Complex Networks in the Climate System

Diplomarbeit in Physik
Angefertigt am
Potsdam Institut für Klimafolgenforschung
und am
Institut für Physik
der Universität Potsdam

Vorgelegt der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

von
Jonathan Friedemann Donges

Potsdam
März 2009

Erstgutachter: Prof. Dr. Dr. h.c. Jürgen Kurths
Zweitgutachter: Dr. Udo Schwarz

Published online at the
Institutional Repository of the University of Potsdam:
URL http://opus.kobv.de/ubp/volltexte/2011/4977/
URN urn:nbn:de:kobv:517-opus-49775
http://nbn-resolving.org/urn:nbn:de:kobv:517-opus-49775

I dedicate this work to my parents,
who inspired me to wonder.

Abstract

Complex network theory provides an elegant and powerful framework to statistically investigate the topology of local and long range dynamical interrelationships, *i.e.*, teleconnections, in the climate system. Employing a refined methodology relying on linear and nonlinear measures of time series analysis, the intricate correlation structure within a multivariate climatological data set is cast into network form. Within this graph theoretical framework, vertices are identified with grid points taken from the data set representing a region on the the Earth's surface, and edges correspond to strong statistical interrelationships between the dynamics on pairs of grid points. The resulting climate networks are neither perfectly regular nor completely random, but display the intriguing and nontrivial characteristics of complexity commonly found in real world networks such as the internet, citation and acquaintance networks, food webs and cortical networks in the mammalian brain. Among other interesting properties, climate networks exhibit the "small-world" effect and possess a broad degree distribution with dominating super-nodes as well as a pronounced community structure.

We have performed an extensive and detailed graph theoretical analysis of climate networks on the global topological scale focussing on the flow and centrality measure *betweenness* which is locally defined at each vertex, but includes global topological information by relying on the distribution of shortest paths between all pairs of vertices in the network. The betweenness centrality field reveals a rich internal structure in complex climate networks constructed from reanalysis and atmosphere-ocean coupled general circulation model (AOGCM) surface air temperature data. Our novel approach uncovers an elaborately woven meta-network of highly localized channels of strong dynamical information flow, that we relate to global surface ocean currents and dub the *backbone of the climate network* in analogy to the homonymous data highways of the internet. This finding points to a major role of the oceanic surface circulation in coupling and stabilizing the global temperature field in the long term mean (140 years for the model run and 60 years for reanalysis data). Carefully comparing the backbone structures detected in climate networks constructed using linear Pearson correlation and nonlinear mutual information, we argue that the high sensitivity of betweenness with respect to small changes in network structure may allow to detect the footprints of strongly nonlinear physical interactions in the climate system.

The results presented in this thesis are thoroughly founded and substantiated using a

hierarchy of statistical significance tests on the level of time series and networks, i.e., by tests based on time series surrogates as well as network surrogates. This is particularly relevant when working with real world data. Specifically, we developed new types of network surrogates to include the additional constraints imposed by the spatial embedding of vertices in a climate network.

Our methodology is of potential interest for a broad audience within the physics community and various applied fields, because it is *universal* in the sense of being valid for any spatially extended dynamical system. It can help to understand the localized flow of dynamical information in any such system by combining multivariate time series analysis, a complex network approach and the information flow measure betweenness centrality. Possible fields of application include fluid dynamics (turbulence), plasma physics and biological physics (population models, neural networks, cell models). Furthermore, the climate network approach is equally relevant for experimental data as well as model simulations and hence introduces a novel perspective on model evaluation and data driven model building. Our work is timely in the context of the current debate on climate change within the scientific community, since it allows to assess from a new perspective the regional vulnerability and stability of the climate system while relying on global and not only on regional knowledge. The methodology developed in this thesis hence has the potential to substantially contribute to the understanding of the local effect of extreme events and tipping points in the earth system within a holistic global framework.

Zusammenfassung

Die Theorie komplexer Netzwerke bietet einen eleganten Rahmen zur statistischen Untersuchung der Topologie lokaler und langreichweitiger dynamischer Zusammenhänge (Telekonnektionen) im Klimasystem. Unter Verwendung einer verfeinerten, auf linearen und nichtlinearen Korrelationsmaßen der Zeitreihenanalyse beruhenden Netzwerkkonstruktionsmethode, bilden wir die komplexe Korrelationsstruktur eines multivariaten klimatologischen Datensatzes auf ein Netzwerk ab. Dabei identifizieren wir die Knoten des Netzwerkes mit den Gitterpunkten des zugrundeliegenden Datensatzes, während wir Paare von besonders stark korrelierten Knoten als Kanten auffassen. Die resultierenden Klimanetzwerke zeigen weder die perfekte Regularität eines Kristallgitters, noch eine vollkommen zufällige Topologie. Vielmehr weisen sie faszinierende und nichttriviale Eigenschaften auf, die charakteristisch für natürlich gewachsene Netzwerke wie z.B. das Internet, Zitations- und Bekanntschaftsnetzwerke, Nahrungsnetze und kortikale Netzwerke im Säugetiergehirn sind. Besonders erwähnenswert ist, dass in Klimanetzwerken das Kleine-Welt-Phänomen auftritt. Desweiteren besitzen sie eine breite Gradverteilung, werden von Superknoten mit sehr vielen Nachbarn dominiert, und bilden schließlich regional wohldefinierte Untergruppen von intern dicht vernetzten Knoten aus.

Im Rahmen dieser Arbeit wurde eine detaillierte, graphentheoretische Analyse von Klimanetzwerken auf der globalen topologischen Skala durchgeführt, wobei wir uns auf das Netzwerkfluss- und Zentralitätsmaß *Betweenness* konzentrierten. Betweenness ist zwar lokal an jedem Knoten definiert, enthält aber trotzdem Informationen über die globale Netzwerktopologie. Dies beruht darauf, dass die Verteilung kürzester Pfade zwischen allen möglichen Paaren von Knoten in die Berechnung des Maßes eingeht. Das Betweennessfeld zeigt reichhaltige und zuvor verborgene Strukturen in aus Reanalyse- und Modelldaten der erdoberflächennahen Lufttemperatur gewonnenen Klimanetzen. Das durch unseren neuartigen Ansatz enthüllte Metanetzwerk, bestehend aus hochlokalisierten Kanälen stark gebündelten Informationsflusses, bringen wir mit der Oberflächenzirkulation des Weltozeans in Verbindung. In Analogie mit den gleichnamigen Datenautobahnen des Internets nennen wir dieses Metanetzwerk den *Backbone* des Klimanetzwerks. Unsere Ergebnisse deuten insgesamt darauf hin, dass Meeresoberflächenströmungen einen wichtigen Beitrag zur Kopplung und Stabilisierung des globalen Oberflächenlufttemperaturfeldes leisten. Wir zeigen weiterhin, dass die hohe Sensitivität des Betweennessmaßes hinsichtlich kleiner Änderungen der Netzwerktopologie die

Detektion stark nichtlinearer physikalischer Wechselwirkungen im Klimasystem ermöglichen könnte.

Die in dieser Arbeit vorgestellten Ergebnisse wurden mithilfe statistischer Signifikanztests auf der Zeitreihen- und Netzwerkebene gründlich auf ihre Robustheit geprüft. In Anbetracht fehlerbehafteter Daten und komplexer statistischer Zusammenhänge zwischen verschiedenen Netzwerkmaßen ist diese Vorgehensweise besonders wichtig. Weiterhin ist die Entwicklung neuer, allgemein anwendbarer Surrogate für räumlich eingebettete Netzwerke hervorzuheben, die die Berücksichtigung spezieller Klimanetzwerkeigenschaften wie z.B. der Wahrscheinlichkeitsverteilung der Kantenlängen erlauben.

Unsere Methode ist *universell*, weil sie zum Verständnis des lokalisierten Informationsflusses in allen räumlich ausgedehnten, dynamischen Systemen beitragen kann. Deshalb ist sie innerhalb der Physik und anderer angewandter Wissenschaften von potentiell breitem Interesse. Mögliche Anwendungen könnten sich z.B. in der Fluiddynamik (Turbulenz), der Plasmaphysik und der Biophysik (Populationsmodelle, neuronale Netzwerke und Zellmodelle) finden. Darüber hinaus ist der Netzwerkansatz für experimentelle Daten sowie Modellsimulationen gültig, und eröffnet folglich neue Perspektiven für Modellevaluation und datengetriebene Modellierung. Im Rahmen der aktuellen Klimawandeldebatte stellen Klimanetzwerke einen neuartigen Satz von Analysemethoden zur Verfügung, der die Evaluation der lokalen Vulnerabilität und Stabilität des Klimasystems unter Berücksichtigung globaler Randbedingungen ermöglicht. Die in dieser Arbeit entwickelten und untersuchten Methoden könnten folglich in der Zukunft, innerhalb eines holistisch-globalen Ansatzes, zum Verständnis der lokalen Auswirkungen von Extremereignissen und Kipppunkten im Erdsystem beitragen.

List of Publications

This thesis is partially based on the following publications:

I J.F. Donges, Y. Zou, N. Marwan and J. Kurths, *The backbone of the climate network*. Submitted to Physical Review Letters (December 2008).

II J.F. Donges, Y. Zou, N. Marwan and J. Kurths, *Complex networks in climate dynamics. Comparing linear and nonlinear network construction methods*. Submitted to European Physics Journal Special Topics (January 2009). Accepted (March 2009).

Acknowledgments

I thank Prof. Jürgen Kurths for supporting me in manifold ways and allowing my ideas to wander freely. I am indebted to all friends at the Nonlinear Dynamics Group at the University of Potsdam and those at the Potsdam Institute for Climate Impact Research (PIK) for freely giving their time, advice and insight. Most particularly, I value the contributions of Yong Zou and Norbert Marwan. Thank you both for reiterating ideas and manuscripts. Many times. I specifically thank André Bergner for sharing his knowledge on time series analysis, Gorka Zamora-López for so well communicating his sharp-witted and well structured understanding of complex network theory, network null-models and significance tests. I also thank Prof. Albert Dìaz-Guilera for the long discussions on random walks on networks and the transitivity problem, Michael Sexton for being the first to explicitly point out the similarity of betweenness backbone structures and ocean currents, and Anders Levermann and Gunnar Schmidt for stimulating discussions and helpful comments. I express my gratitude to Reik Donner and two anonymous referees of paper II for their versatile fruitful suggestions. I thank Udo Schwarz for his very helpful suggestions on time series analysis and significance tests as well as for proof reading this thesis. I furthermore thank Robert Flassig, Norbert Marwan, Torsten Albrecht, Nishant Malik and Hanna Schultz for their roles in correcting parts of the manuscript. I acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multi-model data set. Support of this data set is provided by the Office of Science, U.S. Department of Energy. I also acknowledge the use of NCEP Reanalysis Derived data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at http://www.cdc.noaa.gov/. I am particularly grateful for the time and efforts invested by the authors of the various OpenSource software packages which I essentially relied on during the work on this thesis (Appx. E). Deep appreciation goes to my parents, my grand parents, the German National Academic Foundation (Studienstiftung des deutschen Volkes) and the Fulbright Association for essential life support and intellectual impulses during my studies in Bonn, Potsdam and San Diego. I particularly value the contribution of the German and US American tax payers. "Das kanonische Ensemble", I am grateful for the good times. Thank you all. Thank you, darling, for being there.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Alles ist in Wechselwirkung.

Alexander von Humboldt, "Kosmos" (1845)

During the last decade, the development and application of complex network theory generated a wealth of novel insights into the nature of complex systems in various areas of science, *e.g.*, the internet and world wide web in computer science, food webs, gene expression and neural networks in biology and citation networks in social science (Watts and Strogatz (1998), Newman (2003), Albert and Barabási (2002)). The intricate interplay between the structure and dynamics of real world networks has received considerable attention (Boccaletti et al. (2006)). Particularly, synchronization arising by the transfer of dynamical information in complex network topologies has been studied intensively (Arenas et al. (2008)). The application of complex network theory to climate science is a very young field, where only few studies have been reported recently (Tsonis and Roebber (2004), Tsonis et al. (2006), Tsonis et al. (2008b), Tsonis and Swanson (2008), Yamasaki et al. (2008), Gozolchiani et al. (2008), Donner et al. (2008), Donges et al. (2008), Donges et al. (2009)). The vertices of a climate network are identified with the spatial grid points of an underlying global climate data set. Edges are added between pairs of vertices depending on the degree of statistical interdependence between the corresponding pairs of anomaly time series taken from the climate data set.

When studying networks in the climate system, one has to assume that its dynamics can be approximated reasonably well by a grid of low dimensional nonlinear dynamical systems interacting only with their spatial neighbors according to the locality principle of classical physics. Note that this assumption is made implicitly, when the fundamental partial differential equations of fluid mechanics are discretized and integrated in large scale climate simulations by the coupled atmosphere-ocean general circulation models (AOGCMs) used in weather forecasting and climate science. Due to the continuity of the underlying physical fields, such as temperature or pressure, neighboring grid points are dynamically correlated; these trivial local correlations usually decay quickly within a typical length scale. Additionally, richly structured long range correlations appear, that were named

teleconnections by the climatological community and have been studied extensively since the end of the 19th century.

The climate network approach enables novel insights into the topology and dynamics of the climate system over many spatial scales ranging from local properties as the number of first neighbors of a vertex v to global network measures, such as the clustering coefficient or the average path length. The local degree centrality and related measures have been used to identify super-nodes (regions of high degree centrality) and to associate them with teleconnection patterns in the atmosphere, most notably the North Atlantic Oscillation (NAO) (Tsonis and Roebber (2004), Tsonis et al. (2006), Tsonis et al. (2008b)). On the global scale, climate networks were found to possess "small-world" properties due to long range connections (edges linking geographically very distant vertices), that stabilize the climate system and enhance the information transfer within it (Tsonis and Roebber (2004), Tsonis et al. (2006), Tsonis et al. (2008b)). We stress, that the transfer of information in any complex physical system, *e.g.*, the climate system studied here, will be carried by a flow of matter and energy. By studying the prevalence of long range connections in El Niño and La Niña climate networks (Tsonis and Swanson (2008)) and the time dependence of the number of stable edges (Yamasaki et al. (2008), Gozolchiani et al. (2008)), it has been shown very recently, that the El Niño-Southern Oscillation (ENSO) has a strong impact on the stability of the climate system.

This thesis is organized as follows: First we introduce the necessary *elements of complex network theory* (Chap. 2) and present a refined method of climate network construction and analysis (Chap. 3). In Chap. 4, we describe a hierarchy of *surrogates on the time series and network level*, some of which were developed specifically for this work, and explain how they can be used to test the statistical significance of our results. In the following, we give an account of our central result, the detection of the *backbone of the climate network* formed by a network of channels of high dynamical information flow in the global surface air temperature (SAT) field (Chap. 5). We also touch upon the extension of our methodology to *spatially regional and temporally seasonal* climatological data sets and report interesting results on the seasonality of a regional SAT climate network encompassing the Indian Ocean basin (Chap. 6). Finally, we give some *concluding remarks and an outlook* in Chap. 7.

CHAPTER 2

Elements of complex network theory

Complex network theory in general and inevitably this study of complex networks in the climate system in particular are founded on graph theory. The latter is a branch of mathematics that studies graphs: structures used to model pairwise relations between objects taken from some collection. One of the first works in the domain of graph theory was the famous solution of the Königsberg bridge problem presented by Leonhard Euler in 1736 (Euler (1736)). We use it as an example in Sect. 2.1 to illustrate the fundamental definitions provided there, that we rely on many times in the course of this thesis. For a detailed account on the theory, algorithms and applications of (di)graphs we refer to (Bang-Jensen and Gutin (2006)).

Since Euler's time, graph theory was mainly concerned with the properties of ordered graphs, prominent examples being the graphs used to model crystal lattices, a chessboard or the hexagonal arrangement of combs in the hives of bees. It took more than 200 years until two hungarian mathematicians, Paul Erdős and Albért Rényi, guided the attention of the scientific community towards the opposite end of the spectrum of regularity, ranging from perfect order to complete randomness. From the late 1950's to the mid 1960's they developed the theory of random graphs in eight seminal papers (Erdős and Rényi (1959, 1960, 1961a,b, 1963, 1964, 1966, 1968)). In the spirit of Erdős and Rényi, scientists from disciplines as diverse as social science, electrical engineering and biology now began to regard their subjects of study, *e.g.*, acquaintance networks, power grids and food webs, as entirely random networks [1]. Alas, in the course of the following decades it was realized that this description of natural networks was not satisfactory either. Neither perfect order, nor complete randomness appeared to be promising paradigms to gain a deeper understanding of complex network-structured systems found in the real world.

Finally, Duncan Watts and Steven Strogatz in 1998 published the small-world network model, that in one of its flavors equips a regular network with additional random edges (Watts and Strogatz (1998)). By combining order and randomness elegantly, the model

1 We introduce Erdős-Rényi random graphs as well as various generalizations, and discuss their use as models of real world networks in Chap. 4.

was able to describe some of the properties of complex real world networks that had been perceived as paradoxical before. The work of Watts and Strogatz launched a great effort of research on what came to be known as complex network theory. It is important to note that the networks found in nature are typically too large to simply draw them on a piece of paper and then to extract all properties of interest by just looking at them long enough. Bearing this fact in mind it is straightforward to understand the dominance of physicists in the field of complex network theory: It was necessary to introduce statistical concepts to the study of networks that are abundant in all modern theories of physics, *e.g.*, quantum mechanics, thermodynamics, statistical mechanics or solid state physics. Hence, the present work on complex climate networks also relies heavily on statistical network measures, that we introduce in Sect. 2.2 and Sect. 2.3. The development of complex network theory continues to the time of writing of this thesis and has spawned applications in many different branches of science, including climate science (Tsonis and Roebber (2004), Tsonis et al. (2006), Tsonis et al. (2008b), Tsonis and Swanson (2008), Yamasaki et al. (2008), Gozolchiani et al. (2008), Donner et al. (2008), Donges et al. (2008), Donges et al. (2009)).

Several reports on the (then) current state of complex network theory have been published in the form of review papers (Albert and Barabási (2002), Newman (2003), Boccaletti et al. (2006)). A useful survey of the plethora of network measures is presented by (da F. Costa et al. (2007)). Furthermore there are some noteworthy popular accounts on the emerging field of complex network theory, its history and applications (Barabási (2002), Buchanan (2002), Watts (2003)).

In this chapter we first give a brief introduction into some basic concepts of graph theory in Sect. 2.1. Sect. 2.2 contains definitions of general network measures, whereas Sect. 2.3 presents measures tailored for spatially embedded complex networks.

2.1. Foundations of graph theory

Legend holds that during the early 18th century, the citizens of Königsberg mused on an entertaining question: Could an ambler go on a round trip of old Königsberg, starting and arriving at the same location and crossing each of the seven bridges across the river Pregel exactly once? Leonhard Euler realized that the exact location and shape of the four land areas and seven bridges was not of interest for the solution. In fact, the essential structure underlying the Königsberg bridge problem is the topology of connections of land areas and bridges (Fig. 2.1). To model this situation, one can construct a *graph* by assigning *vertices* to the land areas and *edges* to the bridges.

Definition 2.1.1 (Graph) *An undirected* graph *or network is defined as an ordered pair* $G := (V,E)$ *containing a finite set* $V = \{1,...,N\}$ *of vertices or nodes together with a finite set* E *of edges or links* $\{i,j\}$ *with* $i,j \in V$*, which are 2-element subsets of* V*.* $N = |V|$ *denotes the size (number of vertices) of* G*,* $L = |E|$ *the number of edges of* G*. A graph is called*

dense, *if L is close to the maximum number of edges L_{max}. It is called* sparse, *if $L \ll L_{max}$. The following types of graphs are considered in this work:*

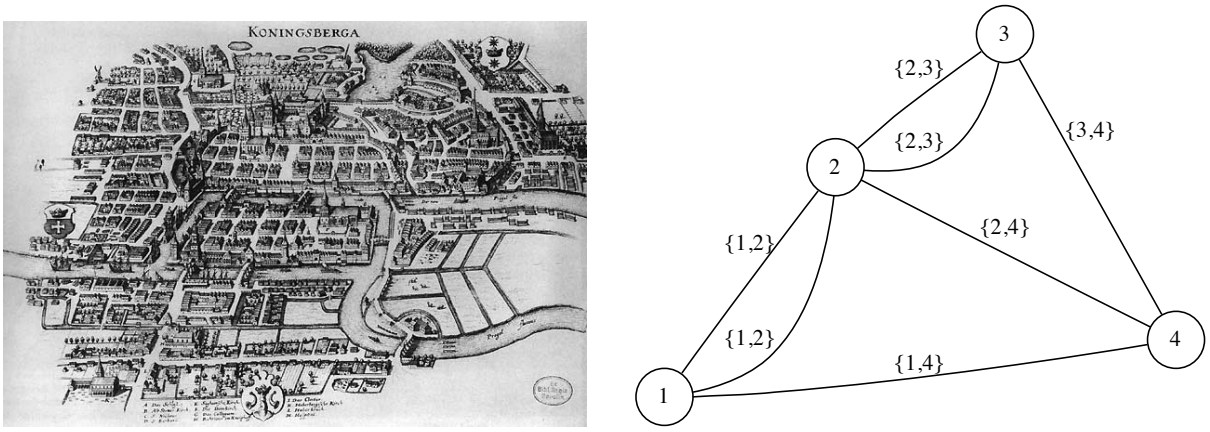
(i) In a simple *graph, one and only one edge $\{i,j\} \in E$ can exist between a pair of vertices and self-loops of the type $\{i,i\}$ are not allowed.*

(ii) A weighted *simple graph possesses an associated mapping $E \to \mathbb{R} : \{i,j\} \mapsto W_{ij}$ assigning a real number $W_{ij} \in \mathbb{R}$ to each edge $\{i,j\} \in E$. $\boldsymbol{W}$ is referred to as the weight matrix of G.*

(iii) For a directed *graph or* digraph, *E is a set of directed edges, directed links or arcs (i,j) that are ordered pairs of elements of V.*

(iv) A spatially embedded *graph carries a mapping $V \to \mathbb{S} : i \mapsto \boldsymbol{r}_i$ assigning each vertex $i \in V$ to an element $\boldsymbol{r}_i \in \mathbb{S}$ of a metric vector space $\mathbb{S}$ and a metric $l : \mathbb{S} \times \mathbb{S} \to \mathbb{R} : (\boldsymbol{r}_i,\boldsymbol{r}_j) \mapsto l_{ij}$. $\boldsymbol{r}_i$ is called the coordinate vector of a vertex i.*

Naturally, graphs can be used to construct mathematical models of any collection of objects equipped with pairwise relations. Their application may range from networks of sexual contacts of university students to a co-ownership network of Silicon Valley high tech companies. In the context of climate networks, vertices are associated to geographical regions, whereas edges represent significant statistical interrelationships of some climatological



(a) (b)

Figure 2.1 (a) Engraving of the prussian city of Königsberg in 1652 by Mattheus Merian-Erben (Merian-Erben (1652)). Four distinct land areas separated by the river Pregel were connected by seven bridges. (b) Representation of the Königsberg graph $G_K := (V_K,E_K)$ of $N = 4$ vertices and $L = 7$ edges. $V_K = \{1,2,3,4\}$, $E_K = \{\{1,2\},\{1,2\},\{1,4\},\{2,3\},\{2,3\},\{2,4\},\{3,4\}\}$. Circles symbolize vertices, the lines connecting them represent edges. The Königsberg graph is undirected. It is *neither* weighted *nor* simple, since multiple edges exist between vertices 1,2 and 3.

observable between a pair of regions. Now let us return to the bridge problem. If the task was to solve it for a much larger city with many rivers, channels and bridges [1], it might be a good idea to divide the city into smaller districts that can be analyzed separately. In the context of graph theory this corresponds to the introduction of *subgraphs*.

Definition 2.1.2 (Subgraph) *A* subgraph $G' := (V',E')$ *of a graph* $G := (V,E)$ *contains a vertex set* $V' \subset V$ *and an edge set* $E' \subset E$*, where* $\forall\{i,j\} \in E' : i,j \in V'$*.*

Note that in the Königsberg picture, a simple graph models a city where every pair of land areas is connected by at most one single bridge. A directed graph could be used to capture the traffic rules of old Königsberg: Some bridges might have allowed traffic to pass in one direction only, others in both (Fig. 2.2(a)). A weighted graph would be able to describe the carrying capacity of the bridges or the duties imposed on any traverser by the city council (Fig. 2.2(b)). We display the relation between the various types of simple graphs used in this thesis in Fig. 2.3.

To perform computations on a simple graph, it can be represented by an *adjacency matrix*. This type of representation is computationally feasible on current desktop computers only for relatively small and dense graphs with $N \leq 10^4$, one major reason being that the memory needed to store the adjacency matrix on a computer grows as N^2. Therefore we use the matrix representation mainly for illustrative purposes in this work, internally most calculations rely on an adjacency list representation of graphs (Bang-Jensen and Gutin (2006) and Appx. E).

Definition 2.1.3 (Adjacency matrix) *A simple directed graph* $G := (V,E)$ *can be represented by an* adjacency matrix $\boldsymbol{A} \in GL_N(\mathbb{R})$ *with elements*

$$A_{ij} = \begin{cases} 0 & (i,j) \notin E \\ 1 & (i,j) \in E, \end{cases} \tag{2.1}$$

where $i,j \in V$*. For an undirected graph,* $\boldsymbol{A}$ *is symmetric.*

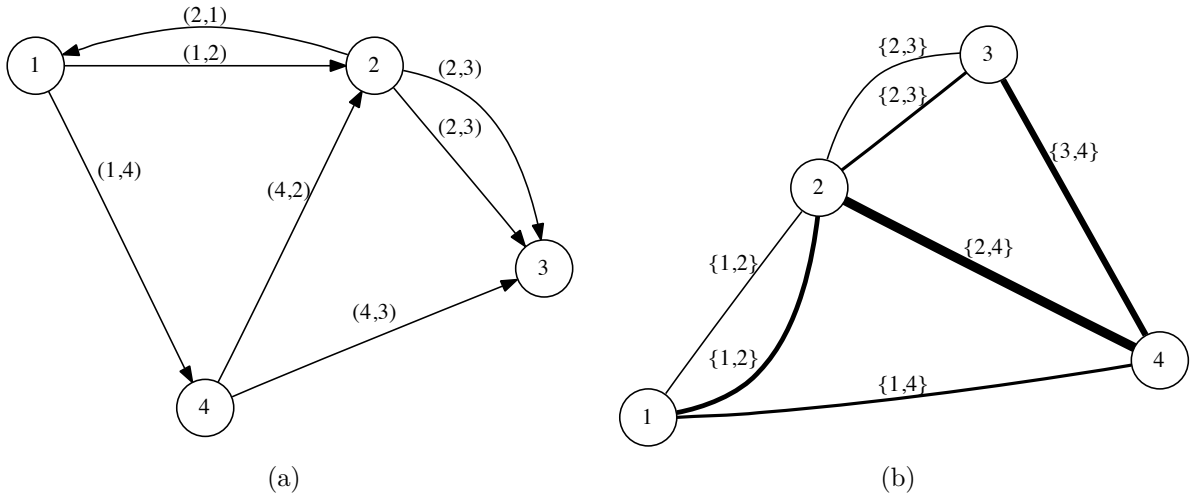
Now that we have defined graphs we are in a position to describe mathematically a stroller's trajectory on the Königsberg graph by introducing the concept of a *walk*.

Definition 2.1.4 (Walk) *A* walk *on the graph* $G := (V,E)$ *is an alternating sequence of vertices and edges, that starts and ends with a vertex. The vertices that precede and follow an edge in the sequence are the end vertices of that edge. A walk is called*

(i) open*, if its start and end vertices are different,*

(ii) closed*, if its start and end vertices are identical,*

1 Just think of Venice or Amsterdam.



(a) (b)

Figure 2.2 In this thesis we use undirected graphs (Fig. 2.1(b)), directed graphs and undirected weighted graphs. We stress that in contrast to the Königsberg example, all climate networks are simple graphs. (a) Illustration of a Königsberg digraph. (b) A weighted version of the Königsberg graph.

(iii) simple*, if no vertex (and therefore no edge) is visited more than once,*

(iv) directed*, if it is defined on a digraph and contains arcs instead of edges.*

In the language of graph theory, Euler was posed the question whether a simple and closed walk existed on the Königsberg graph. Simple walks are referred to as *paths* and play an essential role in the definition of distances between the vertices of a graph. The simple and closed walk of the Königsberg bridge problem is known today as an Eulerian *cycle*, a graph containing an Eulerian cycle is called Eulerian.

Definition 2.1.5 (Path) *A* path $\pi(i,j)$ *is a simple walk on the graph* $G := (V,E)$ *from start vertex* $i \in V$ *to end vertex* $j \in V$*. i is said to be* reachable *from j, if a path containing both vertices exists on G. The length of the path* $|\pi(i,j)|$ *is defined as the number of edges the path contains. We denote the set of all existing different paths from i to j by* $\mathcal{P}(i,j)$*. The* shortest path length *from i to j is then given by* $d_{ij} = \min_{\pi(i,j)\in\mathcal{P}(i,j)} |\pi(i,j)|$*. In this work, we generally use Dijkstra's algorithm to calculate the topological distance matrix* d_{ij} *(Dijkstra (1959)).*

Definition 2.1.6 (Cycle) *A* cycle *is a simple and closed walk on a graph G.*

Euler was able to prove that an Eulerian cycle does not exist on the Königsberg graph. Finally consider a city including some small islands that are only reachable by boat. Some of these islands may be connected by bridges, some of them may not. In this case, the city's connectivity graph contains collections of land areas which are not connected among each other by bridges. We refer to these unconnected subgraphs as *components*. Many graphs

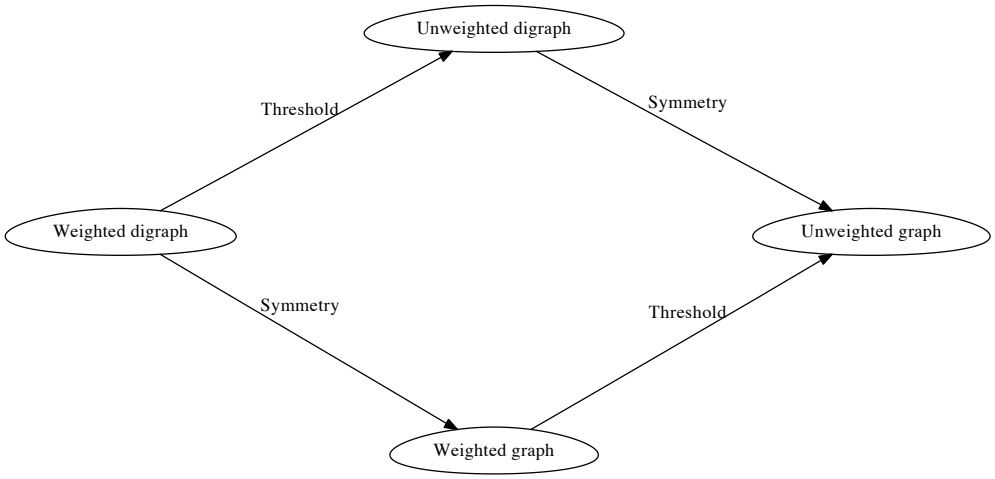


Figure 2.3 This diagram illustrates the structural relationships between the different types of simple graphs considered in this work (da F. Costa et al. (2007)). The operation *symmetry* transforms a simple (weighted) digraph into a simple (weighted) graph by symmetrizing the adjacency matrix $\mathbf{A} \mapsto \Theta\left(\mathbf{A} + \mathbf{A}^T\right)$ (and the weight matrix $\mathbf{W} \mapsto \mathbf{W} + \mathbf{W}^T$), where $\mathbf{M}^T$ denotes the transpose of a square matrix $\mathbf{M}$ and $\Theta(x)$ is the Heaviside function. The operation *threshold* transforms a weighted simple (di)graph into a simple (di)graph by thresholding the weight matrix $A_{ij} \mapsto \Theta\left(W_{ij} - \tau\right)$. All edges with a weight $W_{ij} > \tau$ will be included in the unweighted simple (di)graph, where $\tau \in \mathbb{R}$ is an arbitrary threshold value.

found in nature contain one component that is much larger than all other components and includes nearly all of the graph's vertices. In the case of a city connectivity graph, this *giant component* may correspond to the main land area, while the other small components model islands.

Definition 2.1.7 (Component) *A* component *is a maximally connected subgraph* $G' := (V',E')$ *of a graph* $G := (V,E)$. *That is, all vertices in* V' *are reachable from all other vertices in* V' *and no vertex* $i \in V \backslash V'$ *is reachable from any vertex* $j \in V'$. *We refer to the largest component of* G *with size* $\mathcal{O}(N') = N$ *as the* giant component.

2.2. Topological network measures

The edge density of an undirected network is given by

$$\rho = \frac{L}{\binom{N}{2}} = \frac{\langle k_v \rangle_v}{N}, \tag{2.2}$$

L being the number of edges in the network and $\langle k_v \rangle_v$ the mean vertex degree. $\binom{N}{2} = N(N-1)/2$ gives the maximally possible number of edges. The network measures defined below were selected for this study, because they allow us to compare different aspects of climate network topology on local, mesoscopic and global scales (Table 2.1) and are

Table 2.1 Two dimensional classification of network measures into topological scales vs. fields, distributions and scalar measures.

	local	mesoscopic	global
field	k_v, AWC_v, AED_v, RWB_v	$\mathcal{C}_v$	CC_v, BC_v
distribution	$p(k)$, $p_E(l)$		
scalar	$H(A,B)$	$\mathcal{C}$	$\mathcal{L}$

well established in the literature (Newman (2003), Albert and Barabási (2002), Boccaletti et al. (2006), Freeman (1979)). Degree centrality, the related area weighted connectivity, random walk betweenness and the Hamming distance use only local information on the direct neighborhood of a vertex v. In contrast, closeness and betweenness centrality as well as the average path length include global topological information by relying on shortest paths between pairs of vertices in the network. This is why we refer to the latter three as global measures. On the mesoscopic scale, the local and average clustering coefficient depend only on information about neighbors and next neighbors of vertices. The concept of topological scales is elaborated in greater detail in Zamora-López (2008). We refer to measures assigning a real number $g_v \in \mathbb{R}$ to each vertex $v \in V$ via a mapping $V \to \mathbb{R} : v \mapsto g_v$ as *fields*. *Scalar measures* produce a single real number for the whole graph.

2.2.1. Local measures

2.2.1.1. Degree centrality

The *degree* or *degree centrality* (Freeman (1979)) k_v gives the number of first neighbors of a vertex v and can be calculated from the network adjacency matrix A_{ij} using

$$k_v = \sum_{i=1}^{N} A_{vi}. \tag{2.3}$$

Vertices with exceptionally high degree centrality are usually referred to as *hubs* or *super-nodes*. We extend the use of this term to regions of spatially adjacent vertices with high degree centrality.

2.2.1.2. Degree distribution

The *degree distribution* $p(k)$ is of great interest in the analysis of complex networks and will be used in this work to study the role of super-nodes in seasonal climate networks (Chap. 6). $p(k)$ is an estimator of the probability density function (PDF) of degree centrality k_v, hence it gives the probability to find a vertex with degree k when drawing randomly from the vertex set V. Networks having a power-law degree distribution $p(k) \propto k^{-\gamma}$ with exponent γ are commonly referred to as *scale-free* networks (Newman (2003)). Since there has been

frequent misuse of this term in the literature (Li et al. (2005)), great care should be taken when assigning the scale-free property to real world finite networks. We therefore prefer to speak of a *fat-tailed* degree distribution in the context of this work, even though it has been claimed that climate networks are scale-free (Tsonis and Roebber (2004), Tsonis et al. (2006), Tsonis and Swanson (2008), Tsonis et al. (2008b)).

2.2.1.3. Random walk betweenness

Complementary to the shortest path betweenness centrality (Eq. 2.14), it is interesting to study the centrality of a vertex v with respect to the average flow of random walkers on the network. Here we derive *random walk betweenness* within the framework of the more general search centrality developed by Arenas et al. (2003). Note that an equivalent measure was later independently introduced by Newman (2005).

We focus on a single information packet at vertex i whose destination is vertex k, *i.e.*, a packet searching for k. The probability for the packet to go from i to a different node j on its way to k is denoted by p_{ij}^k. Note that the packet is *removed* as soon as it arrives at its destination k, *i.e.*, $p_{kj}^k = 0, \forall j$. The precise form of p_{ij}^k depends on the search algorithm. For an absorbing random walk we obtain

$$p_{ij}^k = (1 - \delta_{ik}) \frac{A_{ij}}{h_i}, \tag{2.4}$$

where $h_i = \sum_{j=1}^{N} A_{ij}$ denotes the degree of i and δ_{ik} the Kronecker delta. The first factor of the product takes care of the absorption of the randomly walking information packet at the target vertex k, while the second describes that the packet can otherwise proceed to any of the neighbors of i with equal probability h_i^{-1} in the next time step. The probability of the random walker to go from i to j in n steps is given by

$$P_{ij}^k(n) = \sum_{l_1, l_2, \ldots, l_{n-1}} p_{il_1}^k p_{l_1 l_2}^k \ldots p_{l_{n-1} j}^k, \tag{2.5}$$

because all single steps are independent. The sum is taken over all walks $(i,l_1,\ldots,l_{n-1},j)$ of length n between vertices i and j. Defining the matrices $\mathbf{p}^k$ and $\mathbf{P}^k(n)$ with elements p_{ij}^k and $P_{ij}^k(n)$ respectively yields

$$\mathbf{P}^k(n) = \left(\mathbf{p}^k\right)^n. \tag{2.6}$$

We can now consider the average number of times b_{ij}^k, that an information packet generated at i and with destination k passes j. Introducing the matrix $\mathbf{b}^k$ with elements b_{ij}^k,

$$\mathbf{b}^k = \left(\mathbf{1} - \mathbf{p}^k\right)^{-1} \mathbf{p}^k \tag{2.7}$$

holds[1], where **1** represents the identity matrix. The effective random walk betweenness

$$RWB_j = \sum_{ik} b_{ij}^k \tag{2.8}$$

can be calculated from the $\mathbf{b}^k$ by summing over all pairs of source and target vertices i,k. To calculate RWB_j numerically, N matrix inversions of an $N \times N$ matrix have to be performed, each of which scales as $\mathcal{O}(N^3)$ when using Gaussian elimination. The overall computational complexity of the algorithm is thus $\mathcal{O}(N^4)$. Note that RWB_j has to be calculated separately for each component if the network is not connected, because in this case $\mathbf{p}^k$ does not have full rank and hence cannot be inverted.

2.2.1.4. Hamming distance

The *Hamming distance* $H(A,B)$ of two labeled simple graphs with adjacency matrices A_{ij} and B_{ij} measures the fraction of edges that have to be changed to transform one graph into the other (Hamming (1950)). Both graphs must contain the same number of vertices N. Specifically, $H(A,B)$ is given by

$$H(A,B) = \langle XOR(A_{ij}, B_{ij}) \rangle_{ij} , \tag{2.9}$$

where

$$XOR(A_{ij}, B_{ij}) = \begin{cases} 1 & A_{ij} \neq B_{ij} \\ 0 & else. \end{cases} \tag{2.10}$$

Hamming distance is bounded by $0 \leq H(A,B) \leq 1$ and measures the global probability of non-equal entries in the two adjacency matrices. In our application we calculate the Hamming distance of two graphs with approximately equal edge density ρ.

To evaluate the significance of this measurement, we compare it with the expected Hamming distance $H^R(\rho)$ of two independent Erdős-Rényi random graphs of edge density ρ (Erdős and Rényi (1959)). The probability that the entries A_{ij} and B_{ij} differ between the two random graph adjacency matrices is given by $\mathrm{p}\left(A_{ij} \neq B_{ij}\right) = \mathrm{p}\left(A_{ij} = 1\right) \mathrm{p}\left(B_{ij} = 0\right) + \mathrm{p}\left(A_{ij} = 0\right) \mathrm{p}\left(B_{ij} = 1\right) = \rho(1-\rho) + (1-\rho)\rho = 2\rho(1-\rho)$. Since all entries are independent, taking the expectation value reveals the given expression $H^R(\rho) = \langle \mathrm{p}\left(A_{ij} \neq B_{ij}\right) \rangle_{ij} = 2\rho(1-\rho)$. The expected Hamming distance $H^R(\rho)$ gives a reference point against which to judge the similarity of two graphs. We will make use of it in Sect. 3.3.3 to compare the performance of two correlation measures in climate network construction.

