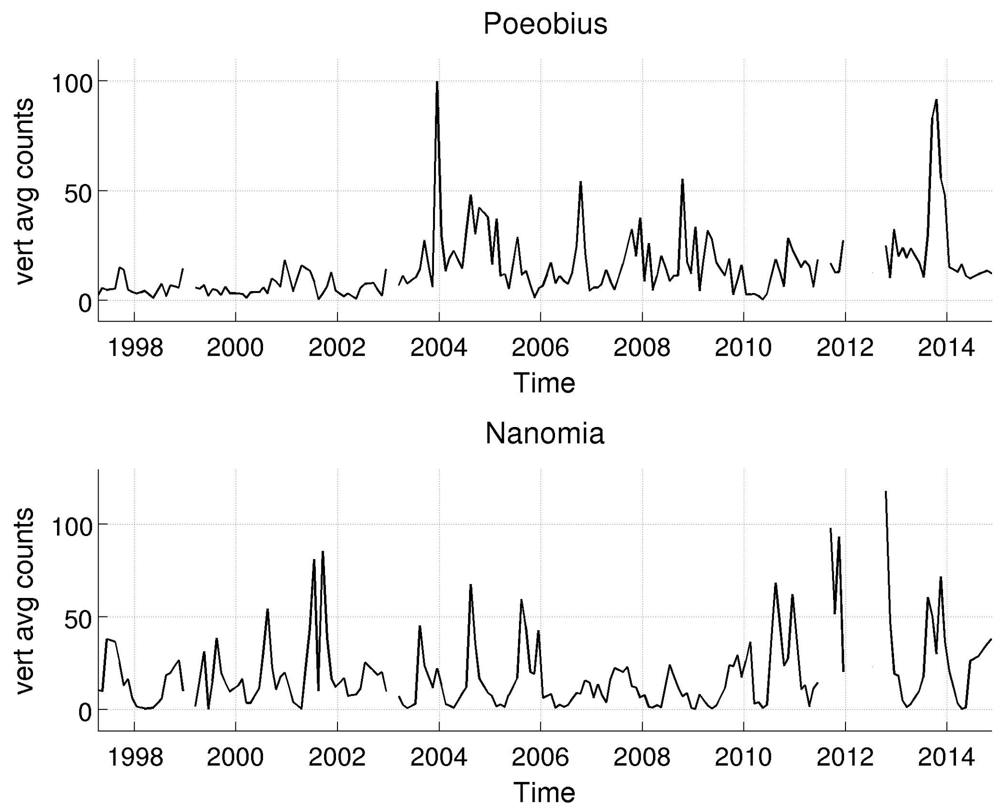
Data synthesis methods

- Principal component analysis (PCA)
  - goal: summarize a large number of input data series into a few ("principal components", PC) that capture most of the variance
  - concept: find the one time series that correlates the most with all of the input data series (PC1), remove the associated variability, repeat.

PCA: simple example (1)

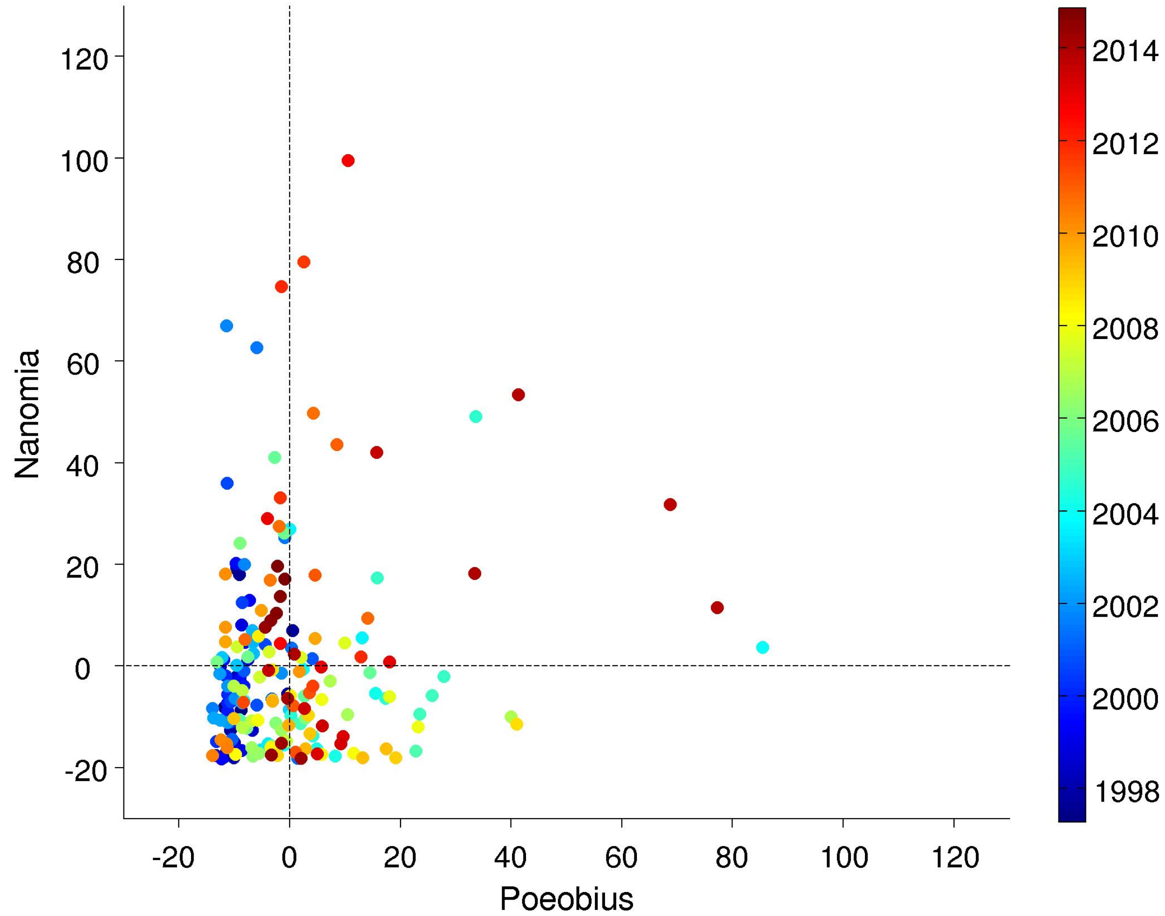
Simplest example: 2 data series (here vertically averaged counts of Poeobius & Nanomia)

PCA will search how these time series are related and extract their **dominant variability**



PCA: simple example (2)

Another way to look at the relationship between the time series



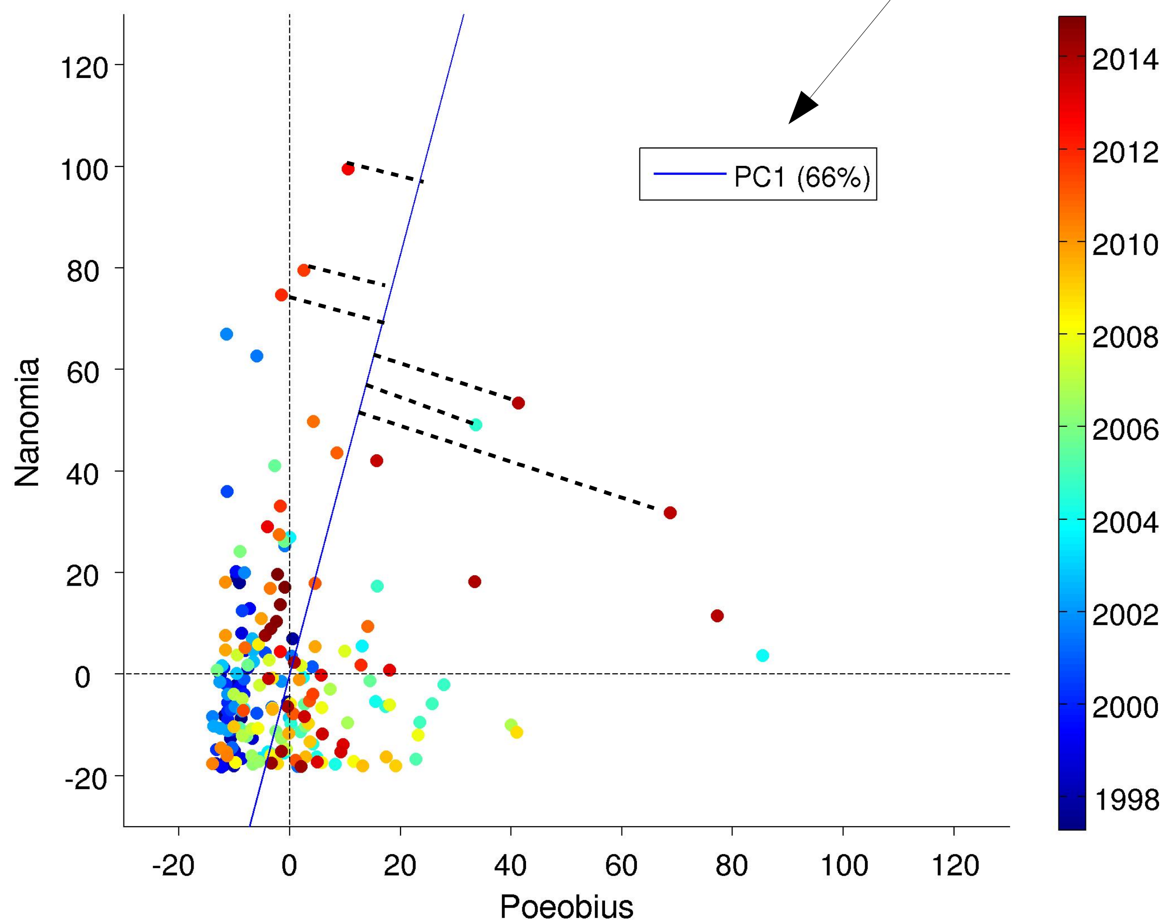
anomaly = data minus mean

PCA: simple example (2)

Another way to look at the relationship between the time series

percentage of variance captured by PC1

PCA finds the axis that maximizes the variance of data points when projected on it (line of **best fit**) – **PC1**

