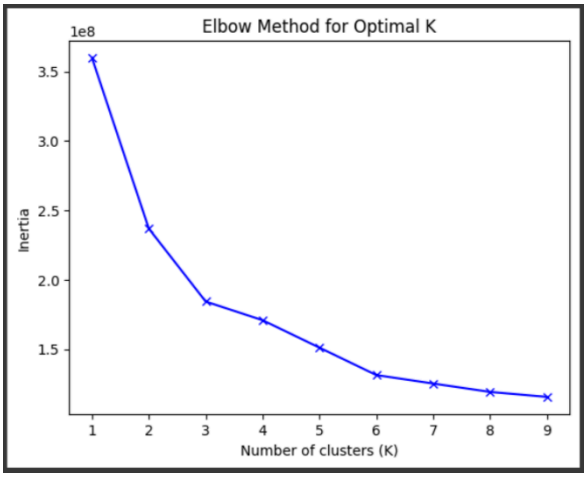
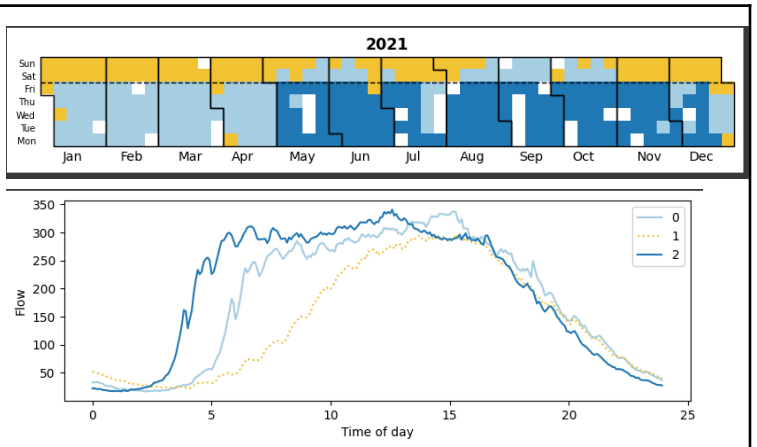
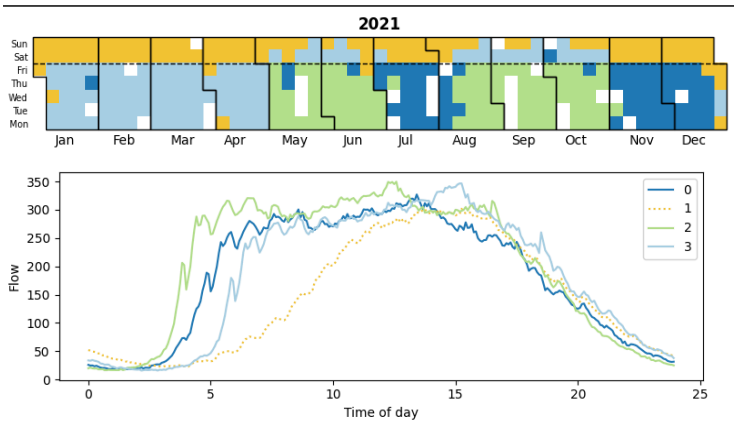
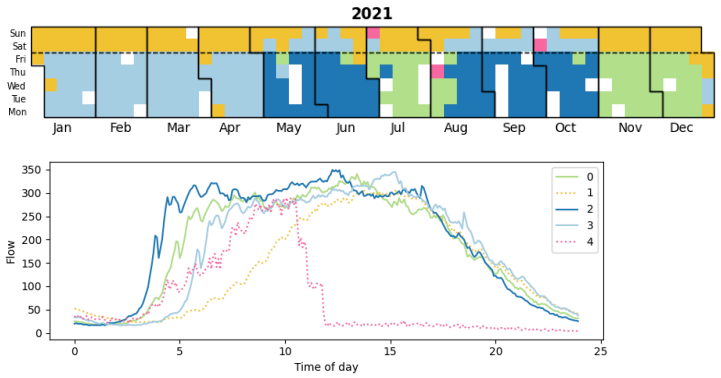
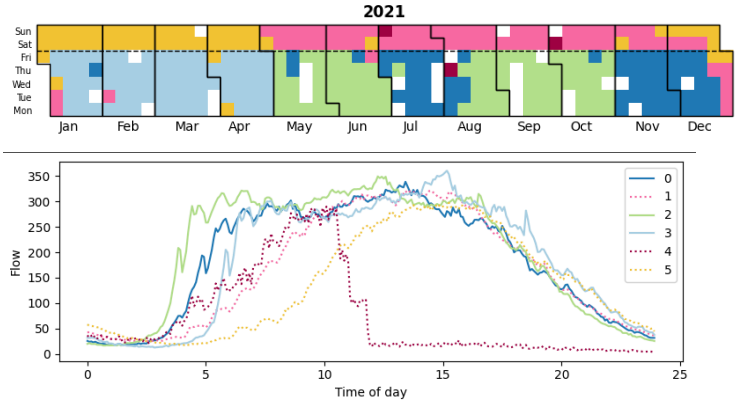
Exercise 5: Clustering

