POTSDAM INSTITUTE FOR CLIMATE IMPACT RESEARCH

Introduction to Functional Network Analysis

Reik V. Donner

Leibniz Gemeinschaft

EOF analysis

Original motivation: extract dominating co-variability from spatio-temporal fields of climate observations records (dimensionality reduction)

Linear PCA:

Diagonalization of lag-zero covariance matrix *C* of multivariate time series (matrix *X* wth standardized components)

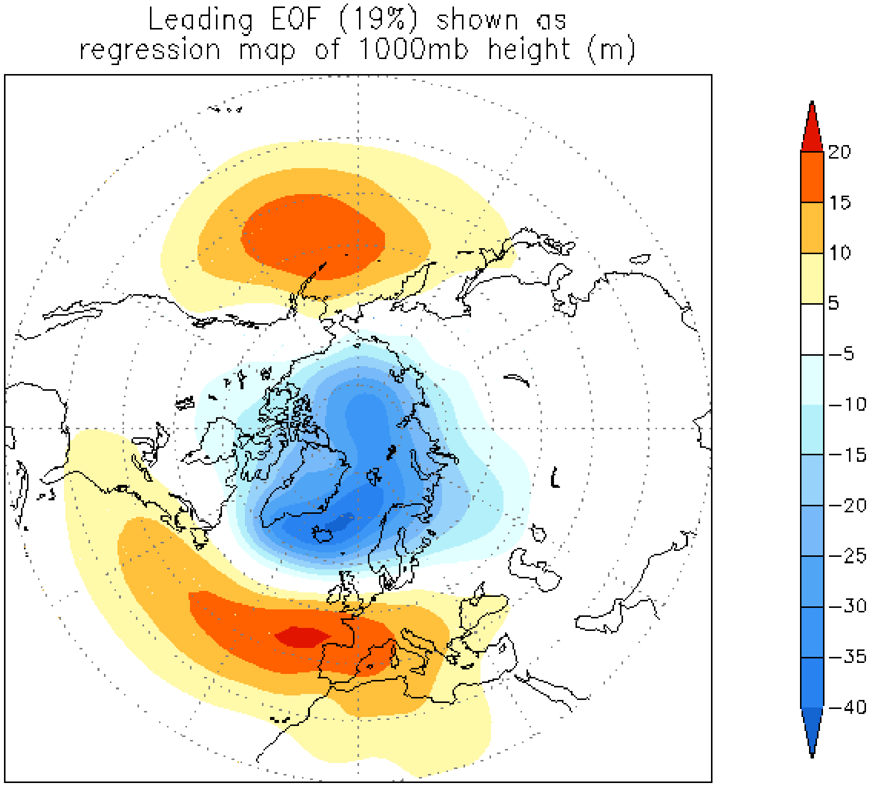
$$C = X^T X \quad \textbf{with} \quad C = U^T \Sigma U \quad \textbf{and} \quad \Sigma = diag(\sigma_1^2, \ldots, \sigma_N^2)$$

- **Compute correlation matrix of all variables**
- **Estimate eigenvalues and eigenvectors**
- **Eigenvectors: additive decomposition into principal components (weighted superpositions of original variables) with individual variances corresponding to associated eigenvalues**

$\Rightarrow$ **spatial EOF patterns + index/score time series describing magnitude and sign of individual EOF modes (characteristic for individual climate oscillations)**

EOF analysis

Example: leading EOF (EOF-1) of near-surface air pressure in Arctic
=> Dipole structure (Arctic Oscillation)

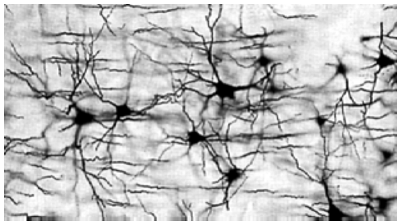


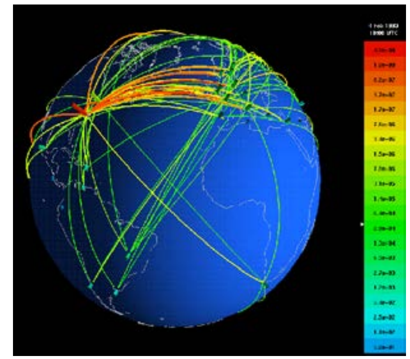
Limitations of EOF analysis

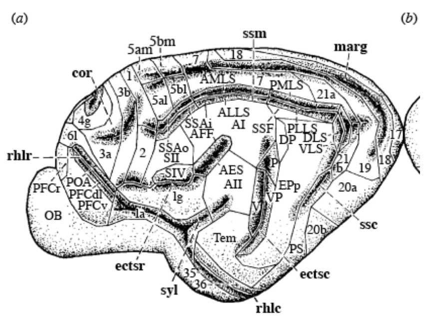
Purpose: extract dominating spatio-temporal (co-)variability modes from fields of climate observations

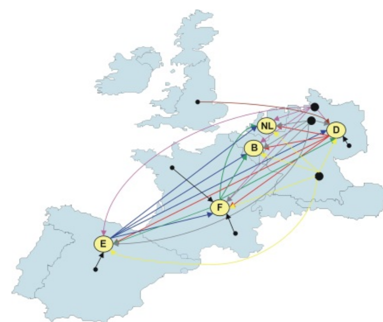
- **Linear decomposition/dimensionality reduction technique**
  - **Potential improvement: nonlinear extensions like kernel PCA, neural network PCA, isometric feature mapping and other nonlinear dimensionality reduction methods**
- **Intrinsic tendency to exhibit dipole (or multipole) structures enforced by orthogonality constraint between modes**
- **EOFs modes do not always coincide with specific climatic mechanisms**
  - **Relevance of EOF modes as dynamical modes (or even proper statistical modes) questionable**
- **Multiple superimposed patterns need to be considered**
- **Spatial patterns = strength of co-variability, unclear relevance of associated temporal patterns in other regions not highlighted by the same EOF**
  - **Integrated view on co-variability, pair-wise co-variability information is lost**

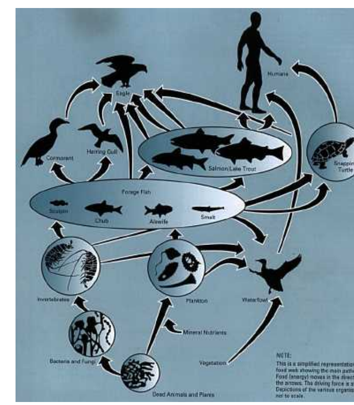
Why network analysis?











Networks are everywhere!

Complex networks appear in various scientific disciplines, including transportation sciences, biology, sociology, information sciences, telecommunication, engineering, economics, etc.

$\Rightarrow$ Solid theory of statistical evaluation

$\Rightarrow$ Efficient numerical algorithms and multiple complementary measures

$\Rightarrow$ Knowledge of interrelations between structure and dynamics

Mathematical foundations

A network (graph) is described by

- **a set of nodes (vertices) V**
- **a set of links (edges) E between pairs of vertices**
- **eventually a set of weights associated with the nodes and/or links**

Basic mathematical structure: adjacency matrix A

$A_{ij}=1 \;\Leftrightarrow\;$ **nodes i and j are connected by a link**

$A_{ij}=0 \;\Leftrightarrow\;$ **nodes i and j are not connected by a direct link**

$\Rightarrow$ **binary matrix containing connectivity information of the graph**

$\Rightarrow$ **undirected graph: A symmetric**

Matrix of link weights W: $A_{ij}=\Theta(W_{ij})$

$\Rightarrow$ **transfer weighted into unweighted graphs by thresholding link weights**

<u>In the following:</u> mainly unweighted and undirected networks

Mathematical foundations

Path: **ordered, mutually exclusive sequence of edges that connects two given vertices.**

Shortest paths (= graph/geodesic distance between two vertices):

$\Rightarrow$ minimum sequence of edges between two given vertices

$\Rightarrow$ for two given vertices, the shortest paths may not be unique

$\Rightarrow$ shortest path length l_{ij}: minimum number of edges between two vertices Note: different conventions for vertices that belong to disjoint network components (l_{ij}=N-1 or infinity, depending on application)

Walk: **general ordered sequence of edges that connects two given vertices (does include possible multiple use of edges)**

Loop: **closed walk of a fixed length that excludes the initial vertex**

Network characteristics

Degree (centrality): number of neighbors of a vertex

$$k_v = \sum_{i=1}^{N} A_{v,i}$$

Closeness centrality: inverse average shortest path length of a vertex

$$c_v = \frac{N-1}{\sum_{i=1}^{N} l_{v,i}}$$

Betweenness centrality: relative fraction of shortest paths on the network that pass a given vertex

$$b_v = \sum_{i,j \neq v}^{N} \frac{\sigma_{i,j}(v)}{\sigma_{i,j}}$$

Local clustering coefficient: relative fraction of neighbors of a vertex that are mutual neighbors of each other

$$\mathcal{C}_v = \frac{2}{k_v(k_v-1)} N_v^{\Delta}$$

Network characteristics

Global clustering coefficient: **mean value of the local clustering coefficient taken over all vertices**

$$\mathscr{C} = \frac{1}{W} \sum_k \frac{\sum_{i,j} A_{ij} A_{jk} A_{ki}}{\sum_{i,j} A_{ki} A_{kj}}$$

Transitivity: **relative fraction of 3-loops in the network**

$$\mathscr{T} = \frac{\sum_{i,j,k} A_{ij} A_{jk} A_{ki}}{\sum_{i,j,k} A_{ki} A_{kj}}$$

=> Both measures characterize closely related properties (asymptotic convergence), but may differ for small graphs

Network characteristics

Average (shortest) path length: **mean of l_{ij} taken over all pairs of vertices (or all pairs of vertices belonging to the same network component)**

$$\mathcal{L} = \langle l_{i,j} \rangle = \frac{2}{N(N-1)} \sum_{i<j} l_{i,j}$$

Network diameter: **maximum graph distance between all pairs of vertices**

$$D = \max_{i,j} l_{i,j}$$

Network radius: **minimum value of maximum distance of a vertex in the network**

$$R = \min_i \max_j l_{i,j}$$

Network characteristics

In general, network measures can be distinguished into different types:

Vertex measures: characterize properties of a single node

Edge measures: characterize properties of a single link

Global network measures: characterize properties of the entire graph

Local properties: measures that take only the adjacency information of a given vertex into account

Meso-scale properties: measures that take adjacency information of a given vertex and its neighbors in the graph into account

