Applied Geostatistics CE369 Final Project Report

Spring 2022

Streamflow Drought in the Colorado River Basin

By: Ali Dadkhah and Matthew Thompson

Introduction

As climate change continues, the way people live is impacted more and more by the increase in frequency of extreme weather. One of the most tangible resources being affected is water. Drought has been a constantly occurring natural hazard in recent years and has devastating socioeconomic, ecological, and even political impacts. Many of drought's impacts are associated with hydrological drought (drought in lakes, rivers, and groundwater), therefore it is important to understand the development and behavior of hydrological drought (van Loon, 2015). Hydrological drought is a broad term related to negative anomalies in surface and subsurface water, thus streamflow drought and groundwater drought are sometimes defined separately (Mishra & Singh, 2010). Streamflow drought is typically thought of as a period in which in-stream flows become depleted beyond a particular flow level, therefore, analysis of streamflow time series during a drought could give a better understanding of drought behavior(Zaidman et al., 2001). Numerous studies leveraged streamflow data to understand and forecast drought behavior (Mishra & Singh, 2011). A good example of utilizing streamflow data for understanding hydrological drought is the study done by D. Peña-Angulo in which changes in the characteristics of hydrological droughts were assessed for different regions of Europe based on the data for 3224 stream gauges spanning 1962-2017 (Peña-Angulo et al., 2022). Spatial scale is a crucial factor in hydrological drought assessment (Yu et al., 2020) thus, in this project we tried to find similar behavior during drought across different watersheds in the Colorado River Basin by clustering them with k-means.

Clustering is a data mining method and a special type of it is time series clustering. Time series clustering is mostly used for finding patterns in a dataset, either frequent or surprisingly rare patterns (anomaly detection) (Aghabozorgi et al., 2015).

The time series data we clustered are temporal semivariograms of days during drought for 425 stream gauges positioned across the study area.

Materials and Methods

Study area

The Colorado River is about 1,450 miles long, with headwaters in Colorado and Wyoming, and it eventually flows across the international border into Mexico. The drainage basin area of about 246,000 square miles includes all of Arizona, and parts of California, Colorado, New Mexico, Nevada, Utah, and Wyoming. The Colorado River is an important water resource for areas outside of the basin, including Denver, Salt Lake City, Albuquerque, Los Angeles, and San Diego for public (municipal) supply, and the Imperial Valley in California for agricultural water supplies. The river and its tributaries provide water to nearly 40 million people, both within and outside of the basin, and irrigates nearly 5.5 million acres of agricultural lands (Colorado River Basin Water Supply and Demand Study (usbr.gov, 2012)). The study area for this project is the Colorado River Basin, including the area within a 100-mile radius around it. Figure 1 shows the study area.

Data

For this project, we used USGS streamflow data collected over 40 years (1980-2021). There are 425 stream gauges located across the study area (Figure 1). The distribution of gauges

over the study area is shown in Figure 2. Different variables for each site are available, from which we chose to use five variables, namely, variable threshold daily streamflow percentile, minimum daily temperature, daily precipitation, snow water equivalent and soil moisture for medium depth (10-40cm). Drought event records for all the 425 sites were provided by USGS as well.

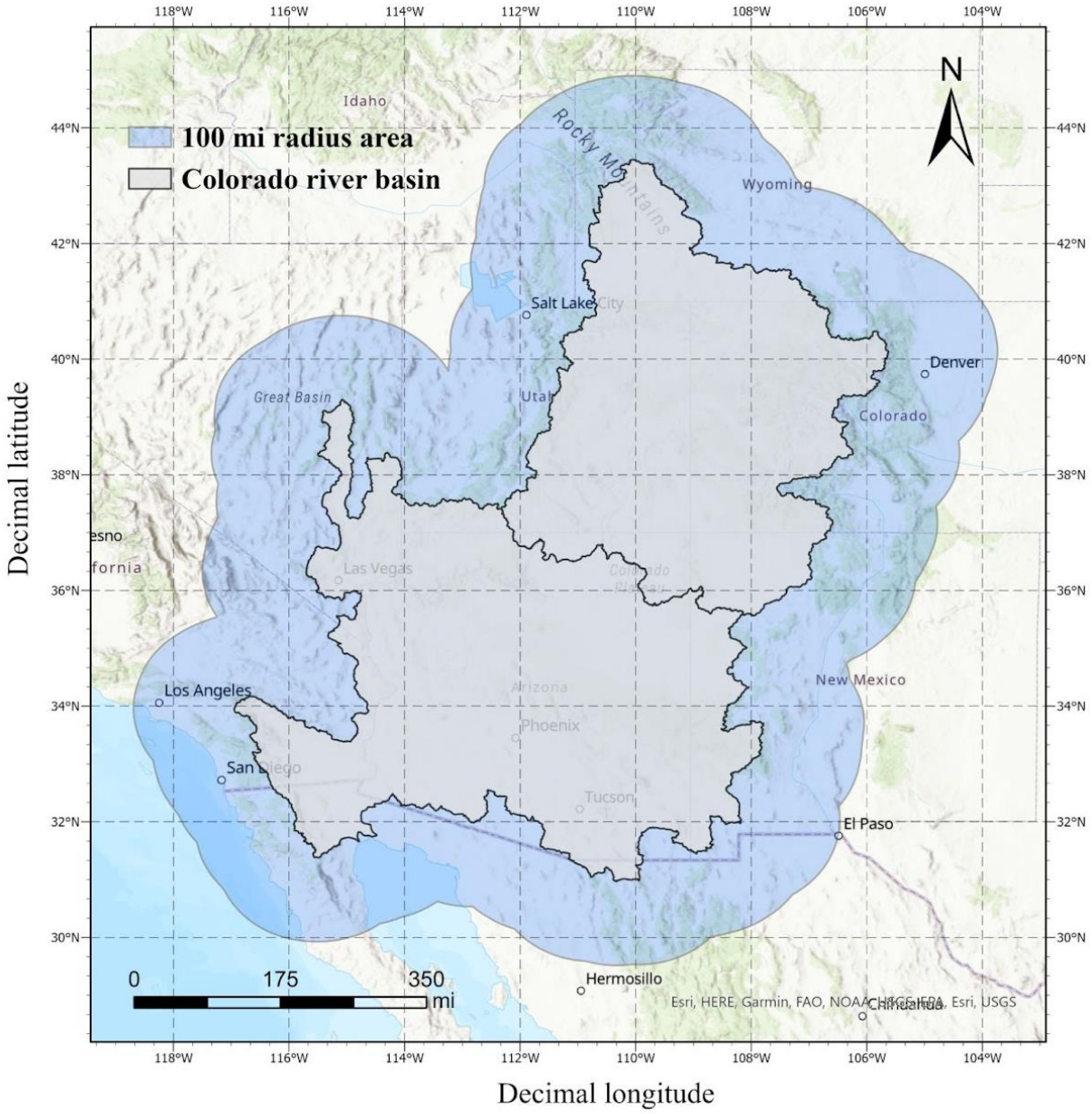


Figure 1: Colorado river basin and 100 miles surrounding area

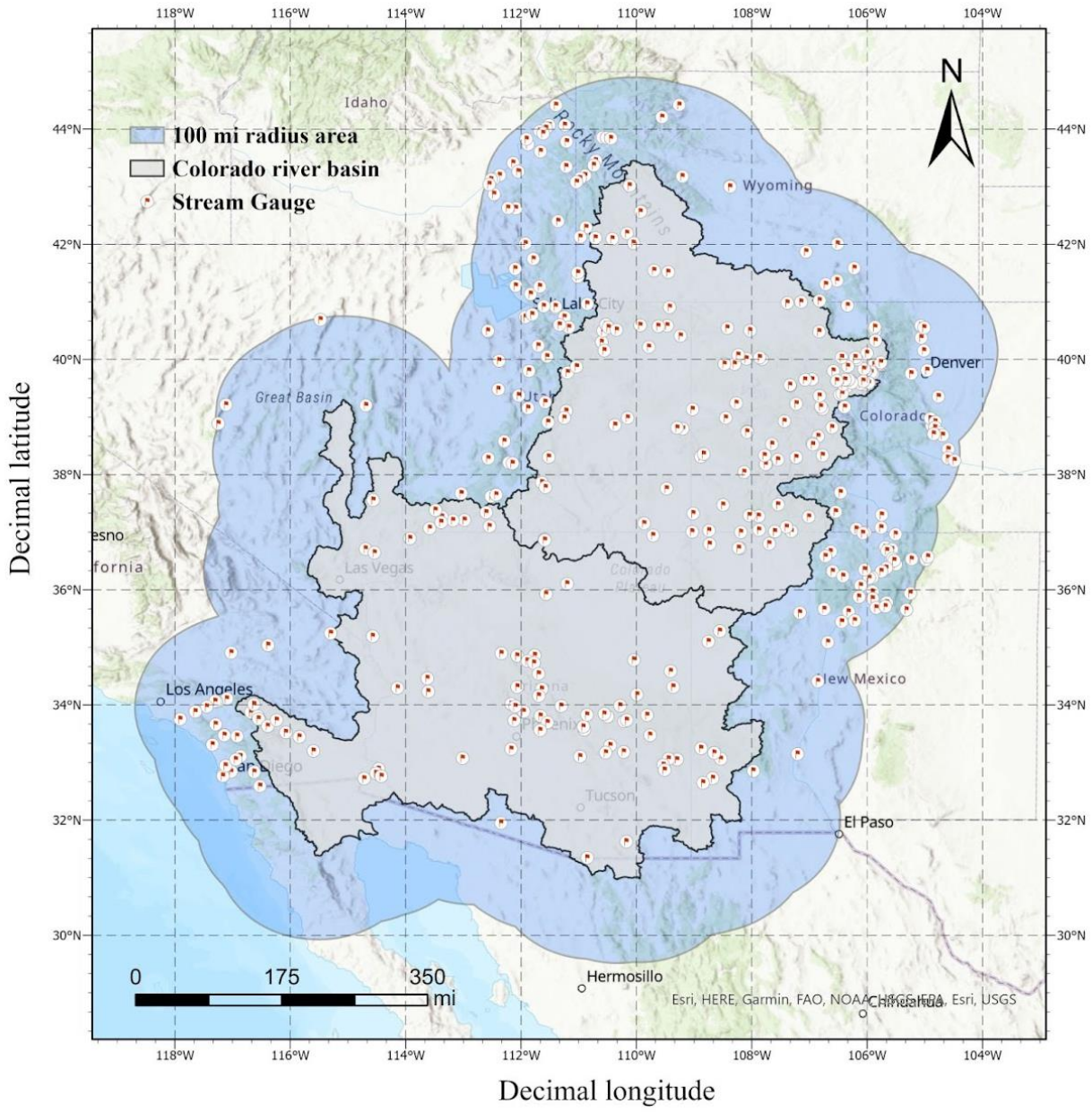


Figure 2: Gauges distribution over the study area

Data preprocessing

The workflow of the data preprocessing was twofold, first, cleaning the daily data in which rows with null values for any of the variables of interest were excluded and unnecessary columns were dropped. Second, gauges with drainage areas greater than 100,000 were pulled out.

