Today's plan….

1. EOFs with actual data – Go through the steps together!!

2. How many EOFs should be retained? Assessing statistical significance of EOF analysis.

3. More examples of EOF analysis from the atmospheric and oceanic sciences literature and beyond

   -Evaluate CESM1 Northern Hemisphere SLP (like Hannachi et al. 2007 Figures we looked at last class, *aside: EOF analysis like this one was used to calculate Arctic Oscillation Index we used in Application Lab* #2)

   -Use EOF analysis to reduce noise and compress data (Barnes example)

EOFs with actual data

Step #1 – Prepare your data – remove the mean

Removing the sampling mean

- it is typical to remove the sampling dimension mean.
- this will make your PC's have a mean of zero
- you may want to remove the seasonal cycle if this is not of interest to you (or diurnal cycle)

Removing the spatial mean

- it is less common to remove the spatial mean since you often want to allow the EOFs to have the same pattern everywhere
- some exceptions: you are interested in the gradients in temperature, not the full temperature field itself
- you are interested in the eddy component of the field, so you might remove the zonal-mean

EOFs with actual data

Step #1 – Prepare your data – standardize your data

Standardizing the data

- sometimes, it is desirable to take the amplitude out of the data before doing EOF analysis, for example, when
  - your data is made of a combination of parameters with different units, and you don't want the parameter with the biggest units to dominate the variance
  - if the variance of the data varies greatly from space to space (and this is not of interest to you)
- the way to do this is after subtracting the mean along the sampling dimension, divide by the standard deviation as well (now you have units of σ)
- note that you will be eigenanalyzing the *correlation matrix* (not the covariance matrix) if you use the eigenanalysis method

EOFs with actual data

Step #1 – Prepare your data – weighting

If your data are gridded, calculate the weight according to the grid area.

If you are calculating EOFs using eigenanalysis of the covariance matrix OR SVD – weight by the square root of the cosine(latitude).

When you multiple the grid points by each other to calculate the covariance matrix – the result is that the data will be weighted by the cosine.

EOFs with actual data

Step #1 – Prepare your data – missing data

Missing data

- most programs will not perform eigenanalysis or SVD with missing values or NaNs
- the best option is to only use locations with $\alpha\%$ of the data present, and calculate the EOFs via the covariance matrix so that $\mathbf{C}$ has no missing data when input into the eigenanalysis function

Note: another common practice is to replace missing data with the mean. If you are working with anomalies (i.e., you removed the mean) – set missing data equal to 0.

EOFs with actual data

Step #1 – Prepare your data – Example from Hannachi et al. 2007

Application

We have applied EOFs to winter monthly SLP over the Northern Hemisphere (NH). The data come from the National Center for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalyses (Kalnay *et al*., 1996; Kistler *et al*., 2001). They are available on a $2.5^\circ \times 2.5^\circ$ regular grid, and span the period January 1948 to December 2000. The mean annual cycle is first calculated by averaging the monthly data over the years, then subtracted from the data to yield SLP anomalies. We are only interested in analysing the winter season defined by December to February (DJF). The data are therefore obtained by concatenating the winter monthly means for all years. Finally a weighting by the square root of the cosine of the corresponding latitude is applied to each grid point to account for the converging longitudes poleward. The data over the NH north of $20\,^\circ$N are used to compute EOFs. Note that the examples presented here have also been used in Hannachi *et al*. (2006).

EOFs with actual data

Step #2 – Calculate EOF using either eigenanalysis or SVD. Compute the principal components.

Step #3 – Plot the eigenvalues

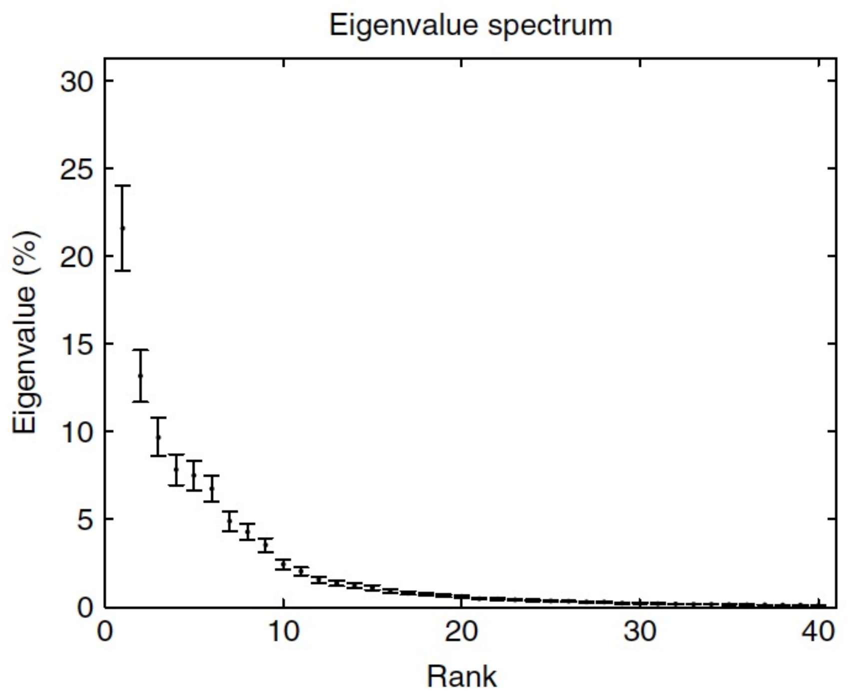


Figure 1. Spectrum, in percentage, of the covariance matrix of winter monthly (DJF) SLP. Vertical bars show approximate 95% confidence limits given by the rule of thumb (18). Only the leading 40 eigenvalues are shown.

