# Statistical Distillation of the Earth System Data Cube

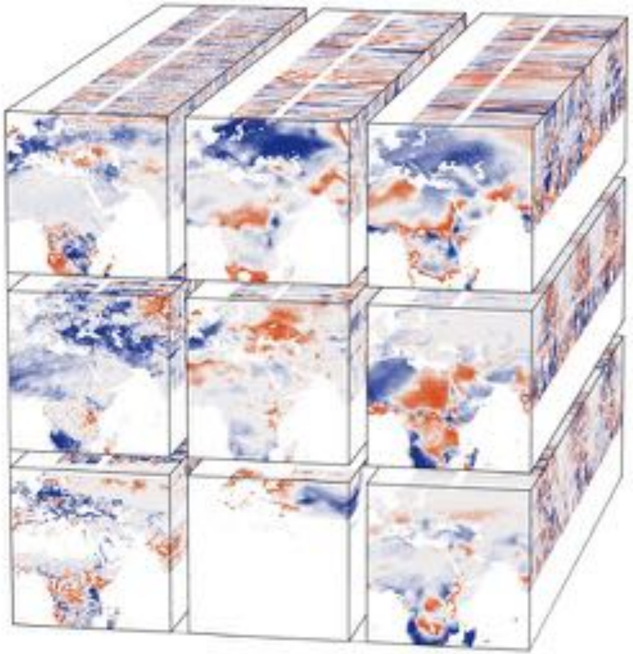
**Gustau Camps-Valls**[1], E. Johnson[1],V. Laparra[1], D. Bueso[1], M. Piles[1]
G. Brandt[2], N. Fomferra[2], H. Permana[2], M. Mahecha[3]

[1]Image Processing Lab (IPL) - Univ. València - **http://isp.uv.es**
[2]Brockmann Consult
[3]MPI Biogeochemistry - Jena

# The Earth as a mathematical object

# Earth System Data Cube (ESDC)

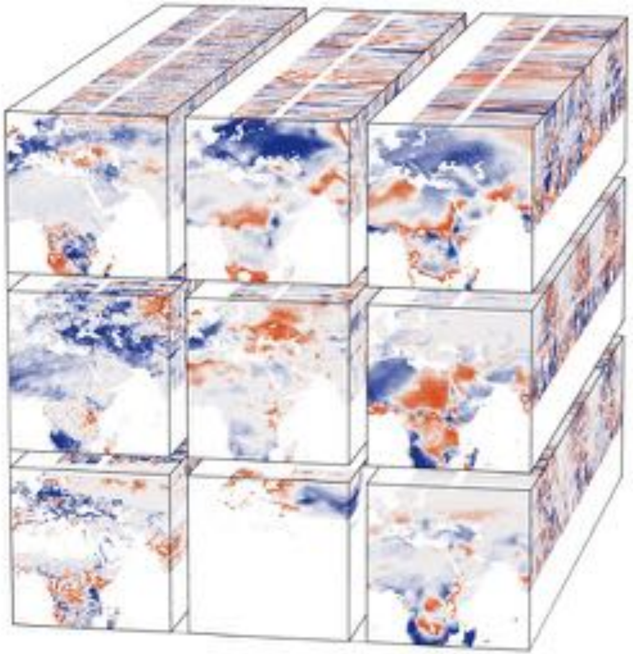
- Global extent
- Nested spatial grids (0.083°, 0.25°)
- Convenient aggregation (e.g. national levels)
- Consistent temporal sampling (8-daily, 2001-2011)
- Many variables: T, W, GPP, Precip, ET, ...
- Priority on the ESA data suite



Earth System Data Cube (source: esdc.net)

# Earth System Data Cube (ESDC)

- Global extent
- Nested spatial grids (0.083º, 0.25º)
- Convenient aggregation (e.g. national levels)
- Consistent temporal sampling (8-daily, 2001-2011)
- Many variables: T, W, GPP, Precip, ET, ...
- Priority on the ESA data suite
- What's the information content of the cubes?
- Which are the interesting, entropic, variables?
- What are the most interesting time and space scales?