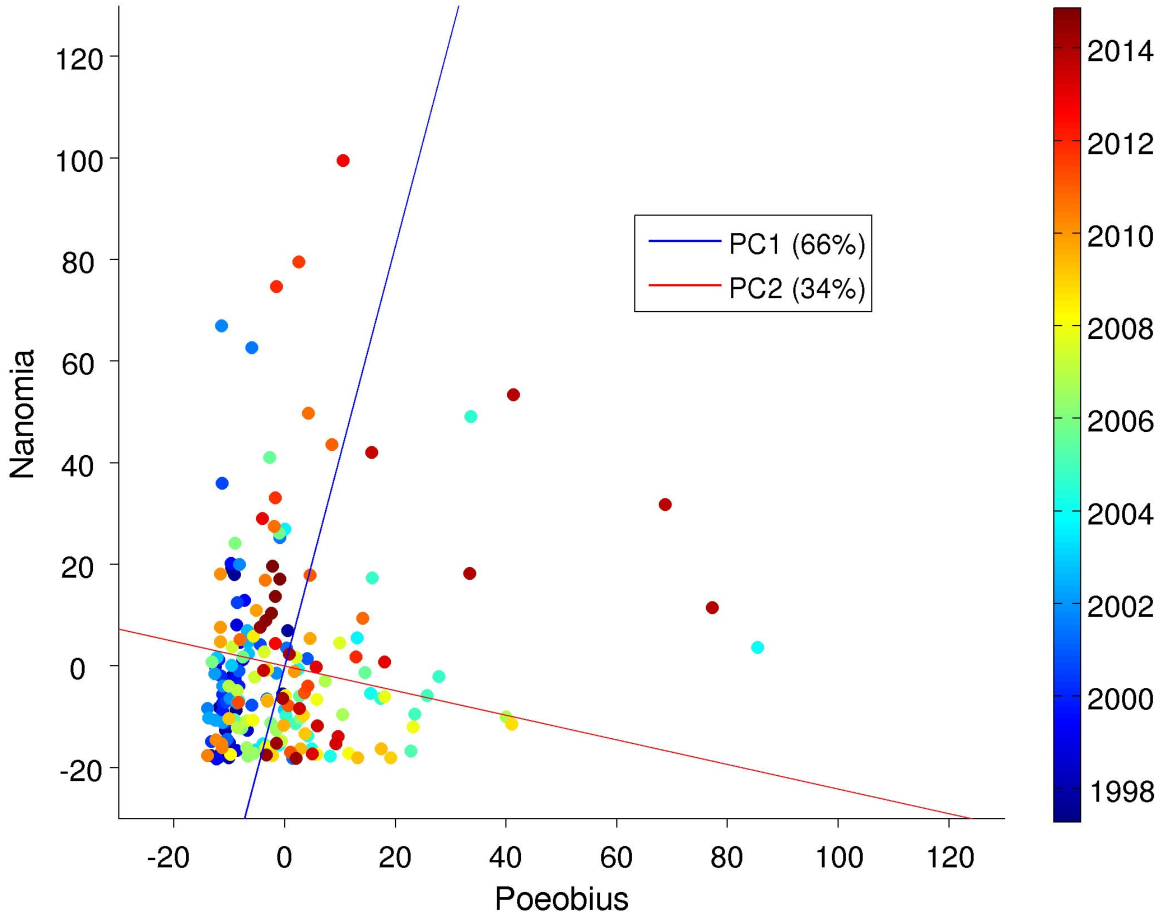


PCA: simple example (2)

Another way to look at the relationship between the time series

PCA finds the axis that maximizes the variance of data points when projected on it (line of **best fit**) – **PC1**

Then the one perpendicular to the first that maximizes the residual variance – **PC2**



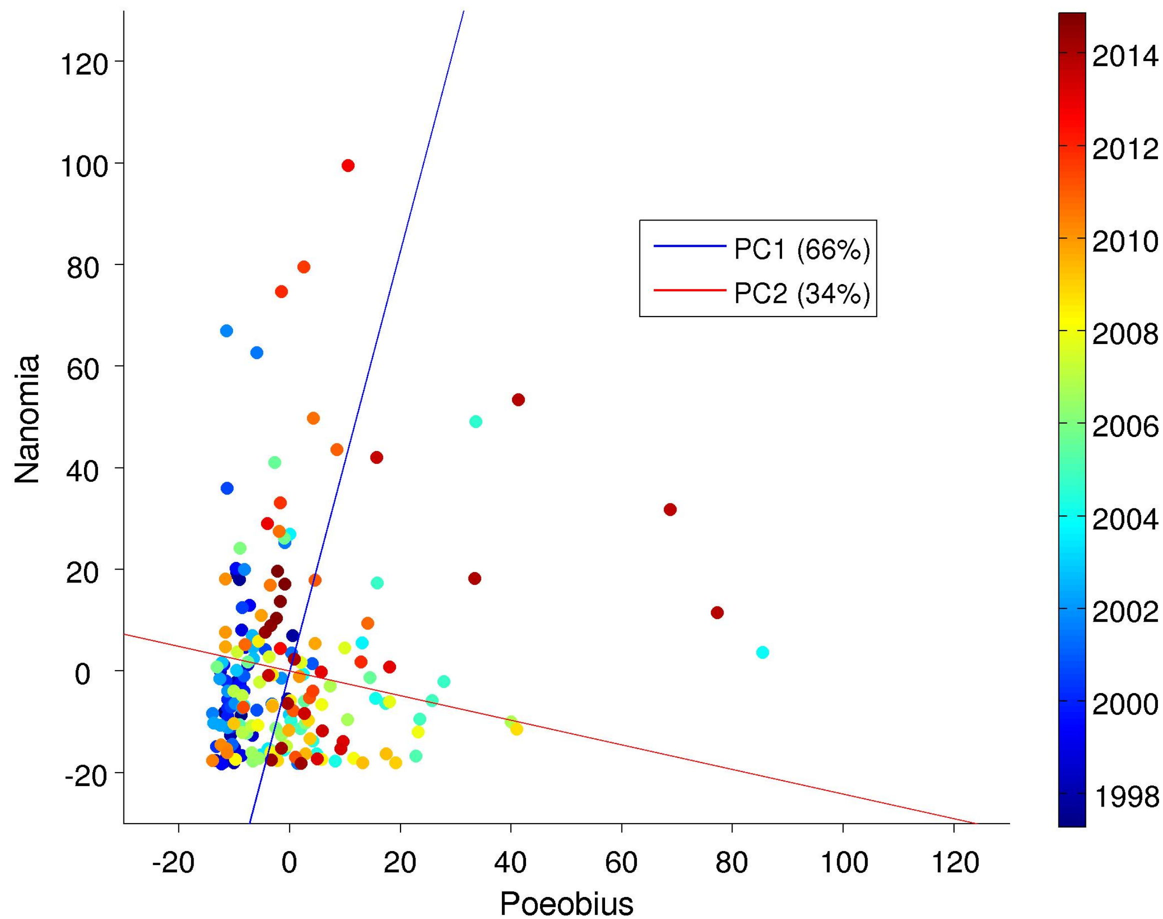
For 2 time series there are only 2 PCs. Mathematically PCs are found by diagonalizing the covariance matrix and finding its eigenvectors. If the shape of the dots was too "round" the modes would not be statistically separable (variance explained would be very close to 50% for both)

PCA: simple example (2)

Another way to look at the relationship between the time series

PCA outputs:

- **principal components** (PC, time series)
- **loadings** (one per species and per mode)
- **percentage of variance** associated with each mode

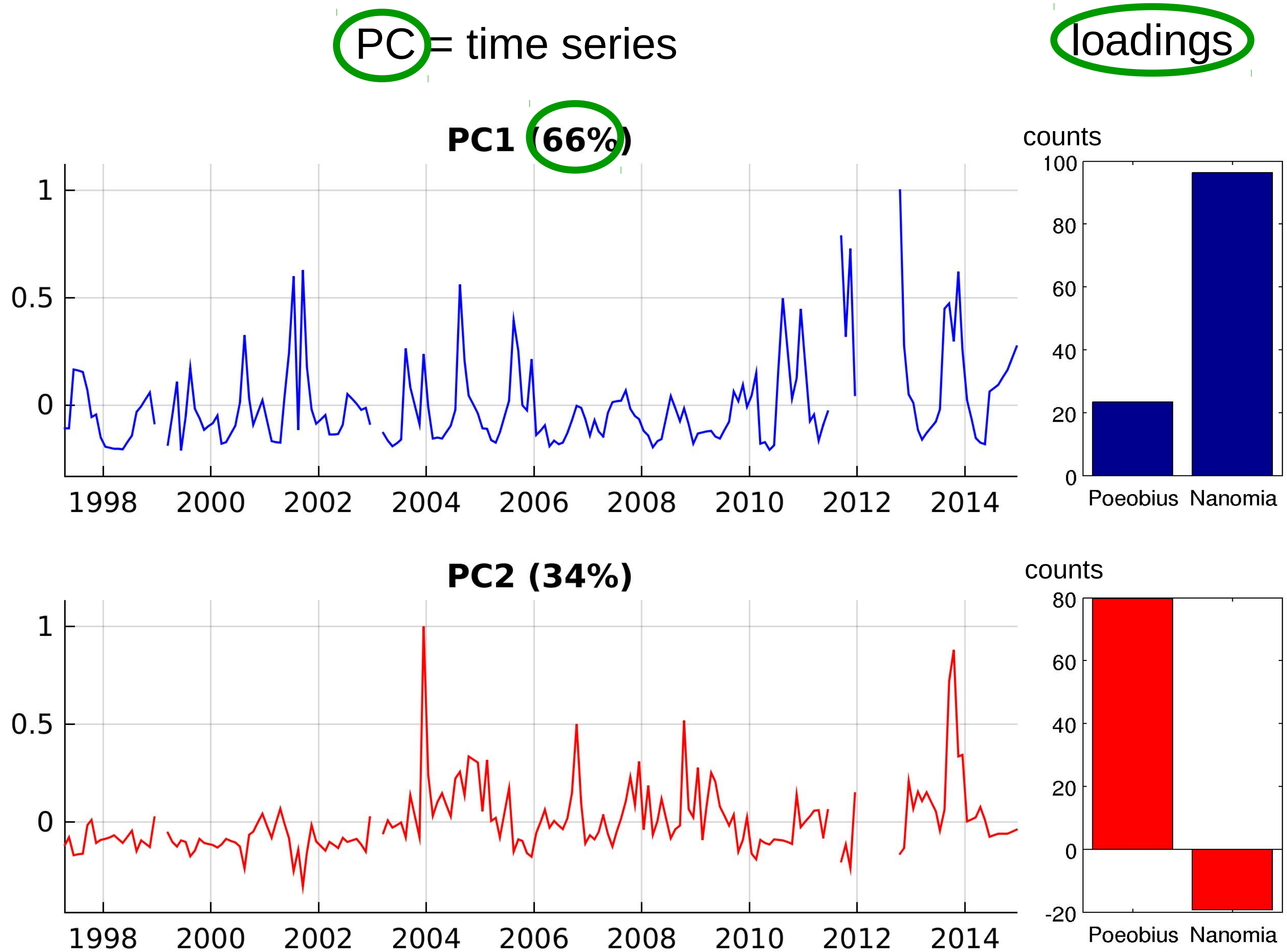


For 2 time series there are only 2 PCs. Mathematically PCs are found by diagonalizing the covariance matrix and finding its eigenvectors. If the shape of the dots was too “round” the modes would not be statistically separable (variance explained would be very close to 50% for both)