Figure from Hannachi et al. 2007

EOFs with actual data

Step #3 – Assess statistical significance. How many EOFs should be retained? North et al. (1982) with N*

EOFs will always give you an answer. For example, EOFs of red noise will give you sines and cosines that look physical. The rate at which the eigenvalues drop off will be a function of the redness of the data.

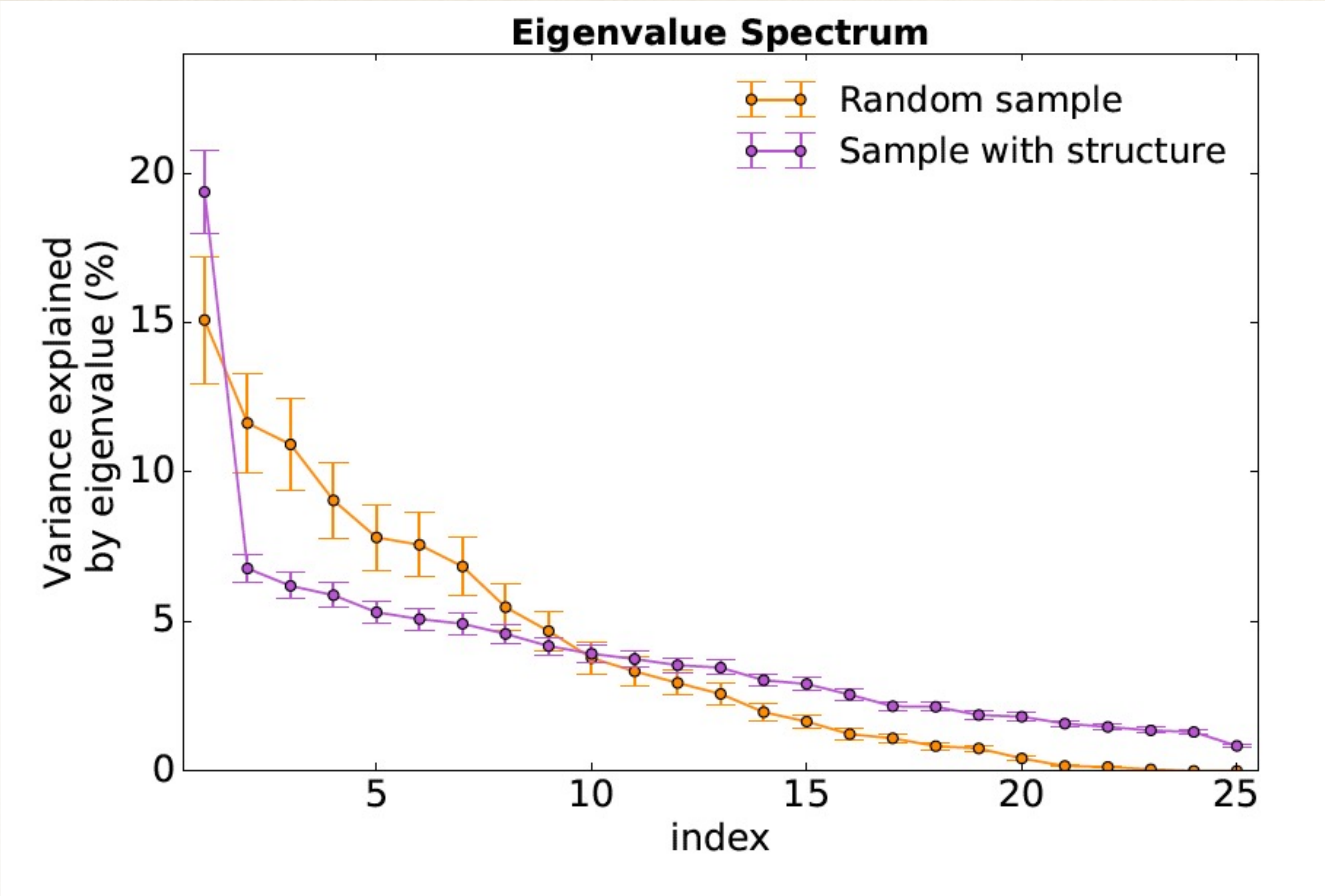
The orthogonality of EOFs impose a constraint on their shape! Higher order EOFs are often trying to represent the noise while still being orthogonal to the other EOFs.

North et al. (1982) argues that the significance of an EOF is a function of the degree of separation between eigenvalues, where the 95% confidence bounds on the eigenvalues is

$$\Delta\lambda = \lambda\sqrt{\frac{2}{N*}} \tag{80}$$

In this case, the N* is our friend the "effective degrees of freedom". This is not obvious how to decide what N* is - you need to come up with a value that is representative of the entire data set!

Step #3 – Assess statistical significance. Example.