Earth System Data Cube (source: esdc.net)

# All is about estimating information, and densities ...

$$\sum_{x} p(x) \log p(x) \sum_{xy} p(x,y) \log p(y|x) \sum_{x} p(x) \log p(x) \sum_{xy} p(x,y) \log p(y|$$

$$\sum_{xy} p(x,y) \log p(y|x) \sum_{x} p(x) \log q(x) \int_{\mathcal{X}} p(x) \log(p(x)/q(x)) \mathrm{d}x$$

$$\sum_{x} p(x) \log p(x) \sum_{xy} p(x,y) \log p(y|x) \sum_{x} p(x) \log q(x) \sum_{x} p(x) \left( \frac{\partial p_\theta(x)}{\partial \theta} \right)^2$$

$$\sum_{x} p(x) \log p(x) \sum_{xy} p(x,y) \log p(y|x) \left( \frac{\partial p_\theta(x)}{\partial \theta} \right)^2 \sum_{x} p(x) \left( \frac{\partial p_\theta(x)}{\partial \theta} \right)^2$$

$$\int_{\mathcal{X}} p(x) \log(p(x)/q(x)) \mathrm{d}x \quad \sum_{x} p(x) \log p(x) \sum_{xy} p(x,y) \log p(y|x)$$

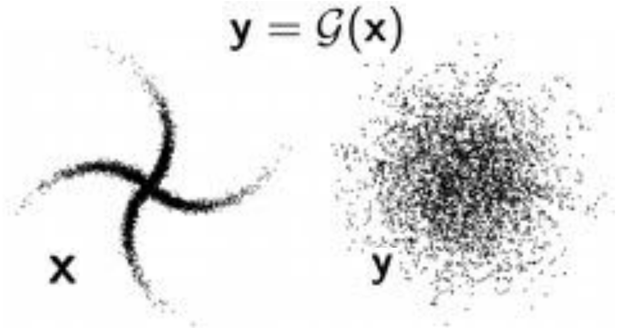
# Estimating densities is challenging ...

- Curse of dimensionality: Multivariate densities are difficult to estimate
- Naive and current methods need millions of points so computationally costly!
- Current approaches are only univariate!

# Estimating densities is ~~challenging~~ easy with RBIG ...

- Curse of dimensionality: Multivariate densities are difficult to estimate
- Naive methods need millions of points as dimension increases, so very costly!
- Current approaches are only univariate!

- Multivariate Gaussianization:
  - Transforms the PDF to a multivariate Gaussian
  - Estimate *p(x)*

