**ATOC7500 – Application Lab #3**

**Empirical Orthogonal Function (EOF) Analysis**

**Note: This application lab requires netcdf4 and cartopy packages. Also – The 2020 python environment is provided – That one works on my computer. The 2022 python environment does not work (I think because xarray has been downgraded). Please check for updates on the python environment. I am providing these files early…**

**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 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**

**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?**

First, I notice that eigenface1 seemingly has the least amount of variability. It’s light and dark features are typically less pronounced than the other faces. This face also has fairly light hair, leading me to believe most of the participants have strong variability across hair color. Perhaps lots of the variability can be explained by the hair alone for this reason. Eigenface2 seems relatively opposite of eigenface1, which is reasonable as they would explain most of the variance together. Eigenface2 has dark hair and dark eyes in direct contrast to eigenface1.

**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?**

For this, we choose face number 59. Qualitatively, it takes about 80 EOFs to reconstruct the face. Most of the face looked good at lower values although the EOFs hadn’t captured that the subject wasn’t smiling yet. The lack of a smile became more apparent after about 80 EOFs. Although the correct person could probably be identified after roughly 40 EOFs.

Now trying face number 73, I think it takes half the amount of EOFs to get a recognizable face. So yes, it depends on the face being used. Faces that are more similar to the EOF that explains the most variance will likely take fewer EOFs total to reconstruct.

**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.***

Since this database contains mostly white men, it will take fewer EOFs to reconstruct the faces of white men since the features with the most variance between white men (say, eyebrow height for example) will explain more variance than skin color. However, if we try to reconstruct the face of someone who is not a white male, then it will likely take more EOFs to reconstruct the face since skin color wasn’t technically explaining most of the variance between white men.

For instance, if we “train” the EOFs with datasets that always have higher pressure over land and lower pressure over the ocean (indicative of wintertime conditions), it will take more EOFs to reconstruct conditions during the summer. So, it is important to use diverse conditions with lots of meteorological events when creating EOFs in atmospheric and oceanic sciences.

**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).

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:**

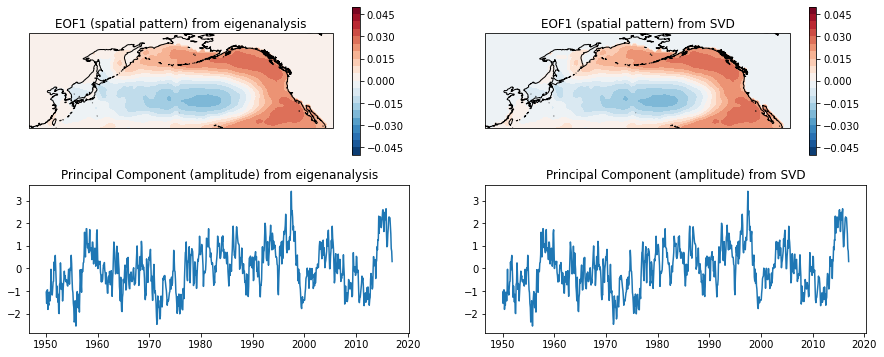
In this notebook, you will analyze observed monthly sea surface temperatures from HadISST (http://www.metoffice.gov.uk/hadobs/hadisst/data/download.html). The data are in netcdf format in a file called HadISST\_sst.nc. *Note that this file is ~500 MB so it might take a bit of time to download.* You will 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?**

