Clim Dyn (2014) 42:879–899
DOI 10.1007/s00382-013-1729-5

# Spatio-temporal network analysis for studying climate patterns

**Ilias Fountalis · Annalisa Bracco · Constantine Dovrolis**

Received: 19 September 2012 / Accepted: 7 March 2013 / Published online: 29 March 2013


**Abstract** A fast, robust and scalable methodology to examine, quantify, and visualize climate patterns and their relationships is proposed. It is based on a set of notions, algorithms and metrics used in the study of graphs, referred to as complex network analysis. The goals of this approach are to explain known climate phenomena in terms of an underlying network structure and to uncover regional and global linkages in the climate system, while comparing general circulation models outputs with observations. The proposed method is based on a two-layer network representation. At the first layer, gridded climate data are used to identify "areas", i.e., geographical regions that are highly homogeneous in terms of the given climate variable. At the second layer, the identified areas are interconnected with links of varying strength, forming a global climate network. This paper describes the climate network inference and related network metrics, and compares network properties for different sea surface temperature reanalyses and precipitation data sets, and for a small sample of CMIP5 outputs.

I. Fountalis
College of Computing, Georgia Tech,
Atlanta, GA 30332-0280, USA
e-mail: fountalis@gatech.edu

A. Bracco (✉)
School of Earth and Atmospheric Sciences,
Georgia Tech, Atlanta, GA 30332-0340, USA
e-mail: abracco@gatech.edu

C. Dovrolis
College of Computing, Georgia Tech,
Atlanta, GA 30332-0280, USA
e-mail: constantine@gatech.edu

**Keywords** Network analysis · Spatial weighted networks · Model validation · Model comparison · Teleconnections

## 1 Introduction

Network analysis refers to a set of metrics, modeling tools and algorithms commonly used in the study of complex systems. It merges ideas from graph theory, statistical physics, sociology and computer science, and its main premise is that the underlying topology or network structure of a system has a strong impact on its dynamics and evolution (Newman et al. 2006). As such it constitutes a powerful tool to investigate local and non-local statistical interactions.

The progress made in this field has led to its broad application; many real world systems are modeled as an ensemble of distinct elements that are associated via a complex set of connections. In some systems, referred to as structural networks, the underlying network structure is obvious (e.g. Internet routers as nodes, cables between routers as edges). In others, the underlying mechanisms for remote connections between different subsystems are unknown *a priori* (e.g. social networks, or the climate system); still, their effects can be mapped into a functional network. An extensive bibliography for applications of network analysis can be found in Newman (2010).

By quantifying statistical interactions, network analysis provides a powerful framework to validate climate models and investigate teleconnections, assessing their strength, range, and impact on the climate system. The intention is to uncover relations in the climate system that are not (or not fully) captured by more traditional methodologies used in climate science (Dijkstra 2005; Corti et al. 1997; Abramov and Majda 2009; Ghil et al. 2002; Ghil and Vautard 1991;

Forest et al. 2002; Allen and Smith 1994; Andronova and Schlesinger 2001), and to explain known climate phenomena in terms of the underlying network's structure and metrics.

Introductions to the application of network analysis in climate science are presented in Steinhaeuser et al. (2010) and Tsonis et al. (2006). We can classify the prior work in this area in three distinct approaches. A first approach assigns known climate indices as the nodes of the network (Tsonis et al. 2007; Swanson et al. 2009; Wang et al. 2009). By studying the collective behavior of these nodes, it has been possible to investigate their relative role over time and to interpret climate shifts in terms of changes in their relative strength. This approach is obviously sensitive to the initial selection of network nodes, and it cannot be used to discover new climate phenomena involving other regions.

A second, and more common, approach represents the nodes of the climate network by grid cells in the given climate field. Specifically, each grid cell is represented by a node, and edges between nodes correspond to statistically significant relations based on linear or nonlinear correlation metrics (Tsonis and Roebber 2004; Donges et al. 2009b). In this approach, it is common to prune edges whose statistical significance is below a certain threshold, and to assume that all remaining edges are equally "strong", resulting in an unweighted network (Tsonis et al. 2008; Donges et al. 2009b; Steinhaeuser et al. 2009). This approach has been used to study teleconnections, uncover interesting global-scale patterns responsible for the transfer of energy throughout the oceans, and analyze relations between different variables in the atmosphere (Tsonis et al. 2008; Tsonis and Swanson 2008; Yamasaki et al. 2008; Donges et al. 2009a, 2011). A limitation of this approach is that it results in a very large number of network nodes (all cells in a spatial grid), and these nodes cannot be used to describe *parsimoniously* any identified climate phenomena.

The third approach focuses on the community structure of the underlying network (Newman and Girvan 2004). A community is a collection of nodes that are highly interconnected, while having much fewer interactions with the rest of the network. Communities can serve as informative predictors in lieu of climate indices (Tsonis et al. 2010; Steinhaeuser et al. 2011a; Pelan et al. 2011), while their evolution and stability has also received some attention (Steinhaeuser et al. 2009, 2011b). Clustering techniques have also been proposed to discover significant geographical regions in a given climate field (again, in lieu of climate indices) (Steinbach and Tan 2003), and to identify dipoles (i.e., two regions whose anomalies are anti-correlated) and to evaluate their significance (Kawale et al. 2011, 2012). These community-based or clustering techniques, however, do not infer a network of teleconnections between different communities (clusters), and they do not quantify the intensity of teleconnections between geographically separated regions within the same community (cluster).

In this work, we propose a new method to apply network analysis to climate science. We first apply a novel network-based clustering method to group the initial set of grid cells in "areas", i.e., in geographical regions that are highly homogeneous in terms of the underlying climate variable. These areas represent the nodes of the inferred network. Links between areas (i.e., the edges of the network) represent non-local dependencies between different regions over a certain time period. These inter-area links are weighted, and their magnitude depends on both the cumulative anomaly of each area and the cross-correlation between the two cumulative anomalies. The similarity of our method to previous community/clustering techniques is that nodes are endogenously determined during the data analysis process. The main differences are that each node corresponds to a distinct geographical region, and these nodes form a weighted network based on the connection intensity that is inferred for each pair of nodes. In other words, the proposed method decouples the identification of the geographical boundary of each network node from the estimation of the connection intensity between different regions.

The proposed method requires a single parameter $\tau$, which determines the minimum degree of homogeneity between cells of the same area. The method is robust to additive noise, changes in the resolution of the given data set, the selection of the correlation metric, and variations in $\tau$. The resulting climate network can be applied, regionally or globally, to identify and quantify relationships between climate areas (or teleconnections) and their representation in models, and to investigate climate variability and shifts. Finally, the proposed method can be extended to investigate interactions between different climate variables.

The rest of this paper is organized as follows: In Sect. 2 we introduce the data sets analyzed in this work. We describe the climate network construction algorithm and the network analysis metrics in Sects. 3 and 4, respectively. The robustness of the climate network inference process is examined in Sect. 5. Applications of the proposed method to a suite of reanalyses and model data sets are presented in Sect. 6. A discussion of the main outcome of this work concludes the paper.

## 2 Data sets

In this section we briefly describe the data sets that are used in the rest of this paper. For sea surface temperatures (SSTs), we construct and compare networks based on the

HadISST (Rayner et al. 2003), the ERSST-V3 (Smith et al. 2008) and the NCEP/NCAR (Kalnay et al. 1996) reanalyses. For precipitation, we rely on CMAP merged data (Xie and Arkin 1997) and ERA-Interim reanalysis (Dee et al. 2011). We also analyze the SST fields generated by two coupled general circulation models chosen from the CMIP5 archive: the NASA GISS-E2H (Hansen et al. 2002) and the Hadley Center HadCM3 (Gordon et al. 2000). We select randomly two runs of each model from the "historical run" ensembles (Taylor et al. 2012).

Because the quality of the measurements contributing to the SST reanalyses deteriorates as we move to higher latitudes, we only consider the latitudinal range of [60°N; 60°S], avoiding sea-ice covered regions. Also, we mostly focus on the period 1979–2005; in the case of HadISST reanalysis, we contrast with the network characteristics during the 1950–1976 interval. Due to space constraints, results are only shown for the boreal winter season (December to February, DJF). When not specified otherwise, all SST data are interpolated (using bilinear interpolation) to the minimum common spatial resolution across all data sets (2° × 2.5°); for precipitation the resolution is 2.5° × 2.5°.

All climate networks are constructed from detrended anomalies derived from monthly averages of the corresponding climate field. The detrending is done using linear regression and the anomalies are computed after removing the annual cycle.

## 3 Climate network construction

The network construction process consists of three steps. First, we compute the "cell-level network" from the detrended anomaly time series of each cell in the spatial grid. Second, we apply a novel *area identification algorithm* on the cell-level network to identify the nodes of the final "area-level network"; an area here represents a geographic region that is highly homogeneous in terms of the given climate field. Third, we compute the weight of the edges between areas, roughly corresponding to teleconnections, based on the covariance of the cumulative anomalies of the two corresponding areas. The following network construction method requires a single parameter, $\tau$, which determines the minimum degree of homogeneity between cells of the same area. In the following we describe each step in more detail.

### 3.1 Cell-level network

Consider a climate field $\mathbf{x}(\mathbf{t})$ defined on a finite number of cells in a given spatial grid. The $i$'th vector of the climate field is a time series $x_i(t)$ of detrended anomalies in cell $i$. The length of each time series is denoted by $T$. We first compute Pearson's cross-correlation $r(x_i, x_j)$[1] between the time series $x_i(t)$ and $x_j(t)$ for every pair of cells $i$ and $j$. We calculate the correlations at zero-lag, assuming that the physical processes linking different cells result from atmospheric wave dynamics and are fast compared to the *1-month averaging time scale* of the input time series. Considering time-lagged correlations is beyond the scope of this paper. Instead of using Pearson's correlation, other correlation metrics could be adopted; in Sect. 5.4 we examine the differences in the resulting network using a rank-based correlation metric.

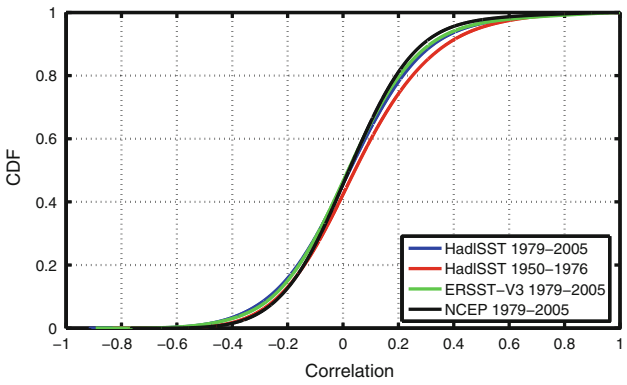


**Fig. 1** Empirical cumulative distribution functions (CDF) of correlations for the HadISST reanalysis during the 1950–1976 and 1979–2005 periods, and for the ERSST-V3 and NCEP reanalyses during the 1979–2005 period

Most of the prior work on climate network analysis applies a cutoff threshold on the correlations $r(x_i, x_j)$ to prune insignificant values and construct a binary (i.e., unweighted) network between cells; for a recent review see Steinhaeuser et al. (2010). Figure 1 shows correlation distributions for four SST reanalyses; note that there is no natural cutoff point to separate significant correlations from noise. We have experimented with methods that first prune insignificant correlations and then construct unweighted networks, and observed that the final area-level network is sensitive to the significance level at which correlations are pruned. Such sensitivity complicates any attempt to make quantitative comparisons between networks constructed from different data sets (for example networks from observations versus models).

For this reason, in the following we present a method that considers *all* pair-wise cell correlations, without any pruning. Thus, the cell-level network is a *complete and weighted* graph, meaning that every pair of cells is connected but with weighted edges between −1 and 1. This cell-level network is the input to the area identification algorithm, described next.

[1] Unless specified otherwise, the term "correlation" will be used to denote Pearson's cross-correlation metric between two time series.

### 3.2 Identification of climate areas

A central concept in the proposed method is that of a *climate area*, or simply *area*. Informally, an area $A$ represents a geographic region that is highly homogeneous in terms of the climate field $\mathbf{x}(\mathbf{t})$.

In more detail, we define as *neighbors* of a grid cell $i$ the four adjacent cells of $i$, and as *path* a sequence of cells such that each pair of successive cells are neighbors. An area $A$ is a set of cells satisfying three conditions:

1. $A$ includes at least two cells.
2. The cells in $A$ form a connected geographic region, i.e., there is a path within $A$ connecting each cell of $A$ to every other cell of that area.
3. The average correlation between all cells in $A$ is greater than a given threshold $\tau$,

$$\frac{\sum_{i \neq j \in A} r(x_i, x_j)}{|A| \times (|A| - 1)} > \tau \tag{1}$$

where $|A|$ denotes the number of cells in area $A$.

The parameter $\tau$ determines the minimum degree of homogeneity that is required within an area. A heuristic for the selection of $\tau$ is presented in "Appendix 1"; we use that heuristic in the rest of this paper.

For the climate network to convey information in the most parsimonious way, *the number of identified climate areas should be minimized*. We have shown elsewhere that this computational problem is NP-Complete, meaning that there exists no efficient way to solve it in practice (Fountalis et al. 2013). Consequently, we have designed an algorithm that aims to *minimize the number of areas* heuristically, based on a so called "greedy" approach (Cormen et al. 2001). The algorithm consists of two parts. First, it identifies a set of areas; secondly it merges some of those areas together as long as they satisfy the previous three area constraints. A pseudocode describing the algorithm is given in "Appendix 2", while the actual software is available at http://www.cc.gatech.edu/~dovrolis/ClimateNets/. An example of the area identification process applied to a synthetic grid is illustrated in Fig. 2.