An example of a plot showing the eigenvalues and their 95% confidence bounds following North et al. (1982) is given in Figure 3. The orange bars are the eigenvalues of a data set that is pure white noise. The purple bars denote the eigenvalues from a data set that has some structure, specifically, the first eigenvector explains substantially more variance than the others.

EOFs with actual data

Step #3 – Assess statistical significance. How many EOFs should be retained? Other options…

– subdividing your sample and comparing EOF structures, are they the same?

– are your results explainable through theory?

– are results sensitive to the size of the domain? (if highly sensitive, this could be sines and cosines being fit inside the boundary)

EOFs with actual data

Step #4 – Plot the EOF patterns (structure) and PC timeseries (how structure varies in sample)

Let $\mathbf{X}$ have dimensions $[\mathrm{M} \times \mathrm{N}]$ such that M is the number of samples and N is the spatial locations.

From now on, to make notation a bit easier, we will only write equations using EOF 1 and PC 1. The steps can be repeated for EOF 2 and PC 2, EOF 3 and PC 3, etc.

In the literature, one sees both the EOFs and the PCs plotted, depending on what is of interest.

Generally, the spatial domain is more interesting since the PC (sampling domain) is usually noisy.

Problem: the EOFs don't have physical units, so we need to figure out how to plot the patterns in a meaningful way

EOFs with actual data

Step #5 - Regress the data (unweighted data if applicable) onto standardize values of the principal component.

Here is how it is done:

1. calculate the leading PC Z (in a variety of ways, either you get it directly from SVD, or using eigenanalysis you project the data and EOF onto each other)
2. standardize Z (the mean is already 0 if you subtracted the mean from the data, so just divide by σ_Z)
3. project Z back onto your *original* anomaly data and divide by the length of Z to get D, the EOF 1 structure in the original units i.e. the unweighted data X. if you standardized X before the computation, you want the non-standardized data

If the original data is $\mathbf{X}$, the weighted data is $\mathbf{X}_w$, following eigenanalysis

$$z_1 = \mathbf{X}_w e_1 \tag{77}$$

Then,

$$\tilde{z}_1 = \frac{z_1 - \mu_{z_1}}{\sigma_{z_1}} \tag{78}$$

Then,

$$d_1 = \frac{1}{M}\tilde{z}_1^\mathsf{T}\mathbf{X} = \frac{1}{\text{length}(z_1)}\tilde{z}_1^\mathsf{T}\mathbf{X} \tag{79}$$

and now $\mathbf{d}_1$ is in physical units of your data, and denotes the anomaly associated with 1 standard deviation variation of z_1.

Back to this Classic Example: EOF Analysis of Northern Hemisphere Sea Level Pressure During Winter Months

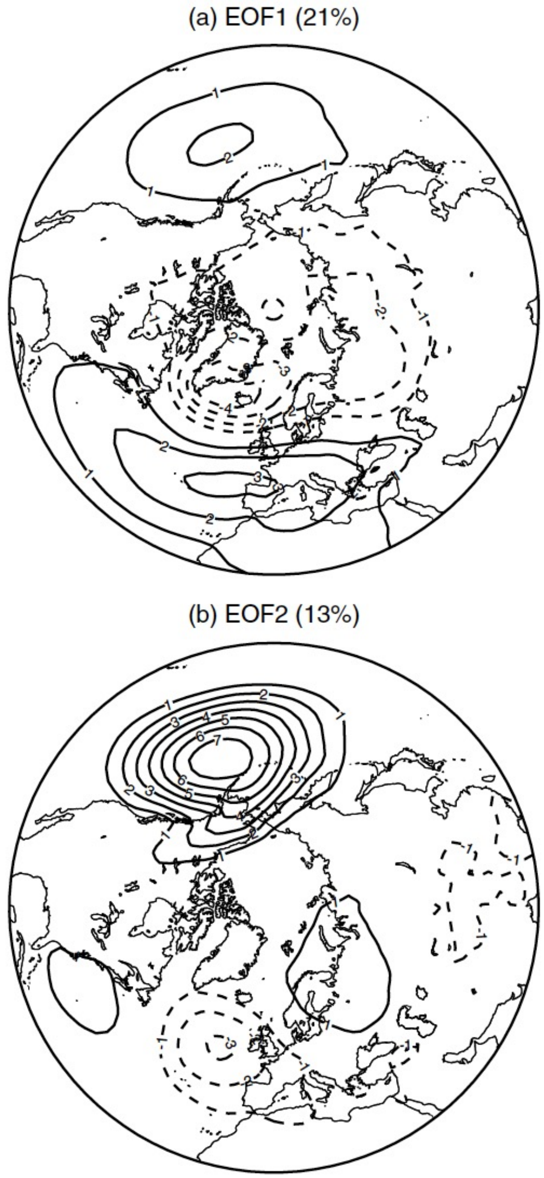


Figure 2. The first (a) and the second (b) EOFs of DJF monthly mean SLP. Positive contours solid, negative contours dashed. EOFs have been multiplied by 100.

