Gina Jozef
3/1/22

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.*

Gina Jozef
3/1/22

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 show you which patterns in the photos explain the most variance between faces (Figure 1). For example, the first EOF looks like a blurred image that generally captures the structure of a human face. Since all of the faces have this general outline, this EOF explains a higher amount of variance in the samples than all other EOFS. However the variance explained is still fairly low (~17.5%; Figure 2). The next EOFs appear to highlight certain areas of a human face where there is variance between different subjects, including the hair, eyes, nose, and mouth. While we don't visualize the PCs in this notebook the PCs would tell you how similar each picture looks like each EOF. If we have a PC value of 1, this means the picture looks verry similar to the EOF. If we have a PC value greater than 1, this means the picture looks like the EOF, but with a greater amplitude. If the PC value is low or negative, then the picture is not very similar to the EOF. I interpret these findings to mean that most of the variance between subjects faces is explained by the overall structure of the human face, as well as the primary facial features (eyes, hair, etc.)

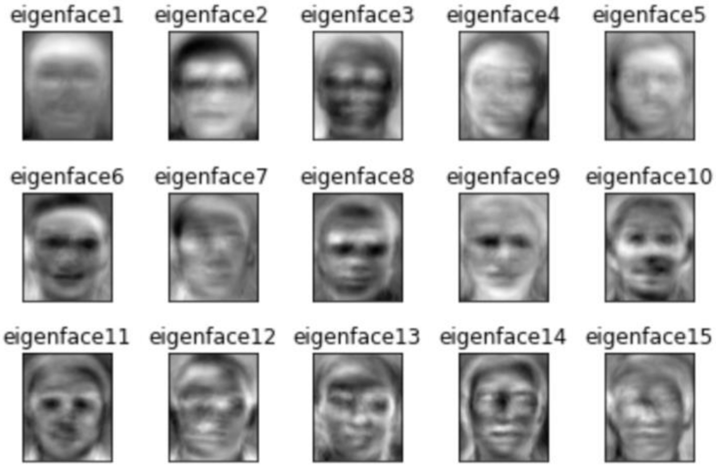

Figure 1: Spatial patterns of the first 15 EOFs (eigenfaces).

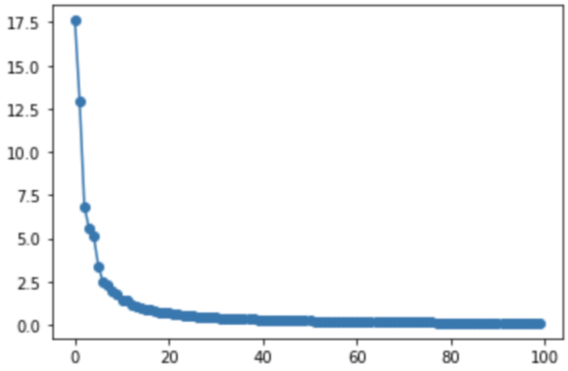

Figure 2: Percent of variance explained (y-axis) by each EOF (x-axis).

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?

Reconstructing face #100:
#100 is a photo of a man whose eyes are closed. The man’s face becomes recognizable when using around 50 EOFs. Below I show the original photo, and the reconstructed photo using the first 50 EOFs. I also show what the face looks like with 40 EOFS, to demonstrate that he is less recognizable, contrasted with what the face looks like with 200 EOFS, which is much more clear.

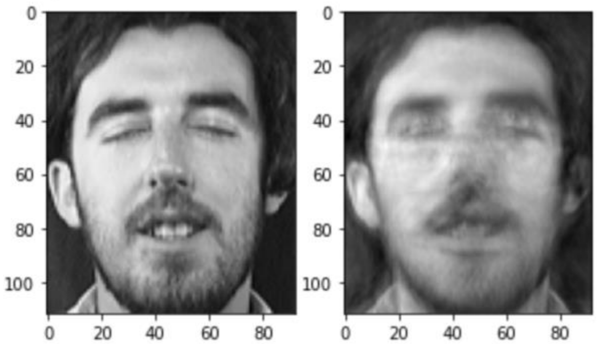

Figure 3: Photo number 100 reconstructed using the first 50 EOFs.