1 Specifically, $\mathbf{b}^k = \sum_{n=1}^{\infty} \mathbf{P}^k(n) = \sum_{n=1}^{\infty} \left(\mathbf{p}^k\right)^n = \sum_{n=0}^{\infty} \left(\mathbf{p}^k\right)^n - \mathbf{1} = \left(\mathbf{1} - \mathbf{p}^k\right)^{-1} - \mathbf{1} = \left(\mathbf{1} - \mathbf{p}^k\right)^{-1} \mathbf{p}^k$. For the geometric series of matrices to converge as $\sum_{n=0}^{\infty} \left(\mathbf{p}^k\right)^n = \left(\mathbf{1} - \mathbf{p}^k\right)^{-1}$, all eigenvalues μ_i of $\mathbf{p}^k$ have to fulfill $|\mu_i| \leq 1$.

2.2.2. Mesoscopic measures

2.2.2.1. Local clustering coefficient

We refer to $\mathcal{C}_v$ as the *local topological clustering coefficient* or *Watts-Strogatz clustering coefficient* (Watts and Strogatz (1998)) of a vertex v. It gives the probability, that two randomly chosen first neighbors of v are also neighbors. With Γ_v being the set of first neighbors of v and $e(\Gamma_v)$ the number of edges connecting the vertices within the neighborhood Γ_v, the clustering coefficient can be written as

$$\mathcal{C}_v = \frac{e(\Gamma_v)}{\binom{k_v}{2}}, \tag{2.11}$$

where the binomial coefficient $\binom{k_v}{2} = \frac{1}{2}k_v(k_v-1)$ gives the maximum number of edges in Γ_v. The local clustering coefficient is normalized to $0 \leq \mathcal{C}_v \leq 1$.

2.2.2.2. Global clustering coefficient

We speak of the *(global) clustering coefficient* $\mathcal{C}$ as the mean Watts-Strogatz clustering coefficient

$$\mathcal{C} = \langle \mathcal{C}_v \rangle_v . \tag{2.12}$$

2.2.3. Global measures

2.2.3.1. Closeness centrality

Closeness centrality CC_v measures the inverse average topological distance of vertex v to all others in the network (Freeman (1979)),

$$CC_v = \frac{N-1}{\sum_{i=1}^{N} d_{vi}}, \tag{2.13}$$

where the topological distance or shortest path length d_{ij} is the minimum number of edges that have to be crossed to travel from vertex i to vertex j ($d_{vv} = 0$ by definition). If i and j are not connected, the maximum topological distance in the graph $d_{ij} = N-1$ is used in the sum. Closeness centrality is normalized to $0 \leq CC_v \leq 1$. Following our definition, CC_v is large, when v is topologically close to the rest of the network. One should bear this in mind, because some researchers have used the inverse of our definition (Freeman (1979), Zamora-López (2008)).

2.2.3.2. Betweenness centrality

Assume that information travels through the network on shortest paths. There are σ_{ij} shortest paths connecting two vertices i and j (Fig. 2.4). We then regard a vertex v to be

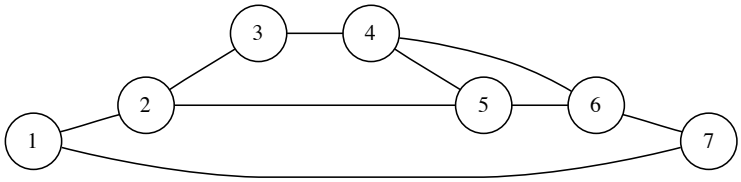


Figure 2.4 We use this simple graph of $N = 7$ and $L = 9$ to illustrate the numbers σ_{ij} and $\sigma_{ij}(v)$, that are important for the calculation of betweenness centrality. There are three shortest paths of length $d_{14} = 3$ connecting vertices 1 and 4, hence $\sigma_{14} = 3$. In contrast, there is only one shortest path of length $d_{16} = 2$ between 1 and 6, *i.e.*, $\sigma_{16} = 1$. Vertex 2 is contained in two of the three shortest paths between 1 and 4, thus $\sigma_{14}(2) = 2$.

an important mediator for the information transport in the network, if it is traversed by a large number BC_v of all existing shortest paths. Mathematically, the *betweenness* BC_v can be expressed by

$$BC_v = \sum_{i,j \neq v}^{N} \frac{\sigma_{ij}(v)}{\sigma_{ij}}, \tag{2.14}$$

where $\sigma_{ij}(v)$ gives the number of shortest paths from i to j, that include v (Freeman (1977, 1979)). Here the contribution of shortest paths is weighted by their respective multiplicity σ_{ij}. For the calculation of betweenness we rely on the fast $\mathcal{O}(LN)$ algorithm introduced by Newman (Newman (2001a,b)).

2.2.3.3. Average path length

The *average or characteristic path length* $\mathcal{L}$ of a graph is defined as the average topological distance between all pairs of vertices,

$$\mathcal{L} = \frac{1}{\binom{N}{2}} \sum_{i<j} d_{ij}. \tag{2.15}$$

Disconnected pairs of vertices are not included in the average, for a detailed discussion see (Newman (2003)).

2.3. Spatially embedded networks and associated measures

The climate networks studied in this work are constructed from two dimensional fields of climatological observables (Chap. 3), that can be approximated to reside on a spherical

surface centered at the Earth's center of mass [1]. They therefore have to be treated as networks spatially embedded on the 2-sphere of radius R_{earth} in $\mathbb{R}^3$, *i.e.*, $\mathbb{S} = S^2(R_{earth})$. Vertex v has an associated two dimensional coordinate vector $\mathbf{r}_v = (\lambda_v,\phi_v)$ with $\lambda_v \in [-90°,90°]$ and $\phi_v \in [0°,360°]$ denoting latitude and longitude respectively. As the metric l we choose the great circle distance

$$l_{ij} = R_{earth} \arccos(\sin(\lambda_i)\sin(\lambda_j) + \cos(\lambda_i)\cos(\lambda_j)\cos(\phi_i - \phi_j)), \tag{2.16}$$

where the edge distance l_{ij} corresponds to the geodesic distance of vertices i,j on S^2.

For the study of complex networks embedded on a spherical surface in three dimensional space it is useful to introduce network measures, that take this geographical nature into account explicitly (Boccaletti et al. (2006)). The embedding induces additional spatial constraints to the network topology and evolution (Rozenfeld et al. (2002), Warren et al. (2002), Kosmidis et al. (2008)). We assess the effect of these spatial contraints, *e.g.*, a given edge distance distribution or average edge distance field, on the topological network measures defined above by designing novel classes of spatially constrained surrogate network models (Sect. 4.2.4). As previous results show, some of the most interesting features of climate networks are detectable only by geographical measures (Tsonis et al. (2006), Tsonis and Swanson (2008)).

Note that for the data sets analyzed here (Sect. 3.1), vertices are not distributed homogeneously on the earth's surface. The density of vertices increases from the equator towards the poles. This induces an inherent bias in the network measures studied, which prompts to use area weighted generalizations of the standard complex network measures, *e.g.*, area weighted connectivity is the generalization of degree centrality. We have performed extensive studies of climate networks constructed from data interpolated to different grids and resolutions and find, that our results (Sect. 3.3) are not altered significantly by the vertex density bias (Sexton et al. (2009)). This holds particularly for the highly interesting path based measures on the global topological scale.

Within the classification scheme for network measures introduced in Sect. 2.2, the geographical measures defined below all belong to the local topological scale (Table 2.1). The reason is, that area weighted connectivity, average edge distance and the edge distance distribution depend exclusively on the lengths of single edges and the local topological adjacency relations between vertices.

1 This is only an approximation, since within the same two dimensional gridded climatological field, grid points v may lie at different distances r_v from the Earth's center of mass. For example, surface air temperature and pressure are defined to be measured two meters above the Earth's surface, with r_v depending on the local surface topography. However, the spherical shell is very good approximation, because the typical vertical separation of grid points is much smaller than their horizontal distance.

2.3.1. Area weighted connectivity

The *area weighted connectivity*

$$AWC_v = \frac{\sum_{i=1}^{N} A_{vi} \cos(\lambda_i)}{\sum_{i=1}^{N} \cos(\lambda_i)}, \tag{2.17}$$

is closely related to the degree centrality k_v of v. It corrects for the fact that in geographical networks defined on a grid, vertices correspond to regions of different area on the earth's surface. For the angularly equidistant grids considered in this work, the corresponding area of vertex v is proportional to the cosine of latitude λ_v (see Sect. 3.1.1). AWC_v can be interpreted as the fraction of the earth's surface area a vertex is connected to (Tsonis et al. (2006)). AWC is thus normalized to $0 \leq AWC_v \leq 1$.

2.3.2. Average edge distance

Average edge distance AED_v measures the average angular great circle distance to the first neighbors of vertex v (Tsonis et al. (2006), Jones (2007)),

$$AED_v = \frac{1}{k_v \frac{1}{N} \sum_{i=1}^{N} l_{vi}} \sum_{i=1}^{N} A_{vi} l_{vi}. \tag{2.18}$$

The average edge distance is normalized such that $ALD_v = 1, \forall v \in V$ for a fully connected graph to correct for geometric biases in regional climate networks.

2.3.3. Edge distance distribution

The *edge distance distribution* $p_E(l)$, the PDF of edge distance l_{ij} calculated for all pairs $\{i,j\}$, allows to assess the tendency of long distance links to arise in a climate network and enables the quantitative comparison of different climate networks concerning this central property. It is important to realize that for a geographical network, $p_E(l)$ has a purely geometric component $p_{geom}(l)$ and a component $p_{net}(l)$ describing the intrinsic properties of the underlying network structure. Assuming independence

$$p_E(l) = p_{geom}(l) p_{net}(l), \tag{2.19}$$

one obtains $p_E(l)$ by dividing the empirically found link distance distribution by the geometric distribution. We speak of $p_{net}(l)$ as the *intrinsic edge distance distribution*. For a global network, $p_{geom}(l)$ can be derived analytically (see below)

$$p_{geom}(l) = \frac{1}{2} \sin\left(\frac{l}{R_{earth}}\right), \tag{2.20}$$

whereas it is calculated numerically from the edge distance matrix l_{ij} of all possible vertex pairs $\{i,j\}$ for regional networks with a lower degree of symmetry. In this work we use $p_{net}(l)$,

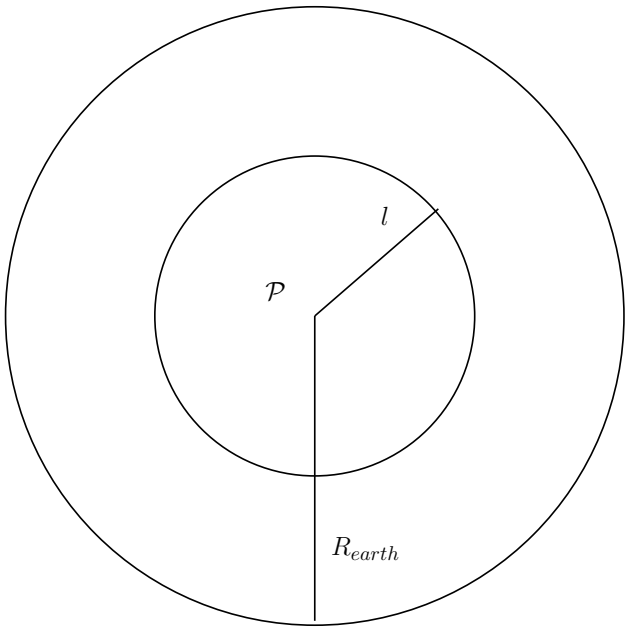


Figure 2.5 Schematic sketch for the derivation of $p_{geo}(l)$ for a global network.

because it enables an objective comparison of climate networks with different geometries, *i.e.*, regional and global, and reflects the network property of interest.

Derivation of $p_{geo}(l)$ for a global network Considering a network embedded on S^2 with a homogenous vertex density, each geodesic distance l is weighted by the circumference of a circle with radius $r(l) = R_{earth} \sin(l/R_{earth})$ (Fig. 2.5). Because of symmetry it is sufficient to perform the calculation for one arbitrary point $\mathcal{P}$ on the spherical surface. One obtains

$$\begin{aligned} p_{geo}(l) &= \frac{2\pi r(l)}{\int_0^\pi dl 2\pi r(l)} \\ &= \frac{1}{2} \sin\left(\frac{l}{R_{earth}}\right). \end{aligned} \tag{2.21}$$

2.4. Summary

In this chapter, we have first outlined the development of complex network theory starting from the solution of the Königsberg bridge problem by Leonhard Euler. We have then defined a minimal set of concepts drawn from graph theory and in the following introduced the required measures of complex network theory for pure as well as spatially embedded networks. Their classification into local, mesoscopic and global topological scales serves to tame the plethora of complex network measures employed in this work by allowing a well structured discussion of our results.

CHAPTER 3

Construction of climate networks

...we seem to live in a universe where orderly structures form whenever there is a flow of energy.

Erich Jantsch, "The Self-organizing Universe" (1980)

In this chapter we present the refined climate network construction methodology developed during the writing of this thesis. In all earlier related works (Tsonis and Roebber (2004), Tsonis et al. (2006), Tsonis et al. (2008b), Tsonis and Swanson (2008), Yamasaki et al. (2008), Gozolchiani et al. (2008), Donner et al. (2008)), researchers have used the linear cross-correlation function of pairs of anomaly time series to quantify the degree of statistical interdependence between different spatial regions. But the highly nonlinear processes at work in the climate system call for the application of nonlinear methods to obtain more reliable results. In a recent work on structures in the betweenness centrality field of climate networks (Donges et al. (2008)), we have introduced mutual information (Kantz and Schreiber (2004)) as a measure of statistical interdependence to climate network construction. The mutual information allows to capture nonlinear relationships between time series. We found that, while many properties of climate networks generated using the Pearson correlation and the mutual information at zero lag are qualitatively and quantitatively similar, the betweenness centrality field shows much greater deviations between the two construction methods. To check the possibility, that these pronounced differences are a signature of nonlinear processes in the climate system, and to bridge the gap between our nonlinear network construction method and the techniques previously used, we present a systematic statistical similarity study of the resulting climate networks. We show, that over a wide range of relevant edge densities (the fraction of the maximum number of possible edges present in the network), a high degree of similarity is maintained on local and mesoscopic topological scales. Furthermore, we address some of the more pronounced differences on the global topological scale, that are uncovered by betweenness centrality, and their possible relation to nonlinear processes in the climate system.

The organization of the chapter is the following: We first describe the data and the filtering and normalization procedures applied to it (Sect. 3.1). We proceed to develop in

detail the method of climate network construction (Sect. 3.2). In Sect. 3.3, we present the systematic comparison of the measures obtained from Pearson correlation and mutual information climate networks, respectively. Furthermore we provide a concise climatological interpretation of our results (Sect. 3.4). The conceptual shortcomings of our method are discussed in some detail in Sect. 3.5, where we particularly focus on the transitivity problem. We also provide an account of the formal relationships of our climate network construction method to classical methods of multivariate climate data analysis (Sect. 3.6). Some conclusions are drawn in Sect. 3.7.

3.1. Data

3.1.1. Description

We utilize the monthly averaged global surface air temperature (SAT) field for climate network construction to maintain consistency with earlier works that analyzed the same field (Tsonis and Roebber (2004), Tsonis et al. (2006), Tsonis et al. (2008b), Tsonis and Swanson (2008), Yamasaki et al. (2008), Gozolchiani et al. (2008), Donges et al. (2008)). The SAT field allows to directly capture the complex dynamics on the interface between ocean and atmosphere due to heat exchange and other local processes[1]. SAT therefore enables us to study atmospheric as well as oceanic dynamics within a common framework. Note that in principle we could use data from any climatological field for climate network construction, *e.g.*, surface air pressure, precipitation, air moisture content, sea surface temperature and salinity.

We use reanalysis data provided by the National Center for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) (Kistler et al. (2001)) and model output from the World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model data set (Meehl et al. (2007)). For optimal comparability with the reanalysis data, we choose a 20th century reference run[2] by the Hadley Centre HadCM3 model (Fig. 3.1). A data set consists of a regular spatiotemporal grid with time series $x_i(t)$ associated to every spatial grid point i at latitude λ_i and longitude ϕ_i. Start and end dates, length of time series $\mathcal{T}$, latitudinal resolution $\Delta\lambda$, longitudinal resolution $\Delta\phi$ and the number of vertices of the corresponding global climate network N are given in Table 3.1. Note that we remove the polar grid points at $\lambda \in \{-90°,90°\}$ from the data sets, since the poles are represented by rows of grid points with identical dynamics.

1 Surface air temperature (SAT) is defined as the air temperature at a height of 2 m above the surface.
2 20c3m, as defined in the IPCC AR4 (Intergovernmental Panel on Climate Change (2007)).

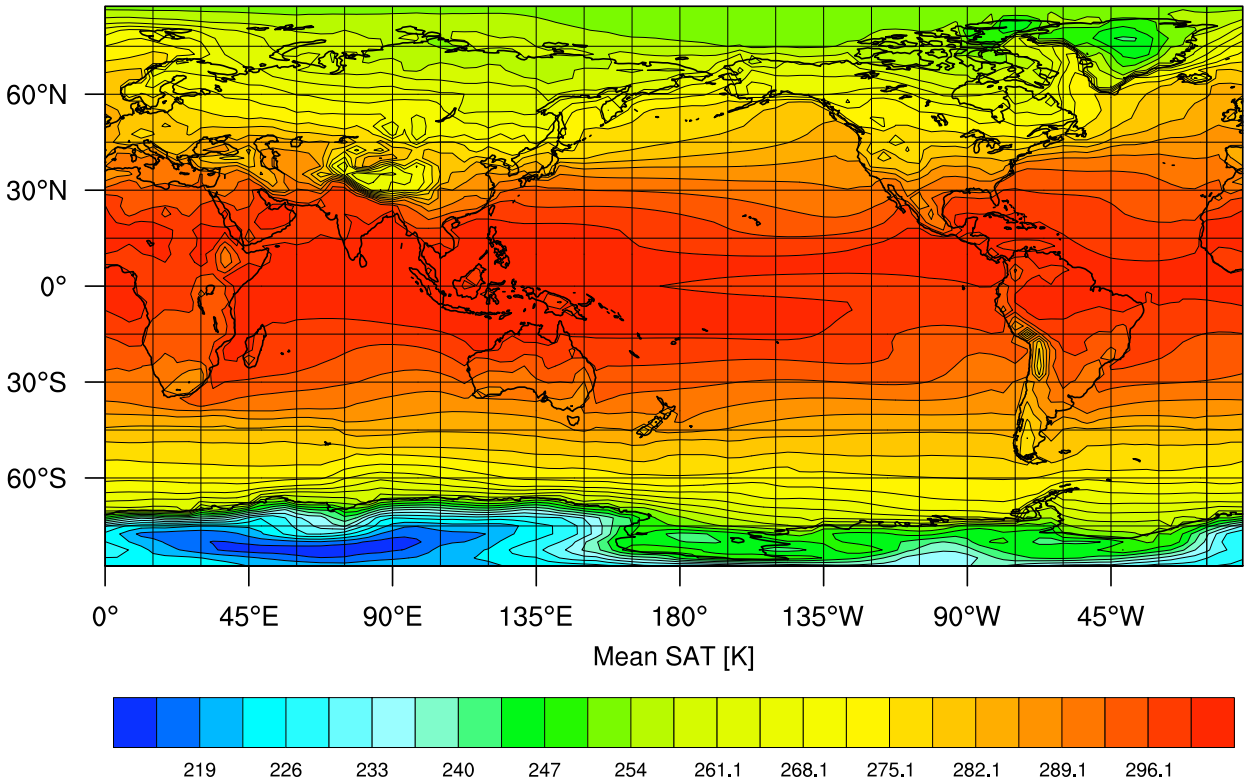


Figure 3.1 The mean surface air temperature field $\langle SAT_i(t)\rangle_t$ calculated from the HadCM3 SAT data set (Meehl et al. (2007)), both taken from the 20th century reference run described in Sect. 3.1.1.

3.1.2. Filtering and normalization

To minimize the bias introduced by the external solar forcing common to all time series in the data set, we calculate anomaly values, *i.e.*, remove the mean annual cycle by phase averaging. Relabeling the time series by month $m \in \{1,\ldots,12\}$ and year y mapping $x_i(t) \to x_i(y,m)$ one obtains anomaly time series $a_i(y,m) = x_i(y,m) - \langle x_i(y,m)\rangle_y$, that are consequently subjected to the inverse mapping $a_i(y,m) \to a_i(t)$. Here and in the following $\langle f(x)\rangle_x$ denotes the expectation value of observable f taken with respect to the variable x. Note that the anomaly time series already have zero mean. We furthermore normalize the anomaly time

Table 3.1 Properties of global model and reanalysis surface air temperature data sets.

	NCEP/NCAR reanalysis	HadCM3
Temporal coverage	Jan 1948 - Dec 2007	Jan 1860 - Dec 1999
$\mathcal{T}$ [months]	720	1680
$\Delta\lambda$ [°]	2.5	2.5
$\Delta\phi$ [°]	2.5	3.75
N	10224	6816

series to unit variance. Up to this point, we follow the method used previously by (Tsonis and Swanson (2008), Yamasaki et al. (2008)). It is known, that the annual cycle induces higher order effects such as seasonal variability of anomaly time series variance. We find that using only data from a particular season to avoid biases due to this effect does not alter our results substantially, so that we choose to use the whole data set for a more accurate evaluation of statistical interdependence.

3.2. Constructing climate networks

To clarify the physical rational behind our method of climate network construction, we discuss it within the framework of synchronization from dynamical systems theory (Pikovsky et al. (2001)). In a discretized model of the climate system, dynamical correlations can be envisioned as arising by (partial) synchronization of nonlinear oscillators on the grid that physically form a locally connected network. Even this simple network topology can generate nontrivial spatial patterns of synchronization (Arenas et al. (2008), Blasius and Tönjes (2005), Tönjes (2007)). The same is true for the synchronization of modes of variability in spatially continuous systems as the underlying fields of fluid- and thermodynamics (Boccaletti et al. (2002)), *e.g.*, SAT. Many measures of synchronization have been proposed and used to infer coupling strength and direction between connected nonlinear oscillators (Pikovsky et al. (2001), Rosenblum et al. (1996)). The Pearson correlation coefficient (Zhou et al. (2007)) and the mutual information (Schmidt et al. (2008)) were successfully employed to retrieve the network topology from the dynamics on the vertices alone.

The concept of synchronization provides a powerful paradigm to guide the enhancement of our understanding of the formation of (nonlinear) teleconnections in the climate system, and to stimulate the development of more advanced measures to detect these effects in measured data (Pikovsky et al. (2001), Boccaletti et al. (2002)). We hence propose that research aiming to construct networks from multivariate climatological data should be embedded within the framework of synchronization in complex networks (Arenas et al. (2008)).

3.2.1. Correlation measures

In the spirit of simplicity facing comparably short time series and desiring consistency with the literature, we choose to first use the standard Pearson correlation coefficient and then cross-check the results by introducing mutual information to climate network construction. The mutual information will allow to investigate nonlinear dynamical relationships (nonlinear teleconnections) that are not fully detectable by using the linear Pearson correlation coefficient (Brockwell and Davis (2002)). Note that we evaluate both measures at zero lag between time series. In principle, one can calculate a time delayed Pearson correlation (the cross correlation function) and mutual information (Kantz and Schreiber (2004)). This is appropriate when studying climate networks on smaller time scales using data sets with (sub-)diurnal resolution

(Yamasaki et al. (2008), Gozolchiani et al. (2008), Donner et al. (2008)). However, in the present work, we intend to study long term structural properties of the climate system on a scale of $\mathcal{O}(10^2)$ years using monthly averaged data. Most physical mechanisms of global information transfer in the SAT field, such as traveling Rossby waves, heat exchange between ocean and atmosphere or the advection of heat by surface currents in the ocean, act on time scales of less than one month. Therefore, it is reasonable to calculate the correlation measures at zero lag between anomaly time series.

3.2.1.1. Pearson correlation coefficient

The parametric empirical Pearson correlation coefficient $R_{ij} = \langle \hat{a}_i(t)\hat{a}_j(t) \rangle_t = R_{ji}$ estimates the strength of a linear relationship between two normalized time series $\hat{a}_i$ and $\hat{a}_j$, given those are normally distributed. It produces spurious results for not normally distributed observables and nonlinear relationships. Consequently it should be used with care when constructing climate networks. The non-parametric Spearman rank order correlation coefficient, that does not depend on the assumption of normally distributed observables, and R_{ij} are found to converge to the same value for nearly all pairs of time series taken from the data sets introduced in Sect. 3.1. The corresponding climate networks hence display close to identical network measures at all topological scales and we conclude, that utilizing the Pearson correlation coefficient to study linear climate networks is statistically justified here.

In contrast to the standard definition of teleconnectivity (Wallace and Gutzler (1981)), we do not limit our analysis to strongly negative correlations. As in earlier works on climate networks, we use the absolute value of Pearson correlation $P_{ij} = |R_{ij}| = P_{ji}$ to construct climate networks, since both large negative and positive values of Pearson correlation are indicative of a strong linear statistical interdependence.

3.2.1.2. Mutual information

In climate science, nonlinear measures of statistical interdependence have been successfully applied to uncover strongly nonlinear relationships of climate observables, *e.g.*, the phase coherence between ENSO and the Indian Monsoon (Maraun and Kurths (2005)). Mutual information from information theory is another nonlinear measure now widely applied in many fields of science, ranging from linguistics (Church and Hanks (1990)) to computational neuroscience (Schmidt et al. (2008)). The mutual information M_{ij} can be interpreted as the excess amount of information generated by falsely assuming the two time series $\hat{a}_i$ and $\hat{a}_j$ to be independent, and is able to detect nonlinear relationships (Kantz and Schreiber (2004)). By definition, M_{ij} is large if the two time series are highly linearly (anti)correlated. In contrast, a strongly nonlinear relationship between $\hat{a}_i$ and $\hat{a}_j$ yields large M_{ij}, but small P_{ij} (see the upper left quadrant in Fig. 3.2(c)). The mutual information can be estimated

using

$$M_{ij} = \sum_{\mu\nu} p_{ij}(\mu,\nu) \log \frac{p_{ij}(\mu,\nu)}{p_i(\mu)p_j(\nu)}, \tag{3.1}$$

where $p_i(\mu)$ is the probability density function (PDF) of the time series $\hat{a}_i$, and $p_{ij}(\mu,\nu)$ is the joint PDF of a pair $(\hat{a}_i,\hat{a}_j)$[1]. By definition, M_{ij} is symmetric, so that $M_{ij} = M_{ji}$. The standard unit of measurement of mutual information is the bit, if logarithms to base 2 are used.

We use a simple histogram approach with equally sized bins for all pairs $\{i,j\}$ to estimate the probability densities. Because the estimator (Eq. 3.1) is known to depend on bin size and partitioning (Schwarz et al. (1993), Hegger et al. (1999), Papana and Kugiumtzis (2008)), we use an identical partitioning for all $\{i,j\}$ to guarantee an optimal comparability of the M_{ij}. We select a bin number of 64, *i.e.*, $\mu,\nu \in \{1,\ldots,64\}$, that meets the Cochran criterion of at least 5 samples per bin for a typical time series length of $\mathcal{O}(10^3)$. The basic algorithm applied here is computationally much less expensive than more advanced methods proposed in the literature (Papana and Kugiumtzis (2008), Kraskov et al. (2004)), which is an important advantage when dealing with up to $\mathcal{O}(10^8)$ pairs in a global climate network. Our algorithm is feasible, since the application to network construction requires only the correct estimation of relative differences of M_{ij} between all pairs of time series. In other words, in our application systematic under- or overestimation of mutual information is not a problem, as long as the error stays approximately constant across all pairs.

3.2.2. Obtaining the network adjacency matrix

We now construct the climate network by thresholding the correlation measure matrix C_{ij} ($C_{ij} = P_{ij}$ or $C_{ij} = M_{ij}$), *i.e.*, only pairs of vertices $\{i,j\}$ that satisfy $C_{ij} > \tau$ are regarded as linked. By definition $C_{ij} \geq 0$, $\forall\{i,j\}$ (see Sect. 3.2.1). Using the Heaviside function $\Theta(.)$, the adjacency matrix A_{ij} of the climate network is then given by

$$A_{ij} = \Theta\left(C_{ij} - \tau\right). \tag{3.2}$$

Note that A_{ij} inherits its symmetry from C_{ij} and the resulting climate network is an undirected and unweighted simple graph.

3.2.3. Choosing the threshold

The last but nontrivial step in climate network construction is the selection of a threshold τ, above which we consider a pair of vertices to be connected. From a statistical point of view

1 The indices μ and ν label the bins of the histograms that we use to estimate the N PDFs $p_i(\mu)$ and the $N(N-1)/2$ joint PDFs $p_{ij}(\mu,\nu)$.

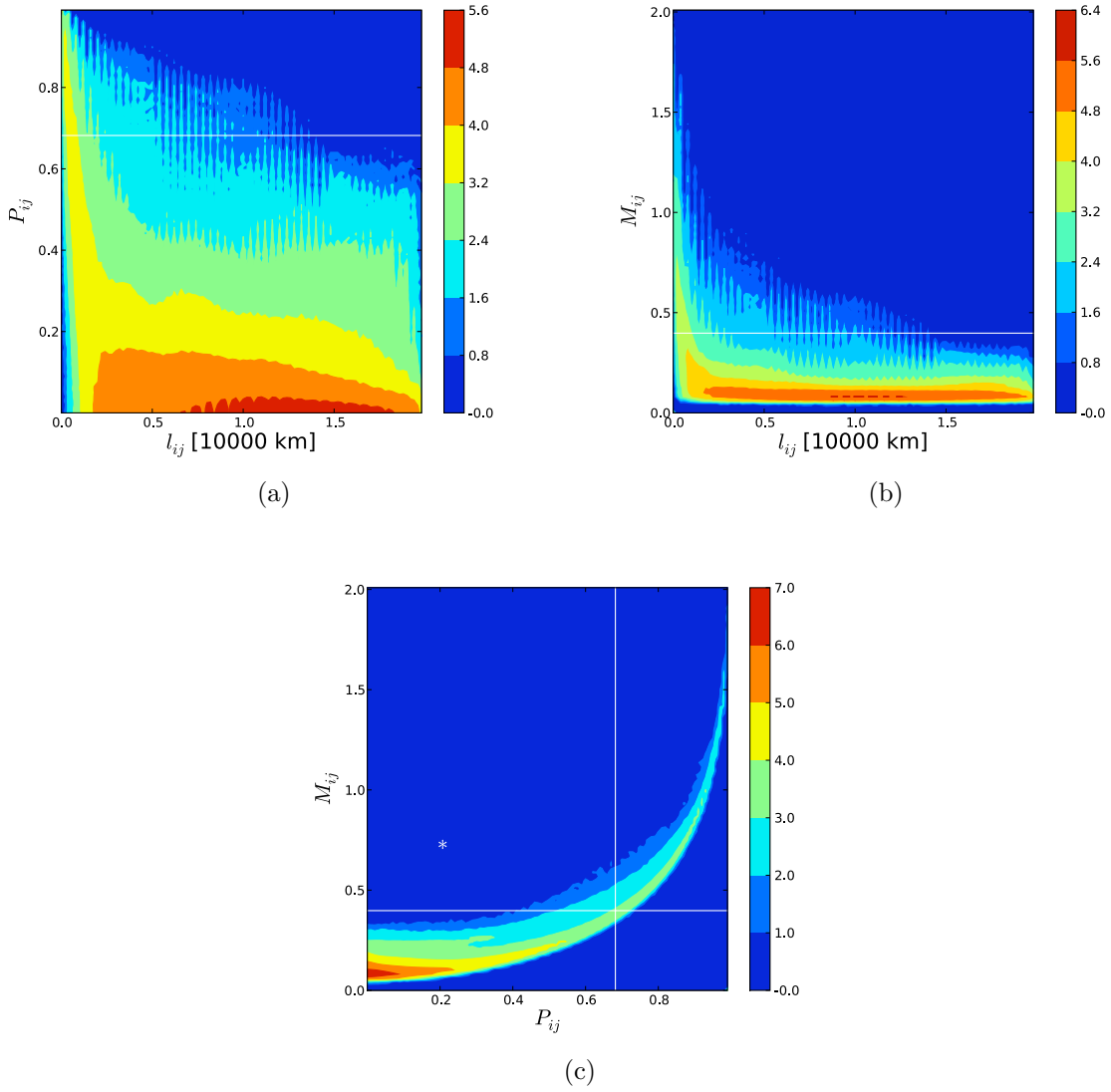


Figure 3.2 (a,b) Frequency plot in the space of correlation measure C_{ij} and edge distance l_{ij} for all $N(N-1)/2 = 23{,}228{,}928$ pairs of time series in the global HadCM3 SAT data set. The apparent oscillations with edge distance are an artifact of the finite spatial resolution of the underlying grid. (c) Frequency plot in the space of Pearson correlation P_{ij} and mutual information M_{ij}. All plots are based on 2D-histograms with 10^4 equally sized rectangular bins. The color bars indicate the common logarithm of frequency. Vertical and horizontal lines mark the thresholds corresponding to edge density $\rho = 0.005$ for P_{ij} and M_{ij} (Fig. 3.3). The asterisk in (c) delineates the quadrant containing edges that exist in the mutual information, but not in the Pearson correlation network of $\rho = 0.005$, and hence are candidates for strongly nonlinear connections.

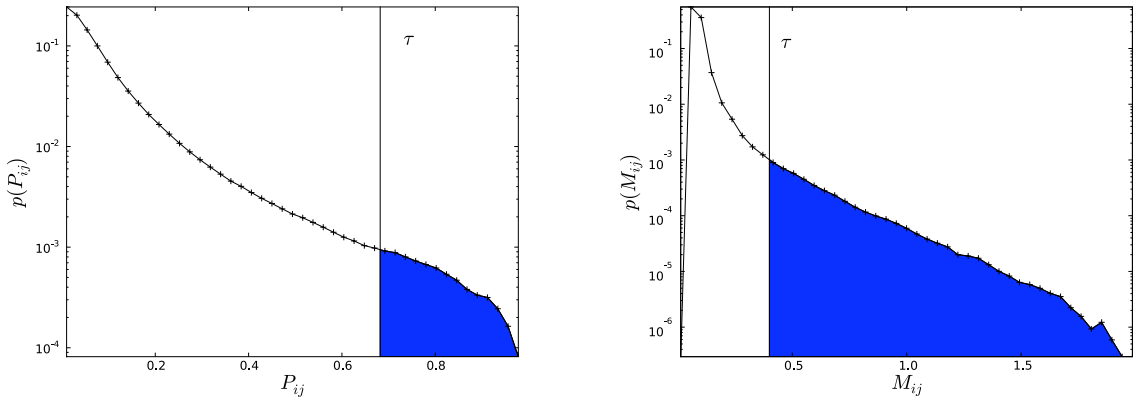
it is desirable to only maintain connections that are statistically significant with respect to some reasonable test and reject those not meeting this criterion. Classical significance tests and randomization experiments have been used to assess the value of τ for climate networks constructed using the Pearson correlation coefficient (Tsonis and Roebber (2004), Tsonis et al. (2006), Tsonis and Swanson (2008)). We built on these results testing against randomly shuffled time series, Fourier surrogates and twin surrogates (Thiel et al. (2006)). Twin surrogates correspond to the null hypothesis of trajectories with random initial conditions on the attractor of the original time series and are found to give the strictest bounds on the significance of network connections detected using Pearson correlation and mutual information (Sect. 4.1.4).

3.2.3.1. On the role of teleconnections

From the perspective of complex network theory, we intend to uncover interesting structures in the topology of the climate network. Different features of the underlying correlation measure matrix C_{ij} will be revealed at different thresholds τ. Consequently, the choice of τ has to reflect a trade-off between the statistical significance of connections and the richness of network structures unveiled. For example, note the potentially interesting long distance edges with high Pearson correlation and mutual information at edge distance $l \gtrsim 15000$km in the global HadCM3 SAT data set (Fig. 3.2(a) and 3.2(b)). They will only be included in the climate network, if the threshold $\tau \lesssim 0.65$ for the Pearson correlation network, or $\tau \lesssim 0.3$ in the case of the mutual information network. Long distance edges with high correlation measure or teleconnections are responsible for all interesting and non-trivial features of climate networks, such as "small-world" behavior, super-nodes or betweenness structures. Without them serving as spatial short cuts in the network, only the locally connected underlying grid remains. Ergo the inclusion of teleconnections must be a necessary criterion in the choice of the threshold in order to obtain interesting results in climate network analysis.

3.2.3.2. Dependence of network measures on edge density

Systematic studies show a smooth dependence of most climate network measures on τ in the range of edge densities considered in this work. This implies that small uncertainties in the choice of the threshold will not lead to strongly deviating results within the complex network framework. Here we discuss the threshold dependence of edge density $\rho(\tau)$, and the edge density dependence of clustering coefficient $\mathcal{C}(\rho)$, average path length $\mathcal{L}(\rho)$, number of components $n_c(\rho)$, relative giant component size $S(\rho)$ and average relative non-giant component size $\langle s(\rho) \rangle$ (Fig. 3.4). Here a component constitutes a maximally connected subset of vertices of the network, *i.e.*, a connected subset of vertices that is not reachable from any other vertex in the network. The term giant component is usually reserved for the largest component containing nearly all of the vertices in the network (Newman (2003)).



(a) (b)

Figure 3.3 PDF $p(C)$ of the correlation measure matrices C_{ij} for the HadCM3 SAT data set. The vertical line indicates the threshold τ yielding an edge density $\rho(\tau) = 0.005$, that is equal to the shaded area. (a) Pearson correlation, $\tau = 0.682$, b) mutual information, $\tau = 0.398$.

$S(\rho)$ in turn always measures the relative size of the largest component, even if its size becomes comparable to that of other components.

The edge density $\rho(\tau)$ decays approximately exponentially due to the shape of the PDF of the absolute value of the correlation measure $p(C)$ (in the following we abbreviate C_{ij} by C),

$$\rho(\tau) = \int_{\tau}^{\infty} dC p(C). \tag{3.3}$$

Note that $\rho(\tau)$ is a monotonic decreasing function of τ. Correlation measure distributions found empirically from climate data generally have a connected support (Fig. 3.3), so that $\rho(\tau)$ is strictly monotonic decreasing and induces a one to one correspondence between threshold τ and edge density ρ (Fig. 3.4(a)).