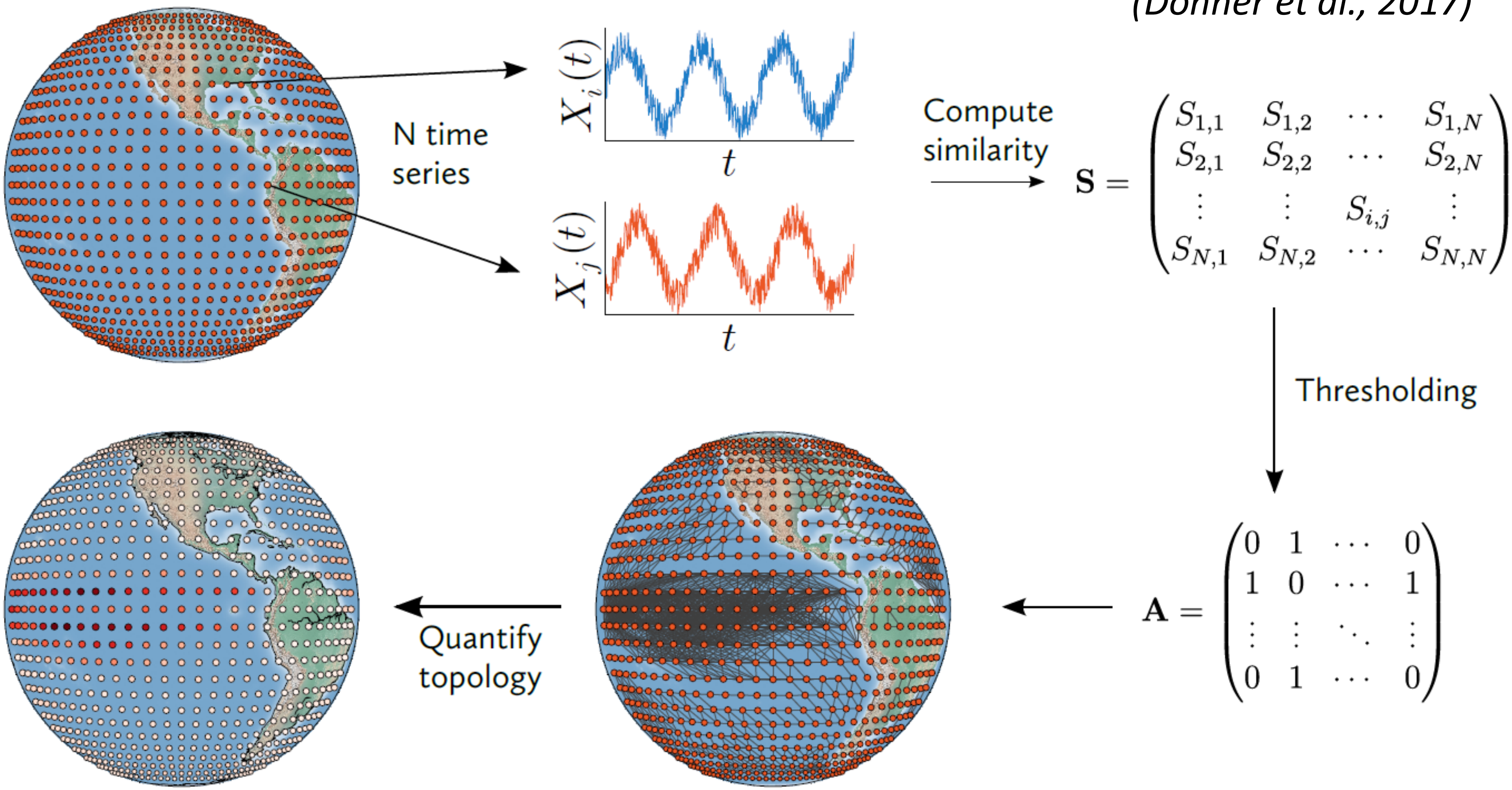
Global properties: measures that take full adjacency information of the whole network into account

Network characteristics

	Vertex measures	Edge measures	Global network measures
Local information	Degree centrality	---	Edge density
Meso-scale information	Local clustering coefficient, Mean nearest-neighbor degree	Matching index (also for non-existing edges)	Global clustering coefficient, Transitivity, Assortativity
Global information	Closeness centrality, Betweenness centrality	Shortest path length, Edge betweenness	Average path length, Diameter, Radius, Efficiency

Climate networks: General workflow

(Donner et al., 2017)



Climate networks

Basic assumptions:

- **Relevant processes in the (continuous) climate system can be approximated by an underlying spatial network structure (spatial coarse-graining is reasonable)**
- **Statistical interdependences between climate variations at different locations reveal corresponding network topology - "functional" network (statistics reflect dynamics) – also used in other fields (e.g., functional brain networks, economics)**

Different possible types of climate networks based on climatological variable and employed similarity measure (e.g. Pearson correlation, different types of mutual information, event synchronization)

Correlation climate networks vs. EOF analysis

Correlation climate networks and EOF analysis based on the same correlation matrix

- **EOF analysis: eigenvector decomposition – linear transformation**
- **Climate network analysis: binarization by thresholding – nonlinear transformation**

If EOF-1 dominates the data set (high fraction of explained variance): approximate relationship between degree field and modulus of EOF-1 (Donges et al., Climate Dynamics, 2015):

$$k_i = \sum_{j=1}^{N} A_{ij} = \sum_{j=1}^{N} \Theta\left(\left|C_{ij}^X\right| - W^*\right) - 1 \qquad C_{ij}^X = \sum_{k=1}^{R} \lambda_k u_{ik} u_{jk}$$

$$k_i \approx \sum_{j=1}^{N} \Theta(\lambda_1 \left|u_{i1} u_{j1}\right| - W^*) - 1$$

Correlation climate networks vs. EOF analysis

Added value of climate network analysis:

- **Commonly, more than just one EOF are statistically relevant – EOF-1 does not tell the whole story**
- **Network allows investigating spatial structure of links (e.g., where are strong correlations with a given location/region located?)**
- **EOF analysis just gives a single spatial pattern and time-dependent score per mode; network analysis provides a multiplicity of characteristics that capture higher-order statistical properties of the spatial correlation structure**
- **Aspects of spatio-temporal organization of climate variability hidden to EOF analysis may be revealed**

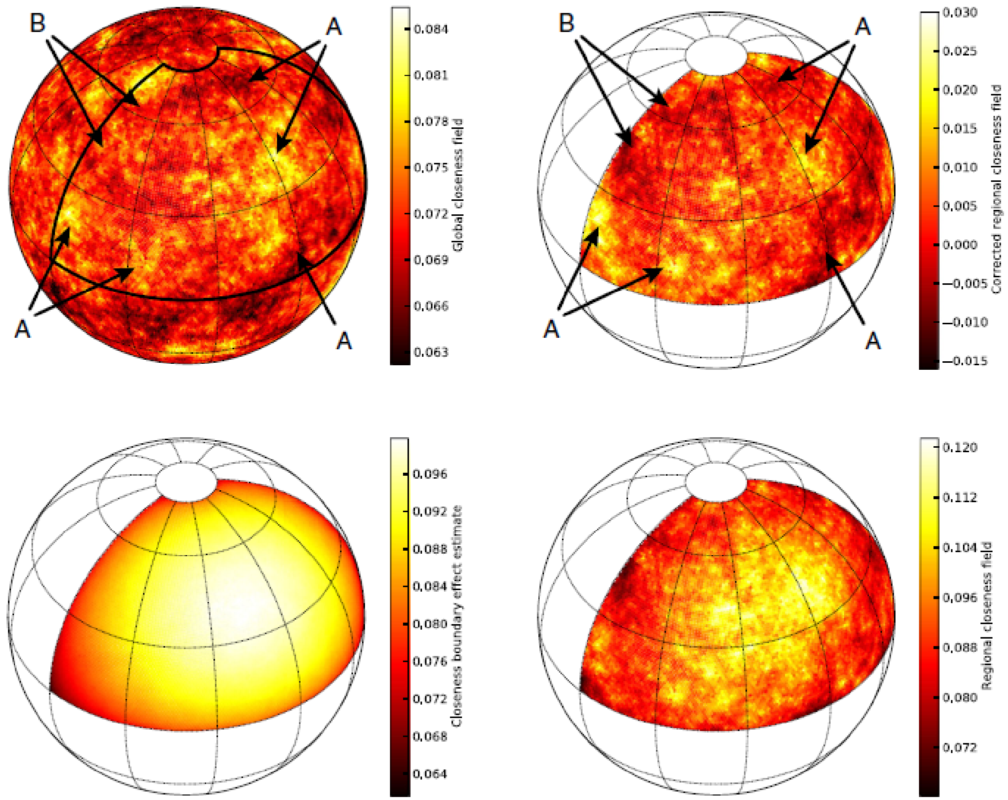
Conceptual challenges

1. Regional climate networks: boundary effects

2. Possibly spatially heterogeneous distribution of vertices

3. Which information is induced by spatial embedding, and which does originate in dynamical properties?

Boundary effects



(Rheinwalt et al., 2012)

Fig. 1: (Color online) Top left: global closeness: closeness centrality of a random network on a sphere. The connection probability depends only on the spatial link length and follows a power law with the exponent −3.5. Top right: corrected regional closeness. Arrows point out areas of strong similarity (A) and dissimilarity (B) in the spatial patterns in the considered region. Bottom left: closeness boundary effects estimate, taken as the median from 1000 surrogates. Bottom right: regional closeness: closeness centrality on a part of the same network as on the whole globe (top left). Nodes in the depicted region are connected if they are connected in the global network.

Boundary effects

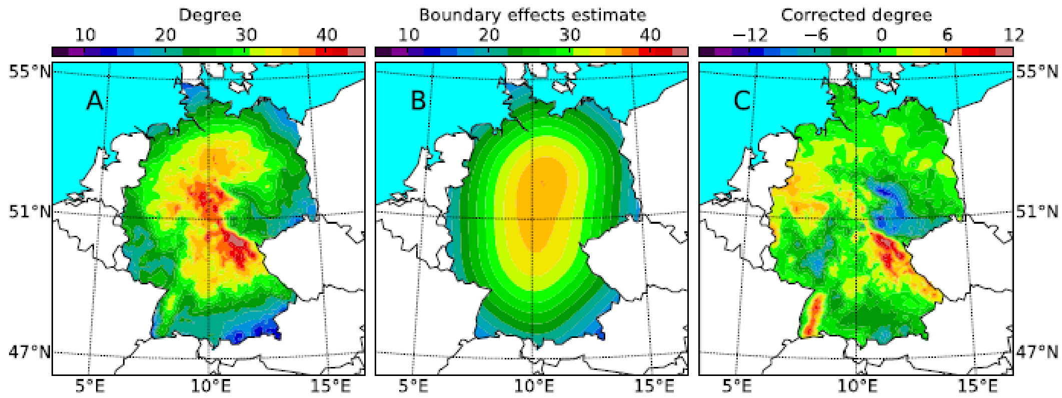


Fig. 5: (Color online) Uncorrected degree field for the network with 50% link density (A), the corresponding boundary effects estimate (B) and the corrected measure (C).

(Rheinwalt et al., 2012)

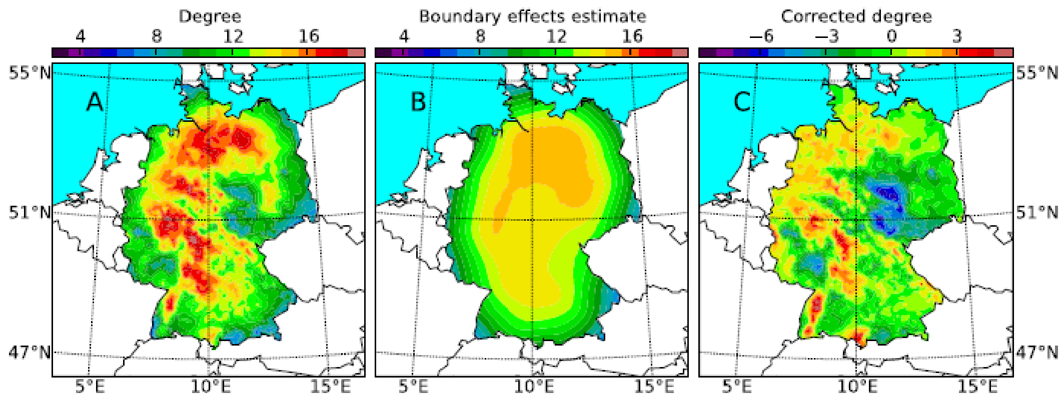


Fig. 6: (Color online) Uncorrected degree field for the network with 25% link density (A), the corresponding boundary effects estimate (B) and the corrected measure (C).

