EOSC 410/510 Assignment 3:

Problem 1 [8 points]:
You are given two timeseries x1 and x2 (*data.mat* or *data.csv*), each one containing 310 points in time. For each timeseries, you'd like to investigate how the characteristic temporal patterns change in time, so you decide to use the singular spectrum analysis (applied on each timeseries separately) with a total lag of **L= 50** days.

Tasks:
1) Plot the timeseries and create the lagged matrix for each of the timeseries. Show (in symbolic matrix form) how your lagged matrix looks like for x1. *[1 point for the plot and 1 point for correct lagged matrix for x1]*

2) Perform SSA. Plot the eigenvectors and PCs of the most important modes (decide yourself how many modes are important) for x1 and x2. Hint: in SSA we are interested in the pairs of modes. *[1 point for each plot: eigenvectors and PCs for x1; eigenvectors and PC2 for x2 -> total 4 points]*

3) How much variance is carried by the dominant signals (signals of different frequencies) in x1 and how much in x2? Note that in SSA, a signal of given frequency is usually captured by two modes. *[1 point for the correct answer for x1; 1 point for the correct answer for x2]*

Problem 2 [17 points]:
You are given a data (*data_problem2.mat* or *data_problem2.csv*) that contains one year of normalized daily streamflow from 194 rivers in Alberta, Canada (i.e. there are 194 stations, each with 365 days of normalized streamflow). The locations of each station are given by a latitude/longitude coordinate pair in *stationLon.mat* and *stationLat.mat* (or *stationLon.csv* and *stationLat.csv*). ABlon.csv and ABlat.csv give coordinates of the Alberta border for plotting (e.g. plt.plot(lon,lat) or figure; plot(lon, lat) will plot the border).
Following the guidelines below, perform two types of clustering to investigate how to cluster these stations across the region on the basis of similarity in their streamflow regimes.

Note: apply PCA on the data first (m=365, n=194) and then perform clustering (hierarchical clustering and SOM) on the first few modes only. Most likely the first 3 modes will be enough to keep. In the final plots, make sure that you reconstruct the data (streamflow) from the clustered PC modes, as was done in the Tutorial example on SST dataset.

Tasks:
1) Perform the hierarchical clustering (if using Matlab, use Ward's method) on the data. Plot the dendrogram. *[1 point for correct dendrogram].*

2) Choose **two** possible options for the optimal number of clusters (k) from the dendrogram and provide some explanation on why you chose those k. [*1 point for reasonable choices of k, 1 point for the explanation*].
For each choice of k:
a) plot the mean streamflow pattern of each cluster [*1 point for each plot = 2 points in total*]
b) plot the clusters on the map of Alberta (lat/lon scatter-plot), i.e. color each station's location (can use a filled circle as a marker) according to the cluster to which it belongs. [*1 point for each plot = 2 points in total*]
Discuss what you think are two key differences between your results for two different choices of k *[2 points for discussion]*

3) Perform clustering using a 3 x 2 SOM, and plot the resulting streamflow patterns as a 3 x 2 SOM. Plot the locations of the stations, coloured according to the cluster (BMU) to which they belong. What is the frequency of each cluster? *[1 point for the patterns plot, 1 point for spatial map of clusters, 1 point for correct frequencies]*

4) Perform clustering using 2 x 2 SOM, and plot the SOM patterns, locations of stations coloured by BMU, and frequency of each cluster as in the case with 3 x 2 SOM. Discuss what you think are two key differences between your results from 3 x 2 SOM and this SOM. *[1 point for the patterns plot, 1 point for spatial map of clusters, 1 point for correct frequencies, 2 points for discussion]*