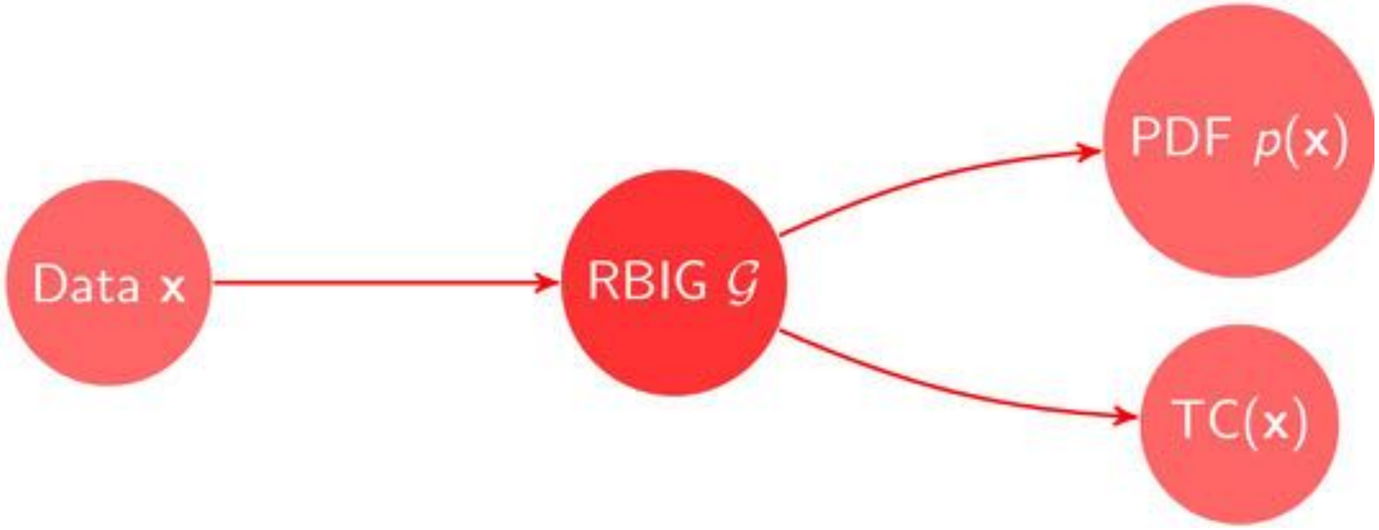



$$p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{y}}(\mathcal{G}(\mathbf{x}))\left|\frac{d\mathcal{G}(\mathbf{x})}{d\mathbf{x}}\right| = p_{\mathbf{y}}(\mathcal{G}(\mathbf{x}))\,|\nabla_{\mathbf{x}}\mathcal{G}(\mathbf{x})|$$

**"Iterative Gaussianization: From ICA to random rotations"**
Laparra, Malo, Camps-Valls, IEEE Trans. Neur. Nets, 2011.

# Spatio-temporal information analysis

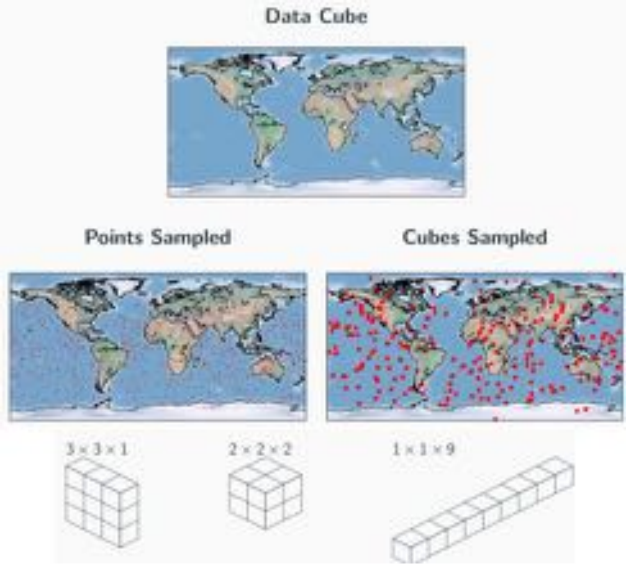
- Multivariate Gaussianization method to estimate *p(x)* and multi-information