The identification part of the algorithm produces areas that are geographically connected by always expanding an area through neighboring cells. Additionally, the algorithm attempts to identify the largest (in terms of number of cells) area in each iteration by selecting, in every expansion step, the neighboring cell that has the highest average correlation with existing cells in that area. The expectation is that this greedy approach allows the area to expand to as many cells as possible, subject to the constraint that the average correlation in the area should be more than $\tau$. It is easy to see that an identified area satisfies the condition given by Eq. 1.

Within the set of areas $V$ identified by the first part of the algorithm, it is possible to find some areas that can be merged further, and still satisfy the previous three constraints. Specifically, we say that two areas $A_i$ and $A_j$ can be merged into a new area $A_k = A_i \cup A_j$ if $A_i$ and $A_j$ have at least one pair of geographically adjacent cells and the average correlation of cells in $A_k$ is greater than $\tau$. The second part of the algorithm, therefore, attempts to merge as many areas as possible (see "Appendix 2").

Figure 3 shows the identified areas before merging (i.e., after Part-1 in "Appendix 2") and after merging (i.e., after Part-2 in "Appendix 2") for the HadISST reanalysis. Figure 3c shows the distribution of area sizes (in number of cells) before and after merging. Area merging decreases substantially the number of small areas (the percentage of areas with less than 10 cells in this example drops from 46 to 10 %).

The identified areas represent the nodes of the inferred climate network. We refer to this network as "area-level network" to distinguish it from the underlying cell-level network.

### 3.3 Links between areas

*Links* (or edges) between areas identify non-local relations and can be considered a proxy for climate teleconnections. To quantify the weight of these links, we first compute for each area $A_k$ the *cumulative anomaly* $X_k(t)$ of the cells in that area,

$$X_k(t) = \sum_{i \in A_k} x_i(t) \cos(\phi_i). \tag{2}$$

The anomaly time series of a cell $i$ is weighted by the cosine of the cell's latitude ($\phi_i$), to account for the cell's relative size.[2] As a sum of zero-mean processes, a cumulative anomaly is also zero-mean.

Figure 4 quantifies the relation between the size of the areas ($\sum_{i \in A_k} \cos(\phi_i)$) identified earlier in the HadISST data set and the standard deviation of their cumulative anomaly. Note that the relation is almost linear, at least excluding the largest 3–4 areas. Exact linearity would be expected if all cells had the same size, their anomalies had the same variance, and every pair of cells in the same area had the same correlation. Even though these conditions are not true in practice, it is interesting that the standard deviation of an area's cumulative anomaly is roughly proportional to its size.

The strength, or weight, of the link between two areas $A_i$ and $A_j$ is captured by the *covariance* of the corresponding cumulative anomalies $X_i(t)$ and $X_j(t)$. Specifically, every

[2] When comparing data sets with different spatial resolution, the anomaly of a cell should be normalized by the size of the cell in that resolution.

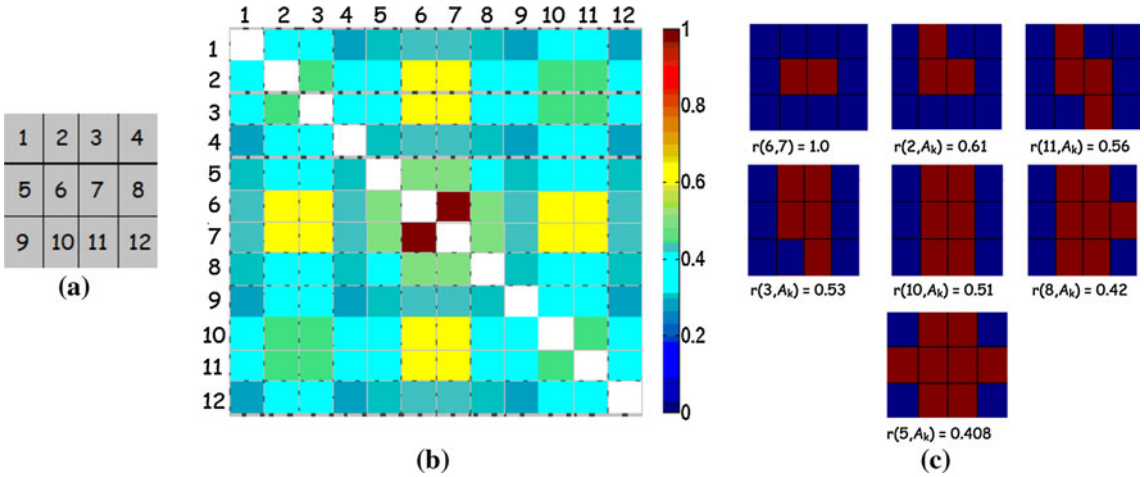


**Fig. 2** An example of the area identification algorithm. **a** 12-cell synthetic grid. **b** The correlation matrix between cells (given as input). **c** The area expansion process for a given $\tau = 0.4$. Cells shown in *red* are selected to join the area (denoted by $A_k$). Cells 1,4,9 and 12 will not join $A_k$ since they do not satisfy the $\tau$ constraint in Eq. 1

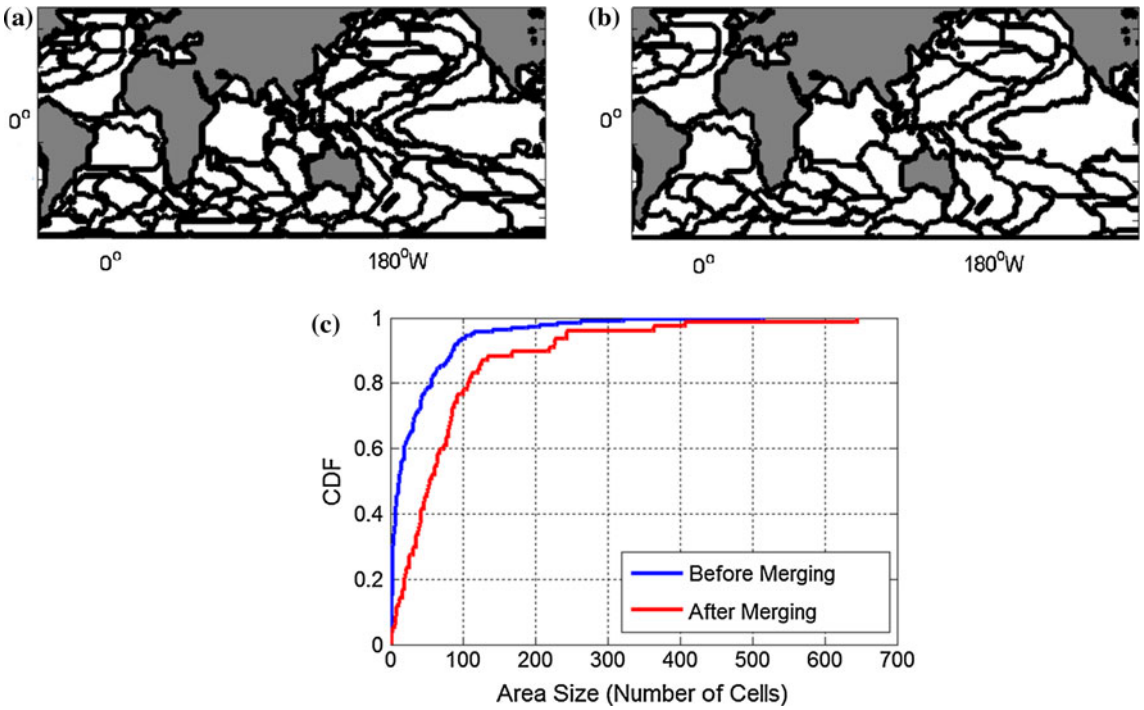


**Fig. 3** Identified areas in the HadISST 1979–2005 data set ($\tau = 0.496$). **a** The 176 areas identified by Part-1 of the area identification algorithm. **b** The 74 "merged" areas after the execution of Part-2. **c** The CDF of area sizes (in number of cells) before and after the merging process

pair of areas $A_i$ and $A_j$ in the constructed network is connected with a link of weight $w(A_i, A_j)$,

$$w(A_i, A_j) \triangleq w(X_i, X_j) = cov(X_i, X_j) = s(X_i)\, s(X_j)\, r(X_i, X_j) \tag{3}$$

where $s(X_i)$ is the standard deviation of the cumulative anomaly $X_i(t)$, while $cov(X_i, X_j)$ and $r(X_i, X_j)$ are the covariance and correlation, respectively, of the cumulative anomalies $X_i(t)$ and $X_j(t)$ that correspond to areas $A_i$ and $A_j$. Note that the weight of the link between two areas does not depend only on their (normalized) correlation $r(X_i, X_j)$, but also on the "power" of the two areas, as captured by the standard deviation of the corresponding cumulative anomalies. Also, recall from the previous paragraph that this standard deviation is roughly proportional to the area's size, implying that larger areas will tend to have stronger connections. The link between two areas can be positive or negative, depending on the sign of the correlation term.

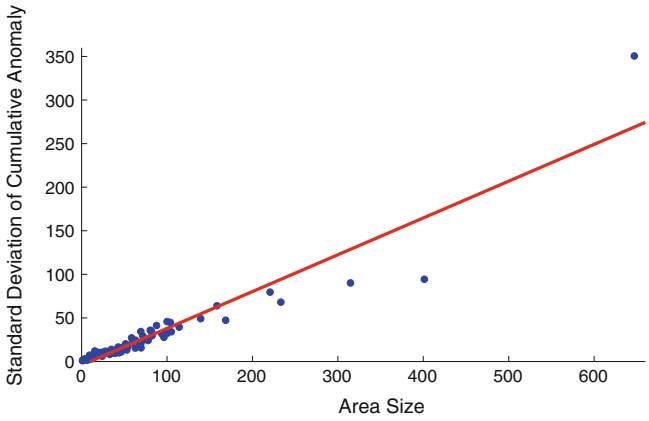


**Fig. 4** The relation between area size and standard deviation of the area's cumulative anomaly ($R^2 = 0.88$) for the HadISST reanalysis during the 1979–2005 period; $\tau = 0.496$

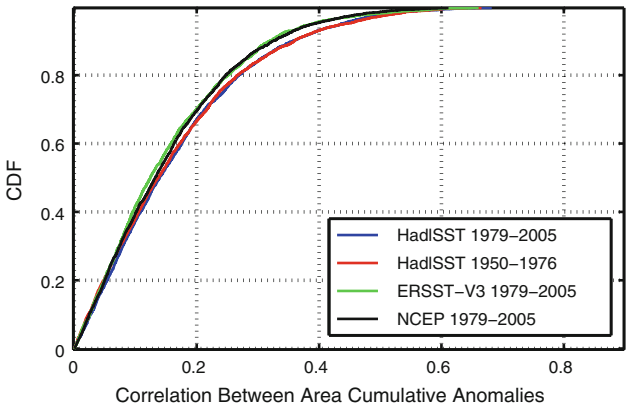


**Fig. 5** CDF of the absolute correlation between area cumulative anomalies for the HadISST reanalysis during the 1950–1976 and 1979–2005 periods, and for ERSST-V3 and NCEP during the 1979–2005 period

Figure 5 presents the cumulative distribution function (CDF) of the absolute correlation between the cumulative anomalies of areas for four SST networks. As with the correlations of the cell-level network, there is no clear cutoff[3] separating significant correlations from noise. For this reason we prefer to not prune the weaker links between areas. Instead, every pair of areas $A_i$ and $A_j$ is connected through a weighted link and the resulting graph is *complete*.

## 4 Network metrics

We now proceed to define a few network metrics that are used throughout the paper. A climate network $N$ is defined by a set $V$ of areas $A_1, \ldots, A_{|V|}$, representing the nodes of the network, and a set of link weights, given by Eq. 3.

[3] Imposing a threshold on the actual strength of the link (computed as the covariance between the cumulative anomalies of two areas) would be incorrect. For example, multiplying low correlations with large standard deviations can produce links of significant weight.

Because the network is a complete weighted graph, basic graph theoretic metrics that do not account for link weights (such as average degree, average path length, or clustering coefficient) are not relevant in this context.

A first representation of the network can be obtained through *link maps*. The link map of an area $A_k$ shows the weight of the links between $A_k$ and every other area in the network. Link maps provide a direct visualization of the correlations, positive and negative, between a given area and others in the system, often related to atmospheric teleconnection patterns. For instance, Fig. 6 shows link maps for the two largest areas identified in the HadISST network in the 1979–2005 period. The first area has a clear correspondence to the El Niño Southern Oscillation (ENSO); indeed, the cumulative anomaly over that area and most common indices that describe ENSO variability are highly correlated (the correlation reaches 0.94 for the Niño-3.4 index). The links of this "ENSO" area depict known teleconnections and their strength. The second largest area covers most of the tropical Indian Ocean and represents the region that is most responsive to interannual variability in the Pacific. It corresponds, broadly, to the region where significant warming is observed during peak El Niño conditions (Chambers et al. 1999).

