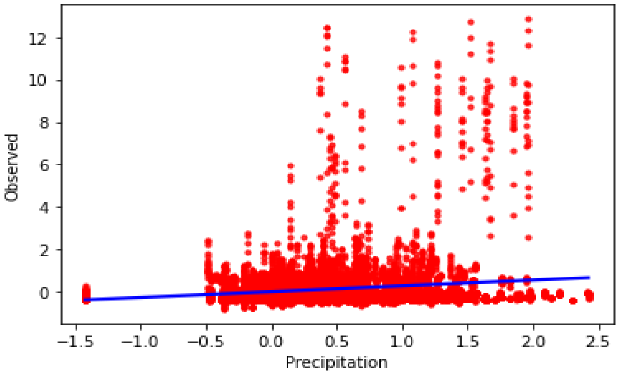
At first, we processed the date column to get the month in which the reading was taken. For that, we subtracted 693963 from the given value to get the number of days passed from January 1, 1990. We then used the datetime package in python to get the current date and from that the current month. We found that the baseflow value varies with the month. It follows something like a sine wave. We also standardized the evapotranspiration, precipitation, irrigation_pumping columns and observed columns. One hot encoding was applied to categorical variables Date and Segment_id. We plotted a series of scatter plots for each of the independent attributes for each unique stream segment to see how the data are associated. We decided to take the stream section which may have a significant association with these attributes. The line plot for baseflow using the overall dataset was plotted. Using the statsmodels.formula.api we tested out multiple linear regression models and a summary of the results was created. We chose the model with the highest R^2 value and p-value less than 0.05 . Results are discussed in the Results section below.

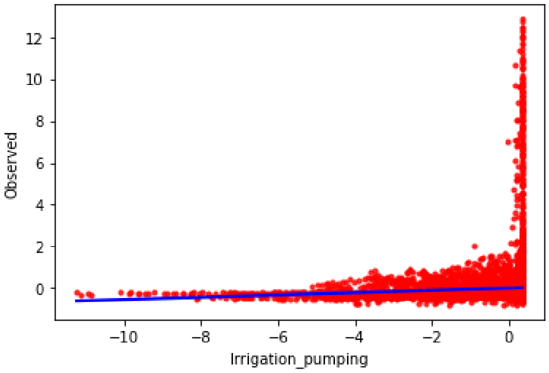
Results

We group the data by months and take the average of evapotranspiration, precipitation, irrigation_pumping and baseflow for the months. The plot is given below. We can see the variation of the values of the parameters over the different months. It gives us an idea of seasonal variation of the parameters. Precipitation however remains almost constant which is a bit strange. The least square lines for evapotranspiration, precipitation, irrigation_pumping are plotted below for the entire data which shows us the overall trend of the parameters.

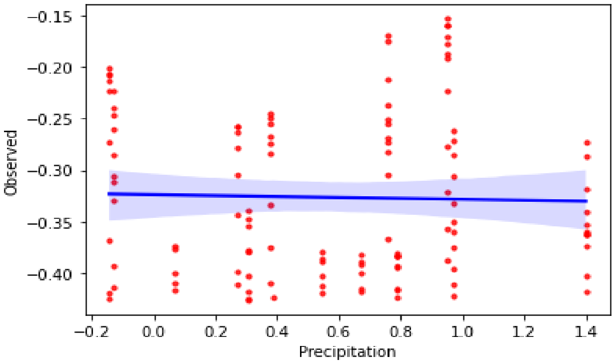


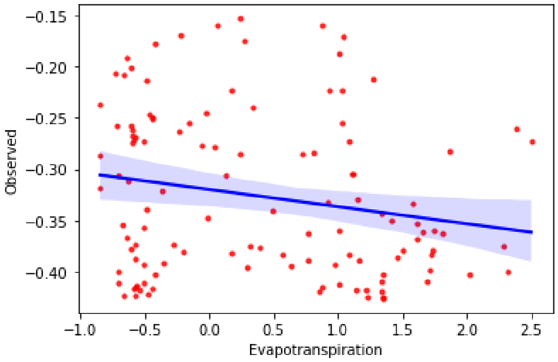


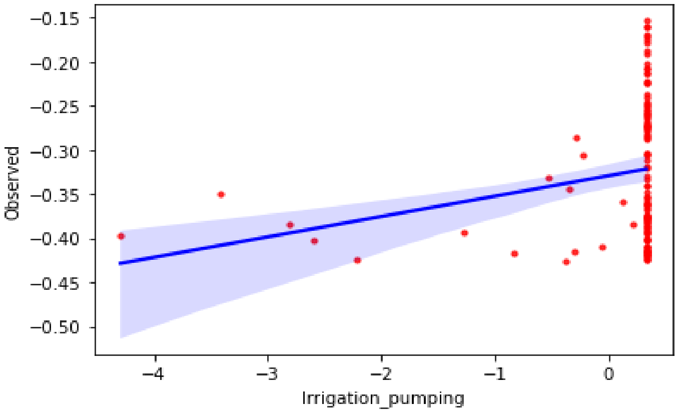




The least square lines for Segement_id 144 is given below.