PCA: simple example (3)

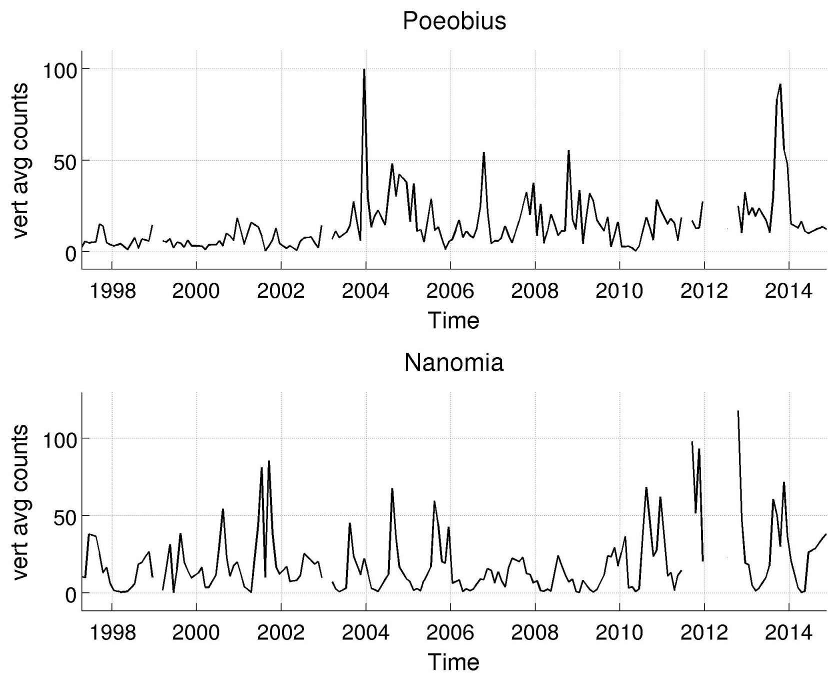
PCA outputs:

- **principal components** (PC, time series)
- **loadings** (one per species and per mode)
- **percentage of variance** associated with each mode



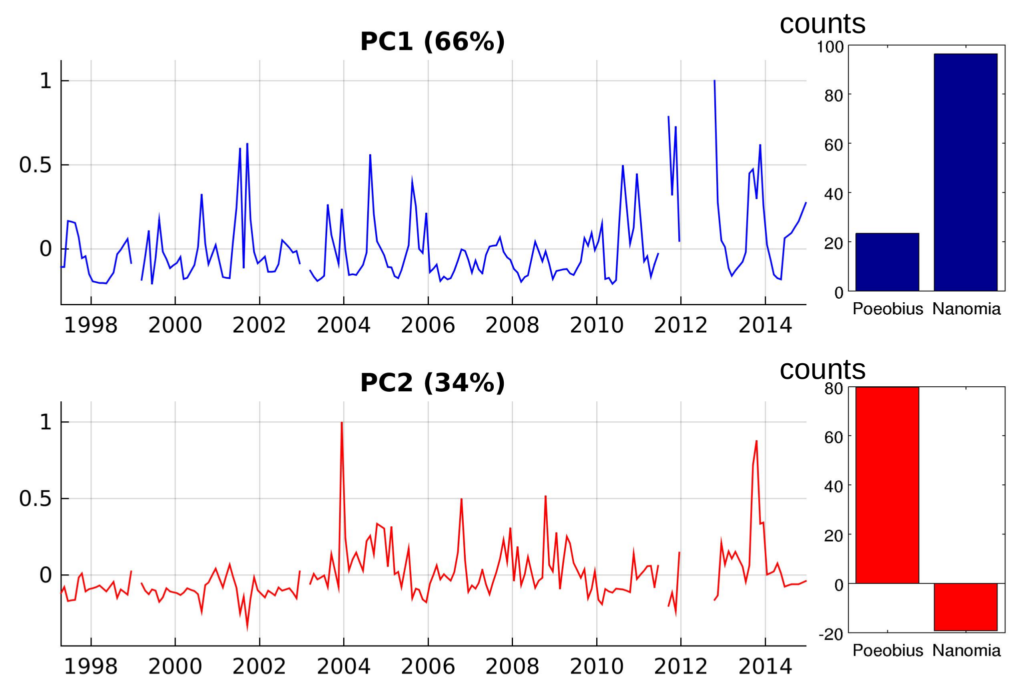
PC1 captures Nanomia most strongly, and shows that Poeobius anomalies are usually weakly positive when Nanomia is high. PC2 captures Poeobius most strongly, and shows that Nanomia anomalies are usually weakly negative when Poeobius is high.

PCA: simple example (3)



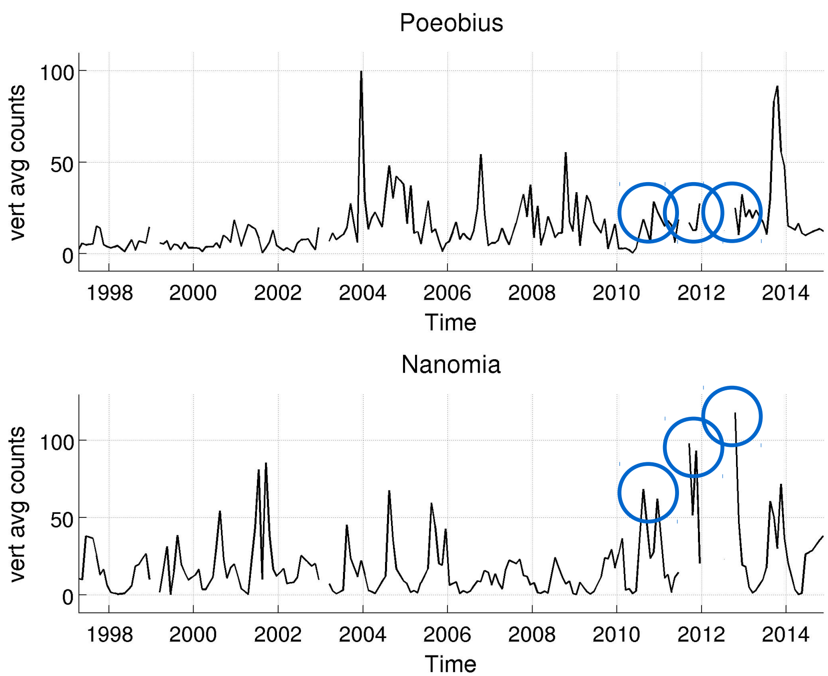
PC = time series

loadings



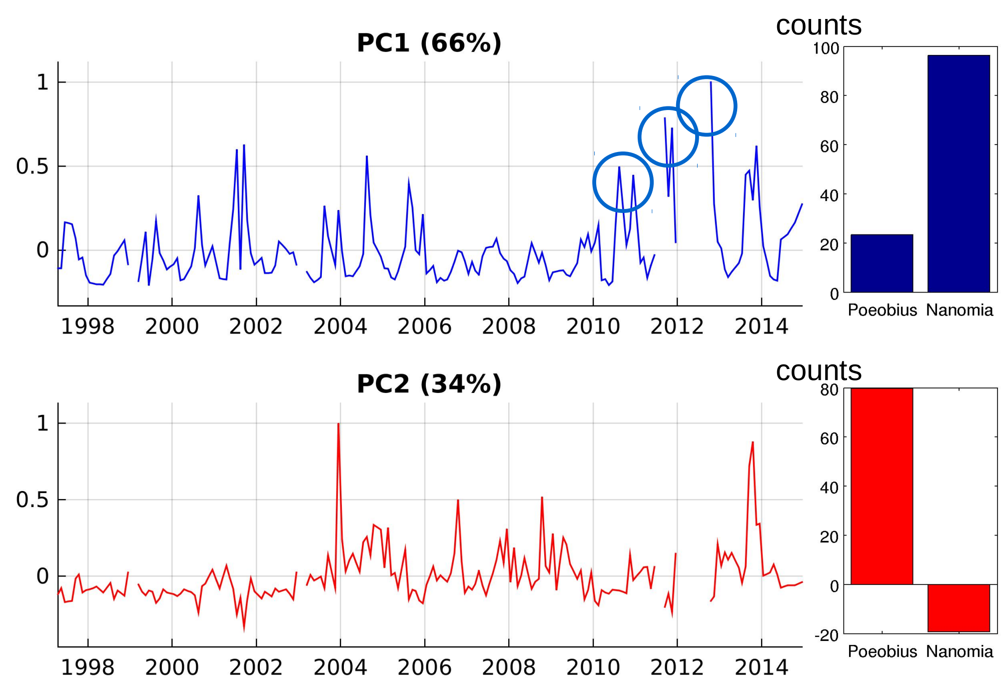
PC1 captures Nanomia most strongly, and shows that Poeobius anomalies are usually weakly positive when Nanomia is high. PC2 captures Poeobius most strongly, and shows that Nanomia anomalies are usually weakly negative when Poeobius is high.

PCA: simple example (3)



PC = time series

loadings



PC1 captures Nanomia most strongly, and shows that Poeobius anomalies are usually weakly positive when Nanomia is high. PC2 captures Poeobius most strongly, and shows that Nanomia anomalies are usually weakly negative when Poeobius is high.