Another metric is the *strength* of an area (also known as weighted degree), defined as the sum of the absolute link weights of that area,

$$W(A_i) = \sum_{j \neq i}^{V} |w(A_i, A_j)| = s(X_i) \sum_{j \neq i}^{V} s(X_j) |r(X_i, X_j)|. \quad (4)$$

Note that anti-correlations (negative weights) also contribute to an area's strength. Figure 7 shows, for example, the strength maps for two HadISST networks covering the 1950–1976 and 1979–2005 periods, respectively. Both the geographical extent of areas and their strength display differences in the two time intervals, particularly in the North Pacific sector and in the tropical Atlantic (Miller et al. 1994; Rodriguez-Fonseca et al. 2009).

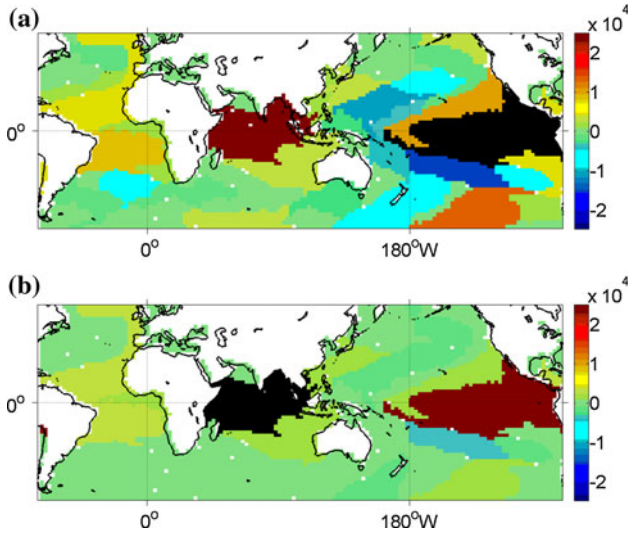


**Fig. 6** Link maps for two areas related to **a** ENSO and **b** the equatorial Indian Ocean in the HadISST 1979–2005 network ($\tau = 0.496$). The *color scale* represents the weight of the link between the area shown in black and every other area in this SST network

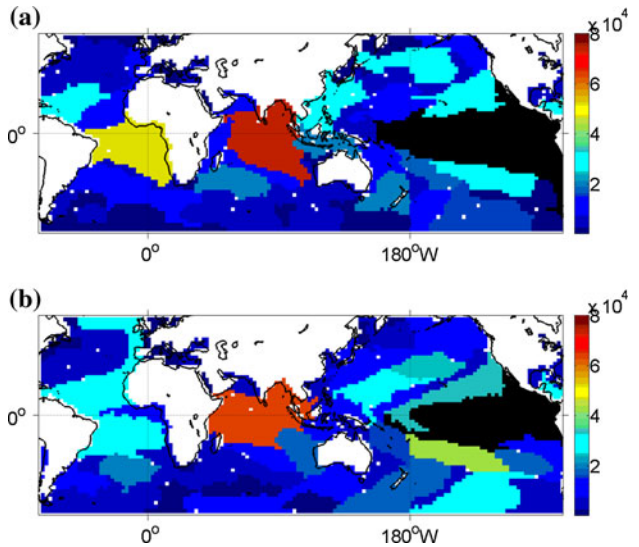


**Fig. 7** Strength maps for two different time periods using the HadISST data set. **a** 1950–1976 network, strength of ENSO area: $20.1 \times 10^4$; **b** 1979–2005 network, strength of ENSO area: $18.8 \times 10^4$

It is often useful to "peel" the nodes of a network in successive layers of increasing network significance. For weighted networks, we can do so through an iterative process referred to as *s-core* decomposition (Van den Heuvel and Sporns 2011). The areas of the network are first ordered in terms of their strength. In iteration-1 of the algorithm, the area with the minimum strength, say $W_{min}$, is removed. Then we recompute the (reduced) strength of the remaining areas, and if there is an area with lower strength than $W_{min}$, it is removed as well. Iteration-1 continues in this manner until there is no area with strength less than $W_{min}$. The areas removed in this first iteration are placed in the same layer. The algorithm then proceeds similarly with iteration-2, forming the second layer of areas. The algorithm terminates when we have removed all areas, say after $K$ iterations. Finally, the $K$ layers are re-labeled as "cores" in inverse order, so that the *first order core* consists of the areas removed in the last iteration (the strongest network layer), while the *Kth order core* consists of the areas removed in the first iteration (the weakest layer). Figure 8 shows the *top five order cores* for two HadISST networks, covering 1950-1976 and 1979–2005, respectively. Again, changes in the relative role of areas are apparent in the North Pacific and in the tropical Atlantic.

Visual network comparisons provide insight but quantitative metrics that summarize the distance between two networks into a single number would be useful. A challenge is that the climate networks under comparison may have a different set of areas, and it is not always possible to associate an area of one network with a unique area of another network.

We rely on two quantitative metrics: the adjusted Rand index (ARI), which focuses on the similarity of two networks in terms of the identified areas, and the *Area Strength Distribution Distance*, or simply *Distance* metric, which considers the magnitude of link weights and thus area strengths.

The (non-adjusted) Rand index is a metric that quantifies the similarity of two partitions of the same set of elements into non-overlapping subsets or "clusters" (Rand 1971).

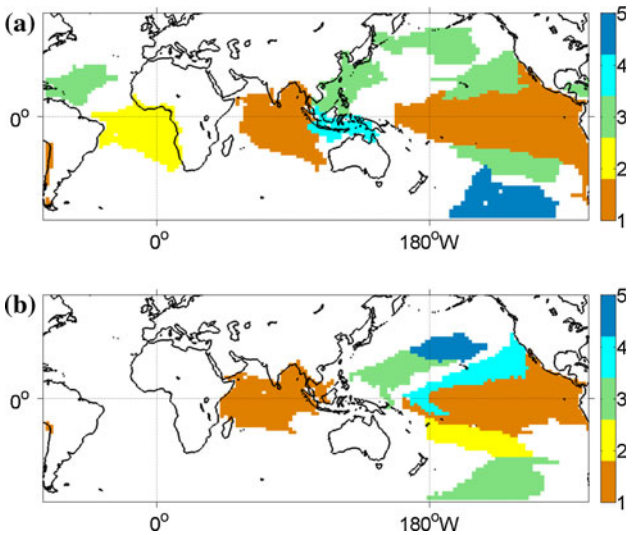


**Fig. 8** Color maps depicting the *top five order cores* for the **a** HadISST 1950–1976, and **b** HadISST 1979–2005 networks

Every pair of elements that belong to the same cluster in both partitions, or that belong to different clusters in both partitions, contributes positively to the Rand index. Every pair of elements that belong to the same cluster in one partition but to different clusters in the other partition, contributes negatively to the Rand index. The metric varies between 0 (complete disagreement between the two partitions) to 1 (complete agreement). A problem with the Rand index is that two random partitions would probably give a positive value because some agreement between the two partitions may result by chance. The adjusted Rand index (Hubert and Arabie 1985; Steinhaeuser and Chawla 2010) ensures that the expected value of ARI in the case of random partitions is 0, while the maximum value is still 1. We refer the reader to the previous references for the ARI mathematical formula.

In the context of our method, the common set of elements is the set of grid cells, while a partition represents how cells are classified into areas (i.e., each area is a cluster of cells). Cells that do not belong to any area are assigned to an artificial cluster that we create just for computing the ARI metric. We use the ARI metric to evaluate the similarity of two networks in terms of the identified areas. This metric, however, does not consider cell anomalies and cell sizes, and so it cannot capture similarities or differences between two networks in terms of link weights and area strengths. Two networks may have some differences in the number or spatial extent of their areas, but they can still be similar if those "ambiguously clustered" cells do not have a significant anomaly compared to their area's anomaly. Also, two networks can have similar areas but the magnitude of their area anomalies can differ significantly, causing significant differences in link weights and thus area strengths. Further, the ARI metric cannot be used to compare data sets with different resolution because the underlying set of cells in that case would be different between the two networks.

For these reasons, together with the ARI, we rely on a distance metric that is based on the area strength distribution of the two networks. The strength of an area, in effect, summarizes the combined effect of the area's spatial scope (which cells participate in that area), and of the anomaly and size of those cells.

Given two networks $N$ and $N'$ with $V$ and $V' \leq V$ areas, respectively, we first add $V - V'$ "virtual" areas of zero strength in network $N'$ so that the two networks have the same number of nodes. Then, we rank the areas of each network in terms of strength, with $A_i$ being the $i$'th highest-strength area in network $N$. Figure 9a shows the ranked area strength distributions for the HadISST networks covering 1950–1976 and 1979–2005 periods. The distance $d(N, N')$ quantifies the similarity between two networks in terms of their ranked area strength distribution,

$$d(N, N') = \sum_{i=1}^{V} |W(A_i) - W(A'_i)| \tag{5}$$

To normalize the previous metric, we introduce the *relative distance* $D(N, N')$. Specifically, we construct an ensemble of randomized networks $N_r$ with the same number of areas and link weight distribution as network $N$, but with random assignment of links to areas. The random variable $d(N, N_r)$ represents the distance between $N$ and a random network $N_r$, while $\overline{d(N, N_r)}$ denotes the sample average of this distance across 100,000 such random networks. The relative distance $D(N, N')$ is then defined as

$$D(N, N') = \frac{d(N, N')}{\overline{d(N, N_r)}}. \tag{6}$$

Note that $D(N, N')$ represents an ordered relation, from network N to N′. A relative distance close to 0 implies that $N'$ is similar to $N$ in terms of the allocation of link weights to areas. As the relative distance approaches 1, $N'$ may have a similar link weight distribution with $N$, but the two networks differ significantly in the assignment of links to areas. The relative distance can be larger than 1 when $N'$'s link weight distribution is significantly different than that of $N$.

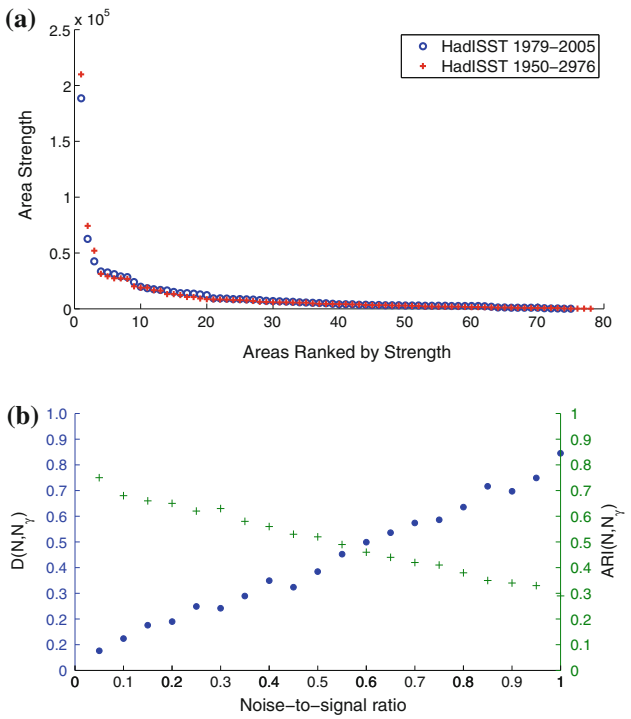


**Fig. 9** **a** Distribution of ranked area strengths for two networks constructed using the HadISST data set over the periods 1950–1976 and 1979–2005, respectively. **b** Distance $D(N, N_\gamma)$ and ARI$(N, N_\gamma)$ between the HadISST 1979–2005 network and networks constructed after the addition of white Gaussian noise in the same data set

Two networks may be similar in terms of the identified areas (high ARI) but with large distance (high $D$) if the strength of at least some areas is significantly different across the two networks (perhaps due to the magnitude of the underlying cell anomalies). In principle, it could also be that two networks have similar ranked area strength distributions (low $D$) but significant differences in the number or spatial extent of the identified areas (low ARI). Consequently, the joint consideration of both metrics allows us to not only evaluate or rank pairs of networks in terms of their similarity, but also to understand which aspects of those pairs of networks are similar or different.

We can also map a distance $D(N, N')$ to an amount of White Gaussian Noise (WGN) that, if added to the climate field that produced $N$, will result in a network with equal distance from $N$. In more detail, let $s^2(x_i)$ be the sample variance of the anomaly time series $x_i(t)$ in the climate field under consideration. We construct a perturbed climate field by adding WGN with variance $\gamma \times s^2(x_i)$ to every $x_i(t)$, where $\gamma$ is referred to as the *noise-to-signal ratio*. Then, we construct the corresponding network $N_\gamma$, and $D(N, N_\gamma)$ is its distance from $N$. A given distance $D(N, N')$ can be mapped to a noise-to-signal ratio $\gamma$ when $D(N, N') = D(N, N_\gamma)$. Similarly, a given ARI value ARI$(N, N')$ can be mapped to noise-to-signal ratio $\gamma$ such that ARI$(N, N')$ = ARI$(N, N_\gamma)$. Figure 9b shows how $\gamma$ affects $D(N, N_\gamma)$ and ARI$(N, N_\gamma)$ when the network $N$ corresponds to the HadISST 1979–2005 reanalysis. As a reference point, note that a low noise magnitude, say $\gamma = 0.1$, corresponds to distance $D = 0.12$ and ARI = 0.68.

Finally, we emphasize that the ARI and $D$ metrics focus on the global scale. Even if two networks are quite similar according to these two metrics, meaningful differences at the local scale of individual areas may still exist. The study of regional climate effects may require an adaptation of these metrics.