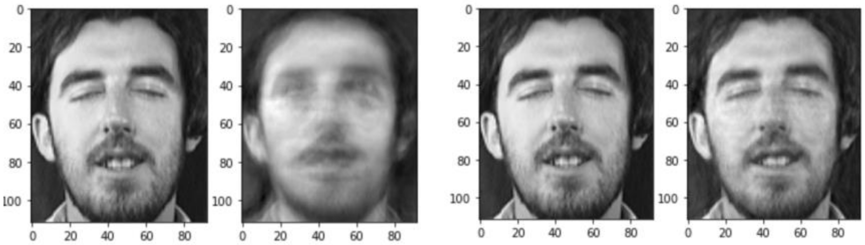

Figure 4: Photo number 100 reconstructed using the first 40 EOFs (left) and the first 200 EOFs (right).

Reconstructing face #130:

#130 is a photo of a man whose eyes are open, but who is wearing glasses. The man's face becomes recognizable when using around 40 EOFs. Below I show the original photo, and the reconstructed photo using the first 40 EOFs. I also show what the face looks like with 30 EOFS, to demonstrate that he is less recognizable, contrasted with what the face looks like with 200 EOFS, which is much more clear. This image is recognizable with fewer EOFs probably because the face is more similar to the average face, and also contains more defining features, such as the glasses.

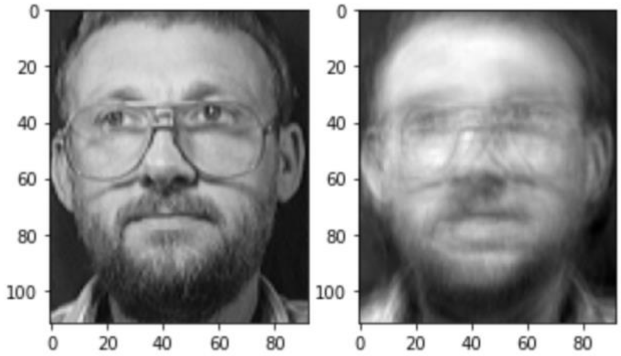

Figure 5: Photo number 130 reconstructed using the first 40 EOFs.

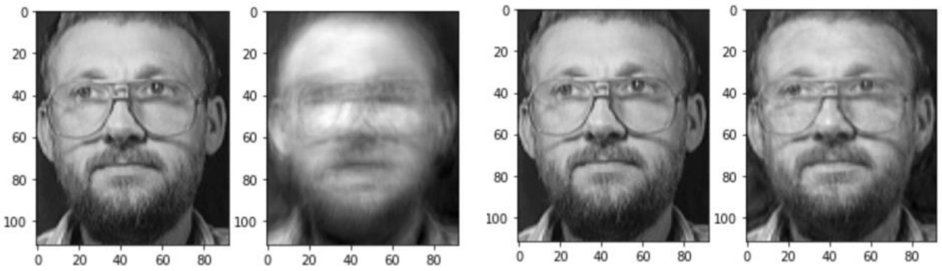

Figure 6: Photo number 130 reconstructed using the first 30 EOFs (left) and the first 200 EOFs (right).

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.*

The fact that the database contains 75% white men likely makes it difficult for the EOFs to be used to recognize faces that are not white men. Therefore, if you wanted to use this EOF analysis for facial recognition in another country in which white men are not the majority, then it would not work very well. An analogy in the atmospheric and oceanic sciences world is if you applied EOF analysis to determine the dominant spatial patterns that explain the most variance in sea surface temperatures in the tropics, and then tried to use those EOFs to learn something about sea surface temperatures in the Arctic. It would not work.