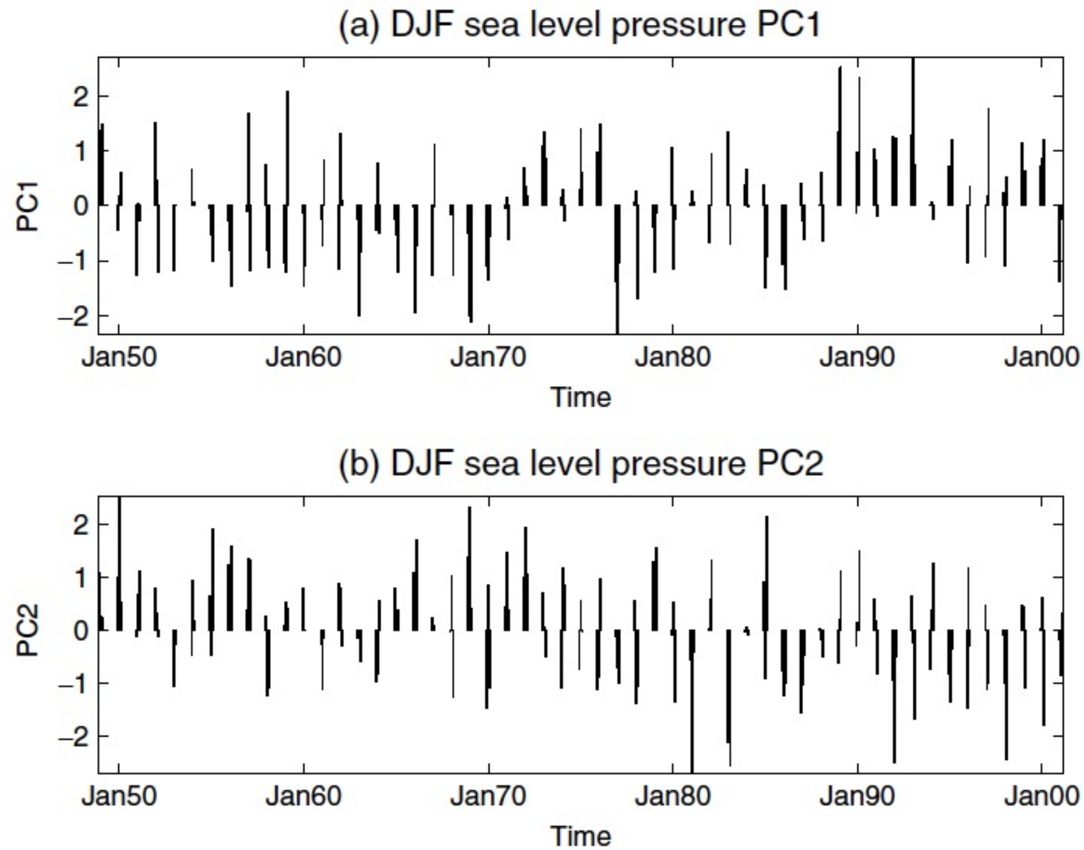
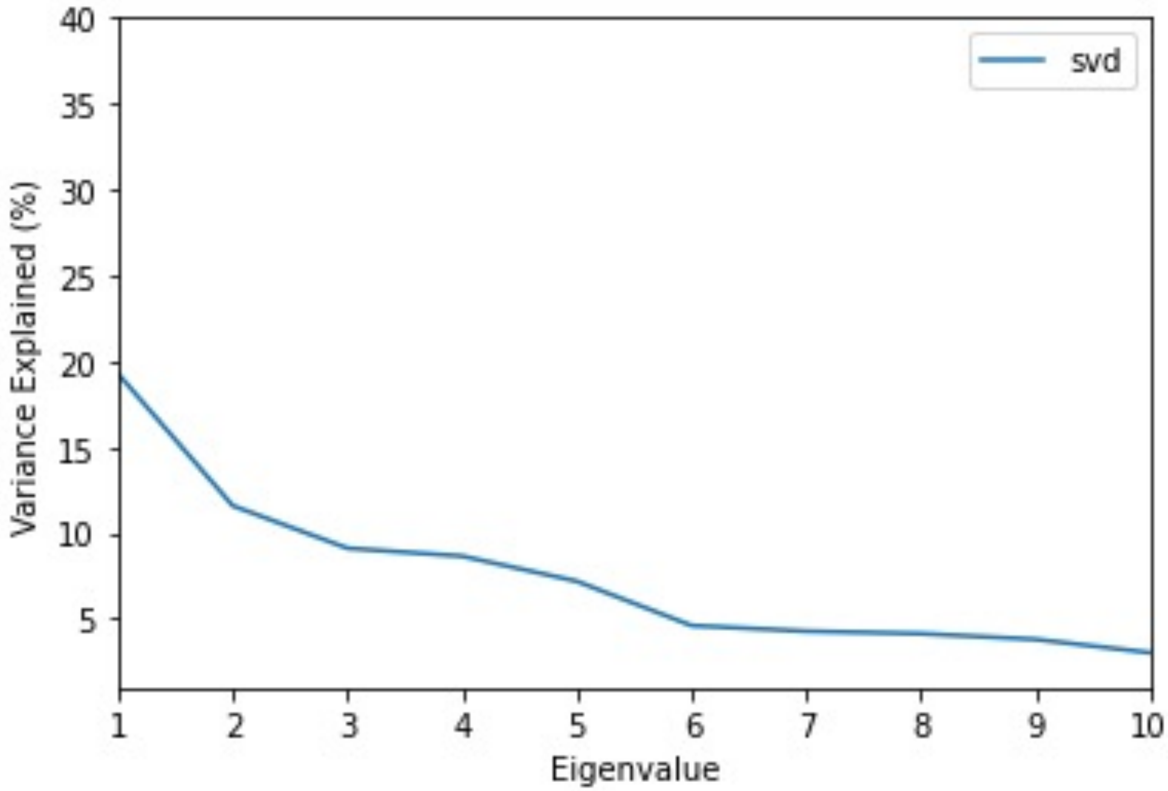


Figure 3. The leading two scaled PCs corresponding to the leading two EOFs of Figure 2.

