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 most obvious pattern is that all the eigenfaces have patterns that are in the general shape of the average face, they are all oval. Distinctive features such as eyes and mouths are often the lightest or darkest areas. I believe this is because after creating the general shape of the face the position of the eyes, nose, and mouth describe the most variance. Additionally, the first EOFs are much more general features, but if you print more and more EOFs they start to resemble individuals much more, or certain features like just the eyes. Perhaps this is a greater problem since there are only 40

subjects, but with more people the EOFs would represent more types of features instead of individuals.

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?

It takes somewhere between 40-100 to go from barely recognizable to fairly clear. Faces that are more centered, looking straight and less unique (more average facial features, no beards, etc.) become are clearer with fewer EOFs.

3) Food for thought: The database contains 75% white men (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.*

This dataset will not be able to recognize and represent faces of non-white men as well since there is much less data on them in this dataset. Since the majority of EOFs are based on describing variance in the faces of white males, these EOFs will be optimized for that. Additionally, the EOFs won't be able to effectively represent something that wasn't in the training// dataset.

EOFs are effective tools for explaining modes of variance in observed samples; however, there can be several issues with using EOFs. As with everything if there are biases in the dataset, there will be biased results which could miss important modes of variance. EOFs are also an artifact of a dataset and are often not useful for extrapolating to new types of data. This means that EOFs need data that is diverse spatially and temporally to be representative of whatever you are looking at. Exactly what that means will vary depending on the application, but it is worth thinking about.

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?

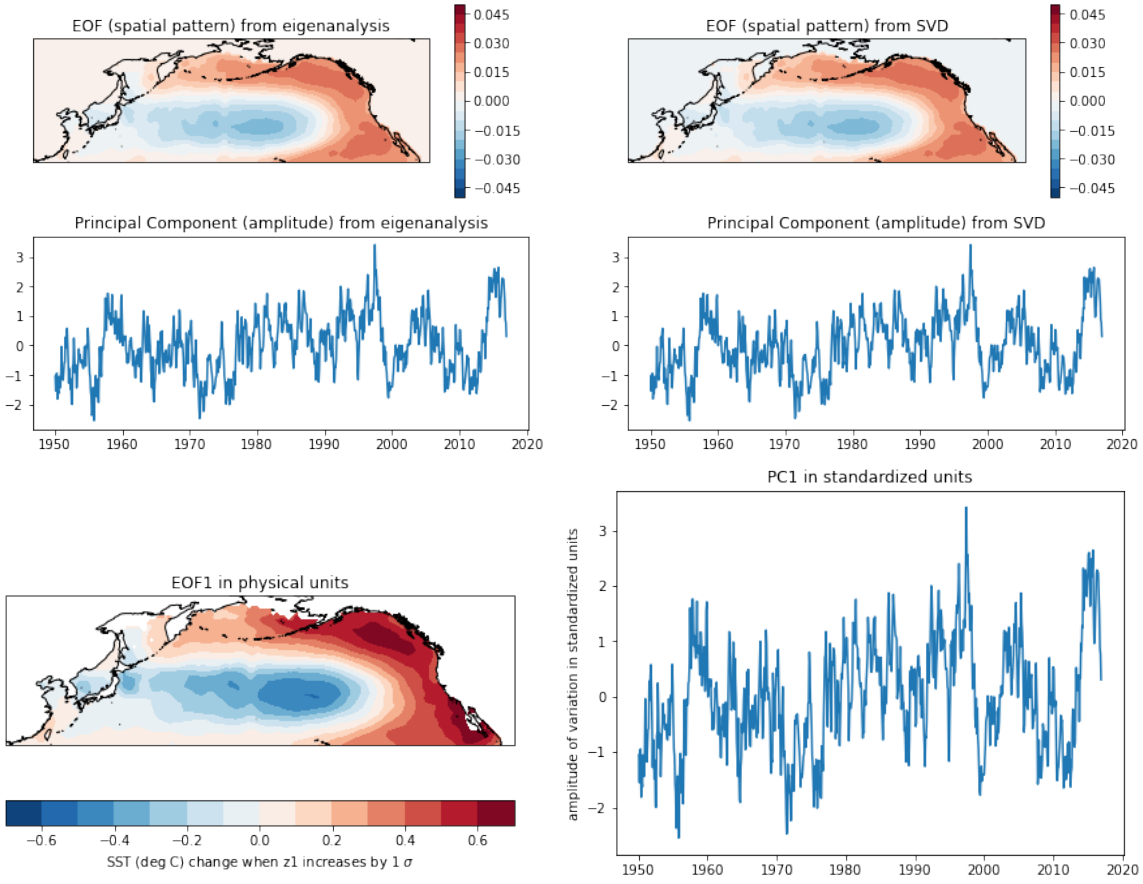
DATA and UNDERLYING SCIENCE:

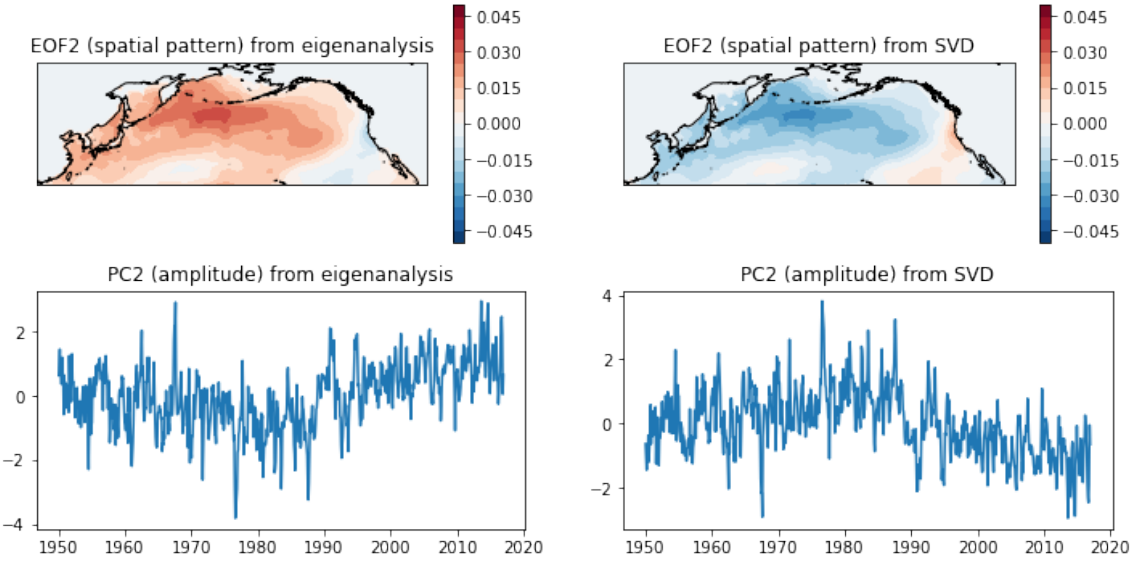
In this notebook, you will analyze observed monthly sea surface temperatures from HadISST (http://www.metoffice.gov.uk/hadobs/hadisst/data/download.html). The data are in netcdf format in a file called HadISST_sst.nc. *Note that this file is ~500 MB so it might take a bit of time to download.* You will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

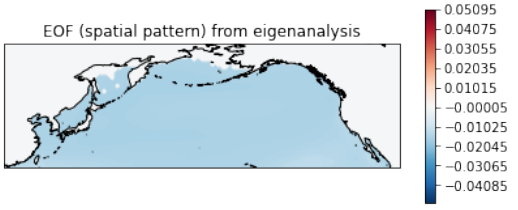
The first EOF is the PDO (Pacific Decadal Oscillation). It describes the often observed patter of the dipoles of the offshore area south of Alaska and west of California compared to the west coast of North America. The second EOF is the possibly NPGO (North Pacific Gyre Oscillation). It has a similar shape to that found in literature. I got the same results for both SVD and eigenanalysis. For EOF two it looks like for some reason both the PC and EOF are inverted for both the SVD, but they will describe the same thing as the PC*EOF will have the same sign. N* at 6% is less than 1/year which seems problematic. Intuitively it makes sense to have at least one independent sample/season which would be 25%.