Boundary effects

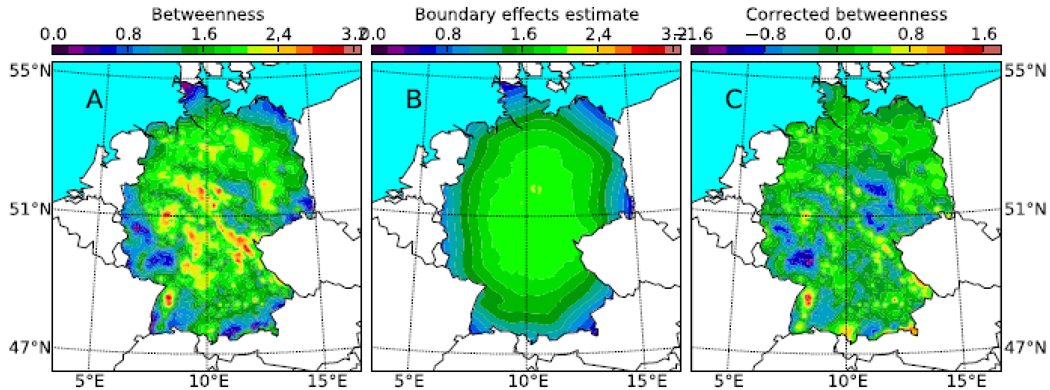


Fig. 7: (Color online) Uncorrected betweenness field for the network with 25% link density (A), the corresponding boundary effects estimate (B) and the corrected measure (C).

(Rheinwalt et al., 2012)

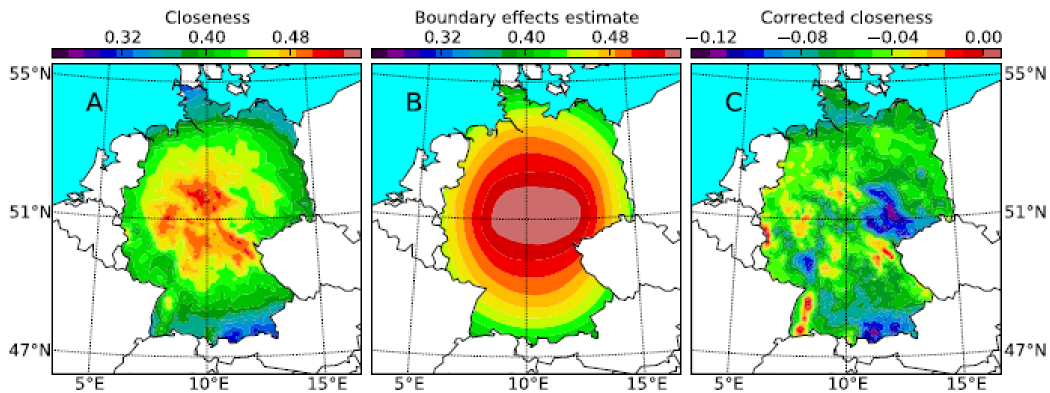


Fig. 8: (Color online) Uncorrected closeness centrality field for the network with 25% link density (A), the corresponding boundary effects estimate (B) and the corrected measure (C).

Correction for represented area

Traditional approach: replace degree by area-weighted connectivity

$$\tilde{C}_i = \sum_{j=1}^{N} \cos\lambda_j \Delta A \Big/ \sum_{\text{over all } \lambda \text{ and } \phi} \cos\lambda \Delta A$$

Generalization: node-splitting invariant measures

$$a_{ij}^{+} = a_{ij} + \delta_{ij}$$

$$W = \sum_{i \in \mathcal{N}} w_i$$

(a) Sum up weights w_v whenever the unweighted measure counts nodes.
(b) Treat every node $v \in V$ as connected with itself.
(c) Allow equality for v and q wherever the original measure involves a sum over distinct nodes v and q.
(d) "Plug in" n.s.i. versions of measures wherever they are used in the definition of other measures.

Correction for represented area

Generalization: node-splitting invariant measures

(Heitzig et al., 2012)

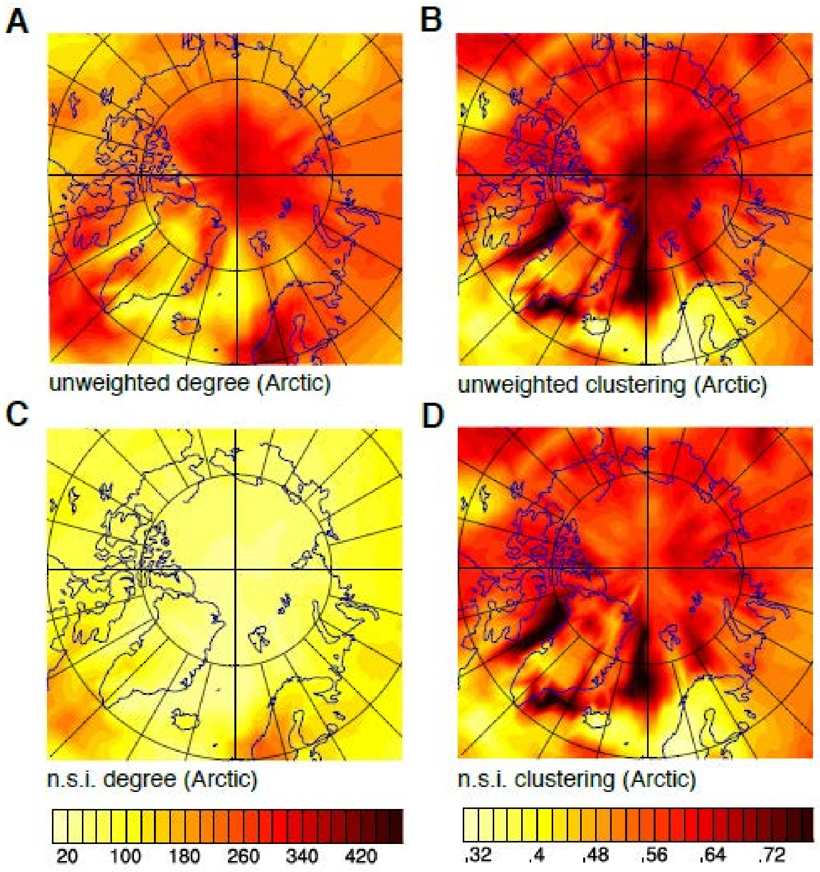
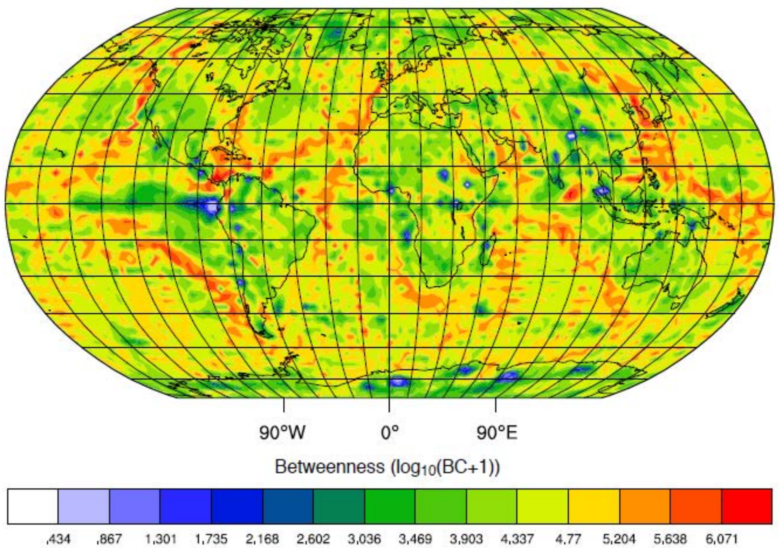


Fig. 3. (Colour online) Comparison of unweighted and weighted (n. s. i.) versions of degree (A,C) and clustering coefficient (B,D) in the northern polar region (Lambert equal area projection) of a global climate network representing correlations in temperature dynamics. The high values at the pole in (A,B) turn out to be an artefact of the increasing grid density toward the pole, as demonstrated by (C,D).

Correction for represented area

Generalization: node-splitting invariant measures



(Donges et al., 2009)

(Heitzig et al., 2012)

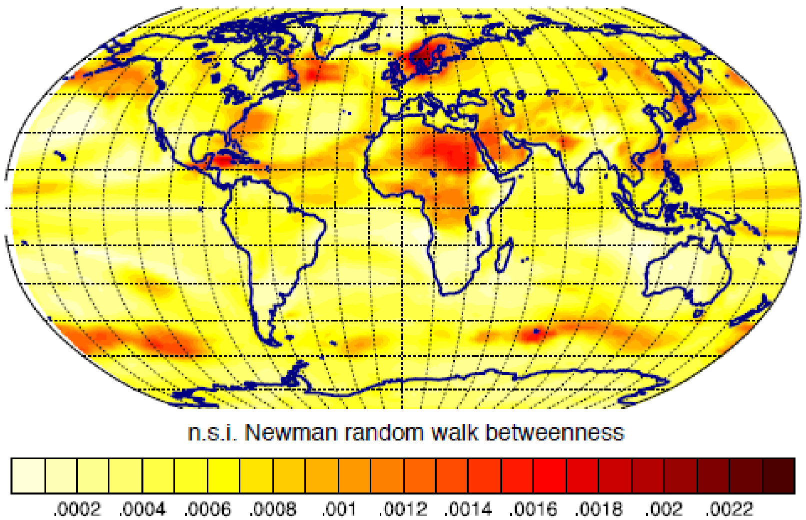


Fig. 13. (Colour online) N. s. i. version NB^*_v of Newman's random walk betweenness in a global climate network representing correlations in surface air temperature dynamics (same network as in Fig. 4, Robinson projection). We can clearly identify the regions of the North Pacific Subpolar Gyre, the North Atlantic Subtropical Gyre including the Gulf Stream and the Canary Current, the North and South Equatorial Currents in the Pacific, and the Antarctic Circumpolar Current. The interpretation of other regions of high values like Scandinavia and Central and North-East Africa remains unclear.

