**ATOC7500 – Application Lab #3**

**Empirical Orthogonal Function (EOF) Analysis**

**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**

**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/eigenfaces give information about the base features that explain the most variance in facial features. The PCs tell us how much a given sample looks like a particular EOF structure, so in this case they tell us how similar an original face is to an eigenface/EOF. To reconstruct an original face, the PCs tell us how much of each eigenface/EOF to use to get back to the original face. From the results, we know that the first 2 EOFs/eigenfaces explain the most variance. Interestingly, they are similar in shape but opposite in hue (one is light where another is dark) but this allows these two states to explain the majority of the variance among the faces.

**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?**

Face: 120

Weights: 200

Using this face and this many weights, I think the face still looks a little weird and spooky.

Weights: 300

The face looks better, more filled in, but still not entirely normal.

Weights: 350

I would say this is the minimum value to use to reconstruct this face – you can definitely make out he face but the skin still looks a little blotchy.

Face: 382

Weights: 100

This is not enough EOFs to reconstruct this face; The face looks very smeared and unclear, though you can tell the image is a face.

Weights: 150

This is still not enough.

Weights: 200

Some facial features (eyes) are more clear but still not great.

Weights: 250:

Slightly better.

Weights: 300

Definitely a face but still not great.

Weights: 350

I would again say this is the minimum value need to reconstruct the face.

Weights: 375

The face is still mildly blotchy but we can definitely make out the original face.

Face: 8

Weights: 100

We can tell this is a face but it looks like a very abstract rendering of a face. This could also be due to the way the face is in the frame (the ears are cut off by the edge of the image and then chin and forehead extend all the way to the edges).

Weights: 150

The eyes are more clear.

Weights: 200

There is a lot more clarity on facial features but the general skin area is pretty blotchy.

Weights: 250

The eyes, mouth, nose, and ears are clear but the space in between is blotchy. We can definitely tell that the image is a face.

Weights: 300

The face is looking much better (more like a face). I would say this is the minimum number of EOFs to use to make out this face. Any more additional EOFs would even out the skin and add more detail.

From all of the above, I would say a minimum of 300 – 350 EOFs (out of 400) are needed to reconstruct a face well. It is possible to make out a few with fewer EOFs but this is the number of EOFs where features start to appear and the skin is a little less blotchy. I do think this number slightly depends on the face but more so on the angle and lighting of the image than on the facial features themselves. Also, this number is subjective depending on how much clarity someone wants in the face to consider it reconstructed.

**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.***

I think there are extreme limitations. For starters, a black background would only work well for lighter skin because the contrast allows the EOF analysis to distinguish between a face and the background. For darker skin, this would have obvious issues. Thus, it would make the most sense to have more of a neutral color for the background (perhaps a gray) or have different colored background that better contrast the skin colors. Also, with 75% white men faces, there is likely a strong bias towards predominately male facial features so more EOFs would probably be needed to reconstruct a face with predominately female facial features. Essentially, this database would work very poorly for faces that are not white male faces and thus is not extremely robust.

This concept runs parallel to atmospheric and oceanic sciences datasets in the sense that it will be hard to use EOF analysis on a dataset for one thing if that thing is not well represented in the data that the EOFs are based on. For example, if I am doing EOF analysis on an ocean sea surface temperature dataset that consists of 85% El Nino years and only 15% La Nina years, the EOFs will be better at explaining the variance in the El Nino years than the La Nina years since those years and patterns make up the majority of the dataset that was used to calculate the EOFs. This all goes to say that one should avoid extrapolating beyond the domain where you have data because your EOFs might not be good representations of the data in this domain.

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

**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 original data consists of the sea surface temperature for the Northern Pacific Ocean. When we pre-process the data, we subtract the mean and divide by the standard deviation, so any variations seen in the temperature are not due to a seasonal or diurnal cycle. Thus, these variations are likely related to the dominant circulation patterns found here. For the first EOF, we can see dominant features on the eastern boundary and in the middle of the northern Pacific Ocean that look very similar to the dominant circulation patterns here. In the middle of the northern Pacific, there is the North Pacific Gyre that rotates clockwise, which would cause downwelling and warmer sea surface temperatures. On the easter boundary, there is the California Current moving from north to south, causing upwelling at the coast and colder sea surface temperatures. While these have opposite signs than those in the figure, the relationship between these two circulation patterns is represented fairly well. As for EOF2, we see the most variation in the northern part of the domain, around the same area where we have a Subpolar Gyre. This gyre rotates counterclockwise, leading to upwelling and cooler water in the middle of the gyre. Thus, these first two EOFs likely represent the dominant circulation going on in the northern Pacific Ocean – the North Pacific Gyre and the subpolar Gyre.

The principal components tell us how much of each EOF is needed to reproduce a certain original spatial pattern. The higher the PC for a particular EOF and spatial pattern, the more that EOF explains that particular pattern. The PCs will change in time because they will be representative of whichever spatial pattern is occurring at that time, yet the EOFs will be the same.

The notebook is currently set up to assume all of the data is independent. From the notebook, we see that the lag-1 autocorrelation is 0.886 which is very strong and suggests the data is fairly red and has a lot of memory, thus, it is not super appropriate to assume all of the data is independent and an effective sample size should be used. There are several different options for calculating N\*. For this scenario, it might be best to use Barnes chapter 2 equation 88 (Wilks) since the data is so red. We do get the same results using eigenanalysis and SVD, though they give opposite signs for the first EOF. I do not think the opposite signs are physically meaningful because they still convey the same spatial information and can be easily flipped.

**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?**

When we remove the seasonal cycle, I think the data will be redder and I think the first EOF will explain more of the variance than before because the dominant variance will be from the seasonal cycle. The first EOF now explains 90.58% of the variance which is a lot more than before. Thus, it is useful to remove the seasonal cycle because this lets you look at variability due to other processes beyond the seasonal cycle.

**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?**

When we detrend the data (remove the linear trend), the EOFs and PCs have the same pattern but the detrended patterns are less extreme (lighter colors). I do not think detrending was incredibly useful because the general spatial patterns in the EOFS are the same. However, in the detrended one, the first EOF explains 6.38% more variance than when we do not detrend.

**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?**

When we do not cosine weight the data, the EOFs and PCs look exactly the same. The first EOF also explains the same amount of variance as before. Thus, I do not think cosine weighting the data was extremely useful in this application but would be useful for even higher latitudes where the cells really start to become smaller, and the longitude lines converge more. For this dataset, the cosine weighting does not have a noticeable impact on the analysis. Conversely, for a dataset in the tropics/near the equator (such as this), I do not think cosine weighting is as important.

**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?**

The first 10 eigenvalues versus variance explained give a very similar plot as before but this time the first EOF explains 22.90% of the variance, which is slightly more than before. Otherwise, the spatial patterns for the first two EOFs look very similar to before but slightly more intense (darker colors with the same color bar). Thus, standardizing the data did not make a huge difference on the EOF pattern results. However, I do suspect that in general it is important to standardize the data to ensure we are looking at anomalies in the data and can better interpret the results (i.e., how does the SST change when the first PC increases by one standard deviation).