**ATOC5860 – Application Lab #3**

**Empirical Orthogonal Function (EOF) Analysis**

**Note: This application lab requires netcdf4 and cartopy packages. Use the culabenv2022clean environment. See included culabenv2022clean.yml file**

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

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

**a) sub-setting 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 physical 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**

**ATOC5860\_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 are essentially “building blocks” for the faces in the dataset. While they are not exactly facial components (eyes, nose, mouth) they represent spatial patterns that when summed with different magnitude can produce a unique face. The PCs are how much of each of these EOFs are in each face. They are essentially the weight on how much of each EOF to include.

1. **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?**

It seems that a good number of EOFs to reconstruct a face is somewhere between 60—100. This depends on the face. It seems that faces looking straight at the camera are easier to construct with fewer EOFs due to their symmetry. This makes sense as the EOFs that we looked at seem roughly symmetric.

**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.***

Machine learning methods (EOF being one of them) tend to be based on interpolation rather than extrapolation. As a result, predicting for samples not represented in the dataset tends to be less reliable. In atmospheric science, I have seen this issue arise when using machine learning methods trained on historical data before large magnitudes of human caused climate change. This rapid change in carbon emissions and global temperatures may not be captured in past data. Care needs to be taken when using this training for predicition.

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

**ATOC5860\_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). Check that they give the same results (They Should!).

2) Assess the statistical significance of the results, including estimating the effective sample size. (Lots more to think about here for estimating the autocorrelation and N\* in data…)

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 #2:**

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?.**

The spatial patterns are the dominant physical structures in the data that explain the most variance. They are orthogonal by construction. The principal components are the “amount of each EOF” at each timestep.

In our EOF1 data could be the north pacific gyre.

The method for determining effective sample size is better than assuming totally independent. We know that there is spatial structure and thus all samples are not independent. The method of spatial averaging is better and probably fairley conservative giving a small effective sample size (~50).

I am curious about finding some kind of spatial (and possibly temporal) autocorrelation to assess the degree of spatial correlation rather than averaging across it.

If we see a difference between methods with a negative sign, as long as both the EOF and PC’s are both flipped, this does not tell us any new information

1. **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?**

If we don’t detrend, I think that we will see the first EOF dominate with an annual / seasonal cycle present. After running this, we see the first EOF explains ~50% of the variance. Essentially, the seasonal cycle dominates (in magnitude) variations from that seasonal cycle. Additionally, the second EOF does not match the first EOF with the data detrended.

**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?**

We notice that the PC1 for the data that was not detrended increased over time. After detrending our original data, PC1 seemed stationary around 0. This essentially removed this increasing SST behavior. We also notice that the magnitude of the EOF from the detrended data is larger than that of the detrended data. This makes sense as the range of the data which was not detrended is larger due to increasing SST rather than smaller fluctuations around this trend.

**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?**

Visually, the cosine weighting does not seem to have much of an effect on the EOFs in this analysis. We did not notice differences in either the EOF or the PC. This could be a bug in the code, but if cosine weighting had a significant effect, we would expect to see it at extremely high and low latitudes.

1. **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?**

When we did not standardize the data, we do not see as much difference in the PC timeseries, however the magnitude of values in the EOF changes slightly. We see that the overall values of the EOF tend to be lower, but with similar physical structure.