PCA: data reconstruction (1)

PCA outputs:
a number of modes, ranked by variance explained, and each consisting of:

- **PC**: the principal component time series
- **loadings**: how much each data series is related to this mode (ie data series projection)
- **varex**: the variance explained by this mode

All modes are orthogonal ie. PCs are uncorrelated with each other (r=0)

Data reconstruction from PCA:

serie X = PC1 * $\text{loading1}_{\text{serieX}}$ + PC2 * $\text{loading2}_{\text{serieX}}$ + ….

time series
common to all data series

serieX loading for mode 1
one number specific to each data series

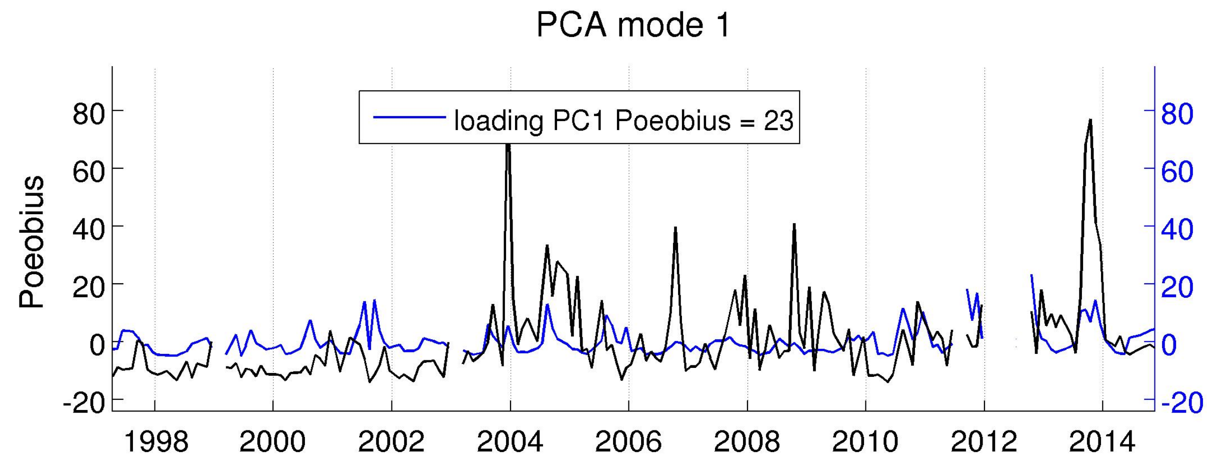
Usually by convention the PCs are normalized between -1 and +1, and loadings keep the units.

PCA: data reconstruction (2)

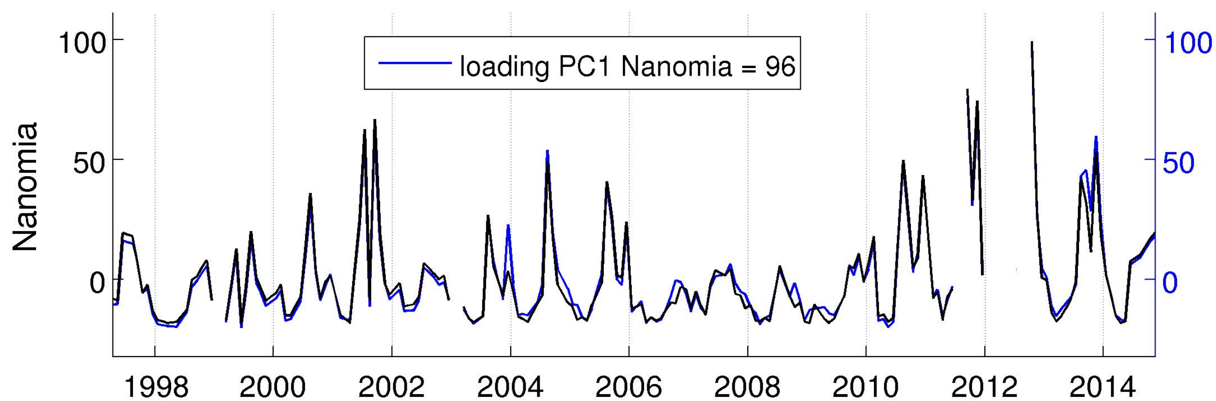
maximizes the information retained from series 1 & 2

$$\text{series X} = \boxed{\text{PC1} * \text{loading1}_{\text{seriesX}}} + \text{PC2} * \text{loading2}_{\text{seriesX}}$$

red line = PC1 * 23



red line = PC1 * 96



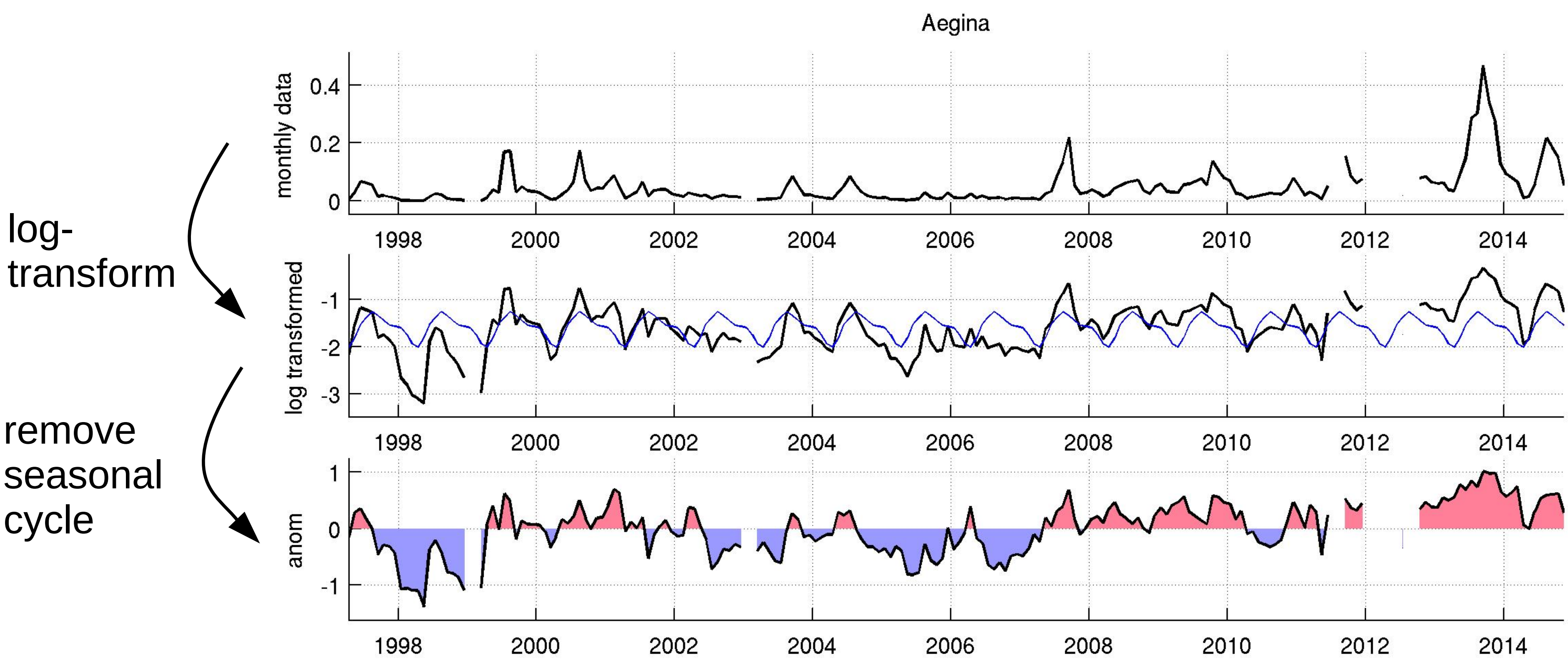
PCA: 3D example

- http://setosa.io/ev/principal-component-analysis/

PCA: more complex example

89 midwater time series (counts per volume), regridded monthly, vertical gaps filled and vertically averaged (200-1000m)

Because PCA maximizes the variance, very high values dominate the analysis → **transformation needed**.

