Sparse Principal Component Analysis as a tool for the exploration of heterogeneous datasets from multidisciplinary field experiments

Sebastian Landwehr, Michele Volpi,

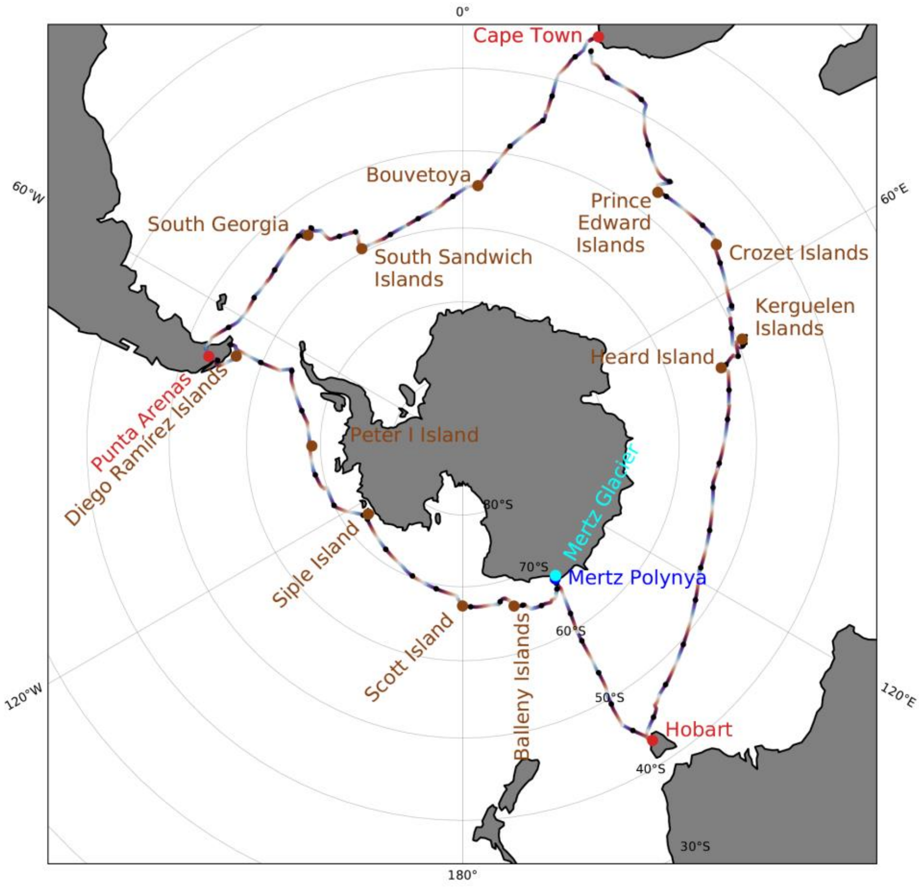
Alexander Haumann, Charlotte M. Robinson, Heather Forrer, Iris Thurnherr, Andrea Baccarini, Valerio Ferracci, Jen Thomas, Yajuan Lin, Nicolas Cassar, Alberto Alberello, Rafel Simo, Irina Gorodetskaya, Silvia Henning, Christian Tatzelt, Alessandro Toffoli, Gang Chen, Moallemi Alireza, Rob Modini, Ruth Airs, Pablo Rodríguez-Ros, Neil Harris, Franziska Aemisseger, Heini Wernli,

Fernando Pérez-Cruz, Julia Schmale

Data Science in Climate and Climate Impact Research

Virtual Workshop at ETH Zurich – 21.08.2020

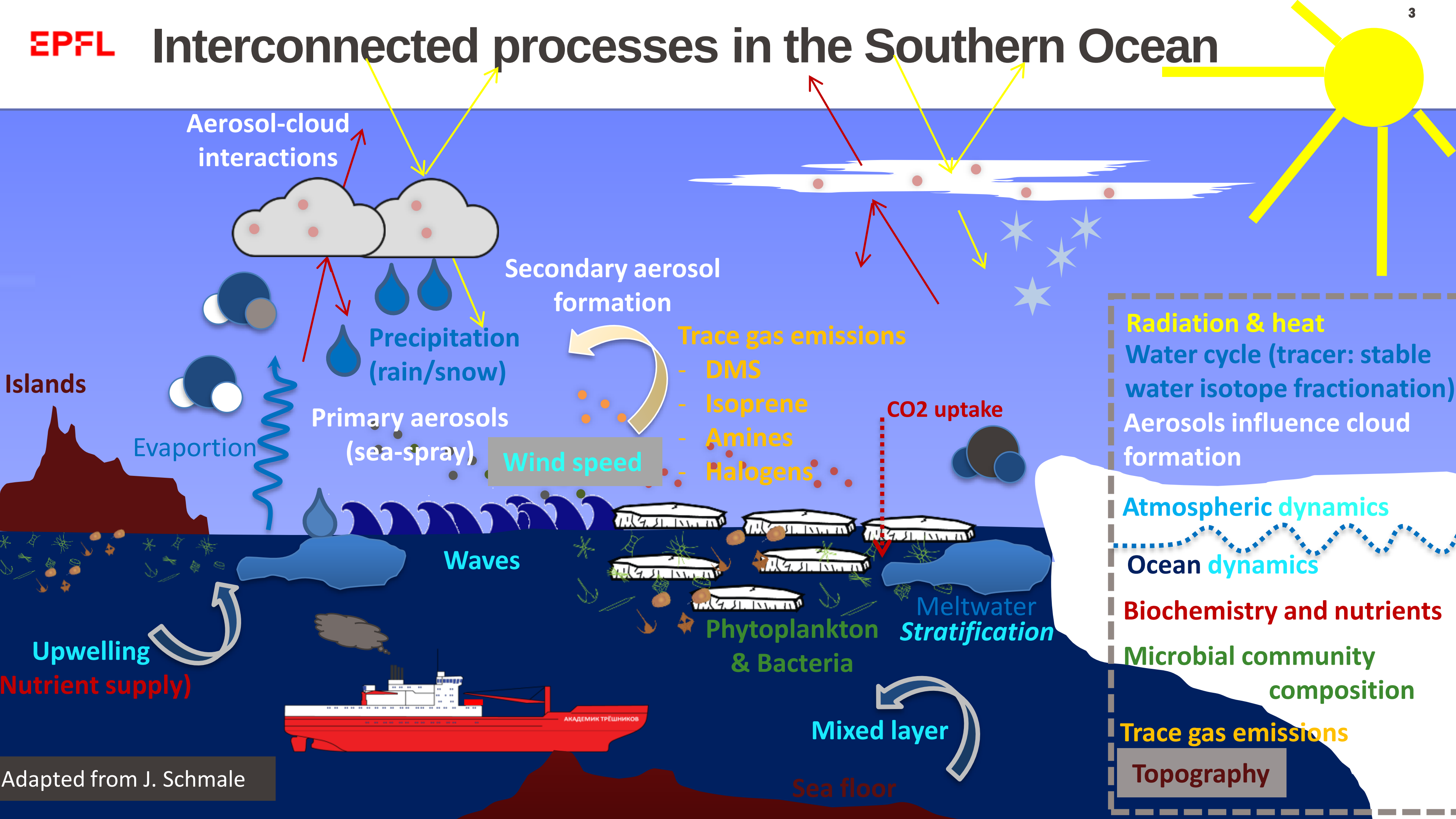
The Antarctic Circumnavigation Expedition