Examples / methodological developments

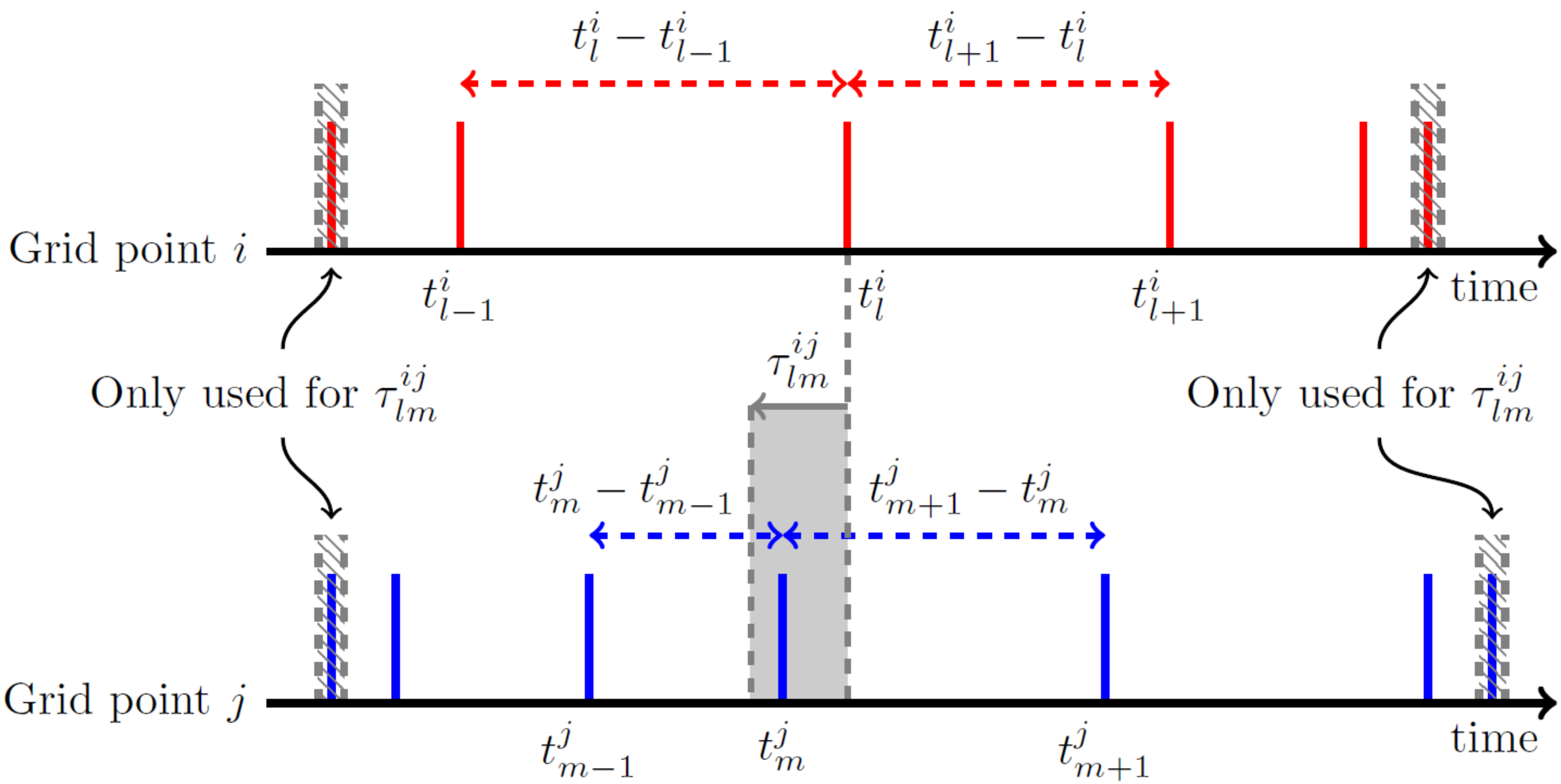
Correlation-based climate networks (yesterday's talk)

Event-based climate networks

Coupled climate networks

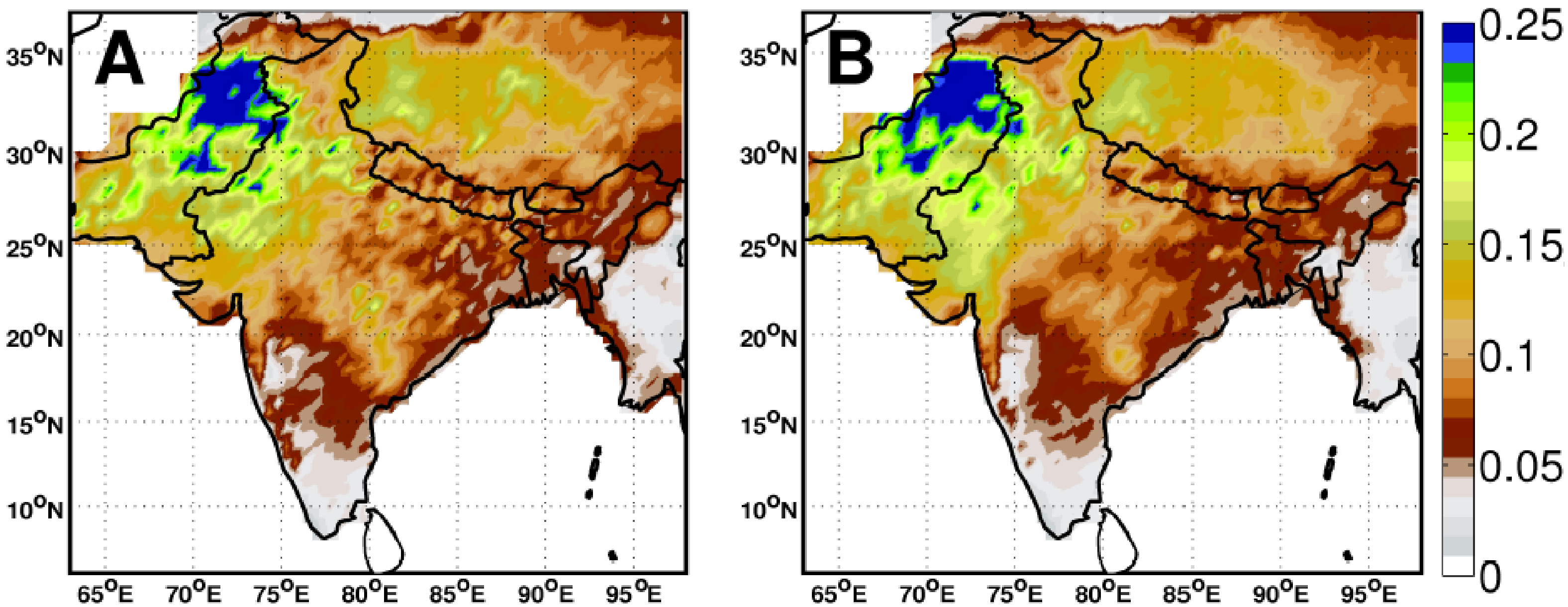
Scale-specific climate networks

Event synchronization



Event synchronization climate networks

Example: Indian summer monsoon precipitation - fraction of grid points with statistically significant event synchronization (event = daily precipitation exceeding (A) 94% and (B) 90% quantile)



(Malik et al., Clim. Dyn., 2012)

Climate networks

(Malik et al., 2012)

Fig. 8 Local clustering coefficient $\mathscr{C}_j$ **a** $\alpha = 94\%$, **b** $\alpha = 90\%$. *Red* colours indicate that the rainfall field is less spatially continuous, i.e., it is fragmented. In contrast, *blue* colours outline more spatially continuous rainfall fields. We observe that $\mathscr{C}_j$ is independent of the spatial scales involved in the rainfall (compare with Fig. 7a)

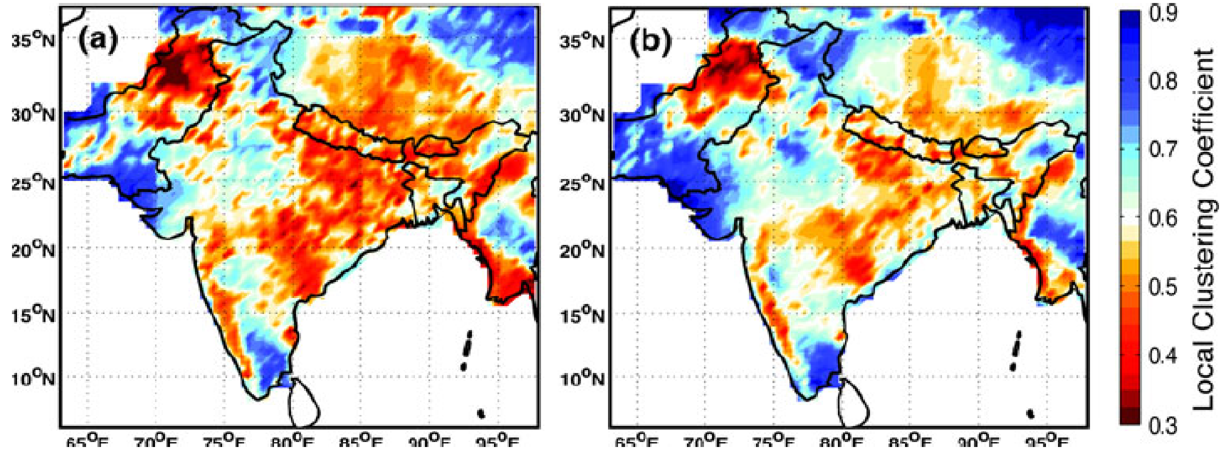
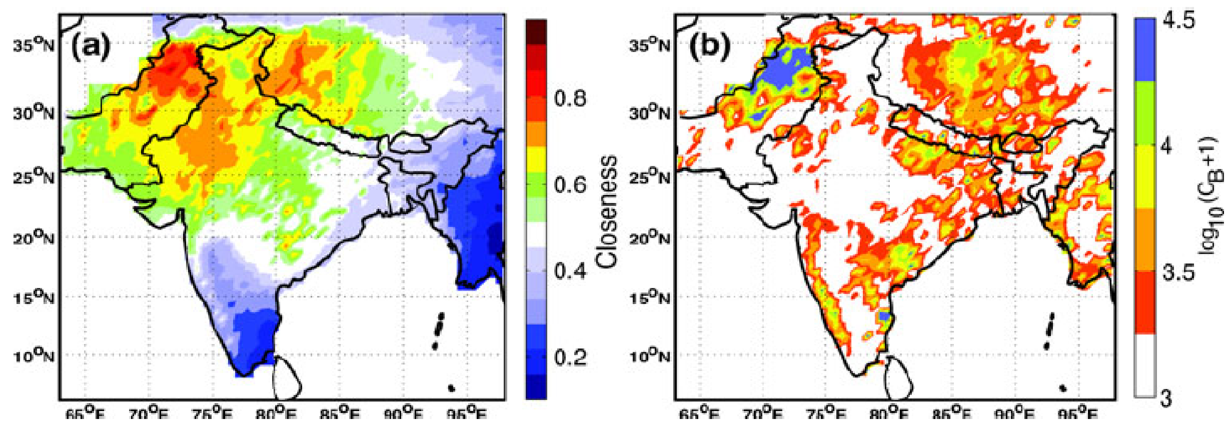


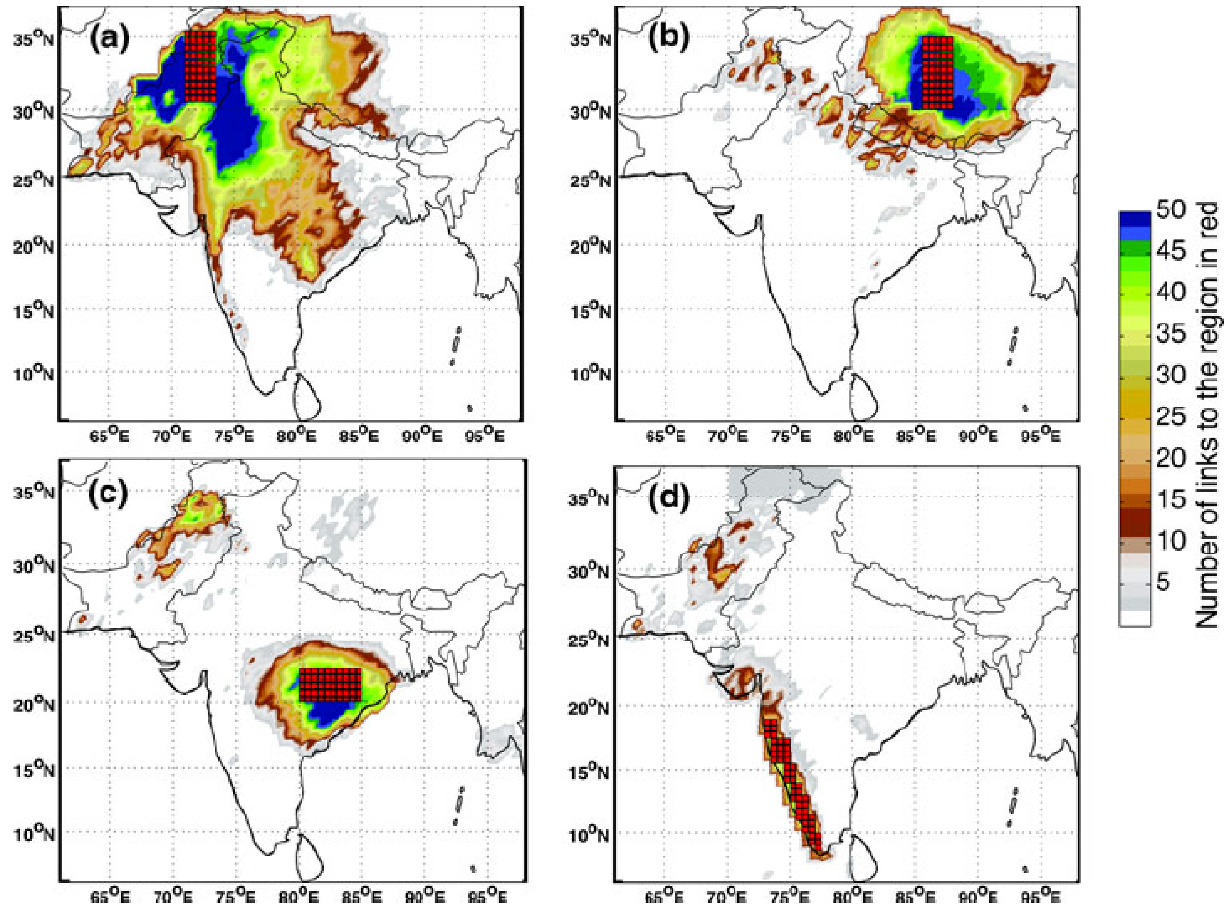
Fig. 9 **a** Closeness centrality C_{Cj}. We observe high C_{Cj} in the northwestern parts of the subcontinent suggesting the importance of atmospheric processes in modulating the ISM activity. **b** Betweenness centrality C_{Bj}. Higher values of C_{Bj} represent the moisture transport pathways over the land during the active phase of the ISM. For both **a** and **b** we chose $\alpha = 90\%$