The clustering coefficient $\mathcal{C}$ is found to stay approximately constant at intermediate values of ρ and decays to zero for small ρ (Fig. 3.4(b)), when the network decomposes into a larger number n_c (Fig. 3.4(d)) of smaller components (Fig. 3.4(e) and 3.4(f)). The average path length $\mathcal{L}$ decays approximately as a power law with growing ρ and has discontinuities at edge densities ρ_μ, where $\tau_\mu = \tau(\rho_\mu)$ equals the correlation measure C_{ij} of edges $\{i,j\}$ with a high edge betweenness centrality Newman (2003), *i.e.*, that lie on many shortest paths between pairs of vertices (Fig. 3.4(c)). When $\tau \geq \tau_\mu$, these shortest paths become considerably longer and components might decouple from the network's giant component. This effect leads to a decrease of $\mathcal{L}$ for small ρ since the network decomposes into smaller disconnected components (Fig. 3.4(f)) and path lengths are measured only within the components. The formation of a giant component encompassing nearly all vertices at $\overline{\rho} \approx 0.0012$, where the giant component size increases from $S \approx 0.5$ to $S \approx 1$ (Fig. 3.4(e)), goes along with discontinuities of $\mathcal{L}$ and $\langle s(\rho) \rangle$. Note that all vertices have joined the giant component at $\rho \approx 0.020$ for the HadCM3

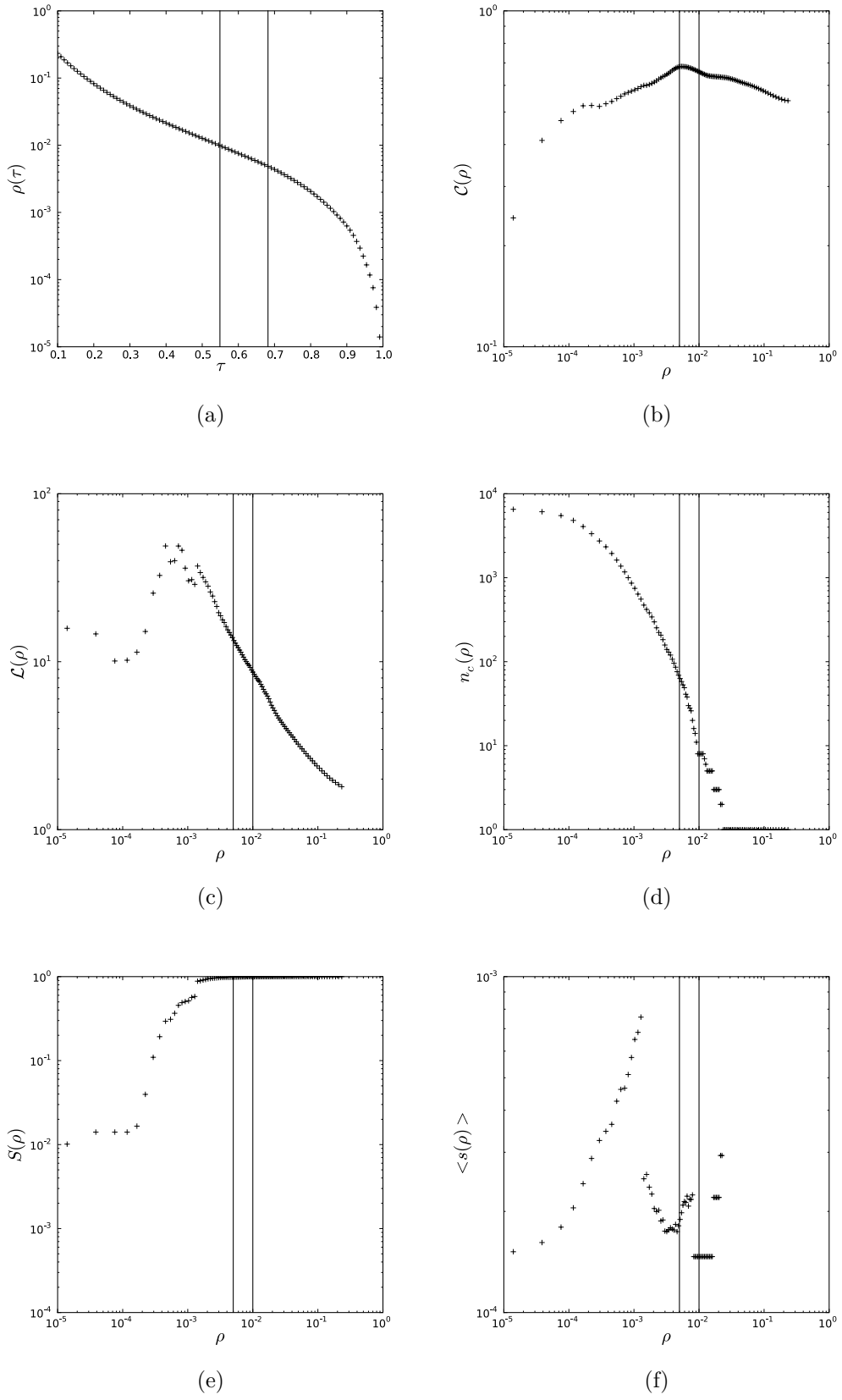


Figure 3.4 Network measures as a function of threshold and edge density for global HadCM3 SAT networks constructed using Pearson correlation. (a) Threshold dependence of edge density $\rho(\tau)$, (b) edge density dependence of clustering coefficient $\mathcal{C}(\rho)$ and (c) average path length $\mathcal{L}(\rho)$. (d) Edge density dependence of the number of components $n_c(\rho)$, (e) giant component size $S(\rho)$ and (f) average non-giant component size $\langle s(\rho) \rangle$. The vertical lines indicate edge densities of $\rho = 0.005$ and $\rho = 0.01$ and corresponding thresholds.

SAT Pearson correlation network (Fig. 3.4(d)) and at $\rho \approx 0.028$ for the corresponding mutual information network (not shown here).

At all edge densities considered in Sect. 3.3 the giant component size is of $\mathcal{O}(1)$. The influence of the non-giant components on measures such as average path length and closeness centrality is therefore negligible in the regime studied here, since larger deviations are only expected for $\rho < \overline{\rho}$. This range of edge densities in turn is not relevant for the conclusions drawn from the comparison presented in Sect. 3.3. To study this regime of very small edge densities in detail, measures more robust to disconnected components such as the local efficiency (related to closeness centrality) and global efficiency (related to average path length) should be considered (Newman (2003)). We chose the definitions given in Chap. 2 to maintain consistency with the existing literature on climate networks.

3.2.3.3. Pragmatic choice of τ

We think that the problem of selecting exactly the right threshold is not as severe as might be thought. Climate network analysis deals with topological properties of correlation measure matrices and aims at gaining new insights heeding this paradigm. In the climate system, it is furthermore not immediately evident which physical entities should take the role of vertices and edges in a complex network. This constitutes the main conceptional difference between our method and attempts of recovering an unknown physically existent network structure from vertex dynamics as in the study of the brain (Zhou et al. (2007), Schmidt et al. (2008), Rabinovich et al. (2006), Zhou et al. (2006), beim Graben et al. (2008)), where one can argue that a more natural identification of neurons and axons with the vertices and edges of a neural network exists. It is known that in the classical local description of geophysical fluid dynamics of atmosphere and oceans, *i.e.*, the Navier-Stokes equations combined with thermodynamic equations, the network of physical interaction has the structure of a regular grid (Vallis (2006)). In a discretized model, the dynamics at each grid point is only coupled to the grid points in the immediate neighborhood. The complex topology observed in climate networks should therefore be treated as a manifestation of structure formation, that allows for uncertainties in the choice of parameters such as τ.

In the spirit of the ideas elaborated in the above paragraphs, we choose to fix the edge density ρ when comparing the properties of climate networks generated using different correlation measures. This will result in different thresholds τ, because the empirical correlation measure distribution $p(C)$ clearly differs between linear Pearson correlation and nonlinear mutual information (Fig. 3.3). The selection of ρ is in each case guided by the principle of balancing between structural richness and statistical significance outlined above.

3.3. Comparison of Pearson correlation and mutual information climate networks

After having introduced our methodology for climate network construction, we proceed to the main aim of this study: A comparison of networks generated using the linear Pearson correlation coefficient and the nonlinear mutual information on local, mesoscopic and global topological scales. The edge density ρ is varied between $\rho_{min} = 0$ and $\rho_{max} = 0.1$ in equally sized steps. Recall, that small edge densities correspond to high thresholds (Sec. 3.2.3). For increasing edge density, edges with decreasing correlation measure are added to the network. Consequently, climate networks with a very high edge density $\rho \geq 0.1$ are not expected to contain meaningful information for climate data analysis, because they contain many connections that are not statistically significant, *i.e.*, that are much more likely to arise by chance. For example, Tsonis et al. use the Pearson correlation coefficient and a threshold of $\tau = 0.5$ in all of their works (Tsonis and Roebber (2004), Tsonis et al. (2006), Tsonis et al. (2008b), Tsonis and Swanson (2008)), which corresponds to an edge density of $\rho \approx 0.01$ for the global HadCM3 SAT data set analyzed here. They report that according to the Student's t test, a value of $P_{ij} = 0.5$ is statistically significant above the 99% level. In our recent work, we use an edge density of $\rho = 0.005$ (Donges et al. (2008)). This larger threshold corresponds to an even higher significance level, because it is less likely to be exceeded by the correlation measures calculated from pairs of one original and one surrogate time series.

We compare the properties of the complex networks obtained at each edge density level on local, mesoscopic and global topological scales. We enable a qualitative discussion of similarity by plotting the fields of area weighted connectivity (Fig. 3.5), local clustering coefficient (Fig. 3.6), closeness (Fig. 3.7) and betweenness centrality (Fig. 3.8) on a world map at fixed edge density $\rho = 0.005$. The local deviations of these fields calculated for Pearson correlation and mutual information climate networks are highlighted by normalized difference fields (Fig. 3.9). For a quantitative comparison at all edge densities considered, we calculated the Spearman rank order correlation coefficient or Spearman's Rho $r_s(\rho)$ of the corresponding fields taken from the Pearson correlation and mutual information networks (Fig. 3.10(d) and Fig. 3.11(d)). We chose to use the Spearman's Rho instead of the Pearson correlation coefficient for this task, because it is known to be more reliable when applied to data with non-Gaussian PDF. This is an important property, considering that some of the fields we are interested in have a highly non-normal frequency distribution (Sect. 3.3.1 and Sect. 3.3.3). Furthermore at each edge density step, we consider the Hamming distance between the networks on the local topological scale, whereas on the mesoscopic and global scale we compare global clustering coefficient and average path length.

In the following we will illustrate the comparison for the HadCM3 SAT data set in detail (Sect. 3.3.1, 3.3.2, 3.3.3 and Fig. 3.5, 3.6, 3.7, 3.8, 3.9, 3.10). Only the quantitative comparison is presented for the NCEP/NCAR reanalysis SAT data set (Fig. 3.11), since we are lead to the same conclusions as for the model data set. Finally we present climatological

interpretations of the observed network structures (Sect. 3.4).

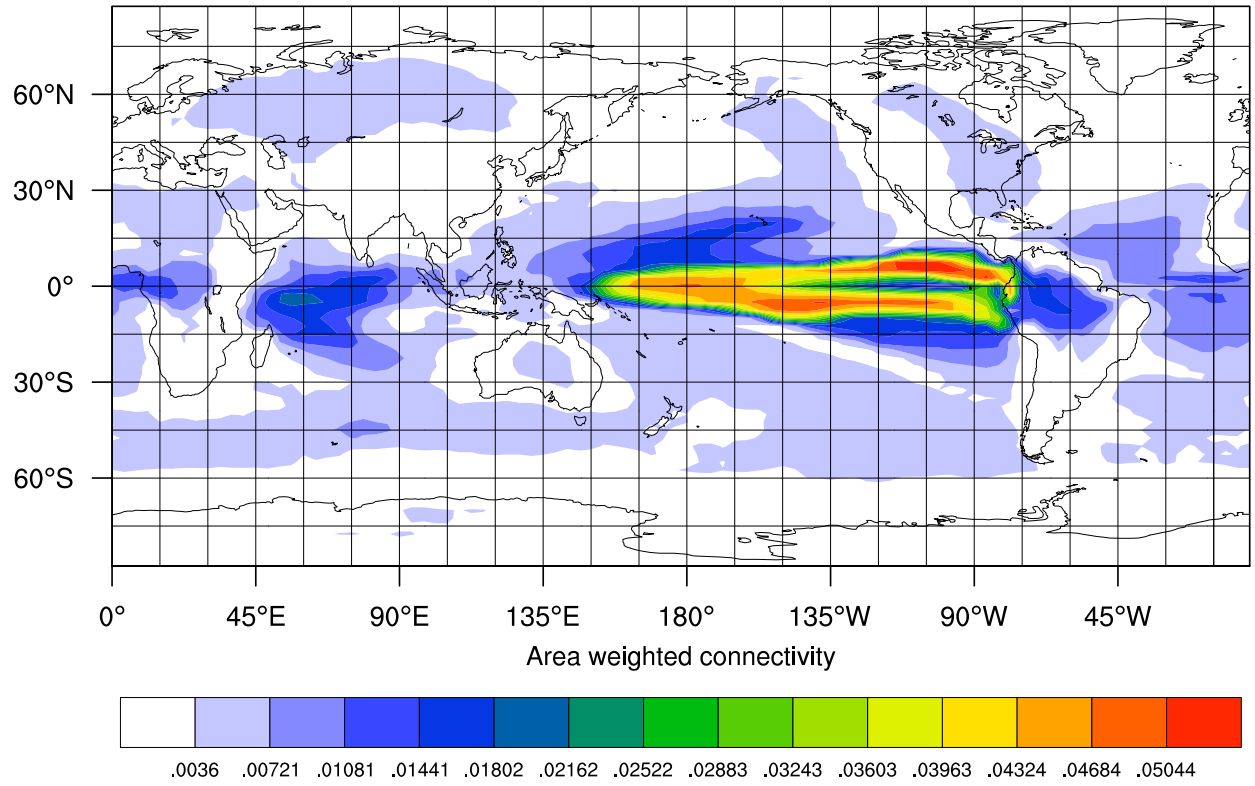
3.3.1. Local comparison

On the local topological scale, we find that Pearson correlation and mutual information climate networks are very similar at low edge densities. At $\rho = 0.005$, the area weighted connectivity (Fig. 3.5) field shows only small deviations by visual inspection, that are most pronounced in the tropics (Fig. 3.9(a)). The rank order correlation coefficient r_s^{AWC} reaches a maximum between $\rho = 0.005$ and $\rho = 0.01$ and decays for larger edge densities (Fig. 3.10(d)). We obtain high values for $\rho = 0.005$ and $\rho = 0.01$ (Table 3.2). Note that for the climate networks studied, area weighted connectivity has a fat tailed PDF (Tsonis et al. (2006)).

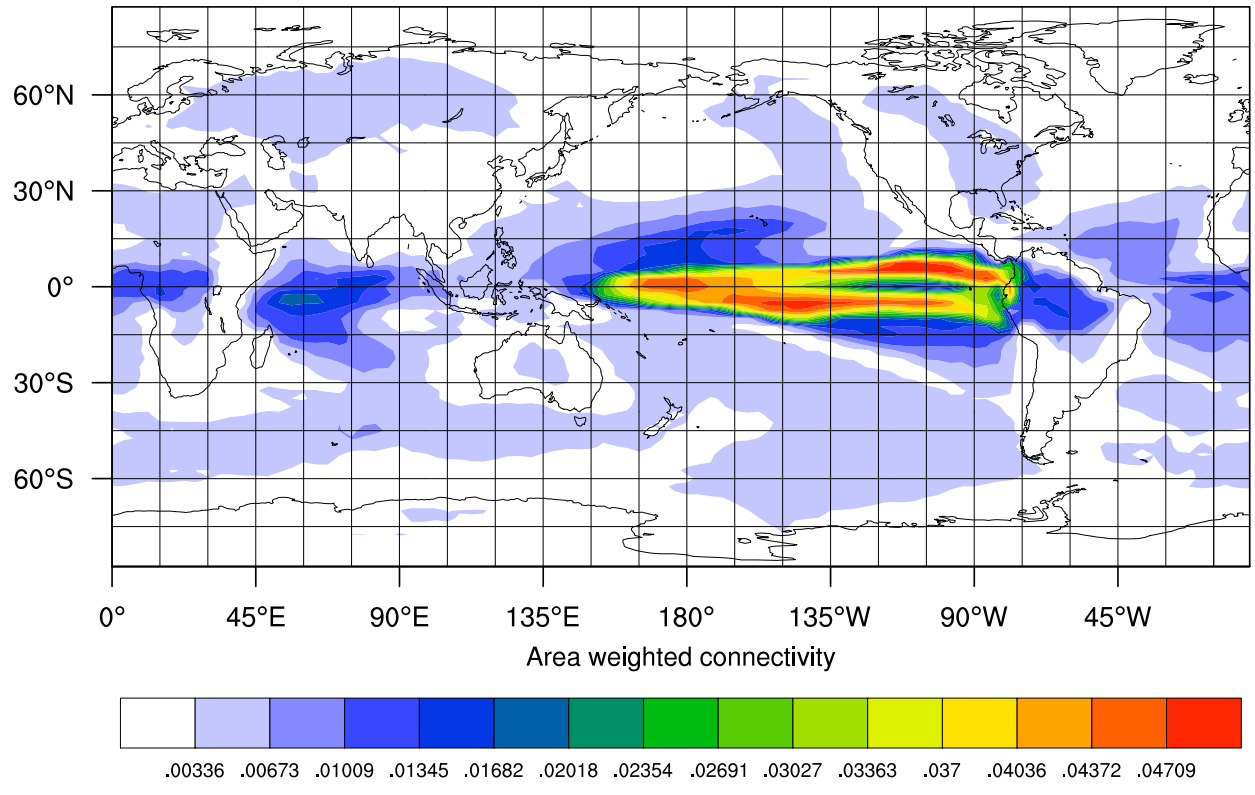
The Hamming distance $H(\rho)$ is always smaller than the expected distance $H^R(\rho)$ of two random networks at edge density ρ (Fig. 3.10(a)). It is notable, that $H(\rho)$ seems to go to zero tangentially to the ρ-axis, *i.e.*, $H'(\rho)|_{\rho=0} \approx 0$, whereas $H^R(\rho)|_{\rho=0} = 2$. Therefore most of the edges with the highest Pearson correlation and mutual information values must coincide. From analytical considerations and Monte-Carlo simulations we find that the standard deviation of the PDF of Hamming distance between the two random networks is of $\mathcal{O}(N^{-1})$ for $N \gg 1$. This means that the expected deviations from the mean $H^R(\rho)$ are of $\mathcal{O}(10^{-4})$ for the climate networks considered here. The difference between measured Hamming distance and $H^R(\rho)$ is by one order of magnitude larger than these expected deviations (Table 3.2). We hence conclude that the observed similarity of Pearson correlation and mutual information networks can be considered statistically significant, with respect to the null hypothesis of random networks of the same size N, at all edge densities considered. Particularly, the results elaborated in this section show, that at the edge densities used in earlier works on climate networks (Tsonis and Roebber (2004), Tsonis et al. (2006), Tsonis et al. (2008b), Tsonis and Swanson (2008), Donges et al. (2008)), Pearson correlation and mutual information give very similar results on the local topological scale.

Table 3.2 Spearman's Rho $r_s(\rho)$ of area weighted connectivity (AWC), local clustering coeffcient (C), closeness centrality (CC) and betweenness centrality (BC) fields and Hamming distances $H(\rho)$ and $H^R(\rho)$ calculated from Pearson correlation and mutual information networks at edge densities $\rho = 0.005$ and $\rho = 0.01$ for the global HadCM3 SAT data set.

	$\rho = 0.005$	$\rho = 0.01$
r_s^{AWC}	0.95	0.88
r_s^{C}	0.80	0.81
r_s^{CC}	0.98	0.95
r_s^{BC}	0.70	0.59
H	0.001	0.003
H^R	0.010	0.02



(a)



(b)

Figure 3.5 Area weighted connectivity fields for global HadCM3 SAT networks at $\rho = 0.005$ (linear color scale) obtained using a) Pearson correlation, b) mutual information. The rank order correlation between the two fields is $r_s^{AWC}(0.005) = 0.95$.

3.3.2. Mesoscopic comparison

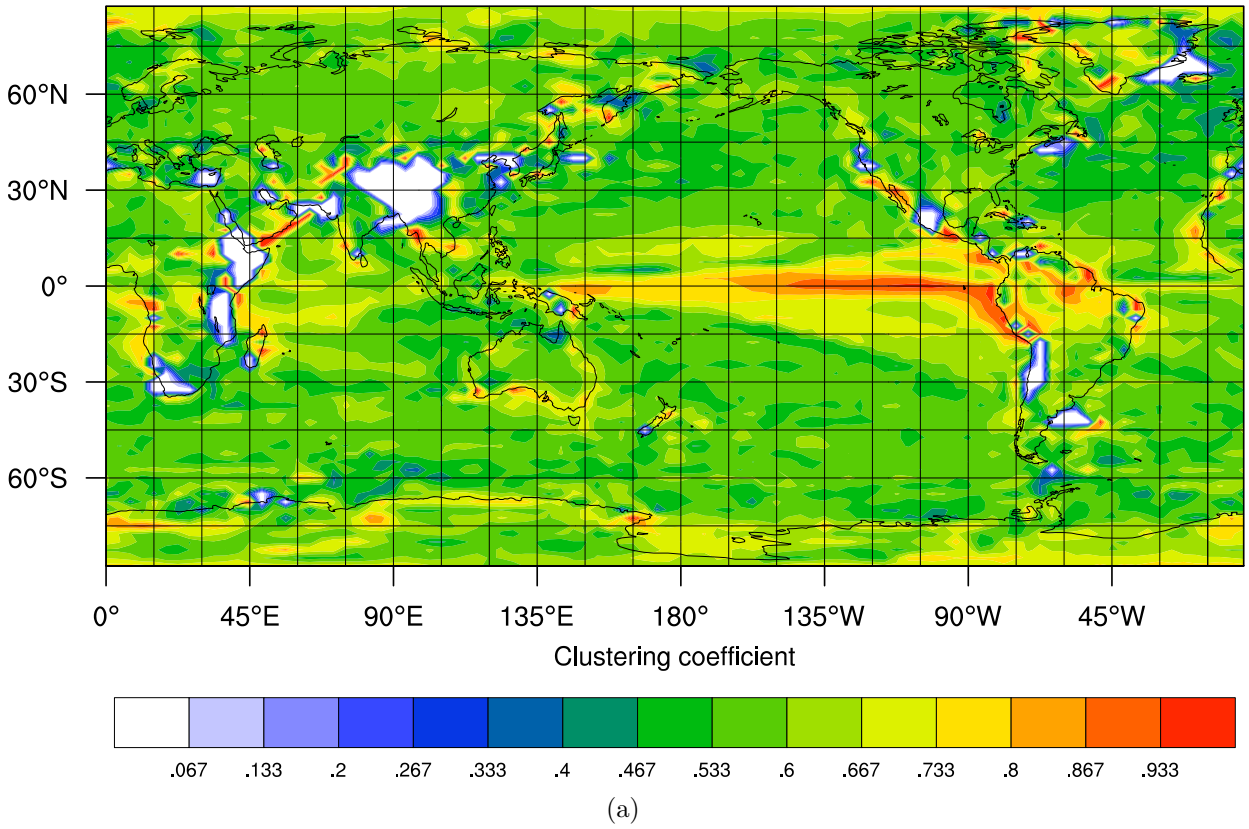
The local and global clustering coefficients also reveal a high degree of similarity on the mesoscopic topological scale. Analogous to AWC, the local clustering coefficient fields are nearly indistinguishable (Fig. 3.6). However, the largest deviations appear to cluster along coastlines (Fig. 3.9(b)). This interesting finding can be understood by considering the qualitatively different dynamics of SAT over oceans and continents, *e.g.*, the on average much larger seasonal variability over continents. Along coastlines, the correlation length of the SAT field is thus smaller than that expected over continents or the ocean away from the coast. Hence Pearson correlation and mutual information have a higher probability to disagree on the existence of edges between spatially adjacent vertices (local edges) along the coastline. These local and mesoscopic deviations in network structure are detected by the local correlation coefficient C_v, that is by design particularly sensitive on the mesoscopic topological scale (Sect. 3.4).

The rank order correlation coefficient reaches a maximum between $\rho = 0.005$ and $\rho = 0.01$ and decays for larger edge densities (Fig. 3.10(d)). We obtain high values for $\rho = 0.005$ and $\rho = 0.01$ (Table 3.2). The global clustering coefficients show only small deviations of $\mathcal{O}(10^{-2})$ at all edge densities considered (Fig. 3.10(b)). We get $\mathcal{C}^P(0.005) = 0.682$, $\mathcal{C}^M(0.005) = 0.678$ and $\mathcal{C}^P(0.01) = 0.657$, $\mathcal{C}^M(0.01) = 0.668$. The local clustering coefficient field is close to normally distributed.

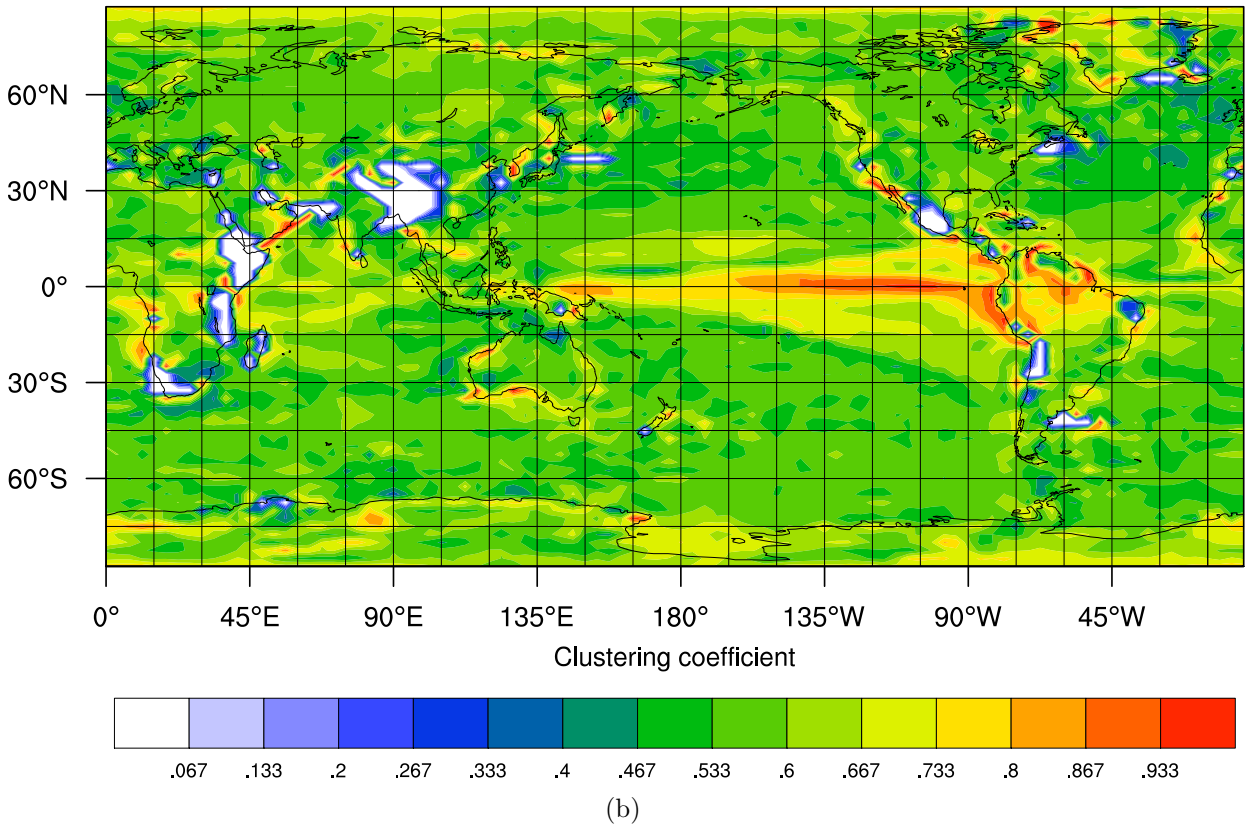
3.3.3. Global comparison

We observe more interesting behavior at the global topological scale. Closeness centrality at $\rho = 0.005$ does not deviate much qualitatively and quantitatively across the two types of networks considered (Fig. 3.7), the largest differences are detected in the tropics with a tendency to decrease with latitude towards the poles, and most notably over South America (Fig. 3.9(c)). The betweenness centrality field shows more pronounced qualitative regional differences (Fig. 3.8). For example, note the differing high betweenness structures over the oceans, particularly over the East Pacific, the North Atlantic and arctic regions (Fig. 3.9(d)). The rank order correlation coefficients r_s^{CC} and r_s^{BC} decay more quickly than the ones on the local and mesoscopic topological scale and fluctuate around values of $r_s^{CC} \approx 0.1$ and $r_s^{BC} \approx 0.4$ for larger edge densities (Fig. 3.10(d)). At $\rho = 0.005$ and $\rho = 0.01$, r_s^{BC} is notably smaller than the Spearman's Rho of the other fields considered, while r_s^{CC} is close to unity (Table 3.2). Confirming earlier studies, we find that betweenness follows a fat tailed PDF (Goh et al. (2002)), whereas the closeness field is normally distributed.

These results indicate, that betweenness centrality may quantify the local differences between networks constructed using Pearson correlation and mutual information at the global topological scale, that could be traces of nonlinear physical processes in the climate system. That the greatest deviations are found between the betweenness centrality fields is plausible, because betweenness is by definition a very sensitive measure and can locally



(a)

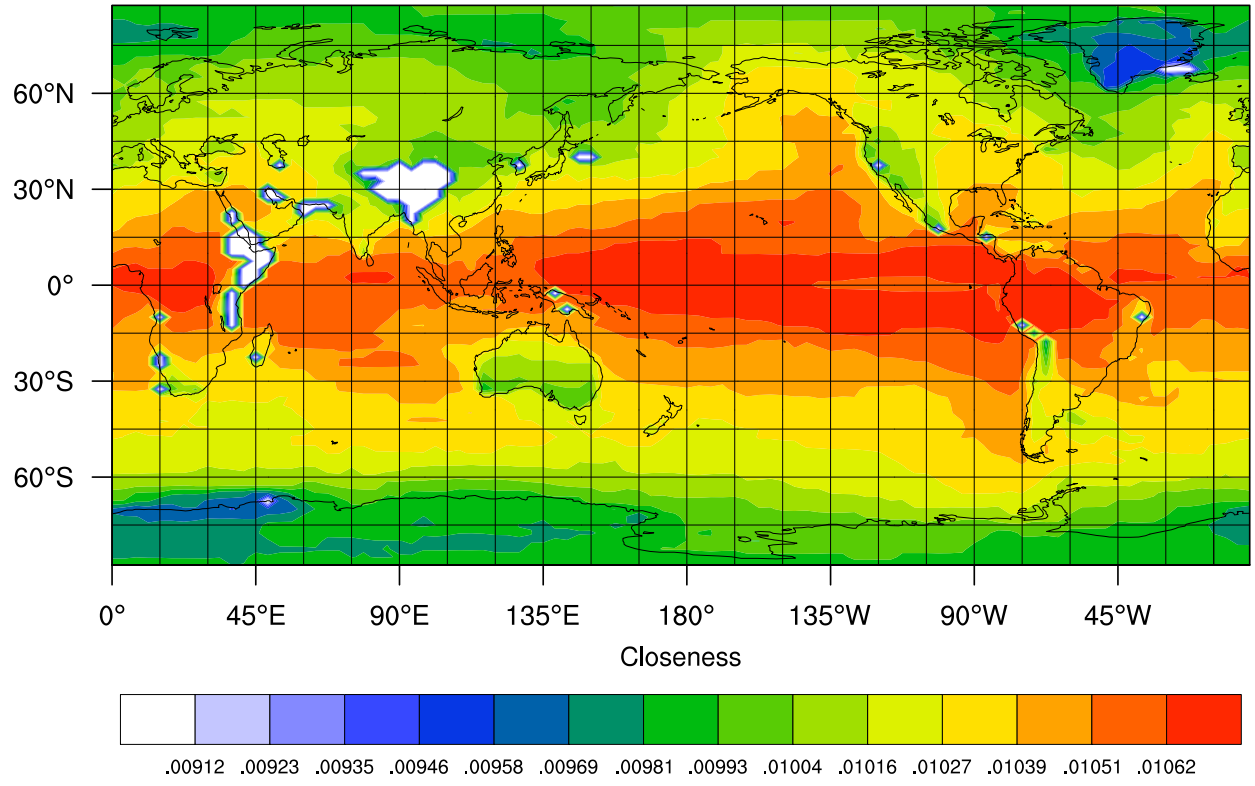


(b)

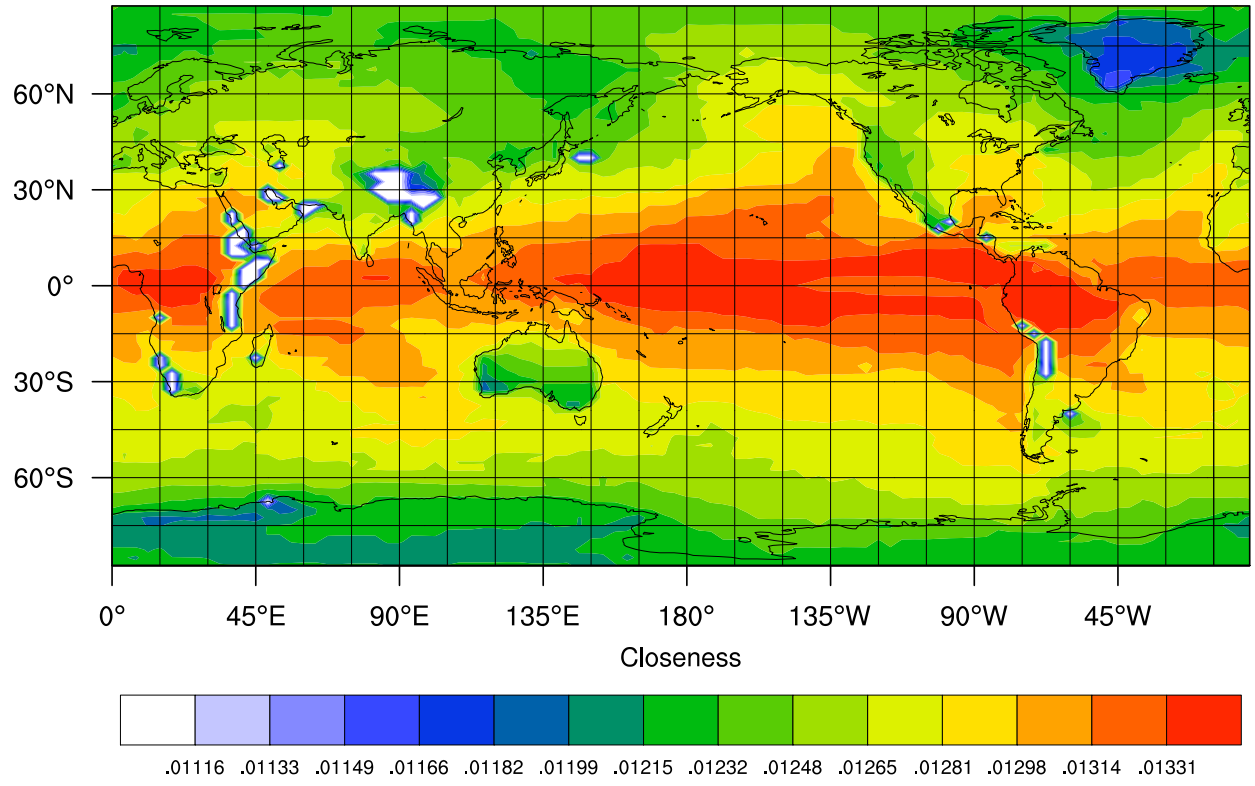
Figure 3.6 Local Watts-Strogatz clustering coefficient fields for global HadCM3 SAT networks at $\rho = 0.005$ (linear color scale) obtained using a) Pearson correlation, b) mutual information. The rank order correlation between the two fields is $r_s^{\mathcal{C}}(0.005) = 0.81$.

depend heavily on the existence or non-existence of a small number of edges in the network (Albert et al. (2004)). Consider for example a small set of edges, that are the only connections between two large communities in a network. The vertices on either end of these edges have a high betweenness centrality, because all shortest paths between the two communities must contain them. If the bridging edges are removed, the betweenness centrality of the beachhead vertices must decrease significantly, since they can now only participate in shortest paths within their own community. This sensitivity of betweenness leads to a large dynamic range of 20 orders of magnitude for the global HadCM3 SAT network, that calls for a logarithmic scale to properly visualize the betweenness distribution (Fig. 3.8).

The average path length (Fig. 3.10(c)) agrees closely, with deviations of $\mathcal{O}(10^{-1})$. We obtain $\mathcal{L}^P(0.005) = 13.4$, $\mathcal{L}^M(0.005) = 13.5$ and $\mathcal{L}^P(0.01) = 8.5$, $\mathcal{L}^M(0.01) = 8.5$.

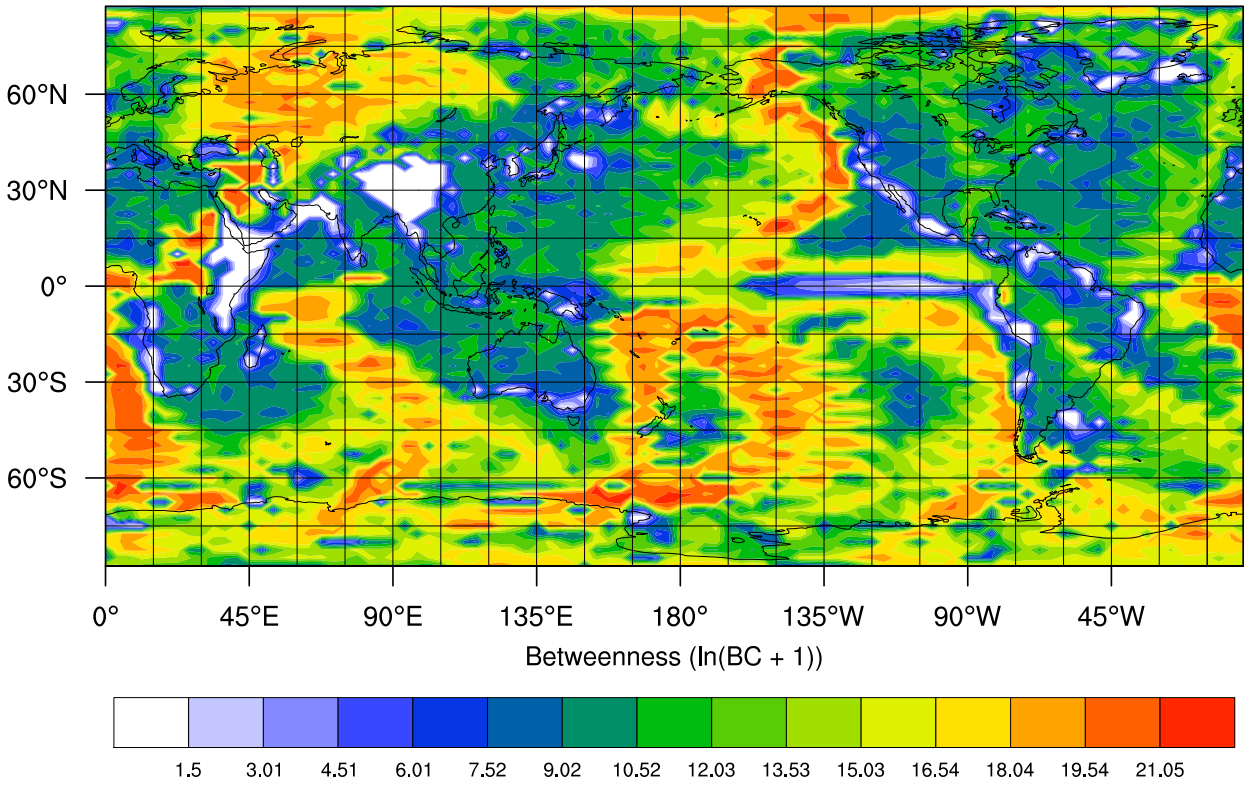


(a)

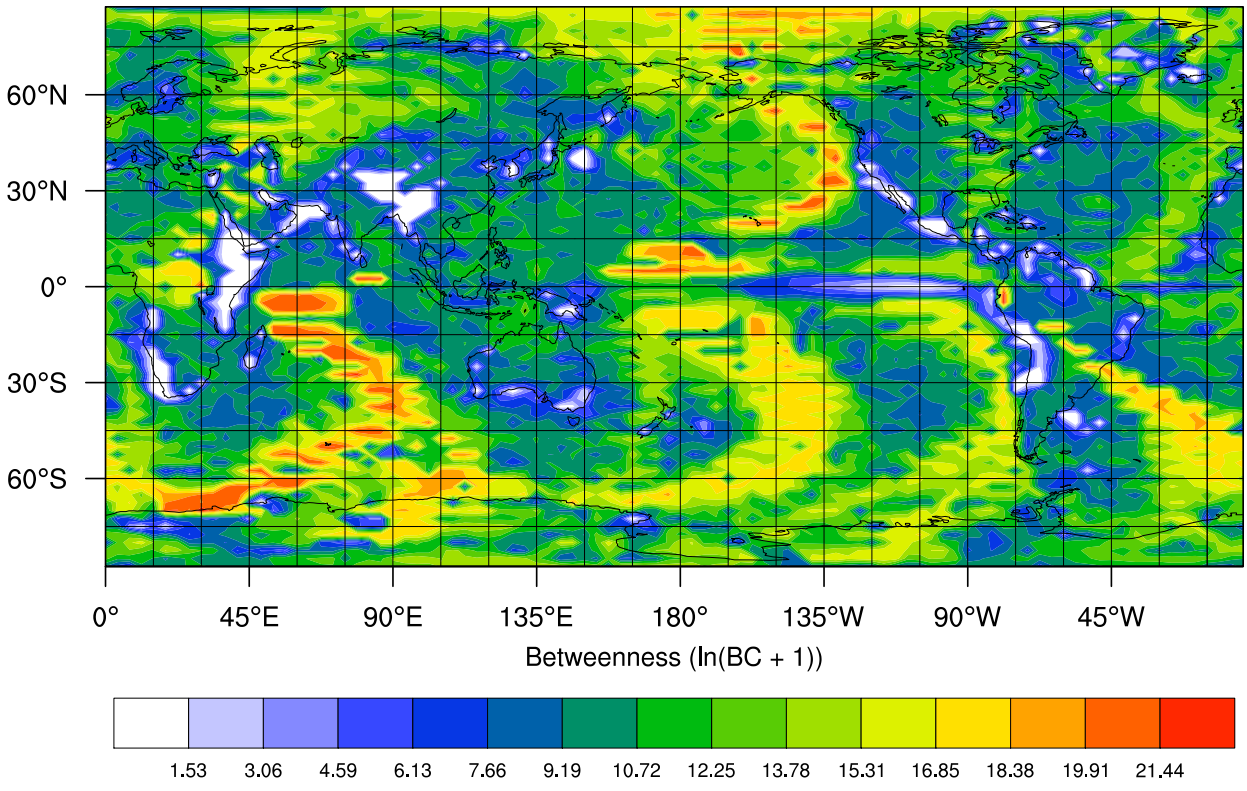


(b)

Figure 3.7 Closeness centrality field for global HadCM3 SAT networks at $\rho = 0.005$ (linear color scale) obtained using a) Pearson correlation, b) mutual information. The rank order correlation between the two fields is $r_s^{CC}(0.005) = 0.98$. The white regions on the map correspond to vertices that are disconnected from the network's giant component.