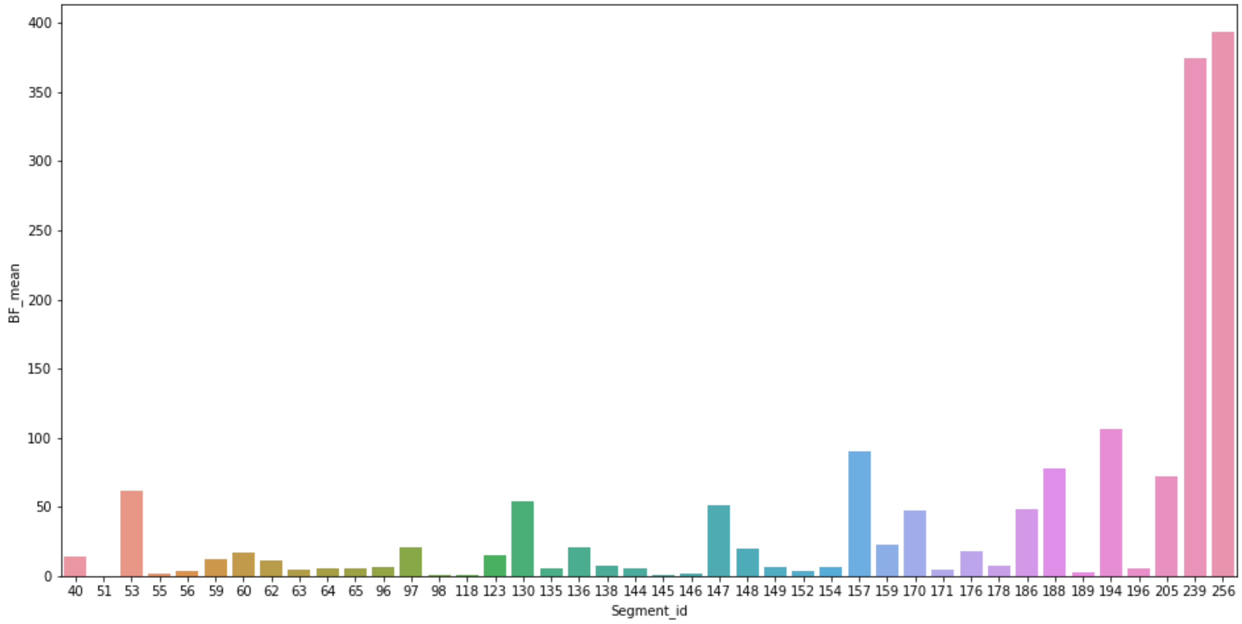






As we can see there is a visible trend in Evapotranspiration and Irrigation pumping data for segment 144. However, there might be a more visible or less prominent trend in other segments.

The figure below gives us an idea of the average baseflow across each segment. It gives us an idea that each segment has a different baseflow and so segment_id could be used as a useful predictor. The (x,y) spatial location of the gaging station can be ignored as they have a given value for a given segment. So, using the segment_id should be good enough.



Model no.	R^2 value	P value
Irrigation_pumping +	0.083	2.91e-291

Evapotranspiration + Precipitation		
Evapotranspiration + Precipitation	0.076	8.20e-270
Irrigation_pumping + Precipitation + Evapotranspiration + Date(one hot encoded)	0.086	1.34e-291
Observed ~ Irrigation_pumping + Precipitation + Evapotranspiration + Segment_id(one hot encoded)	0.718	0.00
Observed ~ Irrigation_pumping + Precipitation + Evapotranspiration + Segment_id(one hot encoded) + Date(one hot encoded)	0.722	0.00

We then use train test split to split our data into training and testing. Fitted the best model with the training data. Then we compute the R^2 value for the training set which comes out to be 0.72

References:

Xu, T., & Valocchi, A. J. (2015). Data-driven methods to improve baseflow prediction of a regional groundwater model. *Computers & Geosciences*, *85*, 124-136.