- 90 days & 33,565 km around Antarctica
- 22 interdisciplinary projects
- 148 scientists
- 73 institutions
- 23 countries
- 11 islands and 1 Glacier
  - 96 CTD stations
  - 3600 events
  - 27500 samples

EPFL

Interconnected processes in the Southern Ocean



Adapted from J. Schmale

A new approach to research cruise data

The classical approach:

- Research questions inspired by prior knowledge
- Study relations between a handful of variables (one or few processes)
- Constrain and advance models
- ***Unexpected relations might be overlooked***
  - We can not test all possible parameter combinations

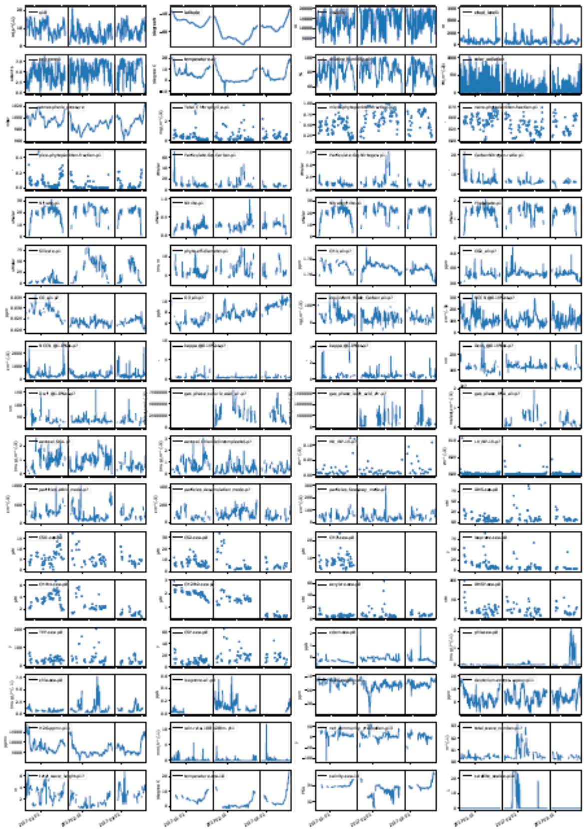
Our approach (data-driven):

- How can we assure not to overlook something interesting that we did not expect?
- Can we dump all our data into an algorithm and get an *unbiased* representation of their relations?

The cruel realities of research cruise data

- n=118 variables (ocean and atmosphere)
  - Physical and dynamical properties
  - Trace gases and isotope composition
  - Bacteria & pytoplancton
  - Nutrients
- At different time resolutions
  - Seconds to days (water/filter samples)
- Missing data
  - Instrument downtime
  - Pollution (the ship`s exhaust plume)

Time series of observed variables

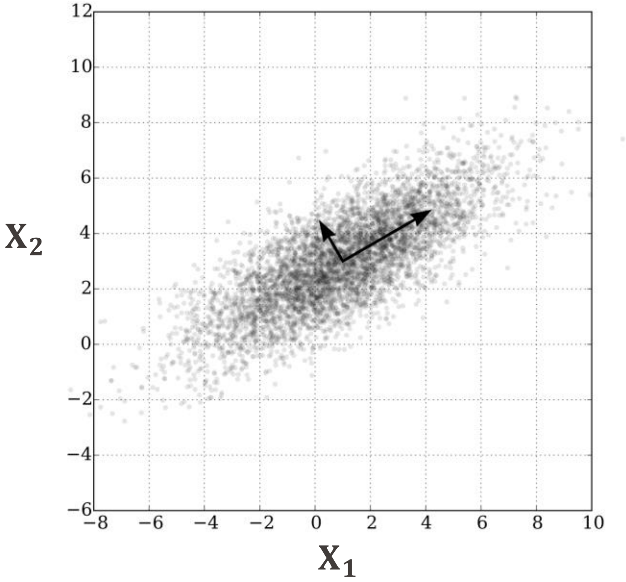


Principle Component Analysis in a nutshell

- Reduce the n-dimensional space of the observations $\mathbf{X_j}(\mathrm{t}) : (\mathbf{1} < \mathbf{j} < \mathbf{n})$ into a lower dimension $\mathbf{k} < \mathbf{n}$
- Find a linear transformation

$$\mathbf{Z_i} = \mathbf{X_j}\ \mathbf{W_{i,j}} : (\mathbf{1} < \mathbf{i} < \mathbf{k})$$

- The reconstruction $\mathbf{X_j} \cong \mathbf{Z_i}\ \boldsymbol{W^T_{i,j}}$
  - maximizes the variance
  - minimizes the reconstruction error



$$\min_{\|\mathbf{w}\|=1} \frac{1}{n} \|\mathbf{X} - \mathbf{X}\mathbf{w}\mathbf{w}^\top\|$$