K-means

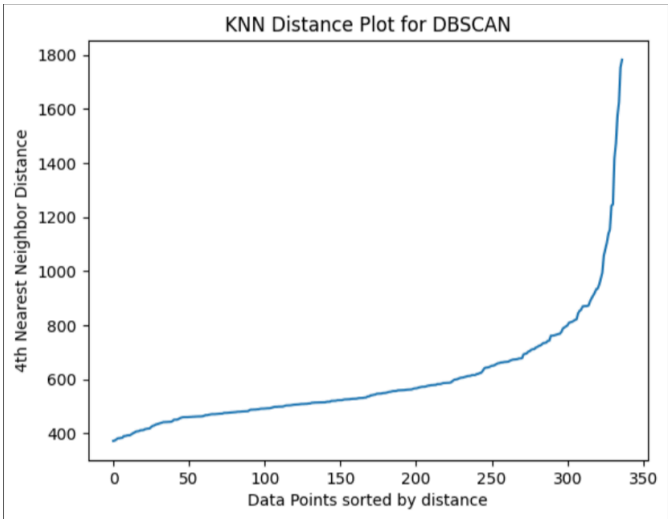
First, the Elbow method was applied to get a direction of how many clusters are optimal for the dataset. It seems by this method that 3 clusters is the optimal number.



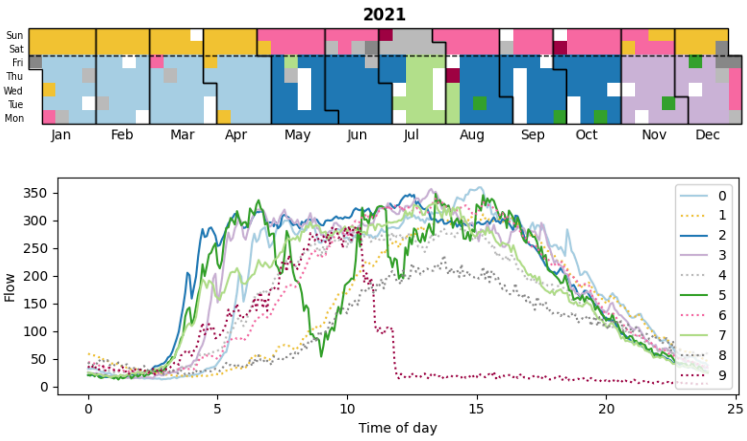
3 clusters Silhouette Score: 0.2692406087798076 Davies-Bouldin Score: 1.3587890766043185 Calinski-Harabasz Score: 159.13421302277044 Prediction accuracy MAE: 31.39955930214274 Prediction accuracy MAPE: 0.40940719318368973	
4 clusters Silhouette Score: 0.22986034355640078 Davies-Bouldin Score: 1.7896752477080542 Calinski-Harabasz Score: 122.80161679824757 Prediction accuracy MAE: 28.907917690806787 Prediction accuracy MAPE: 0.38827596645343576	