Climate networks

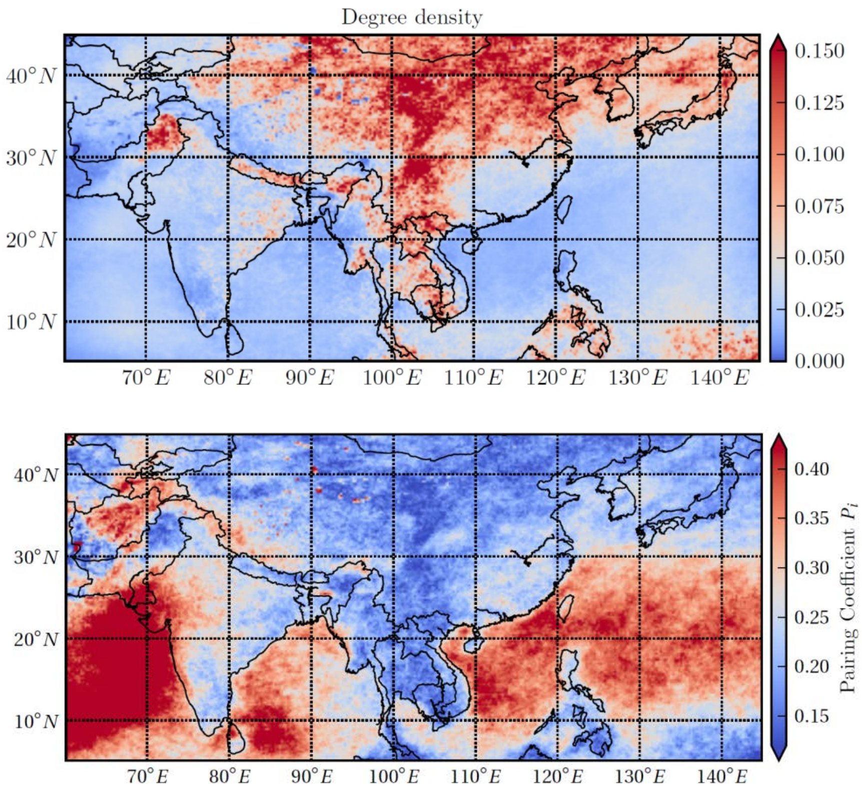
(Malik et al., 2012)

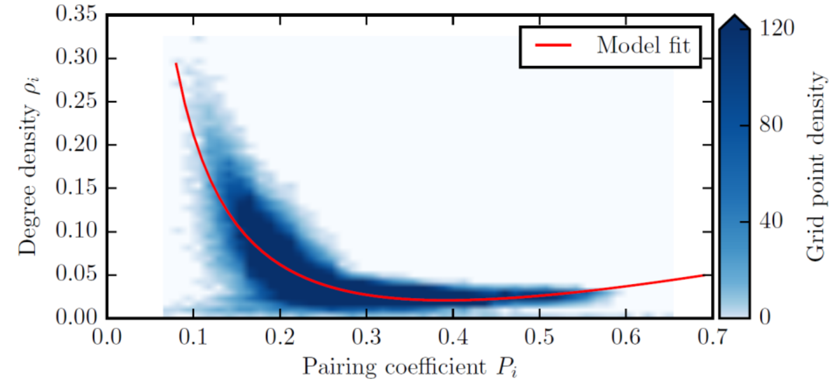
Fig. 10 Links between a set of 50 reference grid points (*gridded red matrix*) to other grid points (*colour bar*) at $\alpha = 90\%$. Note the spatially extensive links for a reference area in northwestern Pakistan (a). In contrast, extreme rainfall linkages in the western Ghats (d) have a limited spatial extent



Problem: heterogeneous waiting times

Degree patterns are largely determined by temporal clustering of events: strong clustering = only short time differences allowed = few ES connections

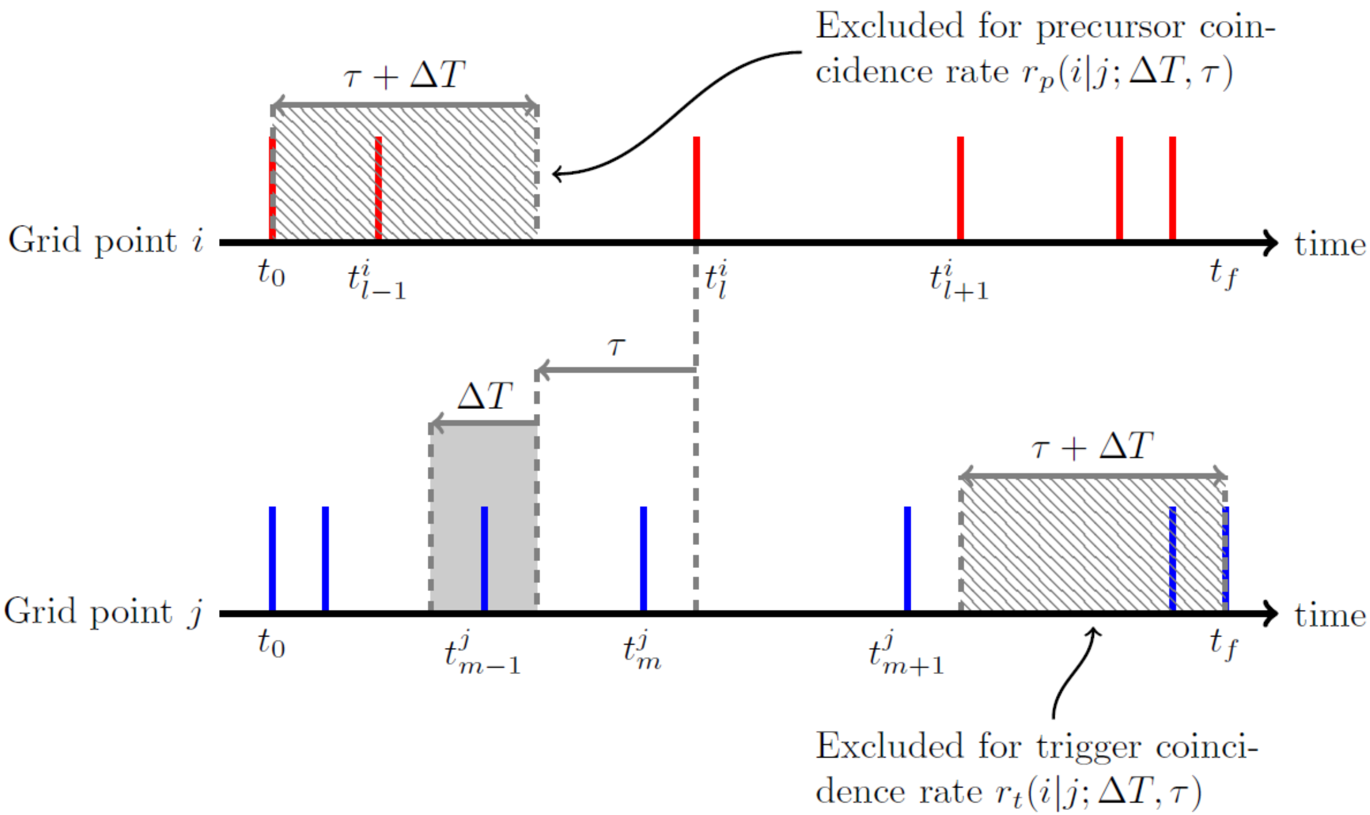




(Odenweller & Donner, in prep.)

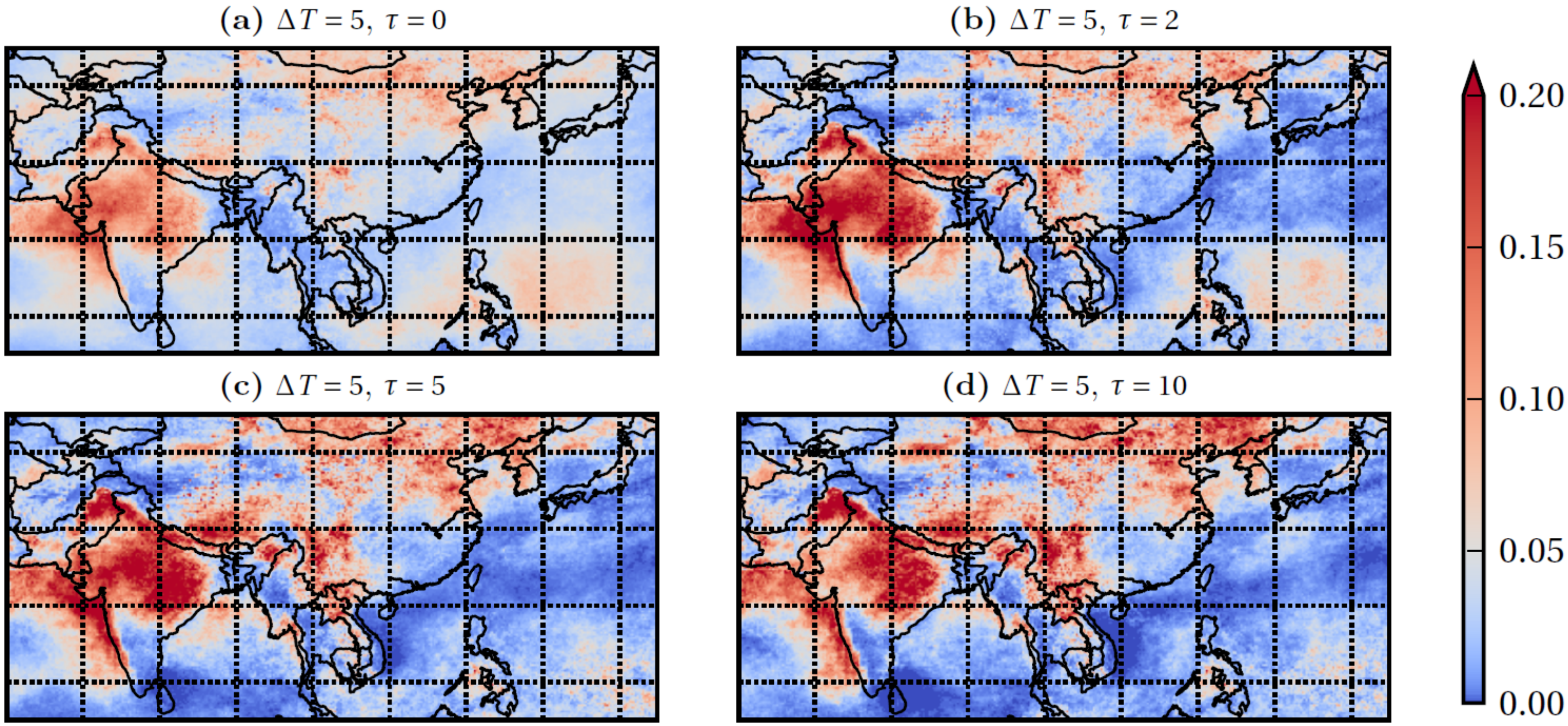
Alternative: event coincidence rate networks