Gina Jozef
3/1/22

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 two EOFs show you the two spatial patterns that explain the most variance in the observed monthly sea surface temperature between 1950 and ~2015. Looking at the first EOF (Figure 1), the spatial pattern shows increasing SST when principal component z1 increases by 1 standard deviation around the edges of the body of water, and decreasing SST in the center of the body of water. The corresponding PC time series is increasing through time, indicating that this pattern is enhanced over the years. The second EOF (Figure 2) shows a spatial pattern with increasing SST when z1 increases by 1 standard deviation in the center and west side of the body of water. This is accompanied by a decreasing PC time series until about 1985, and increasing after that which indicates that this signal is diminished initially and enhanced after 1985.

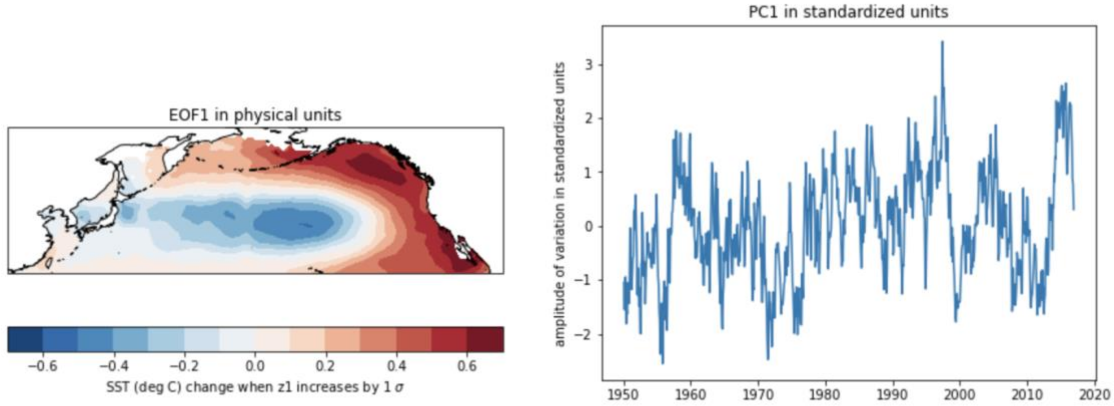

Figure 1: EOF1 pattern associated with 1 standard deviation anomaly of PC1 (left) and corresponding PC1 time series (right).

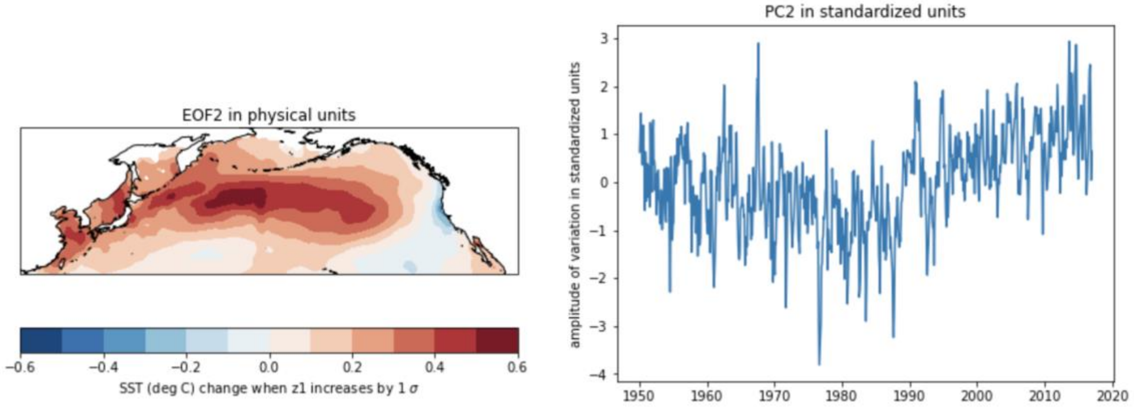

Figure 2: EOF2 pattern associated with 1 standard deviation anomaly of PC2 (left) and corresponding PC2 time series (right).