## 5 Robustness analysis

Analyzing climate data poses many challenges: measurements provide only partial geographical and temporal coverage, while the collected data are subject to instrumental biases and errors both random and systematic. Greater uncertainties exist in general circulation model outputs: climate simulations are dependent on modeling assumptions, complex parameterizations and implementation errors. An important question for any method that identifies topological properties of climate fields is whether it is robust to small perturbations in the input data, the method parameters, or in the assumptions the method is based on. If so, the method can provide useful information on the climate system despite uncertainties of various types. In this section, we examine the sensitivity of the inferred networks to deviations in the input data, the parameter $\tau$, and certain methodological choices. We quantify sensitivity by computing the $D$ and ARI metrics from the original network to each of the perturbed networks.

### 5.1 Robustness to additive white Gaussian noise

As described in Sect. 4, a simple way to perturb the input data is to add white Gaussian noise to the original climate field time series. The magnitude of the noise is controlled by the *noise-to-signal ratio* $\gamma$. The distance $D$ and ARI from the original network $N$ to the "noisy" networks $N_\gamma$ are shown in Fig. 9b for the HadISST reanalysis over 1979–2005. To visually illustrate how noise affects the identified areas, and in particular their strength, Fig. 10 presents strength maps for two values of $\gamma$; the area strengths should be compared with Fig. 7b. Although some differences exist, the ENSO area strength is comparable to that of the original network, and the hierarchy (in terms of strength) in the three basins is conserved.

### 5.2 Robustness to the resolution of the input data set

All data sets compared in this paper have been spatially interpolated to the lowest common resolution. Here we investigate the robustness of the identified network to the resolution of the input data set. To do so, consider the HadISST reanalysis over the 1979–2005 period and compare the network discussed so far, constructed using data

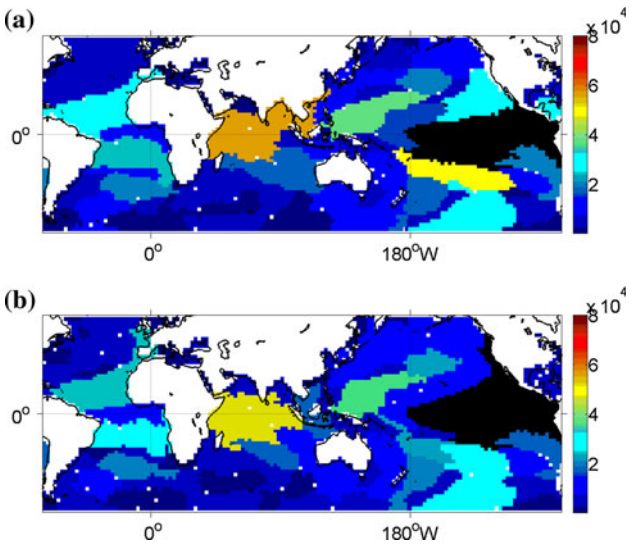


**Fig. 10** Strength maps for two perturbations of the HadISST 1979–2005 data set using white Gaussian noise. **a** $\gamma = 0.05$, strength of ENSO area: $18.0 \times 10^4$. **b** $\gamma = 0.10$, strength of ENSO area: $19.1 \times 10^4$

interpolated on a 2° lat × 2.5° lon grid, with two networks based on a lower (4° lat × 4° lon) and a higher (1° lat × 2° lon) resolution realization of the same reanalysis. Figure 11 shows strength maps for the two new networks. As we lower the resolution the total number of areas decreases, and the areas immediately surrounding the ENSO-related area get weaker. Nonetheless, the hierarchy of area strengths in the three basins is preserved, and differences are small, as quantified by the distance metric. The distance from the default to the high resolution network is $D(N, N') = 0.10$ ($\gamma = 0.07$). The distance from the default to the low resolution network is $D(N, N') = 0.11$ ($\gamma = 0.10$). As previously mentioned, the ARI cannot be used to compare data sets with different spatial resolution.

### 5.3 Robustness to the selection of $\tau$

Recall that the parameter $\tau$ represents the threshold for the minimum average pair-wise correlation between cells of the same area. Even though we provide a heuristic (see "Appendix 1") for the selection of $\tau$, which depends on the given data set, it is important to know whether small deviations in $\tau$ have a major effect on the constructed networks.

Considering again the HadISST 1979–2005 reanalysis, Fig. 12 presents the relative distance and ARI from the original network $N$ constructed using $\tau = 0.496$ (it corresponds to a significance level $\alpha = 0.1$ %), to networks $N_\tau$ constructed using different $\tau$ values. We vary $\tau$ by ± 10 %, in the range 0.45–0.55. This corresponds to a large change, roughly an order of magnitude, in the underlying significance level $\alpha$.

Figure 13 visualizes strength maps for the two extreme values of $\tau$ in the previous range. While some noticeable differences exist, the overall area structure appears robust to the choice of $\tau$. By increasing $\tau$, we increase the required degree of homogeneity within an area, and therefore the resulting network will be more fragmented, with more areas of smaller size and lower strength, and vice versa for decreasing $\tau$.

### 5.4 Robustness to the selection of the correlation metric

The input to the network construction process is a matrix of correlation values between all pairs of cells. So far, we

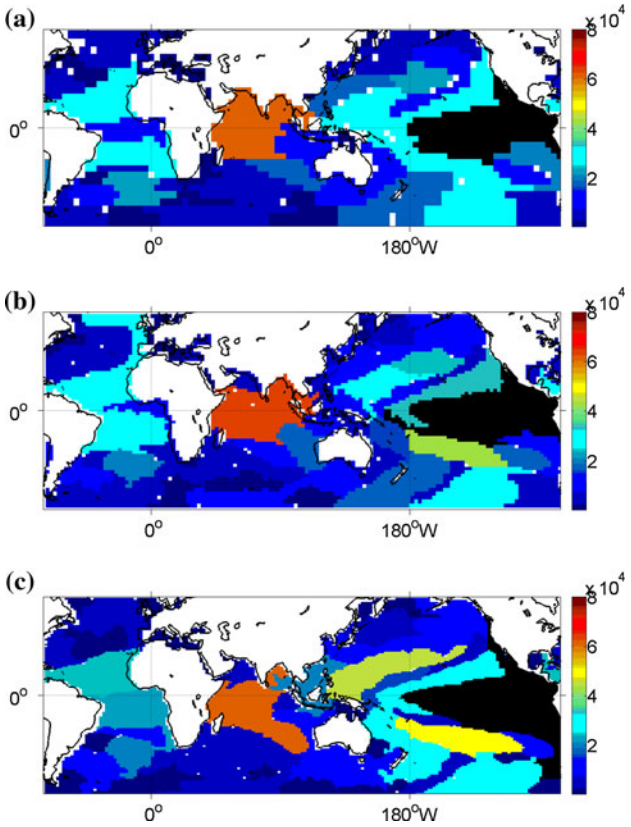


**Fig. 11** Strength maps for the HadISST 1979–2005 network at three different resolutions. **a** Low resolution network, (4° lat × 4° lon), strength of ENSO area: $18.2 \times 10^4$. **b** Default resolution network, (2° lat × 2.5° lon), strength of ENSO area: $18.8 \times 10^4$. **c** High resolution network, (1° lat × 2° lon), strength of ENSO area: $18.2 \times 10^4$

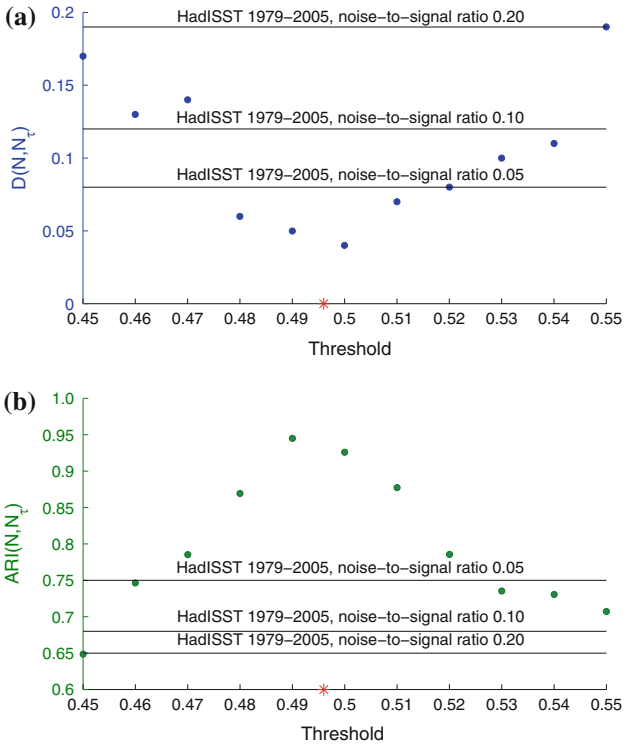


**Fig. 12** **a** Distance $D$ and **b** ARI from the original HadISST 1979–2005 network (marked with * in the x-axis, $\tau = 0.496$) to networks constructed with different values of $\tau$. The *black horizontal lines* correspond to the distance $D(N, N_\gamma)$ and $\mathrm{ARI}(N, N_\gamma)$

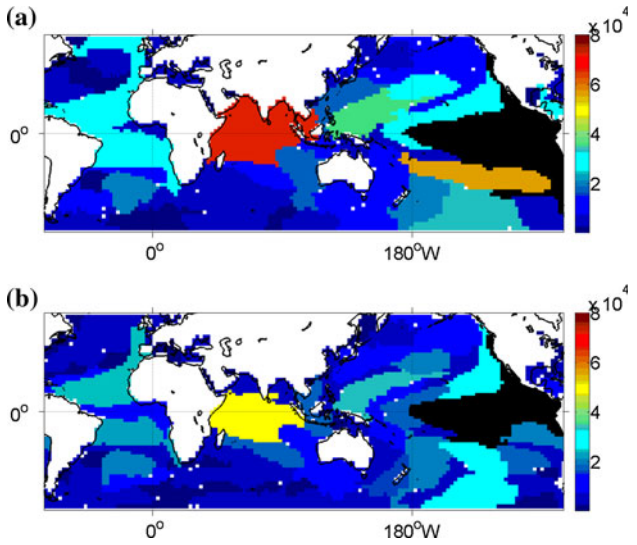


**Fig. 13** Strength maps for the HadISST 1979–2005 network using two values of the parameter $\tau$. The "default" value is $\tau = 0.496$, corresponding to $\alpha = 0.1$ % (see "Appendix 1"). **a** $\tau = 0.45$, strength of ENSO area: $18.7 \times 10^4$. **b** $\tau = 0.55$, strength of ENSO area: $18.6 \times 10^4$

have relied on Pearson's correlation coefficient, which is a linear dependence measure between two random variables. Any other correlation metric could be used instead. To verify that the properties of the resulting network do not depend strongly on the selected correlation metric, we use here the non-parametric *Spearman's rank coefficient* to compute cell-level correlations.

Figure 14 shows the strength map for the HadISST 1979–2005 network using Spearman's correlation metric. Again, while small changes are apparent, the size and shape of the major areas and their relative strength are unaltered. $D(N, N') = 0.08$ and $\mathrm{ARI}(N, N') = 0.76$, where $N$ is the network shown in Fig. 7b; both metrics correspond to $\gamma = 0.05$.

We have performed similar robustness tests using precipitation data obtaining comparable results.

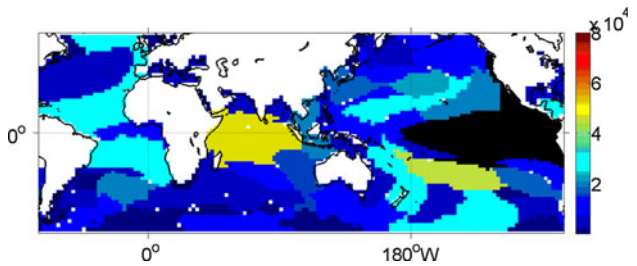


**Fig. 14** Strength map for the HadISST 1979–2005 network using Spearman's correlation; strength of ENSO area: $18.5 \times 10^4$

## 6 Applications

We now apply the proposed method to the climate data sets described in Sect. 2 to illustrate that network analysis can be successfully used to compare data sets and to validate model representations of major climate areas and their connections. We proceed by constructing networks for three different SST reanalyses and two precipitation data sets. We then examine the relation between two different climate fields (SST and precipitation) introducing a *regression of networks* technique. Finally, we analyze the network structure of the SST fields from two coupled climate models participating in CMIP5.

### 6.1 Comparison of SST networks

Here we investigate the network properties and metrics for three SST reanalyses focusing on the 1979–2005 period. Two of them, HadISST and ERSST-V3, use statistical methods to fill sparse SST observations; HadISST implements a reduced space optimal interpolation (RSOI) technique, while ERSST-V3 adopts a method based on empirical orthogonal function (EOF) projections. NCEP/NCAR uses the Global Sea Ice and Sea Surface Temperatures (GISST2.2) from the UK. Meteorological Office until late 1981 and the NCEP Optimal Interpolation (OI) SST analysis from November 1981 onward. The GISST2.2 is based on empirical orthogonal function (EOF) reconstructions (Hurrell and Trenberth 1999). The OI SST analysis technique combines in situ and satellite-derived SST data (Reynolds and Smith 1994). To minimize the possibility of artificial trends, and the bias introduced by merging different data sets, GISST2.2 data are modified to include an EOF expansion based on the OI analysis from January 1982 to December 1993.