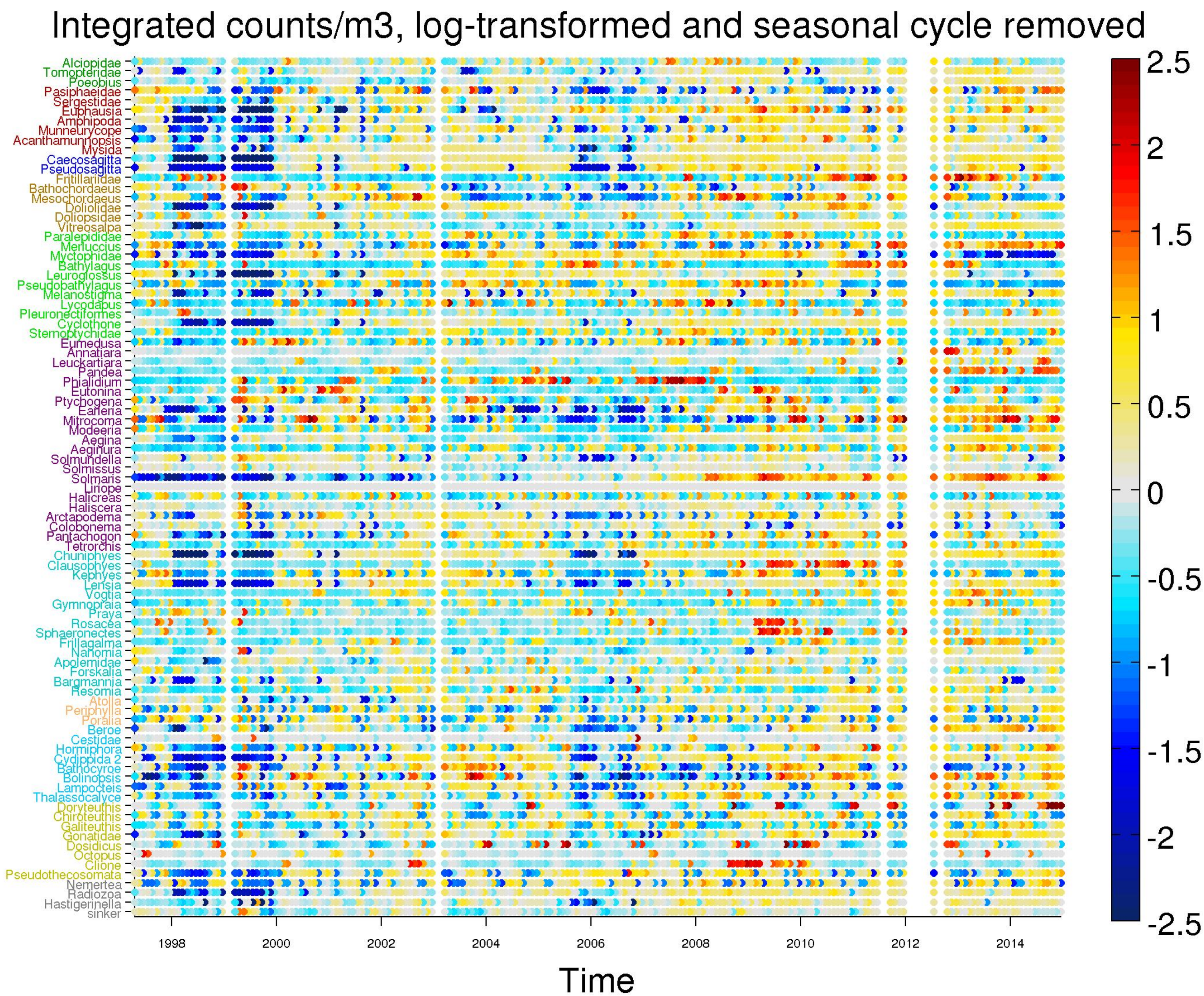


Matrix of all anomalies

Previous example had 2 data series as inputs, and 2 modes as outputs.

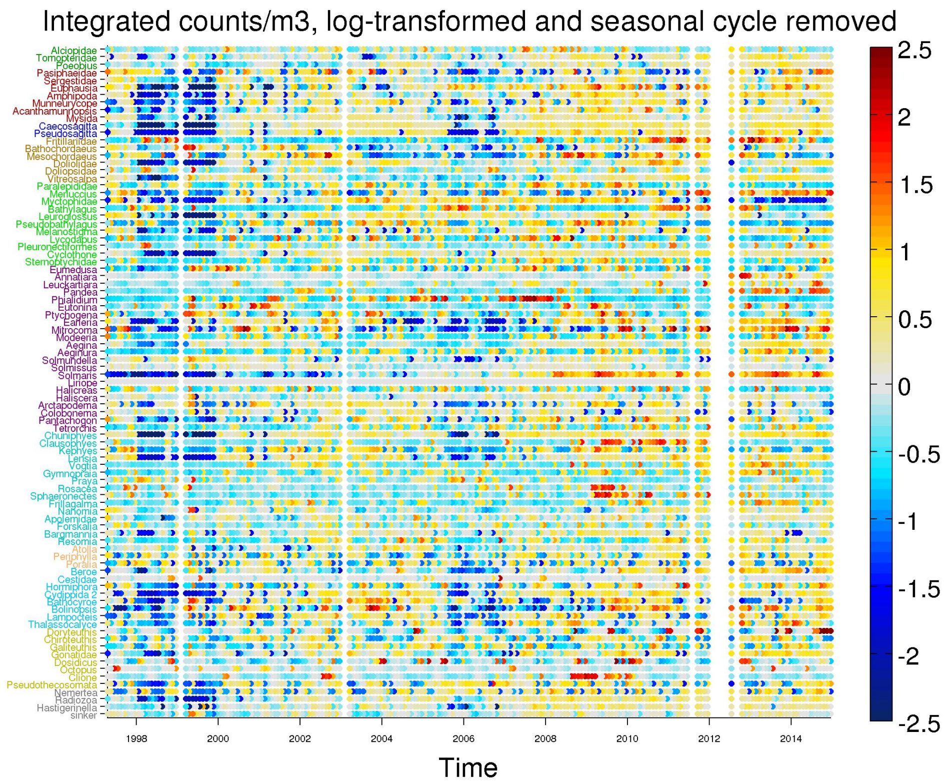
Here 89 data series as inputs, and only keeping the “dominant” modes (highest variance explained) since the goal is to reduce the information to a few time series.

Also modes need to be “statistically separable” which becomes an issues for non-dominant modes.

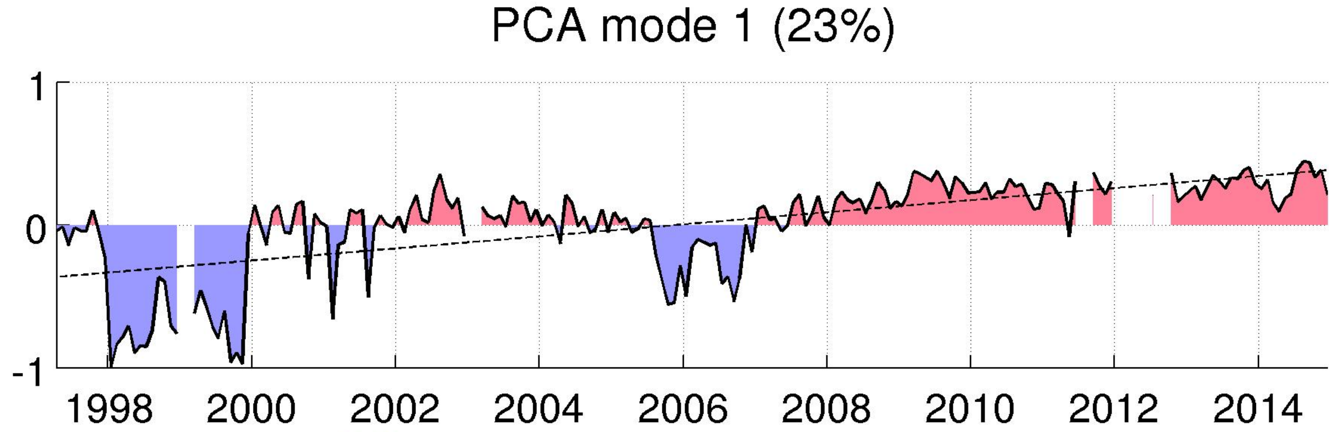


Midwater PCA mode 1

concept: find the one time series that retains the most information from all of the input data series (PC1)

