Project 6 Report

1. Introduction

The aim of this project is to create a multiple Linear Regression model to predict the baseflow based on temporal, categorical, and numerical features named: date, segment ID, irrigation pumping, precipitation, evapotranspiration, observed baseflow, x, and y. with 15591 records. The stakeholders of the analysis can be agriculture Farmers to benefit from predicting baseflow to optimize irrigation scheduling and water usage, ensuring sustainable agricultural practices, or water resource management authorities to predict baseflow to manage water resources effectively. We used 10-fold cross-validation and got an R-squared of 0.806 and a Mean Squared Error of 591.787 on average. Also, we did Ordinary Least Squares without cross-validation which we got an R-squared of 0.816. Furthermore, we looked at the average baseflow over seasons in which summer was the driest one. Also, it sounds like precipitation has a decreasing trend in the meantime irrigation pumping has increased until 1978, and segments 256 (year 1944) and 239 (the year 1949) had the strongest baseflow in the spring season. Also, we saw a negative correlation between baseflow and evapotranspiration and a positive correlation between Precipitation and baseflow, moreover, when irrigation pumping increases the baseflow decreases in most of the random segments. Github repository and presentation links.

2. Dataset

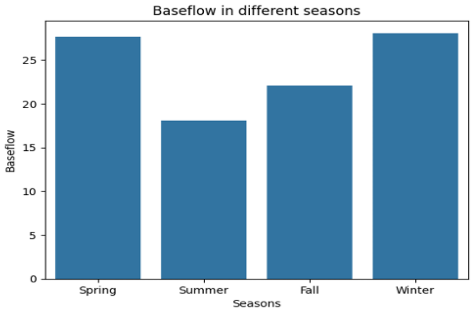
The dataset that is used in this project deals with baseflow. There are some meaningful temporal, numerical, and categorical features such as the date which is the number of days since 01-01-0000 that will be turned from 01-01-1900 then will be changed to year, month, day, and season, segment_id which will be treated as a categorical feature to identify segments of the river, x, and y as a geographical feature and shows the location of observation, evapotranspiration which shows the evapotranspiration amount of an area adjacent to the river segment in the given month, precipitation which shows the precipitation amount of an area adjacent to the river segment in the given month, irrigation pumping which shows the amount of groundwater was pumped out for irrigation, and observed baseflow which shows the base flow of the river. All of these features are important in creating our multiple linear regression model. We can create new columns based on date and use them in analysis like seeing baseflow over seasons, irrigation pumping over years, etc. By applying One-Hot encoding on segments_id and newly created column named season we prepare the dataset to be used in our model. Spatial location can give us useful information about observed baseflow. Also, we can use features like precipitation, evapotranspiration, and irrigation pumping with baseflow to see if there is any correlation between them. This dataset has 8 features and 15591 samples without missing values so none of the rows has been dropped. All the features look related to baseflow and model. The dataset is pretty clean so, we did not drop any samples nor normalize the dataset.

3. Analysis technique

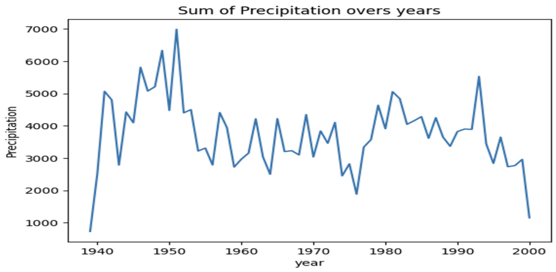
In this project, several techniques were employed to analyze the hydrologic dataset and predict baseflow. The main was linear regression, which was chosen for its simplicity in modeling linear relationships between the dependent variable (baseflow) and independent variables such as evapotranspiration, precipitation, etc. The exploratory data analysis, including scatter plots and regression lines, suggested that linear regression works very well in the data. In addition to linear regression, data preprocessing and feature engineering techniques such as converting date to year, month, day, and season (based on month and day columns), and One-Hot encoding on categorical variables (segment ID and season) were applied to prepare the dataset for analysis. Exploratory data analysis techniques, including data visualization (bar plots, line plots, scatter plots) and summary statistics, were used to gain insights into the relationships between variables to find trends and patterns. Model evaluation, including k-fold cross-validation, splitting the dataset into train and test, and metrics like mean squared error (MSE) and R-squared, were used to assess the performance of the linear regression model. Lastly, more advanced linear regression techniques named Ridge and Lasso regression, were explored to look at intercept and coefficients. All these techniques allowed for a comprehensive analysis and provided insights into the factors influencing baseflow For the linear regression models. We use all the features except year, month, day, and date. We use One-Hot encoded season in the models instead.