- $\mathbf{Z_i}$: *Latent variables*
- $\mathbf{W_{i,j}}$: *Loading vectors*
- $\mathbf{k}$: Hyper parameter

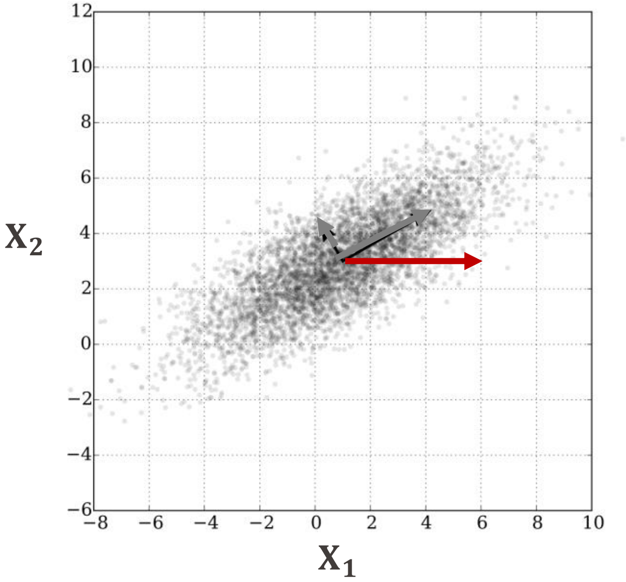
Sparse Principle Component Analysis in a nutshell

PCA:

- Entries of the $\mathbf{W_{i,j}}$ typically non zero
- All $\mathbf{Z_i}$ needed to reconstruct a certain $\mathbf{X_j}$
- Difficult to interpret the results

Sparse PCA:

- Penalizes non-zero weights while still maximizing the variance
- Lower reconstruction accuracy (L2-loss)
- Only few $\mathbf{X_j}$ contribute to each $\mathbf{Z_i}$
- Only few $\mathbf{Z_i}$ needed to reconstruct $\mathbf{X_j}$
- **Easier to interpret**

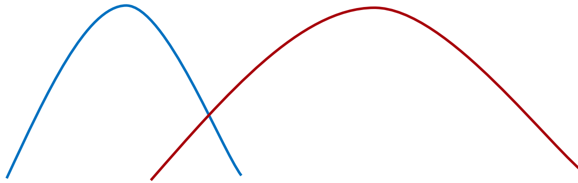


$$\min_{\|\mathbf{v}\|=1} \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{w}\mathbf{v}^\top\|_2 + \alpha\|\mathbf{w}\|_2 + \lambda\|\mathbf{w}\|_1$$

Why sparse Principle Component Analysis for ACE?

- The ACE cruise covered a wide range of environments and weather conditions
- The correlations between the **observed variables** change over time
- Can we describe these changes with a few **latent variables**?
- **We want interpretable results that link a limited number of observed variables**

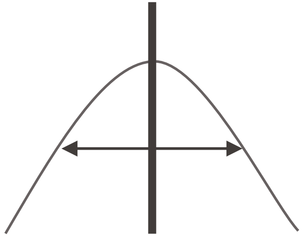
Input data preprocessing



1) Resample to 3 hour resolution -> 730 samples

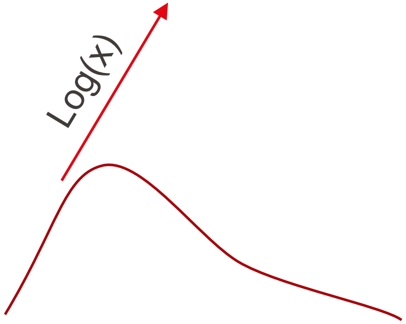
  Averaging if original resolution is higher

  Select nearest point if resolution is lower



3) Normalize to zero mean and unit variance

2) Logtransform if the distribution is *closer to beeing* lognormal than normal distributed



Imputation of missing data

- A priori replace missing data with the global mean
- Use initial sPCA solution to reconstruct the missing data
- … Iterate …

- The reconstruction converges if only a few active variables are missing at timestep t_k
- Otherwise the LV-activation $Z_i(t_k)$ remains close to zero
  - (For each Z_i we exclude poorly covered times from further analysis)

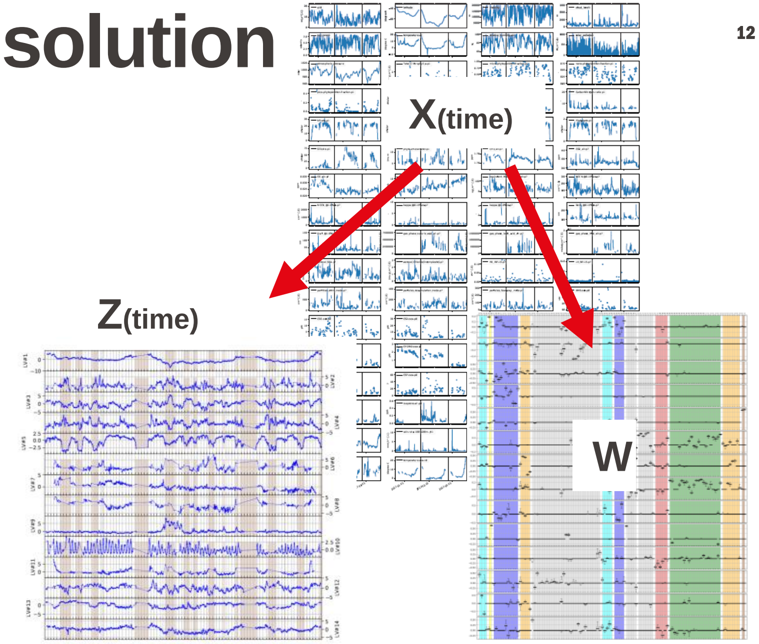
How robust is the sPCA solution?