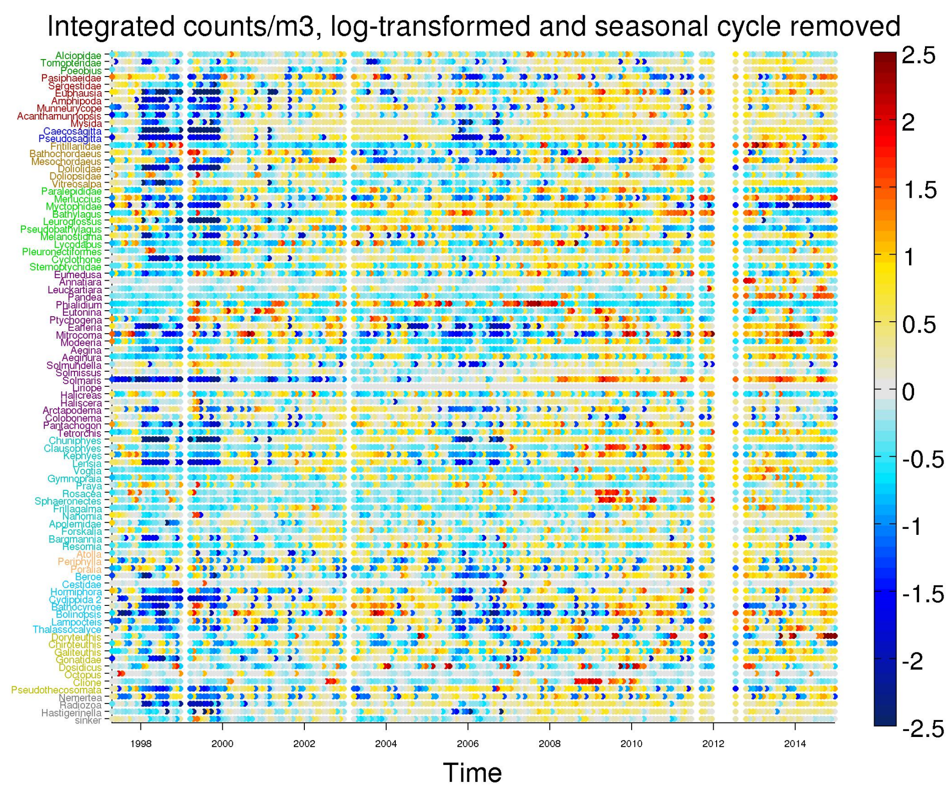


Midwater PCA mode 1

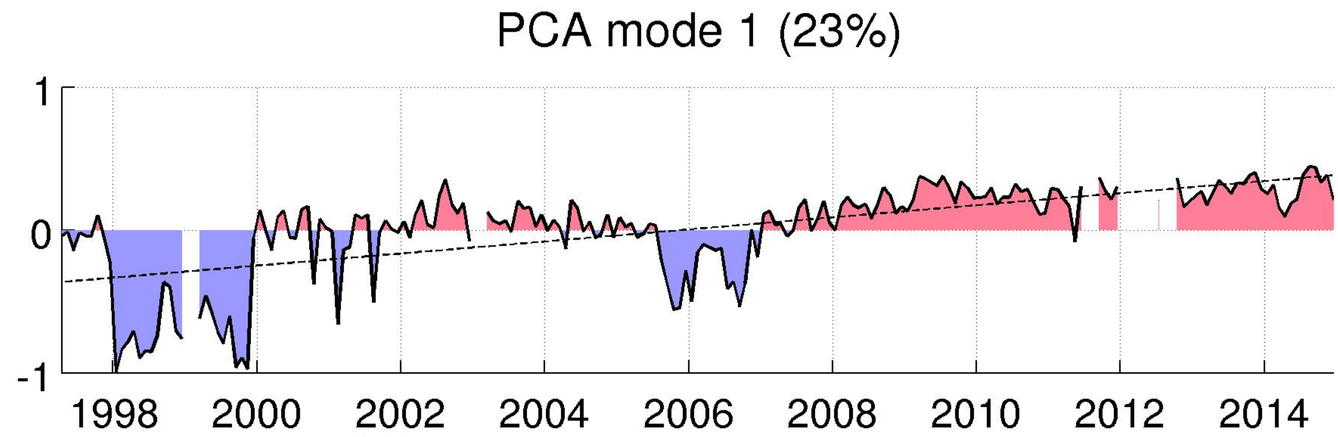


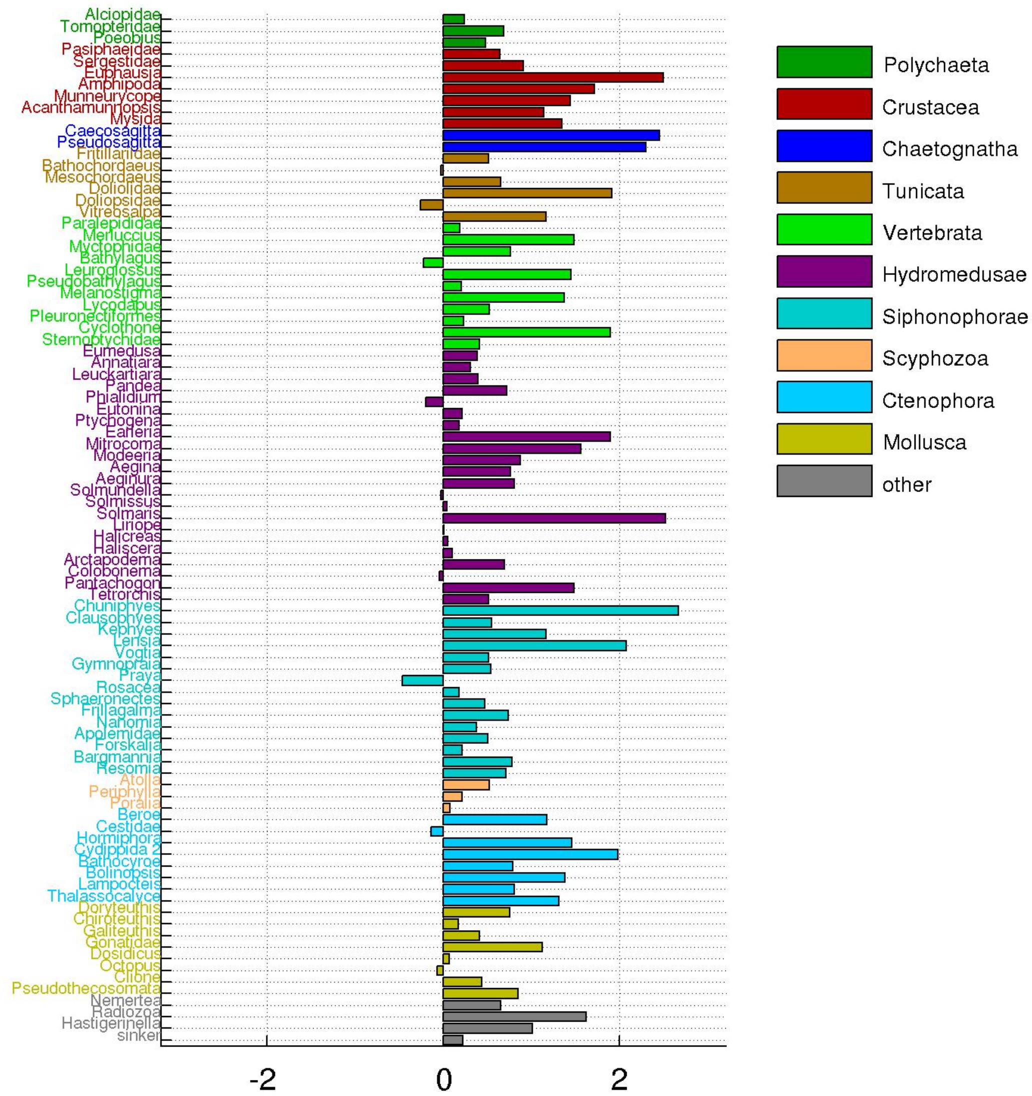
concept: find the one time series that retains the most information from all of the input data series (PC1)

→ PC1 is the time series that captures most of the variance in all data series



Midwater PCA mode 1

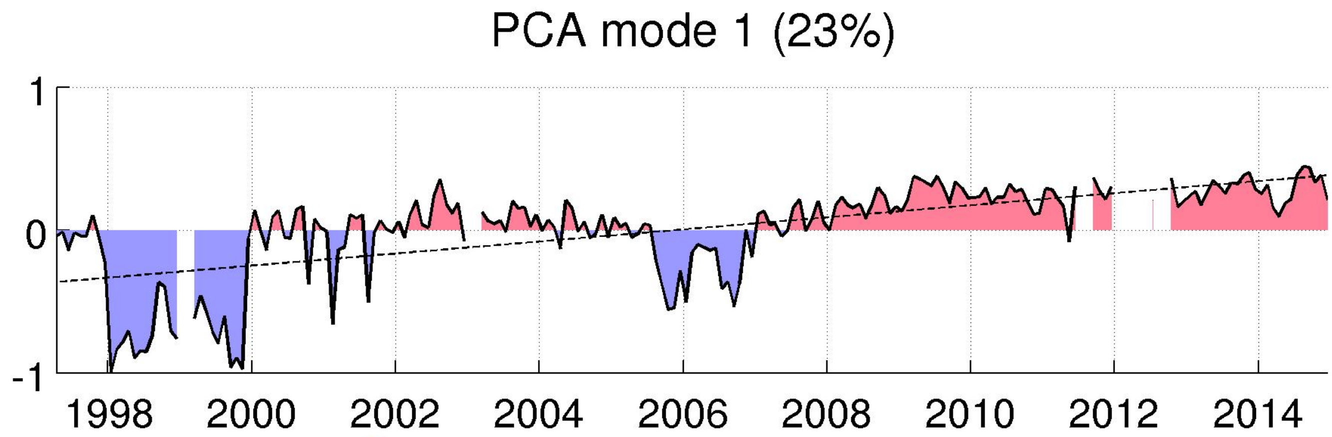


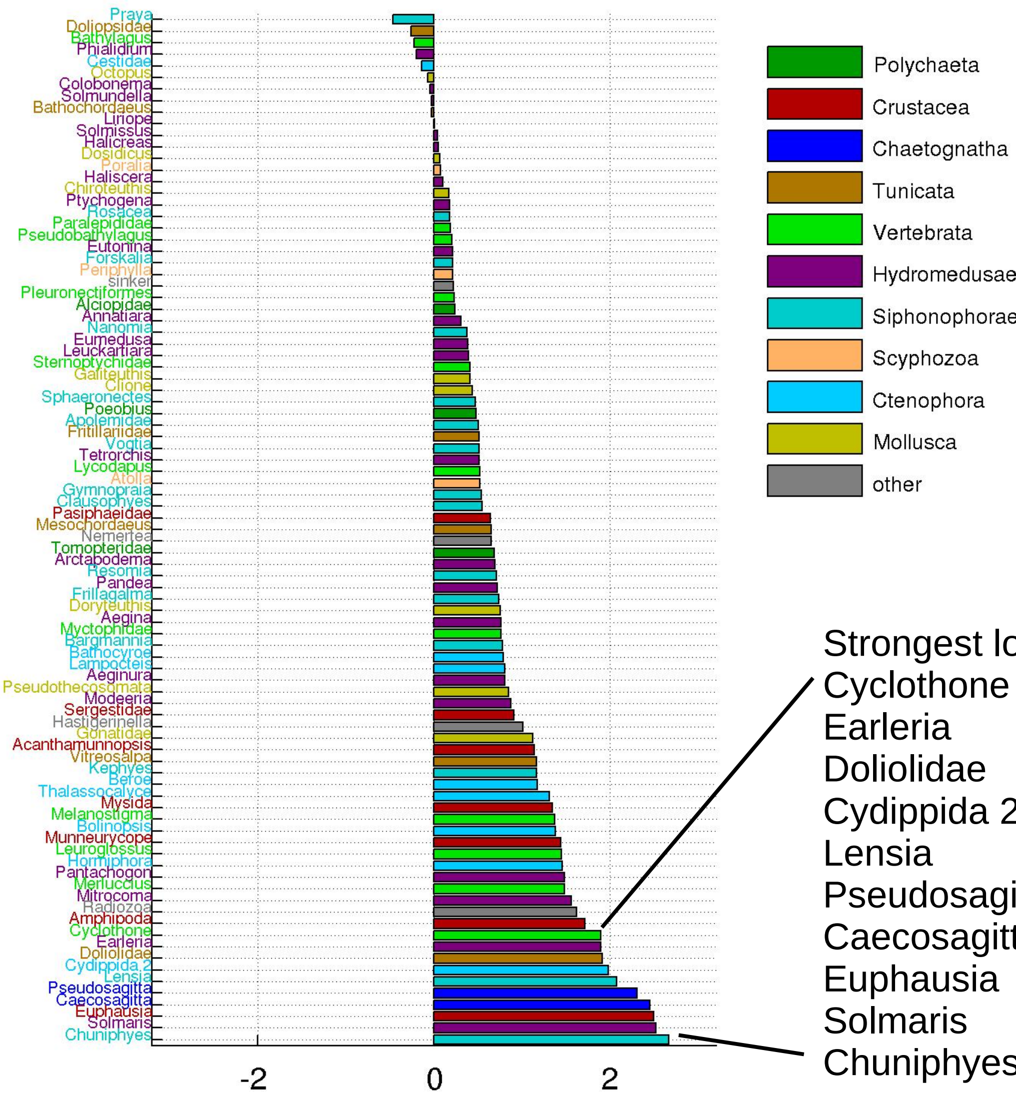


→ PC1 is the time series that captures most of the variance in all data series
→ loadings are taxon-specific and show how strongly PC1 captures each taxon's variability

This mode captures synchronous variations among taxa – taxa with high loadings have a temporal variability similar to PC1.