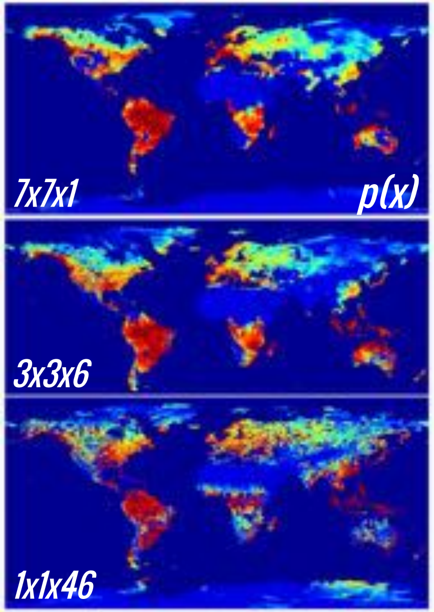


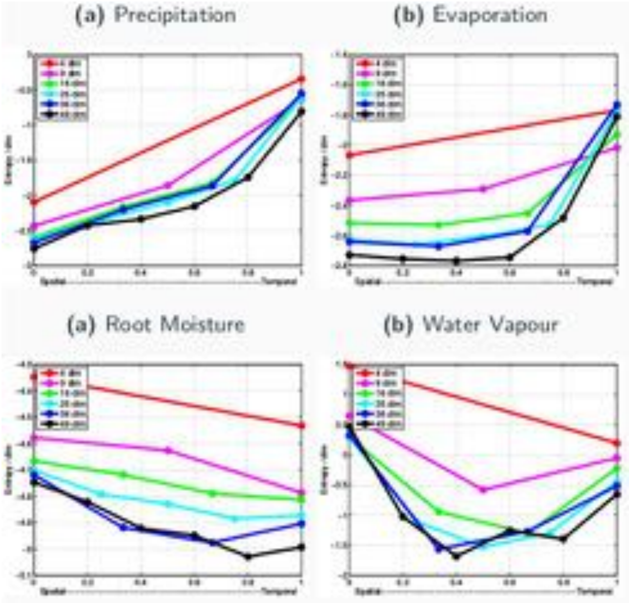

**"Iterative Gaussianization: From ICA to random rotations"**
Laparra, Malo, Camps-Valls, IEEE Trans. Neur. Nets, 2011.

# Spatio-temporal information analysis

- Multivariate Gaussianization method to estimate *p(x)* and multi-information

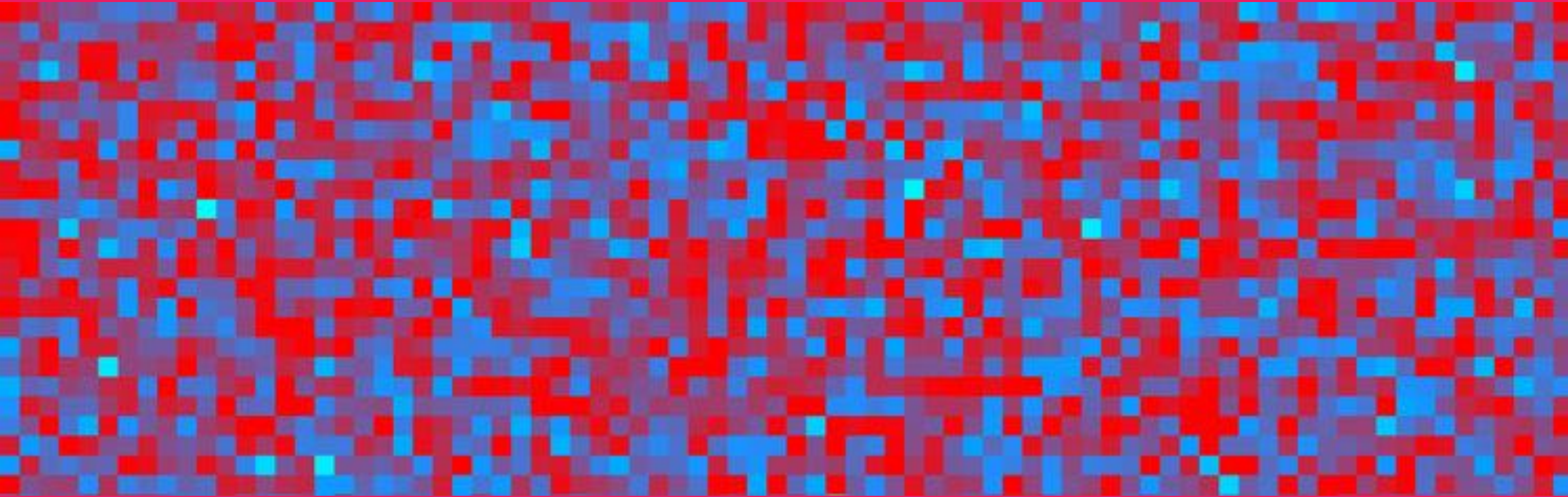


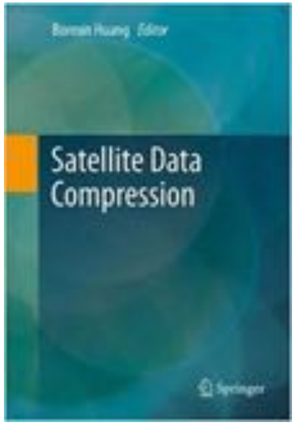

"Iterative Gaussianization: From ICA to random rotations"
Laparra, Malo, Camps-Valls, IEEE Trans. Neur. Nets, 2011.