(a)



(b)

Figure 3.8 Betweenness centrality fields for global HadCM3 SAT networks at $\rho = 0.005$ (logarithmic color scale) obtained using a) Pearson correlation, b) mutual information. The rank order correlation between the two fields is $r_s^{BC}(0.005) = 0.70$.

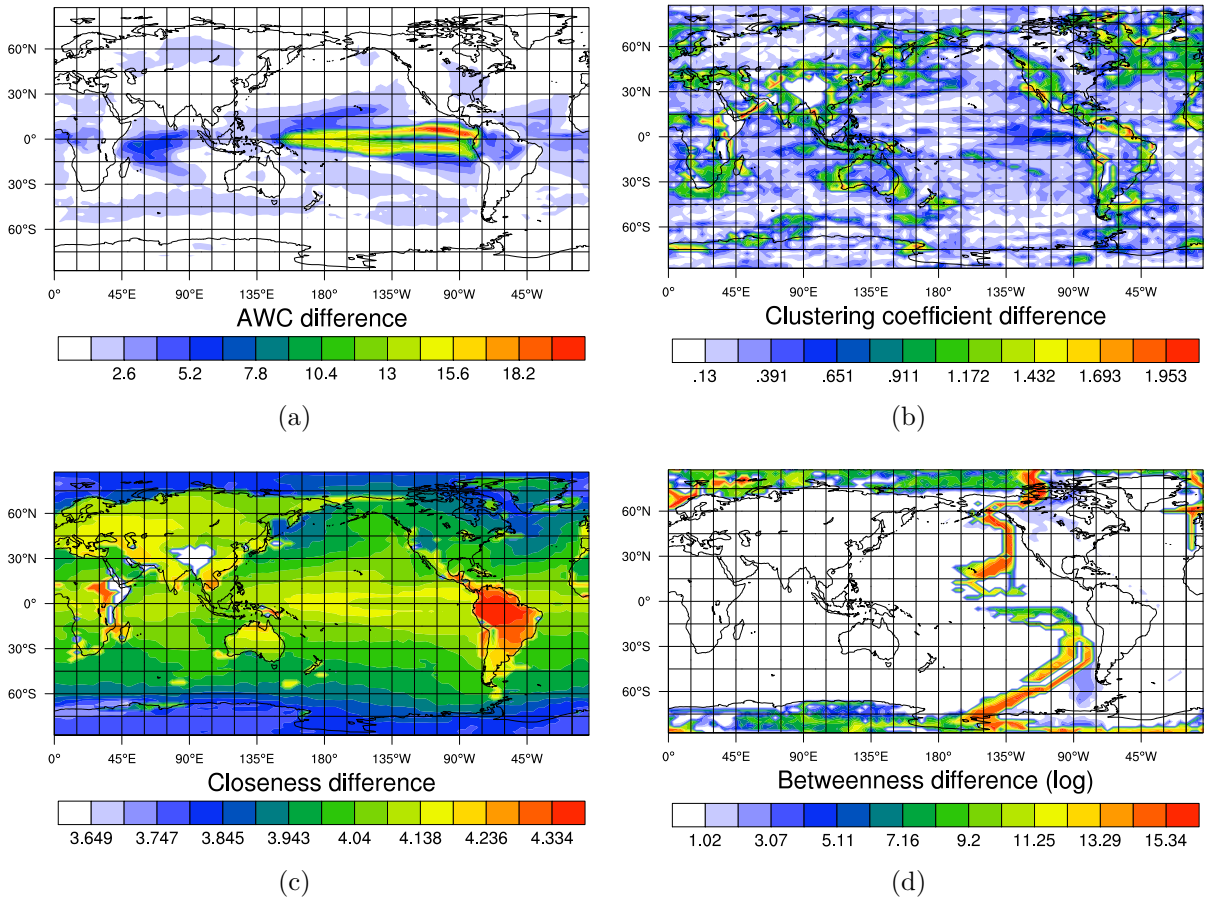


Figure 3.9 Normalized difference fields $\Delta g_v = |g_v^P - g_v^M| / \sqrt{\langle g_w^P \rangle_w \langle g_w^M \rangle_w}$ of network measure fields g_v^P and g_v^M, calculated from Pearson correlation and mutual information HadCM3 SAT climate networks at $\rho = 0.005$. (a) Area weighted connectivity, (b) local clustering coefficient, (c) closeness and (d) betweenness.

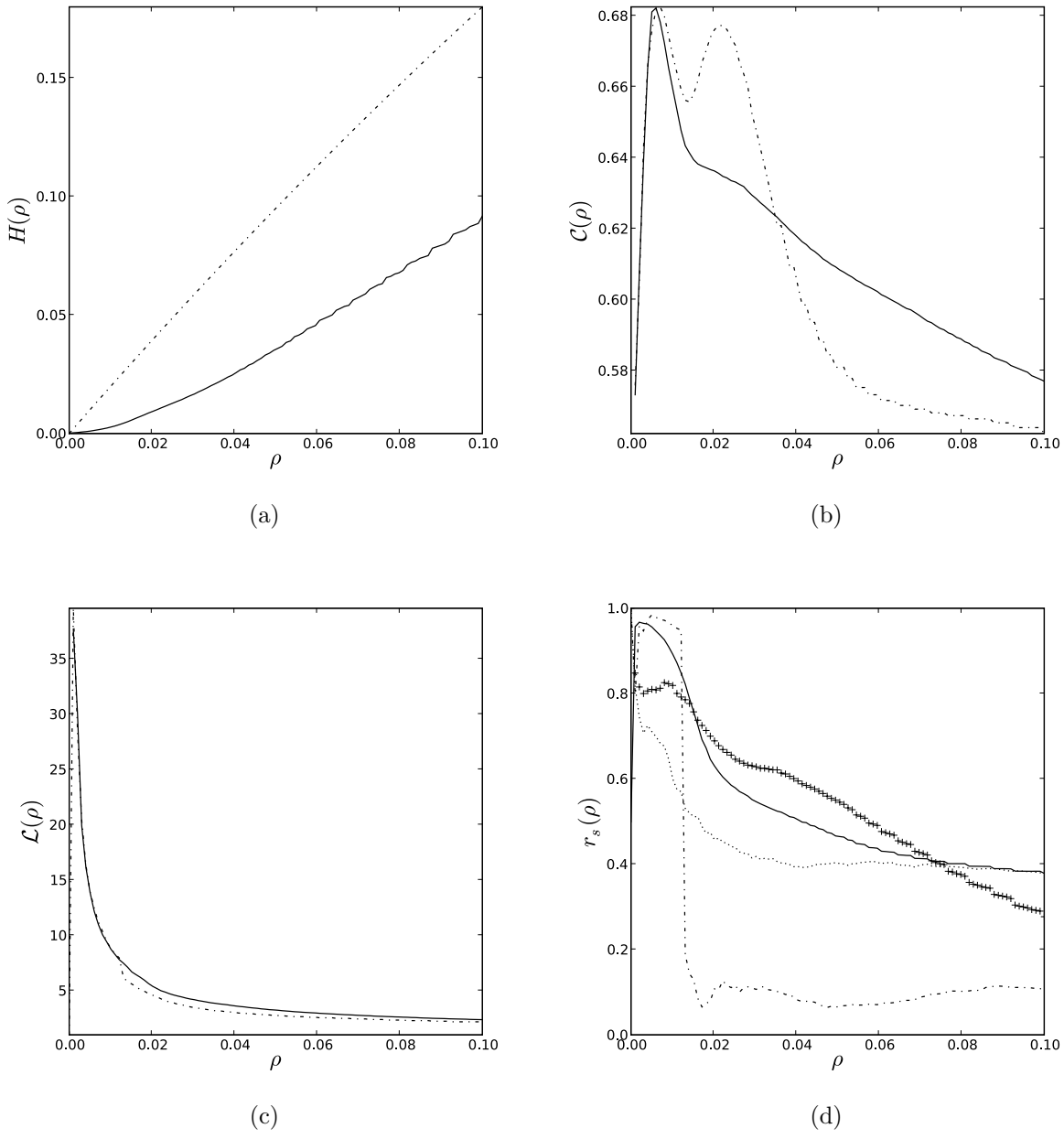


Figure 3.10 Results for the quantitative comparison of Pearson correlation and mutual information climate networks for the global HadCM3 SAT data set, shown as a function of edge density ρ (100 edge density steps). (a) Hamming distance $H(\rho)$ (continuous line) and expected random Hamming distance $H^R(\rho)$ (dashed line) between the two networks. The expected deviations from $H^R(\rho)$ are of $\mathcal{O}(10^{-4})$ (Sect. 3.3.1). (b) Global clustering coefficient $\mathcal{C}^P(\rho)$ of the Pearson correlation (continuous line) and $\mathcal{C}^M(\rho)$ of the mutual information network (dashed line). (c) Average path length $\mathcal{L}^P(\rho)$ of the Pearson correlation (continuous line) and $\mathcal{L}^M(\rho)$ of the mutual information network (dashed line). (d) Spearman rank order correlation coefficients $r_s^{AWC}(\rho)$ for the area weighted connectivity (continuous line), $r_s^{\mathcal{C}}(\rho)$ for the local clustering coefficient (crosses), $r_s^{CC}(\rho)$ for the closeness centrality (dash-dotted line) and $r_s^{BC}(\rho)$ for the betweenness centrality fields (dotted line).

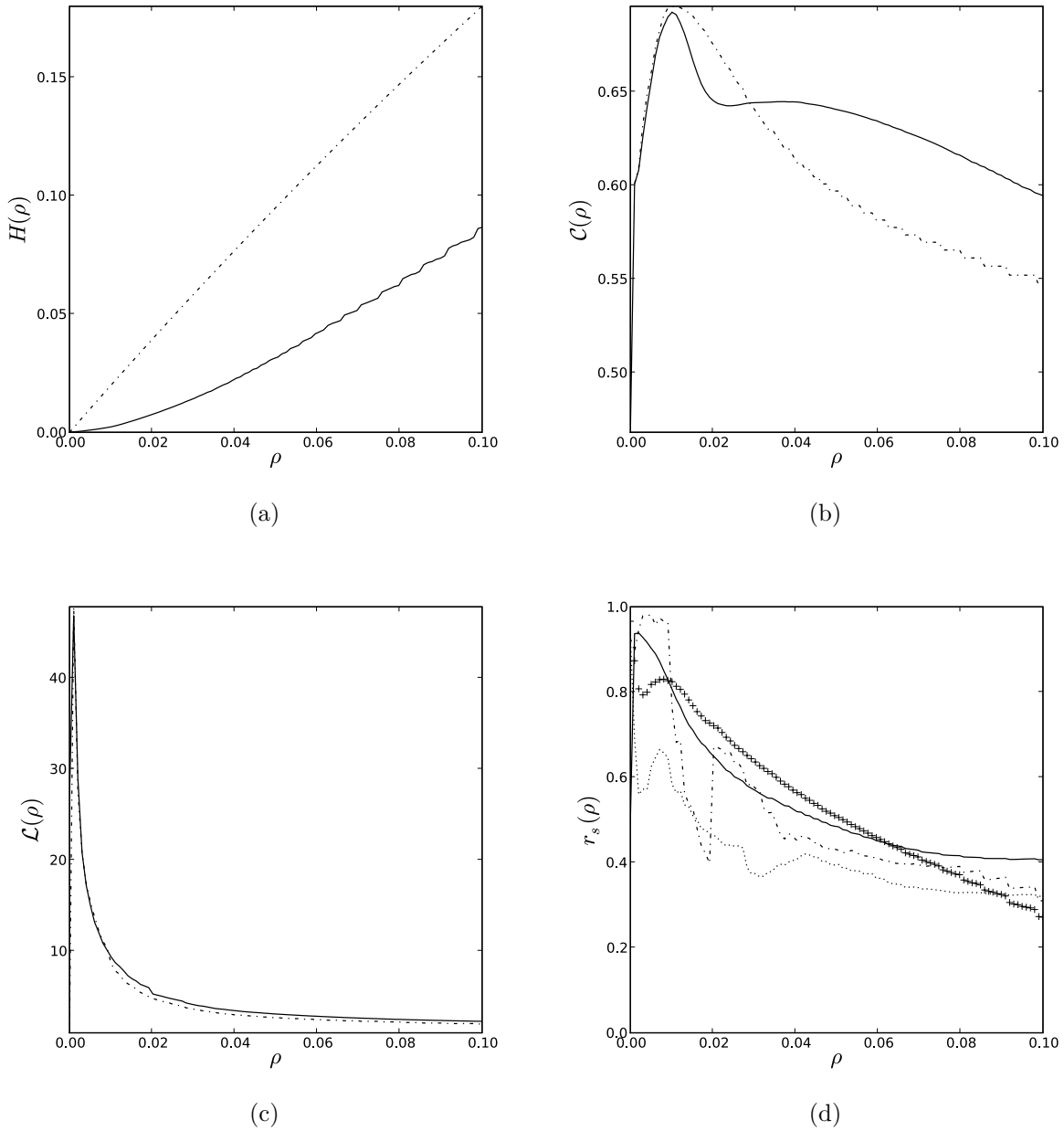


Figure 3.11 This figure shows the same statistics as Fig. 3.10, but evaluated for the global NCEP/NCAR reanalysis SAT data set.

3.4. Climatological interpretation

We give brief climatological interpretations of the network properties unveiled by our approach, since the main aim of this study is the comparison of linear and nonlinear climate network construction methods (Sect. 3.2). Super-nodes found in the AWC field (Fig. 3.5) over the tropics and locally the mid-latitudes, were shown to be related to major atmospheric teleconnection patterns (Tsonis et al. (2008b)). For example, the region of increased AWC in the North East Pacific is associated to the well-known Pacific North-American (PNA) pattern (Wallace and Gutzler (1981)). The El Niño cold tongue in the tropical East Pacific is clearly visible in the AWC field, as well as in all other fields considered (Fig. 3.6, 3.7 and 3.8).

The local clustering coefficient is found to be of $\mathcal{O}(1)$ in a connected region in the equatorial Pacific as well as locally along continental coastlines (Fig. 3.6). The former indicates a high degree of dynamical similarity in the tropical Pacific (Tsonis et al. (2006), Tsonis et al. (2008b)), that is possibly related to ENSO. The latter are more likely to be a signature of our climate network construction method along the coastline and visible on the mesoscopic scale only, that we discuss in Sect. 3.3.2.

The contouring of the closeness field (Fig. 3.7) nicely shows the latitudinally growing influence of the Coriolis force. Pressure gradient forces are balanced by the Coriolis force in the mid-latitudes for large scale atmospheric flows. This balance vanishes in the tropics, because the Coriolis force decays as $\sin(\lambda)$ when latitude λ approaches the equator. The closeness field also shows that the tropics form the center of SAT climate network, the associated vertices being topologically closer to the rest of the network than vertices in the mid-latitudes and arctic regions. This finding can be explained by considering the comparably regular dynamics of the tropical SAT field leading to many edges between tropical vertices, and the more irregular dynamics in the mid-latitudes and arctic regions that results in fewer edges within the mid-latitudes and arctic as well as between these regions and the tropics (von Bloh et al. (2005)). In a global climate network, it is hence more probable to find shorter shortest paths starting from tropical vertices, while shortest paths originating in mid-latitude and arctic vertices are on average longer. Moreover, we point out the lower closeness over Australia and Greenland indicating that these land-masses also form pronounced clusters in the SAT climate network, even though the local clustering coefficient field shows that they are not as highly locally interconnected as the equatorial Pacific. These differences in local connectedness among the detected dynamical clusters are caused by the qualitatively different dynamics over land and oceans (Sect. 3.3.2). The land-sea difference is globally detected by closeness centrality and AWC: Vertices over land masses are found to be on average less well connected and topologically more remote than those over the oceans.

We observe highly localized linear structures in the betweenness field (Fig. 3.8), some of which appear to resemble major surface ocean currents such as the California and Peru currents following the western coastline of the Americas, or the East Greenland, Norwegian

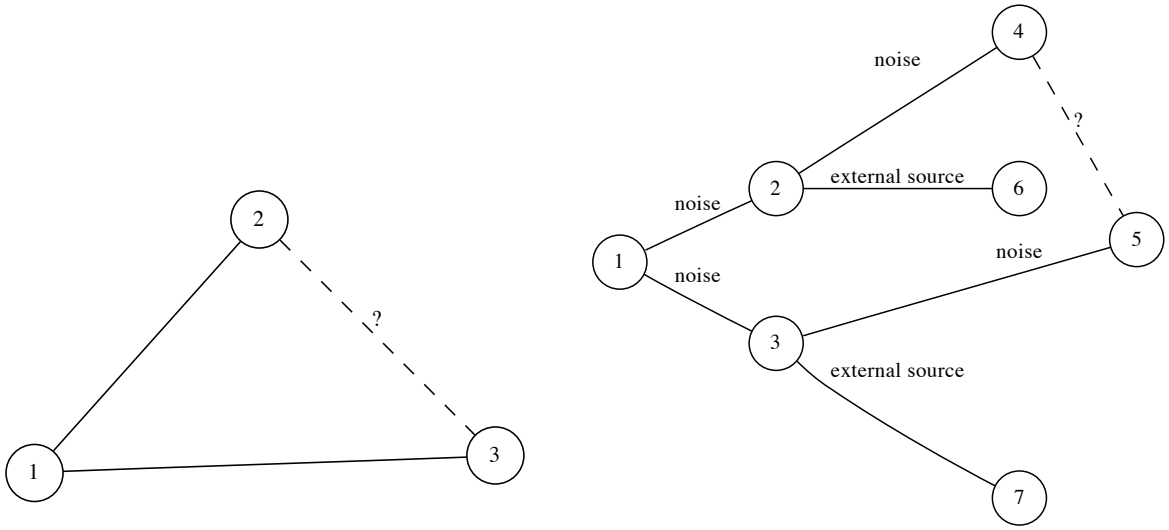
and Canary currents. Note that some of these current resembling structures are particularly visible in the betweenness difference field (Sect. 3.3.3), indicating that nonlinear processes might be involved in the formation of some of the structures. In analogy to the major communication channels of the internet, we refer to these betweenness structures as the backbone of the climate network, because a large fraction of the dynamical information exchanged via topologically shortest paths between all possible pairs of vertices $\{i,j\}$ must pass the high betweenness regions. This is particularly true for information transported by advective processes, where the assumption of information traveling on shortest paths can be substantiated by extremalization principles. In our recent work we report the discovery of the backbone and its possible role in stabilizing the climate system (Donges et al. (2008) and Chap. 5).

Note that the region very close to the equator in the tropical East Pacific has a comparatively low AWC, closeness and betweenness, but a high local clustering coefficient. This indicates that this region forms a internally densely connected cluster in a network sense, *i.e.*, it is dynamically highly interrelated but nearly detached from the rest of the network. We interpret it as a pronounced manifestation of the equatorial Coriolis barrier (Vallis (2006)), that can also be observed weakly over the equatorial Indian and Atlantic Oceans.

In agreement with (Tsonis and Roebber (2004), Tsonis et al. (2006), Tsonis et al. (2008b)) we find that Pearson correlation and mutual information climate networks possess properties of "small-world" networks (Watts and Strogatz (1998), Milgram (1967)), *i.e.*, a small average path length $\mathcal{L} \ll N$ and a large clustering coefficient of $\mathcal{O}(1)$ (Table 3.2, Fig. 3.10 and 3.11). Complex "small-world" networks with comparable global properties are frequently found in nature, *e.g.*, the internet, power grids, social and neural networks, and constitute the subject of study of an equally diverse collection of sciences. The small average path length can be explained by the influence of teleconnections. This indicates that perturbations of the regional dynamics (vertex dynamics) can on average quickly affect the whole globe via paths consisting of statistically highly interrelated pairs of regions (edges). It has been argued that this serves to stabilize the climate system and to enhance the information transfer within it (Tsonis and Roebber (2004), Tsonis et al. (2006), Tsonis et al. (2008b), Tsonis and Swanson (2008)). If the climate network was only locally connected, in other words if all teleconnections were removed from it, the average path length would be of $\mathcal{O}(N)$ as that of a regular grid. The high clustering coefficient is due to the spatial continuity of the underlying physical fields (*e.g.*, SAT), that leads to a prevalence of local triangles (Tsonis et al. (2008a)).

3.5. The transitivity problem

The transitivity problem constitutes one potentially serious conceptual flaw of the climate network construction methodology presented earlier in this chapter. Envision a subgraph of order 3 depicting the topology of physical interactions within some climatological field between three spatially well separated regions on the Earth's surface, that are represented by vertices



(a) (b)

Figure 3.12 Illustration of the transitivity problem. Vertices represent spatially separated regions on the Earth's surface. Straight lines indicate direct physical interactions between regions within some climatological field, *e.g.*, Rossby waves, prevailing winds or surface ocean currents. Dashed lines depict false edges in a climate network not corresponding to any direct physical interaction, that arise due to the transitivity of correlation measures such as Pearson correlation and mutual information. (a) Simplest manifestation of the transitivity problem (first order) in a subgraph of three regions. (b) A higher order version of the transitivity problem setting forth the influence of noise and of dynamical sources external to the "horseshoe" $4-2-1-3-5$.

(Fig. 3.12(a)). While region 1 can exchange dynamical information directly with regions 2 and 3 via Rossby waves, prevailing winds, surface ocean currents or other physical processes, regions 2 and 3 may be separated by a grand mountain chain inhibiting direct interaction. Nevertheless, because of the influence of region 1, a strong statistical interrelationship may be detected between 2 and 3 by Pearson correlation or mutual information. In other words, these two measures possess the property of transitivity: If the time series $\hat{a}_1(t)$ is found to be linearly or nonlinearly related to $\hat{a}_2(t)$ and $\hat{a}_3(t)$, this relationship also holds for the pair $\hat{a}_2(t)$ and $\hat{a}_3(t)$. One can see this geometrically by treating the normalized anomaly time series $\hat{a}_i(t)$ as elements of a $\mathcal{T}$-dimensional vector space $\mathbb{R}^{\mathcal{T}}$. Pearson correlation R_{ij} corresponds to the scalar product of this space and measures the degree of parallelism of the vectors $\hat{a}_i(t)$ and $\hat{a}_j(t)$, where $R_{ij} = \pm 1$ if they are parallel or antiparallel. Hence, if two vectors are close to parallel to a third vector, they also must be close to parallel to each other.

Starting from the triangle (Fig. 3.12(a)), higher order manifestations of the transitivity problem have to be considered, *e.g.*, the "horseshoe" of direct interactions $4-2-1-3-5$ in Fig. 3.12(b). It is reasonable to assume that the addition of noise along each edge and the influence of other regions acting as sources of dynamical information external to the

"horseshoe" lead to a lower probability of falsely adding edge {4,5} to the climate networks as compared to the triangle case. We generally think of the probability of falsely adding edges as decreasing with the length of the path of indirect interaction [1]. In conclusion, the transitivity problem can effect local and mesoscopic network properties, whereas the path-based measures on the global topological scale should be more robust.

To avoid the transitivity problem, more advanced measures have to be introduced to climate network construction, that allow to differentiate between direct and indirect interactions in multivariate dynamical systems. Promising candidates are Granger causality (Granger and Hatanaka (1964)), conditioned transfer entropy (Schreiber (2000)) and partial coherence (Schelter et al. (2006)). The former two concepts have already been applied within the realm of climate science (*e.g.*, Mosedale et al. (2006), Kleeman (2007), Verdes (2005)).

3.6. Relationship to standard methods of teleconnection analysis

Here we briefly address the formal relationship of climate networks to two of the most widely used methods for the determination of teleconnection patterns: Correlation analysis (CA) and empirical orthogonal function (EOF) analysis (von Storch and Zwiers (1999)). We will show that both can be embedded in the framework of complex network theory. Note that the two classical methods allow only to study linear statistical interrelationships within gridded climatological fields and are limited to the local topological scale, since they are concerned with the dynamics of single grid points and pairs of these. Complex network theory allows to naturally extend the ideas behind CA and EOF to include nonlinear relationships between the dynamics on grid points, to the interesting mesoscopic and global topological scales as well as to systematically illuminate the interaction of two or more observables, *e.g.*, climate, vegetation and land use in the domain of Earth system analysis (Schellnhuber and Wenzel (1998)).

Let us consider a fully connected and weighted climate network $G_w := (V_w,E_w)$ constructed from and weighted by Pearson correlation, *i.e.*, $w_{ij} = R_{ij}, \forall\{i,j\} \in V_w \times V_w$. To obtain a fully connected network within our method, it is sufficient to set the threshold to $\tau = \min_{ij} C_{ij}$. The *intensity*

$$h_i = \sum_{j=1}^{N} A_{ij} W_{ij} \tag{3.4}$$

of vertex i is defined as the sum of the weights of the edges attached to it (Arenas et al. (2008)).

1 The length of the path of indirect interaction is 2 for the triangle (Fig. 3.12(a)) and 4 for the "horseshoe" (Fig. 3.12(b)).

3.6.1. Correlation analysis

The jth row or column $\mathbf{r}_j \in \mathbb{R}^N$ of the Pearson correlation matrix $\mathbf{R}$ is referred to as a *correlation map* (von Storch and Zwiers (1999)). There are N such maps containing the estimated linear Pearson correlation from base point j to all other grid points (Fig. 3.13(a)). In the language of complex network theory, the correlation map $\mathbf{r}_j$ corresponds to the ordered set of weights $(W_{ij})_i$ of the edges attached to vertex j in the weighted network G_w. The vertex intensity h_j is proportional to the mean of correlation map j, *i.e.*, $h_j = N \langle \mathbf{r}_j \rangle$.

Correlations between spatially adjacent grid points are usually positive and large due to the continuity of the underlying physical fields, while dipole-like interactions in the atmosphere may give rise to pronounced negative correlations between very distant grid points. A well-known example for a dipole-like oscillation in the pressure field is the North Atlantic Oscillation pattern (NAO) with a center of action over the Azores and one over Iceland (Wallace and Gutzler (1981)). The *teleconnectivity map* $\mathbf{T}$ is particularly useful to visualize these dipole-like patterns and to summarize the information contained in the correlation maps, its local minima revealing the potential centers of action of the teleconnection patterns (von Storch and Zwiers (1999)). Its elements T_j are defined to be the most negative correlation between grid point j and all others, *i.e.*,

$$T_j = \min_i R_{ij} = \min_i (\mathbf{r}_j)_i. \tag{3.5}$$

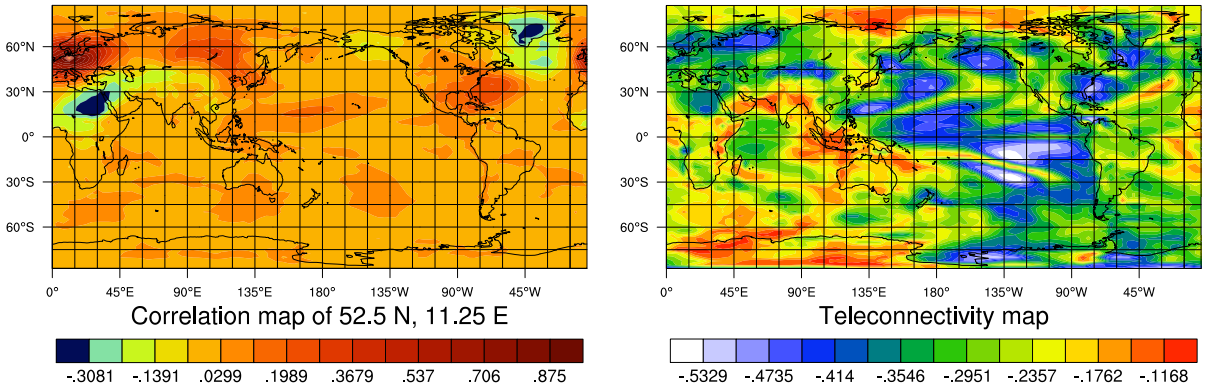
In other words, the teleconnectivity map gives the minimum weight attached to vertex j for all $j \in V_w$, that is to say $T_j = \min_i W_{ij}$.

3.6.2. Empirical orthogonal function analysis

Empirical orthogonal function (EOF) analysis presents an elegant and widely used method for finding the most dominant, linearly uncorrelated spatial and temporal structures within the anomaly field $\mathbf{a}(t)$ of some climatological observable (von Storch and Zwiers (1999)) and hence well suited for the study of teleconnection patterns. EOF analysis is an indispensable linear tool for dimensionality reduction of high dimensional climatological data sets like the ones analyzed in this work, because in many cases the sum of a small set of static orthogonal spatial patterns $\mathbf{e}_j$ with time dependent coefficients $\alpha_j(t)$ can explain most of the variance contained in the data set. We can expand the anomaly field as

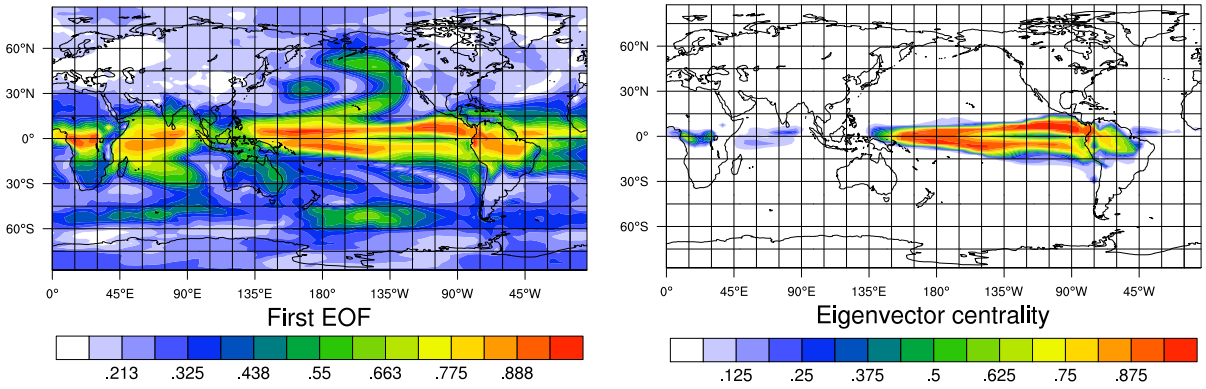
$$\mathbf{a}(t) = \sum_{j=1}^{N} \alpha_j(t) \mathbf{e}_j. \tag{3.6}$$

The EOFs $\mathbf{e}_j$ are chosen to maximize the variance of the anomaly field projected onto the respective subspace spanned by $\mathbf{e}_j$. They are ordered such that $\mathbf{e}_1$ explains most of the total variance of the multivariate data set, $\mathbf{e}_2$ explains less and so on until finally $\mathbf{e}_N$ accounts for the smallest fraction of the total variance. Specifically, for the first EOF this amounts to



(a) (b)

Figure 3.13 Correlation and teleconnectivity maps for the HadCM3 SAT data set. (a) Correlation map of a grid point over Potsdam, Germany at $\lambda = 52.5°N$ and $\phi = 11.25°E$. Note the wave train patterns over the northern hemisphere and the weak structures south of Australia in the mid-latitudes of the southern hemisphere. (b) Teleconnectivity map integrating information from all N correlation maps. For example, the local minimum south of Greenland is a center of action of the North Atlantic Oscillation (NAO), while the three local minima in the North Pacific at approximately $(45°N, 150°W)$, $(30°N, 167°E)$ and $(20°N, 135°E)$ are associated to the Pacific North America Pattern (PNA) (Wallace and Gutzler (1981)).



(a) (b)

Figure 3.14 Comparison of (a) the first EOF (leading eigenvector) of the Pearson correlation matrix $\mathbf{R}$ and (b) Newman's unweighted eigenvector centrality EC_v of the Pearson correlation climate network at $\rho = 0.01$, calculated from the HadCM3 SAT data set. Both fields have been normalized, so that the largest element is equal to unity. Note that as expected, the strongest features of the first EOF are also found in the eigenvector centrality field (Sect. 3.6.2).

minimizing the error

$$\left\langle ||\mathbf{a}(t) - \langle \mathbf{a}(t), \mathbf{e}_1 \rangle ||^2 \right\rangle_t .$$

The EOFs can be calculated as the eigenvectors of the Pearson correlation matrix $\mathbf{R}$, *i.e.*, as solutions of the equation

$$\mathbf{R}\mathbf{e}_j = \nu_j \mathbf{e}_j, \tag{3.7}$$

the associated eigenvalues ν_j giving the variance explained by the EOF $\mathbf{e}_j$. The dominating EOF $\mathbf{e}_1$ is hence the eigenvector corresponding to the largest eigenvalue ν_1.

Interestingly, the first EOF $\mathbf{e}_1$ finds its equivalent in complex network theory in Newman's *weighted eigenvector centrality* EC_v calculated for the fully connected and Pearson correlation weighted network G_w (Newman (2004)). EC_v of vertex v is defined to be proportional to the sum of the weighted eigenvector centralities of the vertex's neighbors multiplied by the weights of the edges connecting them. This allows a vertex to obtain a high EC_v by either being adjacent to many other vertices or by being strongly connected (with high edge weight) to a smaller set of vertices that themselves have a high weighted eigenvector centrality [1]. This consideration yields the eigen-equation

$$\overline{EC}_v = \chi^{-1} \sum_{i=1}^{N} A_{iv} W_{iv} \overline{EC}_i, \tag{3.8}$$

where the desired weighted eigenvector centrality EC_v can be shown to be given by the leading eigenvector of the weighted adjacency matrix with elements $W_{ij}A_{ij}$ (Friedkin (1991)). For G_w, $W_{ij}A_{ij} = R_{ij}$ and therefore $EC_v = (\mathbf{e}_1)_v$ for all $v \in V_w$. The eigenvector centrality field of an unweighted and thresholded climate network at low edge density, *i.e.*, with $W_{ij} = 1$ for all $\{i,j\}$, is still expected to resemble the first EOF, because only the edges with highest weight remain in the network that contribute most to the sum in Eq. 3.8 (Fig. 3.14).

3.7. Conclusions and summary

In summary, we have performed a systematic study of the similarity of climate networks constructed using the linear Pearson correlation and the nonlinear mutual information across local, mesoscopic and global topological scales. First, we have motivated the comparison of the two types of networks at equal edge densities. We have considered only low edge densities, that were shown to yield networks containing statistically highly significant edges as established on the basis of various significance tests. It has been then consistently shown for AOGCM and reanalysis surface air temperature data, that the networks agree well on

1 Note that this is the central idea of the "PageRank" algorithm, that the search engine Google is founded on (Brin and Page (1998)).

the local and mesoscopic topological scales. Using the surface pressure field to construct climate networks also yields qualitatively similar results and identical conclusions on these scales. For the surface air temperature data sets, we have found some interesting qualitative and quantitative deviations at the global scale using betweenness centrality. Even though there still is a high degree of similarity, the deviations are highly localized and structured pointing at a possible involvement of nonlinear processes in their formation.

This work also demonstrates, that our method of calculating mutual information for relatively short time series is reliable at least for the strongly linear interrelations detected by the Pearson correlation coefficient. The global topological scale is of particular interest, since it opens novel perspectives for the understanding of climatological phenomena. For example, as applied to the climate networks discussed in this article, betweenness centrality allows to measure the importance of localized regions on the earth's surface for the transport of dynamical information within a climatological field in the long term mean (Donges et al. (2008) and Chap. 5). Further work is needed to establish, whether the observed deviations on the global topological scale could be due to nonlinear physical processes in the climate system, that are only detectable using mutual information. In the future, we plan to assess this problem by constructing climate networks using a novel method based on statistical significance, *i.e.*, by adding edges to the climate network depending on the significance level of the correlation measure with respect to reasonable null hypotheses. One could then identify candidates for nonlinear interrelationships as edges that have an associated significant mutual information and a Pearson correlation that is not significant.

CHAPTER 4

Surrogate data sets and network models

> *As with EOFs and other patterns, there is a tendency to confuse physical and statistical significance of teleconnection patterns. In general, the patterns are worthy of physical interpretation when the basic structure is not strongly affected by sampling variability (*i.e.*, when there is reproducibility).*
>
> Hans von Storch and Francis W. Zwiers, "Statistical Analysis in Climate Research" (1999)

In essence, out method of climate network construction introduced in Chap. 3 presents a sophisticated nonlinear filter transforming a multivariate data set of N anomaly time series into the complex network domain. The resulting climate networks inevitably depend on systematic and random errors present in the source data set, *e.g.*, inhomogeneous sampling in space and time for reanalysis data or the parameterization of small scale physical processes[1] for AOGCM data constitute systematic errors. Given the presence of these intrinsic uncertainties, we need to assess the robustness of our results, a prototypical example being the delicate structures in the betweenness centrality field described in detail in Chap. 5. In other words, applying Occam's razor we seek the simplest statistical model able to explain the structures extracted from the original data set. Once such a statistical model is found, we only accept physical interpretations of these structures which are compatible with the assumptions made in the construction of the statistical model, *i.e.*, that do not require additional or contradictory presumptions. The statistical models employed in this work are all based on constrained randomization schemes, that is some properties of the observed entity are held (approximately) constant while an as great as possible degree of randomness is maintained otherwise. The conserved properties correspond to the assumptions that we impose on the statistical model and against which we judge the robustness of our results.

In our application we introduce statistical models at two levels of abstraction. Firstly,

1 Notable examples are turbulence and cloud formation.

we can generate model or surrogate data sets on the level of time series analysis, where each time series in the original data set is modeled by a univariate surrogate (Schreiber and Schmitz (2000)). Surrogate data sets can be used to assess the significance of the statistical interrelationships of pairs of time series detected in the original data set and shed light on which model assumptions may be sufficient to explain certain observed climate network properties (Sect. 4.1). Secondly, in the complex network domain we study network models or surrogates that present generalizations of the Erdős-Rényi random graph (Sect. 4.2) and particularly allow us to relate the elaborate betweenness structures found in surface air temperature climate networks (Chap. 5) with the results from earlier studies of climate networks (Tsonis and Roebber (2004), Tsonis et al. (2006), Tsonis et al. (2008b)). We furthermore introduce surrogate network ensembles and related concepts (Sect. 4.3).

It should be noted that while a well established theoretical framework for univariate time series surrogates exists (Schreiber and Schmitz (2000)), to our best knowledge this is so far not the case for network models (Zamora-López (2008)). Furthermore within this first iteration of climate network research, the introduction and study of multivariate time series surrogates appeared to be impractical for conceptional and computational reasons given large data sets of $N \approx 10^4$ time series. We hence followed an explorative approach when comparing the measured properties of data sets (Sect. 4.1) and climate networks (Sect. 4.2) to those of various surrogates. Particularly, giving "numbers" such as confidence intervals or significance levels proved to be out of scope for most measured properties in the context of this thesis. To develop the conceptual foundations to meaningfully be able to provide such "numbers" is highly desirable and should receive increased attention in future research. As a first step in this direction we introduce the concept of surrogate network ensembles and the Z-score (Sect. 4.3) and conclude with a short summary (Sect. 4.4).

4.1. Surrogates for univariate time series

We generate surrogate time series $s_i(t)$ from the normalized anomaly time series $\hat{a}_i(t)$ obtained from a climatological data set (Sect. 3.1.2) relying on a hierarchy of statistical models ranging from shuffled time series surrogates (Sect. 4.1.1) over Fourier surrogates (Sect. 4.1.2) to

Table 4.1 Time series properties (approximately) conserved by shuffled surrogates, Fourier surrogates and twin surrogates. Exactly conserved properties are marked by ♣, those approximately conserved by ♢. $p(\hat{a})$ denotes the PDF of time series $\hat{a}(t)$, while $P_{\hat{a}}(\omega)$ and $M_{\hat{a}}(t)$ respectively indicate its power spectrum and self mutual information.