Temporal semivariogram

A temporal semivariogram of each variable was calculated for each site as follows. Drought events with respect to the 20 percent threshold were extracted from the drought event records for each site. Having the start date and end date of the drought events, for each variable, daily values were pulled out and the semivariance values were calculated with a time step equal to one. The semivariance values and their corresponding lags were stored. At this step, a raw variogram or cloud variogram can be plotted and by binning the values for each day (i.e., bin width equal to 1 day), the final temporal semivariogram which can be interpreted as a time series was calculated.

Time series clustering

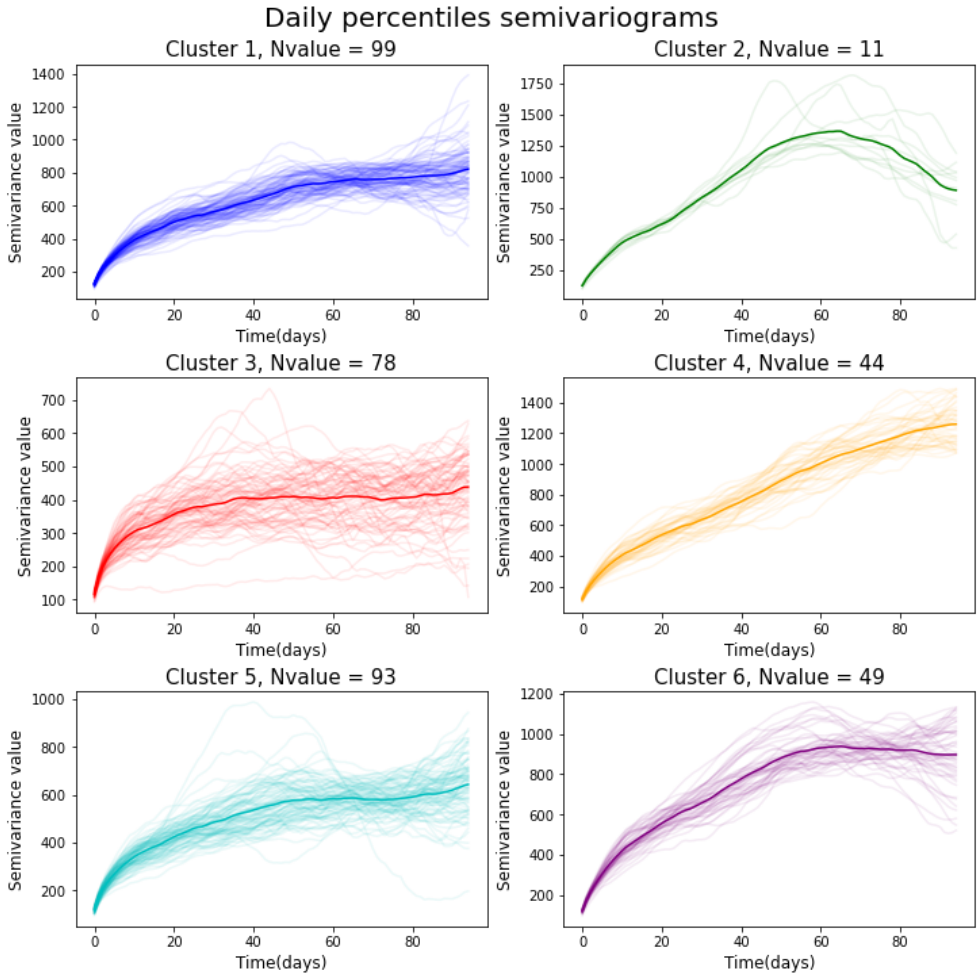
As mentioned above, each final semivariogram can be considered a time series and the k-means clustering algorithm was employed for clustering semivariograms of each variable for all the sites. Euclidean distance was used as the distance metric for clustering the time series data and the number of clusters for all variables were picked by eye. Validating this number is out of the scope of this project. To showing clustered, color-coded gauges on the map, they were separated based on whether they are part of USGS Hydro-Climate Data Network (HCDN) or not.

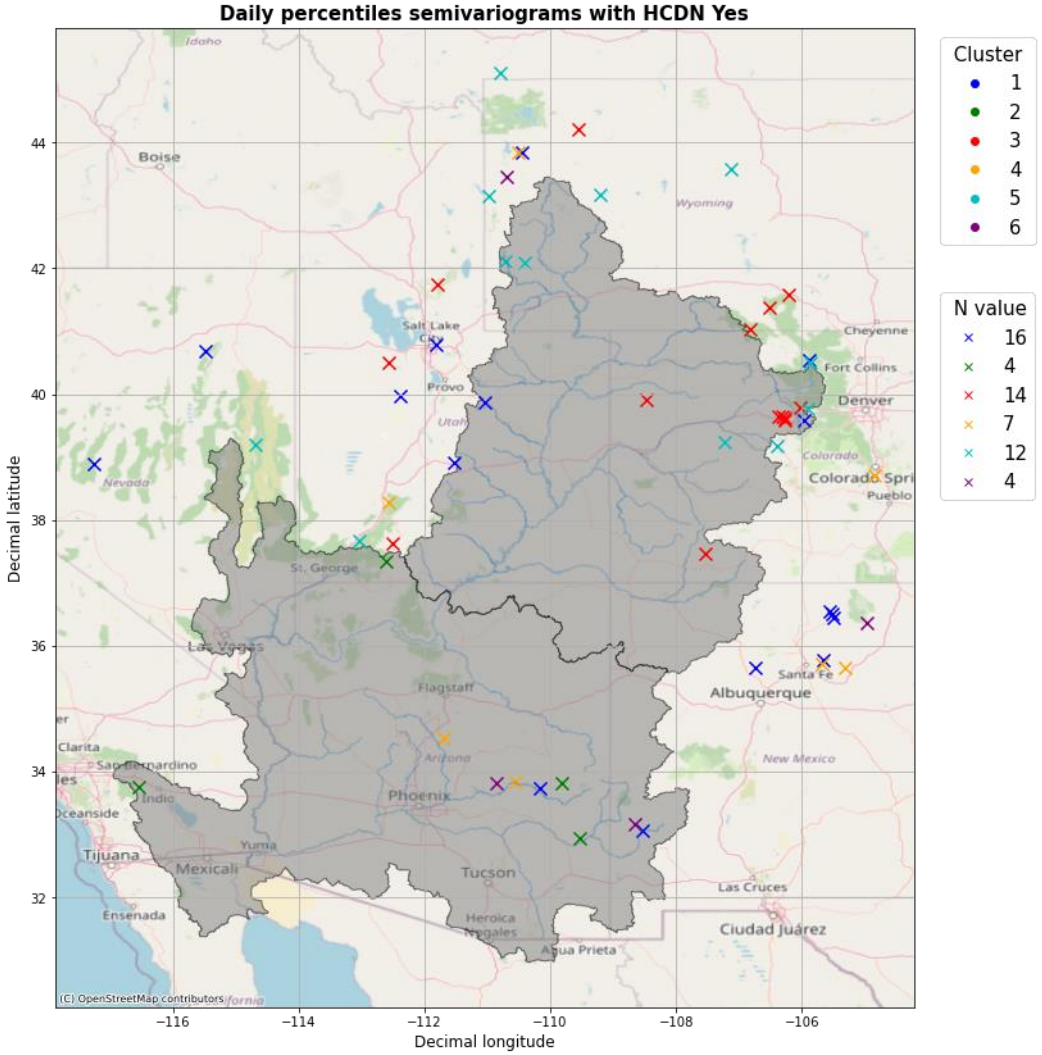
Results

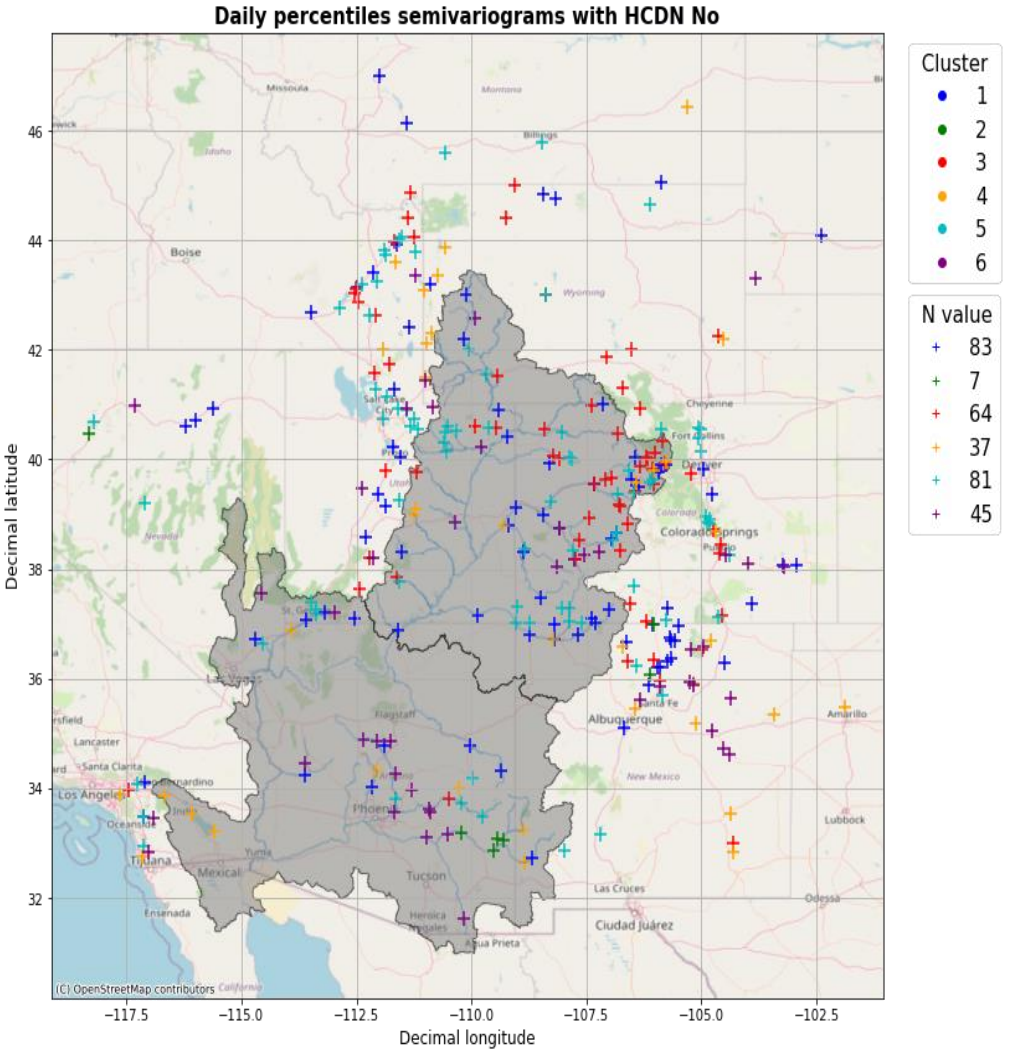
Temporal semivariograms for each site for five different variables were calculated. For each variable, temporal semivariograms were clustered using a k-means clustering algorithm and were color-coded on a map plot based on their cluster number.