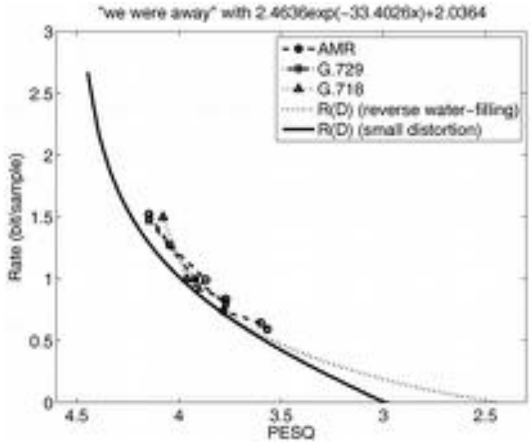
# Compressibility and sparsity of the Earth cube



# Image compressibility with wavelets

- Wavelets excel at spatial analysis
- Rate-distortion: Sparsity vs Reconstruction error


**"On the suitable domain for SVM training in image coding"** Camps-Valls, Gutiérrez, Gómez and Malo. JMLR 2008.

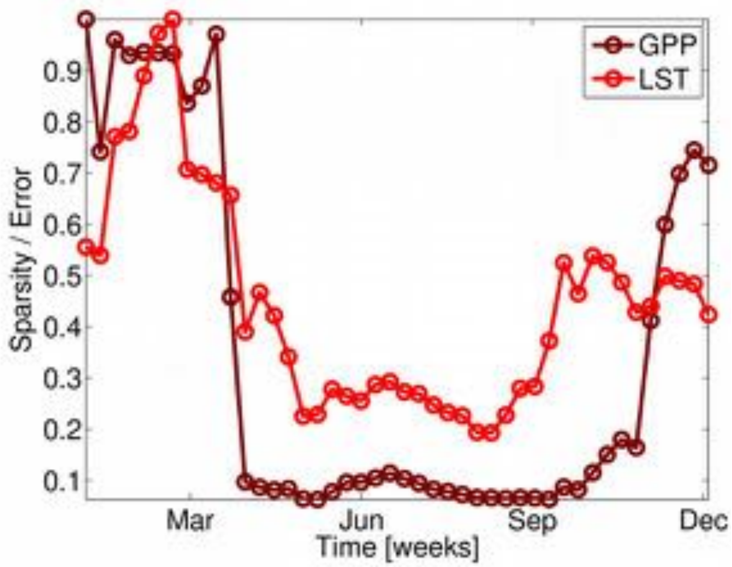
**Check our suite: http://isp.uv.es/software.html**

**"Regression Wavelet Analysis for Loss less Coding of Remote-Sensing Data"** Amrani, Serra-Sagrista, Laparra, Marcellin, Malo. IEEE TGARS 2016

**"Parameter Retrieval Largely Benefits from Spatial-Spectral Image Compression"** Garcia, Serra, Laparra, Calbet, Camps-Valls. IEEE TGARS 2017

# Image compressibility as anomaly index

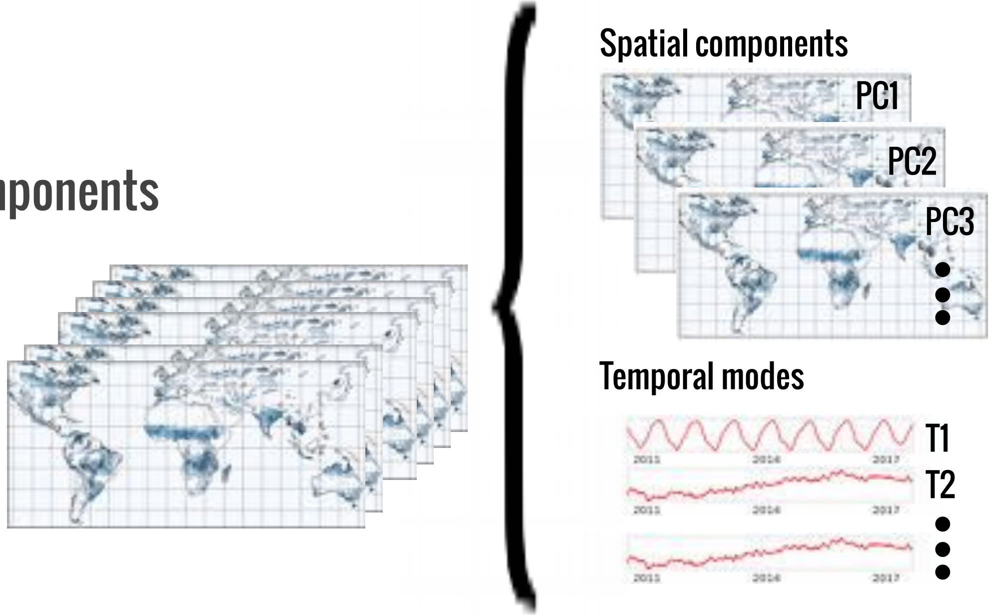
- Anomaly index = Sparsity/Error
- *“Something interesting is going on if you cannot achieve sparsity  and fixing the error”*