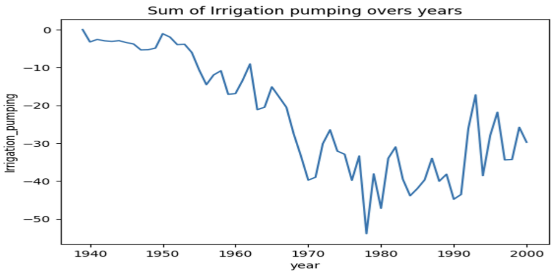
4. Results

Through our comprehensive analysis of the hydrologic dataset, we have gained valuable insights into the factors influencing baseflow in the studied river segments. Changes in base flow have direct implications for water resource management and planning, especially when it comes to agricultural irrigation, reservoir level control, and ecological conservation. As can be seen from the figure, the high base discharge in spring may be due to the replenishment of groundwater by snowmelt, while the decrease in summer may be related to the increase in evaporation rate and the decrease in rainfall. Such information is useful for hydraulic engineers who need to manage river levels and ensure adequate water supplies throughout the year.

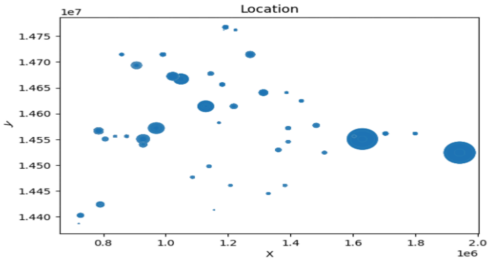


The following graph shows two-time series data: precipitation accumulated over years (left) and irrigation withdrawals accumulated over years (right). In the graph on the left, we see clearly the sum of precipitation over the years had a decreasing trend. The figure on the right shows a general increase in irrigation withdrawals from 1940. As we look, although the precipitation decreased over the years, the irrigation pumping increased from 1940 to 1978. From 1978 it looks like the amount of water going to be pumped was controlled and we do not see an increasing trend in pumping.

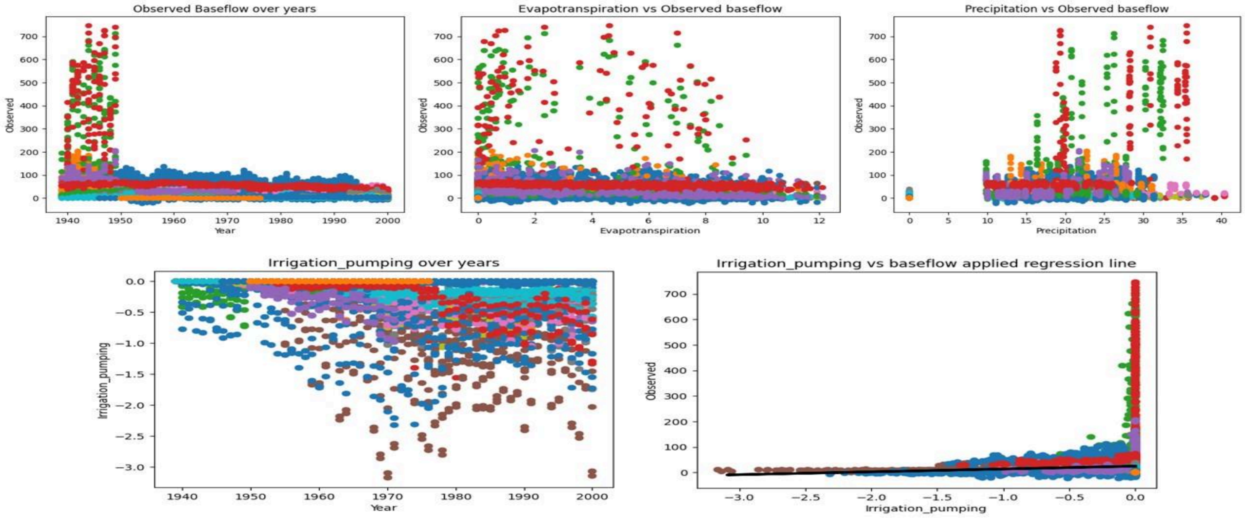




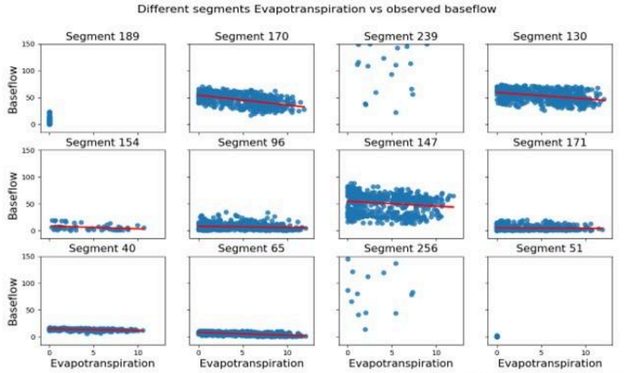
The following scatter plot represents a set of geographic location data, where the (X) and (Y) represent geographic coordinates, and the size of the points represents the size of the base flow observed at those locations. The top two biggest points were in the years 1944 and 1949. Both of them are seen in spring when the irrigation was 0. The spatial locations from the largest baseflow are: (x: 1941550, y: 14524320) and (x:1630030, y:14550720) with base flows: 747.80 and 712.55.