Our bootstrapping approach:

- Run sPCA on the full data set (Master)
- Run sPCA on 20 random sub-samples with 75% of the available data
- Match the weight vectors (W_i) and time-series (Z_i) of the sub-sample solutions to the W_i & (Z_i) of the master solution (most but not all bootstrap LVs match the master)
- Use the median of the bootstrap weight vectors ($\langle W_i \rangle$) to calculate the LV activation time-series (Z_i)
- We can use the ratio of $\langle W_i \rangle$ to the median absolute deviation σW_i as measure of significance

Interpretation of the sPCA solution by domain scientists

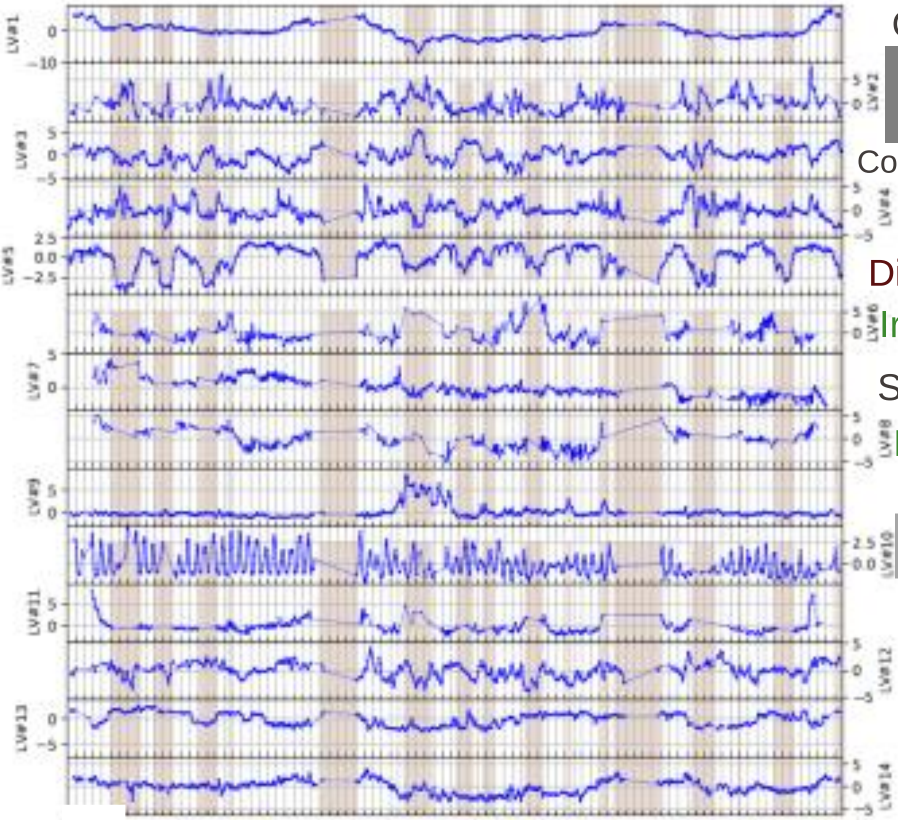
- Discussion of the results in a workshop
- We quickly started to associate the Latent Variables with real world processes
  - Based on the variable composition W_i
  - Based on the activations Z_i when plotted as time series or on the map
- This started a vivid exchange between the participating researchers
- How can we prove our interpretations?
- Some of the parameter combinations where surprising!



EPFL

Results overview

LV activation time series



Time

Climatic zones

Aerosols (condensed gases)

Cold/warm air advection

Precipitation

Distance to land

Iron fertilized blooms

Seasonal signal

Microbial productivity

Sea ice

Solar cycle

Surface nutrient availability

Sea spray

?

