Water Availability Tool Basin Candidates Calculations

Preamble

This document outlines the methods used to calculate the three recommended candidate gauges for each watershed. The methodology is composed of three major steps: retrieving and then computing the physical characteristics of each watershed in the study area, comparing physical characteristics to produce recommended gauges and then re-ordering based on gauges in the downstream watershed. All data sources used to summarize the physical characteristics of watersheds were sourced between January and February 2022.

The base hydrography geospatial dataset used in this project is the NHDPlus high-resolution data set. Detailed documentation for this data set can be accessed at the fol-lowing link https://pubs.usgs.gov/publication/ofr20191096. For further guidance, a user manual is available at https://pubs.usgs.gov/of/2019/1096/ofr20191096.pdf. In particular, the calculations below use the terms “watershed” and “catchment” often, they are defined as follows:

A watershed is a comprehensive geographical area that represents the land drained by a river system or a body of water. It acts as a natural boundary for surface water flow, guiding it towards a common outlet, such as a river, lake, or ocean. Within the Water Availability Tool, a pre-processing step was undertaken to generate watersheds at all stream reaches within the policy area. This was performed using several tables within the NHDPlus high resolution data set. First, all stream reaches (i.e., NHDFlowline feature class table) within the policy area were identified. Watersheds were created for each stream reach by using a linking table that identifies flowing and non-flowing connections (i.e., NHDPlusFlow table) between NHDFlowline features. All identified stream reaches upstream of the stream reach in question were joined via a linking key (NHDPlusID) with the catchment polygon layer (i.e., NHDPlusCatchment polygon feature-class table). The aggregate of catchments identified as upstream of the stream reach in question is the upstream catchment basin or watershed. All NHD layers were sourced from the NHDPlus H <id> GDB gdb files for ids of 1801 HU4, 1802 HU4 and 1805 HU4.

Methodology:

The breakdown of the three major steps is as follows:

1. Computing physical characteristics for each watershed

- Annual precipitation average (1991-2020) from the Parameter-elevation Relation-ships on Independent Slopes Model (PRISM) is retrieved from the following data source https://prism.oregonstate.edu/ and spatially clipped for each catch-ment. Each catchment's annual average precipitation is then summed into its upstream watershed as a weighted average using the percentage of the area that the catchment accounts for within its upstream watershed as the weight. This is done such that the annual average precipitation for a given watershed can be given by:

$$\frac{\sum_{n=1}^{N} a_n \cdot p_n}{\sum_{n=1}^{N} a_n}$$

  Where N is the number of catchments that are contained within the watershed in question and p_n, a_n is the annual average precipitation and area of the n^{th} catchment that are contained within the watershed in question.

- Elevation data from the 10-meter 3D Elevation Program, processed as a data product from the National Hydrography NHDPlus High Resolution found at https://www.usgs.gov/national-hydrography/nhdplus-high-resolution is then spatially clipped and assigned to each catchment and summed up with a similar weighted average as used for calculating the annual precipitation average, to obtain the average elevation for each watershed in the study area. In addition, the minimum and maximum elevations are also recorded for each watershed by comparing all catchment elevations.

- The PRISM model (Parameter-elevation Relationships on Independent Slopes Model) is then sourced again for monthly precipitation (1991-2020) data which can be found at https://prism.oregonstate.edu/. The monthly precipitation is spatially clipped to create average monthly precipitations per catchment. Using the weighted average method above the watersheds in the study area are assigned average monthly precipitation values from the catchments they contain.

- Land Cover And Vegetation data is sourced from CALFIRE-FRAP found at https://map.dfg.ca.gov/metadata/ds1327.html. The data is then spatially clipped to each catchment and describes the land cover/vegetation of each catch-ment as percentages of the following categories: "Agriculture", "Barren/Other", "Conifer Forest", "Conifer Woodland", "Desert Shrub", "Desert Woodland", "Hardwood Forest", "Hardwood Woodland", "Herbaceous", "Shrub", "Urban", "Water", and "Wetland". This information is then summed up for each field as a weighted average similar to the above to obtain the Land Cover and Vegetation for each watershed.

- Elevation data from the 10-meter 3D Elevation Program referenced above is pro-cessed into an aspect raster and is classified into four values, one for each cardinal direction. This data set was spatially clipped to all catchments in the study area. After all the catchments receive values for each cardinal direction they can then

be summed up using the weighted average above to obtain these values for each watershed

- Lastly Geology data is sourced from https://www.sciencebase.gov/catalog/item/598b471de4b09fa1cb0eacfd which is then spatially clipped to each catch-ment and contains data summarizing the geology for each catchment into the following fields as percentages: “Igneous And Metamorphic Undifferentiated”, “Igneous Intrusive”, “Igneous Volcanic”, “Melange”, “Metamorphic And Sedi-mentary Undifferentiated”, “Metamorphic Schist”, “Metamorphic Serpentinite”, “Metamorphic Volcanic”, “Sedimentary Clastic”, “Unconsolidated Undifferenti-ated”, and “Water”. Each of these fields is then summed up for the watershed containing the catchment using the weighted average method above.