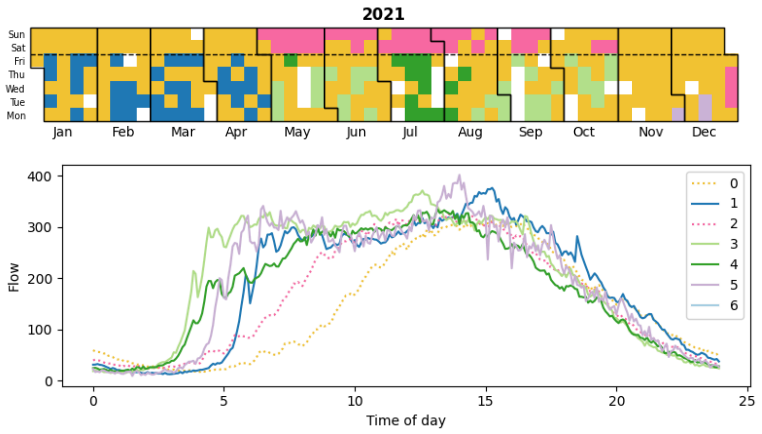
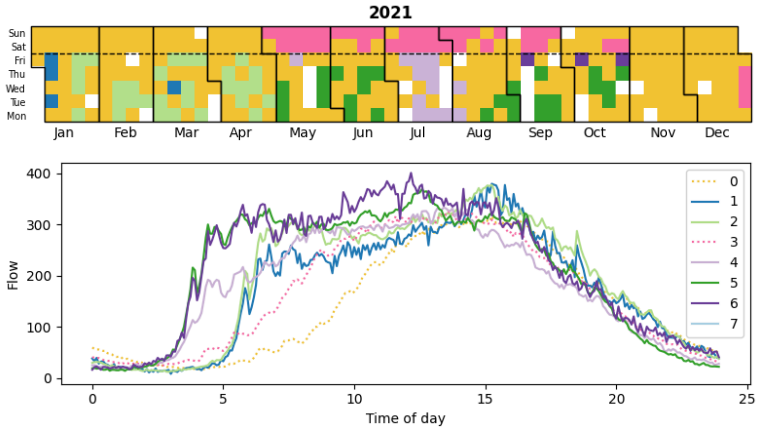
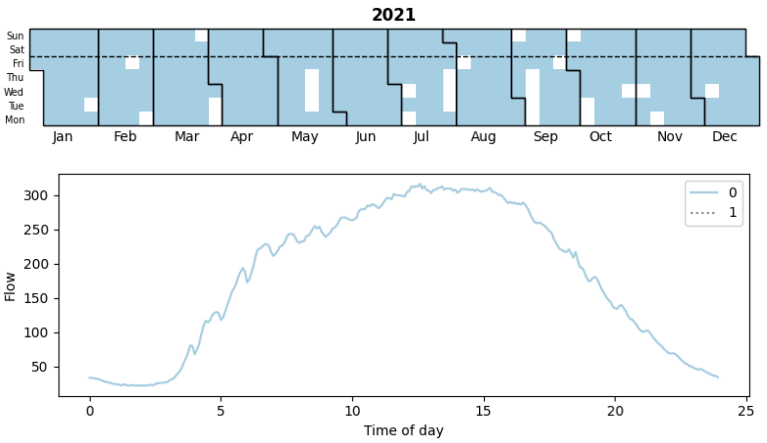
5 clusters Silhouette Score: 0.22969753927464953 Davies-Bouldin Score: 1.5158295908096302 Calinski-Harabasz Score: 114.69473802964785 Prediction accuracy MAE: 27.619317805790065 Prediction accuracy MAPE: 0.2773115163252373	
6 clusters Silhouette Score: 0.23960093549116207 Davies-Bouldin Score: 1.4549397566884494 Calinski-Harabasz Score: 115.18521464607787 Prediction accuracy MAE: 25.954519894405593 Prediction accuracy MAPE: 0.25998120578166506	

DBSCAN



The K-nearest-neighbor plot suggests that the optimal eps lies somewhere between 700 and 900 when min_samples = 4. After plotting the KNN distance plot for min_samples = 3 as well, it is found that epsilon should be within the same range (700-900). Grid search is used to find the optimal epsilon within the range, where the Silhouette score is used to evaluate the hyper parameters. The findings were that the best parameters were eps = 1250 & min_samples = 4. Interestingly, this is not what was suggested by the KNN plot. Unfortunately these hyperparameters only yielded 1 cluster, which is not the goal of this search. It seems from these experiments and methodically searching that DBSCAN might not be suited for this dataset.