Additional parameters: control of acceptable time lags – resolves different scales



Alternative: event coincidence rate networks

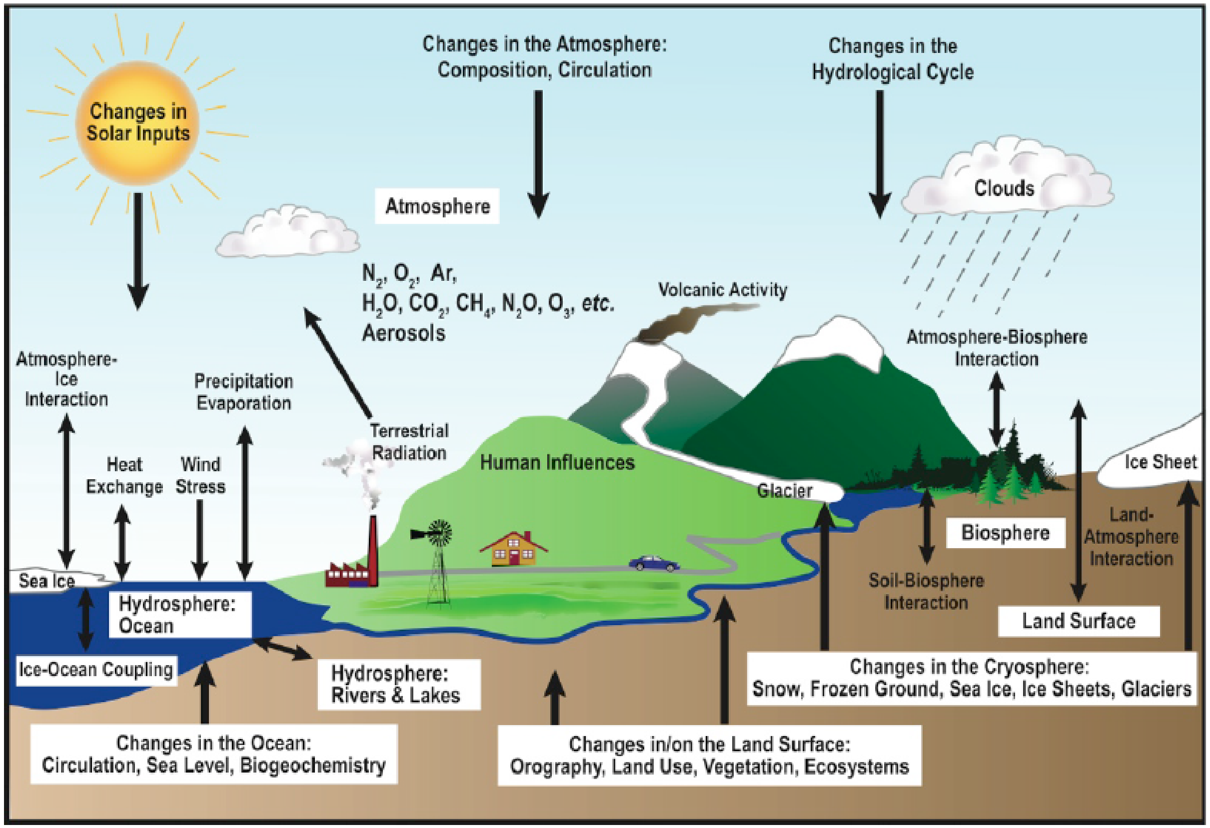
Additional parameters: control of acceptable time lags – resolves different scales



(Odenweller & Donner, in prep.)