Midwater PCA mode 1



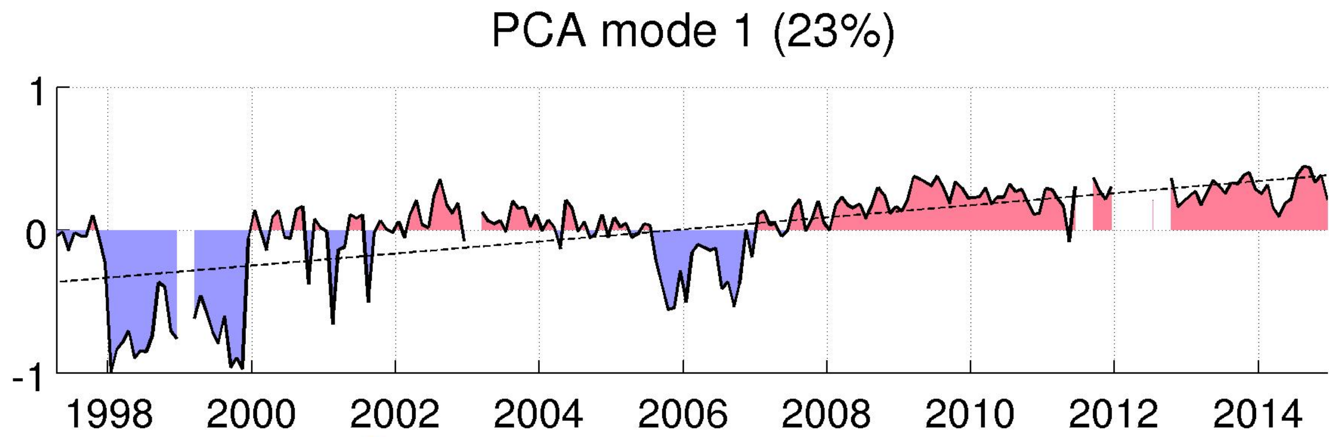


Strongest loadings:
Cyclothone
Earleria
Doliolidae
Cydippida 2
Lensia
Pseudosagitta
Caecosagitta
Euphausia
Solmaris
Chuniphyes

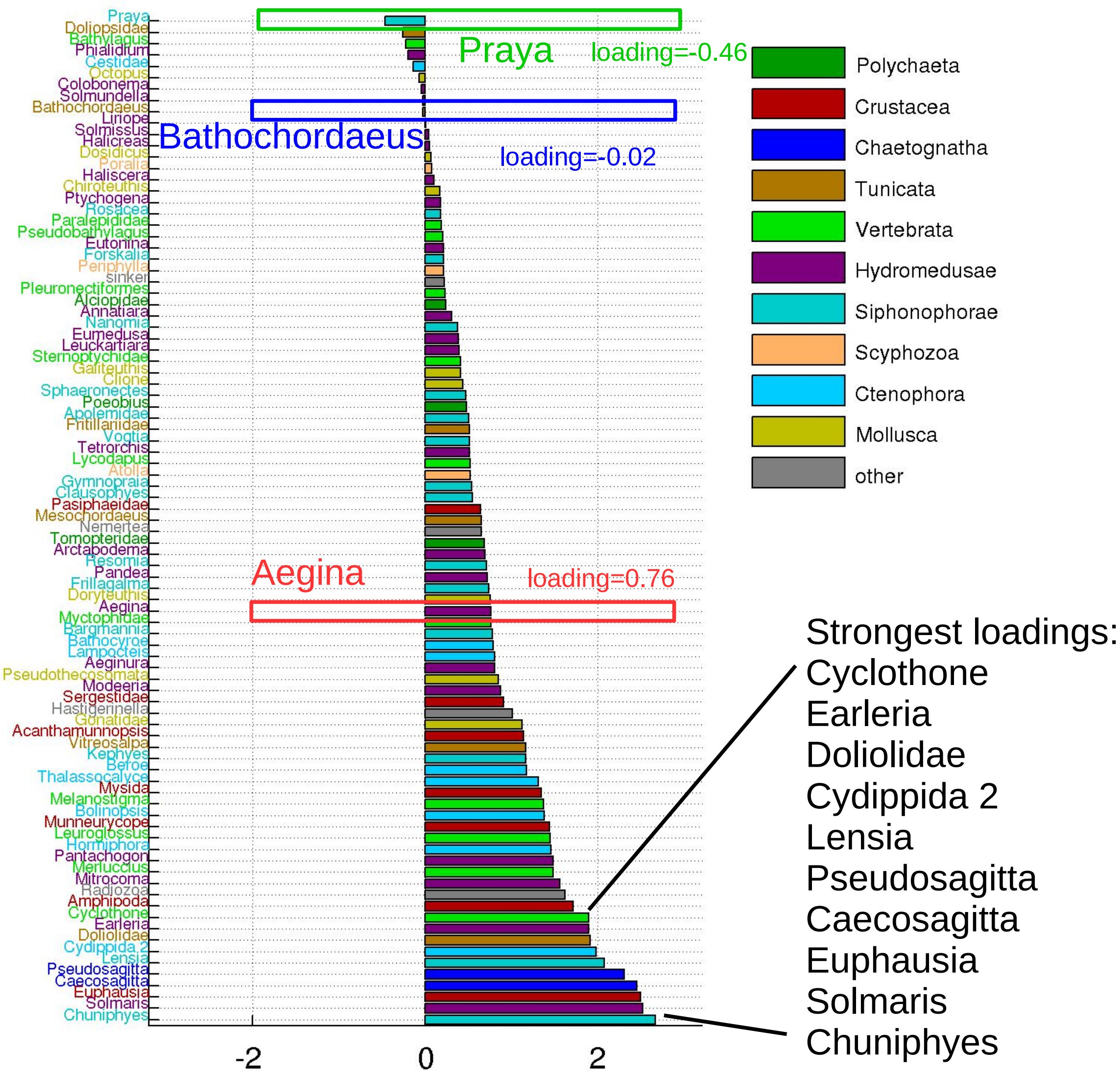
→ PC1 is the time series that captures most of the variance in all data series
→ loadings are taxon-specific and show how strongly PC1 captures each taxon's variability

This mode captures synchronous variations among taxa – taxa with high loadings have a temporal variability similar to PC1.

Midwater PCA mode 1



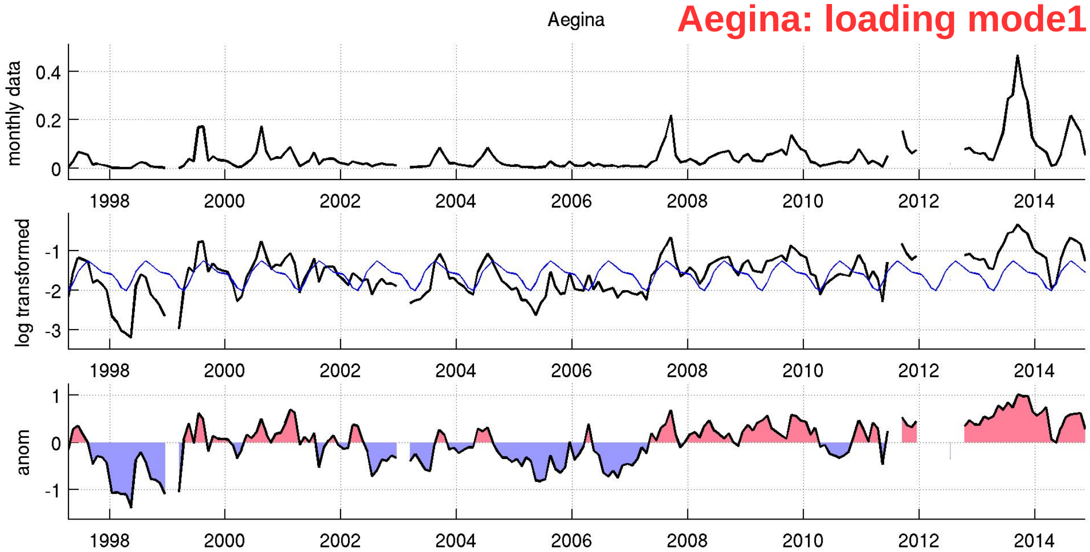
→ PC1 is the time series that captures most of the variance in all data series
→ loadings are taxon-specific and show how strongly PC1 captures each taxon's variability

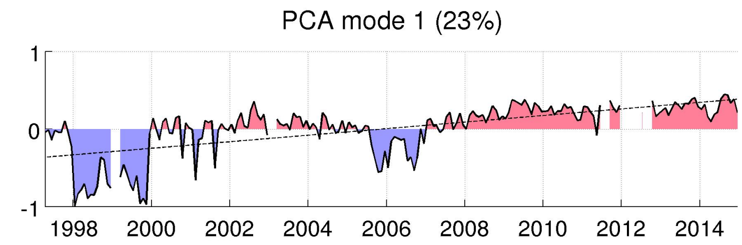


This mode captures synchronous variations among taxa – taxa with high loadings have a temporal variability similar to PC1.