Figures from Hannachi et al. 2007

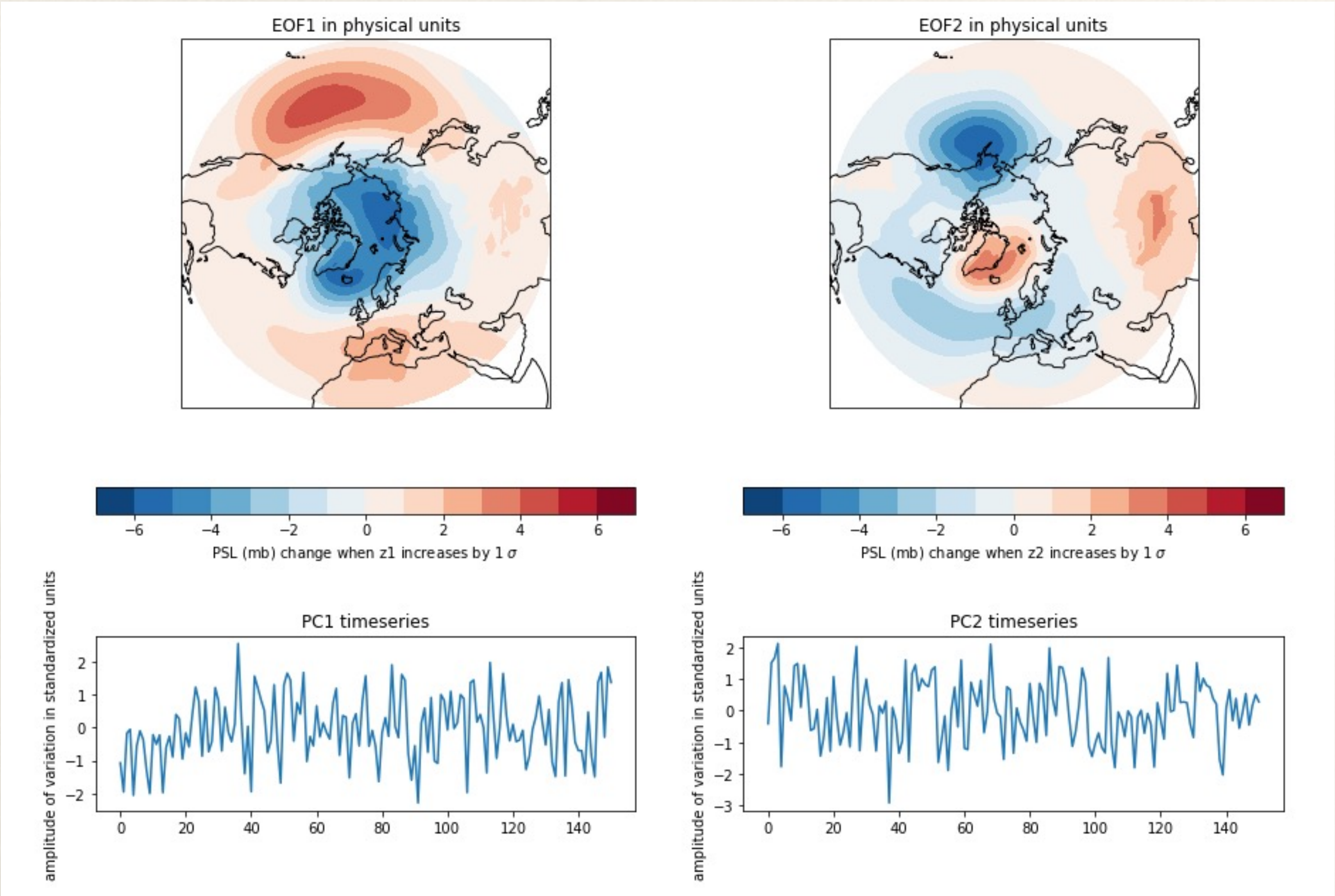
EOFs with actual data – Let's evaluate the CESM by redoing the Northern Hemisphere Sea Level Pressure During Winter Months

```
Percent variance explained by EOF1 in CESM: 19.0 %
Percent variance explained by EOF1 in Obs: 21%
Percent variance explained by EOF2 in CESM: 12.0 %
Percent variance explained by EOF2 in Obs:  13%
```