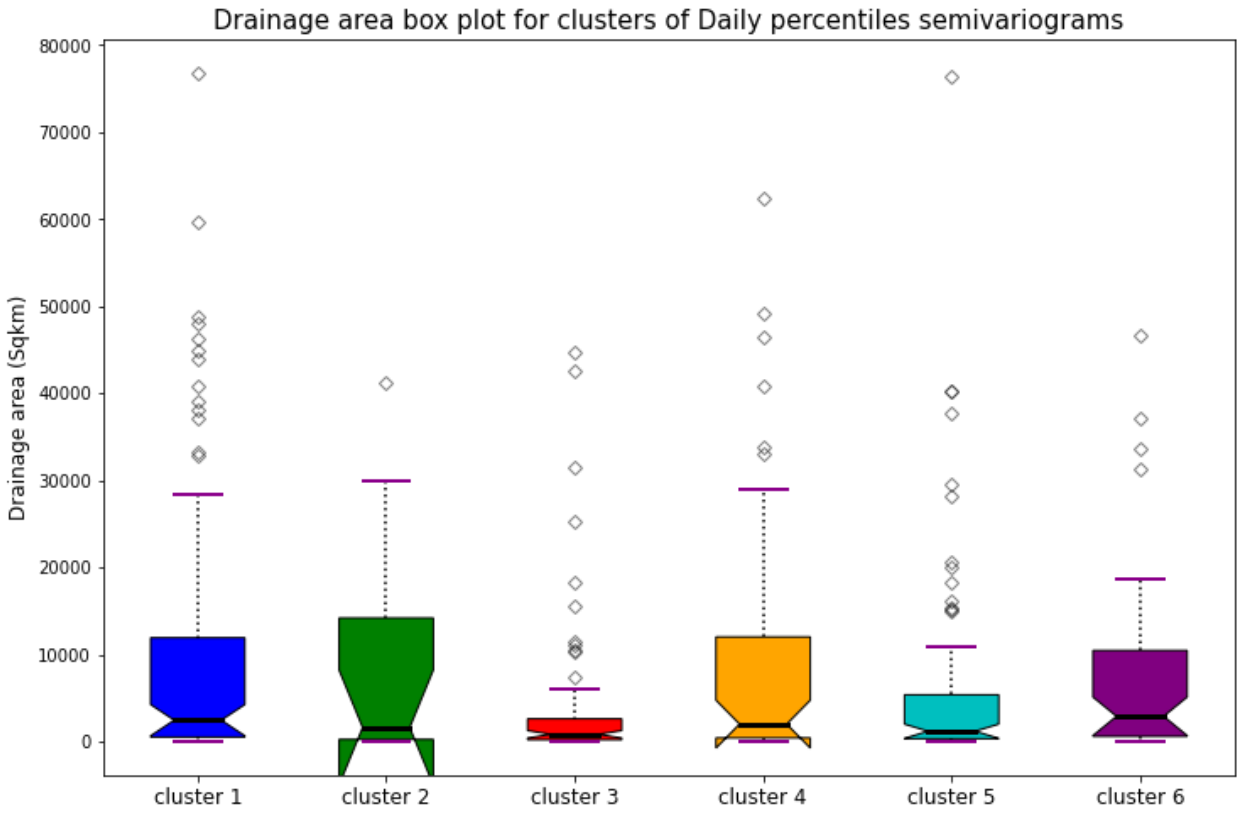
Daily percentile semivariograms

Daily percentile variable is a value between 0 and 100 calculated for each day using (in our case) 20 percent variable threshold averaged over 30 days (thresh_20_jd_30d_avg) and the same day streamflow value. Weibull plotting position method was used for computing the daily percentile values. Figure 3 shows clusters of daily percentile semivariograms. Figure 4 and Figure 5 show the distribution of stream gauges based on whether they are part of USGS HCDN, while also using color-coding to communicate cluster assignment. A boxplot of the drainage area for each site (i.e., watershed corresponding to each gauge) is shown in Figure 6.




Daily percentiles semivariograms with HCDN Yes
Cluster
1
2
3
4
5
6
N value
16
4
14
7
12
4
Decimal latitude
Decimal longitude
Boise
Salt Lake
Provo
Denver
Cheyenne
Fort Collins
Colorado Spri
Pueblo
St. George
Las Vegas
Flagstaff
Albuquerque
Santa Fe
Phoenix
Tucson
Las Cruces
Ciudad Juárez
Tijuana
Mexicali
Ensenada
Yuma
Nevada
Utah
Colorado
Arizona
New Mexico
Wyoming
(C) OpenStreetMap contributors

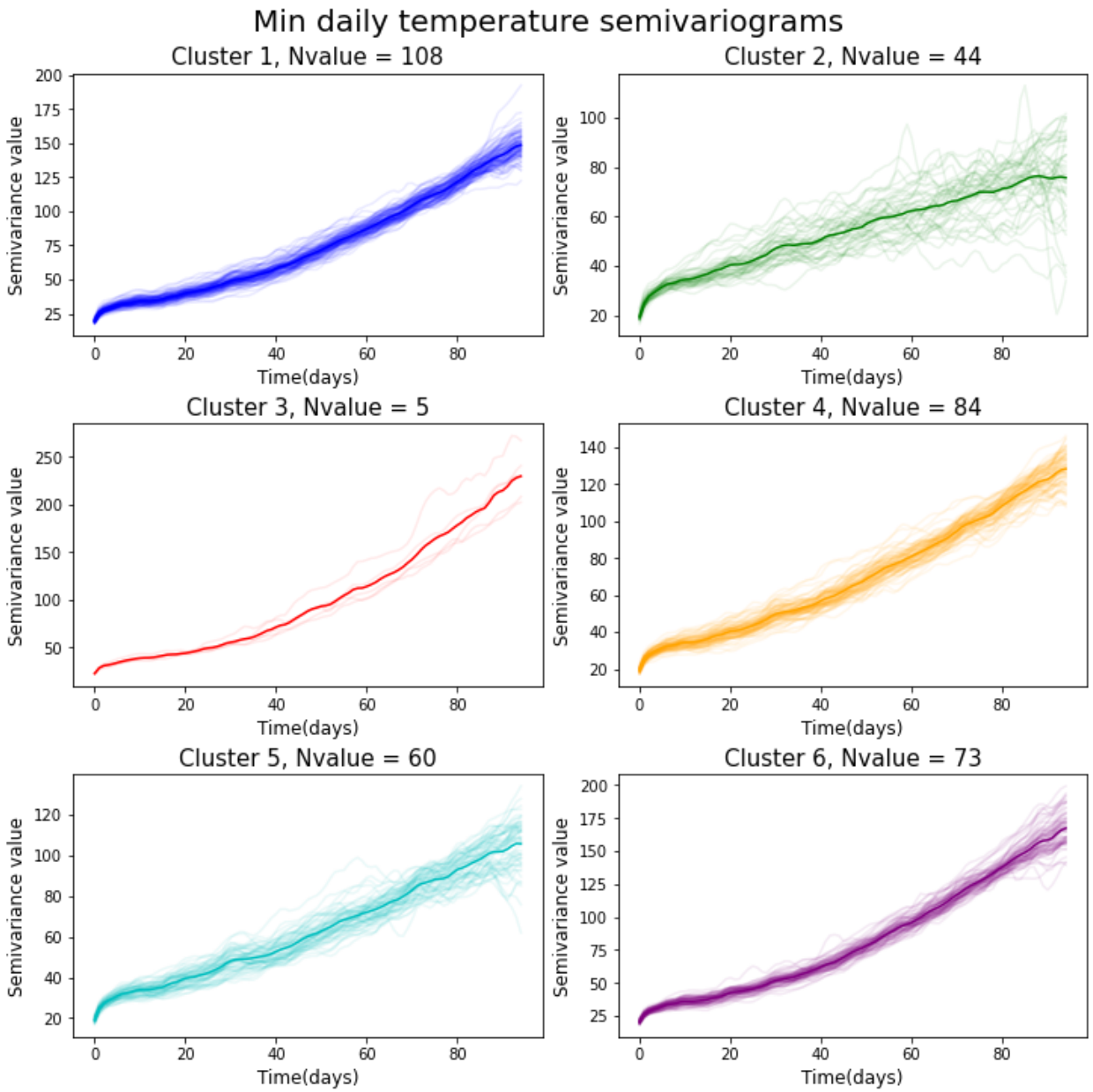

Daily percentiles semivariograms with HCDN No
Cluster
1
2
3
4
5
6
N value
83
7
64
37
81
45
Decimal latitude
Decimal longitude
Boise
Missoula
Billings
Cheyenne
Fort Collins
Denver
Colorado Springs
Albuquerque
Las Vegas
Flagstaff
Santa Fe
Amarillo
Lubbock
Odessa
Los Angeles
Phoenix
Tucson
Las Cruces
Ciudad Juárez
Tijuana
Mexicali
Ensenada
Yuma
Nevada
Utah
Colorado
New Mexico
Wyoming
(C) OpenStreetMap contributors