In Fig. 15, we quantify the differences between the three reanalyses showing correlation maps between the detrended DJF SST anomaly time series for HadISST and ERSST-V3, HadISST and NCEP, and ERSST-V3 and NCEP. The patterns that emerge in the all correlation maps are similar. Correlations are generally higher than 0.9 in the equatorial Pacific, due to the almost cloud free sky and to the in-situ coverage provided since the mid 80s' first by the Tropical Ocean Global Atmosphere (TOGA) program, and then by the Tropical Atmosphere Ocean (TAO)/Triangle Trans-Ocean Buoy Network (TAO/TRITON) program (Vidard et al. 2007). Good agreement between reanalyses is also found in the north-east Pacific, in the tropical Atlantic and in the Indian and Pacific Oceans between 10°S and 30°S. Correlations decrease to approximately 0.7 in the equatorial Indian Ocean and around Indonesia, where cloud coverage limits satellite retrievals, and reach values as small as 0.2–0.3 in the Labrador Sea, close to the Bering

Strait and south of 40°S, particularly in the Atlantic and Indian sectors, due to persistent clouds and poor availability of in-situ data. North of 60°N and south of 60°S the presence of inadequately sampled sea-ice and intense cloud coverage reduce even further the correlations, that attain non-significant values almost everywhere. At those latitudes any comparison between those reanalyses and their resulting networks is meaningless given that it would not possible to identify a reference data set.

The strength maps constructed using these data sets show differences in all basins, and suggest that the network analysis performed allows for capturing more subtle properties than correlation maps (Fig. 16). To begin with the strongest area, corresponding to ENSO, we notice that it has a similar shape in HadISST and NCEP, but it extends further to the west in ERSST-V3. Its strength is about 10 % higher in NCEP compared to the other two reanalyses. In HadISST, the equatorial Indian Ocean appears as the second strongest area, followed by areas surrounding the ENSO region in the tropical Pacific and by the tropical Atlantic. In ERSST-V3 the area comprising the equatorial Indian Ocean has shape and size analogous to HadISST, but 30 % weaker, and it is closer in strength to the area covering the warm-pool in the western tropical Pacific. Also the areas comprising the tropical Atlantic are slightly weaker than in the other two data sets. HadISST and ERSST-V3 display a similar strength hierarchy, with the Pacific Ocean being the basin with the strongest (ENSO-like) area, followed by the Indian, and finally by the Atlantic Ocean. In NCEP all tropical areas (except the area corresponding to the ENSO region) have similar strength and the hierarchy between Indian and Atlantic Oceans is inverted. Also, the equatorial Indian Ocean appears subdivided in several small areas.

Differences in strength maps are also reflected in the *s-core* decomposition (Fig. 17) and in the links between the ENSO-related areas and other areas in the network (Fig. 18). In HadISST and ERSST-V3, the *first order core* is located in the tropical and equatorial Pacific and Indian Ocean, while in NCEP it is limited to the Pacific. As a consequence the strength of the link between the ENSO-related area and the Indian Ocean is much stronger in the first two reanalyses than in NCEP. In HadISST, the ENSO-related and Indian Ocean areas are separated by regions of higher order in the western Pacific, organized in the characteristic "horse-shoe" pattern. In the other two reanalyses the *first order core* extends along the whole Pacific equatorial band and includes the horse-shoe areas. In

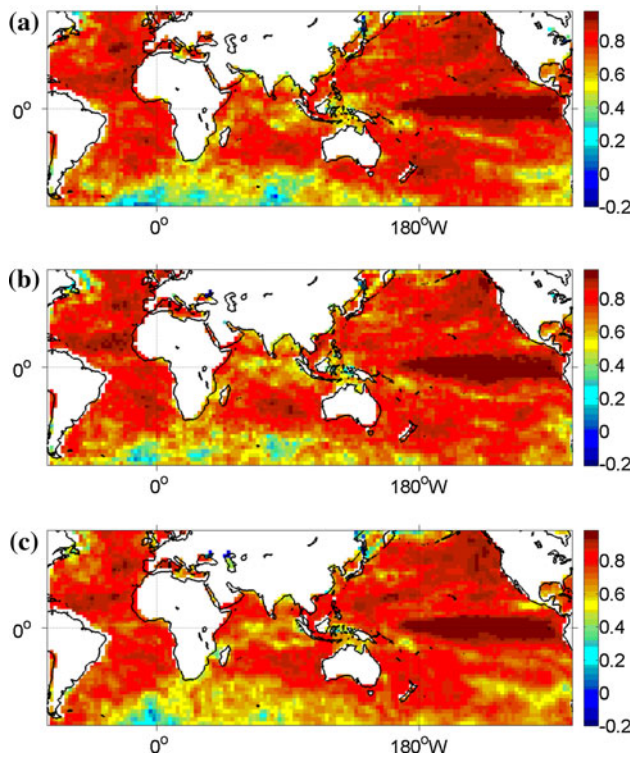


**Fig. 15** Pearson correlation maps between the SST anomaly time series in all pairs of three reanalyses data sets over the 1979–2005 period in boreal winter (DJF). Correlations between **a** HadISST and ERSST-V3; **b** HadISST and NCEP; **c** NCEP and ERSST-V3

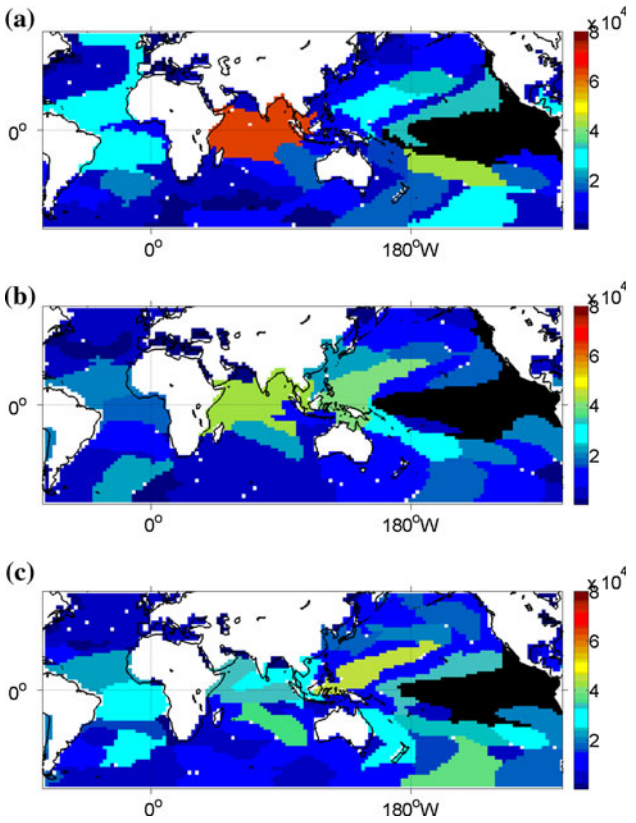


**Fig. 16** Strength maps for networks constructed based on **a** HadISST (ENSO area strength 18.8 $\times$ $10^4$); **b** ERSST-V3 (ENSO area strength 17.6 $\times$ $10^4$); **c** NCEP (ENSO area strength 21.0 $\times$ $10^4$) reanalyses. In all networks the period considered is 1979–2005

correspondence, the links between the ENSO-like and the western Pacific areas are, in absolute value, weaker than the link between ENSO and the Indian Ocean in HadISST, but comparable in ERSST-V3. NCEP shows significantly weaker links overall, but the highest link weights are found between ENSO and the western Pacific.

To conclude the comparison of different SST reanalyses, we measure the distance and ARI values from HadISST to the other two networks. The distance from HadISST to ERSST-V3 is small, $D(N, N') = 0.16$, mapped to a *noise-to-signal* ratio $\gamma = 0.15$. The strongest areas show indeed a good correspondence in strength and size in the two data sets, even if the shapes of the ENSO-related areas differ. The distance from HadISST to NCEP, $D(N, N') = 0.29$ with $\gamma = 0.35$, is greater, as expected from the previous figures, given that all areas except of the ENSO-related one appear significantly weaker, while the ENSO area is stronger than in HadISST. NCEP is also *penalized* because of the differences, compared to HadISST, in the strength (and size) of areas over the Indian Ocean and in the horse-shoe pattern. Recall that $D$ compares areas based on their strength ranking, independent on their geographical location. In this respect, the two strongest areas represented by ENSO and Indian Ocean in HadISST are replaced by ENSO and the North Pacific extension of the horse-shoe region in NCEP.

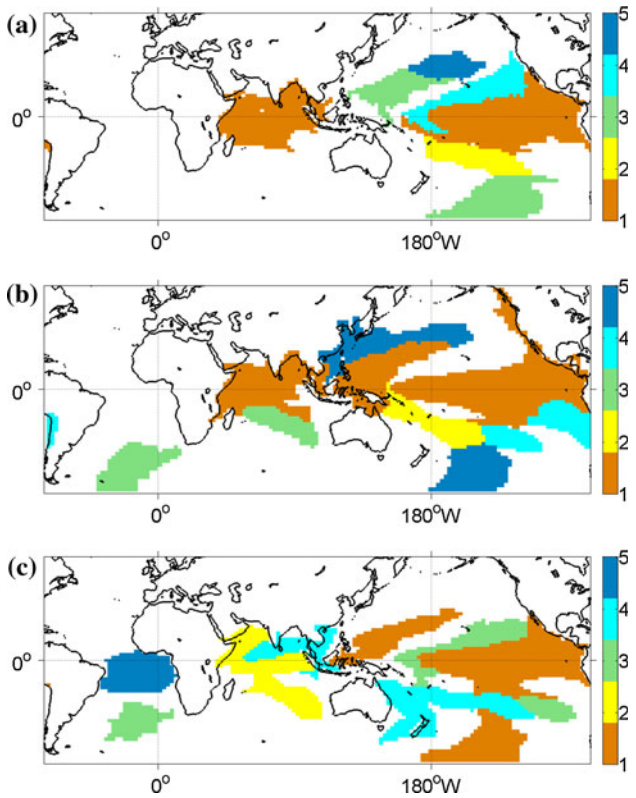


**Fig. 17** *Top five order cores* in **a** HadISST; **b** ERSST-V3; **c** NCEP. The period considered is 1979–2005 in all cases

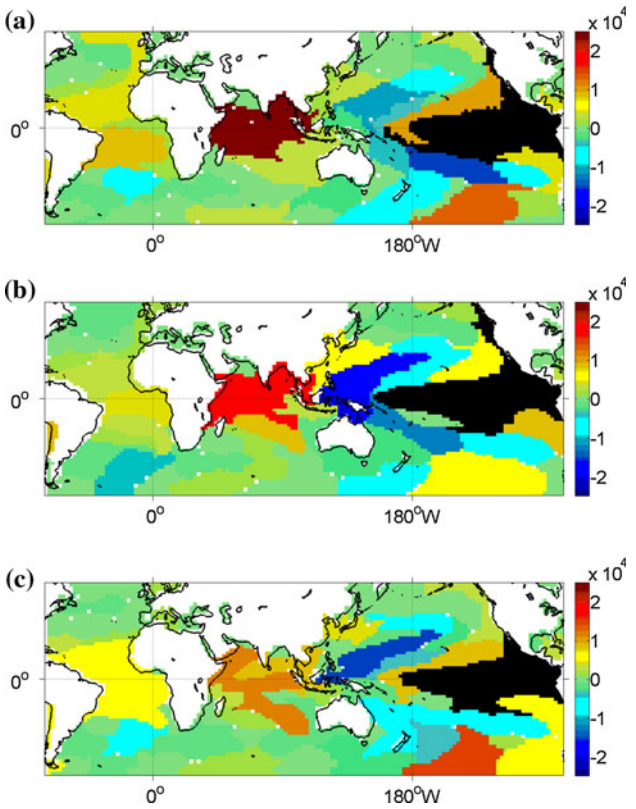


**Fig. 18** Links between the ENSO-like area shown in *black* and all other areas in the three reanalyses. **a** HadISST, **b** ERSST-V3 and **c** NCEP networks

The ARI metric, on the other hand, ranks NCEP closer to HadISST than ERSST-V3 (ARI = 0.59 for NCEP and ARI = 0.54 for ERSST-V3, mapped to $\gamma = 0.35$ and 0.45, respectively). The shape of the ENSO-related area and of areas in the tropical Atlantic and south of 30° S are indeed in better agreement between HadISST and NCEP, despite having different strengths.

The previous discussion illustrates that $D$ and ARI should be considered jointly, as they provide complementary information about the similarity and differences between two networks.

### 6.2 Network changes over time

Network analysis can also be a powerful tool to detect and quantify climate shifts. The insights that network analysis can offer, compared to more traditional time series analysis methods, are related to the detection of changes in network metrics that are associated with specific climate modes of variability, regional or global. Topological changes may include addition or removal of areas, significant fluctuations in the weight of existing links (strengthening and weakening of teleconnections), or variations in the relative significance of different areas, quantified by the area

strength distribution. For instance, Tsonis and co-authors have built a network of four interacting nodes using the major climate indices, the North Atlantic Oscillation (NAO), ENSO, the North Pacific Oscillation (NPO) and the Pacific Decadal Oscillation (PDO), and suggested that those climate modes of variability tend to synchronize with a certain coupling strength (Tsonis et al. 2007). Climate shifts, including the one recorded in the north Pacific around 1977 (Miller et al. 1994), could result from changes in such coupling strength.

Here we compare the climate networks constructed on the HadISST data set over the periods 1950–1976 and 1979–2005 to illustrate that the proposed methodology may also provide insights into the detection of climate shifts. Instead of simply comparing different periods, it is possible to use a sliding window in the network inference process to detect significant changes or shifts without prior knowledge; we will explore this possibility in future work.