	Shuffled	Fourier	Twin
$p(\hat{a})$	♣		♢
$P_{\hat{a}}(\omega)$		♢	♢
$M_{\hat{a}}(t)$			♢

twin surrogates (Sect. 4.1.3). In this order the surrogates preserve an increasing number of statistical time series properties such as the probability density function (PDF) $p(\hat{a}_i)$, the power spectrum $P_{\hat{a}_i}(\omega)$ or the self-mutual information $M_{\hat{a}_i}(t)$ of $\hat{a}_i(t)$ (Table 4.1). We make use of these types of surrogates to assess the significance of statistical interrelationships between pairs of anomaly time series as measured by Pearson correlation and mutual information in Sect. 4.1.4.

4.1.1. Shuffled surrogates

A shuffled surrogate $s^R(t)$ is generated by randomly permuting the anomaly time series $\hat{a}(t)$ in the time direction, *i.e.*, by applying a random permutation matrix Π_{ij} to the anomaly time series,

$$s^R(t) = \sum_{j=0}^{\mathcal{T}-1} \Pi_{tj} \hat{a}(j), \tag{4.1}$$

where $\mathcal{T}$ gives the length of the time series. The $s^R(t)$ therefore conserve only the PDF $p(\hat{a})$ of the original anomaly time series exactly. Shuffled surrogates hence correspond to realizations of white noise with a given PDF $p(\hat{a})$. A shuffled surrogate together with the original time series is shown in Fig. 4.1(a).

4.1.2. Fourier surrogates

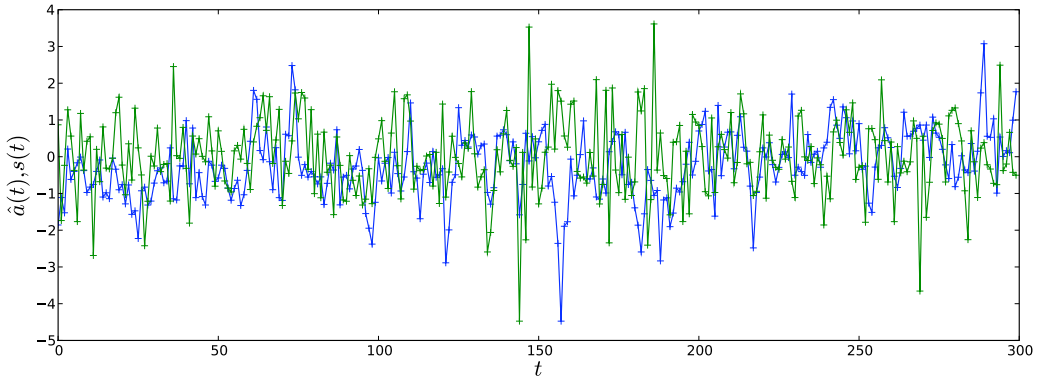
We produce Fourier surrogates $s^F(t)$ by randomizing the phases of the anomaly time series $\hat{a}(t)$, so that the power spectrum $P_{\hat{a}}(\omega)$ or equivalently the linear autocorrelation function remain approximately unchanged. Specifically, we compute the discrete Fourier transform

$$(\mathcal{F}(\hat{a}))(f) = \sum_{t=0}^{\mathcal{T}-1} e^{2\pi i \frac{tf}{\mathcal{T}}} \hat{a}(t) \tag{4.2}$$

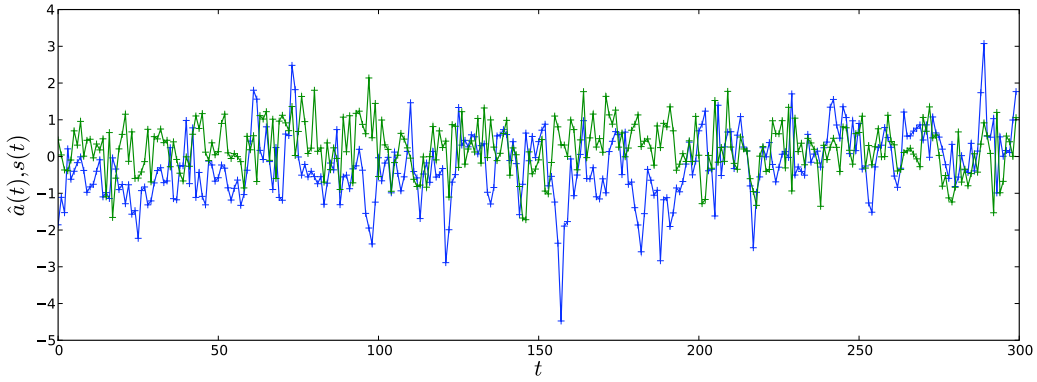
of $\hat{a}(t)$ for all $f \in \{0, \ldots, \mathcal{T}-1\}$. We then add random phases Φ_f drawn from a uniform distribution over the interval $[0,2\pi]$ to each element of the time series in Fourier space and transform back into the time domain to obtain

$$s^F(t) = \frac{1}{\mathcal{T}} \sum_{f=0}^{\mathcal{T}-1} e^{-2\pi i \frac{tf}{\mathcal{T}} + i\Phi_f} (\mathcal{F}(\hat{a}))(f). \tag{4.3}$$

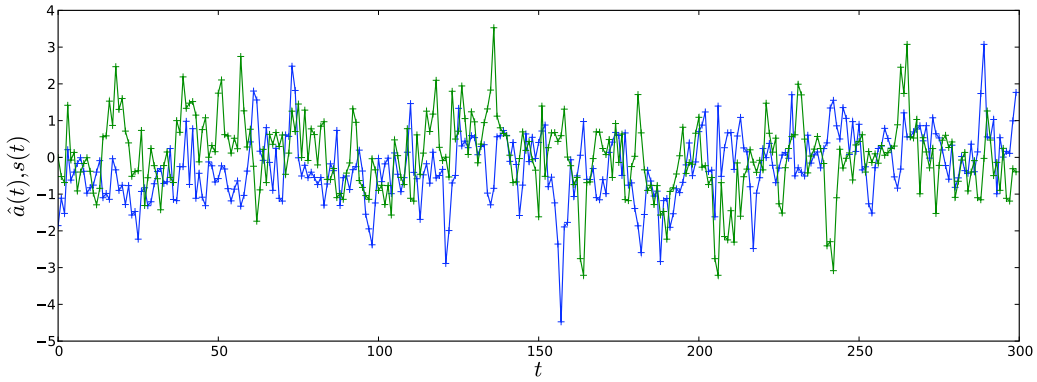
Because of the finite length of the time series involved, the power spectrum is only preserved imperfectly using this simple, but computationally effective approach. More elaborate techniques, that are furthermore able to preserve the PDF of $\hat{a}(t)$ have been proposed in the literature (Schreiber and Schmitz (2000)). The Fourier surrogates introduced above correspond to realizations of colored noise with a given power spectrum $P_{\hat{a}}(\omega)$. A Fourier surrogate together with the original time series is shown in Fig. 4.1(b).



(a)



(b)



(c)

Figure 4.1 Segment of a normalized anomaly time series (blue) and the corresponding surrogate (green) for a grid point at $(-37.5°N, 67.5°E)$, taken from the HadCM3 SAT data set. The panels display the comparison with (a) a shuffled surrogate, (b) a Fourier surrogate and (c) a twin surrogate.

4.1.3. Twin surrogates

We generate twin surrogate time series $s^T(t)$ from the normalized anomaly time series $\hat{a}(t)$ using the twin surrogate algorithm proposed by Thiel et al. (2006). This algorithm is based on the recurrence structure of the original anomaly time series $\hat{a}$, that is captured by the *recurrence matrix* $\mathcal{R}_{ij}$ (Marwan et al. (2007)). Since the anomaly time series provides only a one dimensional representation of the dynamics at a specific grid point, we generally have to try to reconstruct the complete phase space dynamics by time-delay embedding with embedding dimension m and time delay τ. We thus use the reconstructed phase space trajectory $\hat{\mathbf{a}}(t) = (\hat{a}(t-(m-1)\tau), \hat{a}(t-(m-2)\tau), \ldots, \hat{a}(t)) \in \mathbb{R}^m$ to calculate the recurrence matrix

$$\mathcal{R}_{ij} = \Theta(\delta - ||\hat{\mathbf{a}}(i) - \hat{\mathbf{a}}(j)||), \tag{4.4}$$

where $\Theta(.)$ denotes the Heaviside function, $||.||$ the maximum norm and δ a predefined threshold. We then refer to *twins* as points $\hat{\mathbf{a}}(i)$ and $\hat{\mathbf{a}}(j)$ in the embedded time series that correspond to identical columns in the recurrence matrix, *i.e.*, $\mathcal{R}_{ki} = \mathcal{R}_{kj} \forall k$. Note that even though twins share a common neighborhood on the reconstructed attractor, they generally have different pasts and futures. Hence randomness can be introduced by randomly choosing from the future of point $\hat{\mathbf{a}}(i)$ and the futures of its twins when building up the m-dimensional surrogate trajectory $\mathbf{s}^T(t)$. Specifically, following Thiel et al. (2006) and Marwan et al. (2007) we proceed as follows:

(i) Find all pairs of twins $\hat{\mathbf{a}}(i)$,$\hat{\mathbf{a}}(j)$ satisfying $\mathcal{R}_{ki} = \mathcal{R}_{kj} \forall k$ with a minimum temporal separation Δt, *i.e.*, $|i - j| \geq \Delta t$[1].

(ii) Select an arbitrary starting point $\hat{\mathbf{a}}(j)$ and set $\mathbf{s}^T(1) = \hat{\mathbf{a}}(j)$. Set index $i = 2$.

  (A) IF $\hat{\mathbf{a}}(j)$ has no twins, set $\mathbf{s}^T(i) = \hat{\mathbf{a}}(j+1)$,

  (B) IF $\hat{\mathbf{a}}(j)$ has at least one twin, choose between $\hat{\mathbf{a}}(j+1)$ and the futures of the twins of $\hat{\mathbf{a}}(j)$ with equal probability. *E.g.*, if $\hat{\mathbf{a}}(j)$ has one twin $\hat{\mathbf{a}}(k)$, set $\mathbf{s}^T(i) = \hat{\mathbf{a}}(j+1)$ or $\mathbf{s}^T(i) = \hat{\mathbf{a}}(k+1)$ with equal probability.

(iii) Increase $i \to i+1$ and return to step (iii), until $\mathbf{s}^T(t)$ has the same length as $\hat{\mathbf{a}}(t)$.

To obtain a one dimensional twin surrogate $s^T(t)$ it is sufficient to use an arbitrary single component of the m-dimensional surrogate trajectory $\mathbf{s}^T(t)$. Since twin surrogates correspond to the null hypothesis of shadowing typical trajectories on the same attractor as the original data with random initial conditions, they conserve all linear and nonlinear properties of single time series in the limit of infinitely many samples. Therefore, they allow for stronger

1 The minimum temporal separation Δt is introduced to avoid false twins due to oversampling of the attractor along the trajectory.

and more meaningful tests than those based on randomly shuffled time series (corresponding to the null hypothesis of a random process with the same probability distribution as the original time series) and Fourier surrogates (corresponding to the null hypothesis of a random process having the same power spectrum as the original time series).

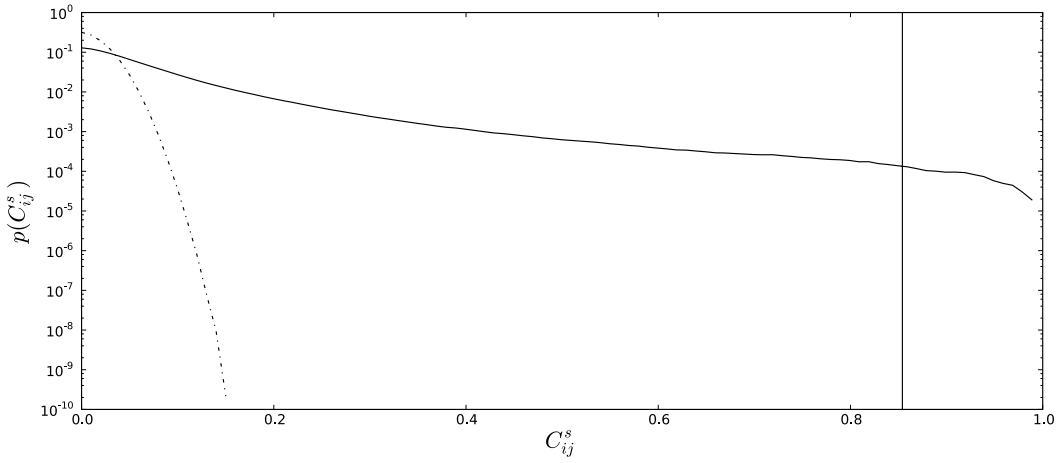
A twin surrogate together with the original time series is shown in Fig. 4.1(c). It should be noted that twin surrogates were used successfully to test for phase synchronization (Thiel et al. (2006)) and coupling asymmetries (Romano et al. (2007)) between two time series by testing one original time series against twin surrogates of the other and vice versa.

Note that considering all statistical tests using twin surrogates performed for this thesis we first did not embed the time series, *i.e.*, chose the trivial embedding parameters $m = 1$ and $\tau = 0$, because the determination of optimal embedding parameters 'by hand' (Kantz and Schreiber (2004)) is not feasible for the large number of time series analyzed here and automated embedding schemes are out of the scope of this work. We have however experimented with other embedding parameters, *e.g.*, $m = 3$ and $\tau = 2$, and found that the test results are robust with respect to reasonable parameter choices. However to further increase the test power, it is desirable to in the future employ algorithms for the automated selection of optimal embedding parameters for each time series separately (Marwan et al. (2007)).

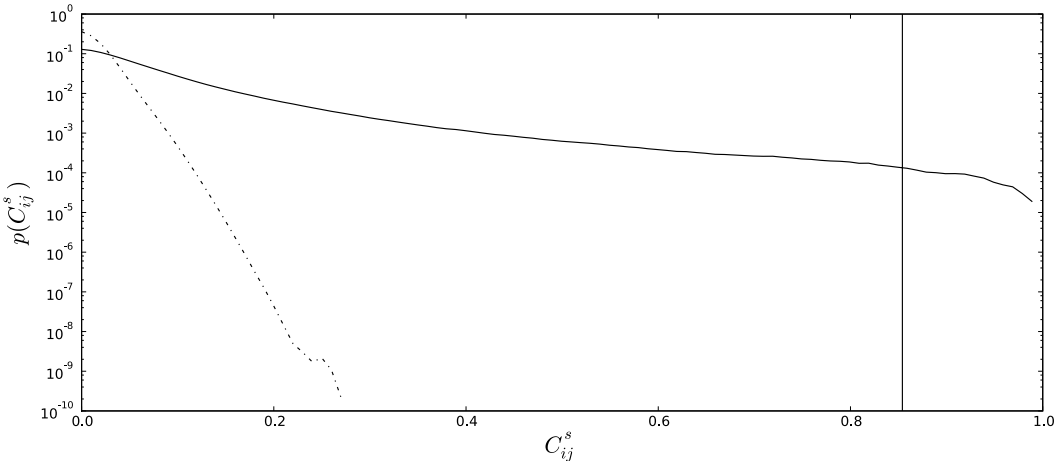
4.1.4. Significance of statistical interrelationships

We would like to evaluate to which degree the strong statistical interrelationships corresponding to edges in our climate networks are statistically significant, *e.g.*, how likely they are to arise by chance. Particularly, our null hypothesis is that the anomaly time series $\hat{a}_i(t)$,$\hat{a}_j(t)$ are statistically independent for all ordered pairs (i,j). We test this null hypothesis by generating n realizations of univariate time series surrogates $s_i^\mu(t)$, $\mu = 1,\ldots,n$, for all anomaly time series $\hat{a}_i(t)$ and calculating the correlation measure C_{ij}^μ between $\hat{a}_i(t)$ and $s_j^\mu(t)$ for all n realizations of surrogates. Then, the PDF $p(C_{ij}^s)$ of C_{ij}^μ over all realizations μ is estimated. If C_{ij} is found to be very unlikely to be drawn from the PDF $p(C_{ij}^s)$, we can reject the null hypothesis for the pair of anomaly time series $\hat{a}_i(t)$,$\hat{a}_j(t)$ with respect to the assumptions implicitly contained in the choice of surrogate type. In this work, the correlation measure C_{ij}^μ can be chosen from Pearson correlation and mutual information at zero lag (Sect. 3.2.1). Note that in general the test matrix C_{ij}^μ will not be symmetric.

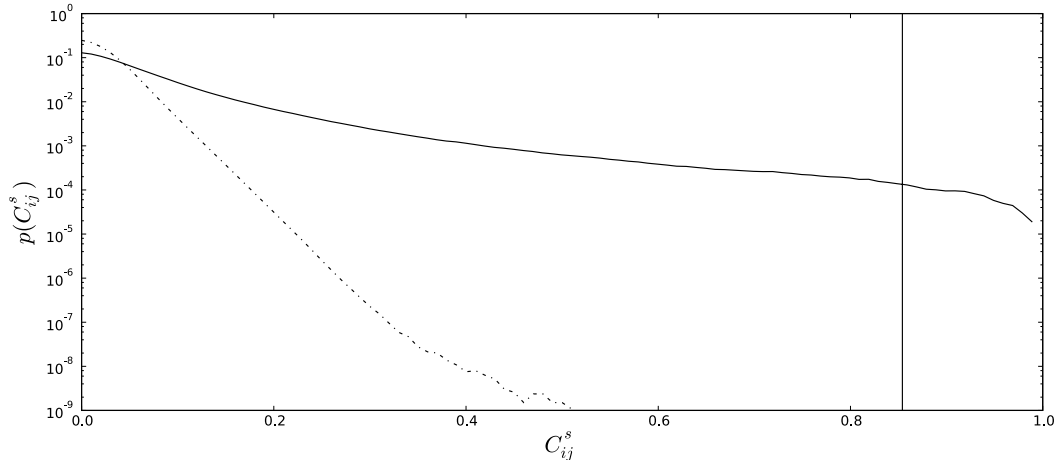
We now estimate the PDF $p(C_{ij}^s)$ of the resulting n test matrices and compare it to the PDF $p(C_{ij})$ of the correlation measure calculated within the original data set. We generally find that the thresholds used for generating the low edge density climate networks studied in this thesis lie far removed from the compact support of $p(C_{ij}^s)$ (Fig. 4.2). This implies that for all of the strong statistical interrelationships included in our climate networks as edges, the null hypothesis of statistical independence can be rejected safely at a high confidence level. Specifically, testing with the three types of surrogates described above shows that high values of the correlation measure C_{ij} do not arise as artifacts of the PDF, the power



(a)



(b)



(c)

Figure 4.2 Comparison of the distributions $p(C_{ij})$ of the correlation measure C_{ij} calculated from the original data set (solid lines) and that of the test matrix $p(C_{ij}^s)$ calculated from $n = 100$ realizations of surrogates (dashed lines) for the HadCM3 SAT data set. Pearson correlation was used as the correlation measure here. (a) Comparison for shuffled surrogates, (b) Fourier surrogates and (c) twin surrogates with $m = 1$, $\tau = 0$, $\delta = 0.1$ and $\Delta t = 7$. The vertical lines indicate the threshold corresponding to an edge density of $\rho = 0.005$. For all types of surrogates, the probability that C_{ij}^s exceeds the threshold is essentially zero.

spectrum or the geometry of the attractor of the single anomaly time series $\hat{a}_i(t)$,$\hat{a}_j(t)$. We hence conclude that on the level of (univariate) time series, the edges included in our climate networks are highly unlikely to arise by chance.

It should be pointed out, that the power of the statistical test against the null hypothesis of independent time series increases within the hierarchy of time series, *i.e.*, the maximum value of C_{ij}^s, $max(C_{ij}^s)$, depends on the type of time series surrogate used (Fig. 4.2). The power of the test increases with $max(C_{ij}^s)$, because a statistical interrelationship C_{ij} has to be larger to be considered significant, when $max(C_{ij}^s)$ is larger. In this sense, Fourier surrogates are more powerful than shuffled surrogates, and twin surrogates are more powerful than shuffled surrogates and Fourier surrogates. This observation is plausible, because an increasing amount of single time series properties is conserved when moving upwards in the surrogate hierarchy (Table 4.1). Since they allow for the most powerful tests, we hence use twin surrogate data sets in all further statistical tests on the time series level (*e.g.*, Chapter 5).

4.1.5. Surrogate data sets

In addition to comparing original and surrogate time series, it is interesting to study which properties of a particular climate network can be explained by generating climate networks from an ensemble of *surrogate data sets*. A surrogate data set μ is a collection of surrogate time series $s_i^\mu(t), i = 1, \ldots, N$ of the same type (Table 4.1), one for each anomaly time series in the original gridded data set. We can then construct an ensemble of n climate networks G^μ from n independently generated surrogate data sets $s_i^\mu(t)$, where $\mu = 1, \ldots, n$. In Sect. 4.3 we show how the resulting surrogate data set network ensemble can be used to assess the significance of any network property of interest, that is observed in the original network.

Table 4.2 Network properties (approximately) conserved by Erdős-Rényi graphs, the configuration model, random link switching and geographical models I and II. Exactly conserved properties are indicated by ♣, those approximately conserved by ♢.

	Erdős-Rényi	Config.	Rand. link switch.	Geo. I	Geo. II
N	♣	♣	♣	♣	♣
L	♢	♢	♣	♣	♣
$p(k)$		♢	♣	♣	♣
k_v		♢	♣	♣	♣
$p_E(l)$				♢	♢
AED_v					♢

4.2. Network models

Analogously to our treatment of univariate time series surrogates, we present a hierarchy of undirected network models or surrogates by imposing an increasing number of constraints on the topology and spatial embedding of the otherwise random networks (Table 4.2). While Erdős-Rényi graph and configuration model belong to the class of constructive network models starting from an empty network of N initially unconnected vertices, random link switching and the geographical models I and II start from the original network and repeatedly make constrained changes until the desired randomness is sufficiently introduced. All network models introduced in this section can be straightforwardly generalized to directed networks (Appx. B).

4.2.1. Erdős-Rényi graphs

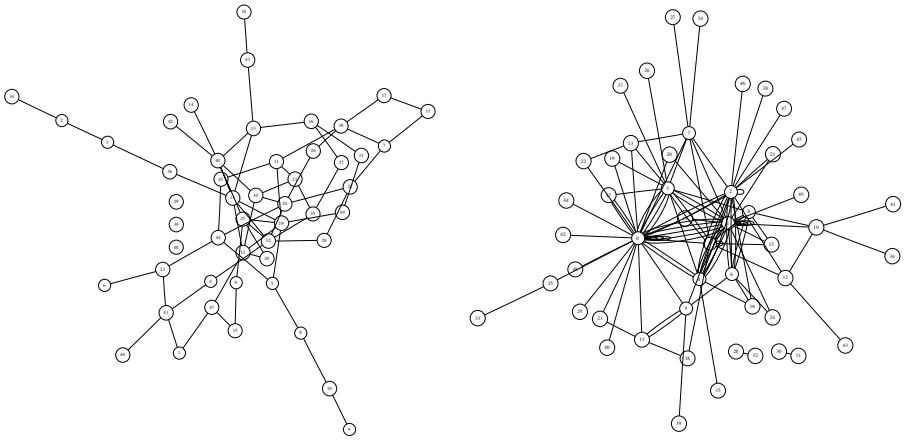
The network model we refer to as the Erdős-Rényi graph was introduced independently by Solomonoff and Rapoport (1951) and Erdős and Rényi (1959). N vertices are initially left unconnected. Each pair of vertices i,j is then independently connected by an edge $\{i,j\}$ with probability p (Fig. 4.3(a)). This leads to an expected number of edges $L = p\binom{N}{2} = p\frac{N(N-1)}{2}$ and a binomial degree distribution $p(k)$, that converges to a poissonian distribution in the thermodynamic limit $N \to \infty$ and $p \to 0$,

$$p(k) = \binom{N}{k} p^k (1-p)^{N-k} \to \frac{z^k e^{-z}}{k!}. \tag{4.5}$$

The degree distribution is therefore sharply peaked, with small fluctuations around the mean degree $z = \langle k_v \rangle_v = L/N = p(N-1)/2$. Because edges are added to the network independently, the expected clustering coefficient is simply given by $\mathcal{C} = p$, thus for fixed mean degree z it is of $\mathcal{O}(N^{-1})$ in the thermodynamic limit. The average path length can be shown to go as $\mathcal{L} \approx \ln N / \ln z$ (Newman (2003)). The most interesting property of the Erdős-Rényi graph is a phase transition at $z = 1$, where a giant component forms in the network[1].

Since the Erdős-Rényi graph conserves only the number of vertices N and the number of links L (or equivalently the edge density $\rho = p$) it presents the simplest network null-model (Table 4.2), that is of limited use when assessing the significance of climate network properties. Particularly, none of the most interesting local network properties, *e.g.*, centrality fields, can be reproduced (Fig. 4.4).

1 Compare this with the begin of the sharp increase in giant component size $S(\rho)$ for the HadCM3 SAT Pearson correlation climate network at $\rho' \approx 0.0002$ (Fig. 3.4(e)). This corresponds to an average degree $z' = N\rho' \approx 6816 \times 0.0002 = 1.36$. The Erdős-Rényi model is hence compatible with this 'phase transition' found in our climate network construction method.



(a) (b)

Figure 4.3 Visualization of two different types of random networks of order $N = 50$ using the Fruchterman-Reingold graph layout algorithm (Di Battista et al. (1998)). (a) Erdős-Rényi random graph with edge density $\rho = 0.05$, its binomial degree distribution is sharply peaked around the mean degree $\langle k_v \rangle_v = \rho N = 2.5$. (b) A configuration model random network with edge density $\rho = 0.10$ following a scale-free degree distribution $p(k) \propto k^{-2}$. Some high degree vertices (hubs) dominate the network, while most others have a very low degree. The network is not simple, it contains multiple edges and self-loops illustrating the greatest problem of the configuration model.

4.2.2. Configuration model

The configuration model allows to construct random surrogate networks with a prescribed number of vertices N and degree centrality field k_v. Each vertex v is first assigned the degree k_v; graphically this results in k_v 'stubs' of edges-to-be with free ends sticking out of vertex v. In the following, pairs of free ends are selected randomly and joined to form edges. It can be demonstrated that this procedure generates every possible topology of a network with given degree centrality field k_v with equal probability (Newman (2003)). Alas, the configuration model ensemble also encompasses networks that are not simple, since 'stubs' are joined independently (Fig. 4.3(b)). To obtain the simple networks needed for our application, we could repeatedly generate configuration model networks until a simple one is produced. Because non-simple networks are highly likely to be drawn from the configuration model ensemble, this recipe proves to be computationally not feasible. We have to resort to *simplifying* the resulting random network, *i.e.*, removing self-loops and joining multiple edges to one single edge. This obviously introduces imperfections: The degree field k_v and thus the degree distribution $p(k)$ and the number of links L of the original graph are not conserved exactly any more (Table 4.2). Particularly, the bias introduced by the simplification increases with degree k, because high degree vertices initially possess more 'stubs' and hence a greater probability to have self-loops and multiple edges attached to them.

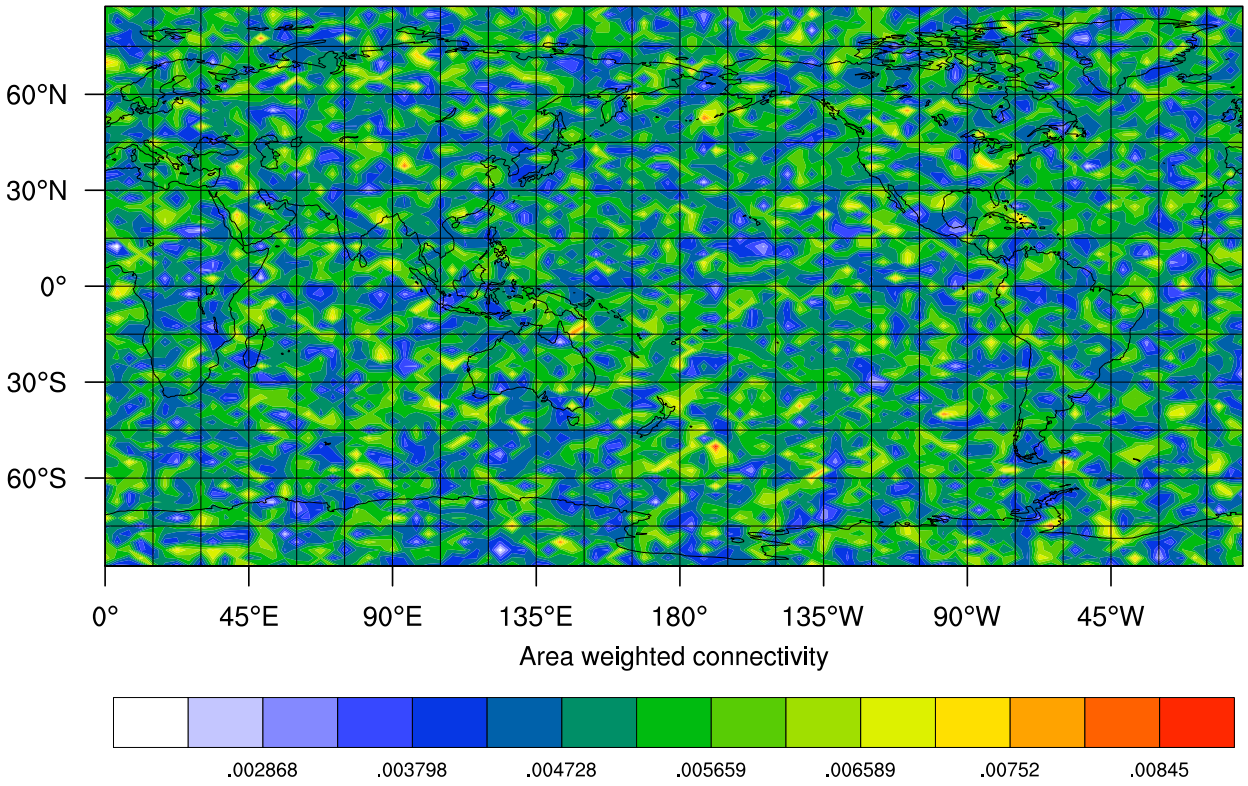


Figure 4.4 The area weighted connectivity field for a realization of the Erdős-Rényi graph with $N = 6816$ and $\rho = 0.005$, *i.e.*, the same parameters as those of the HadCM3 SAT climate network studied throughout this thesis. It is obvious that the Erdős-Rényi graph is not suitable for the study of local properties of climate networks.

4.2.3. Random link switching

Using random link switching to generate networks with a given number of vertices N and degree centrality field k_v overcomes the limitations of the configuration model (Table 4.2). In each switching step, two edges $\{i_1,j_1\}$ and $\{i_2,j_2\}$ are selected randomly from the original network and rewired as $\{i_1,j_2\}$ and $\{i_2,j_1\}$, if the new edges did not already exist before or introduce self-loops (Fig. 4.5). The price one has to pay for the exact conservation of k_v, $p(k)$ and L is an increased computational cost as compared to the configuration model. We have to rewire at least $n_r = L$ times to generate a *maximally random* network with given k_v, *i.e.*, to destroy any additional internal structure (Zamora-López (2008) and Fig. 4.5).

4.2.4. Surrogates for spatially embedded networks

When studying centrality fields, *e.g.*, closeness and betweenness, on the global topological scale (Sect. 2.2.3) in spatially embedded networks (Sect. 2.3) it is desirable to have surrogate networks at hand, that conserve the most basic spatial network statistics, *i.e.*, the edge distance distribution $p_E(l)$ and the average edge distance field AED_v. For this purpose we propose a generalization of the random link switching procedure outlined above (Sect. 4.2.3)

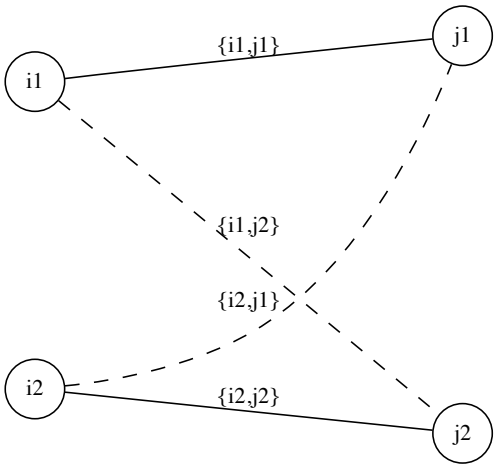
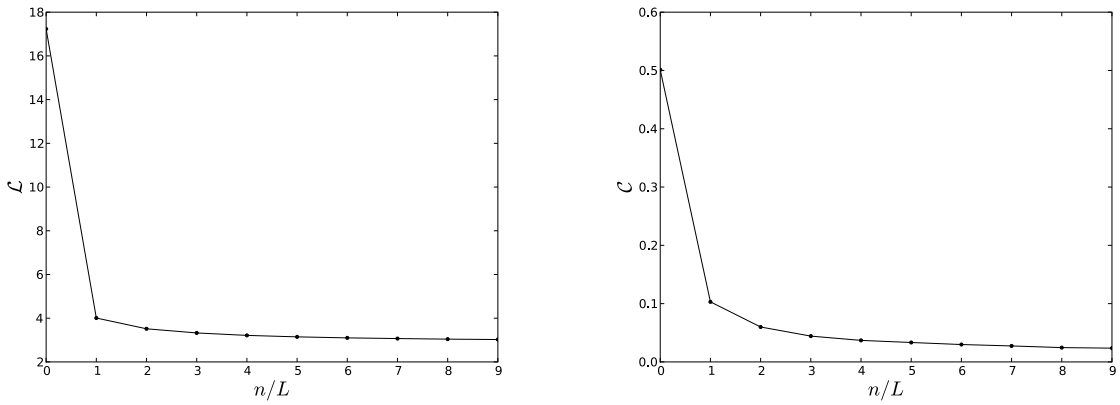


Figure 4.5 Illustration of the random link switching algorithm. The original edges $\{i_1,j_1\}$ and $\{i_2,j_2\}$ (continuous lines) are replaced by new edges $\{i_1,j_2\}$ and $\{i_2,j_1\}$ (dashed lines) in one rewiring step.



(a) (b)

Figure 4.6 Convergence of (a) average path length $\mathcal{L}$ and (b) clustering coefficient $\mathcal{C}$ to the maximally random state with number of rewiring steps over number of links n_r/L using geographical model I with $\varepsilon = 0.1$. While $\mathcal{L}$ and $\mathcal{C}$ decrease sharply during the first L rewiring steps, they approach their limits much more slowly for $n_r/L > 1$. The test network is a regional Pearson correlation climate network encompassing the mid-latitudes of the Southern Hemisphere (Fig. 5.5), that was constructed from the HadCM3 SAT data set at $\rho = 0.005$.

coming in two flavors. The geographical model I conserves $p_E(l)$ additionally to the random link switching method, while geographical model II conserves $p_E(l)$ as well as AED_v (Table 4.2). Again, more than $n_r = L$ edges have to be rewired to obtain a maximally random network, with a correspondingly large computational cost[1] (Fig. 4.6).

4.2.4.1. Geographical model I

To approximately conserve the edge distance distribution $p_E(l)$ we have to ensure that the rewired edges $\{i_1,j_2\}$ and $\{i_2,j_1\}$ contribute to the same bins in the edge distance histogram as the original edges $\{i_1,j_1\}$ and $\{i_2,j_2\}$. Specifically, this amounts to imposing the condition

$$C_1 := (|l_{i_1j_1} - l_{i_1j_2}| < \varepsilon \wedge |l_{i_2j_2} - l_{i_2j_1}| < \varepsilon) \vee (|l_{i_1j_1} - l_{i_2j_1}| < \varepsilon \wedge |l_{i_2j_2} - l_{i_1j_2}| < \varepsilon)\,,$$

and is achieved by the following algorithm:

(i) Start with a spatially embedded network $G := (V,E)$. Set index $k = 0$.

(ii) IF $k < n_r$:

  (A) Randomly choose two edges $\{i_1,j_1\}$ and $\{i_2,j_2\}$ from V.

  (B) IF edges $\{i_1,j_2\}$ and $\{i_2,j_1\}$ do not exist AND $i_1 \neq i_2 \neq j_1 \neq j_2$ AND C_1 is TRUE: Rewire the edges $\{i_1,j_1\}, \{i_2,j_2\} \to \{i_1,j_2\}, \{i_2,j_1\}$ and increase the index $k \to k + 1$.

  (C) Jump to step (ii).

(iii) Return the rewired network G'.

Note that the only parameter this algorithm depends on is the tolerance ε, determining the quality of conservation of $p_E(l)$. The run-time of this algorithm increases sharply with decreasing tolerance ε, since the probability of finding edges that are suitable for rewiring in each step also decreases.

4.2.4.2. Geographical model II

To additionally conserve the average edge distance field AED_v, we have to demand that the rewired edges $\{i_1,j_2\}$ and $\{i_2, j_1\}$ have approximately the same length, *i.e.*,

$$C_2 := |l_{i_1j_2} - l_{i_2j_1}| < \varepsilon'.$$

1 One can think of a variant of the configuration model here, where the free ends of the 'stubs' attached to randomly selected vertices i and j are joined with probability $p = p_E(l_{ij})$. Repeating this operation until no free stubs remain would allow to generate random networks with approximately conserved $p(k)$ and $p_E(l)$ at lower computational cost, but with the same problems as the configuration model.

C_2 together with C_1 corresponds to the condition, that all of the involved edges $\{i_1,j_1\}$, $\{i_2,j_2\}$, $\{i_1,j_2\}$ and $\{i_2,j_1\}$ have approximately the same length. We impose C_2 at step (iiB) of the algorithm described above, before condition C_1 is evaluated. This is more efficient since checking C_1 is computationally more costly than evaluating C_2. Hence, if C_2 is FALSE, we can skip testing for C_1 and return directly to step (ii). For convenience, we set $\varepsilon' = \varepsilon$.

4.3. Ensembles of surrogate networks

We refer to an *ensemble of surrogate networks* as a collection of n complex networks G^μ, $\mu = 1,\ldots,n$, independently generated from surrogate data sets (Sect. 4.1.5) or by a particular network model (Sect. 4.2). Any network property ξ^μ of interest can then be calculated for each ensemble member G^μ. Subsequently, the PDF of ξ^μ and its moments are estimated from the network ensemble using a Monte-Carlo approach[1]. For example, we can use the *ensemble mean* $m_r(\xi) = \langle \xi^\mu \rangle_\mu$ and the *ensemble standard deviation* $\sigma_r(\xi) = \sqrt{\langle (\xi^\mu - m_r(\xi))^2 \rangle_\mu}$ of ξ^μ to calculate the Z-score

$$Z(\xi) = \frac{\bar{\xi} - m_r(\xi)}{\sigma_r(\xi)}, \tag{4.6}$$

that allows us to assess the statistical significance of the measurement $\bar{\xi}$ from the original network with respect to the network model[2] that was used to generate the ensemble (Zacharias et al. (2002)). In other words, the Z-score $Z(\xi)$ quantifies by how many ensemble standard deviations the original network property and its ensemble mean differ. If therefore $|Z(\xi)| \gg 1$, we can consider the measured network property to be significant with respect to the chosen network model. It should be pointed out, that the Z-score $Z(\xi)$ necessarily has the same dimensionality as the measurement ξ itself, *i.e.*, it will either be a scalar, a 1D function or a 2D field. We will use the Z-score in later chapters to test our results from real networks against various types of surrogate networks.

4.4. Summary

We have introduced a hierarchy of three types of univariate time series surrogates: Shuffled surrogates, Fourier surrogates and twin surrogates. We have used them to show that within the power of the statistical test used, all edges in the low edge density climate networks considered in this thesis correspond to statistically significant interrelationships between the

1 Using Monte-Carlo sampling is necessary, since in general the distribution of some property ξ of a certain type of network model is not derivable analytically. Considering the network models presented here, some analytical results are known for the Erdős-Rényi graph and the configuration model (Newman (2003)).