Atkin mode particles

Atmosphere

Ocean

Water cycle

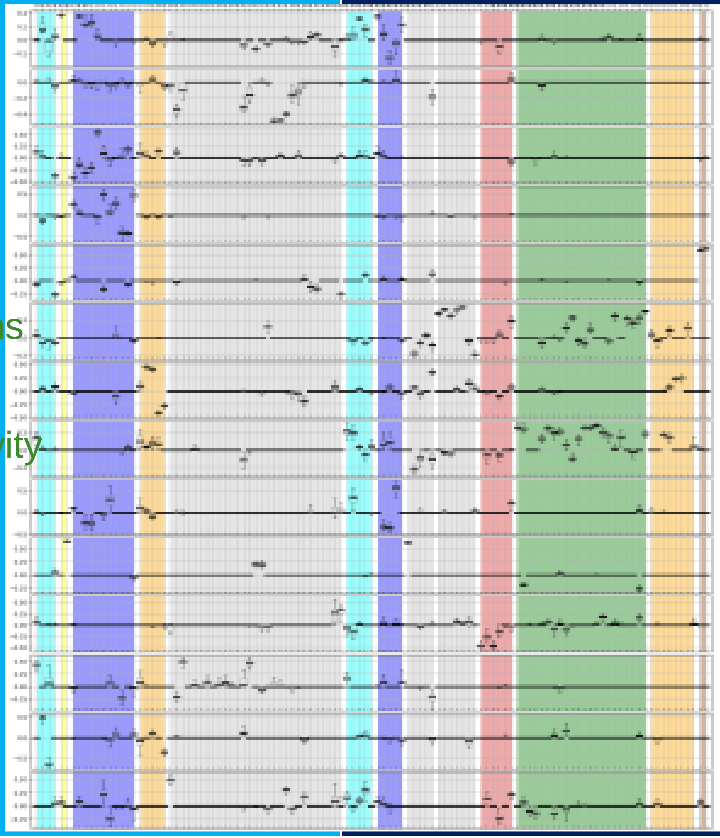
Dynamics

Trace

Chemistry

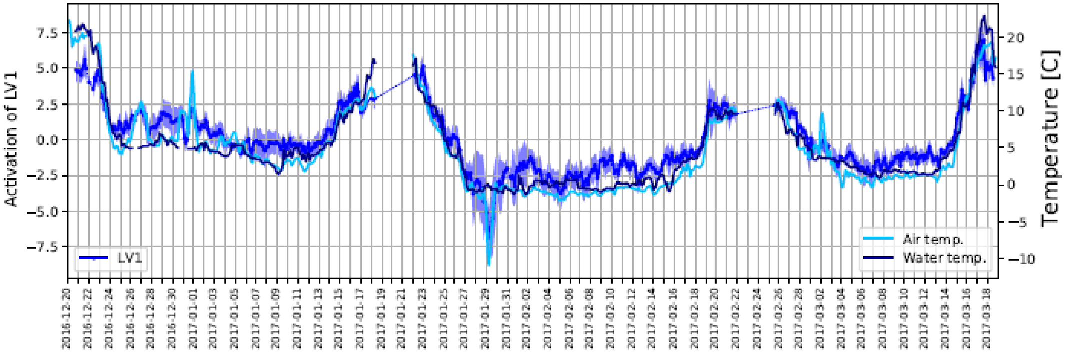
Topography

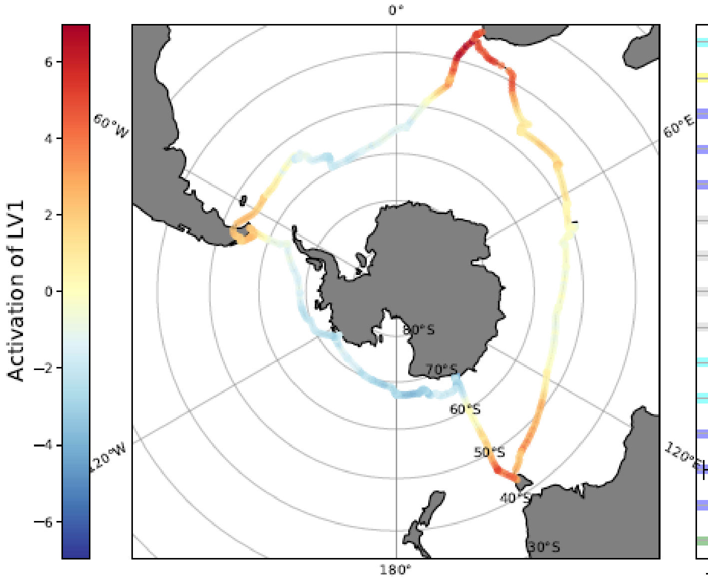
Light gases Aerosols / particulate Biology

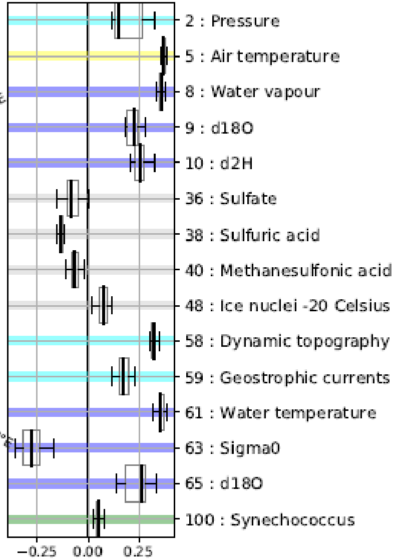


Observed variables

LV1 - Climatic Zones

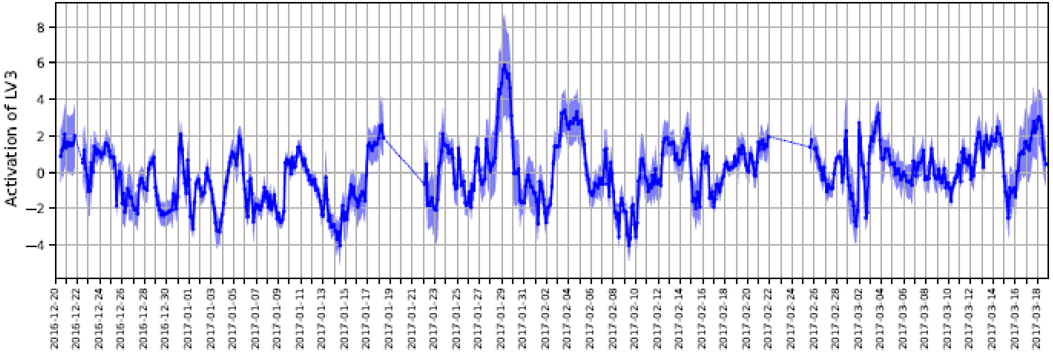


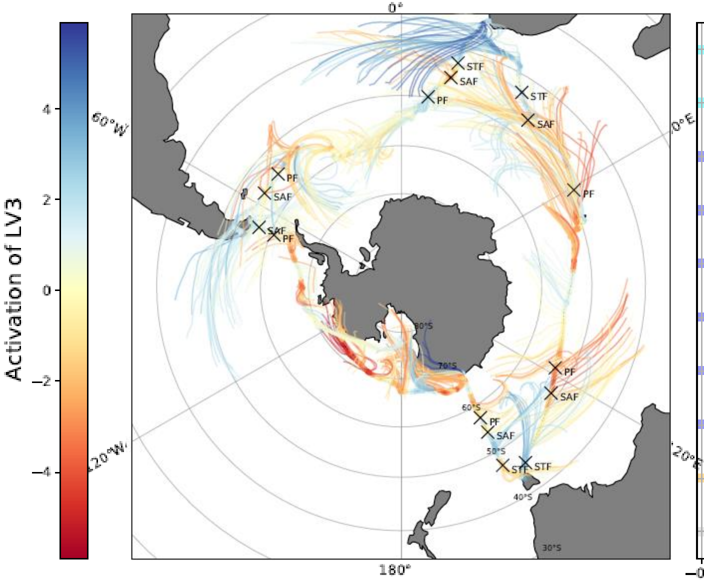


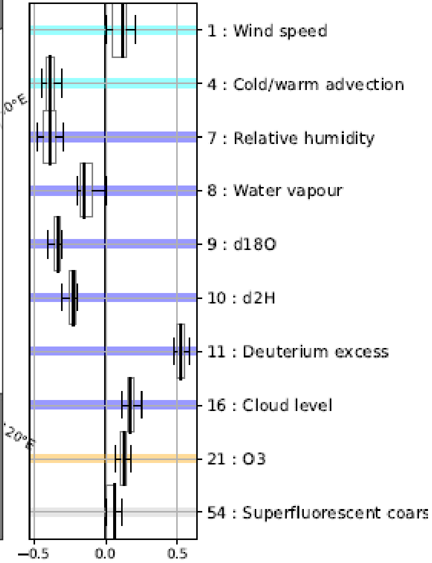


- Explains 10% of the total variability
  - (All LV together explain 60%)
- Depicts the effect of the ships locations (warmer air/water further north)