The method for estimating effective sample size is from the below equation:

$$N^* \cong \frac{1-\rho(\Delta t)}{1+\rho(\Delta t)} N \qquad \text{(Wilks method; from Barnes Ch 2. Eq. 88)}$$

This results in an effective sample size of 49. Since we are using monthly data with a total of 804 times, this indicates an independent sample approximately every 16 months. I would expect an independent sample maybe 2-4 time per year, so I am skeptical of this method for estimating N*. Another method for determining N* that could be used is that from Leith (Journal of Applied Meteorology, 1973)

$$N^* \cong \frac{N\Delta t}{2T_e} = \frac{total\ length\ of\ record}{two\ times\ the\ e-folding\ time\ of\ autocorrelation}$$

(Leith method; from Barnes Ch. 2. Eq. 89)

The results from eigenanalysis and SVD are the same, except the sign is swapped for the second EOF (see Figure 3 below). Both methods, however, result in the same conclusion. Since the egenanalysis shows a general warming signal that increases in magnitude through time based on the PC time series, this is the same thing as the cooling signal shown by the SVD method which decreases in magnitude through time.

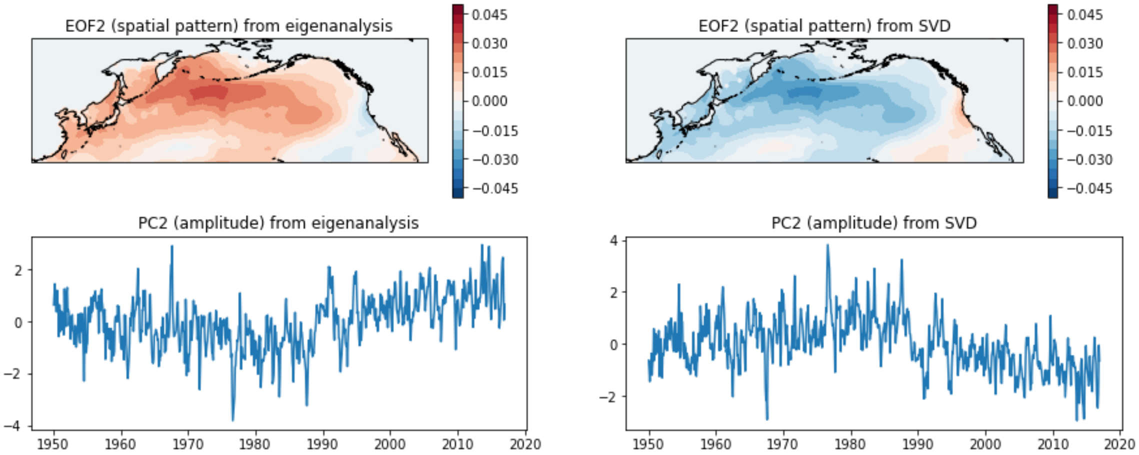

Figure 3: EOF2 (top) and PC2 (bottom) from eigenanalysis (left) and SVD (right).

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?
ATOC5860_applicationlab3_eof_analysis_cosineweighting_cartopy_keepSC.ipynb

Removing the seasonal cycle, I would expect to see a PC time series that oscillates every half year between its maximum and minimum values, and this is what we do see. This oscillation indicates that the spatial pattern observed wavers between being essentially the same as and the opposite of the EOF pattern shown in Figure 4. This oscillation is simply due to the seasonal cycle of SST. Therefore, removing the seasonal cycle is not useful when we want to know the trend in SSTs over time that occurs due to processes other than the seasonal cycle. However, if we want to know how much of the variance is explained by the seasonal cycle, it is useful to leave it in. Removing the seasonal cycle impacts my analysis by revealing that the seasonal cycle explains a high percent of variance, and higher order EOFs which reveal patterns separate from the seasonal cycle only explain a very low percent of the variance.