2 Zacharias et al. (2002) in their seminal paper use random link switching (Sect. 4.2.3) to create the surrogate network ensemble.

dynamics on pairs of grid points. In addition we have developed a hierarchy of surrogate networks and particularly proposed a two-flavored algorithm designed to create surrogates for spatially embedded networks, such as climate networks. Finally, we have commented on the possibility to use surrogate network ensembles together with the Z-score to assess the statistical significance of any measured network property ξ with respect to the null hypothesis implicitly given by the choice of the type of surrogate network. Surrogate networks constructed from surrogate data sets correspond to the null hypothesis that the time series of the original data set are statistically independent under the constraint of given PDF $p(\hat{a})$ (shuffled surrogates), power spectrum $P_{\hat{a}}(\omega)$ (Fourier surrogates) and attractor geometry (twin surrogates). In similar fashion, on the network level surrogate networks reflect the null hypothesis that the original network is random under the constraint of given number of vertices N and edges L (Erdős-Rényi graph), degree field k_v and degree distribution $p(k)$ (Configuration model and random link switching), edge distance distribution $p_E(l)$ (Geographical model I) and average edge distance field AED_v (Geographical model II). The latter models also maintain the properties conserved by the former ones (Table 4.2).

CHAPTER 5

The backbone of the climate network

Here we report on how by combining our linear and nonlinear network construction techniques with advanced topologically global centrality measures, we uncover peculiar wave-like structures in the betweenness[1] fields of climate networks constructed from monthly averaged reanalysis and atmosphere-ocean coupled general circulation model (AOGCM) surface air temperature (SAT) data (Sect. 5.1). Akin to the homonymous data highways of the internet, these betweenness structures form the *backbone* of the SAT network, bundling most of the information flow between remote regions. Some major features of the backbone appear to be closely related to surface ocean currents pointing to an essential role of the oceanic surface circulation in stabilizing the climate system by promoting the global flow of dynamical information (Sect. 5.2). Note that these insights are conceptually new and cannot be obtained using classical methods of climatology such as principal component analysis (PCA) or singular spectrum analysis (SSA) of anomaly fields (von Storch and Zwiers (1999)), because these are by design local in a network sense and are not suitable to study local flow measures depending on a global network topology (Sect. 3.6).

We describe our results (Sect. 5.1) and give a physical interpretation of betweenness and the related concepts of dynamical information and information in the context of climate networks (Sect. 5.2). We have performed intensive statistical tests with various types of surrogates to ensure the robustness of our results (Sect. 5.3) and finish with our conclusions and an outlook (Sect. 5.4).

5.1. Results for AOGCM and reanalysis data

Following the method outlined in Chap. 3, we uncover peculiar wave-like structures of high betweenness in maps of both reanalysis and model SAT climate networks (Fig. 5.1). In analogy with the internet, we call the network of these channels of high information flow the *backbone* of the climate network. We observe that prominent mainly meridional features of

1 Whenever we speak simply of betweenness, we refer to the shortest path betweenness (Eq. 2.14).

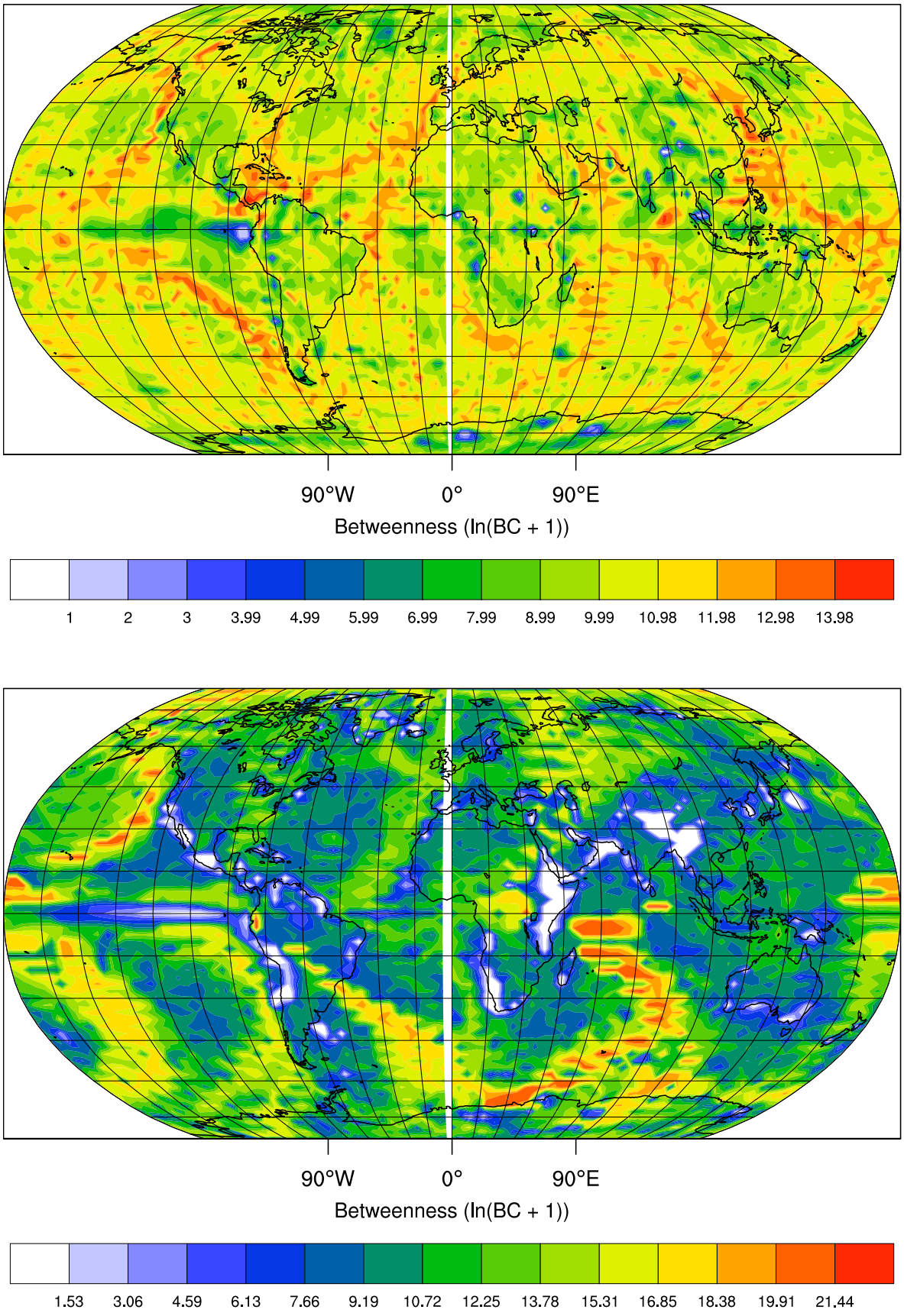


Figure 5.1 a) Betweenness for the NCEP/NCAR reanalysis SAT network at $\rho = 0.004$. b) Betweenness for the HadCM3 SAT network at $\rho = 0.005$. Note that some features of the HadCM3 backbone in b) correspond closely to ocean surface currents shown in Fig. 5.2, *e.g.*, the California, Peru and Canary currents.

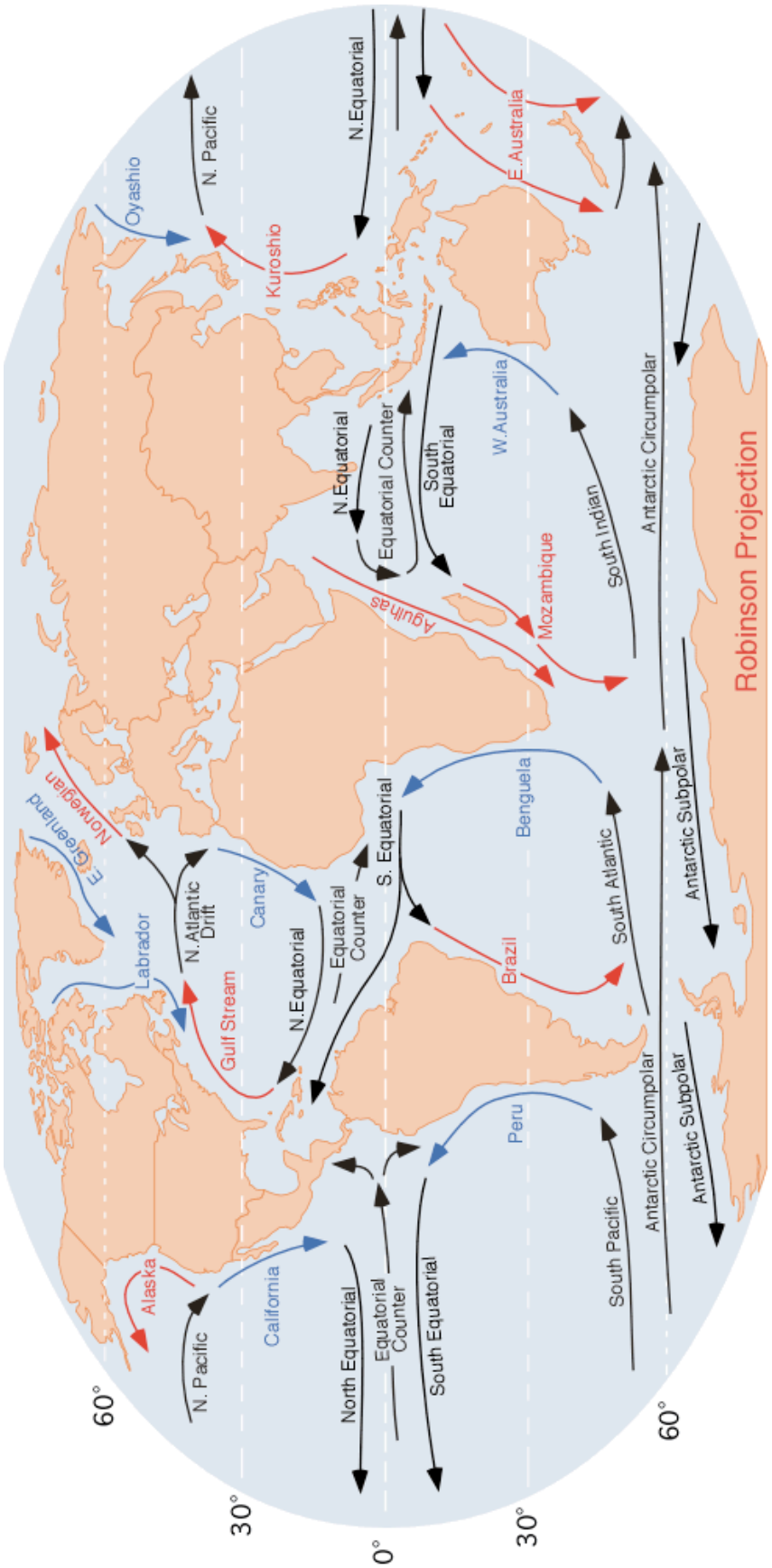


Figure 5.2 A schematic map of global surface ocean currents (Pidwirny (2006)).

the backbone tend to approach the equator tangentially, as one would expect from modes of the atmospheric and oceanic general circulation due to the vanishing coriolis force at the equator (Vallis (2006)). There is also a qualitative agreement on the location of major backbone structures for both reanalysis (Fig. 5.1(a)) and model networks (Fig. 5.1(b)), *e.g.*, the high betweenness channel over the Atlantic Ocean and the backbone structures over the eastern Pacific Ocean, both connecting the Arctic with the Antarctic. It is important to stress that some backbone structures, *e.g.*, the channel over the South Atlantic, disappear in climate networks constructed from the same model and reanalysis data sets, but using the linear Pearson correlation. This indicates, that nonlinear physical processes are important for the information transport within the climate system which are captured using the mutual information for network construction, but not by linear measures (Sect. 3.3).

Note that the strongest backbone structures lie mainly over the ocean and avoid to cross the land in both model and reanalysis climate networks. Therefore a physical mechanism involving an atmosphere-ocean coupling might be responsible for the information transport in the SAT field measured by betweenness. Indeed, some of the strongest features found in the HadCM3 betweenness field (Fig. 5.1(b)) as well as in the NCEP/NCAR betweenness field (Fig. 5.1(a)) resemble closely major surface ocean currents (Fig. 5.2). For example, note the strong betweenness structures off the west coast of North and South America that resemble the Alaska and Peru current, and the backbone feature along the west coasts of Africa and Europe following the path of the Canary and Norwegian currents. These observations can be understood considering the strong coupling between sea surface temperature (SST) and SAT over the ocean via heat flux (Stewart (2005)). Temperature anomalies in SST are advected by the surface ocean currents and transfered to the SAT field via heat flux coupling. Therefore, ocean currents provide a physical mechanism for the transport of dynamical information on localized linear structures over large distances. However, no clear traces of the strong western boundary currents (WBCs) such as the Gulf Stream or the Kuroshio are visible in the backbone structure (Fig. 5.1(b)). This might be due to the fact, that WBCs are much narrower than the eastern boundary currents discussed above (Vallis (2006)), so that the effect of WBCs is not resolved by the grid underlying the HadCM3 climate network (see Table 3.1). Using higher resolution SAT data taken from the AOGCM ECHAM5 developed by the Max Planck Institute for Meteorology in Hamburg (Meehl et al. (2007)), we find that our method does indeed seem to detect WBCs (Appx. D).

To exclude the possibility that the observed backbone structures over the ocean might be simply due to local anomalies in the SST-SAT gradient caused by surface currents, we have calculated the gradient field from the model run that we used to construct the HadCM3 climate network, and found that the SST-SAT gradient and betweenness are not correlated strongly (Fig. 5.3). Because of the questionable quality of measured oceanic data for the period of time considered, we did not attempt the corresponding analysis for reanalysis data.

Furthermore, the backbone is neither seen in fields of degree nor closeness centrality (Sect. 3.3) and no clear statistical relationship between these centrality measures and betweenness

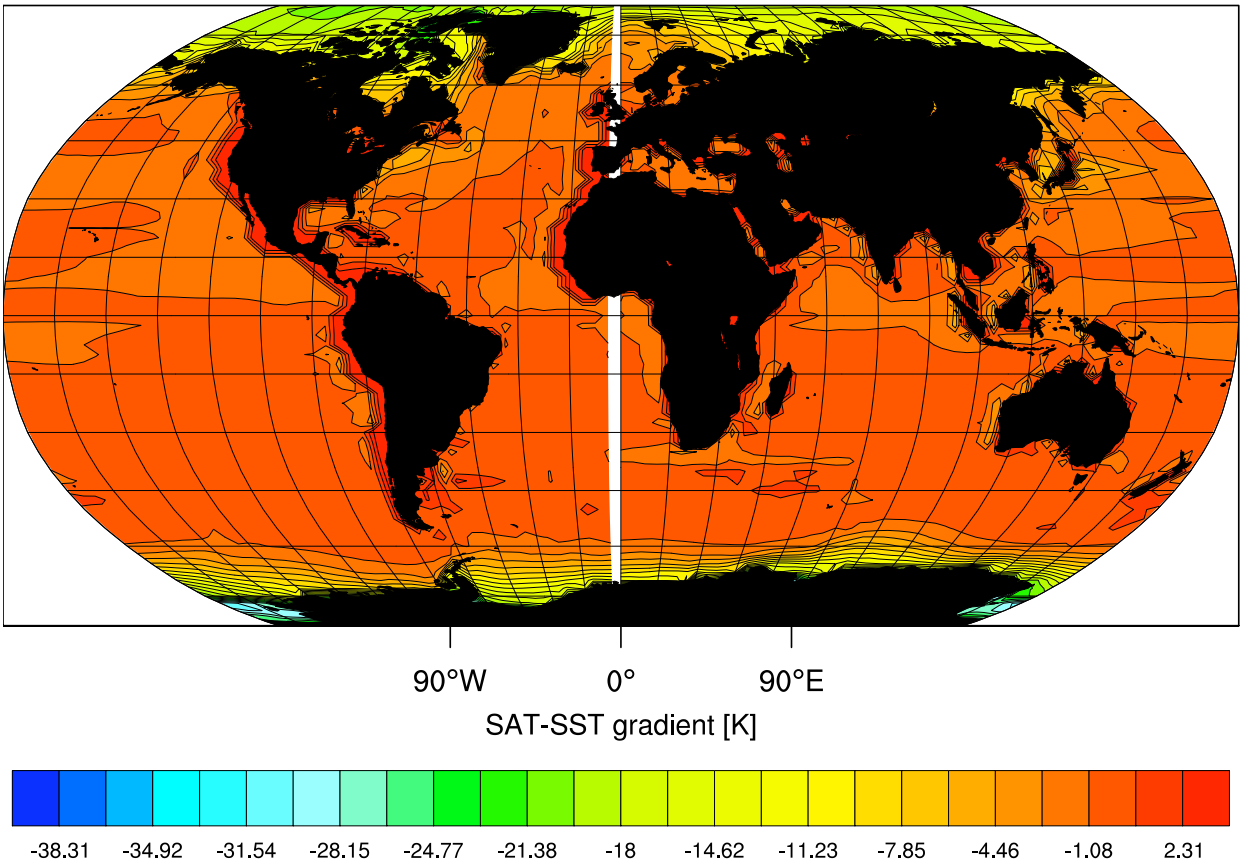
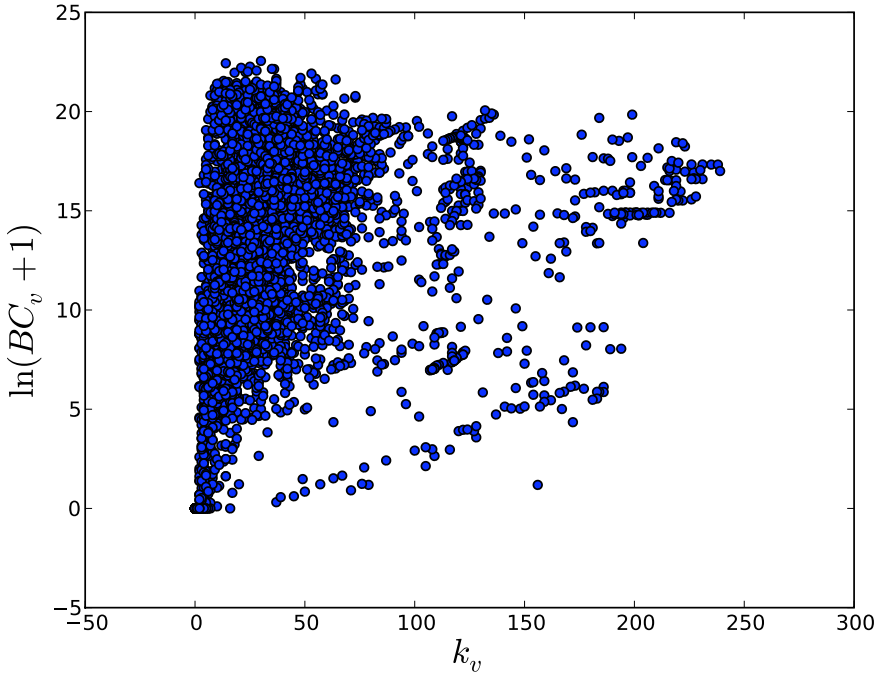


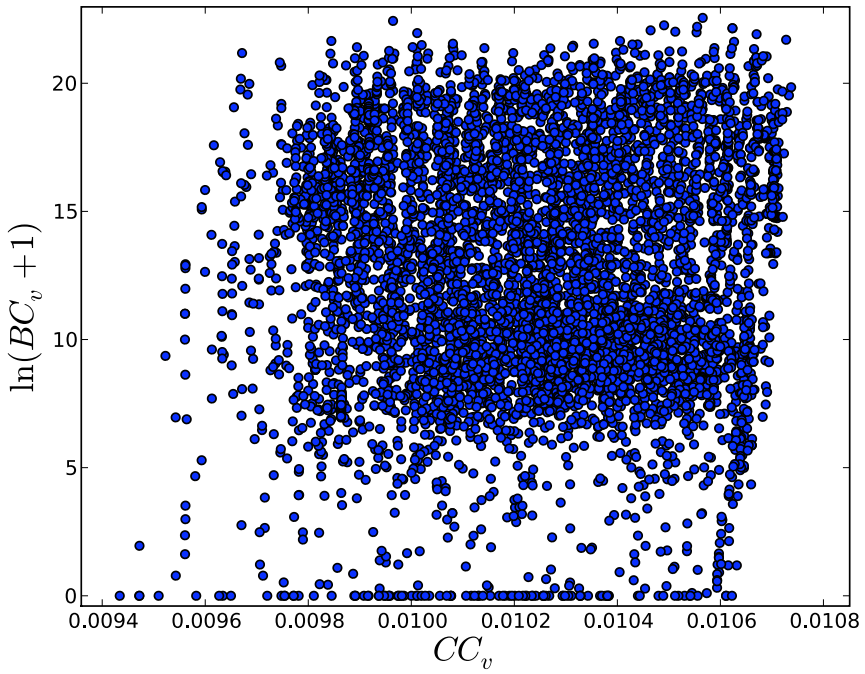
Figure 5.3 The mean SAT-SST gradient field $\langle \Delta T_v(t) \rangle_t = \langle SAT_v(t) \rangle_t - \langle SST_v(t) \rangle_t$ calculated from the HadCM3 SAT and SST data sets (Meehl et al. (2007)), both taken from the 20th century reference run described in Sect. 3.1.1. $\langle \Delta T_v(t) \rangle_t$ shows no structure remotely resembling the backbone, but we can interpret tongues of negative SAT-SST gradient, *e.g.*, off the east coast of Canada and in the Norwegian Sea, as traces of surface ocean currents carrying warm water, particularly the North Atlantic Current (or North Atlantic Drift). The mean SST field $\langle SST_v(t) \rangle_t$ was interpolated to match the grid of the SAT data set for this analysis. Note that the anomalously large positive gradient along the land-sea interface is an artifact of the interpolation and hence does not have a physical meaning.

can be detected (Fig. 5.4). Therefore we conclude that the backbone structures observed in model and reanalysis networks are neither a trivial response to local anomalies in the SST-SAT gradient nor artifacts of chains of super-nodes with high degree and closeness centrality. In contrast, the vertices of highest betweenness are found in the range $0 \leq k_v \lesssim 50$ of low degree centrality (Fig. 5.4(a)). This indicates that not the super-nodes, but inconspicuous vertices with a comparably small number of neighbors are most essential for the efficient information transport in the climate network and could be referred to as *information hubs*. Comparing Fig. 5.4(a) with Fig. 5.1(b) we can see that the vertices forming the backbone are information hubs, *i.e.*, have a relatively low degree centrality. The remarkable presence

of information hubs is a hint to the existence of a community structure in the SAT climate network (Appx. A), since it is plausible that low degree hubs topologically lying between strongly connected communities should have a high betweenness centrality (Sect. 3.3.3).



(a)



(b)

Figure 5.4 Scatter plots of betweenness BC_v against degree k_v and closeness CC_v for the HadCM3 SAT Pearson correlation climate network at $\rho = 0.005$, where betweenness is plotted on a logarithmic scale. There is no obvious apparent statistical relationship that could be read off the scatter plots. Specifically, the Spearman's rho of the centrality fields are $r_s(k_v,BC_v) = 0.4594$ and $r_s(CC_v,BC_v) = 0.0069$.

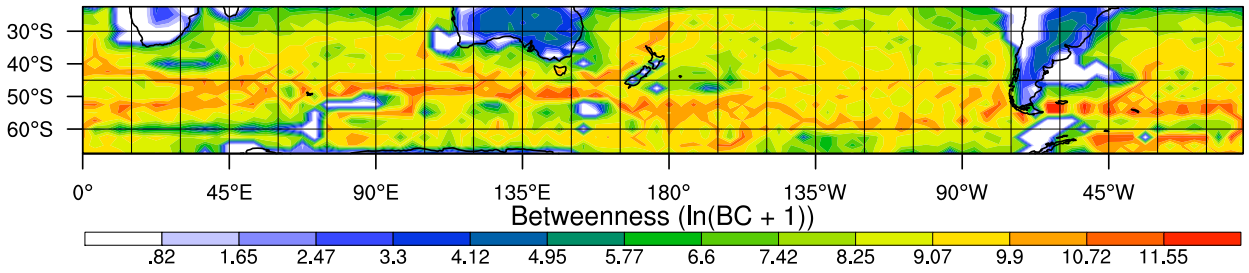
5.2. Physical interpretation of betweenness

When seeking a physical interpretation of betweenness in the context of climate networks, we first have to clarify what we mean by *information* and *information flow*. Consider two regions, the dynamics of which are found to be significantly statistically interrelated. In order for region i to "know" what region j is doing, there has to be some physical mechanism for the transfer of dynamical information in the climatological field. We can think of this information flow as the spread of perturbations in a spatiotemporal system (Vastano and Swinney (1988)). Now one can ask, how a perturbation of the dynamics of i would spread in the complex network until it reaches j. Depending on the physical mechanisms involved it could preferentially travel on topologically shortest paths, that due to our method network construction correspond to the most direct connection between i and j in terms of average dynamical interrelationships. It is plausible that this optimal mode of information transport is approximately realized by advective processes in the climate system, *e.g.*, ocean currents or jet streams in the atmosphere. In contrast, diffusive modes of the spread of perturbations can be envisioned as randomly walking packages of dynamical information on the complex network. When studying global climate networks, turbulence is the most likely candidate for a diffusive physical process promoting the spread of perturbations, while molecular diffusion is negligible (Vallis (2006)).

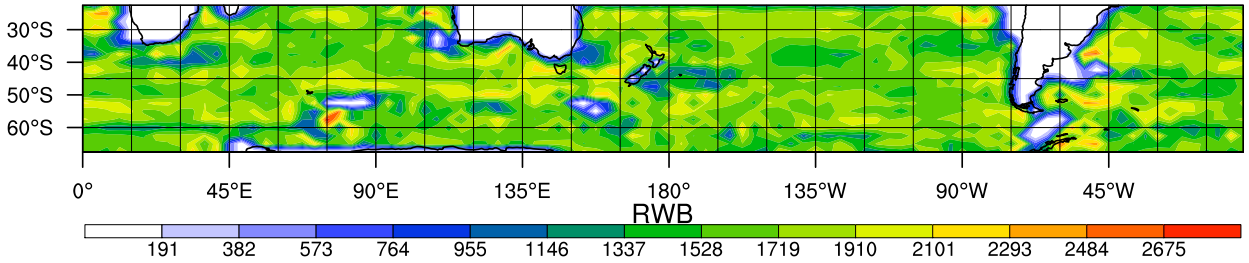
Following this reasoning it is plausible that the shortest path betweenness field BC_v (Eq. 2.14) displays backbone structures related to advective physical processes, *i.e.*, surface ocean currents, since BC_v is a measure of information flow centrality. While it is thus reasonable to regard shortest path betweenness as a measure of advective information flow centrality, random walk betweenness (Sect. 2.8) quantifies the importance of a region v for diffusive information flow in the climate network. In other words, shortest path betweenness assumes a local knowledge of the global network topology[1], whereas random walk betweenness just presumes that the local topology, *i.e.*, the neighbors, are locally known at vertex i. Ergo, the two betweenness measures give complementary information and should both be studied and compared to yield a more comprehensive picture of the spatial distribution of information flow within a climate network.

Because the computational complexity of our algorithm to calculate random walk betweenness scales as $\mathcal{O}(N^4)$, whereas Newman's shortest path betweenness algorithm scales

1 This assumption need not lead to unphysical results, since similar situations arise in many physical systems obeying some extremalization principle. For example, the refraction of light on the interface between two transparent media can be described by minimizing the time needed for the light to travel from medium A to medium B. Of course this does not imply, that a single photon locally 'knows' the optimal trajectory of minimal travel time and follows it. Even though the physical system as a whole behaves as if this was the case, we cannot project the globally appropriate extremalization principle to its local constituents. It is hence justified to use shortest path betweenness as a measure of information flow mediated by physical processes in a climate network, that do not violate the locality principle of classical physics.



(a)



(b)

Figure 5.5 Comparison of betweenness measures for a HadCM3 SAT Pearson correlation climate network at $\rho = 0.005$. The network is geographically restricted to the mid-latitudes of the southern hemisphere with $-66.6° \leq \lambda \leq -23.5°$ and $N = 1824$. (a) shows the shortest path betweenness field, while (b) displays the spatial distribution of random walk betweenness. Note the high shortest path betweenness band in the Southern Ocean resembling the Antarctic Circumpolar Current, that is not visible in the random walk betweenness field. This points to an advective physical process involved in the formation of the shortest path betweenness band.

as $\mathcal{O}(LN)$, we were yet unable to provide a comparison of both measures for global climate networks. Therefore we present some results for regional climate networks here. Considering the mid-latitudes of the southern hemisphere, we observe a band of high shortest path betweenness over the Southern Ocean resembling the Antarctic Circumpolar Current (Fig. 5.5(a)). That this structure is not seen in the random walk betweenness field gives further evidence, that an advective physical process such as a surface ocean current could be involved in its formation (Fig. 5.5(b)).

5.3. Significance testing

Here we present tests of the robustness of the observed backbone structures using twin surrogate networks on the time series level as well as the configuration model and geographical model I on the network level. While only the results considering the model SAT network are discussed below, we have performed the statistical tests for both the reanalysis and model SAT climate networks and came to the same conclusions.

5.3.1. Twin surrogate network ensemble

To test the statistical robustness on the time series level we develop the null hypothesis that the time series of the SAT data set are pairwise independent. Specifically, we generate 100 twin surrogates from the original time series at each grid point (Sect. 4.1.3), that were shown to allow for the most powerful tests of this null hypothesis among the types of surrogates considered here (Sect. 4.1.4). We then construct an ensemble of 100 networks from the surrogate data sets (Sect. 4.1.5), fixing the edge density of the original climate network, and compute the ensemble mean AWC field and ensemble mean betweenness field. While interestingly, the ensemble mean AWC field closely resembles the AWC field of the climate network (Fig. 5.6(a)), the ensemble mean betweenness field is again highly correlated to the ensemble mean AWC field and contains no backbone structures (Fig. 5.6(b)). The corresponding Z-score is $Z(r_s(k_v,BC_v)) \approx -422$. This highlights that while the network ensemble generated from twin surrogate data sets is able to explain the local network topology to some degree, it does not account for the delicate structures on the global topological scale. Based on these observations, we reject the null hypothesis that the time series of the SAT data set are pairwise independent and infer, that the backbone indeed characterizes the intrinsic complex topology of dynamical interrelationships.

5.3.2. Configuration model ensemble

To ensure the statistical robustness of our results on the network level, we test the null hypothesis, that the climate network is a random graph with a given degree field. Using the configuration model (Sect. 4.2), we generate a Monte Carlo ensemble of 100 surrogate networks, that have approximately the same degree field as the original climate network. We find that in sharp contrast to the original network, the ensemble mean betweenness field is highly correlated to the degree field, and does not display the backbone structures observed in the original climate network. More precisely, we introduce the Spearman's rho $r_s(k_v,BC_v)$ of the degree and betweenness fields as a network observable[1]. We obtain $\bar{r}_s(k_v,BC_v) = 0.4594$ for the original network and $r_s(k_v,BC_v) = 0.9813 \pm 0.0011$ from the configuration model ensemble. The corresponding Z-score $Z(r_s(k_v,BC_v)) \approx -474$ shows impressively that the low correlation observed in the original network cannot be explained by the configuration model ensemble. Calculating the Z-score field $Z(BC_v)$ of betweenness with respect to the configuration model ensemble gives a similar picture (Fig. 5.7). For most regions or vertices, $Z(BC_v)$ is so large that we need a logarithmic color scale to visualize it properly. The betweenness structures have a particularly large Z-score and can hence be seen clearly in the Z-score field. Based on this evidence we reject the null hypothesis that the climate network is random under the constraint of a given degree field and conclude,

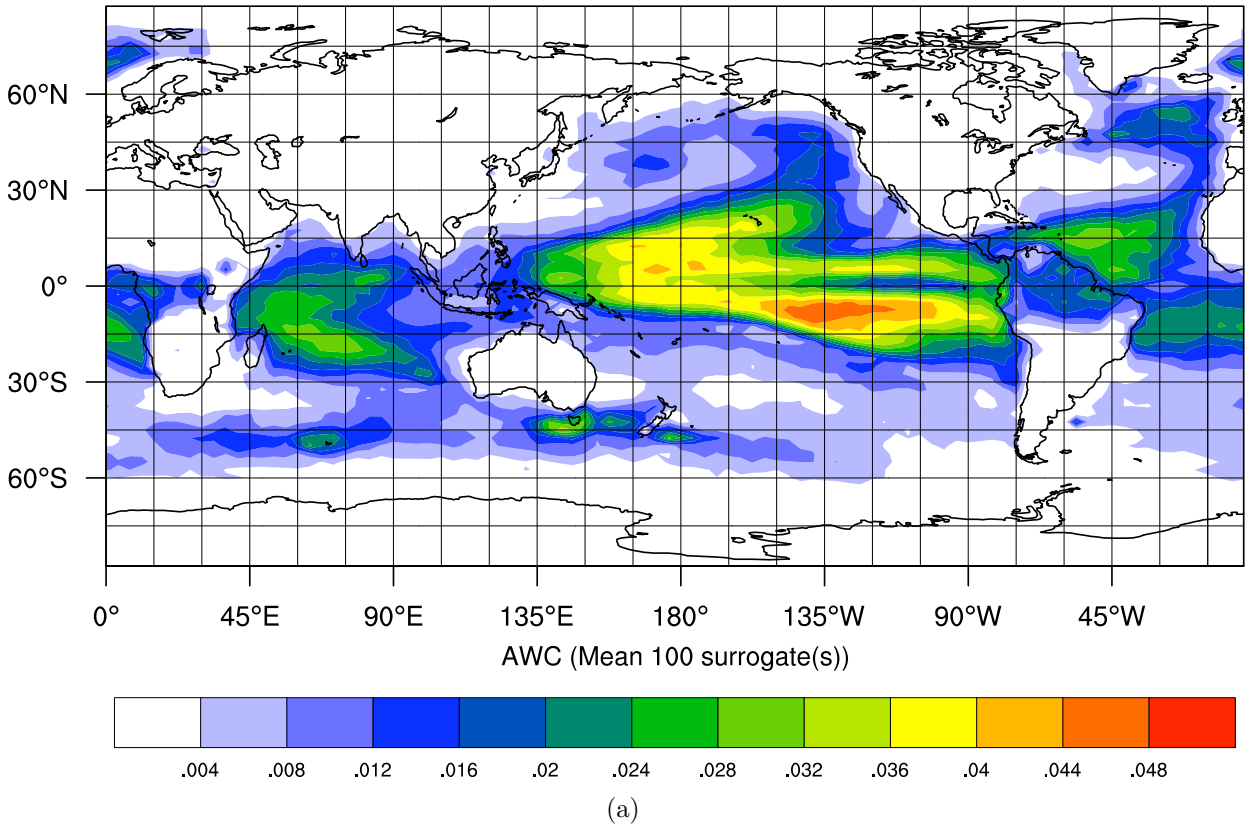
1 Again we choose the Spearman's Rho instead of the Pearson correlation coefficient because of its robustness with respect to the non-normal PDF's of k_v and BC_v generally found in our climate networks (Sect. 3.3).

that the backbone is unlikely to be a trivial consequence of the degree field.

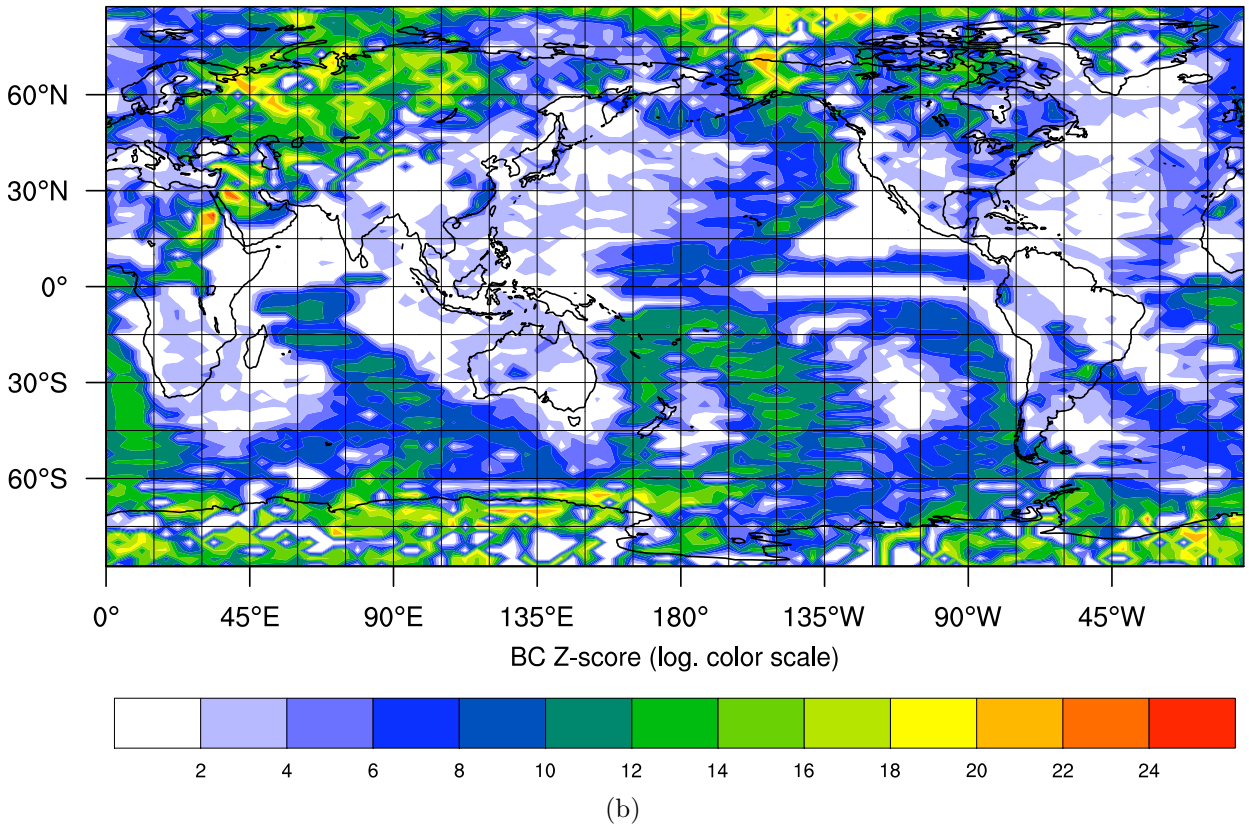
5.3.3. Geographical model I ensemble

On the network level, we can furthermore test the null hypothesis that the HadCM3 Pearson correlation climate network is random with a prescribed degree field and edge distance distribution using geographical model I (Sect. 4.2.4.1). Because of the high computational cost of rewiring $n_r > L$ times for a sufficient number of realizations of geographical model I, we demonstrate this test for a regional network encompassing the mid-latitudes of the southern hemisphere (Fig. 5.8). We find that even though appearing to be blurred, the ensemble mean betweenness field (Fig. 5.8(a)) still contains structures resembling those observed in the original betweenness field (Fig. 5.5(a)). The betweenness Z-scores (Fig. 5.8(b)) are seen to be smaller with respect to the dynamic range of the original betweenness field as compared to the betweenness Z-scores calculated from a configuration model for the global HadCM3 SAT climate network (Fig. 5.7). These findings indicate that some aspects of the backbone structures found in the betweenness field might be explained by the simplest spatial statistics of the climate network, *i.e.*, the edge distance distribution $p_E(l)$ together with the degree field k_v. Building on these preliminary results, much more work is needed to shed light on the connections between the characteristics of the spatial embedding and global path based measures of centrality.

After the $n_r = 3L$ rewiring steps performed for each realization here, average path length $\mathcal{L}$ and clustering coefficient $\mathcal{C}$ are found to be already very close to their maximally random limiting values (Fig. 5.5). Nevertheless we cannot guarantee that the network property of interest, *i.e.*, the betweenness field, has already converged to its hypothetical maximally random state (Zamora-López (2008)). Therefore it is highly desirable to in the future parallelize the surrogate generation and analysis code for an efficient generation of geographical model I and II as well as random link switching network surrogates with a larger number of rewiring steps $n_r \gg L$ and a lower tolerance parameter ε. Furthermore theoretical considerations leading to rigorous lower bounds for n_r would be very helpful.



(a)



(b)

Figure 5.6 Statistics of an ensemble of $n = 100$ Pearson correlation climate networks constructed from twin surrogate data sets ($m = 1$, $\tau = 0$, $\delta = 0.1$ and $\Delta t = 7$) of the HadCM3 SAT data set at $\rho = 0.005$. (a) The ensemble mean AWC field resembling the original AWC field (Fig. 3.5(a)). (b) The Z-score field $Z(BC_v)$ of betweenness, where the backbone is again seen to be highly significant.