2. Statistical Analysis of Physical Characteristics

- Now that the physical characteristics have been gathered for each watershed they must be used to generate the three recommended candidate gauges. First, the precipitation data was reduced in dimensionality from twelve parameters (e.g. monthly precipitation) to two to reduce noise and emphasize the meaningful pat-terns. This was done by assuming p^m_n represents the n^{th} watershed’s monthly precipitation at the m^{th} month then we can obtain each watershed’s monthly standardized z-score using:

$$z_n^m = \frac{p_n^m - \mu^m}{\sigma^m}$$

where μ^m and σ^m are the population mean and standard deviation of the precip-itation of all watersheds in the study area at the m^{th} month. Calculating these z-scores ensures that the mean of the data is 0 and the standard deviation is 1. Using the new set of z-scores a principal component analysis can be performed to reduce the dimensionality to 2 vectors of data using a LAPACK-based Singular Value Decomposition implementation. This leaves us with two principal compo-nents that summarize the 12 months of precipitation data for each watershed.

- The same process is then followed for the geology and land cover/vegetation data as the geology data is composed of eleven vectors and the land cover/vegetation data is composed of 13 vectors. After the data sets are standardized to z-scores and have a principal component analysis run on them we are left with two principal components for each data set.

- Now each watershed has the following data associated with it: Elevation min, Ele-vation Max, Elevation Average, North Facing %, South Facing %, East Facing %, West Facing %, Annual Average Precipitation, Area(mi^2), principal Component 1 Geology, principal Component 2 Geology, principal Component 1 Monthly Pre-cipitation, principal Component 2 Monthly Precipitation, principal Component 1 Land Cover / Vegetation, principal Component 2 Land Cover / Vegetation, lon-gitude, and latitude. Longitude and latitude are computed as the centroid of the most downstream catchment polygon (i.e.,

the mouth of the watershed). For the purposes of this document, they will be referred to as the centroid of the water-shed. Now all vectors of data are transformed into a standard z-score as monthly precipitation was above except for longitude and latitude which are transformed into the following non-standard z-scores:

$$z_n^{\text{long}} = \frac{(\text{long}_n - \mu^{\text{long}}) \cdot 10}{\sigma^{\text{long}}} \text{ or } z_n^{\text{lat}} = \frac{(\text{lat}_n - \mu^{\text{lat}}) \cdot 10}{\sigma^{\text{lat}}}$$

where z_n^{long} is the z-score for the centroid longitude of the n^{th} watershed, μ^{long} is the population mean centroid longitude over all watersheds in the study area, and similarly, σ^{long} is the population standard deviation over all the watershed's centroid longitude. The variables for latitude represent the same characteristics but for latitude instead of longitude. This is a non-standard z-score and has a standard deviation of 10 instead of the expected unit length standard deviation. This adjustment is made to give more weight to the centroid longitude and latitude in the subsequent calculations.

- Next the watersheds that have had gauges on them for 10 or more years are determined so that a Euclidean pairwise distance matrix can be made using them. The Euclidean pairwise matrix is an m by n matrix where n is the number of watersheds in the study area and m is the number of watersheds with gauges on them. The $(i, j)^{th}$ element of the Euclidean pairwise distance matrix can be calculated by:

$$\sqrt{\sum_{c=1}^{17} (f_i^c - f_j^c)^2}$$

where f_i^c is the c^{th} z-scored physical characteristic of the i^{th} gauged watershed such that $0 < i \leq m$, f_j^c is the c^{th} z-scored physical characteristic of the j^{th} watershed within the study area such that $0 < j \leq n$. The sum is 17 as after the principal component analyses that is the number of physical characteristics each watershed vector has.

- A process of re-weighting is then completed on the outputted data using the years of record and whether the candidate gauge is active or inactive. The squared sum above is re-weighted using the following logic. This is done so users will more likely see their recommended gauges having more data and more active than otherwise. (n => number of years of record)

$$\text{if } n \leq 20$$

$$distance = distance/n$$

$$\text{else}$$

$$distance = distance/20$$

Then if the gauge is reporting data actively (up to imported date):

$$distance = distance/2$$

- Lastly to populate the recommended candidate gauges for the watershed y the Euclidean pairwise distance matrix's column y is inspected and the gauges are ordered, smallest to largest, by their associated euclidean distance

3. Adjust Based on Overlap of Downstream watershed

- Lastly, the list generated in step 2 above is partitioned into two lists, maintaining the order from above. These lists are based on whether the given watershed contains the point representing the gauge. The gauges in the downstream watershed are placed at the top of the list, followed by the other gauges.
- The top 3 candidate gauges are then selected as the first three elements of the combined list