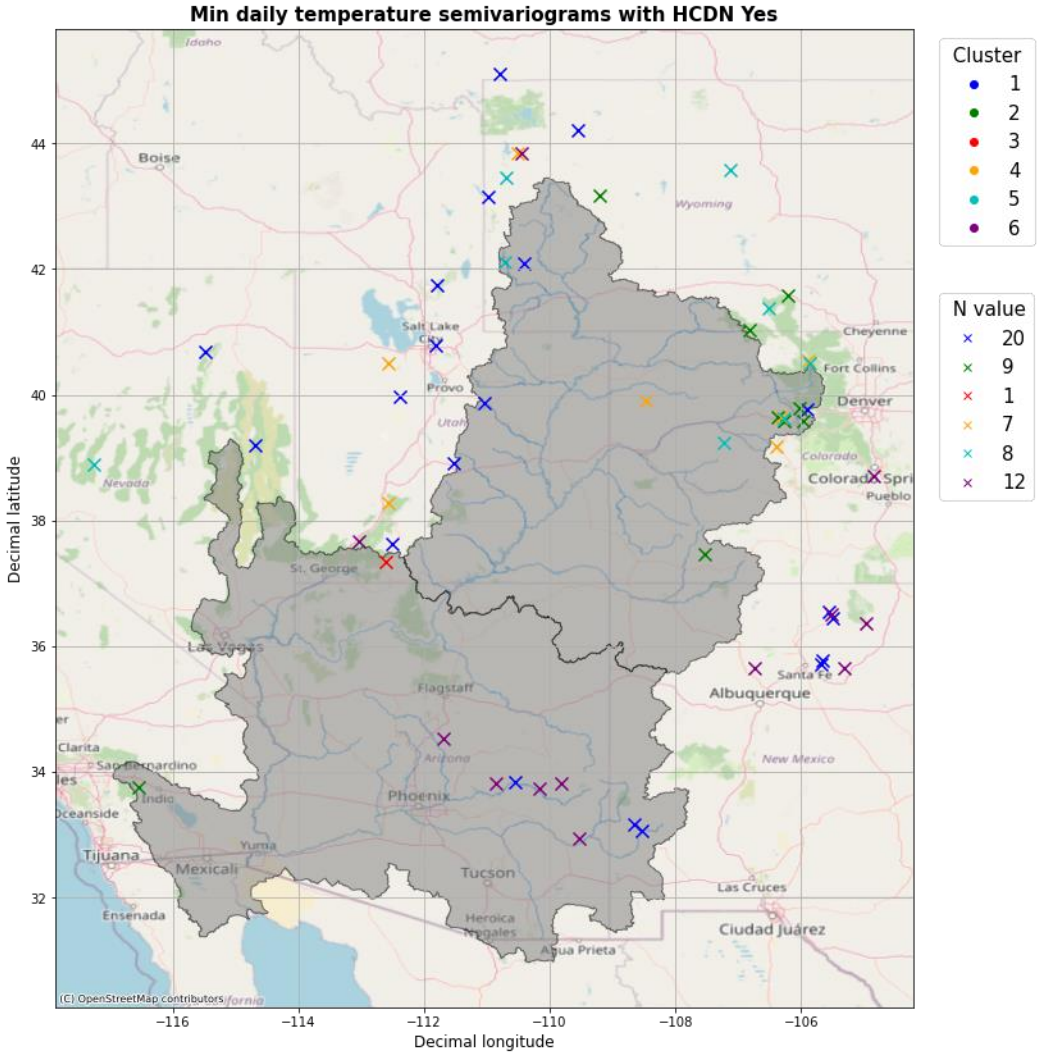

Drainage area box plot for clusters of Daily percentiles semivariograms
Drainage area (Sqkm)
80000
70000
60000
50000
40000
30000
20000
10000
0
cluster 1
cluster 2
cluster 3
cluster 4
cluster 5
cluster 6

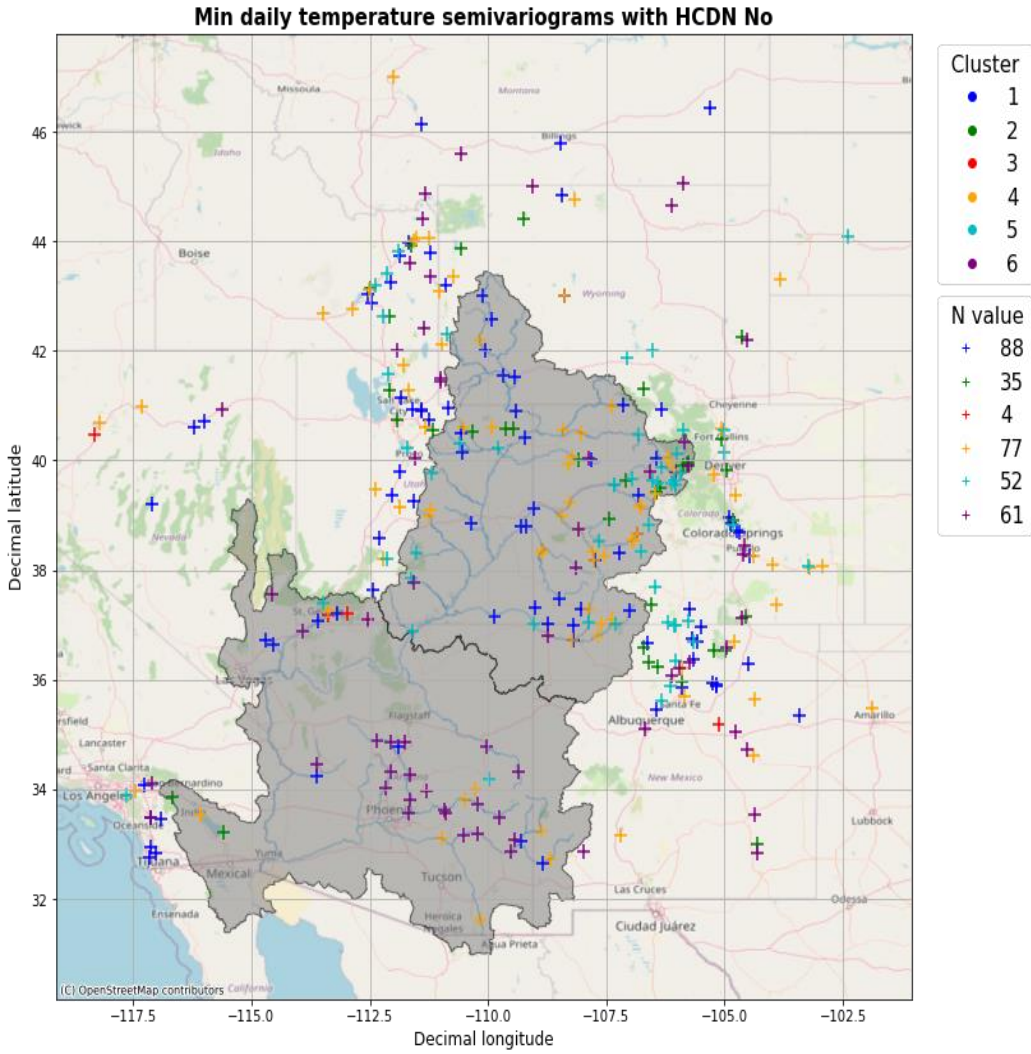
Min daily temperature semivariograms

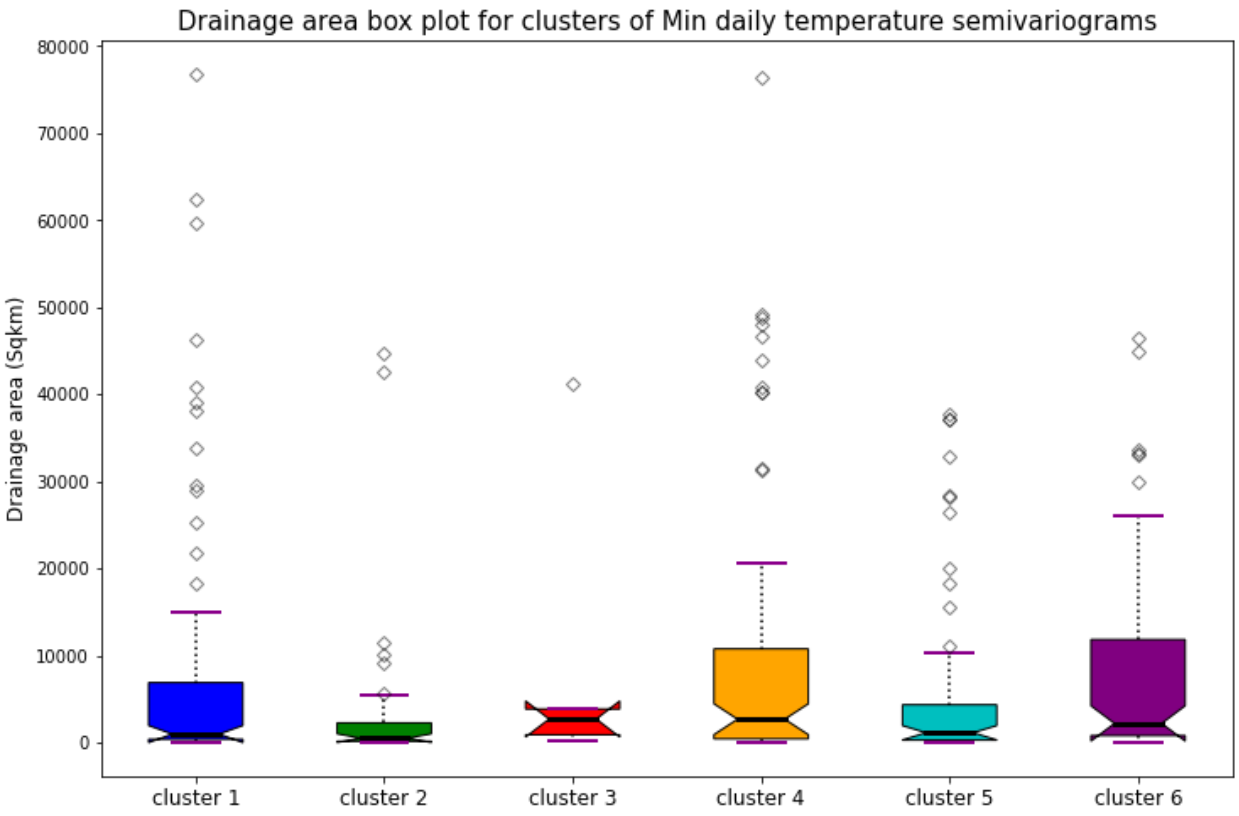
A plot of clusters, the distribution of gauges in each cluster, and a boxplot of drainage sizes are shown in Figures 7, 8 and 9, respectively.