Strength maps for the two networks were shown in Fig. 7, while the *top five order cores* were shown in Fig. 8. The links from the ENSO-related area and from the equatorial Indian Ocean during the 1950–1976 period are presented in Fig. 19, and they can be compared with Fig. 6. When the 1979–2005 period is compared to the earlier period, we note a substantial strength decrease for the area covering the south tropical Atlantic and a significant weaker link between this area and ENSO. This suggests an alteration in the Pacific-Atlantic connection, which indeed has been recently pointed out by Rodriguez-Fonseca et al. (2009) and may be linked to the Atlantic warming (Kucharski et al. 2011). Additionally, there is a change in the sign of the link weight between the ENSO area and the area off the coast of Alaska in the north Pacific, which is related to the change in sign of the PDO in 1976–1977 (Miller et al. 1994; Graham 1994).

Despite those differences, the distance from the 1979–2005 HadISST network to the 1950–1976 network is less than the distance from the former to any of the other reanalyses investigated earlier: $D(N, N') = 0.13$ with noise $\gamma = 0.10$. The ARI, on the other hand, is 0.55 ($\gamma = 0.40$). The ARI value reflects, predominantly, the changes in shape and size of the ENSO-related areas and of the areas over the North Atlantic and North Pacific.

### 6.3 Comparison of precipitation networks

One of the advantages of the proposed methodology is its applicability, without modifications, to any climate variable. As an example, in the following we focus on precipitation, chosen for having statistical characteristics very different from SST due to its intermittency. We investigate the network structure of the CPC Merged Analysis of Precipitation (CMAP) (Xie and Arkin 1997) and ERA-Interim reanalysis (Dee et al. 2011). Both data sets are available from 1979 onward. CMAP provides gridded, monthly averaged precipitation rates obtained from satellite estimates. ERA-Interim is the outcome of a state-of-the-art data assimilative model that assimilates a broad set of observations, including satellite data, every 12 hours. As in the case of SSTs, we present the precipitation networks focusing on boreal winter (December to January) based on detrended anomalies from 1979 to 2005. Figure 20 shows the map of area strengths for both data sets, Fig. 21 presents the *top five order cores*, while Fig. 22 depicts links from the strongest area in the two networks.

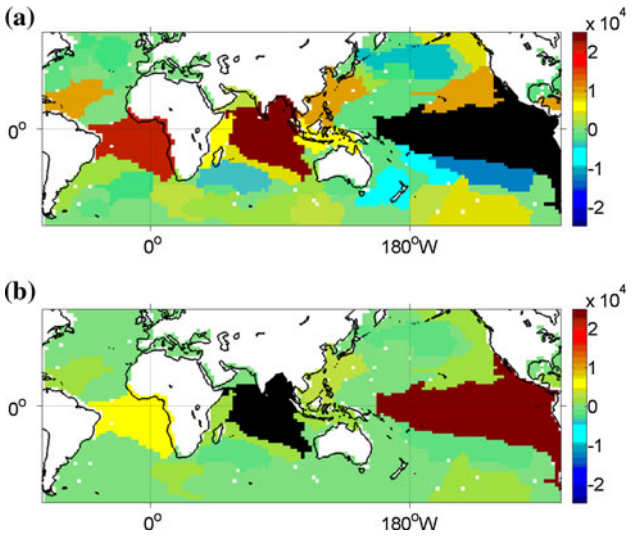


**Fig. 19** Links for the HadISST network over 1950–1976 from the (**a**) ENSO-related area, and **b** the equatorial Indian Ocean area (*in black* in the two panels)

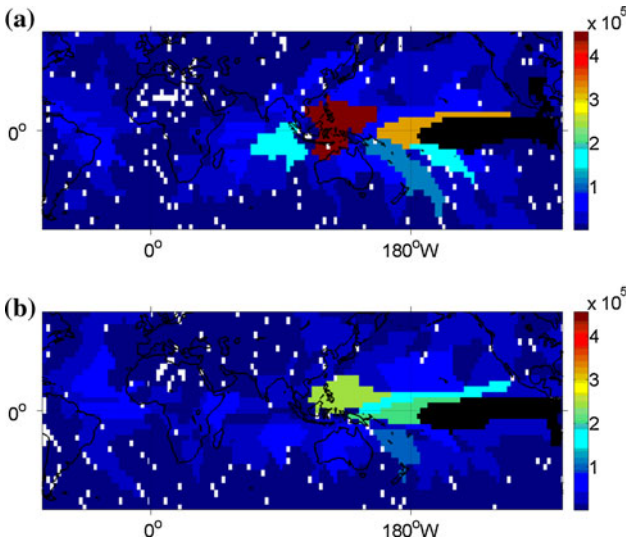


**Fig. 20** Precipitation networks. Area strength map in **a** CMAP (equatorial Pacific area strength $49.4 \times 10^4$), and **b** ERA-Interim (equatorial area strength $41.0 \times 10^4$)

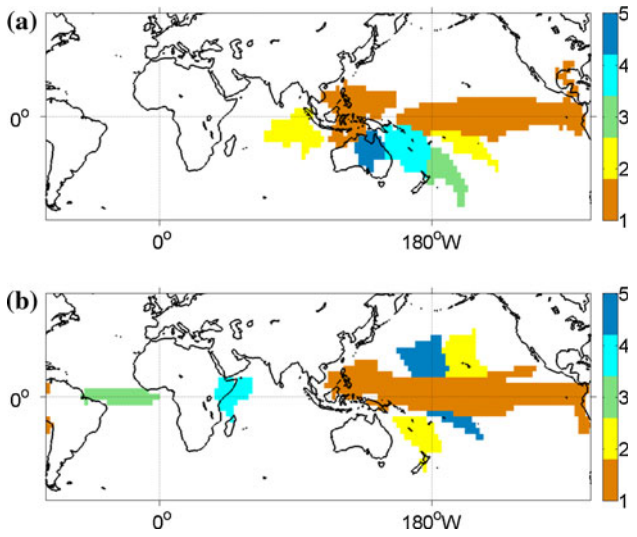


**Fig. 21** *Top five order cores* in **a** CMAP, and **b** ERA-Interim

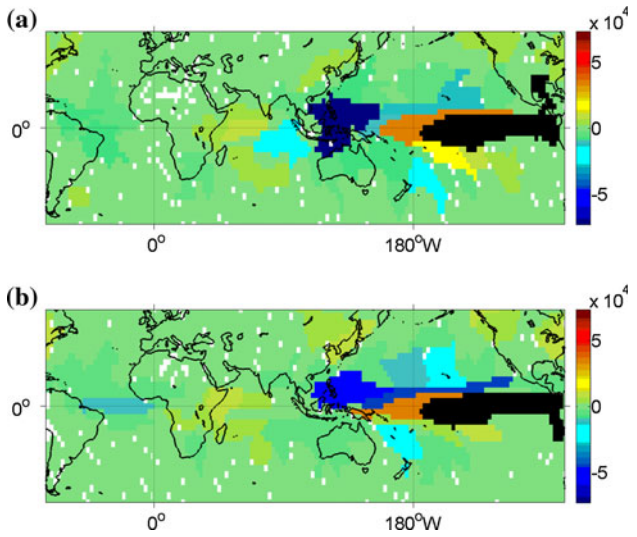


**Fig. 22** Link maps from the strongest area (*in black*) for the two precipitation reanalysis data sets. **a** CMAP; **b** ERA Interim

The precipitation network is, not surprisingly, characterized by smaller areas, compared to SSTs. Precipitation time series are indeed highly intermittent, resulting in weaker correlations between grid cells. The areas with the highest strength are concentrated in the tropics, where deep convection takes place. The strongest area is located in the equatorial Pacific in correspondence with the center of action of ENSO. In CMAP, this area is linked with strong negative correlation to the area covering the warm-pool region, and together they represent the *first order core* of this network. The *second order core* covers the eastern part of the Indian Ocean and eastern portion of the South Pacific Convergence Zone (SPCZ). Both those regions are strongly affected by the shift in convection associated with ENSO events. In the reanalysis, the warm-pool area extends predominantly into the northern hemisphere, and its strength and size, as well as the weight of its link with the ENSO-related area, are reduced. Additionally, the Indian Ocean is subdivided in small areas all of negligible strength, similarly to what seen for NCEP SSTs, indicating that the atmospheric teleconnection between ENSO and the eastern Indian Ocean that causes a shift in convective activity over the Indian basin (see e.g. Klein et al. (1999); Bracco et al. (2005) is not correctly captured by ERA-Interim. The *s-core* decomposition does not include in the *second order core* any area in the Indian Ocean, but is limited to two areas to the north and to the south of the ENSO-related one.

The distance from the CMAP network to the ERA-Interim network is $D(N, N') = 0.21$, with $\gamma = 0.25$, while the ARI value is 0.49, with $\gamma = 0.45$. These values reflect larger differences compared to the SST networks we presented earlier, but precipitation is known to be one of the most difficult fields to model, even when assimilating all available data, due to biases associated with the cloud formation and convective parameterization schemes (Ahlgrimm and Forbes 2012). In particular $D$ is affected by the significant difference in the strength and size of the area over the warm-pool, and of the one between the ENSO-related area and the warm-pool, while the ARI is affected by the difference in the partitions over the warm-pool and most of the Atlantic and Indian basins.

### 6.4 Regression between networks

So far we have shown applications of network analysis considering one climate variable at a time. In climate science it is often useful to visualize the relations between two or more variables to understand, for example, how changes in sea surface temperatures may impact rainfall. A simple statistical tool that highlights such relations is provided by regression analysis. Here we apply a similar approach using climate networks.

Consider two climate networks $N_x$ and $N_y$, constructed using variables $\mathbf{x}(\mathbf{t})$ and $\mathbf{y}(\mathbf{t})$, respectively. The relation between an area of $N_x$ and the areas of $N_y$ can be quantified based on the cumulative anomaly of each area, using the earlier link weight definition (see Eq. 3). Similarly, a link map for an area $A_i \in V_x$ can be constructed based on the link weights between the area $A_i$ and all areas $A_j \in V_y$.

For instance, we construct a network linking the area that corresponds to ENSO in the HadISST reanalysis to the areas of the CMAP precipitation network for the period 1979–2005 in boreal winter. Both networks are dominated by the ENSO area and it is expected that this exercise will portrait the ENSO teleconnection patterns. Results are shown in Fig. 23. The *regression* of the rainfall network

onto the ENSO-related area in the SST reanalysis visualizes the well known shift of convective activity from the warm-pool into the central and eastern equatorial Pacific during El Niño. For positive ENSO episodes, negative precipitation anomalies concentrate in the warm-pool and extend to the SPCZ and the eastern Indian Ocean. Weak, positive correlations between SST anomalies in the equatorial Pacific and precipitation are seen over the western Indian Ocean and east Africa, part of China, the Gulf of Alaska and the north-east USA. This approach is only moderately useful on reanalysis or observational data, where known indices can be used to perform regressions without the need of constructing a network. Its extension to model outputs, however, is advantageous compared to traditional methods, because it does not require any ad-hoc index definition, but relays on areas objectively identified by the proposed network algorithm.

### 6.5 CMIP5 SST networks

We now compare the HadISST network with networks constructed using SST anomalies from two coupled models participating in CMIP5. Our goal is to exemplify the information that our methodology can provide when applied to model outputs. We do not aim at providing an exhaustive evaluation of the model performances, which would be beyond the scope of this paper. We analyze the SST fields of two members of the CMIP5 historical ensemble from the GISS-E2H and HadCM3 models over the period 1979–2005. Historical runs aim at reproducing the observed climate from 1850 to 2005 including all forcings. We show strength maps (Fig. 24), *top five order cores* (Fig. 25), and link maps for the area that is related to ENSO (Fig. 26).