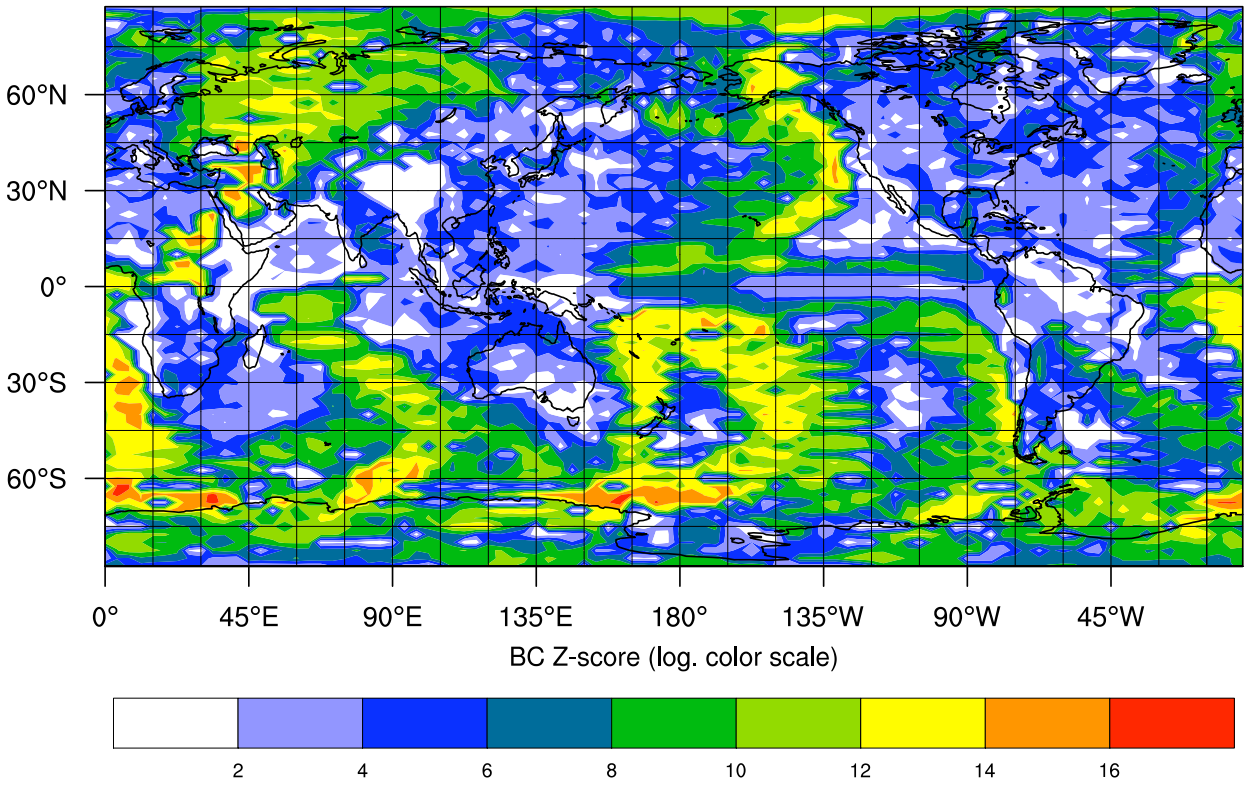
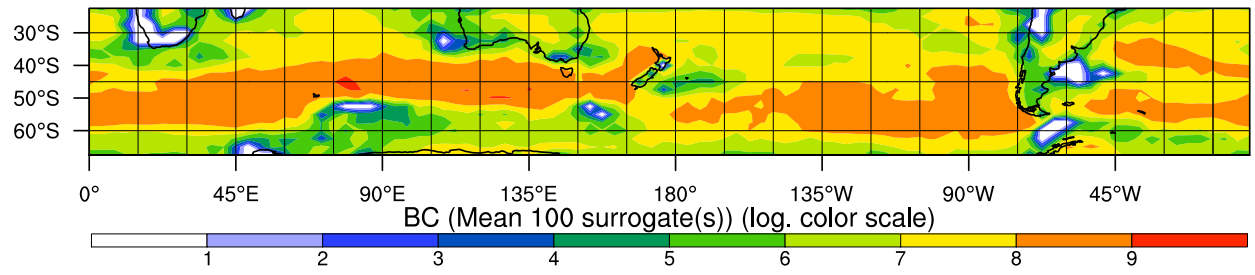
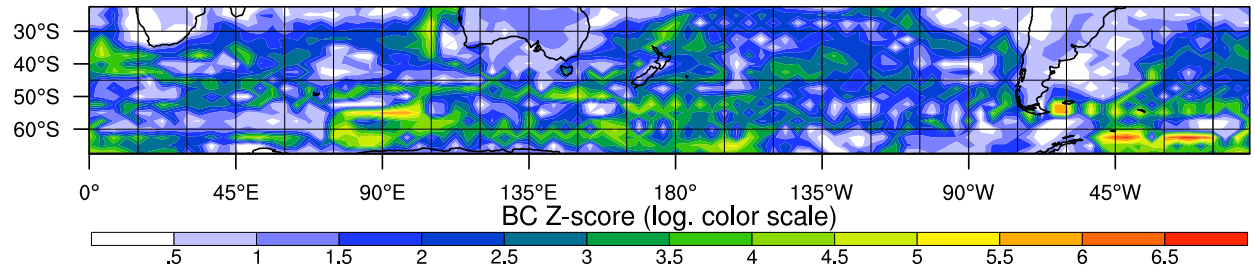


Figure 5.7 Z-score field $Z(BC_v)$ for the betweenness field with respect to a configuration model ensemble with $n = 100$ members calculated for the HadCM3 SAT Pearson correlation climate network at $\rho = 0.005$. A logarithmic color scale is needed to properly visualize the pronounced deviations of the ensemble betweenness field from that of the original network. The backbone structures are clearly recognizable with a large Z-score, indicating a high statistical significance with respect to the configuration model ensemble.



(a)



(b)

Figure 5.8 Ensemble mean and Z-score of the betweenness field with respect to 100 realizations of geographical model I ($n_r/L = 3$, $\varepsilon = 0.1$) for a HadCM3 SAT Pearson correlation climate network at $\rho = 0.005$. The network is geographically restricted to the mid-latitudes of the southern hemisphere. Compare to the betweenness field calculated from the original climate network (Fig. 5.5(a)).

5.4. Summary

In summary, using mutual information from nonlinear time series analysis and betweenness from complex network theory, we have uncovered novel pathways of global information flow in the climate system, that we call the backbone of the climate network.Several conceptually independent types of tests reveal that the backbone does not arise by chance and is not a trivial consequence of the degree centrality field studied in previous works on climate networks (Tsonis and Roebber (2004), Tsonis et al. (2006), Tsonis et al. (2008b), Tsonis and Swanson (2008)), but on the contrary represents a statistically significant feature of the underlying SAT data set. Surface ocean currents appear to play a major role in the information transfer and hence in the dynamical stabilization of the climate system in the long term mean (140 years for the HadCM3 model run and 60 years for the reanalysis data). We observe similar backbone structures in AOGCM model output and reanalysis data giving confidence that the backbone is not a model artifact.

It is important to realize that our complex network approach is an essential ingredient in the discovery of the backbone. The main advantage of betweenness is that it takes into account the global network topology of pairwise interrelationships between regions. However, the classical linear methods (*e.g.*, PCA, SSA, see von Storch and Zwiers (1999)) widely applied to disclose teleconnection patterns in climatology use information about next neighbors at each grid point, and are therefore only local within the complex network framework.

Our method is promising to next study the impact of extreme events such as strong El Ninos (Appx. C), extreme Monsoons or volcanic eruptions on the topology of climate networks. In the future it will thereby allow us to obtain new insights into the individual local signature of changes in the information flow structure and stability of the climate system. Our method may also be valuable to illuminate differences in the backbone structure in different climate states of earth's history, *e.g.*, holocene, glacial and cretaceous, and to assess the impact of global warming on the stability of the climate system from a different perspective.

Until now there is to our best knowledge no other method, that is able to extract the localized structure of (dynamical) information flow in spatially extended systems from time series data alone. Our method offers a qualitatively new level of understanding the dynamics of complex spatially extended systems, because by relying on paths in the network, betweenness subsumes information on the global network topology of physical interactions in a locally defined flow measure. The methodology presented in this chapter can hence be considered to be *universally* valid for all spatially extended dynamical systems, offering intriguing prospects for future research in other fields of physics.

CHAPTER 6

Seasonal and monsoon climate networks

As was already demonstrated in the previous chapter, climate networks can be used to uncover interesting structure in climate dynamics by studying spatial and temporal subsets of the complete global data set of some climatological observable. For example, Tsonis and Swanson (2008) studied the impact of ENSO on global teleconnection patterns by constructing two global climate networks from a monthly averaged surface air temperature reanalysis data set: An El Niño network using concatenated data from El Niño months and likewise a La Niña network from La Niña months. Yamasaki et al. (2008) and Gozolchiani et al. (2008) in turn studied regional climate networks constructed from complete time series to investigate the impact of ENSO on the number of edges or equivalently the edge density of these regional networks. In the context of this chapter it is important to remind the reader that Tsonis *et al.* have used only data from winter months in all of their studies (Tsonis and Roebber (2004), Tsonis et al. (2006), Tsonis et al. (2008b), Tsonis and Swanson (2008)) while we processed the complete time series containing data from all months in all analyses presented above (Sect. 3.1).

After briefly describing the methodology (Sect. 6.1), we present some interesting results on the seasonal and monsoon variability of the spatial characteristics of a regional SAT climate network encompassing the Indian Ocean basin (Sect. 6.2) and significance tests on the time series and network level (Sect. 6.3). Selected results are displayed for both model and reanalysis data. Table 6.1 summarizes information on the regional climate networks used in this study.

6.1. Methodology

A *seasonal climate network* G_S is constructed by calculating the correlation measure C_{ij} for a temporally ordered subset of the two anomaly time series $\hat{a}_i(t)$ and $\hat{a}_j(t)$ involved, *i.e.*, by concatenating all samples that belong to one of the four seasons, $S =$

Table 6.1 Properties of regional Indian Ocean data set used for generating seasonal and monsoon climate networks.

	Indian Ocean basin
$(\lambda_{min},\lambda_{max})$ [°]	(−45,45)
(ϕ_{min},ϕ_{max}) [°]	(30,140)
$N_{NCEP/NCAR}$	1665
N_{HadCM3}	1110

$\{MAM, JJA, SON, DJF\}$[1].

Likewise, we generate a *monsoon climate network* G_m by concatenating data from the summer months June, July and August, when the Indian summer monsoon is strongest (Wang (2006)). The non-monsoon state is characterized by the *non-monsoon climate network* $G_{\bar{m}}$ constructed from concatenated anomaly time series of the winter months December, January and February.

To highlight the localized impact of the monsoon circulation on the spatial structure of regional teleconnections within the Indian Ocean basin SAT climate network, we introduce the notion of *exclusive climate networks*. Let $\mathbf{A}_m$ and $\mathbf{A}_{\bar{m}}$ be the adjacency matrices of monsoon and non-monsoon climate networks, respectively. The *exclusive monsoon climate network* contains only edges that are present in the monsoon network, but not in the non-monsoon network. Its adjacency matrix $\mathbf{A}_m^{excl.}$ is given by

$$\mathbf{A}_m^{excl.} = \mathbf{A}_m - \Theta\left(\mathbf{A}_m + \mathbf{A}_{\bar{m}} - 1\right), \tag{6.1}$$

where $\Theta(.)$ denotes the Heaviside function (to be applied elementwise) and the second term on the right side of Eq. 6.1, $\Theta\left(\mathbf{A}_m + \mathbf{A}_{\bar{m}} - 1\right)$, can be interpreted as the adjacency matrix of the network of coinciding edges. Similarly, the *exclusive non-monsoon climate network's* adjacency matrix $\mathbf{A}_{\bar{m}}^{excl.}$ is obtained by replacing $m \to \bar{m}$ in Eq. 6.1.

6.2. Results

6.2.1. Seasonal climate networks

Comparing the degree and intrinsic edge distance distributions of seasonal climate networks constructed from AOGCM and reanalysis SAT data, some pronounced deviations appear (Fig. 6.1). Most notably, the summer HadCM3 network possesses more high degree hubs and long range edges than the other seasonal HadCM3 networks. For the reanalysis data set

1 This corresponds to the standard definition of seasons in meteorology, *i.e.*, spring comprises all of March, April and May, summer all of June, July and August, fall all of September, October and November and winter all of December, January and February. Note that maintaining our eurocentric bias we refer to northern hemisphere summer as summer etc.

this observation cannot be confirmed in general, but the summer network has more high degree vertices and long range connections than the winter network. These deviations can be thought of to partly be an imprint of higher order effects of the annual cycle that have not been filtered out by phase averaging during the data preprocessing stage (Sect. 3.1).

Seasonal climate networks additionally show marked differences in their area weighted connectivity (Fig. 6.2) and average edge distance fields (Fig. 6.3), where we only show those for the HadCM3 SAT data set here. The super-nodes present in the seasonal networks change their position, shape and intensity, *e.g.*, the super-node with long range connectivity over the Philippines that is only unambiguously present in the winter climate network.

6.2.2. Monsoon climate networks

To find traces of the monsoon circulation in our SAT climate networks, we compare the monsoon (summer) and non-monsoon (winter) state by constructing exclusive monsoon and non-monsoon networks and comparing their characteristics. The exclusive monsoon (summer) network is found to host remarkably more high degree vertices and long range edges than the exclusive non-monsoon (winter) network for both model and reanalysis data (Fig. 6.4).

Comparing the AWC and AED fields of exclusive monsoon and non-monsoon climate networks, the shifting super-nodes can be visualized very clearly. The HadCM3 exclusive monsoon network shows pronounced centers of action of high long range connectivity, *i.e.*, high AWC and AED, south of India and east of Africa (Fig. 6.5(a), 6.5(a)). While some of these features are also found in the AWC and AED fields of the corresponding exclusive non-monsoon network, new centers of action appear over Pakistan, the southern Indian Ocean, over the Philippines and south of Korea over the East China Sea. It is important to point out, that the centers of action observed in exclusive networks are build up entirely of edges that exist exclusively in either the monsoon or the non-monsoon state. Ergo, the edge density ρ of exclusive networks is generally smaller than that of the seasonal climate networks from which they are constructed (Table 6.2).

Comparing the NCEP/NCAR (Fig. 6.6) and HadCM3 (Fig. 6.5) exclusive networks, we observe a similar seasonal shift of the "center of mass" of super-nodes from the Northern (summer) to the the Southern Hemisphere (winter). While some more localized structures

Table 6.2 Edge densities of common and exclusive monsoon and non-monsoon climate networks for a HadCM3 SAT Pearson correlation data set encompassing the Indian Ocean basin. The seasonal network used for their construction were fixed at edge density $\rho = 0.01$.

ρ	HadCM3	NCEP/NCAR
Common	0.0051	0.0059
Excl. monsoon	0.0039	0.0034
Excl. non-monsoon	0.0039	0.0034

in the AWC and AED fields also coincide to some degree, *e.g.*, the center of action south of Korea in the exclusive non-monsoon networks, we observe some manifest deviation. For example, the NCEP/NCAR exclusive monsoon network possesses a strong super-node over the Bay of Bengal, that is not seen in the HadCM3 exclusive monsoon network. The marked localized deviations suggest that exclusive monsoon and non-monsoon climate networks might be used in the future to evaluate the performance of AOGCMs in capturing seasonal and monsoon variability in the tropics, but also other climate zones.

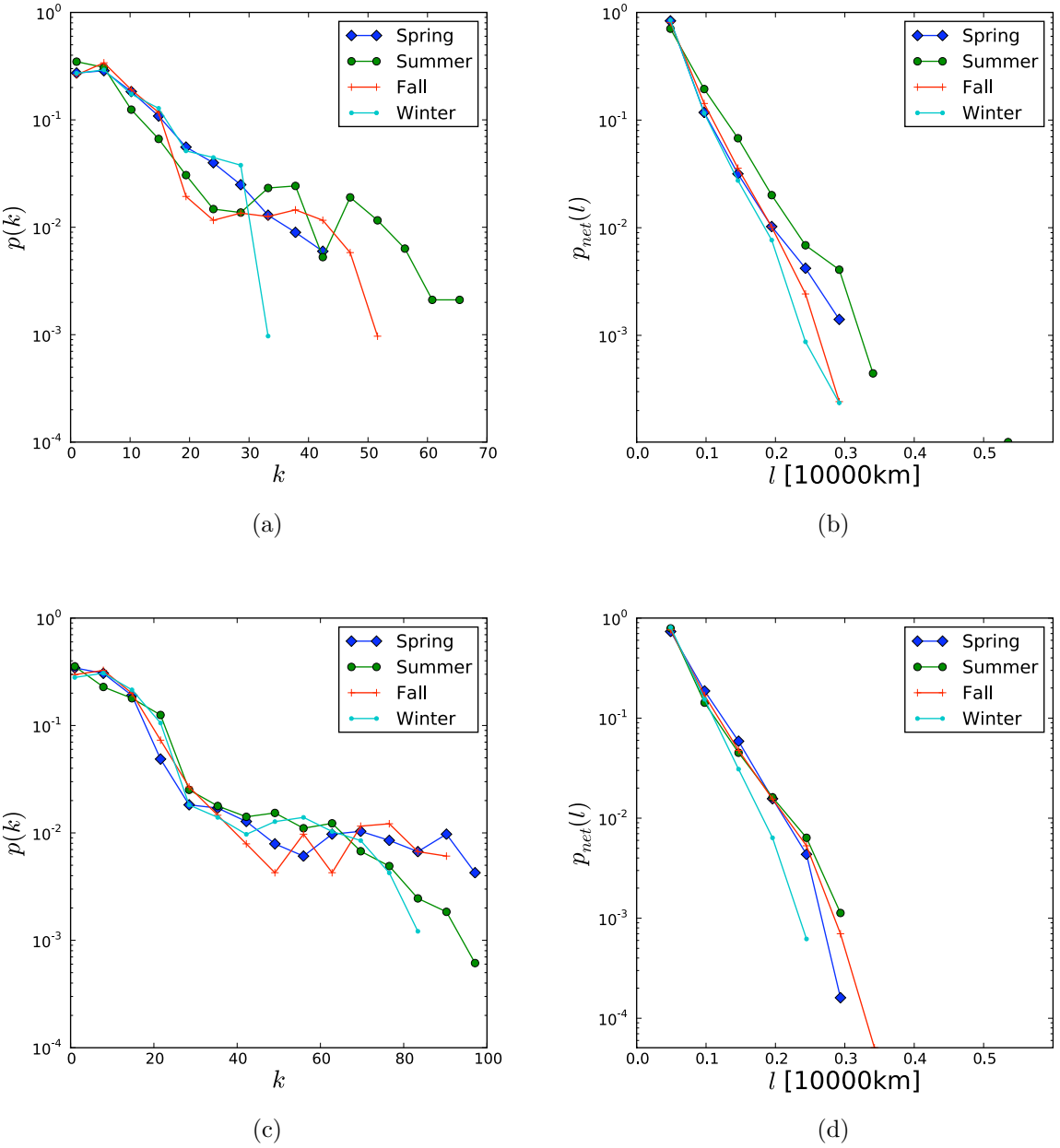
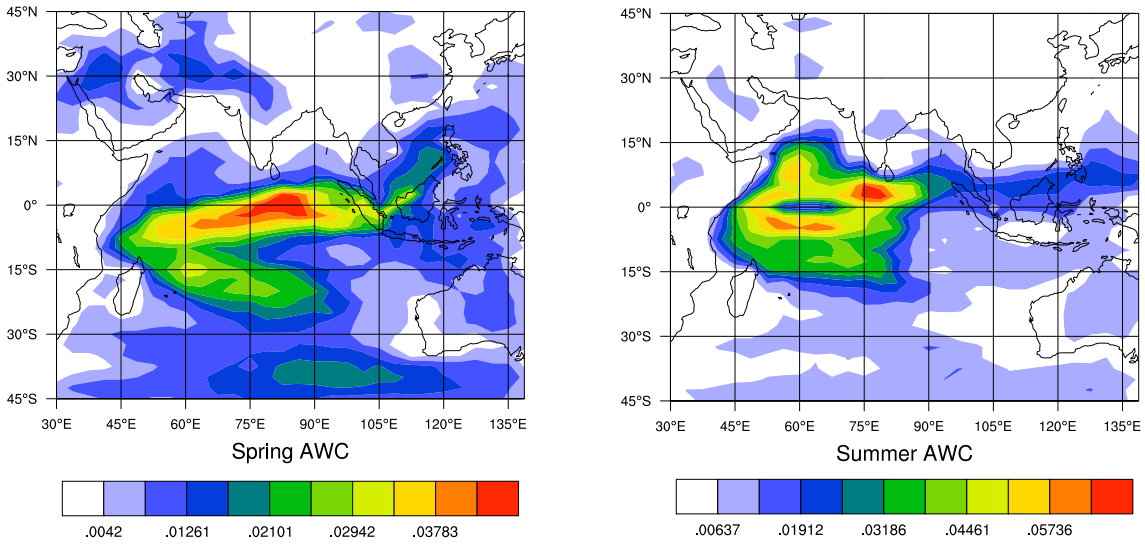
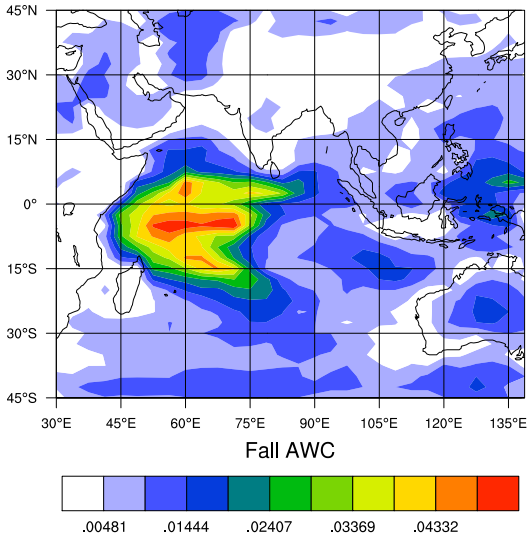


Figure 6.1 (a,c) Degree distribution $p(k)$ and (b,d) intrinsic edge distance distribution $p_{net}(l)$ for spring (blue diamonds), summer (green circles), fall (red plus signs) and winter (turquoise dots) Indian Ocean basin climate networks at $\rho = 0.01$ constructed from (a,b) HadCM3 and (c,d) SAT data sets using Pearson correlation. Considering the HadCM3 data set, the summer network is special in the sense that among all seasonal networks it contains the hubs of highest degree and a notably larger number of long range connections. Compare also the different color bar scales in Fig. 6.2 and Fig. 6.3.

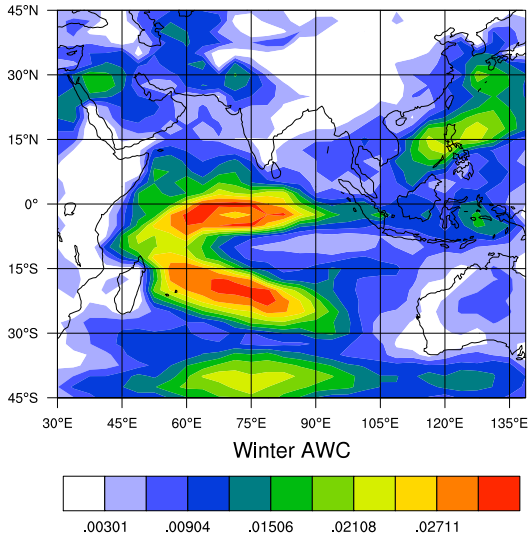


(a)

(b)



(c)



(d)

Figure 6.2 AWC fields for (a) spring, (b) summer, (c) fall and (d) winter HadCM3 SAT Pearson correlation climate networks at $\rho = 0.01$, encompassing the Indian Ocean basin. Note that the scale of color bars it not the same for all panels.

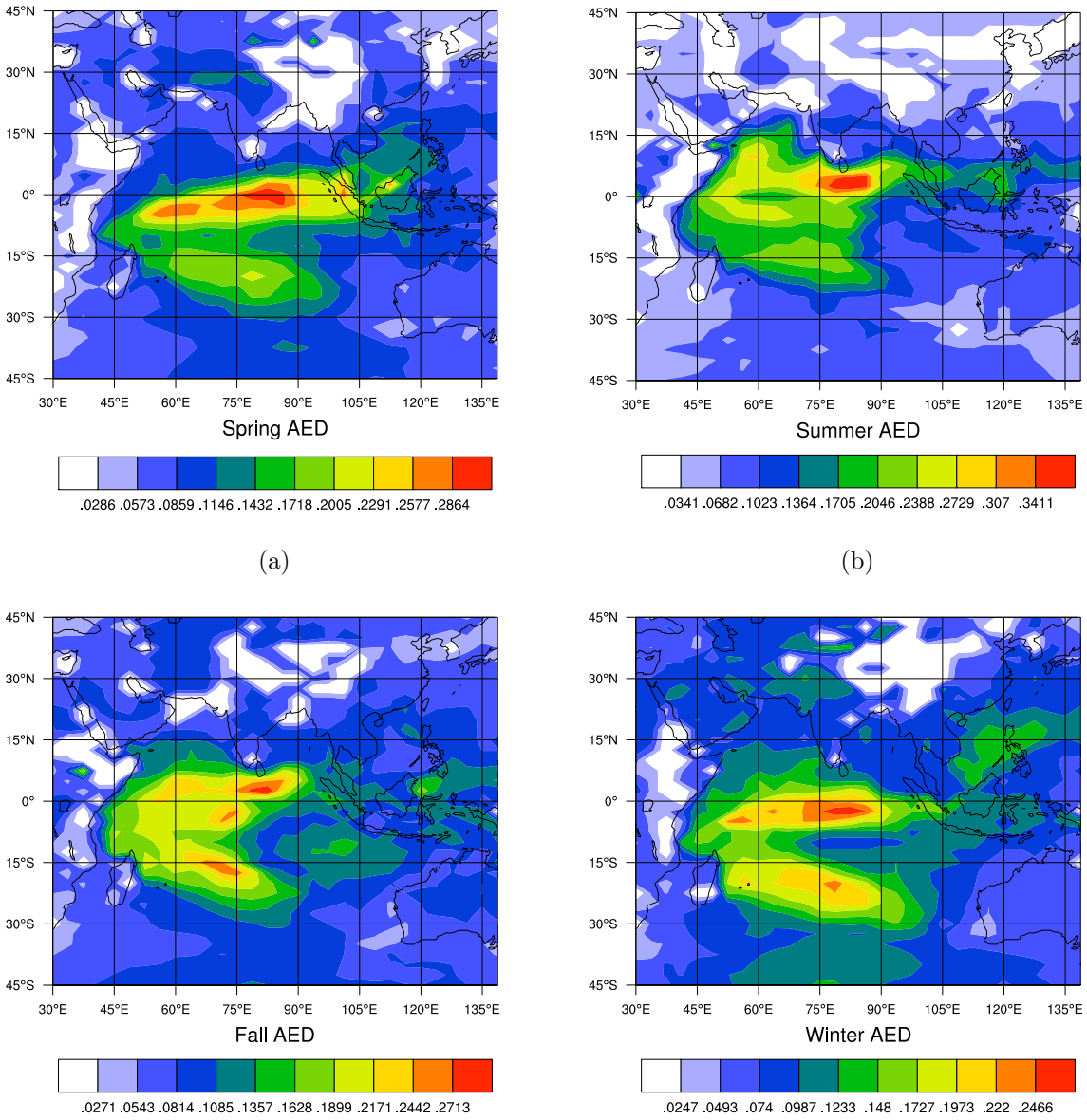


Figure 6.3 AED fields for (a) spring, (b) summer, (c) fall and (d) winter HadCM3 SAT Pearson correlation climate networks at $\rho = 0.01$, encompassing the Indian Ocean basin. Note that the scale of color bars it not the same for all panels.

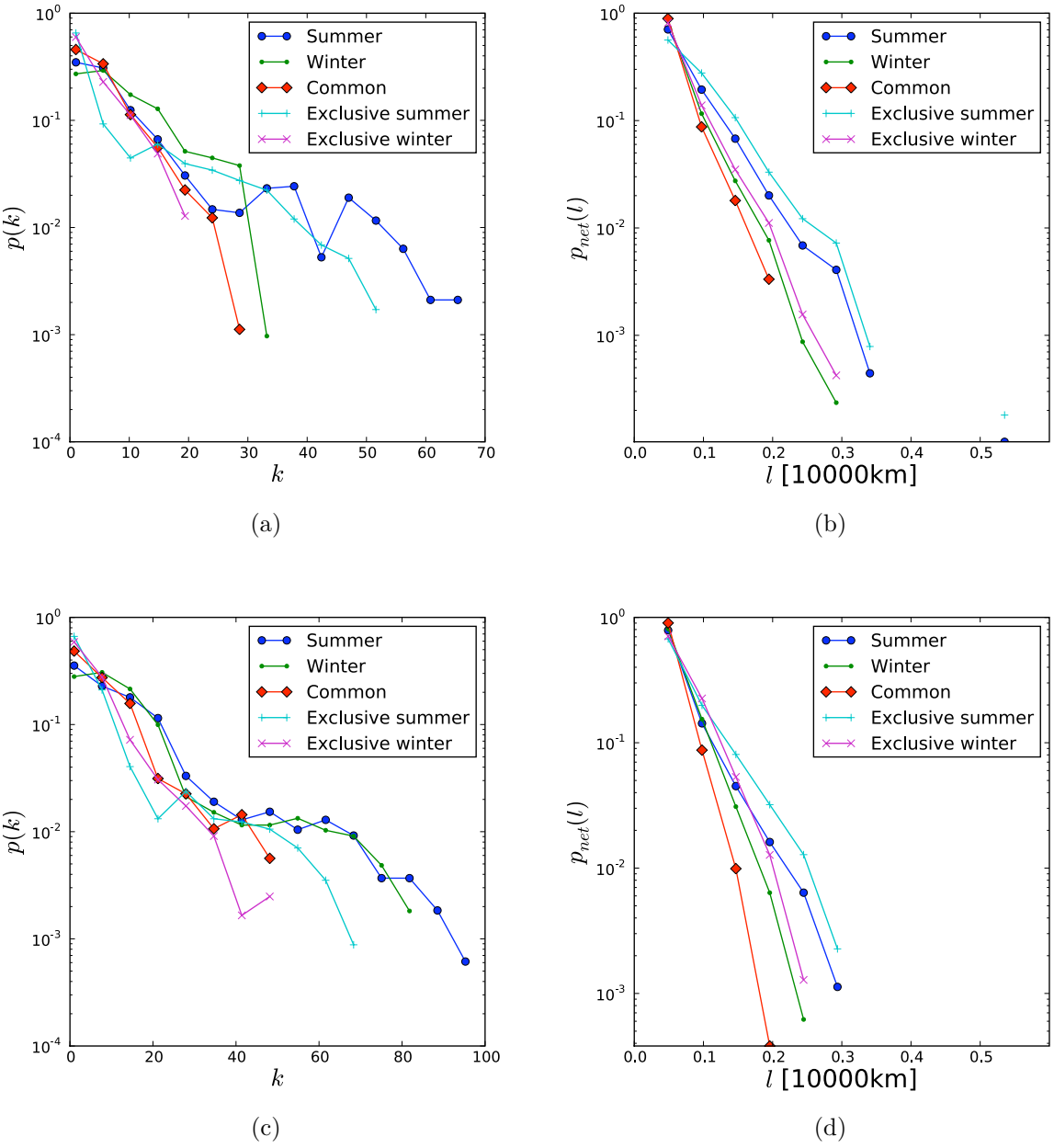


Figure 6.4 (a,c) Degree distribution $p(k)$ and (b,d) intrinsic edge distance distribution $p_{net}(l)$ for monsoon (blue circles), non-monsoon (green dots), common edges (red diamonds), exclusive monsoon (turquoise plus signs) and exclusive non-monsoon (lilac crosses) Indian Ocean basin climate networks constructed from (a,b) the HadCM3 and (c,d) the NCEP/NCAR SAT Pearson correlation matrix. The edge density of the monsoon and non-monsoon networks was fixed at $\rho = 0.01$, resulting in lower edge densities in the range of $0.003 \leq \rho \leq 0.006$ for the common and exclusive networks (Table 6.2). For AOGCM and reanalysis data, the monsoon (summer) climate network possess hubs of higher degree and more long range edges than the non-monsoon (winter) network.

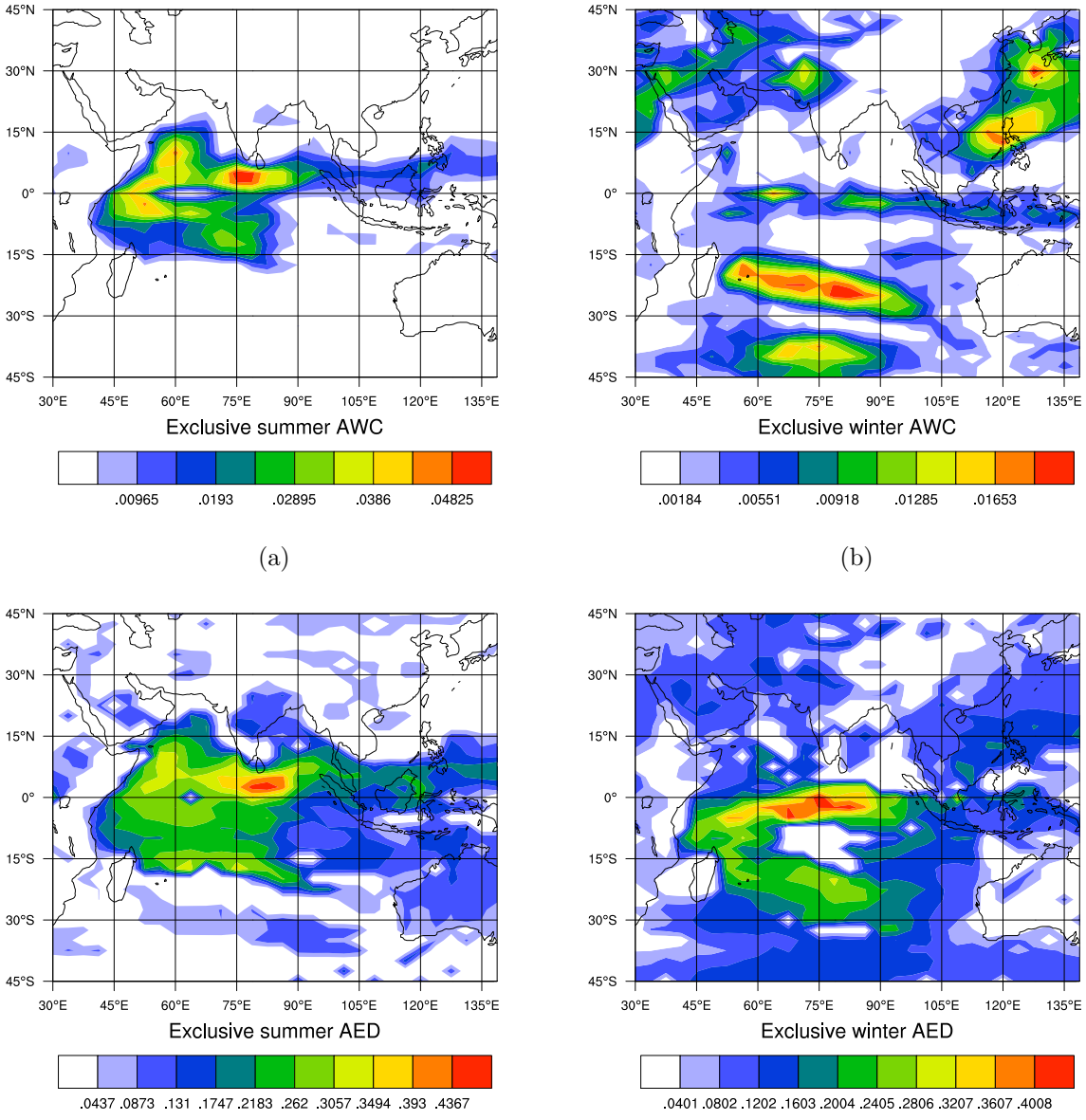
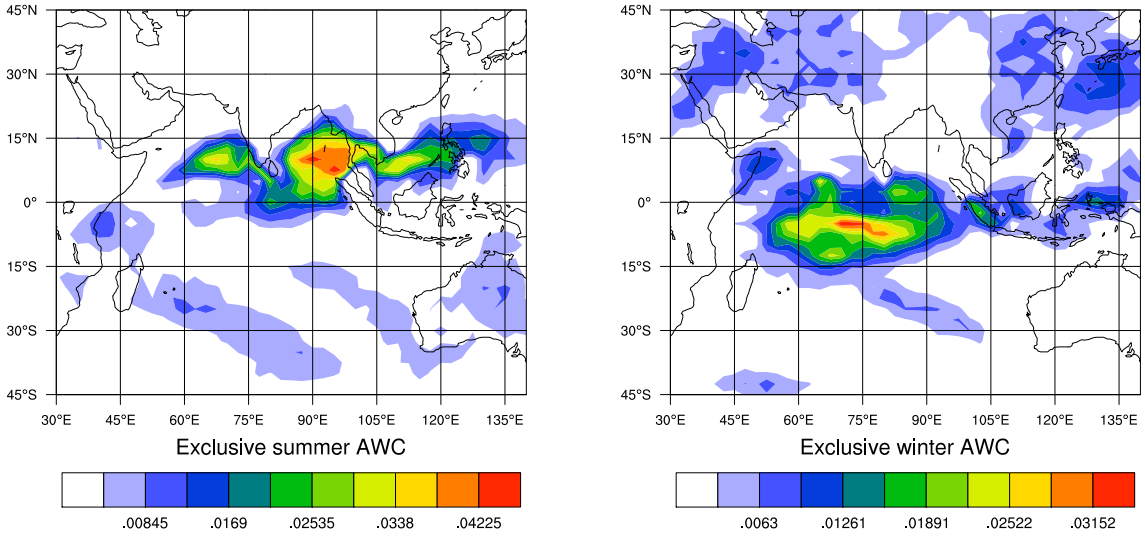
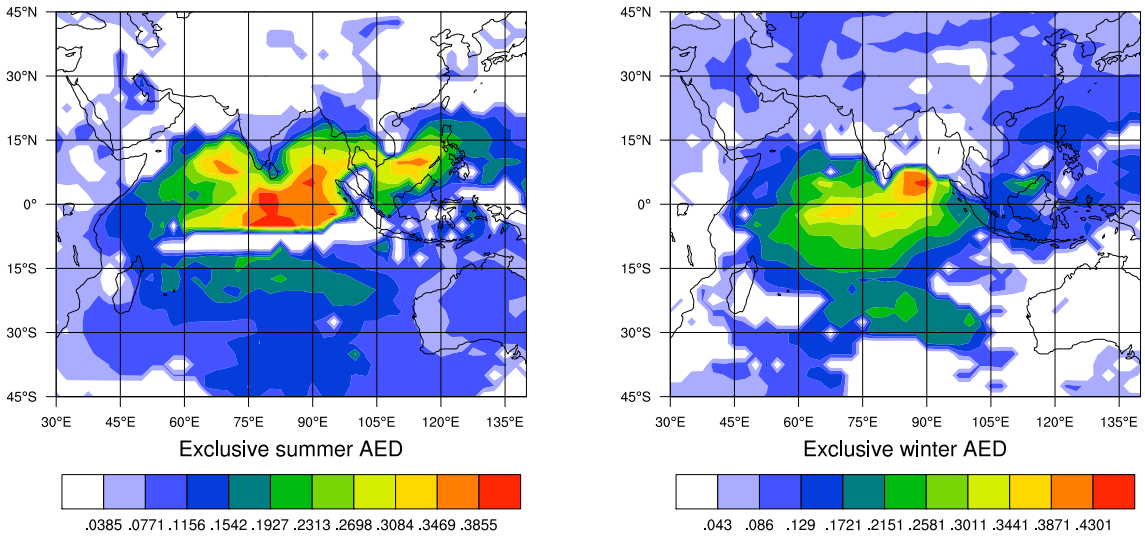


Figure 6.5 AWC fields for exclusive (a) monsoon and (b) non-monsoon networks as well as AED fields for exclusive (c) monsoon and (d) non-monsoon HadCM3 SAT Pearson correlation climate networks, encompassing the Indian Ocean basin. The edge density of the monsoon and non-monsoon networks was fixed at $\rho = 0.01$, resulting in lower edge densities in the range of $0.003 \leq \rho \leq 0.006$ for the common and exclusive networks (Table 6.2). Note that the scale of color bars it not the same for all panels.