eps = 500 min_samples = 2 Silhouette Score: 0.22969753927464953 Davies-Bouldin Score: 1.5158295908096302 Calinski-Harabasz Score: 114.69473802964785	

eps = 500 min_samples = 3 Silhouette Score: -0.02772100330028316 Davies-Bouldin Score: 2.3795720914163345 Calinski-Harabasz Score: 35.327906331751436	
eps = 500 min_samples = 4 Silhouette Score: -0.08626688979370234 Davies-Bouldin Score: 2.167100015154176 Calinski-Harabasz Score: 29.590313849579786	
eps = 1250 min_samples = 4 Silhouette Score: 0.4593859647121622 Davies-Bouldin Score: 1.3491896589311563 Calinski-Harabasz Score: 15.698219086119039	

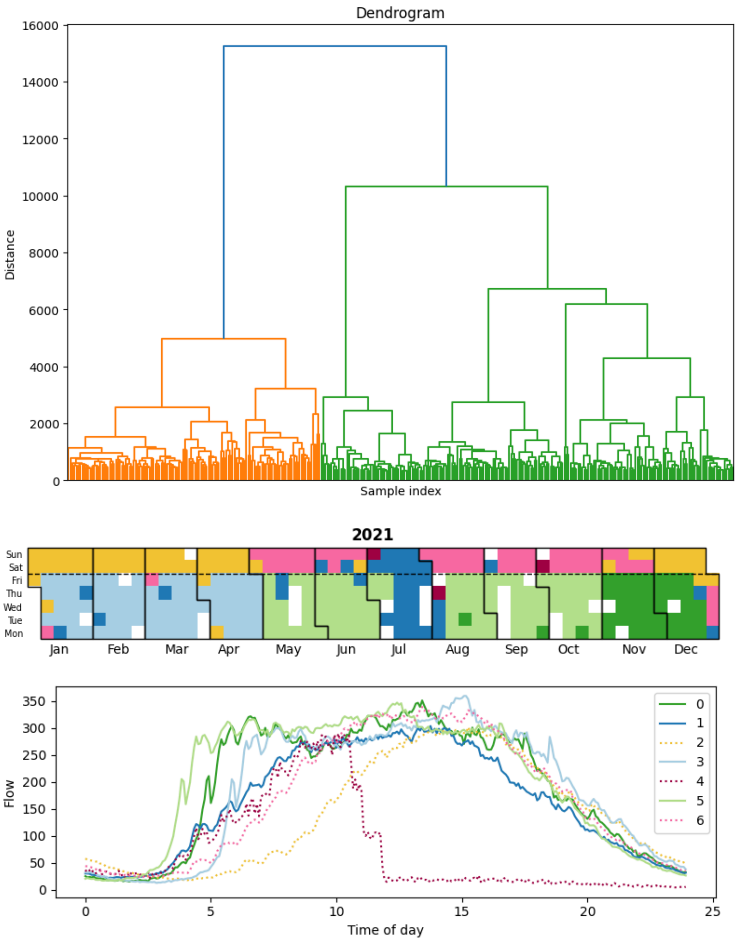
Agglomerative clustering

First, dendrograms were plotted to visualize the hierarchical structure of the data for each linkage parameter and possibly be a guide to how to tune the hyperparameters of the clustering method.

Dendrogram: linkage = “ward” → n_clusters = 7

Silhouette Score: 0.24315720694085063
Davies-Bouldin Score: 1.5136059986976242
Calinski-Harabasz Score: 100.74968399031134