What do modes represent? …back to the data

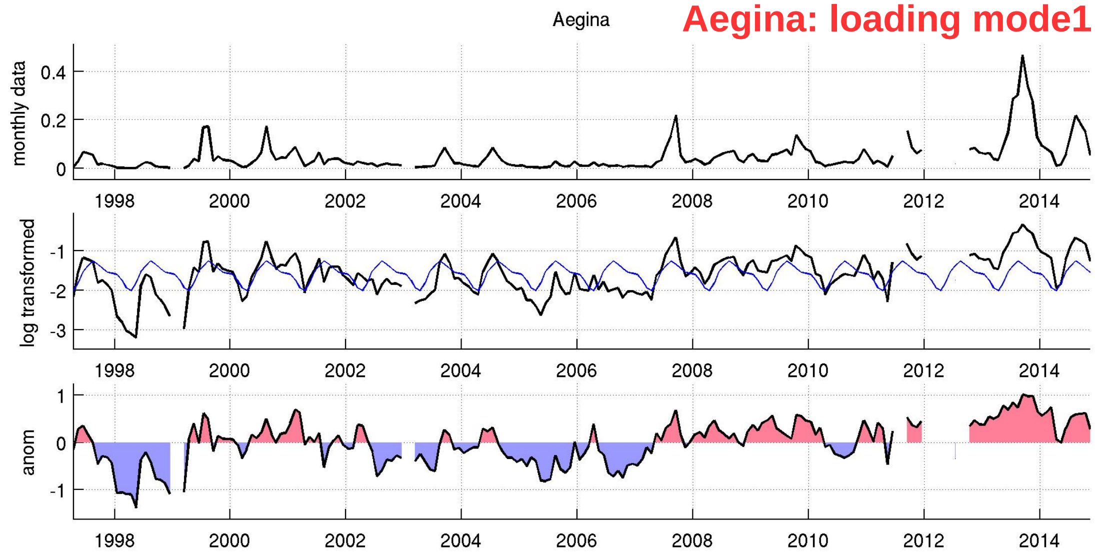
Aegina: loading mode1 = 0.76

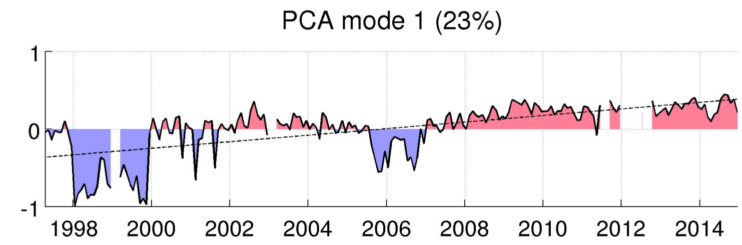




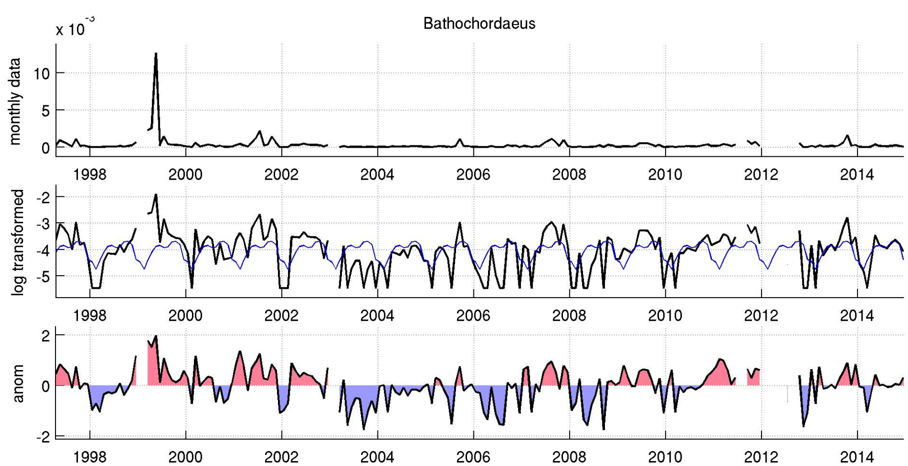
What do modes represent?
…back to the data

Aegina: loading mode1 = 0.76

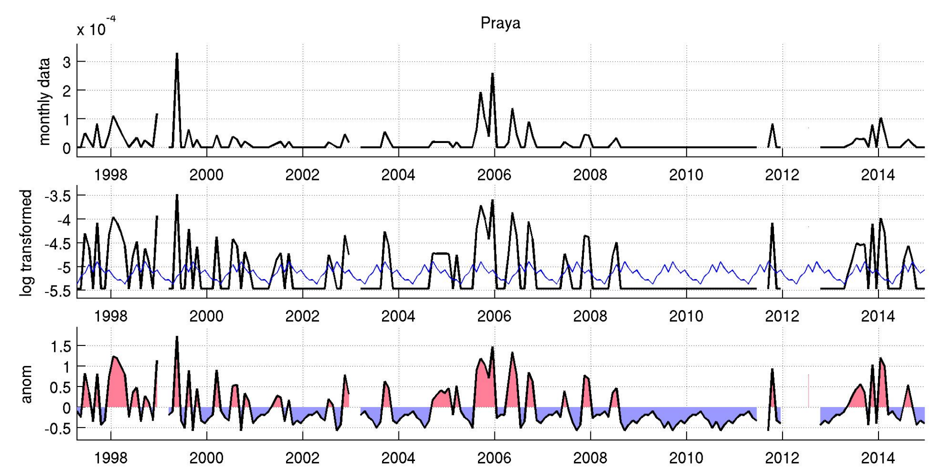




Bathochordaeus:loading mode1 ~ 0



Praya: loading mode1 = -0.46



EOFs (Empirical Orthogonal Functions)

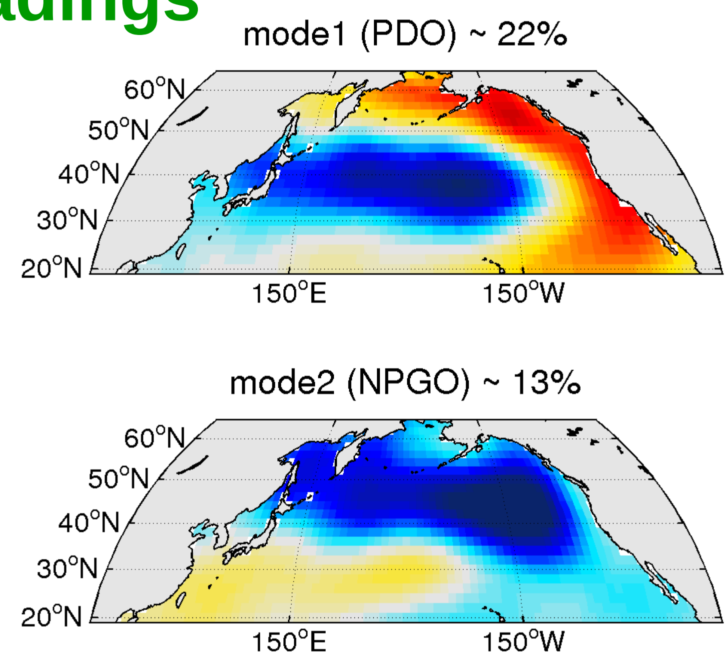
Simply another name for PCA... applied to spatial datasets.

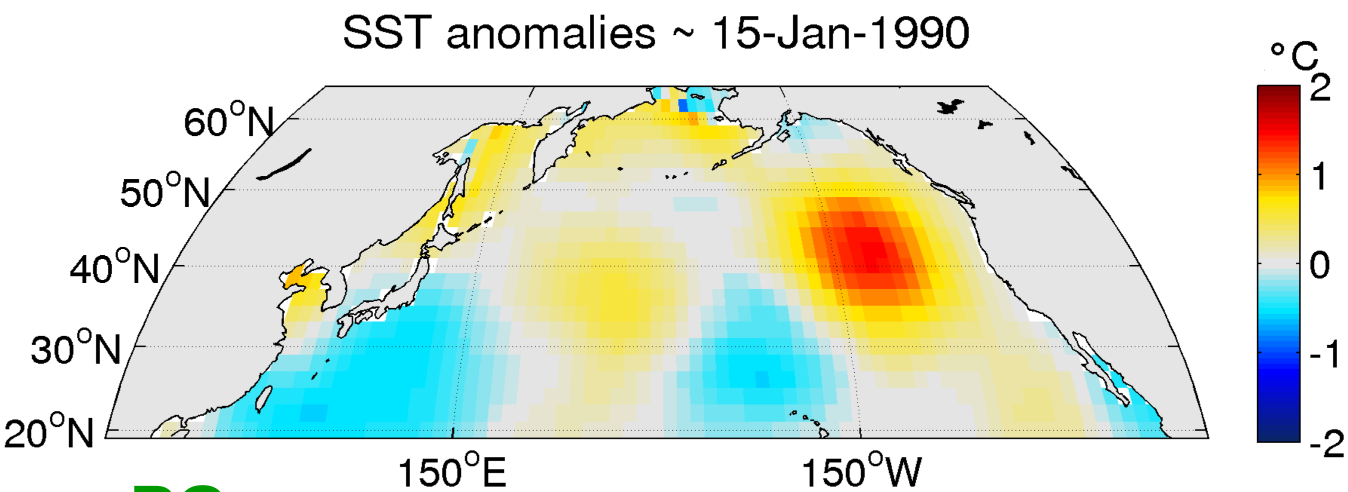
	PCA	EOFs
variable	counts, biomass...	SST, Chl...
dimensions	species, groups	pixels (lon/lat)
output (PC)	time series	time series
output (loading)	loadings for each species	loadings for each pixel = map

Example: PDO & NPGO

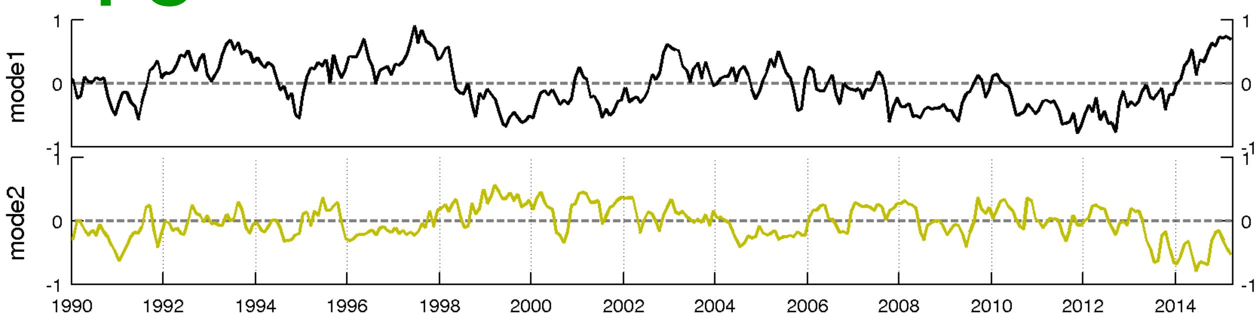
	PCA	EOFs
variable	counts, biomass...	SST, Chl...
dimensions	species, groups	pixels (lon/lat)
output (PC)	time series	time series
output (loading)	loadings for each species	loadings for each pixel = map

loadings





PC



PDO = Pacific Decadal Oscillation
NPGO = North Pacific Gyre Oscillation

Issues with PCA

- Very sensitive to inputs & preprocessing. In particular, the variance of the input data series conditions their relative *weight*.
- Captures *linear* relationships between variables.
- By definition all modes are *orthogonal*, whereas natural modes of variability can be correlated.
- Adding more data points changes the result.
- It is a statistical method that can capture real phenomena as well as *noise*.

BUT remains a very powerful tool to **synthesize** a huge number of time series into a few meaningful ones.

Next time...

- Clustering and ordination: find which taxa / which years are similar to each other
- PCA as an ordination method
- Multidimensional scaling (MDS) vs. PCA
- What is the question?
  - generate time series describing ecosystem changes?
  - or only identify “types” of ecosystems and associated time periods?