EOSC 510/410 Assignment 3:

Note: Please do **not** submit your code; only submit your assignment as a PDF with the figures/results/tables embedded inside the document. **Include your name(s) in the document name**, e.g. Assignment1_Anderson.pdf

Please submit the assignment to the TA (Geena Littel): glittel@eoas.ubc.ca

Problem 1:
Following the guidelines below, perform hierarchical clustering of the data from data_problem1.mat (or data_problem1.csv) in the space of first few eigenvectors (decide yourself how many modes to use).
Guidelines:

a) Perform PCA on the data (the data has 57 variables and 672 observations in time) and decide how many modes to keep. *[1 point for correct PCA, 1 point for a reasonable choice of modes to keep].*

b) Perform hierarchical clustering with Ward's method on the data in the PC space of the modes you kept. Plot the dendrogram. *[1 point for correct dendrogram].*

c) Chose three possible options for the optimal number of clusters (k) and plot the results (clustered data in PC space) for those options. *[1 point for the correct choices of k, 1 point for the plots]*

d) For only one of the cluster options above (on choice of k): plot, on the same graph, the mean pattern (57 variables) of each cluster (using the reconstructed data according to the selected number of PC modes). Plot the time-series (672 points) of occurrences of these clusters. *[1 point for the mean patterns plot, 1 point for the time-series plot]*

Problem 2:
Following the guidelines below, perform clustering using self-organizing maps from data_problem2.mat (or data_problem2.csv). The data is made up of normalized seasonal 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).

a) Perform clustering using a 3 x 2 SOM. Plot the 6 SOM patterns. Plot the locations of the stations, coloured according to the cluster to which they belong. What is the frequency of each cluster? *[1 point for correct SOM patterns, 1 point for map of clusters, 1 point for correct frequencies]*

b) Perform clustering a differently sized SOM, and plot the SOM patterns, locations of stations coloured by BMU, and frequency of each cluster as in a). Discuss what you think are two key differences between your results from a) and b). *[1 point for plots, 2 points for discussion]*

c) Calculate quantization error and topographic error for a range of SOM sizes (e.g.: 1x2, 2x2, 2x3, 3x3, 3x4, 4x4, 4x5, 5x5) and discuss what you find. In what circumstance is it more important to minimize quantization error, versus in what circumstance is it more important to minimize topographic error? *[1 point for discussion of QE and TE with map size, 1 point for discussion on circumstances to minimize QE/TE]*

d) Calculate quantization error and topographic error for pairs of SOMs which have the same number of nodes but different map sizes and discuss what you find (e.g.: are QE and TE the same for a 2x3, 3x2 map, and 1x6 map? A 3x4 and 4x3 map? A 4x5 and 5x4 map? A 1x2 and 2x1 map?). *[1 point for identifying if QE/TE are the same for maps with the same number of nodes and different shape, 1 point for discussion]*