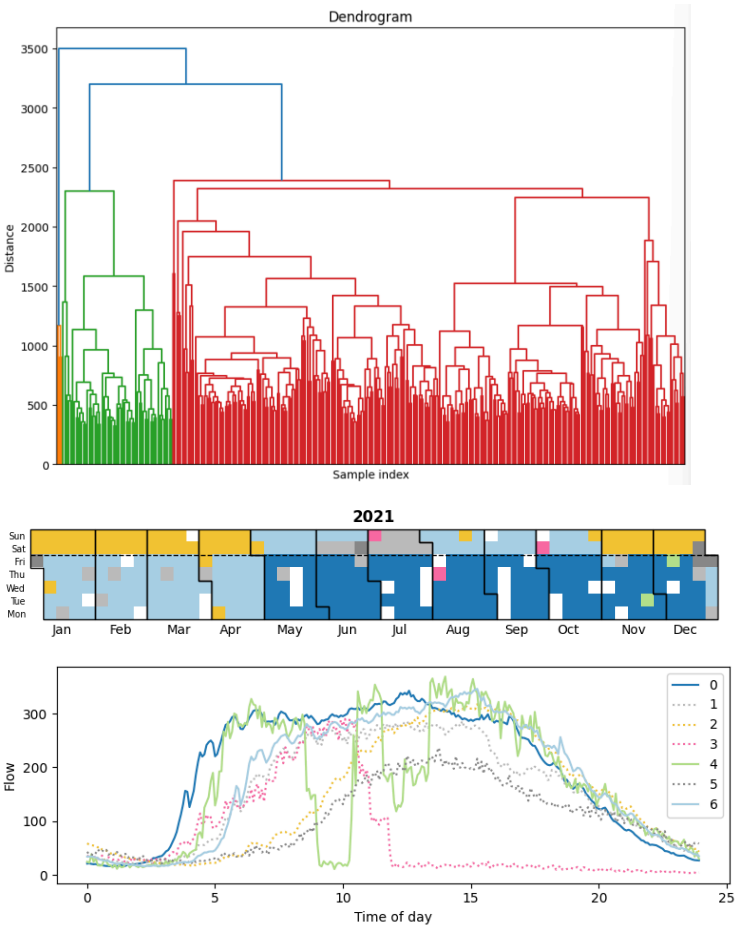
Prediction accuracy MAE: 24.767895406358022
Prediction accuracy MAPE: 0.25017379103102816



Dendrogram: linkage = “complete” → n_clusters = 7

Silhouette Score: 0.2715320814283351
Davies-Bouldin Score: 1.0652968726117817
Kalinski-Harabasz Score: 85.69304002012917

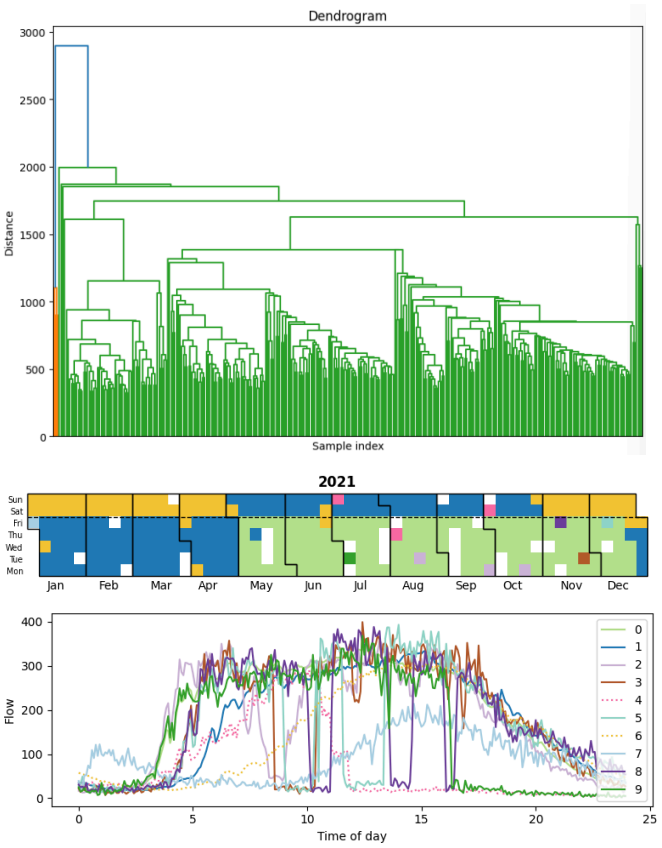
Prediction accuracy MAE: 29.614943170022855
Prediction accuracy MAPE: 0.28774933269845826



