\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

---------------------------------------------------------------------------------------------------------------------

**Leveraging USGS stream gage records to quantify channel change in response to extreme precipitation events in California**

---------------------------------------------------------------------------------------------------------------------\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Date: 06/09/2023

Western Washington University

This document includes instructions and descriptions for programs and related files for DATA-493 senior project.

File formats: ipynb, csv, py, txt.

Contents:

Section 1: Data scraping.

Section 2: Model training and testing.

Section 3: Graphics generation and further analysis.

========================================================================

**1. Data scraping**

========================================================================

**CaliforniaSites.ipynb** - This is a small notebook whose purpose is to convert the document with stream gauge numbers from txt to csv format to be used in other notebooks of this project. It contains detailed instructions on how to obtain a txt file with stream gauge info from usgs.gov. This notebook is the key starting point of this project.

**water-year summary.ipynb** - Takes a list of stream gauges and returns a database of all gauges’ water-year summaries with Location, Gage, and Remarks info. Data is scraped from the link “https://waterdata.usgs.gov/nwis/wys\_rpt/?site\_no=" + site\_no. Some gauges might not have water-year summaries or certain parts of it that we pull, so it is simply replaced with an NA string in the database. The script works for the layout of the current water-year summary page, but it is relatively easy to modify in case this changes. Inspecting HTML tags for the webpage could be helpful in that (in Chrome, HTML could be accessed with the Developer Tools function, instruction here: https://developer.chrome.com/docs/devtools/open/). Note that running the script may take several hours, depending on the number of gauges in the list. For California gauges, we had 158 gauges and it took about 2 hours to scrape. At the end of the notebook, there is also some text preprocessing code to remove unnecessary tags that were pulled with the data.

**field\_measurements.ipynb** - Takes a list of stream gauges and returns a folder with field measurement data for each gauge in separate files. The notebook contains a script to scrape each gauge’s field measurements from the link in such format: "https://waterdata.usgs.gov/nwis/measurements?site\_no=" + site\_no + "&agency\_cd=USGS&format=rdb\_expanded". The script prints out success status when the data exists, and an error where it does not.

**combine data.ipynb** - This is a small notebook whose purpose is to combine all the field measurement files obtained in field\_measurements.ipynb and produce a single database of all measurements from all available gauges.

**rating\_curve.ipynb** - Takes a list of stream gages and returns the rating curve information about those gages as provided by the USGS. Data is scraped from the following link, 'https://nwis.waterdata.usgs.gov/nwisweb/data/ratings/exsa/USGS.' + gage\_number + '.exsa.rdb', and may not contain a rating curve. Code is currently running for any input of stream gages, but can be changed for other purposes quite easily.

========================================================================

**2. Model training and testing**

========================================================================

**water\_year\_summary\_model.ipynb** - Takes water-year summary data that needs to be labeled, training data of manually labeled WA, OR, AZ, NV water-year summaries, and outputs a labeled dataset of the input file. This model has already produced an output of the labeled California gauges, and for the future use on other datasets the model will need to be run again on the new input file. The name of the output and input file can be changed in User Input, however, the thrain file should not be replaced.

========================================================================

**3. Graphics generation and further analysis**

========================================================================

**rating\_curve\_plotter.ipynb** - Takes the field measurement data, rating curve data, and a water-year summary dataset that has been labeled by the above model. Creates a series of visualizations, such as plots of stage-discharge data per each gauge with its’ rating curve, plots of stage residuals per each gauge, and a time series plot with residuals of all gauges next to each other. The plots of stage-discharge data with rating curves can be found in the “Rating curves” folder, plots of residuals for valid gauges can be found in the “Stage residuals” folder, and residuals of “maybe” gauges in “MB Stage residuals” folder, and single time series plots are located in the same directory with the notebook. All the plots can be customized to add new features. Note: when plotting stage-discharge with rating curves, the code plots for all 158 gauges that were available in the list of stream gauges. There is an option included in the comments to plot images only for the valid gauges instead.

**Map\_maker.ipynb** - Takes a list of stream gauges, their location data, and their validity as described by the above model, and creates a interactive HTML map using folium. All aspects of the map are easily changeable for other purposes, and folium allows for much personalization of the map and related HTML.

**std\_residuals.ipynb** - Takes valid gages as described by the above model, the rating curve data, and field measurements. Residuals are calculated and the standard deviation for the stage residual value is saved for that gage. Output is the a CSV file of the gage number and respective standard deviation.

**openaiForRemark** - Documentation and code file of the prompt engineering research section of this project. Code is currently not working, but is set up to show usage of openai’s NLP engines. A new API code is needed, and will be included with any openai account.

**modelTesting.txt** - Documentation file of call and response to GPT-3 for research into prompt engineering. Though not used in the scope of this project, it provides general research as to the usage of GPT-3 for NLP through prompts. File includes test information about stream gages, as well as actual stream gage information for comparison.

**Directory outline**

| Medium | Input | Output |
| --- | --- | --- |
| USGS Website | User input | 500PtSitesCAGauges.txt |
| CaliforniaSites.ipynb | 500PtSitesCAGauges.txt | california gages.csv |
| field\_measurements.ipynb | california gages.csv | field\_measurement\_data folder |
| combine data.ipynb  (inside field\_measurement\_data folder) | csv files in field\_measurement\_data folder | field\_measurements.csv |
| water-year\_summary.ipynb | california gages.csv | clean\_csv\_ water-year summary.csv |
| water\_year\_summary\_model.ipynb | clean\_csv\_ water-year summary.csv  train\_gages.csv | CA\_valid\_gages.csv |
| rating\_curve.ipynb | CA\_valid\_gages.csv | rating\_curve\_wa.csv |
| rating\_curve\_plotter.ipynb | rating\_curve\_wa.csv, CA\_valid\_gages.csv,  field\_measurements.csv | Stage Residuals folder  MB Stage residuals folder  Residual time series.jpg  Residual time series (maybe gages).jpg |
| std\_residuals.ipynb | CA\_valid\_gages.csv,  rating\_curve\_wa.csv,  field\_measurements.csv | std.csv |
| mapMaker.ipynb | CA\_valid\_gages.csv,  std.csv,  california gages.csv | map.html |
| openaiForRemark.ipynb | openai\_test\_data.csv | No current output. |