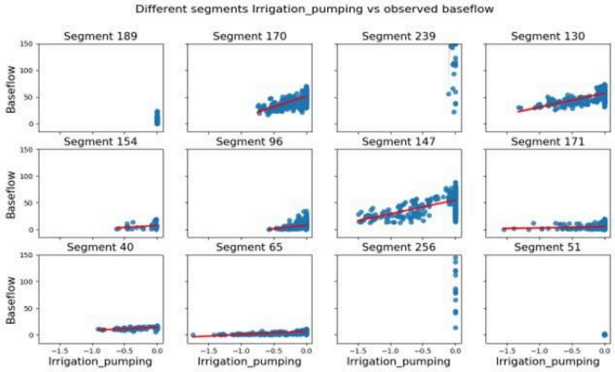


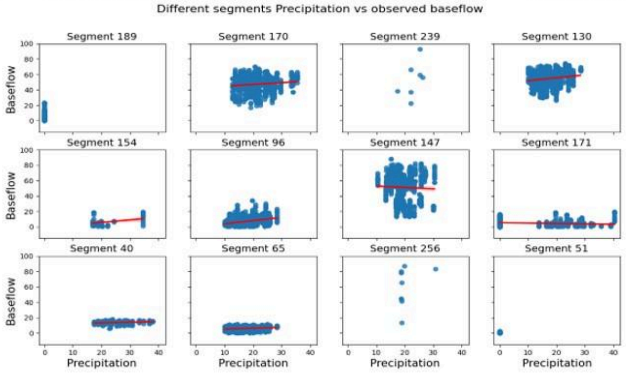
The following graphs show the scatter plots of baseflow versus year, evapotranspiration, precipitation, irrigation pumping, and irrigation pumping over the years of all segments. As we see the baseflow decreased over the years and irrigation pumping increased drastically from 1950 to 2000. Also, we created a linear regression model only on irrigation pumping vs baseflow and tested that, and plotted the regression line (black line on figure bottom right). We found 24.675 for intercept and coefficient of 11.176, and an R-squared of 0.002 which is not good at all. So, it means that we have to use other features as well for creating a more complex model.



Because the above figures do not show a clear correlation of baseflow versus evapotranspiration, precipitation, and irrigation pumping we randomly selected 12 different segments which are 189, 170, 239, 130, 154, 96, 147, 171, 40, 65, 256, 51 to analyze the correlation between the factors. As seen, in most of the segments there is a negative correlation between baseflow and evapotranspiration (left charts). Furthermore, in most of the segments by increasing irrigation pumping we can see that the baseflow decreased (right charts). In the meantime, we cannot clearly talk about precipitation because we saw a slightly negative correlation as well in segment 147 but generally, by looking at charts we can see in most of the segments precipitation had a positive correlation with baseflow. As the precipitation increased the baseflow increased as well (bottom charts). In general, we can say these three factors are important for the baseflow. Therefore, as we saw there are a lot of useful features like these three and other features like One-Hot encoded segments_id, season, x, and y which are important elements of the creation of the model. We will use all the features (not including date, year, month, and day) to predict accurately the baseflow to inform the stakeholders about controlling the water usage for farmers to irrigate or to manage water resources effectively so water resource management authorities can try to stop wasting water.







Based on the R-squared values from each fold shown in the following table, most are above 0.78, indicating that the model can fit the data relatively well. The average R-squared value of the model under 10-fold cross-validation is 0.806, reaching a high level. This suggests that the model can robustly capture the relationships within the data and possesses good generalization ability.

R-Squared in fold	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CV R-Squared	**0.821**	**0.809**	**0.838**	**0.632**	**0.824**	**0.780**	**0.864**	**0.785**	**0.863**	**0.839**

The average Mean Squared Error (MSE) is 591.787, reflecting the average deviation between the model's predicted values and the actual values. Additionally, the average intercept estimated on the 10 folds is 10626.618, representing the predicted value of the dependent variable when all independent variables are set to zero.

In our last analysis, we got 0.816 for R-squared and an intercept of 8345.216 for the Ordinary Least Squares model without using cross-validation[1].

[1] The coefficients are provided in the code because we had a lot of features based on applying One-Hot encoding on segments_id and seasons. Some of the coefficients are as evapotranspiration: -1.05, precipitation: 0.57, irrigation: 1.78, x: 0.0003, y: -0.0007, etc.