The plots below depict the results for the other 3 variables.

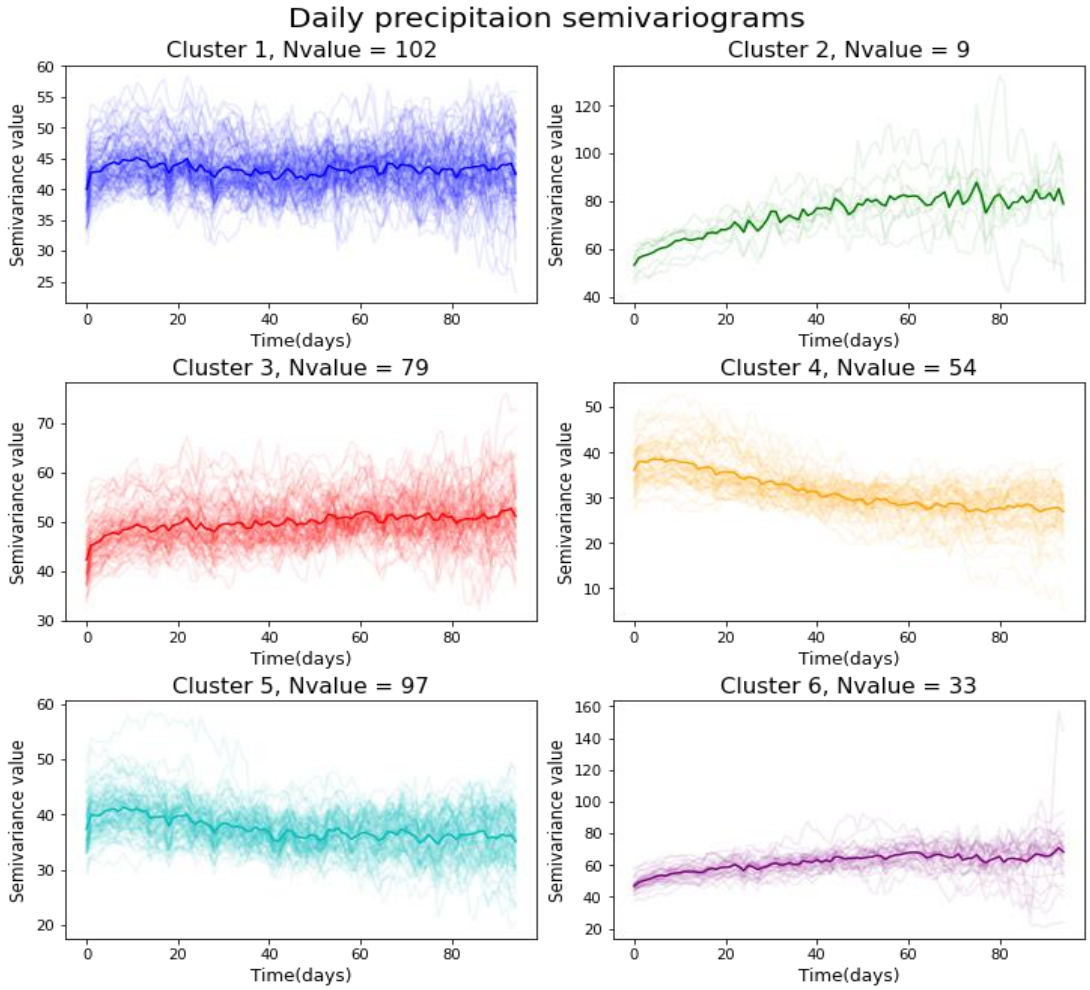


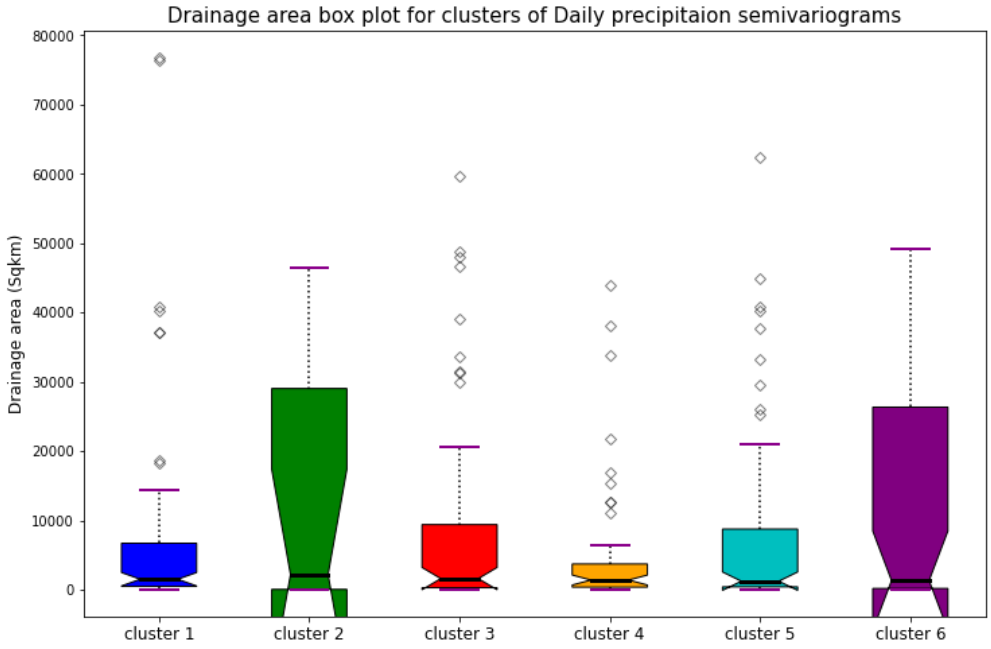

Min daily temperature semivariograms with HCDN Yes
Cluster
1
2
3
4
5
6
N value
20
9
1
7
8
12
Decimal latitude
Decimal longitude
Boise
Salt Lake City
Provo
Cheyenne
Fort Collins
Denver
Colorado Springs
Pueblo
St. George
Las Vegas
Flagstaff
Albuquerque
Santa Fe
Phoenix
Tucson
Tijuana
Mexicali
Ensenada
Las Cruces
Ciudad Juárez
Agua Prieta
(C) OpenStreetMap contributors

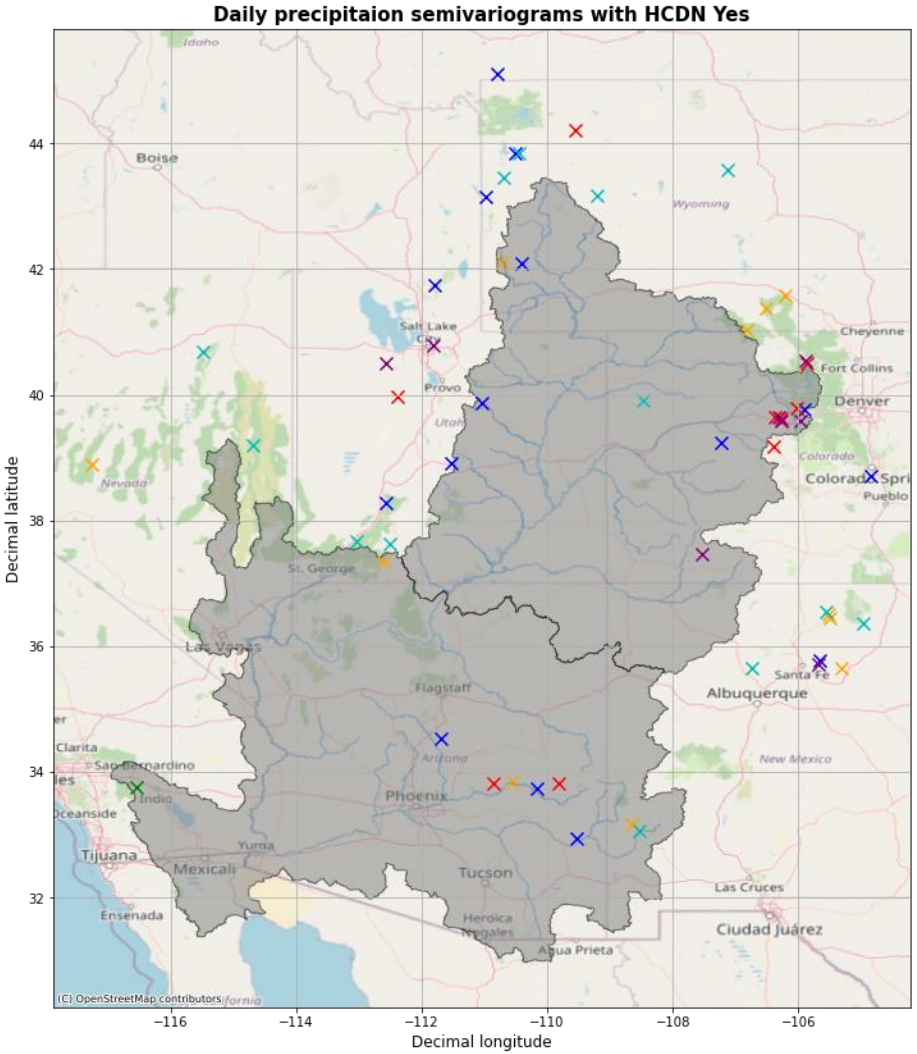

Min daily temperature semivariograms with HCDN No
Cluster
1
2
3
4
5
6
N value
88
35
4
77
52
61
Decimal latitude
Decimal longitude
Boise
Missoula
Billings
Cheyenne
Fort Collins
Denver
Colorado Springs
Albuquerque
Santa Fe
Amarillo
Lubbock
Odessa
Las Cruces
Ciudad Juárez
Tucson
Phoenix
Los Angeles
Lancaster
Tijuana
Mexicali
Ensenada
(C) OpenStreetMap contributors