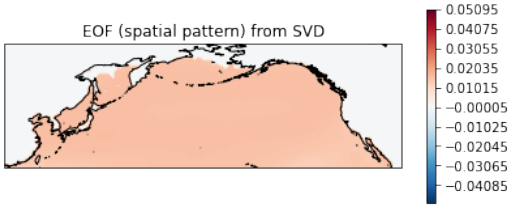


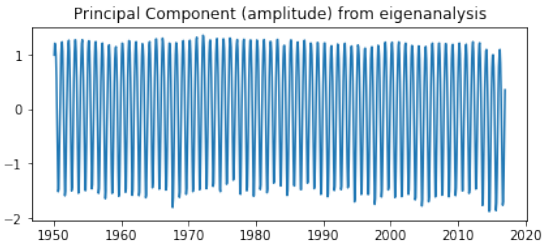


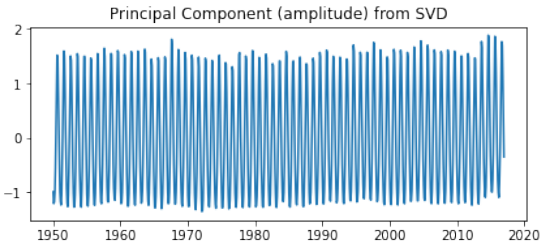
2) <u>Save a copy of the notebook, rename it. Repeat the analysis but this time do not remove the seasonal cycle.</u> What do you think you will see? Discus your results with your neighbor. How do the EOFs and PC change? Was removing the seasonal cycle from the data useful? What impacts does removing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

I thought that EOF 1 will describe a large amount of variance since the seasons are very important for controlling SSTs. That ended up being true as EOF 1 described 90.5% of the variance and its pattern is consistent with seasons since it is just a more or less constant amount over the north Pacific. This is further supported by looking at PC 1, which has and annual cycle that has more or less the same maximum and minimum every year just like temperatures during the seasons. Removing the seasonal cycle in is common practice in data analysis because it allows you to see other interesting patterns better. Everyone know the seasons are important, and EOFs are interesting because they can reveal otherwise hidden patterns.



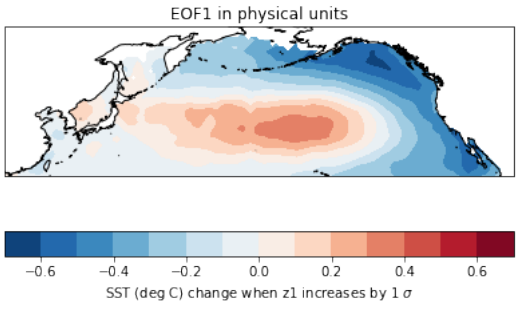


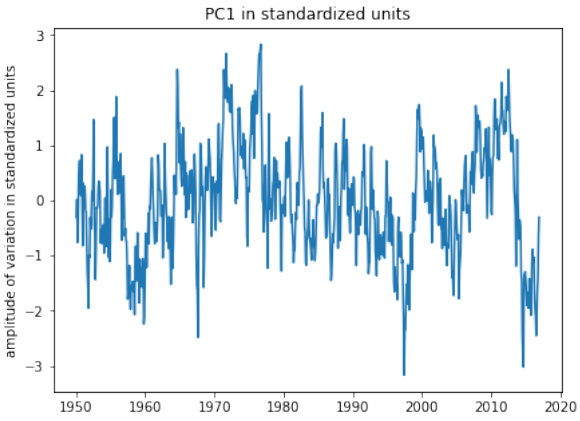




3) <u>Save a copy of the notebook, rename it. Repeat the analysis but this time detrend the data.</u> Discus your results. How do the EOFs and PC change? Was detrending the data useful? What impacts does detrending have on your analysis?

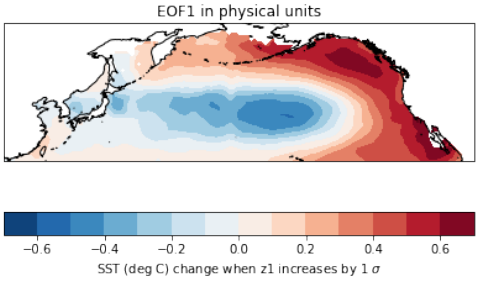
The trend was removed from PC1 which makes sense because the largest source of variance having a trend would likely be the signal/slope that is removed from the detrending process. Additionally, the detrending process does alter the EOFs slightly, but they retained their general shapes. This makes sense as the original dataset is slightly altered, but overall similar. I don't think there is a good way to imagine exactly how the EOFs are effected since the detrending is more directly impacting the time dimension versus the space dimension.