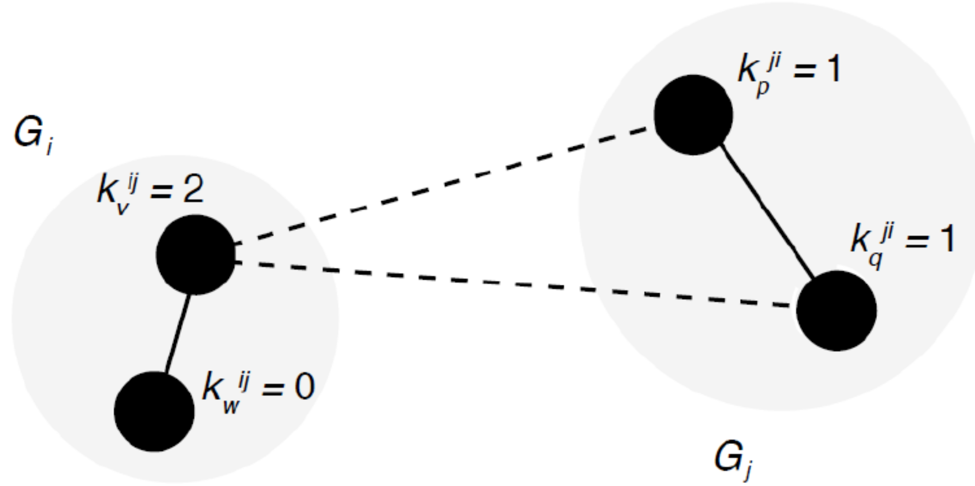
Coupled climate networks

Hierarchical and multi-domain interactions in Earth system

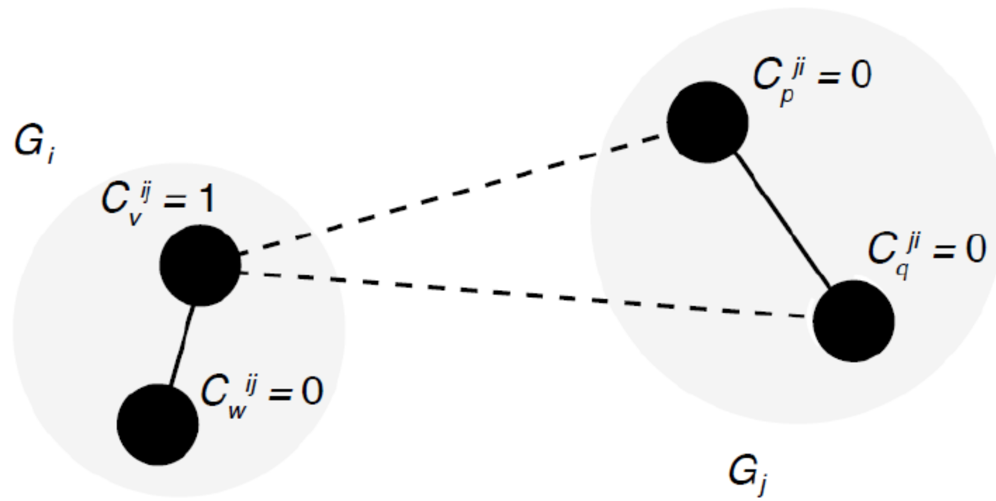


Coupled climate networks

Cross-degree

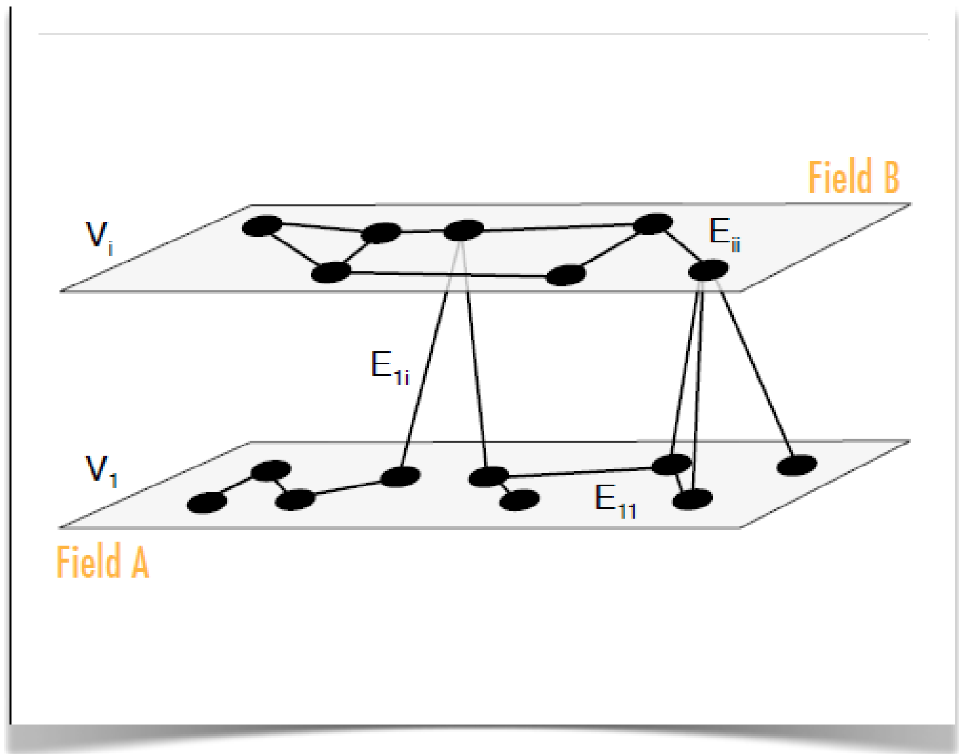


Cross-clustering



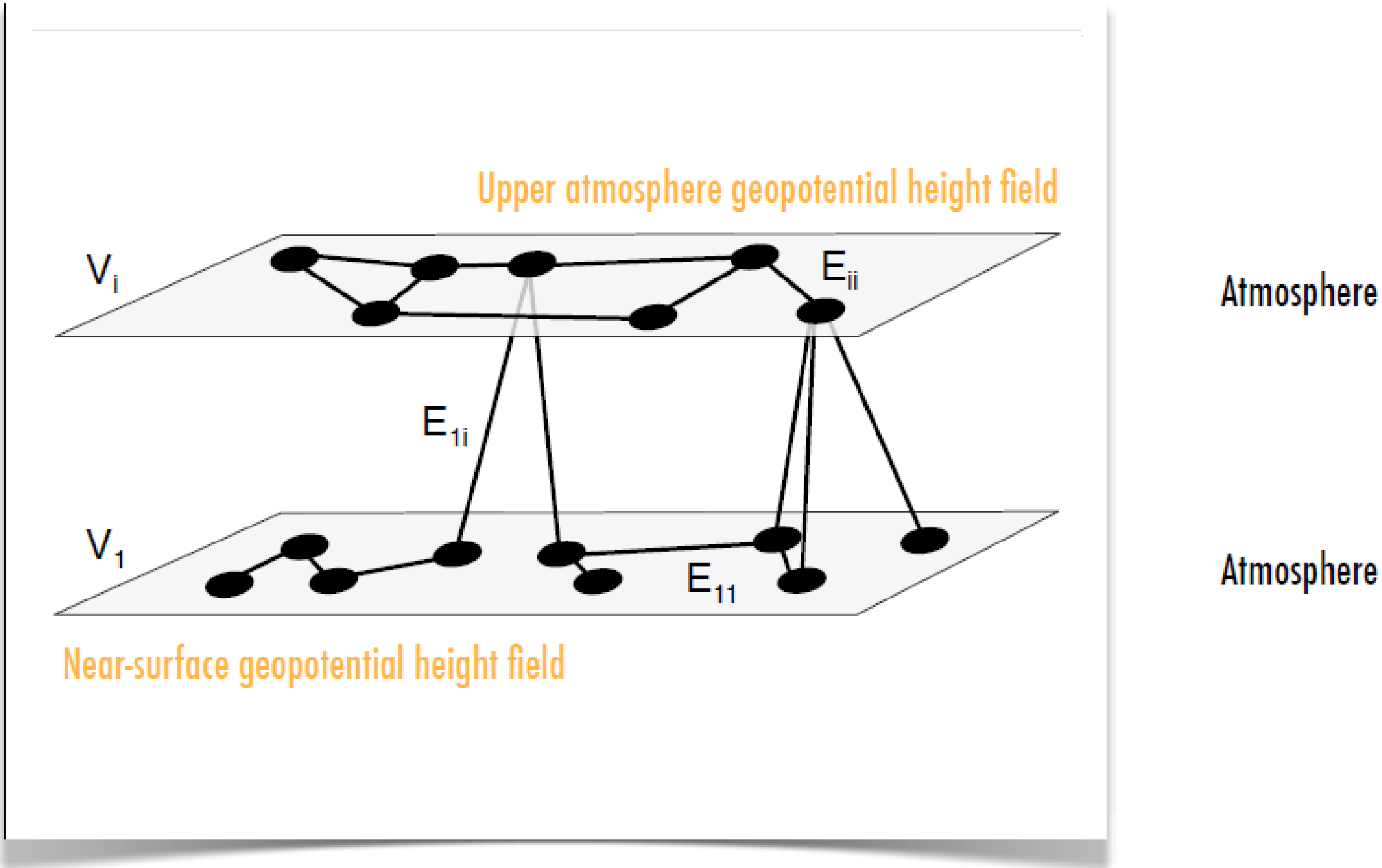
Coupled climate networks

- Study statistical interdependency structure between climatological fields
- Conceptually and formally linked to maximum covariance analysis (SVD of cross-covariance matrix)



Donges et al., Eur. Phys. J. B **84** (2011)
Donges et al., Climate Dynamics (2015)

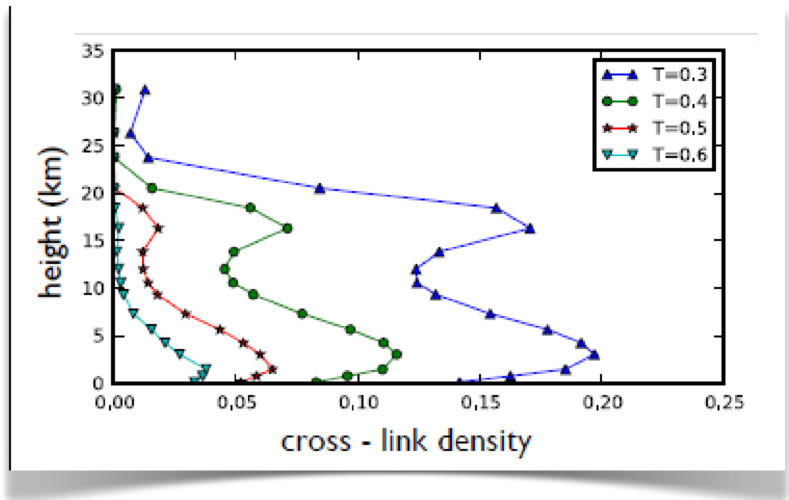
Atmospheric general circulation



Donges et al., Eur. Phys. J. B **84** (2011)

Atmospheric vertical dynamical structure

- Cross-link density shows extrema close to
  - Planetary boundary layer
  - Tropopause

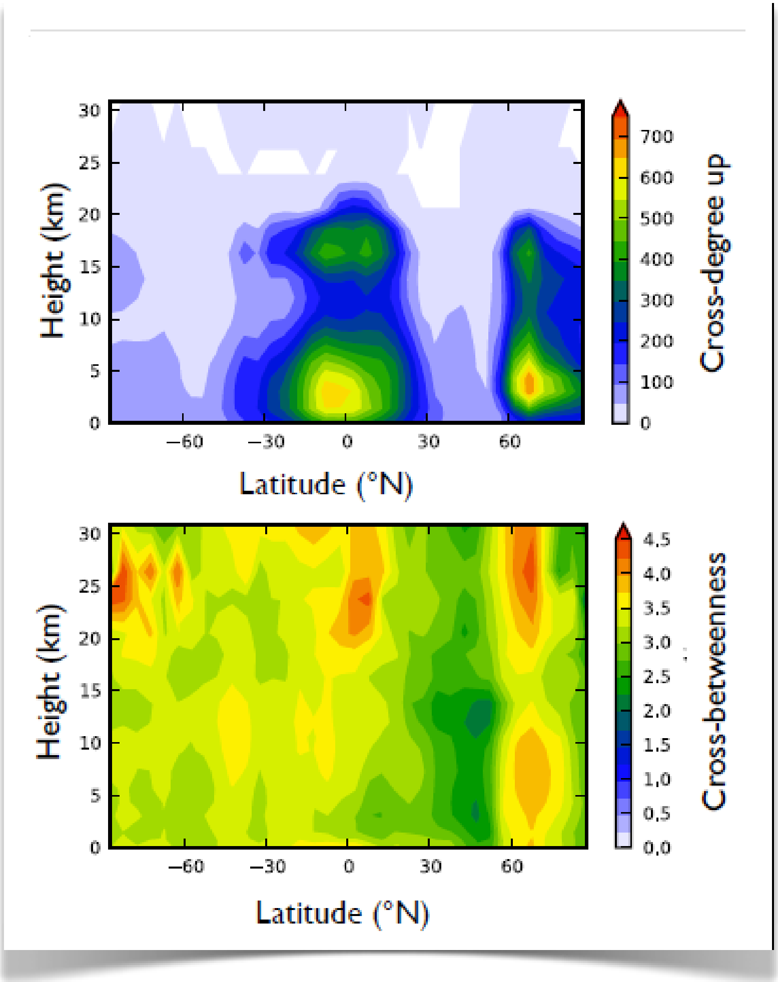


Donges et al., Eur. Phys. J. B **84** (2011)

Atmospheric general circulation

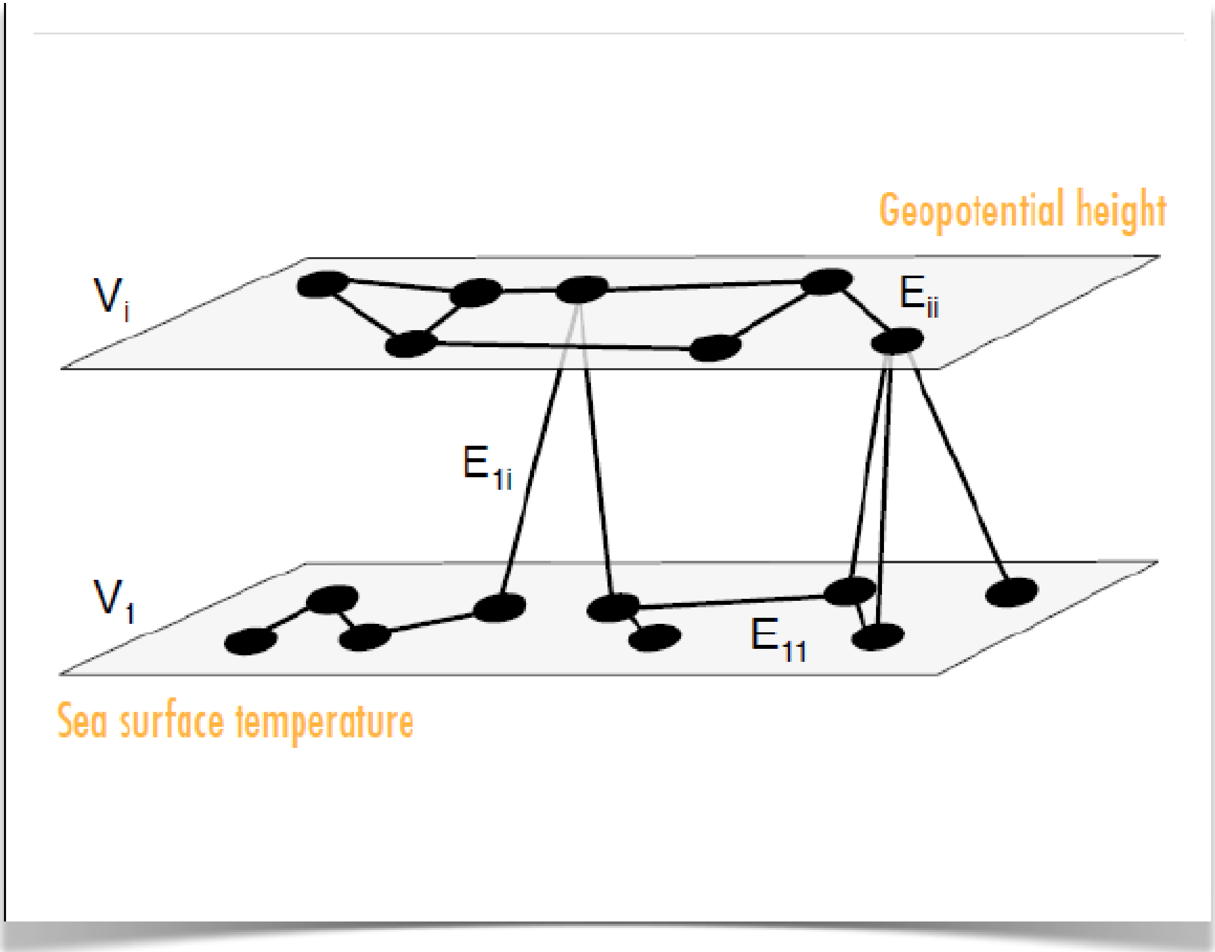
- Cross-degree shows convection cells and meridional circulation

- Cross-betweenness highlights importance of arctic polar vortex



Donges et al., Eur. Phys. J. B **84** (2011)

Atmosphere – ocean interactions



Atmosphere

Ocean

Hierarchical coupling structure

- Introduce node weights to network statistics (node splitting invariant measures)
- Relationship between cross-degree and clustering indicates hierarchical structures in ocean-atmosphere interactions