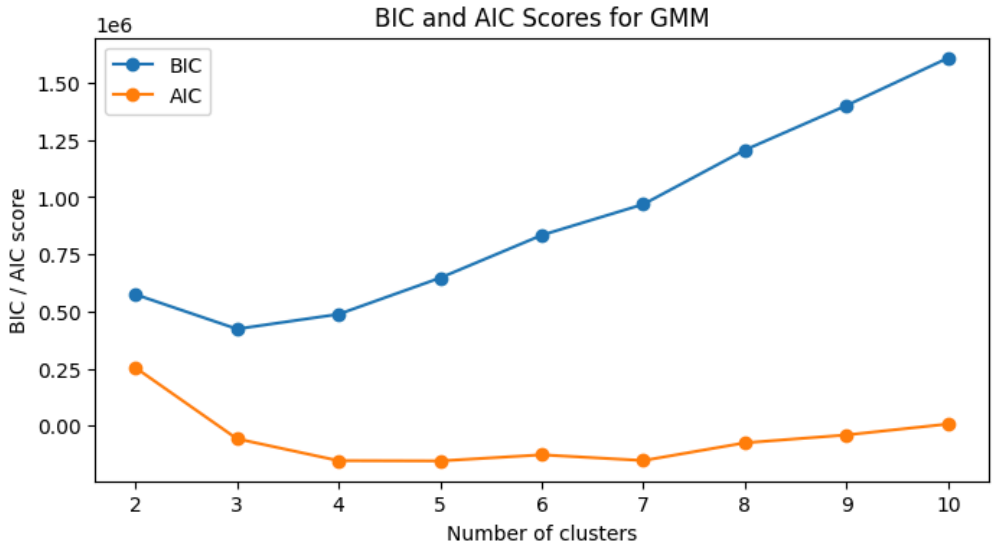
Dendrogram: linkage = “average” → n_clusters = 10

Silhouette Score: 0.26804898521189274
Davies-Bouldin Score: 0.7634697993367069
Calinski-Harabasz Score: 51.205896114225624

Prediction accuracy MAE: 29.785638482102204
Prediction accuracy MAPE: 0.2945594157657579



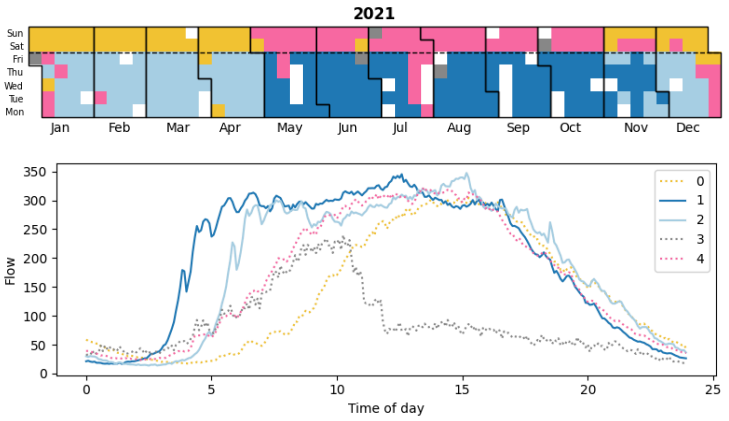
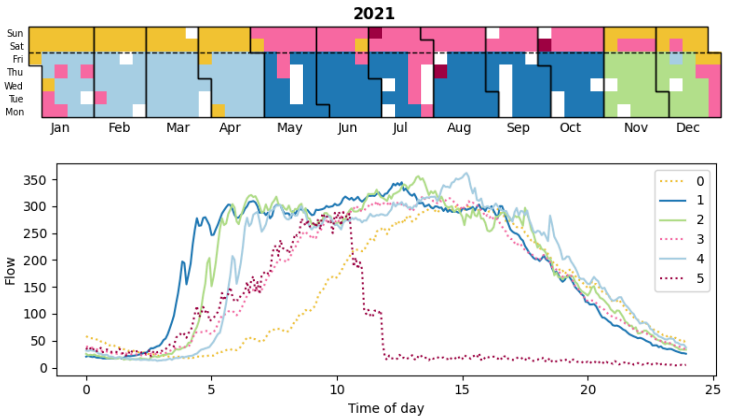
Gaussian Mixture