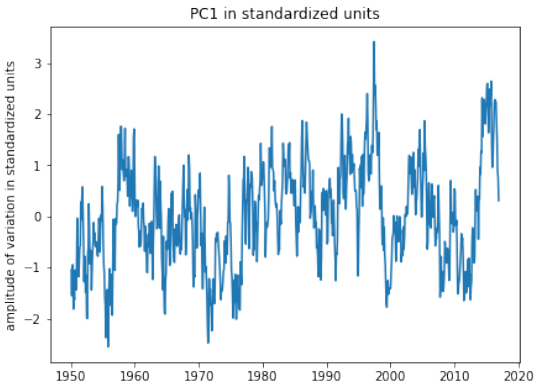




4) <u>Save a copy of the notebook, rename it. Repeat the analysis but this time do not apply the cosine weighting.</u> Discus your results. How do the EOFs and PC change? Was cosine weighting the data useful? What impacts does cosine weighting have on your analysis? What are examples of analyses where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

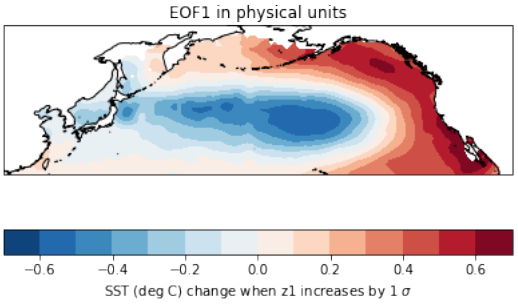
Cosine weighting is done to correct for the spherical shape of the earth compared to treating every grid of latitude and longitude the same like in a basic rectangular projection. I would expect that the result of this is that it will distort the EOFs in a similar way that the differences in the projections distort the landmass. Since this area is relatively far north, I would expect this to have a large impact. The data found however, showed little impact. I believe that some amount of the filtering through time by location removed some of the signal of the cosine weighting, but I'm not sure.

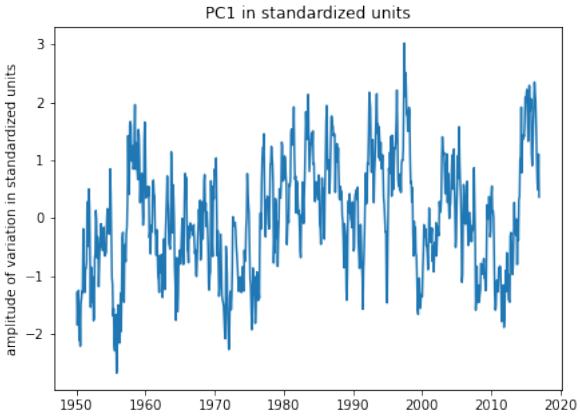




5) <u>Save a copy of the notebook, rename it. Repeat the analysis but this time do not standardize the data (i.e., comment out dividing by standard deviation).</u> Discus your results. How do the EOFs and PC change? Was standardizing the data useful? What impacts does standardizing the data have on your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

I think standardizing the data will create a distribution of similar standardized temperatures everywhere instead of retaining the original distribution of temperatures everywhere. I believe this will make areas of smaller temperature variation equally as important as areas of large temperature variation. Since this is the ocean with relatively minimal variation overall, I don't expect this to be as impactful as if this was done over a landmass. This appears to be true in the sense that there are minimal changes observed, but the "cold center" is overall more defined. This is the opposite of what I hypothesized, however the "warm poles" are muted which is what I hypothesized. I am unsure as to why this is observed.


EOF1 in physical units
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
SST (deg C) change when z1 increases by 1 σ


PC1 in standardized units
amplitude of variation in standardized units
3
2
1
0
−1
−2
1950
1960
1970
1980
1990
2000
2010
2020