Chart, line chart

Description automatically generated



Graphical user interface, application

Description automatically generated

Executing this code without any modifications, the feature that is most dominant in EOF1 is cooler sea surface temperatures in the north Pacific that extend to the western basin and warmer temperatures near eastern basin. This pattern is indicative of a negative phase of Pacific Decadal Oscillation (PDO). One may be able to eyeball the PC analysis and see spikes in positive PCs roughly every 10 years, although a Fourier stransform and power spectrum abalysis would be required to verify this. In EOF2, we see that our methods return the opposite results. In essence, where the eigenanalysis returns positive anomalies in central basin, the SVD method returns negative anomalies in the central basin. However, these EOF structures multiplied by their respective PCs return the same results. EOF2 explains the second most dominant phase of SST in the North Pacific – the North Pacific Gyre Oscillation (NPGO). This method for calculating averages across the entire spatial dimension to create one timeseries that can be autocorrelated. I would probably calculate individual Nstars for each grid cell and take the minimum to ensure that significance is not overcompensated. Otherwise, the Nstar value is an average. Both the eigenanalysis and SVD methods create the same results. For EOF1, the signs are the same. For EOF2 the signs are different. However, flipped signs are accomadated for in the PCs where their signs are also 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?**

**Shape

Description automatically generated**

Graphical user interface

Description automatically generated with low confidence

I predict that when the seasonal cycle is kept, the EOF will show a map that just explains a typical temperature gradient across the North Pacific, with warmer temperatures to the south and cooler temperatures to the north.

When the seasonal cycle is kept, the first EOF explains nearly 90% of the variance by itself. This is because the EOF no longer shows temperature anomalies, but general increases and decreases of temperature across the annual cycle. So, as the entire continent warms up, for example, only the PC coefficient increases since the warming is felt everywhere in the same hemisphere. Removing the seasonal cycle from the data is not very useful, as the PCs no longer explain anomalies. The PCs increase and decrease with seasonal temperature which could already be inferred from the timeseries of temperature. However, the PCs do show the differences in heat capacity between the surface and ocean, where the surface may warm more in the summer months while the ocean remains relatively cool.

**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?**

Chart

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated

When the values are detrended, we see that the first EOF explains much less variance than when the seasonal cycle is included, at only 26%. However, the first EOF now explains more of the variance than it did without detrending the data, which was only 20%. The first EOF still correctly captures the strong temperature maximum extending from the central basin to the eastern basin although there is no longer a bullseye of the opposite temperature anomaly in the northern basin. This means it could be likely that the warming in the Northern Pacific Basin is due to global warming, since it does not show up in detrended data. Furthermore, when detrending, PC1 has reduced its slope from roughly 0.3 to 0. Furthermore, the EOF has changed sign after detrending the data. However, this approach has not significantly changed the results from the original code. The PCs for this version tend to reach higher relative maxima, indicating that fewer EOFs may be required to accurately depict the anomalies. This is further shown by the first eigenvalue containing a larger percent of variance explained. Detrending is useful for the analysis because it removes EOFs that explain, for example, global warming. While global warming is important, it is not the variability we are looking for in this type of analysis.

**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?**

Chart, line chart

Description automatically generated

Graphical user interface

Description automatically generated with low confidence

When the data is no longer weighted by the cosine of the latitude, the percent variance explained by the first eigenvalue reduces to about 21. In this case, we see that neither the EOF anomaly or PC timeseries have changed in relation to the original code. Thus, cosine weighting will likely be more important further north in the northern hemisphere or near the south pole. Cosine weighting may also be more important depending on the projection that is used for the reanalysis dataset.

**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?**

**A picture containing chart

Description automatically generated**

**Graphical user interface

Description automatically generated with low confidence**

When we repeat the analysis without standardizing the data, the amount of variance explained by the first eigenvector has increased to about 23%. As before, the first EOF shows a cold tongue in the central Pacific basin that extends towards the eastern basin. The western basin has warmer temperatures. This EOF is similar to those created by detrending the data as well as the original code. This version shows larger maxima for the negative temperature anomalies but smaller maxima for the positive temperature anomalies. The PC timeseries in this case is similar to the original code although the overall maximum has reduced somewhat. The PC timeseries has nearly the same pattern. Although the EOF and PC timeseries are similar to the original code in this case, the percent of explained variance is higher which makes this EOF more useful.