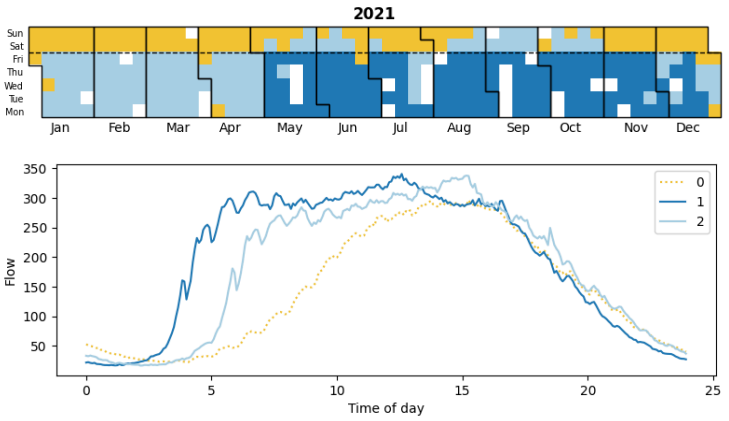
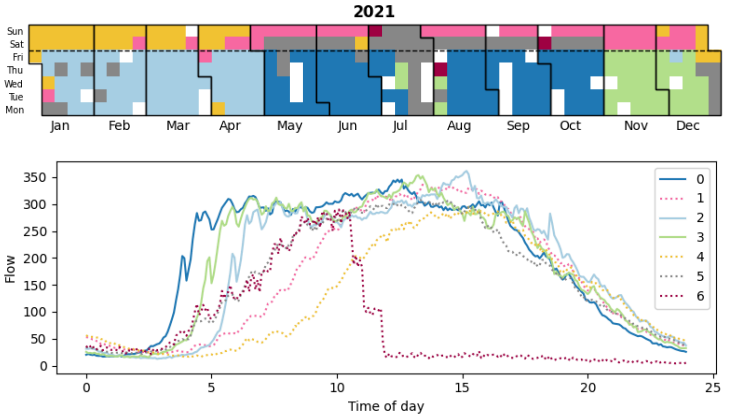


Akaike information criterion and Bayesian information criterion are two methods for balancing model fit and number of parameters, which prevents overfitting. They are based on the likelihood function. Here they are used to figure out a good value for number of clusters. We can see that 3 for BIC, and 4 or 7 for AIC are good values to try.

$BIC = -2 \cdot \log\text{-likelihood} + k \cdot \log(N)$

$AIC = 2 \cdot k - 2 \cdot \log\text{-likelihood}$

5 clusters Silhouette Score: 0.25590833834410825 Davies-Bouldin Score: 1.317908136249299 Calinski-Harabasz Score: 121.9237459568086 Prediction accuracy MAE: 27.611920664472656 Prediction accuracy MAPE: 0.28904106642841115	
4 clusters Silhouette Score: 0.24413737962687762 Davies-Bouldin Score: 1.4038663284411552 Calinski-Harabasz Score: 113.41804850233356 Prediction accuracy MAE: 25.8730825331614 Prediction accuracy MAPE: 0.24737840184762833	

3 clusters Silhouette Score: 0.2694490844214001 Davies-Bouldin Score: 1.3594285802342858 Calinski-Harabasz Score: 159.12748390756323 Prediction accuracy MAE: 31.421506957199632 Prediction accuracy MAPE: 0.4094173399007	
7 clusters Silhouette Score: 0.2141650065816657 Davies-Bouldin Score: 1.6046060601975143 Calinski-Harabasz Score: 102.13536855296255 Prediction accuracy MAE: 25.702894605737736 Prediction accuracy MAPE: 0.2729065937807956	

Silhouette score:
range -1 to 1.
-1 is incorrect clustering
0 is overlapping clusters, or not well separated
1 perfect clustering

Davies-Bouldin score:
Measures ratio of the distance between clusters to the size of the clusters. How far the clusters are from each other and how compact they are.
0 is a perfect score.
Sensitive to cluster shape

Calinski score
ratio of dispersion between clusters to the dispersion within clusters
Higher values are better. There is no upper limit
Good forclusters with a clear structure and separation between clusters