**ATOC7500 – Application Lab #3**

**Empirical Orthogonal Function (EOF) Analysis**

**in class October 5 and October 7, 2020**

**Note: This application lab requires netcdf4 and cartopy packages.**

**A reminder of the EOF/PCA Analysis Recipe – 5 steps**

**1) Prepare your data for analysis. Examples might include:**

**a) subsetting the global data to a smaller domain**

**b) subtract the mean**

**b) standardizing the data (divide by the standard deviation)**

**d) cosine weighting (Account for the decrease in grid-box area as one approaches the pole (i.e. weight your data by the cosine of latitude)**

**e) detrend the data**

**f) remove the seasonal or diurnal cycle**

**g) remove NaN – EOF analysis does not work with missing data.**

**2) Calculate the EOFs and PCs using one of the two methods discussed in class: a) Eigenanalysis of the covariance matrix**

**b) Singular Value Decomposition (SVD).**

**3) Plot the first 10 eigenvalues (scaled as the percent variance explained) in order of variance explained. Add error bars following North et al. 1982. Describe how you determined the effective degrees of freedom N\*. How many statistically significant EOFs are there?**

**4) Plot EOF patterns and PC timeseries (usually just the first three or so unless you want to look at more).**

**5) Regress the data (unweighted data if applicable) onto standardize values of the 3 leading PCs. In other words, project the standardized principal component onto the original anomaly data X to get the EOF in pjysical units. You should have one regression pattern for each PC – i.e., the EOF pattern associated with a 1 standard deviation anomaly of the PC. *Note: The resulting patterns will be similar to the EOFs but not identical.***

**Notebook #1 – EOF analysis using images of people**

**ATOC7500\_applicationlab3\_eigenfaces.ipynb**

**LEARNING GOALS:**

1) Complete an EOF analysis using Singular Value Decomposition (SVD).

2) Provide a qualitative description of the results. What are the eigenvalues, the eigenvectors, and the principal components? What do you learn from each one about the space-time structure of your underlying dataset?

**DATA and UNDERLYING SCIENCE:**

In this notebook, you apply EOF analysis to a standard database for facial recognition: the At&t database.

<https://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>

*“Our Database of Faces, (formerly 'The ORL Database of Faces'), contains a set of face images taken between April 1992 and April 1994 at the lab. The database was used in the context of a face recognition project carried out in collaboration with the Speech, Vision and Robotics Group of the Cambridge University Engineering Department.*

*There are ten different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement).”*

The goal is to think a bit “out of the box” of Atmospheric and Oceanic Sciences about potential applications for the methods you are learning in this class for other applications.

**Questions to guide your analysis of Notebook #1:**

**1) Execute all code without making any modifications. What do the EOFs (spatial patterns) tell you? What do the PCs tell you? How do you interpret what you are finding?**

The EOFs tell us the different modes of variability faces can take on. The PC tells us the degree to which each individual face “agrees” or “disagrees” with that specific structure. The first two EOFs suggest that the two most important spatial structures of faces are face shape (EOF 1) and hair (EOF 2).

**2) Reconstruct a face. How many EOFs do you need to reconstruct a face from the database? Does it depend on the face that it used?**

Between 170 and 200 EOFs are needed to reconstruct a face from the database. While some faces look better than others within this range, all of the faces sampled look reasonable within this range. The biggest variable controlling the picture quality

**3) Food for thought: The database contains 75% white men (**[**https://www.cl.cam.ac.uk/research/dtg/attarchive/facesataglance.html**](https://www.cl.cam.ac.uk/research/dtg/attarchive/facesataglance.html)**). How do you think this database limitation impacts the utility of the database for subjects who are not white men? What are some parallels that you might draw when analyzing atmospheric and oceanic sciences datasets? *Hint: Think about the limitations of extrapolation beyond the domain where you have data.***

The fact that the database is not diverse means that the eigenfaces are not representative facial modes of the entire population, but instead just representative of an individual sample.

**Notebook #2 – EOF analysis of Observed North Pacific Sea Surface Temperatures**

**ATOC7500\_applicationlab3\_eof\_analysis\_cosineweighting\_cartopy.ipynb**

**LEARNING GOALS:**

1) Complete an EOF analysis using the two methods discussed in class: eigenanalysis of the covariance matrix, Singular Value Decomposition (SVD).

2) Assess the statistical significance of the results, including estimating the effective sample size.

3) Provide a qualitative description of the results. What are the eigenvalue, the eigenvector, and the principal component? What do you learn from each one about the space-time structure of your underlying dataset?