In all model integrations the ENSO-like area extends too far west into the warm-pool region, and is too narrow in the simulated width, in agreement with the recent analysis by Zhang and Jin (2012). The warm-pool is therefore not represented as an independent area anticorrelated to the ENSO-like one. In the GISS-E2H model the strength of the ENSO area is underestimated compared to the HadISST (see Fig. 16a), but the overall size of the area is larger than observed. Both the extent and strength of the Indian Ocean area around the equator and of the areas forming the horse-shoe pattern are reduced with respect to HadISST. Links in GISS-E2H are overall weaker than in the reanalysis (see Fig. 18a), the role of the Atlantic is slightly overestimated, and the high negative correlations between the ENSO region and the areas forming the horse-shoe patterns are not captured. In HadCM3, on the other hand, the strength of the ENSO area is comparable or greater than in the observations. In this model, areas are more numerous and fragmented than in the reanalysis, and in several cases

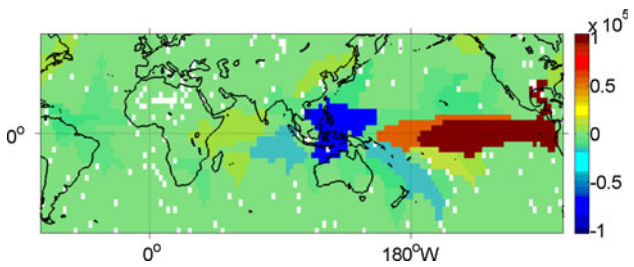


**Fig. 23** Link maps from the ENSO-like area in HadISST data set to all areas in the CMAP data set, considering the 1979–2005 period. Values greater than $|1 \times 10^5|$ are saturated

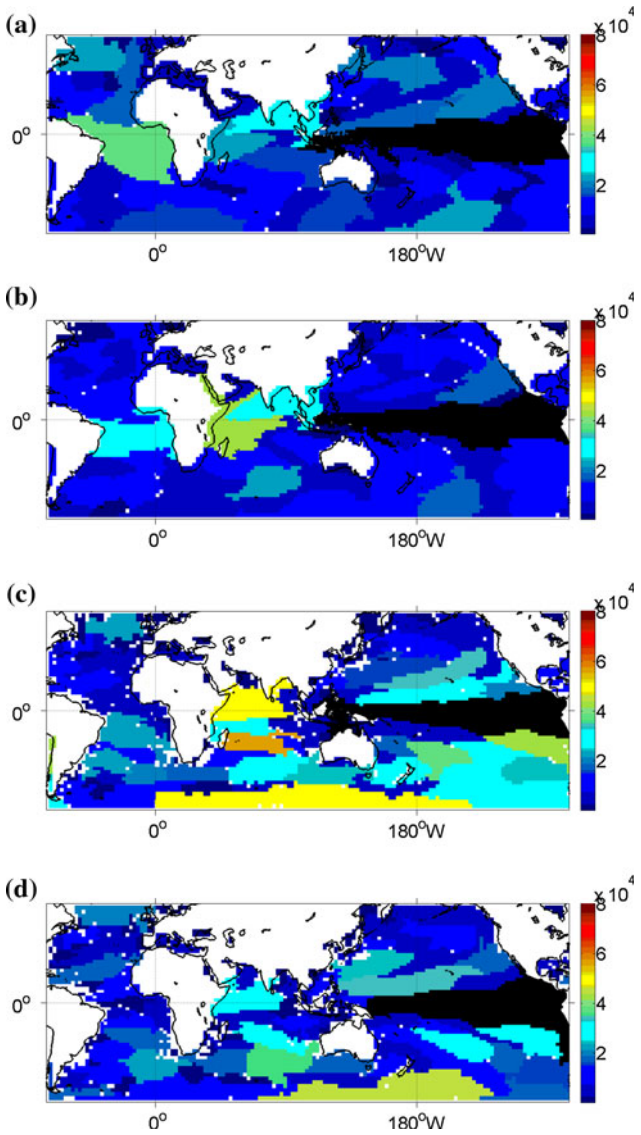


**Fig. 24** Strength maps for two members of the GISS-E2H and HadCM3 "historical" ensemble. **a** GISS-E2H run 1 (ENSO area strength $9.8 \times 10^4$); **b** GISS-E2H run 2 (ENSO area strength $10.0 \times 10^4$); **c** HadCM3 run 1 (ENSO area strength $23.3 \times 10^4$) and **d** HadCM3 run 2 (ENSO area strength $16.9 \times 10^4$)

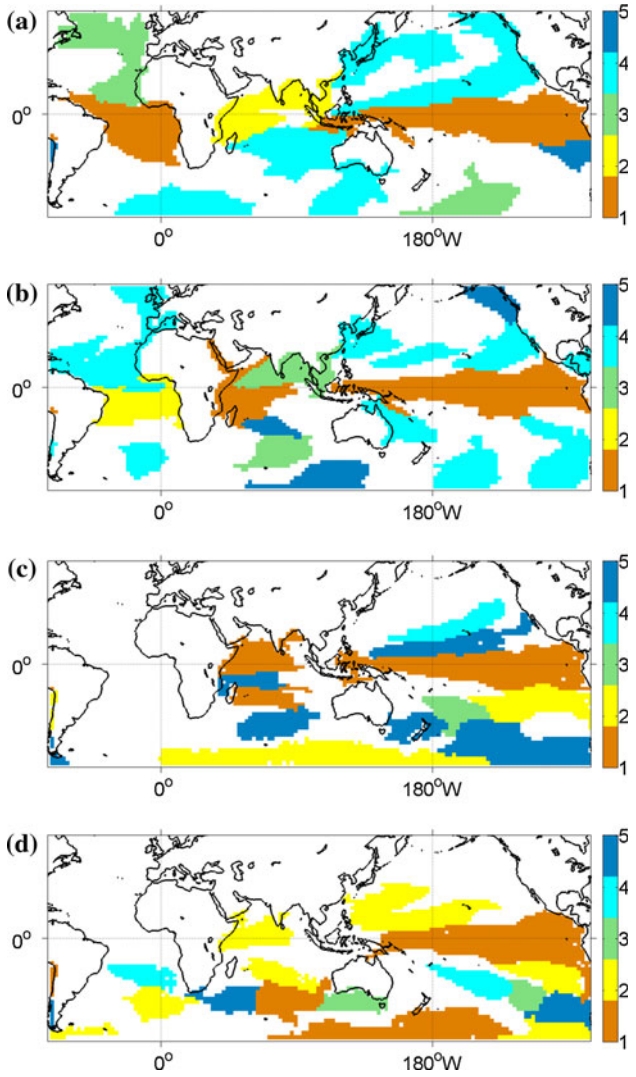


**Fig. 25** *Top five order cores* identified in the SST anomaly networks for **a**, **b** two GISS-E2H ensemble members and **c**, **d** two HadCM3 integrations

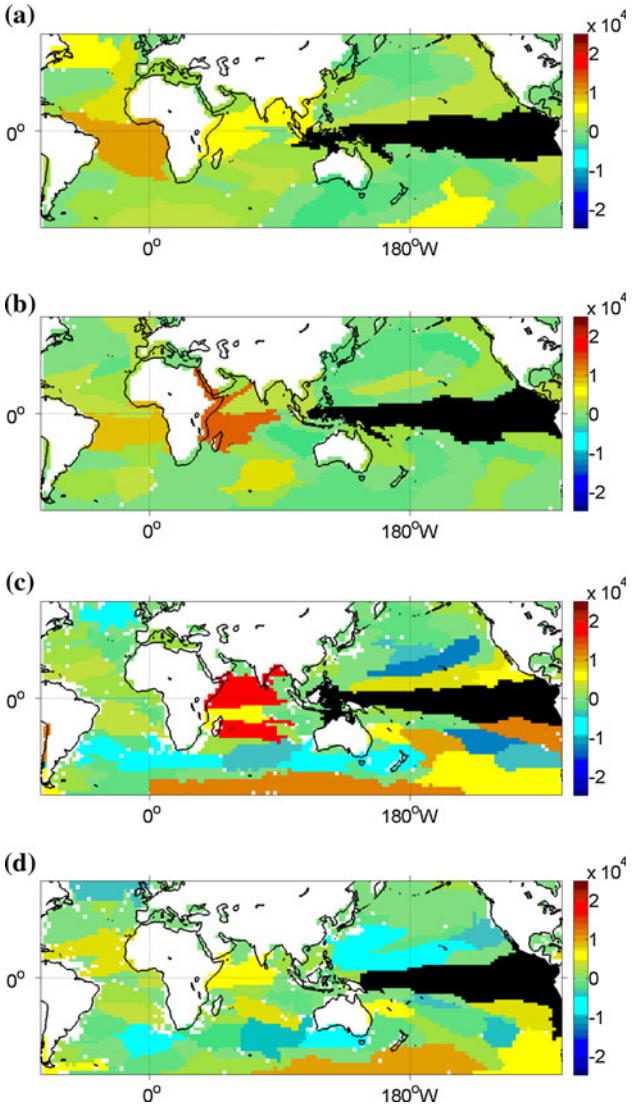


**Fig. 26** Link maps from the ENSO-like area in the **a**, **b** GISS-E2H and **c**, **d** HadCM3 models

confined within narrow latitudinal bands. This bias may result from too weak meridional currents and/or weak trade wind across all latitudes, as suggested by Zhang et al. (2012). HadCM3 shows also erroneously strong links between the modeled ENSO area and the Southern Ocean, particularly in the Pacific and Indian sectors, as evident in the *s-core* decomposition and link maps. The link strengths in HadCM3 are closer to the observed, but some areas in the southern hemisphere play a key role, unrealistically.

To conclude this comparison we present the distance from the HadISST reanalysis to those two models, and the corresponding ARI values. Table 1 summarizes this comparison. $D(N, N')$ from HadISST to the two GISS-E2H integrations is 0.29 and 0.37, with $\gamma = 0.35$ and $\gamma = 0.45$, respectively. $D(N, N')$ from HadISST to the two HadCM3 runs is 0.56 and 0.35, with $\gamma = 0.70$ and $\gamma = 0.40$. One of the GISS member networks displays a significantly smaller distance from HadISST than both networks build on the HadCM3 runs. This is due to the fact that in all networks considered the ENSO-like area overpowers all others in terms of strength and, furthermore, there exist a few other strong areas (weaker than the ENSO-related area by less than one order of magnitude). Focusing on the extent of the areas in the GISS member with smaller $D$ we observe striking differences relative to the base HadISST network: the GISS model is unable to reproduce the horse-shoe pattern, and it splits the tropical Indian Ocean in two areas. However, it reproduces quite well the overall size of most areas, and the strength of the largest two in the tropics, despite inverting the relative strengths of the Indian Ocean and of the south tropical Atlantic. The latter in GISS and the former in HadISST have comparable size and strength,

and $D$ cannot account for their different location. The HadCM3 networks, on the other hand, are too fragmented and characterized by unrealistically strong areas in the Southern Ocean, and they are penalized by $D$ for not capturing properly the size of the strongest areas. The ARI values are 0.46 and 0.48 for the two GISS members, and 0.43 and 0.45 for the two HadCM3 integrations. GISS again outperforms HadCM3 due to better representation of the shape of most areas.

As already mentioned, the relative distance and ARI metrics, while individually unable to quantify the differences and similarities between networks, can be jointly used successfully to rank several networks with respect to a common reference. Two networks are similar if both ARI is large and $D$ is small. If any of these two conditions is not met, a further analysis of the other metrics we have introduced can provide useful information on the differences between the data sets under consideration.

## 7 Discussion and conclusions

We developed a novel method to analyze climate variables using complex network analysis. The nodes of the network, or areas, are formed by clusters of grid cells that are highly homogeneous to the underlying climate variable. These areas can often be mapped into well known patterns of climate variability.

The network inference algorithm relies on a single parameter $\tau$ that determines the degree of homogeneity between cells in an area. The requirement of only one parameter, combined with the fact that no link pruning in the underlying cell-level network is imposed, adds robustness to the area-level network's structure and makes the comparison of different networks more reliable.

The constructed climate networks are complete weighted graphs. In effect, our network framework allows for investigating and visualizing the relative strength of node interactions, which can be associated with teleconnection patterns. The inferred networks are robust under random perturbations when adding noise to the anomaly time series of the climate variable under investigation, to small changes in the selection of $\tau$, to the choice of the correlation metric used in the inference algorithm, and to the spatial resolution of the input field.

In this paper we constructed networks for a suite of SST and precipitation data sets, and we analyzed them with a set of weighted metrics such as link maps, area strength and *s-core* decomposition. Link maps enable us to visualize all statistical relationships between areas, while strength maps highlight the relative importance of those relationships, identifying major climate patterns. The *s-core* decomposition, on the other hand, identifies the backbone structure of a network, clustering areas into layers of increasing significance. Finally, we quantified the degree of similarity between different networks using the the adjusted Rand index metric and a newly introduced "distance metric", based on the area strength distribution.

After analyzing three SST reanalyses and two precipitation data sets, we investigated the network structure of the SST fields generated by two CMIP5 models, GISS-E2H and HadCM3, focusing on SST anomalies. We visualized model biases in the underlying network topology and in the spatial expression of patterns, and we quantified the distance between model outputs and reanalyses. We found significant differences between model and observational data sets in the shape and relative strength of areas. The most striking biases common to both models are the excessive longitudinal extension of the area corresponding to ENSO, and the inability to represent the horse-shoe pattern in the western tropical Pacific. Links are generally weaker than observed in GISS-E2H, but the relative strength, shape and size of the main areas are in reasonable agreement with the reanalyses. The HadCM3 network, on the other hand, is closer to observations in the absolute strength of its areas, but the areas are too numerous in the tropics and unrealistically strong nodes are found in the South Pacific. In the near future, we aim at providing a comprehensive comparison of CMIP5 outputs to the climate community by extending our analysis to a much larger number of models.

In this work we limited our analysis to linear and zero-lag correlations. The methodology presented, however, could be generalized to include the analysis of nonlinear phenomena and non-instantaneous links, by introducing nonlinear correlation metrics, such as mutual information or the maximal information coefficient (Reshef et al. 2011), and time-lags. Additionally, the set of metrics proposed can be enhanced to capture more complex relationships in the underlying network.

**Table 1** $D$ and ARI from HadISST (1979–2005) to reanalyses, GISS-E2H and HadCM3, and corresponding noise-to-signal ratios $\gamma$

| Dataset | $D$ | $\gamma$ ($D$) | ARI | $\gamma$ (ARI) |
|---|---|---|---|---|
| HadISST 1950-1976 | 0.13 | 0.10 | 0.55 | 0.40 |
| ERSST-V3 | 0.16 | 0.15 | 0.54 | 0.45 |
| NCEP | 0.29 | 0.35 | 0.59 | 0.35 |
| GISS run 1 | 0.29 | 0.35 | 0.46 | 0.60 |
| GISS run 2 | 0.37 | 0.45 | 0.48 | 0.55 |
| HadCM3 run 1 | 0.56 | 0.70 | 0.43 | 0.70 |
| HadCM3 run 2 | 0.35 | 0.40 | 0.45 | 0.60 |

**Acknowledgments** This work was made possible by a grant from the Department of Energy, Climate and Environmental Sciences Division, SciDAC: Earth System Model Development. We thank the anonymous reviewers for the insightful comments that helped improve the paper.