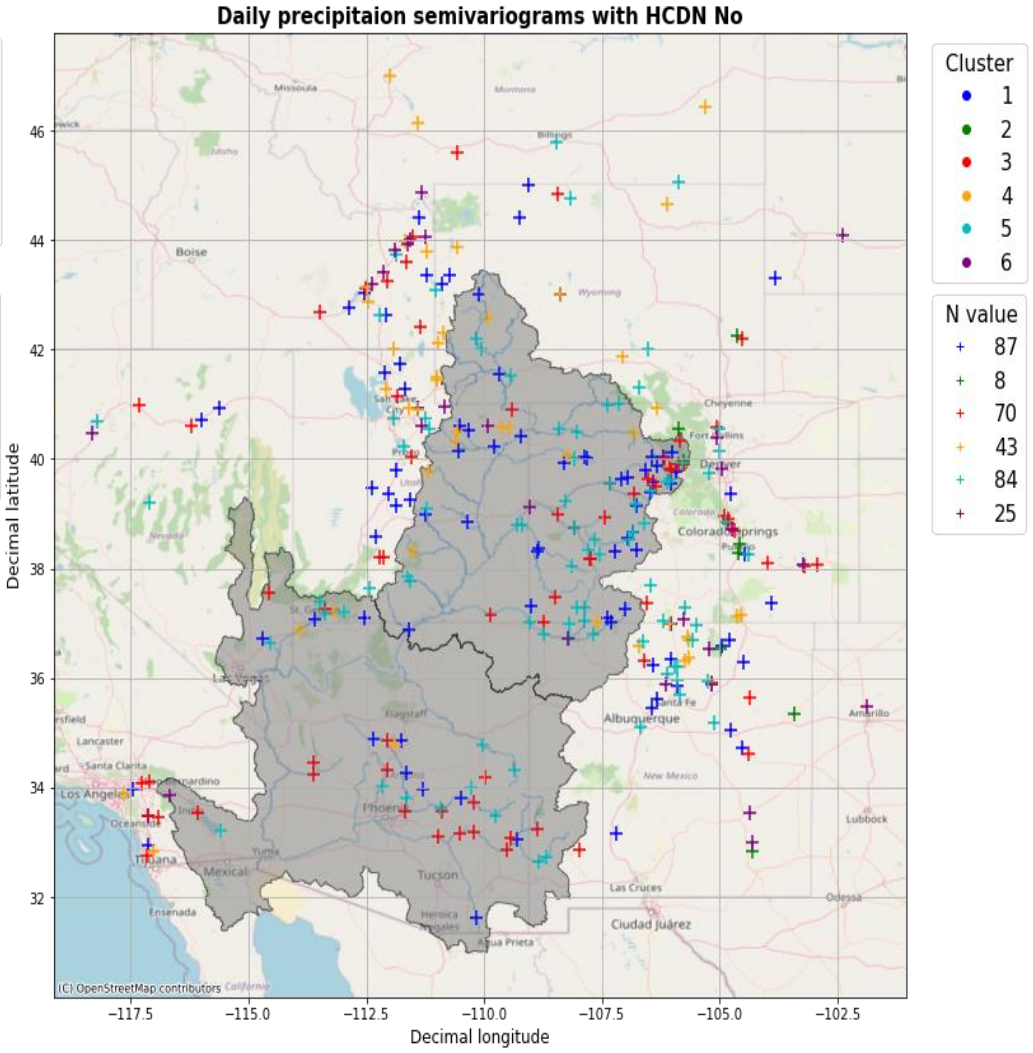

Drainage area box plot for clusters of Min daily temperature semivariograms
Drainage area (Sqkm)
80000
70000
60000
50000
40000
30000
20000
10000
0
cluster 1
cluster 2
cluster 3
cluster 4
cluster 5
cluster 6