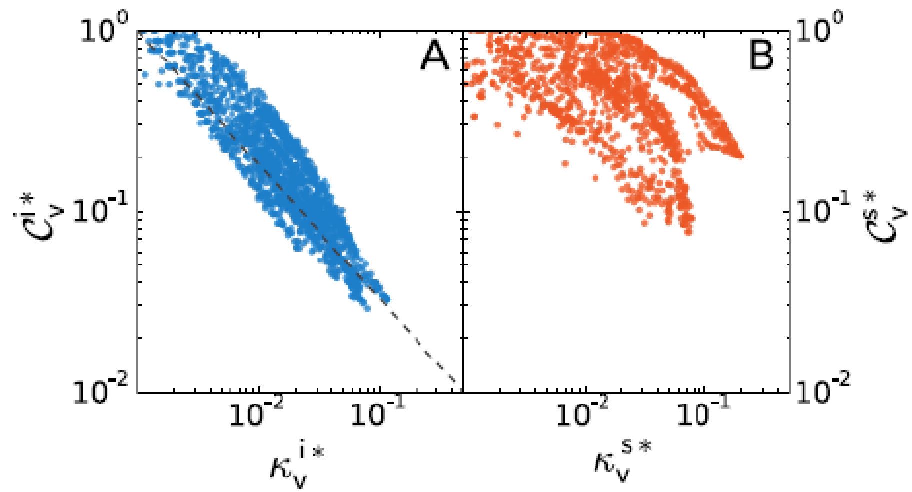
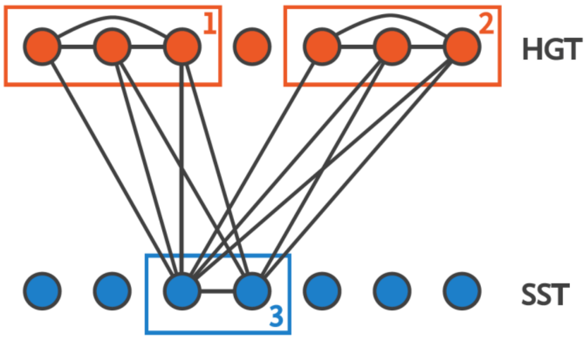
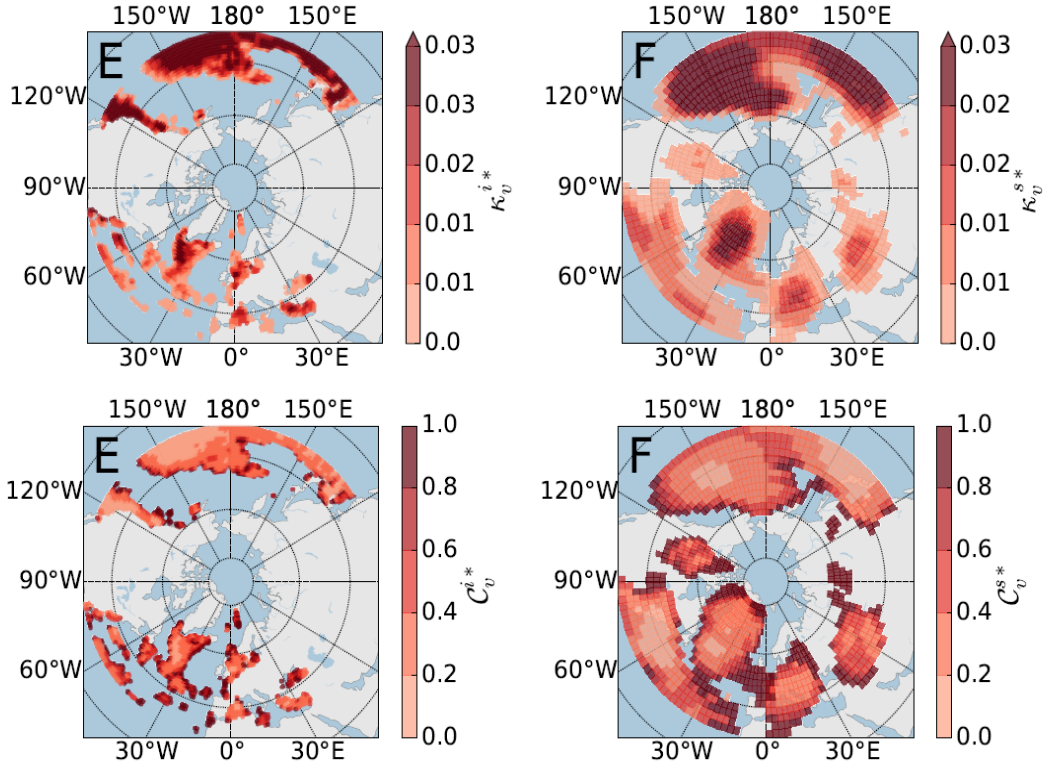


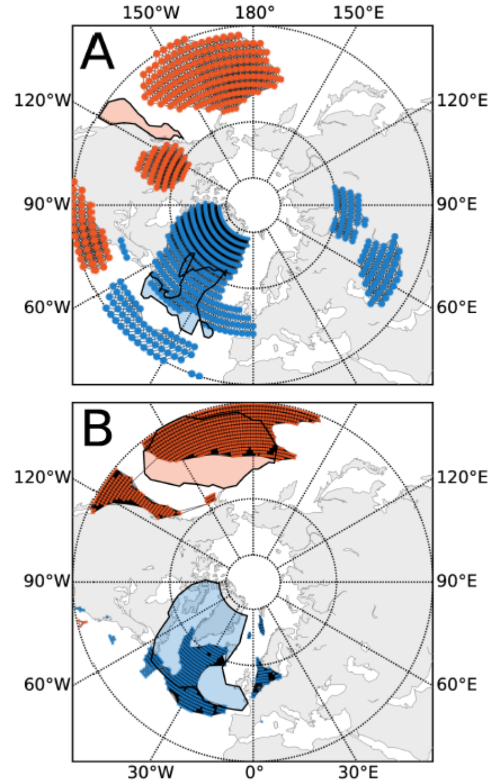
Fig. 6 N.s.i. local cross-clustering coefficients $\mathcal{C}_v^{i*}(\kappa_v^{i*})$ for nodes in the SST field (A) and $\mathcal{C}_v^{s*}(\kappa_v^{s*})$ for nodes in the 100 mbar HGT field (B) as functions of the respective n.s.i. cross-degree densities. The dashed line in (A) indicates the relationship $\mathcal{C}_v^{i*}(\kappa_v^{i*}) \sim (\kappa_v^{i*})^{-1}$ expected for traditional network measures $\mathcal{C}_v(k_v)$ in the case of hierarchical network structures (Ravasz and Barabási, 2003; Ravasz et al, 2002).

(Wiedermann et al., Int. J. Climatol., 2017)

Hierarchical coupling structure



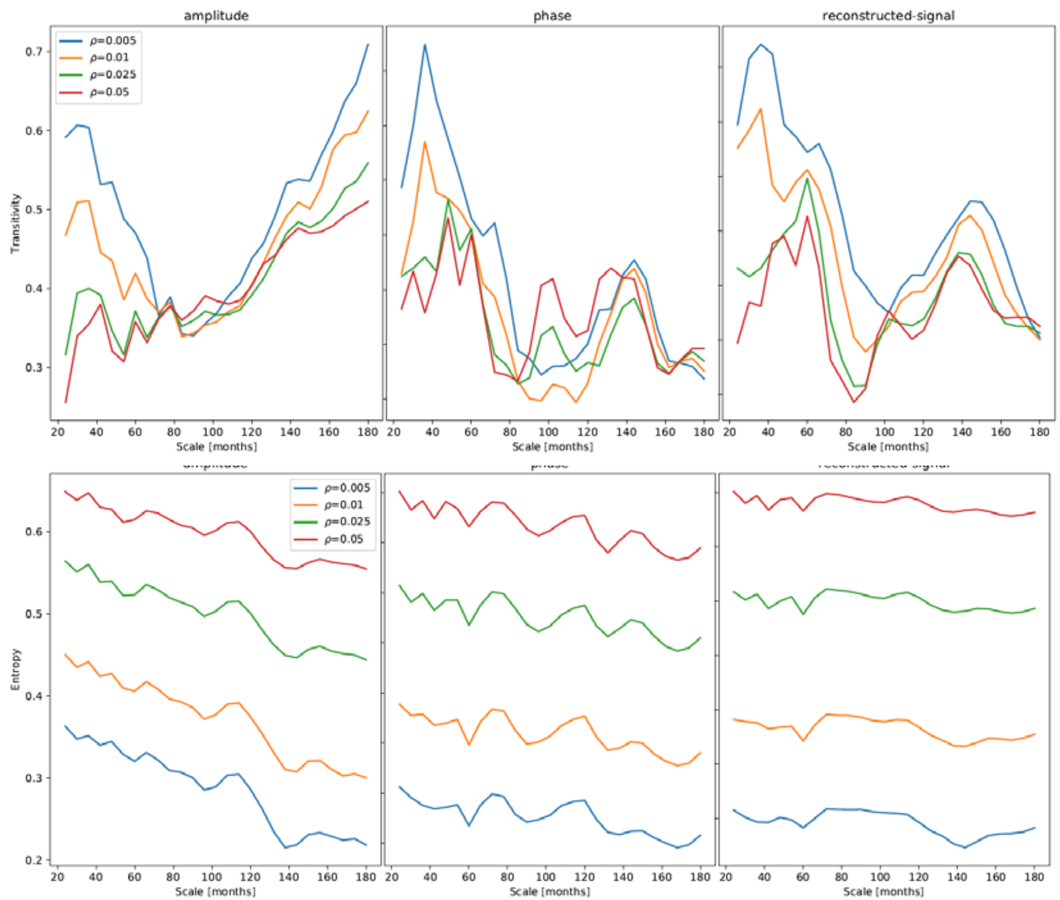




(Wiedermann et al., Int. J. Climatol., 2017)

Scale-dependent climate networks

Climate variability patterns depend on temporal scale of dynamics – reflected in scale-dependent networks obtained for filtered climate data



transitivity

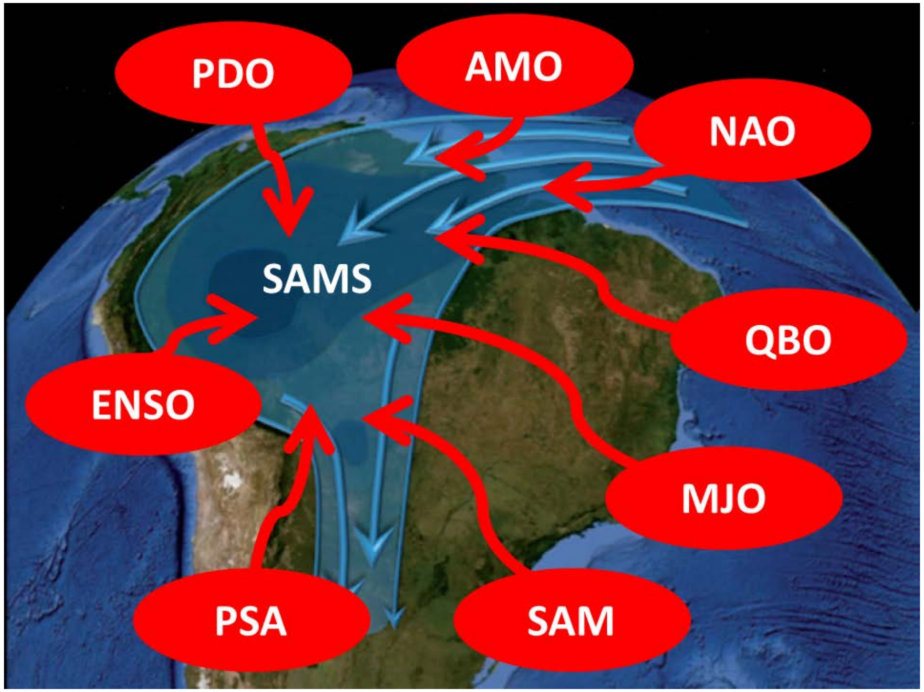
entropy

(Donner et al., in prep.)

Interactions among different climate patterns

Idea: identification of key "modes" via dimensionality reduction/community detection in scale-dependent climate networks (and cross-scale coupled climate networks)

Example: South American Monsoon System (new project)



Conclusions

- (Correlation) climate network analysis as extension and complement of classical EOF analysis
- Extensions to other (nonlinear) similarity measures (e.g. event-related statistics) and coupling structures among different subsystems
- Coupled climate network analysis allowing to resolve previously unknown structural organization features among different climate subsystems

Python package *pyunicorn* for climate network analysis at GitHub (Donges et al., Chaos, 2015)