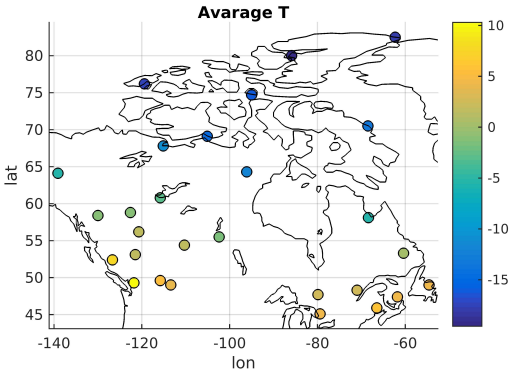
EOSC 410/510 Assignment 2:

Problem 1 [total of 7 points]:
You'd like to analyze the given data (**PCA.mat** or **PCA.csv**) using principal component analysis (PCA). The dataset contains time series of four variables: x1, x2, x3 and x4, each containing 40 observations (points in time). You'd like to investigate whether you can decrease the number of dimensions (variables) in this dataset.

Tasks/Questions:

1) Plot the time series for each variable at the same graph. [*1 point for the plot*]
2) Perform PCA on the data and plot the fraction of variance explained per mode. [*1 point for the plot*]
3) Determine how many modes you want to keep in order to reconstruct the data and provide a rationale for your selection. [*1 point for the answer, and 1 point for the rationale*].
4) Plot the PCs of the significant modes (i.e. those that you decided to keep). Briefly discuss the results (e.g. are the PCs showing some trend, or oscillations or something else) [*1 point for the plot(s) and 1 point for the discussion.*]
5) Plot PC1 vs PC2, i.e. plot the data in the space of the first two eigenvectors [*1 point for the plot*]

Problem 2 [total of 13 points]:



You are given a dataset containing annual temperature timeseries (1951-2000) from 28 stations across Canada. Figure on the right shows average annual temperature over these 50 years for each station (plotted as a heat map). Data is given in **Temp_data.csv** (or **Temp_data.mat** where **T** contains temperature data -> 50 rows and 28 columns, **Lat** is latitude of each station, **Lon** is longitude of each station, and **years** is the array of years). You'd like to analyze characteristic spatial patterns of temperature data across these stations, and evolution of these spatial patterns in time. To do so, you apply PCA on the data.

Tasks/Questions:

1) Plot the fraction of variance explained by each mode. How many modes would you decide to keep if you want these modes to explain in total >90% of variance in the dataset? [*1 point for the plot, 1 point for the answer.*]

2) For each of the first three modes: plots its spatial pattern (as a heat map) and its evolution in time. [2 *points for the plots*]

3) Focus on the results from the first mode only and answer the following:
a) Describe briefly the spatial pattern (e.g. are there any spatial gradients and how are they oriented, for example are they oriented North-South or West-East or something else?) [*1 point for the answer*]
b) Describe briefly the temporal pattern (e.g. is it revealing any long-term trends, decadal fluctuations, or anything else?) [*1 point for the answer*]
c) What years was the spatial pattern most pronounced and what years was the opposite (flipped) spatial pattern most pronounced? [2 *points for the answer*]

4) Reconstruct the data from the first three modes (k=3). Plot the original and reconstructed timeseries for the following two stations: station #6 (lat=45.1, lon=-79.4) and station #26 (lat=64.1, lon=-139.1). Calculate the correlation coefficient between the original and reconstructed timeseries for each of these two stations. Based on the correlations, at which station of the two is the reconstruction better (i.e. better resembling the original data) and why is this so (why is the reconstitution better over one station than other)? [2 points for the plots; 1 point for the correlations, 2 points for the answer]