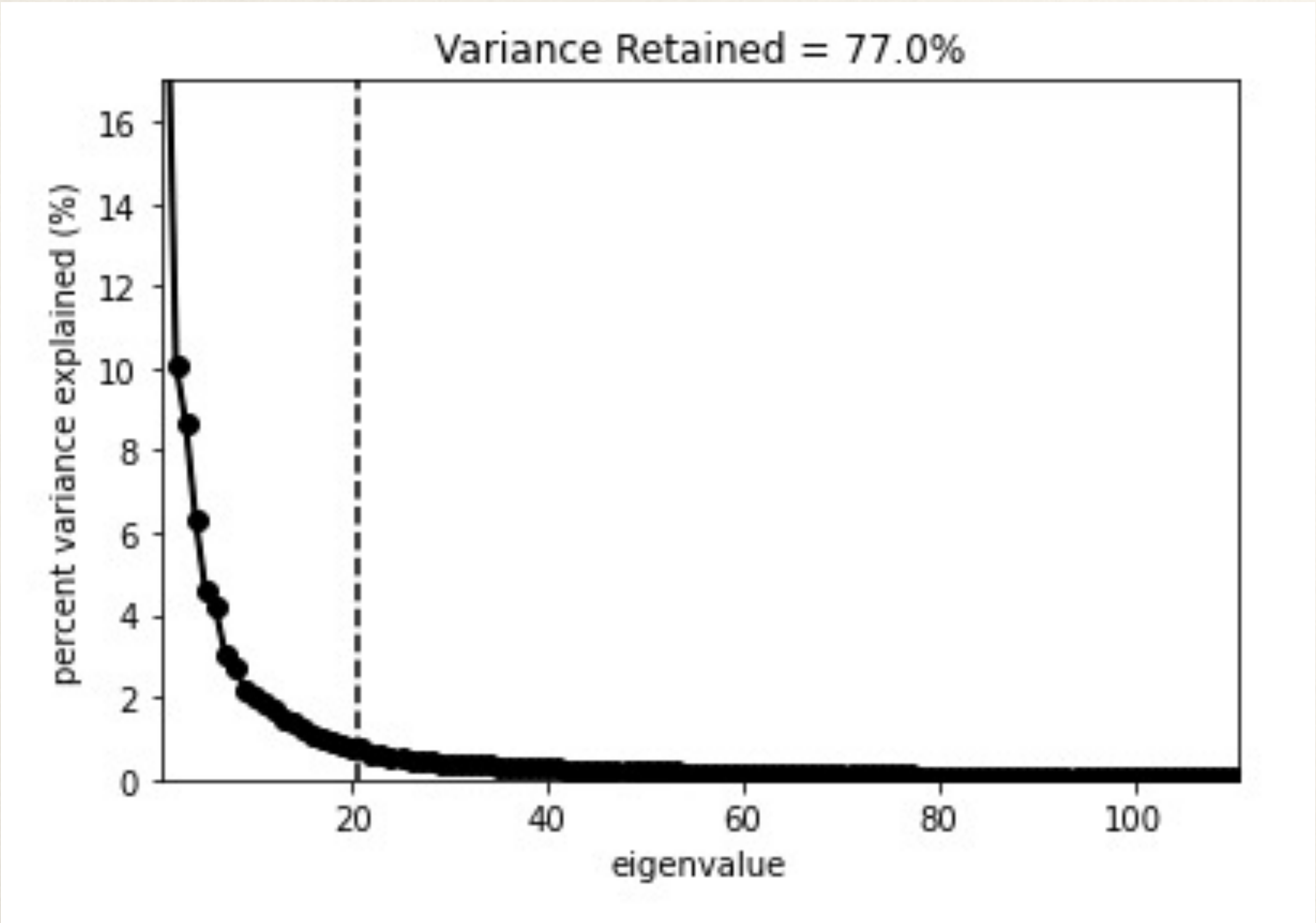
Try the python code –
eof_example_CESM_NH_PSL_cartopy_cosweightLAST.ipynb

EOFs with actual data – Let's evaluate the CESM by redoing the Northern Hemisphere Sea Level Pressure During Winter Months



Try the python code –
eof_example_CESM_NH_PSL_cartopy_cosweightLAST.ipynb

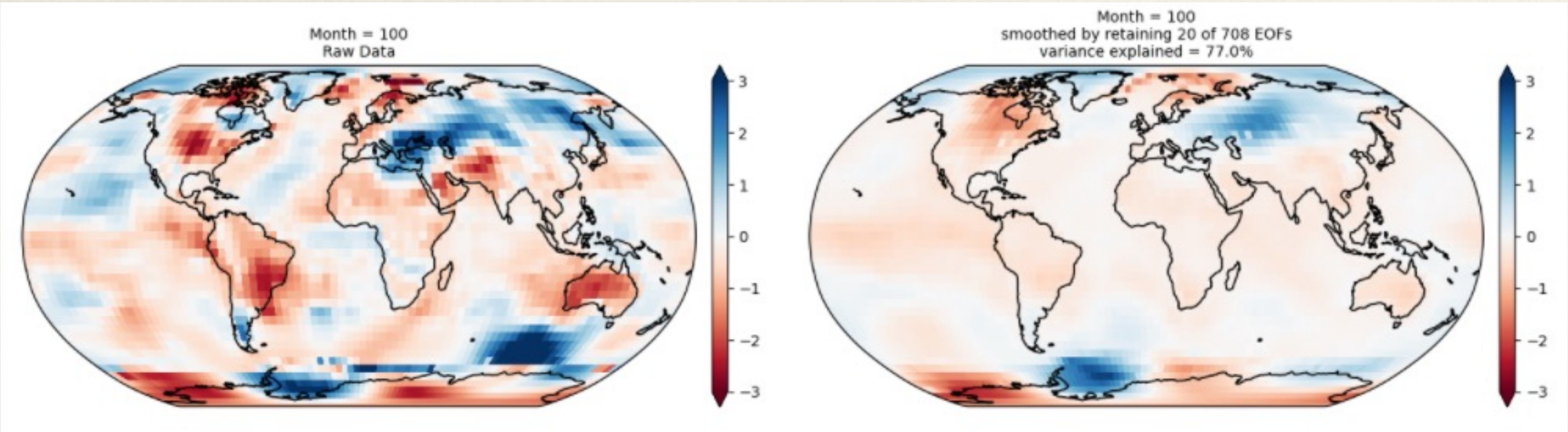
EOFs with actual data – Application of data compression/noise reduction using monthly surface air temperatures