4) Assess influence of data preparation on EOF results. What happens when you remove the seasonal cycle? What happens when you detrend? What happens when you cosine weight by latitude? What happens when you standardize your data (divide by standard deviation)? What happens when you compute anomalies?

**DATA and UNDERLYING SCIENCE:**

In this notebook, you will analyze observed monthly sea surface temperatures from HadISST (http://www.metoffice.gov.uk/hadobs/hadisst/data/download.html). The data are in netcdf format in a file called HadISST\_sst.nc. *Note that this file is ~500 MB so it might take a bit of time to download.* You will subset the data to only look at the North Pacific. Depending on how you prepare your data for analysis – you might expect to see different spatial patterns (eigenvectors) and different time series (principal components). Some things you might look for in your results are the Pacific Decadal Oscillation, “global warming”, the seasonal cycle, …. Depending on your data preparation – your hypothesis for what you should see in your EOF analysis should change. Note: In this dataset - land is NaN, sea ice is -999 – the notebook sets all values over land and sea ice to 0 for the EOF analysis.

**Questions to guide your analysis of Notebook #1:**

**1) Your first time through the notebook – Execute all code without making any modifications. Provide a physical interpretation for at least the first two EOFs and principal components (PC). What do the EOFs (spatial patterns) tell you? What do the PC time series for the EOFs tell you? What do you think of the method for estimating the effective sample size (Nstar)? Can you propose an alternative way to estimate Nstar? Do you get the same results using eigenanalysis and SVD? If you got a different sign do you think that is meaningful?.**

Both EOFs are different modes of SST variability in the North Pacific, broadly related to the PDO and potentially the Aleutian Low. PCs tell us the degree to which a given month follows a given mode of variability or not. Using eigenanylsis vs SVD, we got the same results. In SVD however, the signs of the principle components and the EOFs were switched relative the eigenalaysis results. This however, did not affect the results, as when the PCs and EOFs are recombined, they are exactly the same.

In terms of estimating the effective sample size, there are multiple ways to approach this problem. For one, there is the option of equation to use as you can use either the Leith or Wilks equations, which will give slightly different, but similar results. There is also the consideration of using a uniform autocorrelation for the entire domain, as some areas of the domain might have more red noise than others. One approach could be to calculate the lag one autocorrelation at every grid point, and use the maximum autocorrelation when calculating the effective sample size.

**2) Save a copy of the notebook, rename it. Repeat the analysis but this time do not remove the seasonal cycle. What do you think you will see? Discus your results with your neighbor. How do the EOFs and PC change? Was removing the seasonal cycle from the data useful? What impacts does removing the seasonal cycle have on your analysis?**

Without the seasonal cycle removed, the first EOF clearly shows the seasonal SST, explaining over 80% of total SST variability throughout the domain. As a result, both the EOF and PC magnitudes are larger. Removing the seasonal cycle was thus really useful as it sheds light on additional modes of variability that are less obvious.

**3) Save a copy of the notebook, rename it. Repeat the analysis but this time detrend the data. Discus your results. How do the EOFs and PC change? Was detrending the data useful? What impacts does detrending have on your analysis?**

With the linear trend removed, the first EOF explains more variance than the EOF with the linear trend, reflected in the magnitude of PC1. The negative SST anomaly region also moves northward in the first EOF. The fact that the first EOF is stronger without the long term warming trend makes sense because the linear trend was uniformly influencing all years, thus muddying the interannual variability, and by extension, modes of variability.

**4) Save a copy of the notebook, rename it. Repeat the analysis but this time do not apply the cosine weighting. Discus your results. How do the EOFs and PC change? Was cosine weighting the data useful? What impacts does cosine weighting have on your analysis? What are examples of analyses where cosine weighting would be more/less important to do?**

Without cosine weighting, the PC component and variance explained by each EOF stays roughly the same, but the structure of the EOFs change, with much more weight on the northern parts of the domain. This is because removing the cosine weighting overemphasized variability in the northern parts of the domain.

**5) Save a copy of the notebook, rename it. Repeat the analysis but this time do not standardize the data (i.e., comment out dividing by standard deviation). Discus your results. How do the EOFs and PC change? Was standardizing the data useful? What impacts does standardizing the data have on your analysis?**

There is very little difference between EOFs and PCs of standardized vs non-standardized SST fields. This makes sense because all standardizing the dataset does is make the standard deviation equal to 1. This makes a difference when your field is a combination of datasets with very different variances, but with a field like SSTs, where this isn’t a huge variance, there won’t be a huge difference in EOFs and PCs.