LV3 - Meridional cold and warm air advection

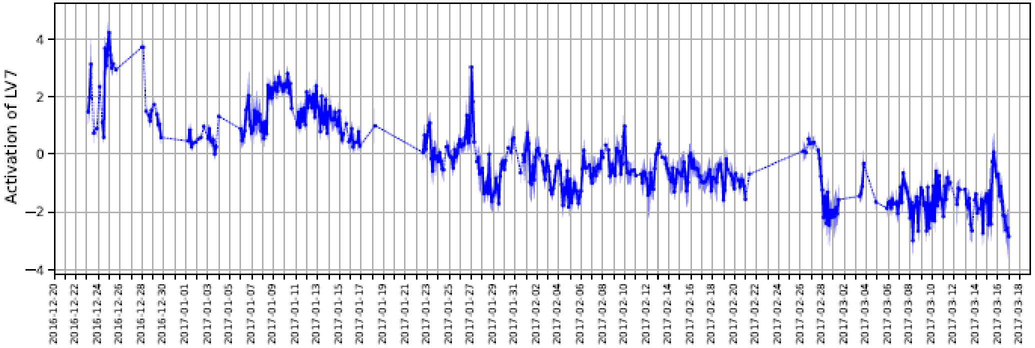


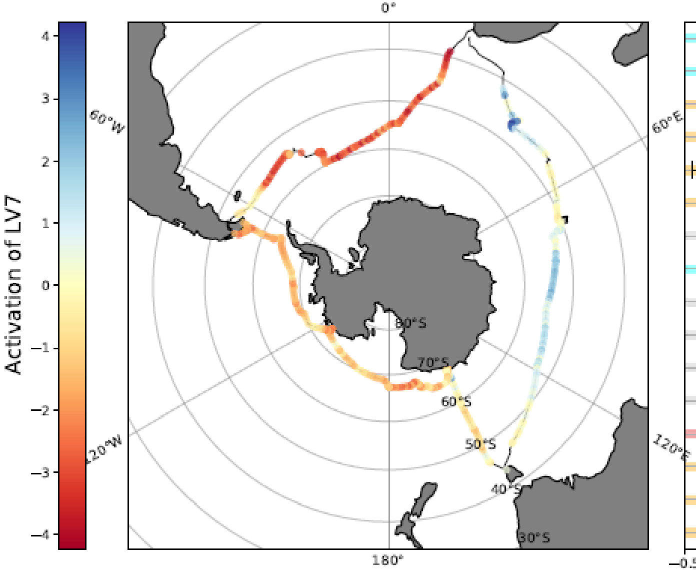


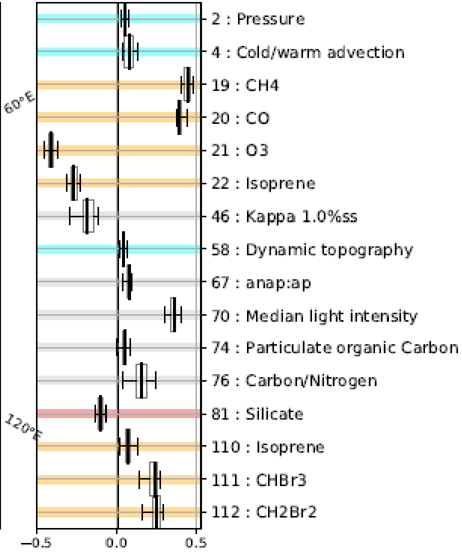


- Relates mostly to the air mass origin
  - Warm northerly air passing over colder ocean
  - Cold southerly air passing over warmer ocean
- Most relevant for the water cycle

LV7 – Seasonal signal



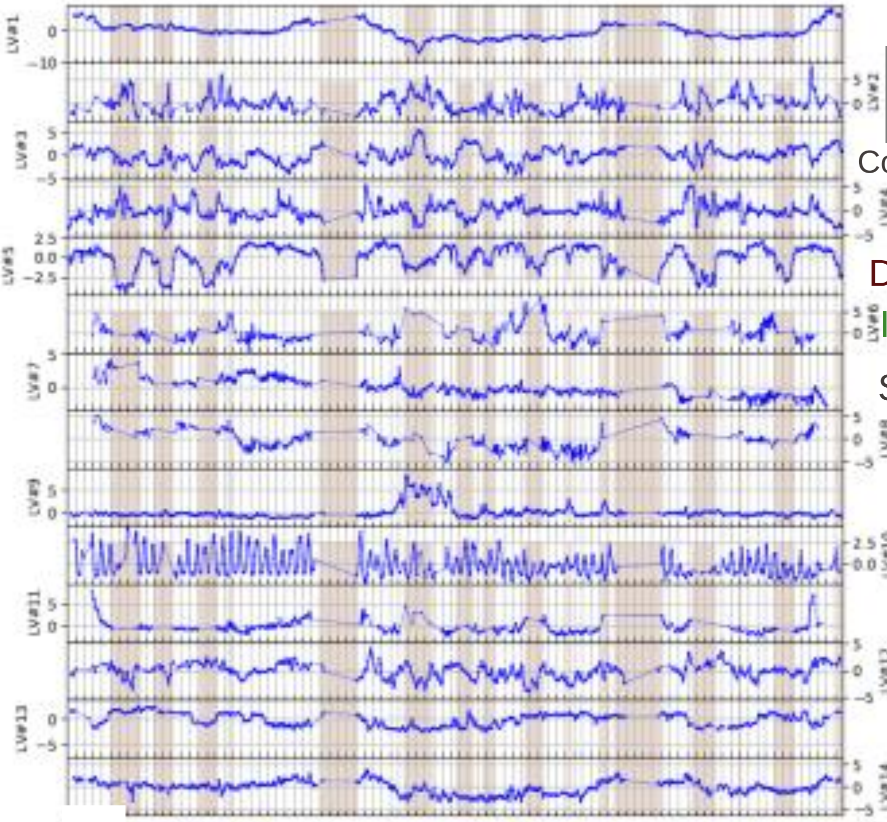




- Depicts the ending summer (ACE took place December to March)
  - Reduction in median light intensity in the ocean mixed layer and reduced bio activity:
  - Lower concentration of dissolved halogenated trace gases (CH_2Br_2, $CHBr_3$) and Isoprene
  - Why does atmospheric Isopren react differently?
- Change in atmospheric oxydation capacity
  - Increasing O3 (Ozon)
  - Decreasing CO (Carbon monoxid) and CH4 (Methane)