Gina Jozef
3/1/22

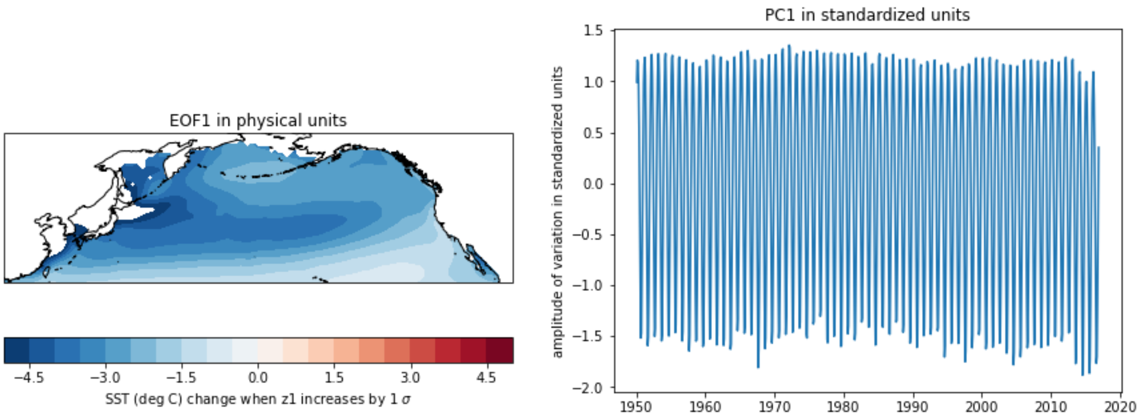

Figure 4: EOF1 pattern associated with 1 standard deviation anomaly of PC1 (left) and corresponding PC1 time series (right), when seasonal cycle is not removed.

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?
ATOC5860_applicationlab3_eof_analysis_cosineweighting_cartopy_detrend.ipynb

When detrending the data, we get a similar first EOF pattern as we do when we don't detrend the data, though the sign is flipped (Figure 5). What is different is the PC time series which on average remains about constant, rather than increasing with time. This results in an analysis that the SST is not changing significantly through time according to this EOF pattern. Since a goal of this activity is to understand how SST is changing through time, and which patterns explain this variance, then detrending the data is not very useful, as it removes this change through time.

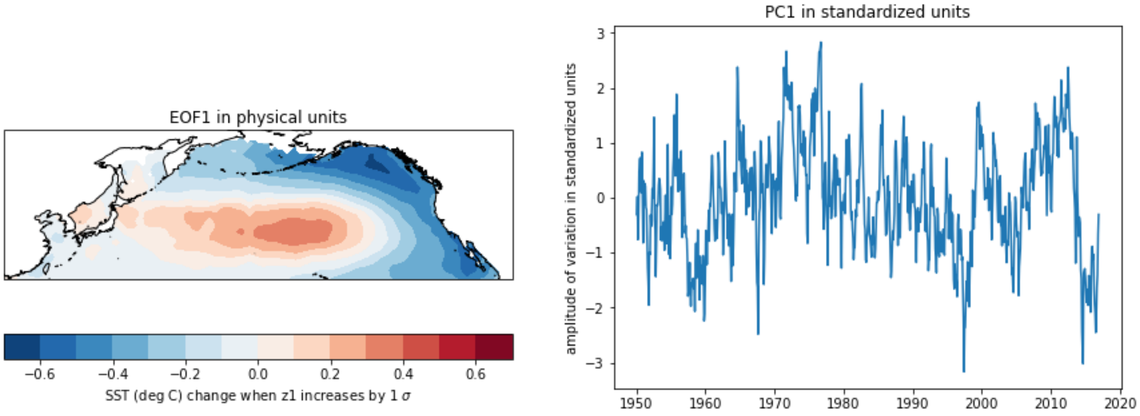

Figure 5: EOF1 pattern associated with 1 standard deviation anomaly of PC1 (left) and corresponding PC1 time series (right), when data are detrended.

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?
ATOC5860_applicationlab3_eof_analysis_cosineweighting_cartopy_nowgt.ipynb

Figure 6 below shows the first EOF pattern when cosine weighting is not applied. Since this spatial pattern looks very similar to when cosine weighting is applied, then cosine weighting does not significantly impact the analysis in this case. This is likely because the sample location is far from the poles, so the distances between lines of longitude are approximately the same throughout the domain. Cosine weighting would probably be much more important when working with a sample area close to the poles.