# Extracting components from spatial-temporal data

# Spatio-temporal analysis of the Earth cubes

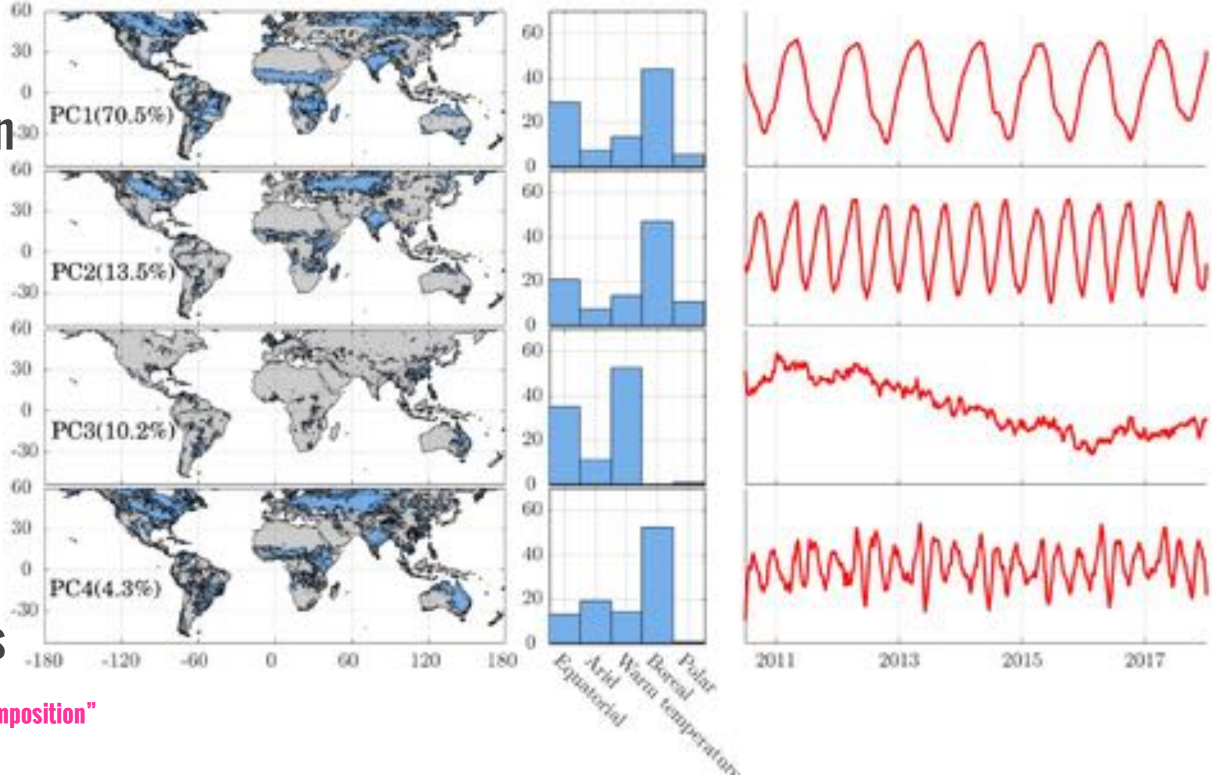
- **PCA/EOF** is popular, yet cannot cope with nonlinear spatio-temporal relations
- **ROCK PCA**
  - copes with nonlinearities
  - extracts spatial and temporal components
  - as many as components as pixels
  - very fast and parallel versions