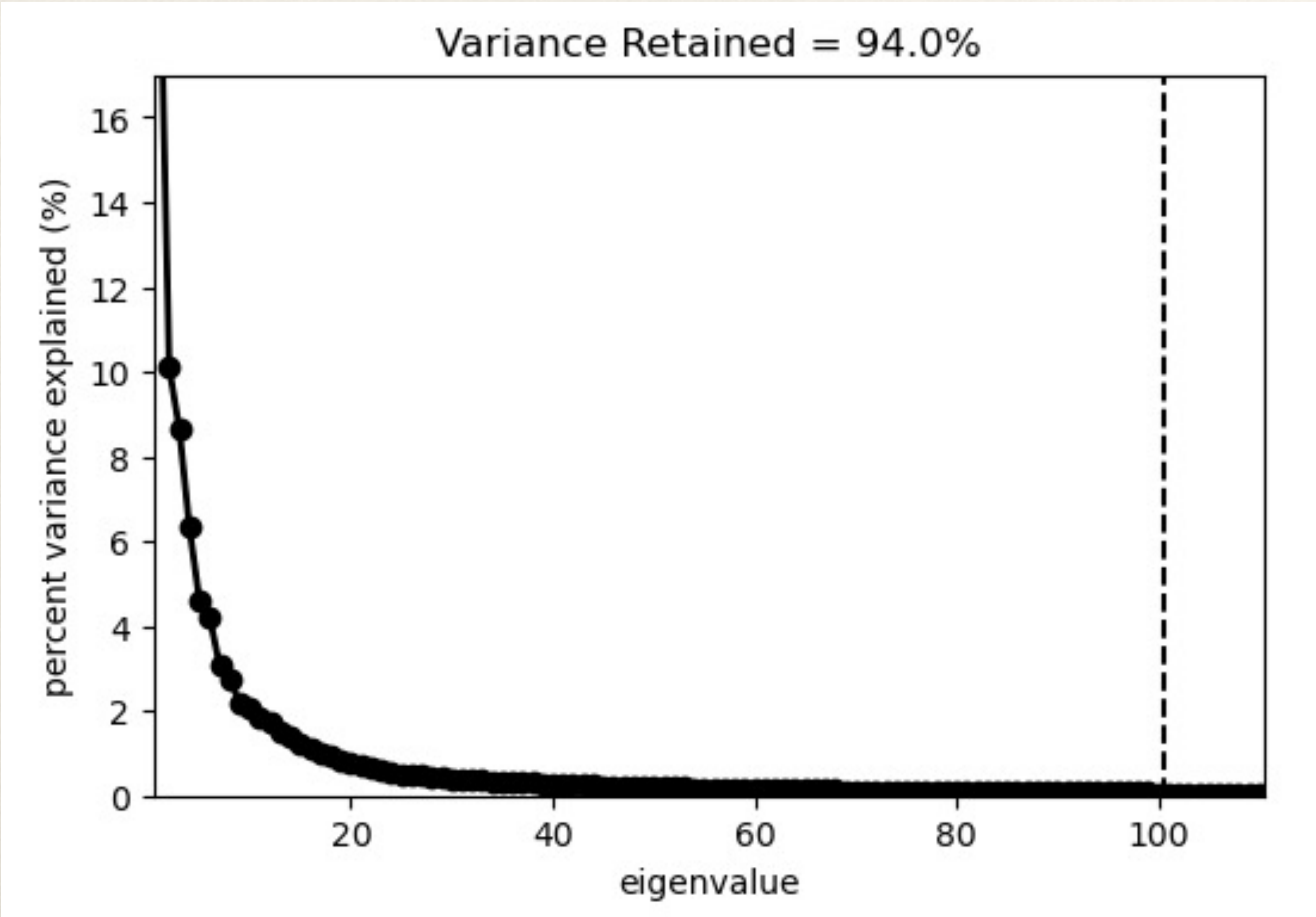
Try the python code – eof_noise_reduction_cartopy.ipynb

EOFs with actual data – Application of data compression/noise reduction using ERA Interim 2 meter temperatures.