Daily precipitation semivariograms

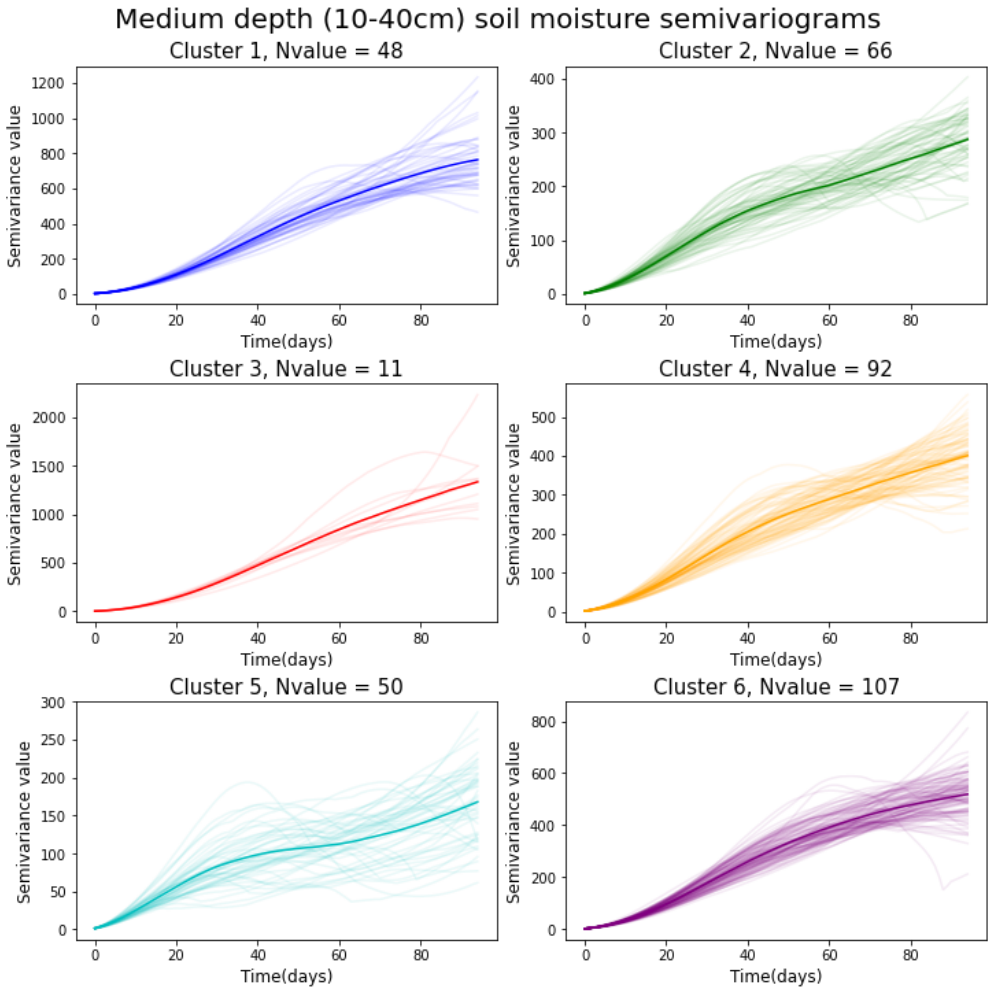


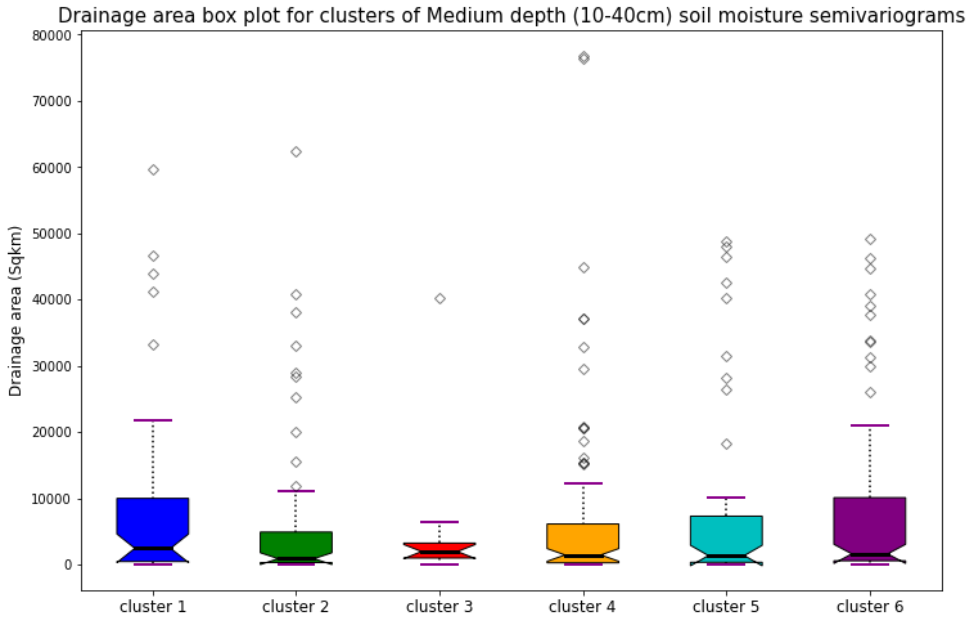


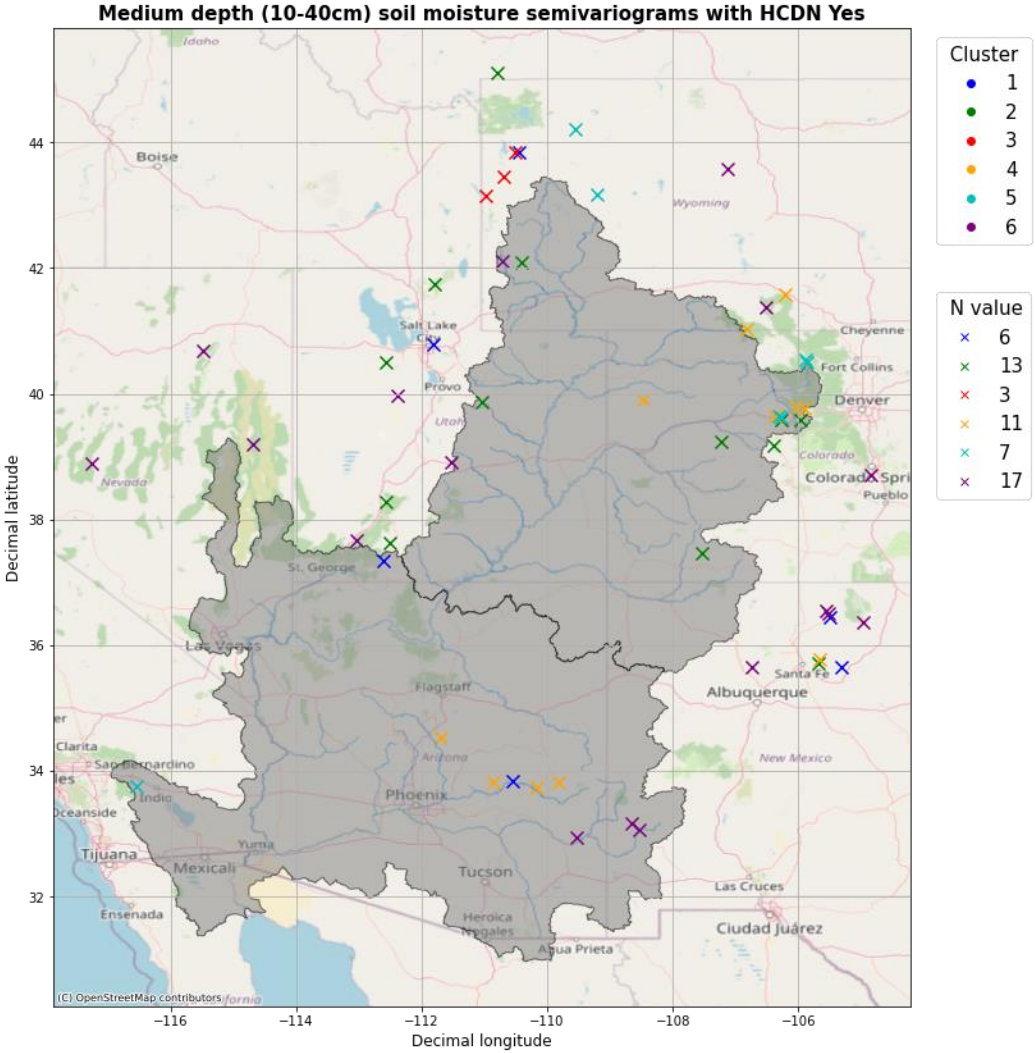

Daily precipitaion semivariograms with HCDN Yes
Cluster
1
2
3
4
5
6
N value
15
1
9
11
13
8
Decimal latitude
Decimal longitude
Boise
Salt Lake City
Provo
Cheyenne
Fort Collins
Denver
Colorado Springs
Pueblo
St. George
Las Vegas
Flagstaff
Albuquerque
Santa Fe
Phoenix
Tucson
Las Cruces
Ciudad Juárez
Tijuana
Mexicali
Ensenada
Yuma
Agua Prieta
(C) OpenStreetMap contributors

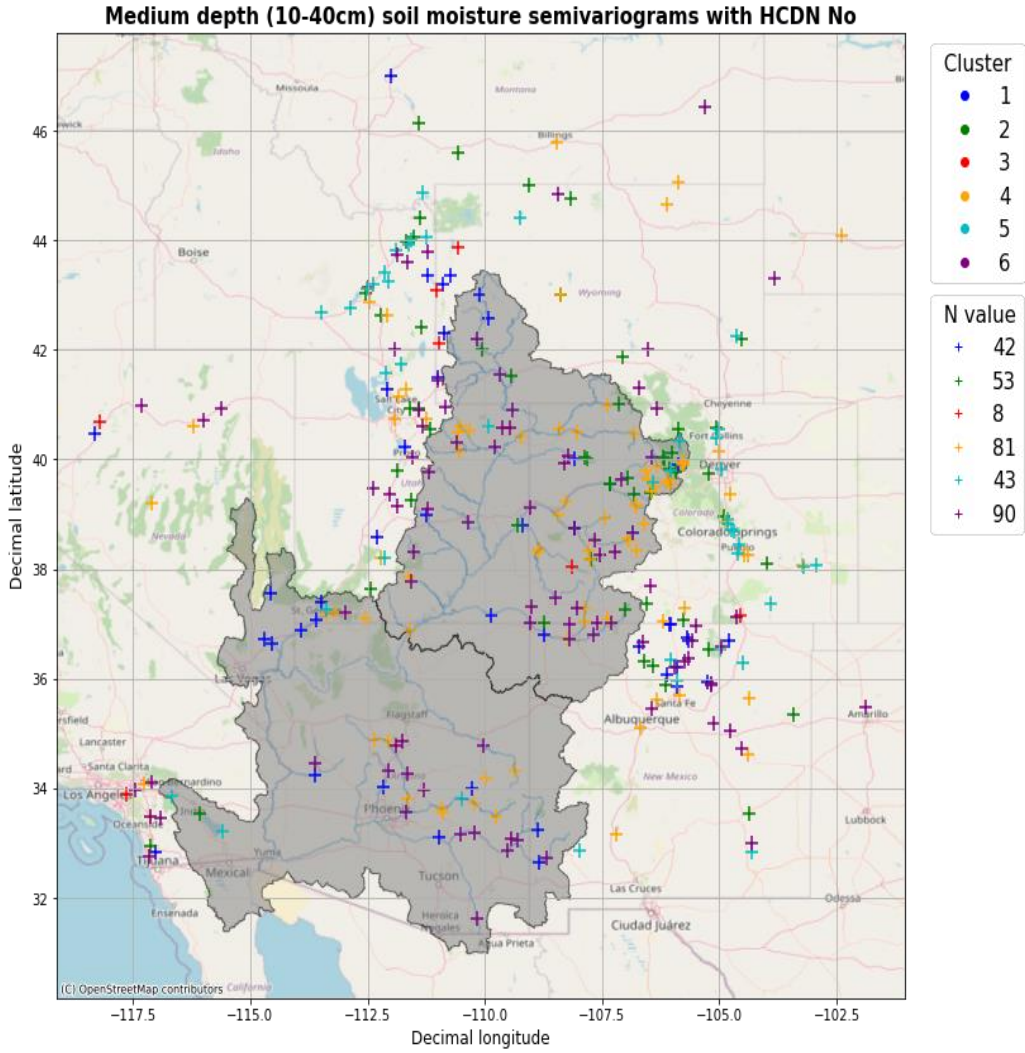

Daily precipitaion semivariograms with HCDN No
Cluster
1
2
3
4
5
6
N value
87
8
70
43
84
25
Decimal latitude
Decimal longitude
Boise
Missoula
Billings
Cheyenne
Colorado Springs
Albuquerque
Santa Fe
Amarillo
Lubbock
Odessa
Las Cruces
Ciudad Juárez
Tucson
Phoenix
Tijuana
Mexicali
Ensenada
Yuma
Agua Prieta
(C) OpenStreetMap contributors

