

**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 standardized 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 EOFs highlight the variance in facial features explained by each set. The principal components are the individual sets of facial features and can be combined to perform an EOF analysis to explain ‘x’ amount of variance. The first few eigenfaces suggest that a good number of employees are bald, that a strong browline is present, and that the light source is generally to the right of the image.*

**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?

*The number of EOFs needed depends on how accurate you want the face to be. Increasing the number of EOFs will decrease unexplained variance. The number of EOFs needed depends on how closely the face resembles the average face.*

**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 database limitation decreases its utility when attempting to recognize faces that are not explained by the largest PCs. A parallel in atmospheric sciences might be trying to apply an EOF analysis of tropical cyclone path across the Atlantic to tropical cyclone paths across the Pacific.*

## **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?.

*The first EOF/PC indicates a pool of generally cool water over the western and central Pacific, with a band of warmer water wrapping around along the Alaskan/Canadian border and then to the south of the colder water. The PC time series shows how closely correlated each year's signal is with the PC in question. The Nstar estimation method is the standard method, however for the purposes of calculating higher moments we might*

*instead use the Bretherton method. The results are identical between the eigenanalysis and SVD for the first order EOF but the sign is switched for the second order. I do not think the different sign is meaningful because the PC timeseries magnitude is also reversed.*

**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? Discuss 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?

*It becomes immediately evident that the primary driver of SST is the seasonal signal. The first EOF explains 95% of the variance, and the PC timeseries oscillates with a seasonal signal. Also, the spatial pattern becomes inexplicable in physical units, as the “noise” of the seasonal signal drowns out almost all other signal. Removing the seasonal trend was not useful.*

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

*The sign of EOF 1 in physical units is reversed, such that the coastal regions are relatively cold while the central ocean is relatively warm. The first PC timeseries no longer shows any trends (as expected), which seems to limit its utility as an analysis tool. Detrending allows us to compare years without an observed climate signal.*

**4) Save a copy of the notebook, rename it. Repeat the analysis but this time do not apply the cosine weighting.** Discuss 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 had no effect on out EOF/SVD analysis.*

**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).** Discuss 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 EOF explains more variance in the non-standardized analysis. The EOFs and PC timeseries remain largely the same, though there are subtle differences in the location and magnitude of the spatial signal pattern. Standardizing the data is useful from a general statistics standpoint as it allows us to expand our analysis toolkit.*