Let’s compare the original data(left) reconstructed data with the first 20 EOFs (right)

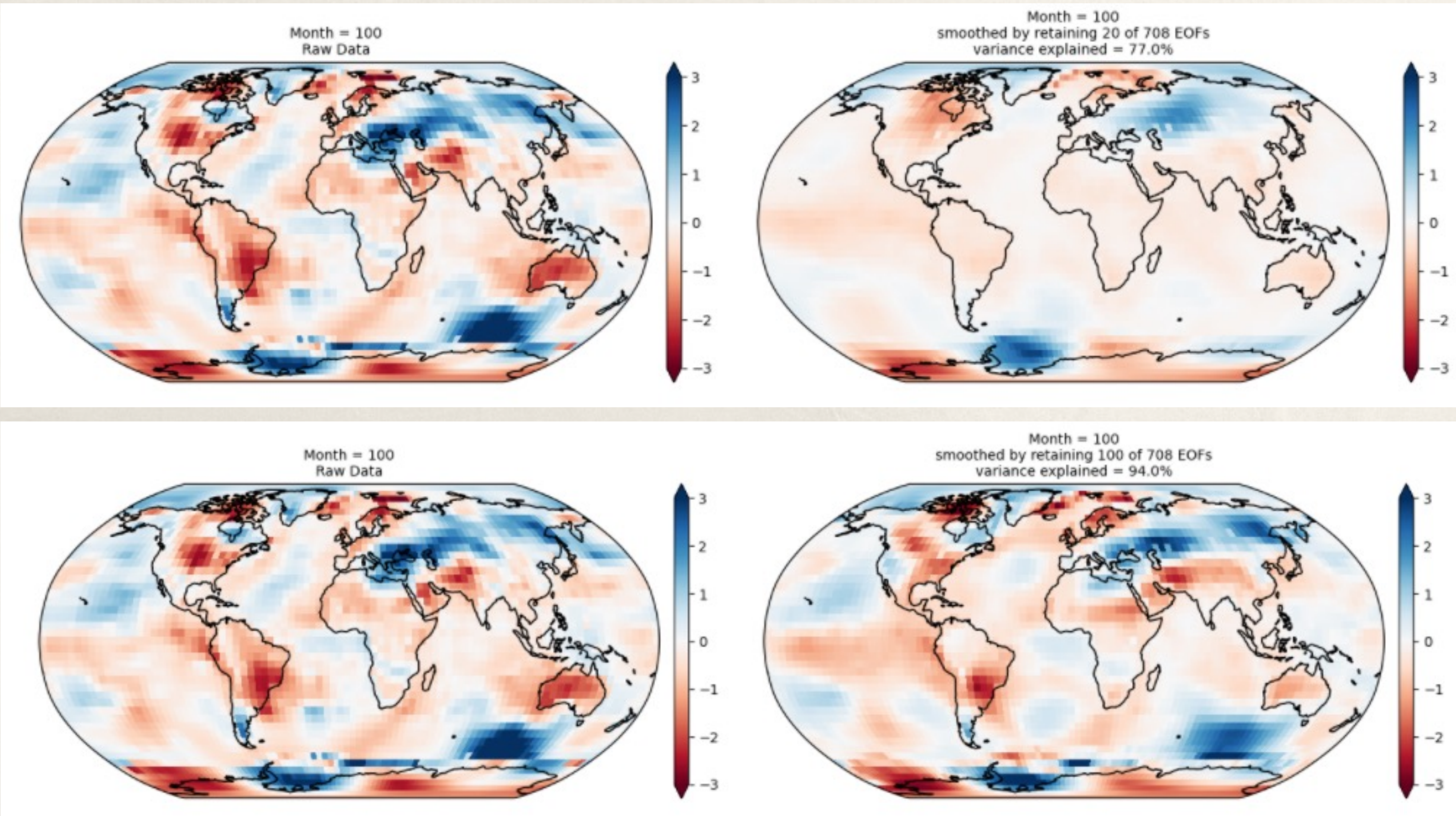


take2 – this time with 100 EOFs retained



Try the python code – eof_noise_reduction_cartopy.ipynb

EOFs with actual data – Application of data compression/noise reduction using monthly surface air temperatures.
Let's compare the reconstructed data with the first 20 EOFs (top) and the first 100 EOFs (bottom)



Power of EOF analysis (standard)

(from Hannachi et al. 2007)

- Identifies main orthogonal modes of variance
- Can be used in data analysis to identify main spatial structures and their temporal evolution
- Can be used to compress data
- Can be used to reduce noise
- Can be used to efficiently compare models and observations

Limitations of EOF analysis (standard)

(from Hannachi et al. 2007) and others

- ***Geometric constraint not very physical:***
  -physical processes are not independent, physical modes are not expected in general to be orthogonal.
  -spatial orthogonality can cause structures to exist over most of the domain with significant amplitude when the real signal is more localized.
  -domain dependence (something you can check!)
  -wave features challenging to interpret

EOF Analysis – Limitations (from Hannachi et al. 2007)

"Orthogonality constraint can cause EOFs to have structure over most of the domain and with significant amplitude, when in fact one expects the patterns to be more localized"

"Physical modes are expected in general to be non-orthogonal."

There are ways to get around these limitations:

Rotating EOFs and sacrificing orthogonality to simplify/localize the structure of the patterns (discussed at length in Hannachi et al. 2007 and also in the Hartmann notes pages 95-97). ***Perhaps you will try to code for yourself in Homework #3?***