Results overview

LV activation time series



Time

Climatic zones

Aerosols (condensed gases)

Cold/warm air advection

Precipitation

Distance to land

Iron fertilized blooms

Seasonal signal

Microbial productivity

Sea ice

Solar cycle

Surface nutrient availability

Sea spray

?

Atkin mode particles

Atmosphere

Ocean

Water cycle

Dynamics

Trace

Chemistry

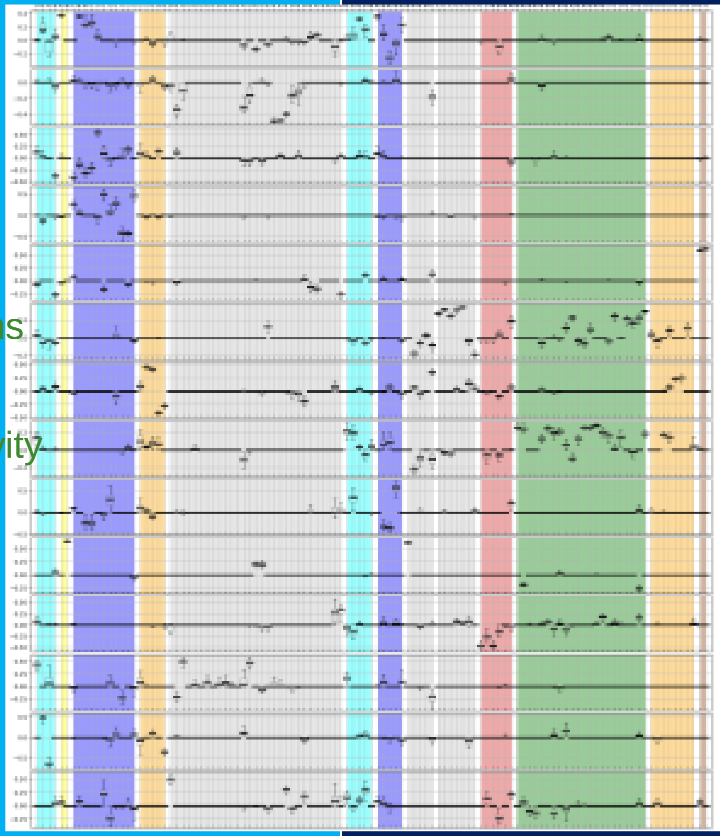
Topography

Light

gases

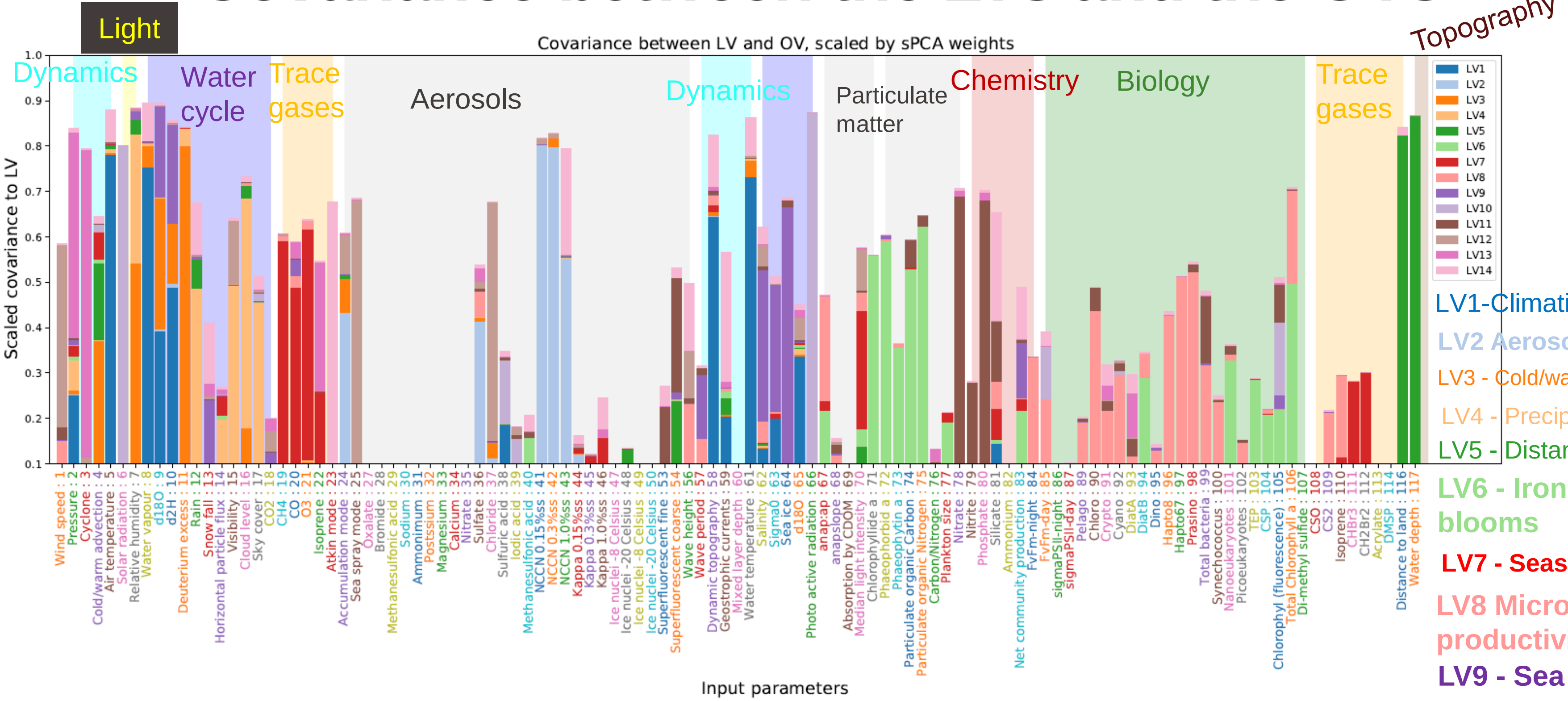
Aerosols / particulate

Biology



Observed variables

Covariance between the LVs and the OVs



LV1-Climatic zones

LV2 Aerosols (gases)