Medium depth (10-40cm) soil moisture semivariograms

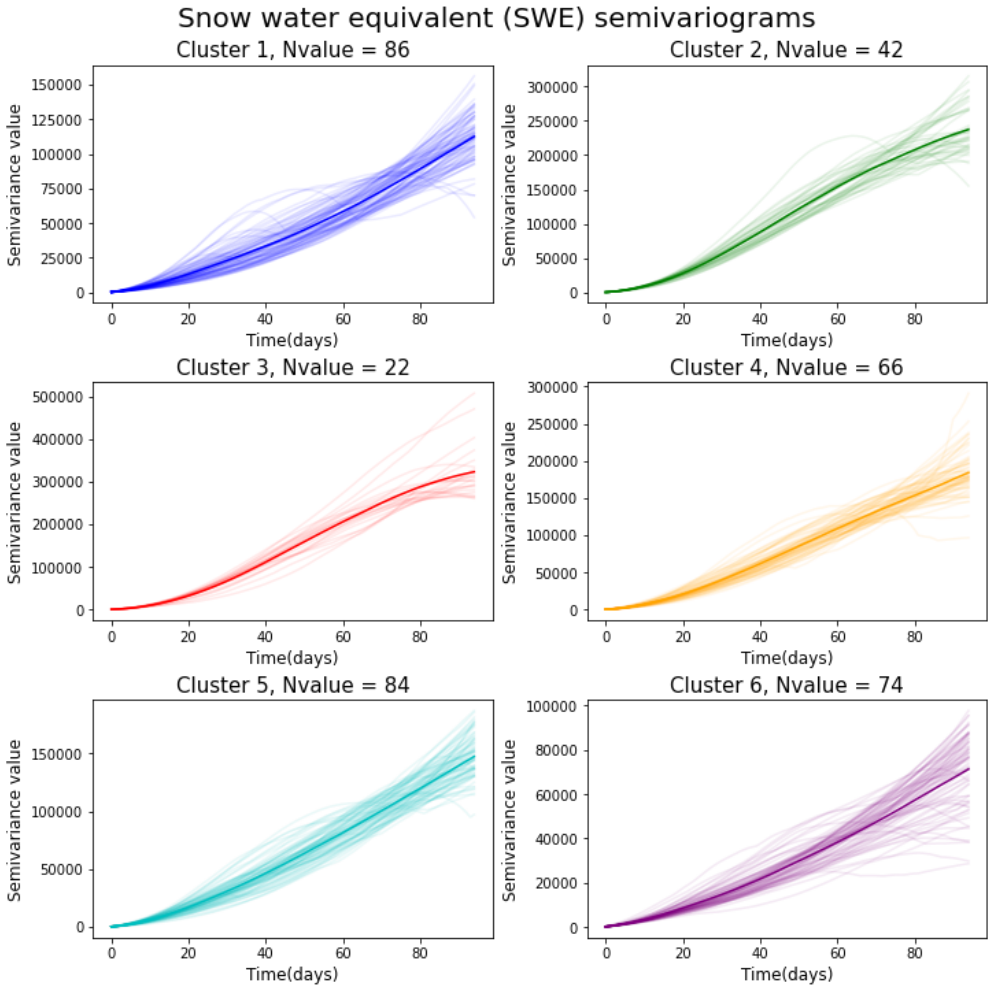


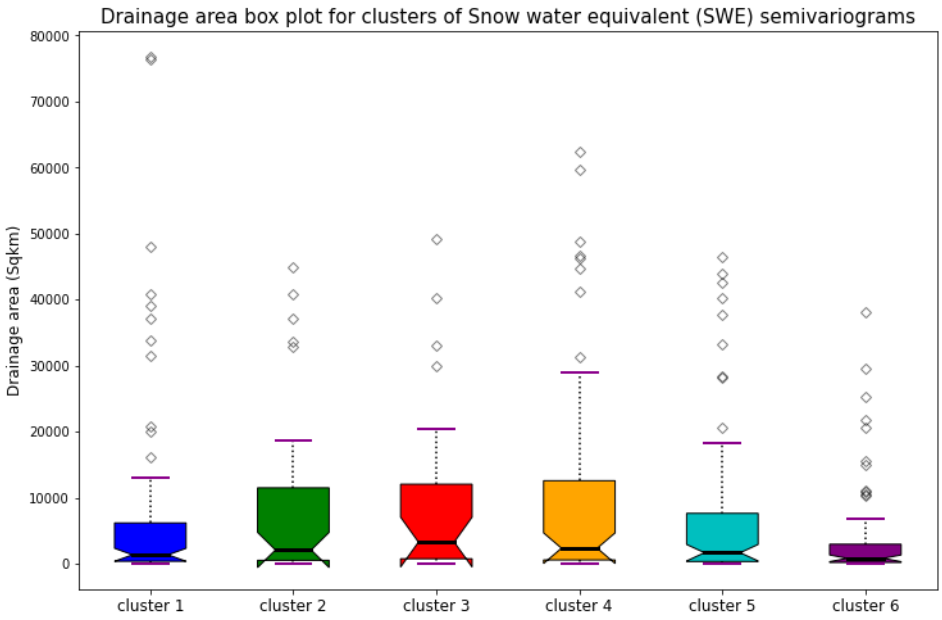


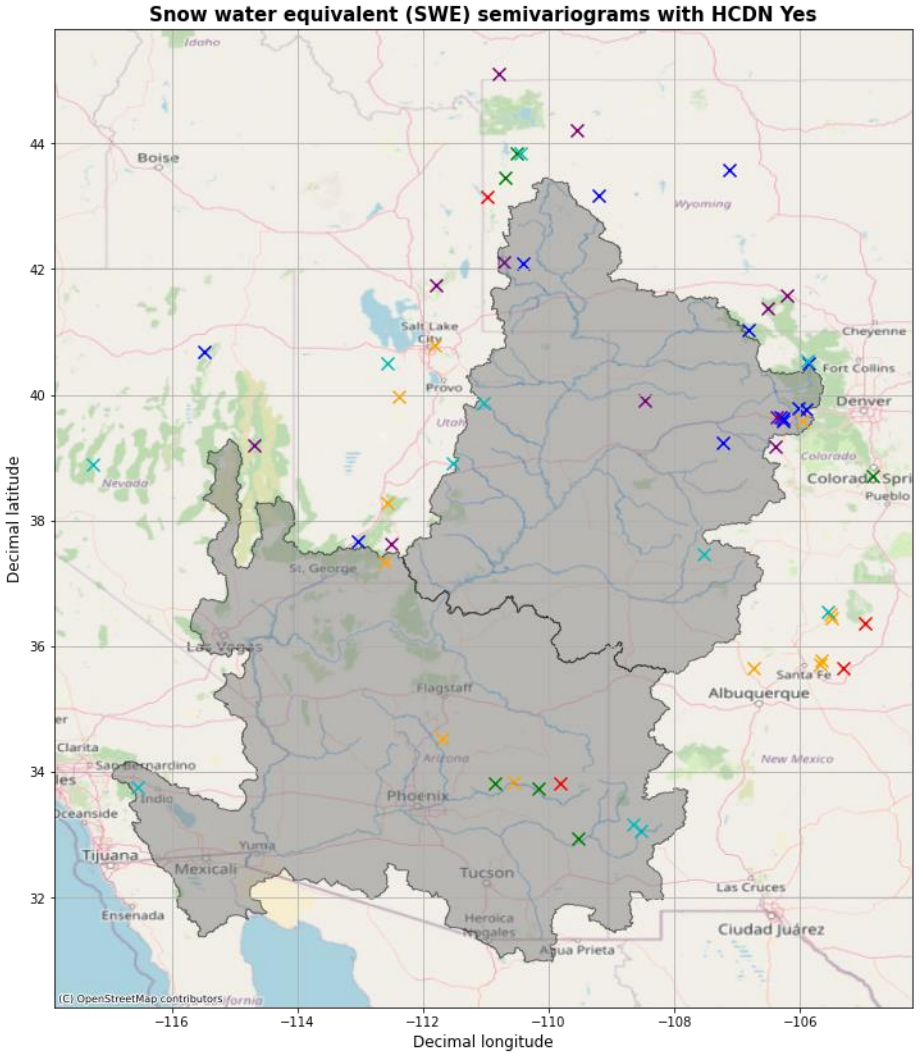

Medium depth (10-40cm) soil moisture semivariograms with HCDN Yes
Cluster
1
2
3
4
5
6
N value
6
13
3
11
7
17
Decimal latitude
Decimal longitude
Boise
Salt Lake City
Provo
Cheyenne
Fort Collins
Denver
Colorado Springs
Pueblo
St. George
Las Vegas
Flagstaff
Albuquerque
Santa Fe
Phoenix
Tucson
Las Cruces
Ciudad Juárez
Tijuana
Mexicali
Ensenada
Yuma
Agua Prieta
(C) OpenStreetMap contributors

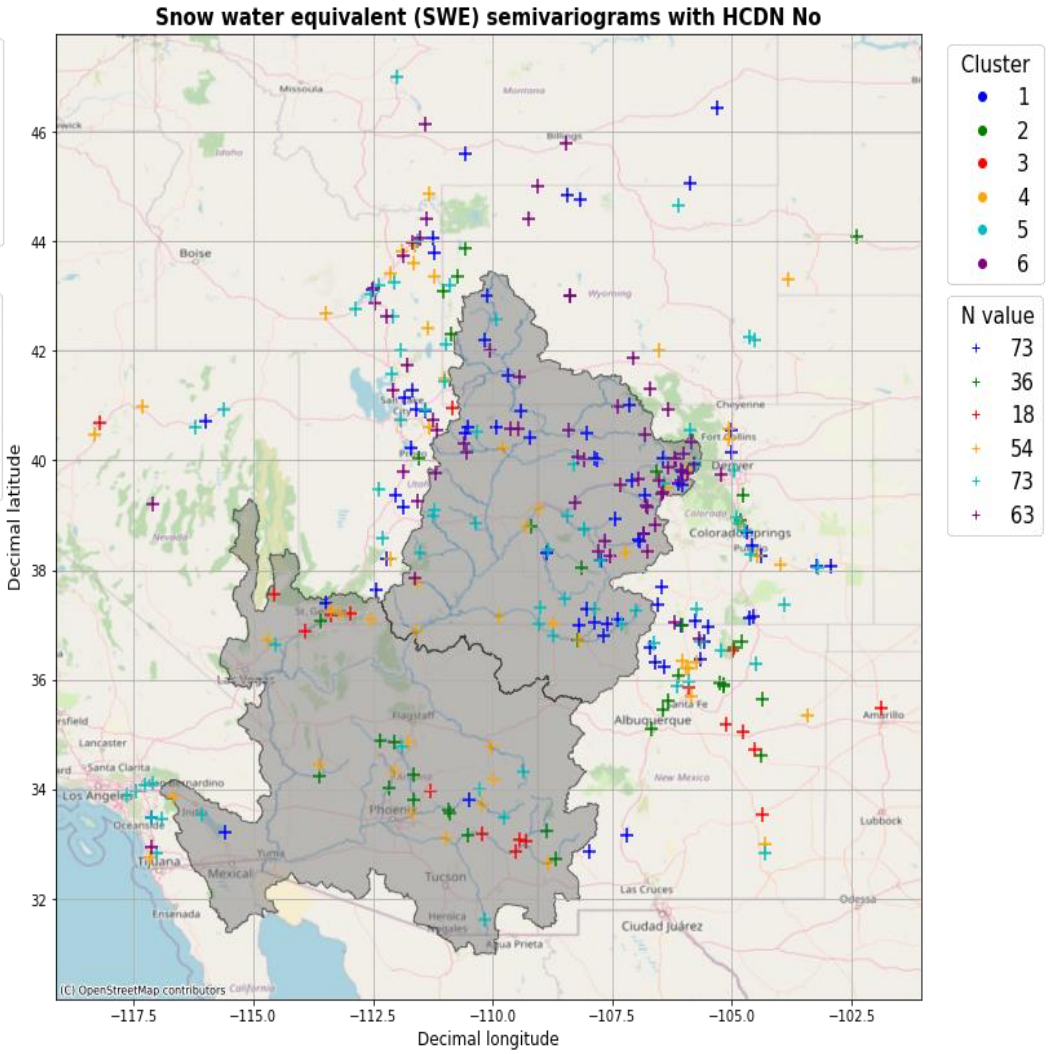

Medium depth (10-40cm) soil moisture semivariograms with HCDN No
Cluster
1
2
3
4
5
6
N value
42
53
8
81
43
90
Decimal latitude
Decimal longitude
Boise
Missoula
Billings
Cheyenne
Fort Collins
Denver
Colorado Springs
Albuquerque
Santa Fe
Las Vegas
Flagstaff
Tucson
Las Cruces
Ciudad Juárez
Phoenix
Lubbock
Odessa
Amarillo
(C) OpenStreetMap contributors

Snow water equivalent (SWE) semivariograms






Snow water equivalent (SWE) semivariograms with HCDN Yes
Cluster
1
2
3
4
5
6
N value
13
6
4
12
11
11
Decimal latitude
Decimal longitude
Boise
Salt Lake City
Provo
Cheyenne
Fort Collins
Denver
Colorado Springs
Pueblo
St. George
Las Vegas
Flagstaff
Albuquerque
Santa Fe
Phoenix
Tucson
Las Cruces
Ciudad Juárez
Tijuana
Mexicali
Ensenada
Yuma
Agua Prieta
(C) OpenStreetMap contributors


Snow water equivalent (SWE) semivariograms with HCDN No
Cluster
1
2
3
4
5
6
N value
73
36
18
54
73
63
Decimal latitude
Decimal longitude
Boise
Missoula
Cheyenne
Fort Collins
Denver
Colorado Springs
Albuquerque
Santa Fe
Las Vegas
Flagstaff
Phoenix
Tucson
Las Cruces
Ciudad Juárez
Tijuana
Mexicali
Ensenada
Agua Prieta
Los Angeles
Lubbock
Odessa
(C) OpenStreetMap contributors

Discussion

In our project, we clustered semivariograms of 5 variables of interest during drought events for each of the 425 USGS stream gauges. We used a k-means clustering algorithm and specified the number of clusters by eye, however, there are metrics to evaluate the quality of clustering and to determine the optimal number of clusters. For the purpose of this project the number of clusters for each variable were determined by eye, therefore, for some variables it may seem like four groups is too many due to the similarity between clusters.

Another issue that could be revisited in future work is the clustering process. We used a fixed random seed for the k-means clustering and this gave us the same results on each model run for a specific variable.

Separating unregulated gauges (gauges with HCDN yes value) is another future work that may give us more insights and understanding of the clustered areas. Also assessing longer droughts (e.g., longer than 100 days) and shorter droughts (e.g., shorter than 50 days) might be informative.