ATOC7500 – Application Lab #3
Empirical Orthogonal Function (EOF) Analysis
in class October 5 and October 7, 2020

Note: This application lab requires netcdf4 and cartopy packages.

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

1) Prepare your data for analysis. Examples might include:
- **a) subsetting 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
ATOC7500_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?

EOFs are the orthogonal spatial (structural) patterns which explain the temporal (sample) variance in our dataset. In this example, EOFs represent the unique facial features (facial hair, eyebrows, etc.) which explain the difference between faces. Principal components tell us the amplitude of each EOF across the sample dimension. So in our case EOFs tell us the unique facial features and the PCs tell us how strong each of those features are across the different photos.

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 a man (face 10) I need about 50 EOFs to recreate the face, whereas for a woman (face 340) I need about 80 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.*

Not including woman or people of color means that features which explain variance in white men are emphasized in the EOFs over features which explain variance in all people!

In my field, many people are now trying to use artificial intelligence to predict snow properties over Antarctica. However, because most in-situ observations which are used to train their models, are collected in summer, near stations, and in pleasant weather, the sample is biased away from more typical weather conditions which lead to different snow properties.

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

ATOC7500_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?.

The first EOF/PC is the Pacific Decadal Oscillation (PDO) and is manifested as a cold or warm patch in the midlatitude (roughly 20-40 degrees north) Pacific Ocean as well as a dipole feature at the British Columbia/South East Alaska Coast. The second EOF is likewise as warm/patch in the North Pacific, but there is not accompanying dipole, except for the western coast of Washington, Oregon and California.

The EOFs and PCs tell us the spatial patterns which explain variance in Pacific SSTs as well as the associated strength of these patterns in time.

Currently we use the number of years as the effective sample size. To improve this, we could calculate the lag1 autocorrelation of SSTs, which would inform us of the memory in our SST time series. We could then calculate a new effective sample size N* = (1-AR1) / (1+AR1).

Indeed, I do get the same results using eigenanlaysis and the SVD. However, the EOFs have different signs. This is not meaningful because we also get opposite sign principal components therefore when we project the PC onto the EOF, the recreated signal amplitudes are identical. I.e. -PC * -EOF = PC * EOF.

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?

The first EOF is now explains a lot of variance. I expect it to represent the seasonal cycle because that is a large source of variance that is now not removed.

Because the first EOF has a nearly uniform pattern across the domain, it is clearly associated with the seasonal cycle! The first PC now looks like a sine wave with a 12 month period. Now the first EOF explains the majority of the variance, whereas the next EOFs explain much less. This supports the idea that seasonal variability is very important! However the second EOF still resembles the PDO so it does not seem to swamp our analysis.

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?

After detrending the data, the first EOF still resembles the PDO which suggests that detrending does not significantly change our analysis. I think this may be the case because linear trend existing in the data is a small magnitude signal compared to the PDO itself, as well as seasonal variability which is captured in the first EOF when I do not remove the seasonal cycle. In fact, it looks like the average increase in surface temperature over the time period is ~0.3 degrees C, which is smaller than the PDO signal as measured by EOF 1.

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?

Removing the cosine weighting does not seem to affect the EOFs and PCs in a significant way, thus cosine weighting doesn't seem to be very useful. However, if I was also considering data that spans the equator to the poles, then the grid cells would be very different in size and thus cosine weighting would be more important!

Note that this find is quite surprising to me because the latitude range was already quite significant!

4) <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?

There are modest changes when I do not standardize the data. For example the cool anomaly associated with the first EOF deepens. It is probably still a good idea to standardize, because grid cells with larger magnitude variability are now being over emphasized and can potentially drown out the signal in grid cells with lower magnitude variability. But to be clear the general pattern remains unchanged.

What I take away from this analysis is that the preprocessing should be done in such a way that you isolate the signal you're interested.