LV3 - Cold/warm advection

LV4 - Precipitation

LV5 - Distance to land

LV6 - Iron fertilized blooms

LV7 - Seasonal signal

LV8 Microbial productivity

LV9 - Sea ice

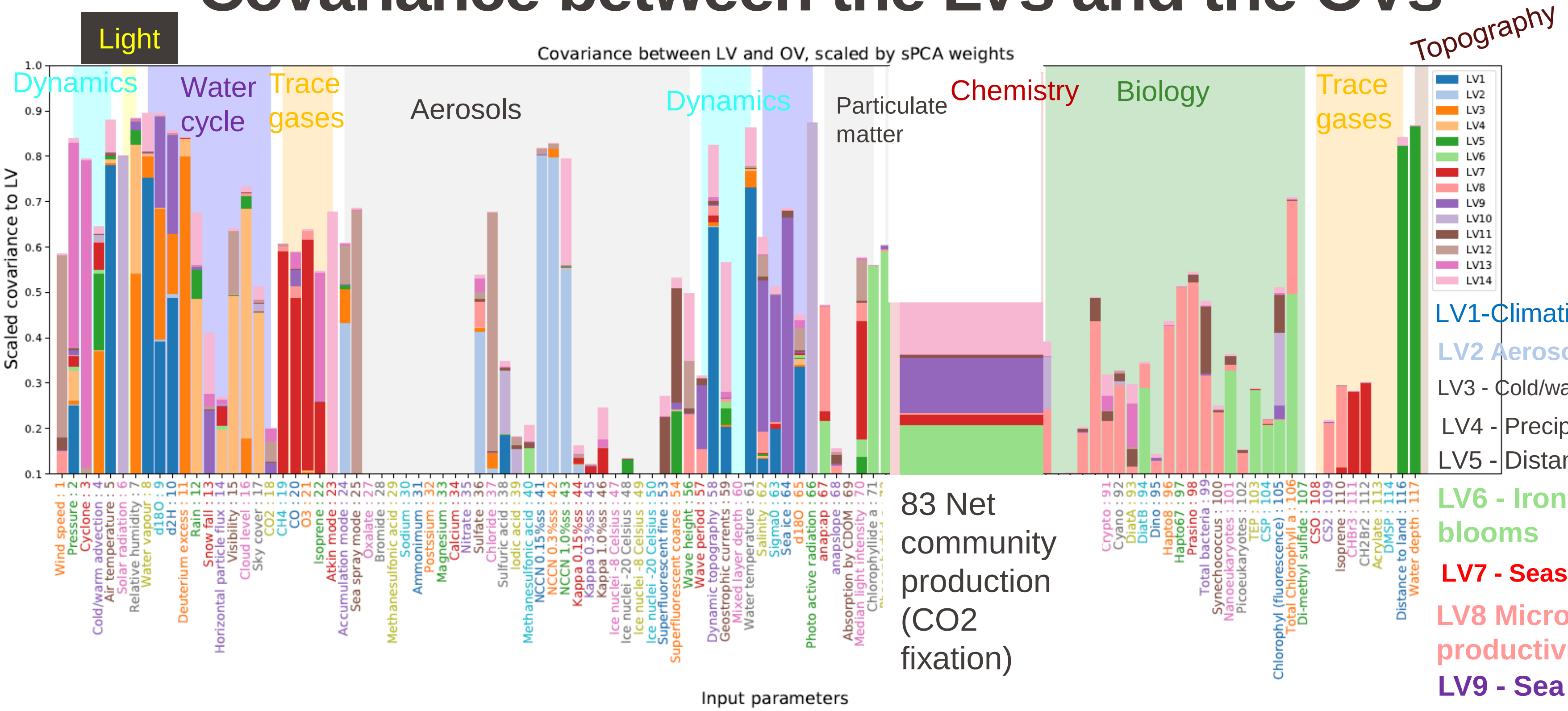
LV10 - Solar cycle

LV11 Surface nutrient availability

LV12 Sea spray

- The sPCA solution provides a mean to separate process influences
- Sparse time series are not well explained by the sPCA solution because they contribute only little variance

Covariance between the LVs and the OVs



LV1-Climatic zones

LV2 Aerosols (gases)

LV3 - Cold/warm advection

LV4 - Precipitation

LV5 - Distance to land

LV6 - Iron fertilized blooms

LV7 - Seasonal signal

LV8 Microbial productivity

LV9 - Sea ice

LV10 - Solar cycle

LV11 Surface nutrient availability

LV12 Sea spray

- The sPCA solution provides a mean to separate process influences
- Sparse time series are not well explained by the sPCA solution because they contribute only little variance

Conclusions & Future Work

Sparse Principle Component Analysis

- Provides a condensed representation of a diverse multi-variable data set
- Gives a decomposition of processes and their effect on the observed variance
- Is applicable to real world data
  - Sparsity of weight vectors → only few relevant parameters per process
  - Missing data reconstruction (there is room for improvement)
  - Significance test via bootstrapping of multiple sPCA solutions
- Limited to linear relations <=> fully traceable and easy to interpret
- Biased towards denser time series – nothing we can do about this
- **Is a useful tool for the interdisciplinary exploration of data sets from field experiments**
- Tool to find unexpected relations?

Developments that would improve the method:

- Accounting for temporal and spatial component (e.g. autoregressive)
- Introduction of non-linearity