“Rotated Complex Kernel PCA for spatio-temporal data decomposition”
Bueso, Piles, Camps-Valls, IEEE TGARS, 2018

# Spatio-temporal analysis of the Earth cubes

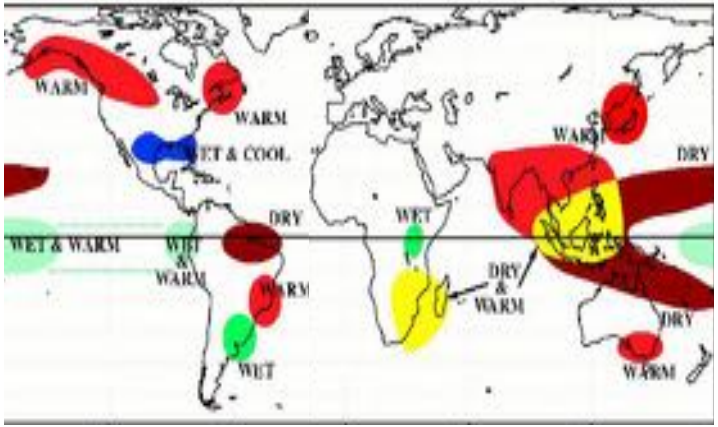
- SM decomposition
  - Meaningful compression
  - Climate-specific modes of variability
  - Boreal and Equatorial modes of SM variability dominate
  - Seasonal and ENSO related temporal modes

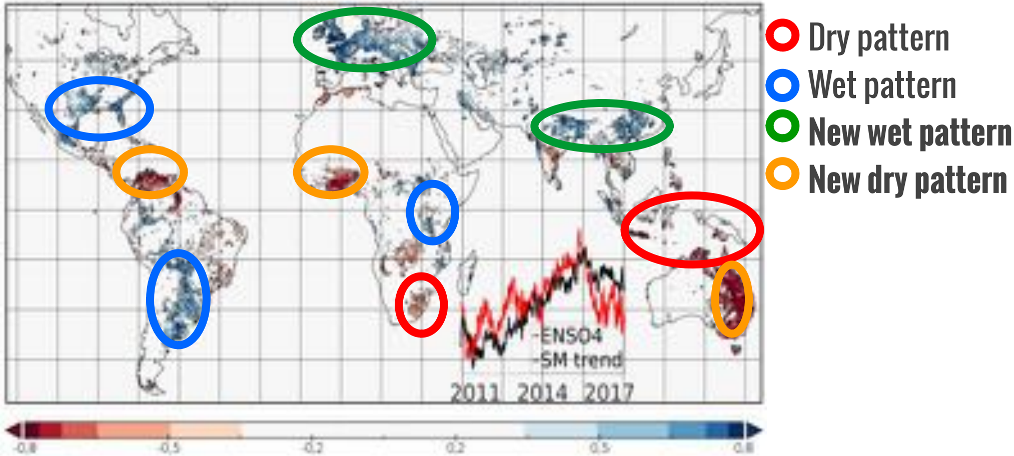



"Rotated Complex Kernel PCA for spatio-temporal data decomposition"
Bueso, Piles, Camps-Valls, IEEE TGARS, 2018

# Spatio-temporal analysis of the Earth cubes

- PC3 highly correlates with ENSO + new spatial patterns uncovered




- Nonlinear cross-correlation uncovers unreported SM-ENSO lags

| | ENSO 1.2 | ENSO 3 | ENSO 3.4 | ENSO 4 |
|---|---|---|---|---|
| Lag [days] | 60 | 30 | 25 | 5 |
| Max Corr | 0.56 | 0.68 | 0.66 | 0.8 |

**"Rotated Complex Kernel PCA for spatio-temporal data decomposition"** Bueso, Piles, Camps-Valls, IEEE TGARS, 2018

# Conclusions

# Conclusions

- Earth data cubes pose great opportunities
    - Research
    - Applications
- Information theoretic approaches
    - Density estimation and multi-information
    - Compression and sparsity
- New machine learning methods
    - Multivariate data
    - Multisource data
    - Structured spatio-temporal relations
    - Nonlinear features
    - Discover new patterns