## Appendix 1: Selection of threshold $\tau$

The threshold $\tau$ is the only parameter of the proposed network construction method. It represents the *minimum average pair-wise correlation between cells of the same area*, as shown in Eq. 1. Intuitively, $\tau$ controls the minimum degree of homogeneity that the climate field should have within each area. The higher the threshold, the higher the required homogeneity, and therefore the smaller the identified areas.

Throughout this paper, we select $\tau$ based on the following heuristic. First, we apply the one-sided $t$ test for Pearson correlations at level $\alpha$ and with $-2$ degrees of freedom (recall that $T$ is the length of the anomaly time series) to calculate the minimum correlation value $r_\alpha$ that is significant at that level (Rogers 1969). For example, with $\alpha = 1\ \%$ and $T = 81$ (corresponding to 27 years of SST montly DJF averages), we get $r_\alpha = 0.34$.

Instead of prunning any correlations $r(x_i, x_j)$ that are below $r_\alpha$, we estimate the expected value of only those correlations that are larger than $r_\alpha$,

$$\bar{r}_\alpha \triangleq E[r(x_i, x_j), r(x_i, x_j) > r_\alpha] \tag{7}$$

For a set of $k$ randomly chosen cells that have statistical significant correlations (at level $\alpha$) between them, $\bar{r}_\alpha$ is approximately equal, for large $k$, to their average pair-wise correlation. A climate area, however, is not a set of randomly chosen cells, but a geographically connected region. So, we require that the average pair-wise correlation of cells that belong to the same area should be higher than $\bar{r}_\alpha$, i.e.,

$$\tau = \bar{r}_\alpha \tag{8}$$

Note that $\tau$ is independent of the size of an area, but it depends on both $\alpha$ and on the distribution of pair-wise correlations $r(x_i, x_j)$.

## Appendix 2: Pseudocode of area identification algorithm

Below we present the pseudocode for the area identification algorithm used in this paper.

```
function PART-1
    Mark all cells as available
    k ← 0
    V ← ∅
    while true do
        Identify the two available and neighboring cells (i, j) with the maximum correlation
        if r(x_i, x_j) < τ then
            exit                                   ▷ No additional areas can be identified
        else
            Area A_k ← i, j
            i, j ← unavailable
            V ← EXPAND(A_k)
            k = k + 1
        end if
    end while
end function

function EXPAND(Area A_k)
    Construct set Nei(A_k): all available neighboring cells to area A_k
    while true do
        if Nei(A_k) = ∅ then
            return A_k
        else
            m = arg max_{m∈Nei(A_k)} r̂(x_m, A_k),   with r̂(x_m, A_k) = Σ_{i∈A_k} r(x_m, x_i)/|A_k|   ▷
Identify the cell m in Nei(A_k) that has maximum average correlation with existing cells in
A_k.
            if r̂(x_m, A_k) > τ then
                A_k ← m
                m ← unavailable
                Include available neighbors of m in Nei(A_k)
            else
                return A_k
            end if
        end if
    end while
end function
```

```
function PART-2(Areas V = {A_1, ..., A_|V|})
    Mark all areas A_i ∈ V as available
    while true do
        A_k = arg max_{A_i ∈ V} |A_i|        ▷ Identify the largest available area A_k ∈ V in terms of number of cells.
        if A_k = ∅ then
            exit                                                    ▷ No additional available areas.
        else
            Construct set Nei(A_k): all geographically adjacent areas to A_k
            if Nei(A_k) = ∅ then
                A_k ← unavailable
            else
                Identify area A_j ∈ Nei(A_k) such that average correlation of all cells in A_k ∪ A_j is maximum
                if r̂(A_j, A_k) > τ then
                    Remove A_j from V
                    A_k = A_k ∪ A_j
                else
                    Mark A_k as unavailable
                end if
            end if
        end if
    end while
end function
```

## References


Abramov R, Majda A (2009) A new algorithm for low-frequency climate response. J Atmos Sci 66(2):286–309

Ahlgrimm M, Forbes R (2012) The impact of low clouds on surface shortwave radiation in the ecmwf model. Mon Weather Rev 140

Allen M, Smith L (1994) Investigating the origins and significance of low-frequency modes of climate variability. Geophys Res Lett 21(10):883–886

Andronova N, Schlesinger M (2001) Objective estimation of the probability density function for climate sensitivity. J Geophys Res 106(22):605–22

Bracco A, Kucharski F, Molteni F, Hazeleger W, Severijns C (2005) Internal and forced modes of variability in the indian ocean. Geophys Res Lett 32(12):L12707

Chambers D, Tapley B, Stewart R (1999) Anomalous warming in the indian ocean coincident with el niño. J Geophys Res 104(C2):3035–3047

Cormen T, Leiserson C, Rivest R, Stein C (2001) Introduction to algorithms. Section 24:588–592

Corti S, Giannini A, Tibaldi S, Molteni F (1997) Patterns of low-frequency variability in a three-level quasi-geostrophic model. Climate Dyn 13(12):883–904

Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Hlm EV, Isaksen L, Kllberg P, Khler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette JJ, Park BK, Peubey C, de Rosnay P, Tavolato C, Thpaut JN, Vitart F (2011) The era-interim reanalysis: configuration and performance of the data assimilation system. Q J R Meteorol Soc 137(656):553–597

Dijkstra H (2005) Nonlinear physical oceanography: a dynamical systems approach to the large scale ocean circulation and El Niño, vol 28. Springer, Berlin

Donges JF, Zou Y, Marwan N, Kurths J (2009a) The backbone of the climate network. EPL (Europhys Lett) 87(4):48007

Donges JF, Zou Y, Marwan N, Kurths J (2009b) Complex networks in climate dynamics. Eur Phys J Spec Top 174(1):157–179

Donges JF, Schultz H, Marwan N, Zou Y, Kurths J (2011) Investigating the topology of interacting networks. Eur Phys J B Condens Matter 84(4):635

Forest C, Stone P, Sokolov A, Allen M, Webster M (2002) Quantifying uncertainties in climate system properties with the use of recent climate observations. Science 295(5552):113–117

Fountalis I, Dovrolis C, Bracco A (2013) Efficient algorithms for the detection of homogeneous areas in spatial and weighted networks. Tech Rep Coll Comput Georgia Tech

Ghil M, Vautard R (1991) Interdecadal oscillations and the warming trend in global temperature time series. Nature 350(6316): 324–327

Ghil M, Allen M, Dettinger M, Ide K, Kondrashov D, Mann M, Robertson A, Saunders A, Tian Y, Varadi F et al (2002) Advanced spectral methods for climatic time series. Rev Geophys 40(1):1003

Gordon C, Cooper C, Senior C, Banks H, Gregory J, Johns T, Mitchell J, Wood R (2000) The simulation of sst, sea ice extents and ocean heat transports in a version of the hadley centre coupled model without flux adjustments. Clim Dyn 16(2):147–168

Graham N (1994) Decadal-scale climate variability in the tropical and north pacific during the 1970s and 1980s: observations and model results. Clim Dyn 10(3):135–162

Hansen J, Sato M, Nazarenko L, Ruedy R, Lacis A, Koch D, Tegen I, Hall T, Shindell D, Santer B et al (2002) Climate forcings in goddard institute for space studies si2000 simulations. J Geophys Res 107(10.1029)

Van∼den Heuvel M, Sporns O (2011) Rich-club organization of the human connectome. J Neurosci 31(44):15775–15786

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218

Hurrell JW, Trenberth KE (1999) Global sea surface temperature analyses: multiple problems and their implications for climate analysis, modeling, and reanalysis. Bull Am Meteorol Soc 80(12):2661–2678

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J et al (1996) The ncep/ncar reanalysis 40-year project. Bull Am Meteorol Soc 77(3):437–471

Kawale J, Liess S, Kumar A, Steinbach M, Ganguly A, Samatova NF, Semazzi FHM, Snyder PK, Kumar V (2011) Data guided discovery of dynamic climate dipoles. In: CIDU. pp 30–44

Kawale J, Chatterjee S, Ormsby D, Steinhaeuser K, Liess S, Kumar V (2012) Testing the significance of spatio-temporal teleconnection patterns. In: ACM SIGKDD conference on knowledge discovery and data mining

Klein SA, Soden BJ, Lau NC (1999) Remote sea surface temperature variations during enso: evidence for a tropical atmospheric bridge. J Clim 12(4):917–932

Kucharski F, Kang IS, Farneti R, Feudale L (2011) Tropical pacific response to 20th century atlantic warming. Geophys Res Lett 38(3)

Miller A, Cayan D, Barnett TP, Graham NE, Oberhuber JM (1994) The 1976-1977 climate shift of the pacific ocean. Oceanography 7:21–26

Newman M (2010) Networks: an introduction. Oxford University Press, Inc, Oxford

Newman M, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026113

Newman M, Barabasi A, Watts D (2006) The structure and dynamics of networks. Princeton University Press, Princeton, NY

Pelan A, Steinhaeuser K, Chawla NV, de Alwis Pitts D, Ganguly A (2011) Empirical comparison of correlation measures and pruning levels in complex networks representing the global climate system. In: Computational intelligence and data mining (CIDM), 2011 IEEE symposium on, IEEE. pp 239–245

Rand WM (1971) Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 66(336):846–850

Rayner N, Parker D, Horton E, Folland C, Alexander L, Rowell D, Kent E, Kaplan A (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J Geophys Res 108(D14):4407

Reshef D, Reshef Y, Finucane H, Grossman S, McVean G, Turnbaugh P, Lander E, Mitzenmacher M, Sabeti P (2011) Detecting novel associations in large data sets. Science 334(6062):1518–1524

Reynolds RW, Smith TM (1994) Improved global sea surface temperature analyses using optimum interpolation. J Clim 7(6):929–948

Rodriguez-Fonseca B, Polo I, Garca-Serrano J, Losada T, Mohino E, Mechoso CR, Kucharski F (2009) Are atlantic nios enhancing pacific enso events in recent decades? Geophys Res Lett 36(20)

Rogers G (1969) A course in theoretical statistics. Technometrics 11(4):840–841

Smith T, Reynolds R, Peterson T, Lawrimore J (2008) Improvements to noaa's historical merged land-ocean surface temperature analysis (1880–2006). J Clim 21(10):2283–2296

Steinbach M, Tan PN (2003) Discovery of climate indices using clustering. In: Proceedings of the 9th ACM SIGKDD intel conference on knowledge discovery and data mining. pp 24–27

Steinhaeuser K, Chawla NV (2010) Identifying and evaluating community structure in complex networks. Pattern Recogn Lett 31(5):413–421

Steinhaeuser K, Chawla NV, Ganguly AR (2009) An exploration of climate data using complex networks. In: KDD workshop on knowledge discovery from sensor data. pp 23–31

Steinhaeuser K, Chawla NV, Ganguly AR (2010) Complex networks in climate science: progress, opportunities and challenges. In: CIDU. pp 16–26

Steinhaeuser K, Chawla NV, Ganguly AR (2011a) Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science. Stat Anal Data Min 4(5):497–511

Steinhaeuser K, Ganguly A, Chawla NV (2011b) Multivariate and multiscale dependence in the global climate system revealed through complex networks. Clim Dyn 1–7

Swanson K, Tsonis A et al. (2009) Has the climate recently shifted? Geophys Res Lett 36(6):L06711

Taylor K, Stouffer R, Meehl G (2012) An overview of cmip5 and the experiment design. Bull Am Meteorol Soc 93(4):485

Tsonis A, Roebber P (2004) The architecture of the climate network. Phys A Stat Mech Appl 333:497–504

Tsonis A, Swanson K (2008) Topology and predictability of el nino and la nina networks. Phys Rev Lett 100(22):228502

Tsonis A, Wang G, Swanson K, Rodrigues F, Costa L (2010) Community structure and dynamics in climate networks. Clim Dyn 1–8

Tsonis AA, Swanson KL, Roebber PJ (2006) What do networks have to do with climate? Bull Am Meteorol Soc 87(5):585–595

Tsonis AA, Swanson K, Kravtsov S (2007) A new dynamical mechanism for major climate shifts. Geophys Res Lett 34:L13705

Tsonis AA, Swanson KL, Wang G (2008) On the role of atmospheric teleconnections in climate. J Clim 21(12):2990–3001

Vidard A, Anderson D, Balmaseda M (2007) Impact of ocean observation systems on ocean analysis and seasonal forecasts. Mon Weather Rev 135(2):409–429

Wang G, Swanson K, Tsonis A et al. (2009) The pacemaker of major. Geophys Res Lett 36(7):L07708

Xie P, Arkin P (1997) Global precipitation: a 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. Bull Am Meteorol Soc 78(11):2539–2558

Yamasaki K, Gozolchiani A, Havlin S (2008) Climate networks around the globe are significantly effected by el niño. Arxiv Preprint arXiv:08041374

Zhang W, Jin F (2012) Improvements in the cmip5 simulations of enso-ssta meridional width. Geophys Res Lett 39(23)

Zhang W, Jin F, Zhao J, Li J (2012) On the bias in simulated enso ssta meridional widths of cmip3 models. J Clim 25