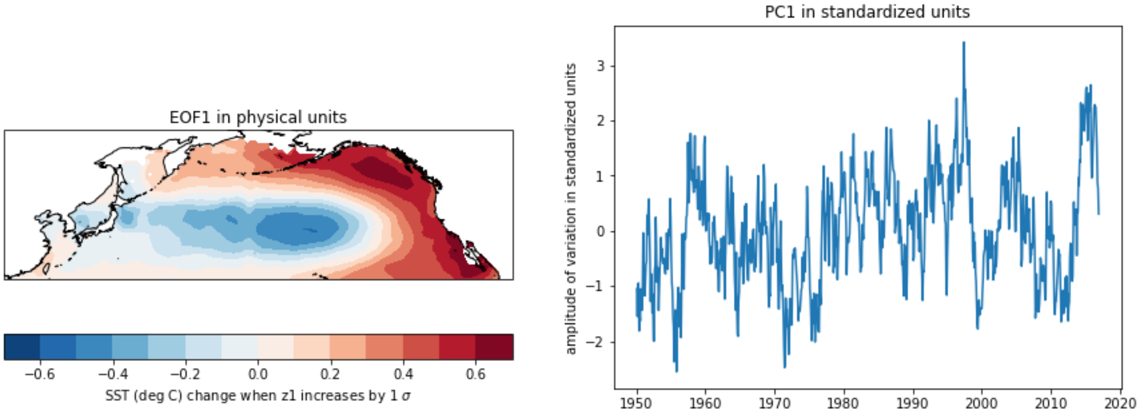

Figure 6: EOF1 pattern associated with 1 standard deviation anomaly of PC1 (left) and corresponding PC1 time series (right), when data are not cosine weighted.

4) 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?
ATOC5860_applicationlab3_eof_analysis_cosineweighting_cartopy_nostand.ipynb

Figure 7 below shows the first EOF pattern when the data are not standardized. Again, this spatial pattern looks very similar to when standardization is applied. The main difference I notice is enhanced negative SST in the center of the body of water and diminished positive SST near the coast on the east of the body of water. Therefore, standardization seems to be important for highlighting the warming signal near the coast. However, in this case, it does not make a major difference in the results.

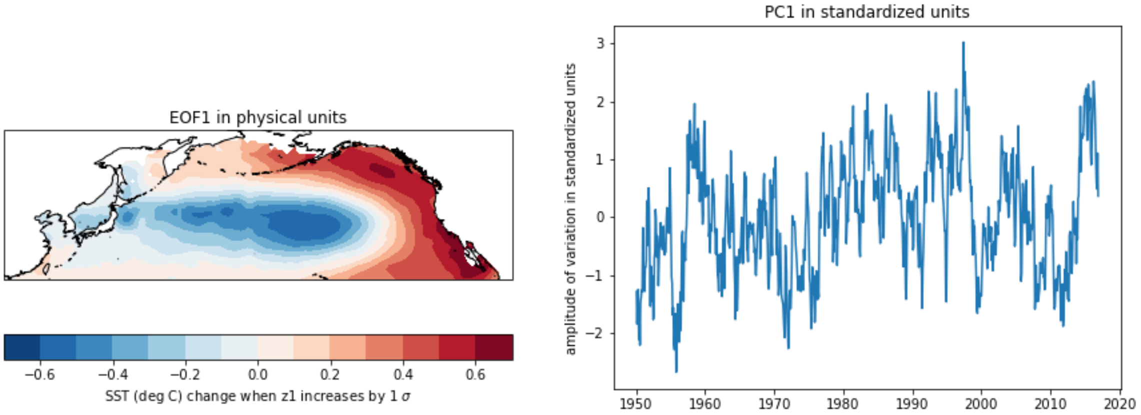

Figure 7: EOF1 pattern associated with 1 standard deviation anomaly of PC1 (left) and corresponding PC1 time series (right), when data are not standardized.