(a)

(b)



(c)

(d)

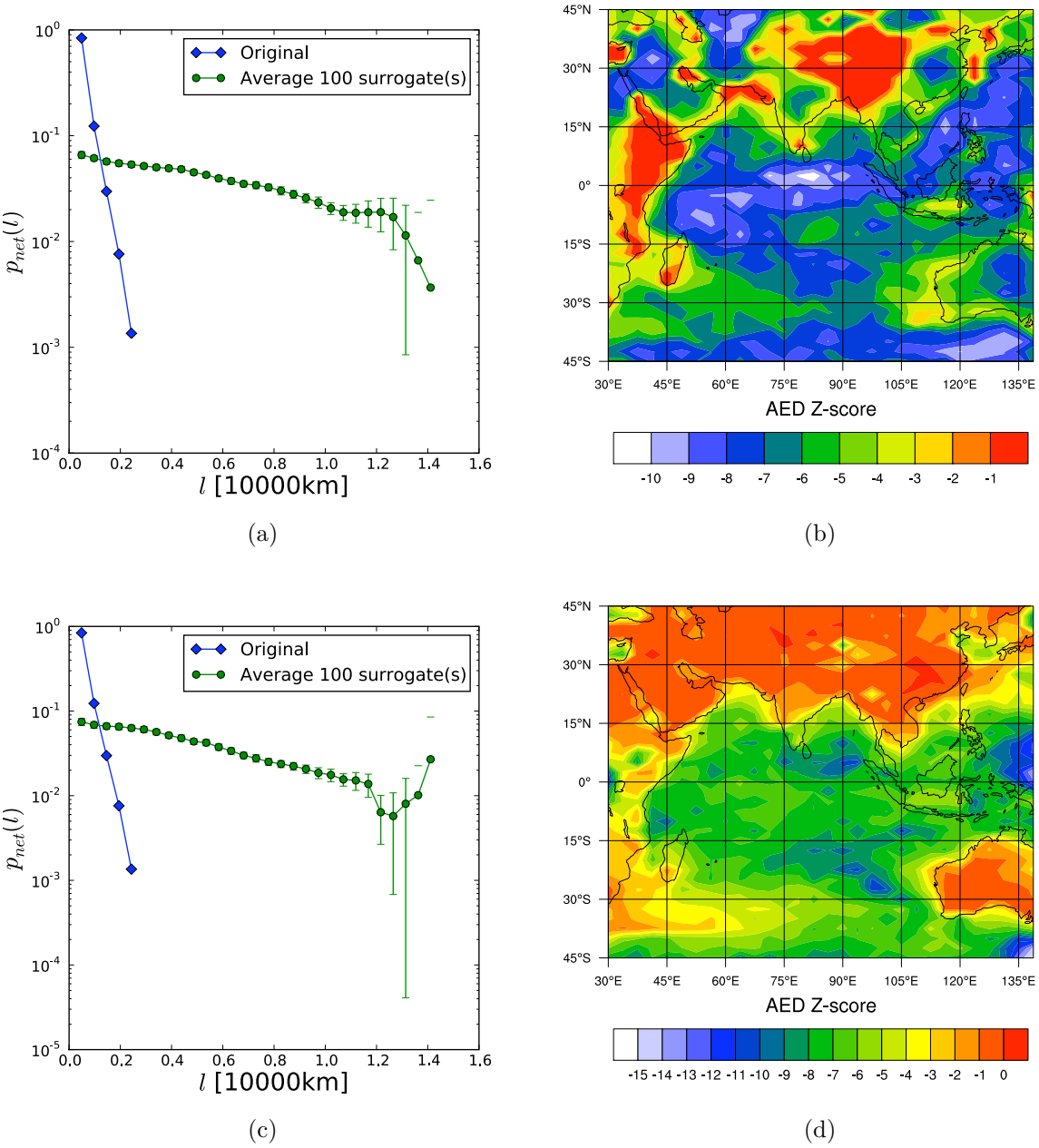
Figure 6.6 AWC fields for exclusive (a) monsoon and (b) non-monsoon networks as well as AED fields for exclusive (c) monsoon and (d) non-monsoon NCEP/NCAR SAT Pearson correlation climate networks, encompassing the Indian Ocean basin. The edge density of the monsoon and non-monsoon networks was fixed at $\rho = 0.01$, resulting in lower edge densities in the range of $0.003 \leq \rho \leq 0.006$ for the common and exclusive networks (Table 6.2). Note that the scale of color bars it not the same for all panels.

6.3. Significance tests

For brevity, we demonstrate significance tests of the spatial characteristics for the all year HadCM3 SAT Pearson correlation climate network encompassing the Indian Ocean basin (Chap. 4). To test if the spatial network properties observed in Sect. 6.2 can be explained by the degree field alone[1], we develop the null hypothesis that the regional Indian Ocean basin SAT network is random with a fixed degree field. Employing a configuration model network ensemble to test this null hypothesis, we find that the intrinsic edge distance distribution $p_{net}(l)$ (Fig. 6.7(a)) and average edge distance field AED_v (Fig. 6.7(b)) of the original climate network deviate significantly from those of the configuration model ensemble. The approximately exponential decay of $p_{net}(l)$ is much slower for the ensemble than for the original network. The Z-scores are particularly large over the ocean, indicating a higher significance of the observed AED patterns there. We can hence reject the null hypothesis that the network is random with a given degree field with high confidence.

By similar arguments, we can reject the null hypothesis that the anomaly time series of the underlying SAT data set are pairwise independent using a twin surrogate network ensemble (Fig. 6.7(c) and 6.7(d)). This null hypothesis corresponds to the assumption that the observed spatial properties can be explained by single time series properties such as the PDF, power spectrum and self mutual information alone.

1 The degree and area weighted connectivity fields are nearly equivalent here, since the Indian Ocean basin network is centered on the equator and the vertex density is nearly homogenous, *i.e.*, $\cos(\lambda)$ is comparable to one everywhere ($\cos(\lambda_{max} \approx 0.7)$).



(a) (b)

(c) (d)

Figure 6.7 Significance tests for (a,c) intrinsic edge distance distribution $p_{net}(l)$ and (b,d) average edge distance field AED_v for the HadCM3 SAT Pearson correlation climate network at $\rho = 0.01$ over the India Ocean basin. The ensemble mean $p_{net}(l)$ and AED Z-scores were obtained from 100 realizations of (a,b) the configuration model and (c,d) twin surrogate networks ($m = 1$, $\tau = 0$, $\delta = 0.1$ and $\Delta t = 7$). The error bars in (a,c) indicate the respective ensemble standard deviation for each bin.

6.4. Summary and outlook

In summary, we have presented a case study of the application of the climate network approach to regional and seasonal data sets. For a regional Indian Ocean basin climate network we have uncovered interesting seasonally changing structures in AOGCM and reanalysis surface air temperature data, some of which could be related to the monsoon circulation. Finally we have demonstrated how the network models introduced in Chap. 4 could be applied to test the significance of the well defined spatial structures that are observed in the original regional climate network.

It is desirable in the future to develop a well founded climatological interpretation of the observed spatial network structures. Further work is also needed to investigate the possible role of nonlinear physical processes involved in the monsoon circulation within the framework of climate network analysis, *e.g.*, by comparing mutual information and Pearson correlation climate networks (Sect. 3.3). Finally, we suggest that coupled climate networks building on coupled pattern analysis from climatology might be more suitable to study the monsoon dynamics (Bretherton et al. (1992)), because they allow to simultaneously the dynamical interrelationships between several climatological fields within a unified framework. For example, a 3-coupled regional network constructed from surface air temperature, surface air pressure and precipitation or air moisture content presents a promising candidate for studying monsoon dynamics.

CHAPTER 7

Conclusions and outlook

> *The universe is a much more intricate place than we can imagine. I often think our conscious minds will never encompass more than a tiny fraction of it all and that our comprehension of the Earth is no better than an eel's comprehension of the ocean in which it swims.*
>
> James Lovelock, "The Revenge of Gaia" (2006)

The work on this thesis commenced as an experimental undertaking. It was a priori not very clear which questions to ask and what to look for on the pursuit of a deeper understanding of climate system dynamics using time series analysis and complex network theory as the tools at hand. Especially during the early phase of this endeavor the resulting freedom allowed us to touch upon many ideas, some of which have found their way into the appendix in a condensed and concise format.

While iteratively refining our methods and theoretical insight, it became apparent that *complex networks in the climate system* must be understood on two levels of abstraction. First, we envision the intricate topology of physical interactions between the different components of the climate system, *e.g.*, hydrosphere, atmosphere, cryosphere and biosphere, mediated by processes such as currents, winds, waves and diffusion to form a *physical complex network*. This view does not imply the denial of the locality principle of classical physics, in contrast the physical climate network on the deepest level is formed by elementary local interactions following the basic laws of physics. As an illustrative analogy to this idea consider a complex electrical circuit, for example a high-end computer chip. In principle it is possible to describe its behavior drawing on the basic laws of quantum mechanics by calculating the time-dependent spatial probability density distribution of electrons in this elaborately crafted system of isolating, semi-conducting and conducting solids. While the solution of this elementary problem might be feasible in the remote future, delicately and most probably by relying on computing machines, it would not deepen the physical understanding of the chip as a whole. The network paradigm proves to provide a more natural description of chip dynamics and design principles, even though it is not set at the basic level of physical theory.

Yet usually nobody would claim that to view a computer chip as a network of interacting electronic building blocks is unphysical. Hence it is our conviction that it is legitimate to view the network paradigm as a valid physical description of the climate system, as it is complementary to the classical local description relying for example on the Navier-Stokes equations, thermodynamics and radiation balances, and has the potential to provide a new level of understanding.

This said, we secondly turn to the more pragmatic method of climate network analysis developed in this thesis. We think of the application of complex network theory to climate data analysis as a computational laboratory, where the various network measures introduced in Chap. 2 take the roles of scientific instruments. Each instrument is designed to reveal selected aspects of the climate system's complex dynamics that are not directly perceivable and quantifiable by the limited human senses, *e.g.*, by looking at a movie of the time evolution of the surface air temperature field we can neither "see" the underlying backbone of high dynamical information flow detected by the instrument betweenness centrality nor is it a priori obvious that surface ocean currents seem to act as the physical carrier of this flow. The climate network laboratory is hence crafted to trace the underlying physical climate network's footprints which lie implicitly hidden in multivariate time series data. Akin to physical scientific instruments, network measures are prone to systematic and random errors. We have tried to avoid systematic errors in network construction by carefully comparing different climate network construction techniques and particularly by providing several criteria for the choice of a reasonable threshold for the correlation measure matrix (Chap. 3 and paper II). The robustness of our results against the more subtle effects of systematic and random errors was evaluated using a hierarchy of statistical tests on the level of time series analysis and complex network theory, where we developed and utilized novel types of network surrogates incorporating spatial constraints (Chap. 4).

While our climate network laboratory produces highly interesting and robust measurements like the backbone structures discussed in Chap. 5, one has to be aware that it is difficult to directly relate our laboratory climate networks to the physical climate network described above[1]. In this context, the largest issues are presented by the limitations of the network construction method, *e.g.*, the transitivity problem and the missing power to discriminate between causal relationships and mere correlations. Furthermore, the interpretation of our laboratory climate networks is complicated by the fact, that their vertices and edges do not correspond to clearly defined physical entities obeying known laws, as is the case for the vertices (transistors, capacitors, inductors, resistors etc.) and edges (wires) of the computer chip network. Hence, climate networks should at this stage mainly be considered as a novel paradigm in climate data analysis, that has a great potential to yield new insights into climate system dynamics and particularly creates new perspectives for assessing the stability and vulnerability of the climate system.

1 A similar problem is faced when electroencephalogram (EEG) data is used for the study of brain dynamics

In this spirit, the major contribution of this work is the development of a method capable of extracting the localized structure of dynamical information flow within a climatological field from data alone, *e.g.*, the backbone of the surface air temperature network (Chap. 5 and paper I), employing a refined climate network construction method and particularly sensitive centrality measures on the global topological scale. While we mainly relied on the surface air temperature field as a training case, our methodology is valid for other climatological fields, *e.g.*, surface air pressure or moisture content as well. More generally, we emphasize that this methodology is *universal* in the sense that it can in principle be applied to study the information flow within any spatially extended dynamical system. Our results are hence of interest for a broad audience within the physics community and various applied fields. Possible fields of application include among others fluid dynamics (turbulence), plasma physics, biological physics (population models, neural networks, cell models). Furthermore, our method is equally relevant for experimental data as well as model simulations and hence introduces a novel perspective on model evaluation and data driven model building. Note that technical suggestions on how our methods could be extended and developed in future research are given in the summary sections of the respective chapters.

Our work is timely in the context of the current debate on climate change within the scientific community and hence of broad interest, since it allows to assess from a new perspective the vulnerability and stability of the climate system regionally while relying on global and not only on regional knowledge. The method introduced in paper I therefore has the potential to substantially contribute to the understanding of the local effect of extreme events and tipping points in the earth system within a holistic global framework. In the future, it is hence desirable to extend the idea of climate network analysis to Earth system network analysis. The notion of coupled pattern networks constitutes a first step in this direction (Sect. 6.4). Even though this generalization may seem natural, it is expected to be conceptually and methodologically very challenging because the subsystems involved, *i.e.*, atmosphere, hydrosphere, biosphere, lithosphere and anthroposphere, are qualitatively distinct. For example, it is not a priori obvious how to measure the relationship between the dynamics of entities belonging to different spheres when aiming to construct a meaningful network that reflects the topology of complex interactions within the Earth system. Nevertheless we think that the application of complex network approaches to Earth system analysis presents a worthwhile objective for future research. In essence, complex network theory has the potential to significantly enhance our understanding of Earth system dynamics by integrating large amounts of data in a mathematically well-defined framework and providing the tools to extract conceptional models in the spirit of dimensional reduction, *i.e.*, by outlining the essential entities and their interaction. These data driven conceptual models could in turn be validated against independent data and finally be used to make predictions.

Bibliography

R. Albert and A.L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

R. Albert, I. Albert, and G.L. Nakarado. Structural vulnerability of the North American power grid. *Physical Review E*, 69(2):025103, 2004.

A. Arenas, A. Cabrales, A. Díaz-Guilera, R. Guimerà, and F. Vega-Redondo. Search and congestion in complex networks. In A. Díaz-Guilera, R. Pastor-Satorras, and J.M. Rubí, editors, *Statistical Mechanics of Complex Networks*. Springer, 2003.

A. Arenas, A. Dìaz-Guilera, J. Kurths, Y. Moreno, and C.S. Zhou. Synchronization in complex networks. *Physics Reports*, 469(3):93–153, 2008.

AT&T Research and Bell Labs. Graphviz – graph visualization software, 2004–2009. URL `http://www.graphviz.org/`.

J. Bang-Jensen and G. Gutin. *Digraphs: Theory, Algorithms and Applications*. Springer, 2006.

A.L. Barabási. *Linked: The New Science of Networks*. Perseus Publishing, 2002.

P. Barrett, J.D. Hunter, and P. Greenfield. Matplotlib - A portable Python plotting package. In *Astronomical Data Analysis Software & Systems XIV.*, 2004. URL `http://matplotlib.sourceforge.net/`.

P. beim Graben, C.S. Zhou, M. Thiel, and J. Kurths, editors. *Lectures in Supercomputational Neuroscience: Dynamics in Complex Brain Networks*. Springer Complexity: Understanding Complex Systems. Springer, 2008.

A. Bergner, R. Meucci, K. Al Naimee, M.C. Romano, M. Thiel, J. Kurths, and FT Arecchi. Continuous wavelet transform in the analysis of burst synchronization in a coupled laser system. *Physical Review E*, 78:016211, 2008.

B. Blasius and R. Tönjes. Quasiregular concentric waves in heterogeneous lattices of coupled oscillators. *Physical Review Letters*, 95(8):084101, 2005.

S. Boccaletti, J. Kurths, G. Osipov, D.L. Valladares, and C.S. Zhou. The synchronization of chaotic systems. *Physics Reports*, 366:1–101, 2002.

S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, February 2006.

C.S. Bretherton, C. Smith, and J.M. Wallace. An intercomparison of methods for finding coupled patterns in climate data. *Journal of Climate*, 5(6):541–560, 1992.

S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

P.J. Brockwell and R.A. Davis. *Introduction to Time Series and Forecasting*. Springer, 2002.

M. Buchanan. *Nexus: Small Worlds and the Groundbreaking Theory of Networks*. WW Norton & Co., 2002.

K.W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

A. Clauset, M.E.J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004.

Computational & Information Systems Laboratory at the National Center for Atmospheric Research (NCAR). Python interface to the ncar command language, 2004–2008. URL `http://www.pyngl.ucar.edu/`.

G. Csárdi and T. Nepusz. The igraph software package for complex network research. *InterJournal Complex Systems*, 1695, 2006. URL `http://igraph.sourceforge.net/`.

L. da F. Costa, F.A. Rodrigues, G. Travieso, and P.R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances In Physics*, 56:167, 2007.

G. Di Battista, P. Eades, R. Tamassia, and I.G. Tollis. *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall PTR Upper Saddle River, NJ, USA, 1998.

E.W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

J.F. Donges, Y. Zou, N. Marwan, and J. Kurths. The backbone of the climate network. Submitted, 2008.

J.F. Donges, Y. Zou, N. Marwan, and J. Kurths. Complex networks in climate dynamics. Comparing linear and nonlinear network construction methods. *European Physical Journal Special Topics*, 174:157–179, 2009. In Press.

R.V. Donner, T. Sakamoto, and N. Tanizuka. Complexity of spatio-temporal correlations in japanese air temperature records. In R.V. Donner and S.M. Barbosa, editors, *Nonlinear Time Series Analysis in the Geosciences: Applications in Climatology, Geodynamics and Solar-Terrestrial Physics*, pages 125–154. Springer, 2008.

P. Erdős and A. Rényi. On random graphs I. *Publicationes Mathematicae Debrecen*, 6: 290–297, 1959.

P. Erdős and A. Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

P. Erdős and A. Rényi. On the strength of connectedness of a random graph. *Acta Mathematica Academiae Scientiarum Hungaricae*, 12(1):261–267, 1961a.

P. Erdős and A. Rényi. On the evolution of random graphs. *Bulletin of the International Statistical Institute*, 38:343–347, 1961b.

P. Erdős and A. Rényi. Asymmetric graphs. *Acta Mathematica Academiae Scientiarum Hungaricae*, 14:295–315, 1963.

P. Erdős and A. Rényi. On random matrices. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 8:455–461, 1964.

P. Erdős and A. Rényi. On the existence of a factor of degree one of a connected random graph. *Acta Mathematica Academiae Scientiarum Hungaricae*, 17:359–368, 1966.

P. Erdős and A. Rényi. On random matrices II. *Studia Scientiarum Mathematicarum Hungarica*, 3:459–464, 1968.

L. Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 8:128–40, 1736.

L.C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1): 35–41, 1977.

L.C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3): 215–239, 1979.

N.E. Friedkin. Theoretical foundations for centrality measures. *American Journal of Sociology*, pages 1478–1504, 1991.

K.I. Goh, E. Oh, H. Jeong, B. Kahng, and D. Kim. Classification of scale-free networks. *Proceedings of the National Academy of Sciences*, 99(20):12583–12588, 2002.

A. Gozolchiani, K. Yamasaki, O. Gazit, and S. Havlin. Pattern of climate network blinking links follows El Niño events. *Europhysics Letters*, 83:28005, 2008.

C.W.J. Granger and M. Hatanaka. *Spectral Analysis of Economic Time Series*. Princeton University Press, 1964.

R. W. Hamming. Error detecting and error correcting codes. *Bell System Technical Journal*, 26(2):147–160, 1950.

R. Hegger, H. Kantz, and T Schreiber. Practical implementation of nonlinear time series methods: The tisean package. *CHAOS*, 9:413–435, 1999.

Intergovernmental Panel on Climate Change. *IPCC Fourth Assessment Report*. 2007.

E. Jantsch. *The Self-organizing Universe: Scientific and Human Implications of the Emerging Paradigm of Evolution*. Pergamon Press Oxford, 1980.

E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–2009. URL `http://www.scipy.org/`.

E.W. Jones. On the Topology of El Niño and La Niña Climate Networks. Master's thesis, University of Wisconsin-Milwaukee, May 2007.

H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, 2nd edition, 2004.

R. Kistler, E. Kalnay, W. Collins, S. Saha, G. White, J. Woollen, M. Chelliah, W. Ebisuzaki, M. Kanamitsu, V. Kousky, et al. The NCEP–NCAR 50–Year Reanalysis: Monthly Means CD–ROM and Documentation. *Bulletin of the American Meteorological Society*, 82(2): 247–268, 2001. URL `http://www.cdc.noaa.gov/`.

R. Kleeman. Information flow in ensemble weather predictions. *Journal of the Atmospheric Sciences*, 64(3):1005–1016, 2007.

K. Kosmidis, S. Havlin, and A. Bunde. Structural properties of spatially embedded networks. *Europhysics Letters*, 82(4):48005, 2008.

A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004.

L. Li, D. Alderson, R. Tanaka, J.C. Doyle, and W. Willinger. Towards a theory of scale-free graphs: definition, properties, and implications (extended version). *Arxiv preprint cond-mat/0501169*, 2005.

J. Lovelock. *The Revenge of Gaia: Why the Earth is Fighting Back – and How We Can Still Save Humanity*. Allan Lane, 2006.

D. Maraun and J. Kurths. Epochs of phase coherence between El Niño/Southern Oscillation and Indian monsoon. *Geophysical Research Letters*, 32(15):15709, 2005.

N. Marwan, M. Carmen Romano, M. Thiel, and J. Kurths. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5-6):237–329, 2007.

G.A. Meehl, C. Covey, T. Delworth, M. Latif, B. McAvaney, J.F.B. Mitchell, R.J. Stouffer, and K.E. Taylor. THE WCRP CMIP3 Multimodel Dataset: A New Era in Climate Change Research. *Bulletin of the American Meteorological Society*, 88(9):1383–1394, 2007. URL `https://esg.llnl.gov:8443/`.

M. Merian-Erben. Engraving of Königsberg, 1652. URL `http://www.preussen-chronik.de/_/bild_jsp/key=bild_kathe2.html`.

S. Milgram. The small world problem. *Psychology Today*, 2(1):60–67, 1967.

T.J. Mosedale, D.B. Stephenson, M. Collins, and T.C. Mills. Granger causality of coupled climate processes: Ocean feedback on the North Atlantic Oscillation. *Journal of Climate*, 19(7):1182–1194, 2006.

M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2): 167–256, 2003.

M. E. J. Newman. Analysis of weighted networks. *Physical Review E*, 70(5):56131–56131, 2004.

M.E.J. Newman. Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64(1):16131, 2001a.

M.E.J. Newman. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1):16132, 2001b.

M.E.J. Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54, 2005.

T.E. Oliphant. *A Guide to NumPy*. Trelgol Publishing, 2006. URL `http://www.numpy.org/`.

A. Papana and D. Kugiumtzis. Evaluation of mutual information estimators on nonlinear dynamic systems. *eprint arXiv: 0809.2149*, 2008.

M. Pidwirny. Surface and subsurface ocean currents: Ocean current map. In *Fundamentals of Physical Geography*. Department of Geography, Okanagan University College, 2nd edition, 2006. URL `http://www.physicalgeography.net/fundamentals/8q_1.html`.

A.S. Pikovsky, M.G. Rosenblum, and J. Kurths. *Synchronization: A Universal Concept in Nonlinear Sciences*. Cambridge University Press, 2001.

M. I. Rabinovich, P. Varona, A. I. Selverston, and H. D. I. Abarbanel. Dynamical principles in neuroscience. *Reviews of Modern Physics*, 78:1213–1265, October 2006.

M.C. Romano, M. Thiel, J. Kurths, and C. Grebogi. Estimation of the direction of the coupling by conditional probabilities of recurrence. *Phys. Rev. E*, 76:036211, 2007.

M.G. Rosenblum, A.S. Pikovsky, and J. Kurths. Phase synchronization of chaotic oscillators. *Physical Review Letters*, 76(11):1804–1807, Mar 1996.

A.F. Rozenfeld, R. Cohen, D. Ben-Avraham, and S. Havlin. Scale-free networks on lattices. *Physical Review Letters*, 89(21), 2002.

H.J. Schellnhuber and V. Wenzel. *Earth Systems Analysis: Integrating Science for Sustainability*. Springer, 1998.

B. Schelter, M. Winterhalder, R. Dahlhaus, J. Kurths, and J. Timmer. Partial phase synchronization for multivariate synchronizing systems. *Physical Review Letters*, 96(20): 208103, 2006.

G. Schmidt, G. Zamora-López, and J. Kurths. Structure function relationship in complex cat's brain network unveiled by synchronization. International Journal of Bifurcation and Chaos. In press, 2008.

T. Schreiber. Measuring Information Transfer. *Physical Review Letters*, 85(2):461–464, 2000.

T. Schreiber and A. Schmitz. Surrogate time series. *Physica D: Nonlinear Phenomena*, 142 (3-4):346–382, 2000.

U. Schwarz, A.O. Benz, J. Kurths, and A. Witt. Analysis of solar spike events by means of symbolic dynamics methods. *Astronomy and Astrophysics*, 277(1):215–224, 1993.

M. Sexton, J.F. Donges, and N. Marwan. The spatial resolution dependence of complex climate networks. Internal report, 2009.

R. Solomonoff and A. Rapoport. Connectivity of random nets. *Bulletin of Mathematical Biology*, 13(2):107–117, 1951.

R.H. Stewart. *Introduction to Physical Oceanography*. Texas A & M University, 2005.

M. Thiel, M.C. Romano, J. Kurths, M. Rolfs, and R. Kliegl. Twin surrogates to test for complex synchronisation. *Europhysics Letters*, 75(4):535–541, 2006.

R. Tönjes. *Pattern Formation Through Synchronization in Systems of Nonidentical Autonomous Oscillators*. PhD thesis, University of Potsdam, 2007.

K.E. Trenberth et al. The definition of El Niño. *Bulletin of the American Meteorological Society*, 78(12):2771–2777, 1997.

A.A. Tsonis and P.J. Roebber. The architecture of the climate network. *Physica A*, 333: 497–504, 2004.

A.A. Tsonis and K.L. Swanson. Topology and predictability of El Niño and La Niña networks. *Physical Review Letters*, 100(22):228502, 2008.

A.A. Tsonis, K.L. Swanson, and P.J. Roebber. What do networks have to do with climate? *Bulletin of the American Meteorological Society*, 87:585–595, May 2006.

A.A. Tsonis, K.L. Swanson, and G. Wang. Estimating the clustering coefficient in scale-free networks on lattices with local spatial correlation structure. *Physica A: Statistical Mechanics and its Applications*, 387(21):5287–5294, 2008a.

A.A. Tsonis, K.L. Swanson, and G. Wang. On the role of atmospheric teleconnections in climate. *Journal of Climate*, 21(12):2990–3001, 2008b.

A.A. Tsonis, G. Wang, K.L. Swanson, F.A. Rodrigues, and L. da F. Costa. Community structure and dynamics in climate networks. Status unknown, 2009.

G.K. Vallis. *Atmospheric and Oceanic Fluid Dynamics: Fundamentals and Large-scale Circulation*. Cambridge University Press, 2006.

G. van Rossum et al. Python language website, 1991–2009. URL `http://www.python.org/`.

J.A. Vastano and H.L. Swinney. Information transport in spatiotemporal systems. *Physical Review Letters*, 60(18):1773–1776, 1988.

P.F. Verdes. Assessing causality from multivariate time series. *Physical Review E*, 72:026222, 2005.

W. von Bloh, M.C. Romano, and M. Thiel. Long-term predictability of mean daily temperature data. *Nonlinear Processes in Geophysics*, 12:471–479, 2005.

A. von Humboldt. *Kosmos. Entwurf einer physischen Weltbeschreibung.* Cotta'scher Verlag, Stuttgart & Tübingen, 1845.

H. von Storch and F.W. Zwiers. *Statistical Analysis in Climate Research*. Cambridge University Press, 1999.

J.M. Wallace and D.S. Gutzler. Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Monthly Weather Review*, 109(4):784–812, 1981.

B. Wang. *The Asian Monsoon*. Springer, 2006.

C. P. Warren, L. M. Sander, and I. M. Sokolov. Geography in a scale-free network model. *Physical Review E*, 66(5):056105, 2002.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393: 440–442, 1998.

D.J. Watts. *Six Degrees: The Science of a Connected Age*. WW Norton & Co., 2003.

K. Yamasaki, A. Gozolchiani, and S. Havlin. Climate networks around the globe are significantly affected by El Niño. *Physical Review Letters*, 100(22):228501, 2008.

D.A. Zacharias, J.D. Violin, A.C. Newton, and R.Y. Tsien. Partitioning of lipid-modified monomeric GFPs into membrane microdomains of live cells. *Science*, 296(5569):913–916, 2002.

G. Zamora-López. *Linking Structure and Function of Complex Cortical Networks*. PhD thesis, University of Potsdam, 2008.

G. Zamora-López, C.S. Zhou, V. Zlatic, and J. Kurths. The generation of random directed networks with prescribed 1-node and 2-node degree correlations. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224006, 2008.

C.S. Zhou, L. Zemanová, G. Zamora-López, C. C. Hilgetag, and J. Kurths. Hierarchical organization unveiled by functional connectivity in complex brain networks. *Physical Review Letters*, 97(23):238103, 2006.

C.S. Zhou, L. Zemanová, G. Zamora-López, C. C. Hilgetag, and J. Kurths. Structure function relationship in complex brain networks expressed by hierarchical synchronization. *New Journal of Physics*, 9:178, 2007.

APPENDIX A

Community structure in climate networks

Here we present some hints pointing to a pronounced *community structure* present in the HadCM3 SAT climate network (Sect. 3.1.1). Roughly, a subgraph of a network forms a community, if it is internally notably more densely connected than with external vertices. In this sense the components of a network constitute a trivial community structure. Generally, which subgraphs we consider to form communities is not clearly defined, but depends on the choice of some parameter or ad-hoc criterion (da F. Costa et al. (2007)). Within the framework of climate networks, communities correspond to sets of geographical regions carrying a similar dynamics with respect to some climatological field, *e.g.*, surface air temperature. Studying the community structure of climate networks we can accordingly expect to find some known climate zones, but there is also the hope to reveal yet unknown dynamically coherent structures. We suggest that it could in the future be particularly interesting to extract community structure from climate networks generated using sophisticated measures of synchronization (Sect. 3.2) and from networks extending the notion of coupled patterns between two (or more) climatological observables (Bretherton et al. (1992)). Very recently we came to know of a paper also studying community structure in climate networks which was not yet published by the date of submission of this thesis (Tsonis et al. (2009)).

Average nearest neighbor degree The average nearest neighbor degree

$$k_v^{nn} = \frac{1}{k_v} \sum_{i=1}^{N} A_{vi} k_i \tag{A.1}$$

of vertex v is a measure of degree-degree correlations in a complex network. For the HadCM3 SAT climate network we find that vertices with similar degree tend to be connected preferentially among each other (Fig. A.1), *i.e.*, the climate network is assortative. Assortativity hints at an identifiable community structure, since vertices with a low degree can be thought of connecting only to vertices within in the same community whereas high degree hubs link the communities by connecting to other hubs in different communities. Only for the highest degree nodes with $k \geq 200$ the SAT climate network appears to be disassortative, *i.e.*, k^{nn} decays with k in this region. To reassure these results on degree-degree correlations, proper

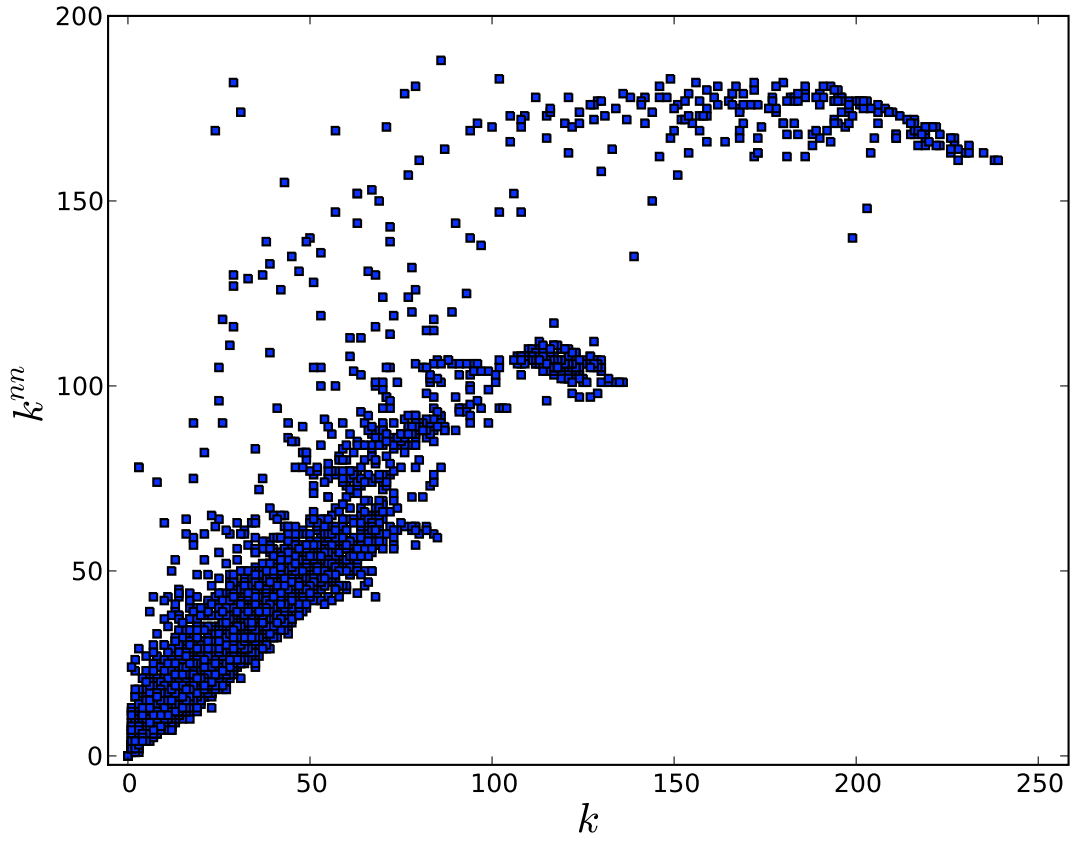


Figure A.1 Average nearest neighbor degree k_v^{nn} plotted against degree k_v. The HadCM3 SAT Pearson correlation climate network at $\rho = 0.005$ clearly shows assortative behavior. The Spearman's Rho of k_v and k_v^{nn} is $r_s(k_v,k_v^{nn}) = 0.9834$.

significance tests on the network theoretical level have to be performed (Zamora-López et al. (2008), Zamora-López (2008)).

Community structure based on modularity maximization To obtain a first impression of the network's community structure, we employ a fast algorithm based on greedy maximization of modularity (Clauset et al. (2004)). The ten largest communities detected already show some interesting characteristics (Fig. A.2). The remaining communities are very small and most of the associated vertices are disconnected from the network's giant component.

The communities tend to be oriented zonally, reflecting the zonal alignment of the major climate zones. They furthermore show a preference to follow continental boundaries again demonstrating the land-sea difference in SAT dynamics. The largest community encompasses Africa, the Indian Ocean, the Southern Ocean and large expanses of the Atlantic Ocean. While the Antarctic forms an isolated community, the Arctic is connected to Greenland, Europe, large parts of Asia and North America. The community extending from the North

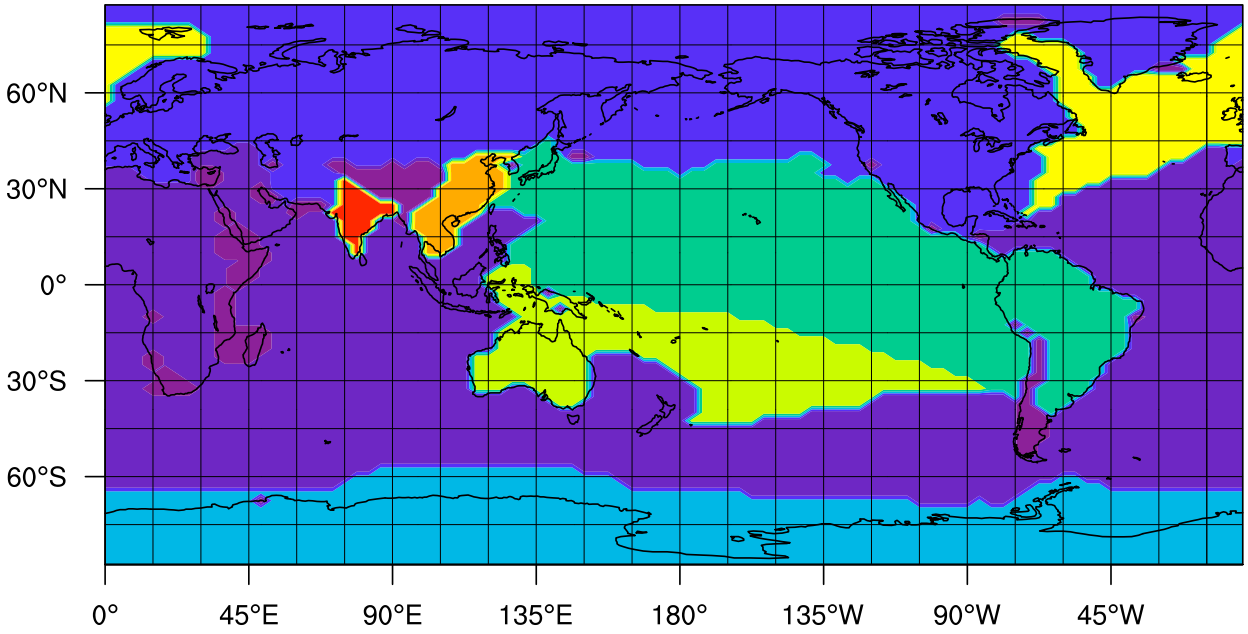


Figure A.2 The ten largest communities of the HadCM3 SAT Pearson correlation climate network at $\rho = 0.005$ obtained by fast and greedy maximization of modularity.

Atlantic to the Baffin, Labrador and Norwegian Seas could be an indicator of the oceanic realm of influence of the North Atlantic Current. The Pacific Ocean is dominated by two communities that also include Australia and South America, respectively. Note also the more localized communities over India and South East Asia.

APPENDIX B

Towards directed climate networks

Directed networks (digraphs) extracted from multivariate time series data promise to yield highly interesting novel insights into the dynamics of spatially extended systems, *e.g.*, into that of the climate system. To explore the potential of this approach we extend the climate network construction method presented in Chap. 3. Particularly we build on the undirected Pearson correlation climate network by experimentally employing the average phase shift $\Delta\Phi_{ij} = \langle \Phi_i(\omega) - \Phi_j(\omega) \rangle_{\omega>0}$[1] between anomaly time series $\hat{a}_i(t)$,$\hat{a}_i(t)$ as a measure of directionality, where $\Phi_i(\omega) = \arg(\mathcal{F}(\hat{a}_i))(\omega)$ and $\mathcal{F}(\hat{a}_i))(\omega)$ denotes the discrete Fourier transform of $\hat{a}_i(t)$ (Eq. 4.2). It should be pointed out that $\Delta\Phi_{ij}$ is only easily interpretable as an average phase shift, if the power spectra of $\hat{a}_i(t)$,$\hat{a}_i(t)$ are small banded.

Starting with an unconnected set of vertices, we assemble the climate digraph in the following fashion. If $\hat{a}_i(t)$ on average lags behind $\hat{a}_j(t)$, *i.e.*, if $\Delta\Phi_{ij} < 0$, and $P_{ij} > \tau$ we create an arc (j,i). If in turn $\hat{a}_i(t)$ on average runs ahead of $\hat{a}_j(t)$, that is $\Delta\Phi_{ij} > 0$, and $P_{ij} > \tau$ an arc (i,j) is established. Specifically, using the Pearson correlation P_{ij} and average phase shift $\Delta\Phi_{ij}$ matrices the adjacency matrix A_{ij} is given by

$$A_{ij} = \Theta(P_{ij} - \tau)\Theta(\Delta\Phi_{ij}), \tag{B.1}$$

with $\Theta(.)$ the Heaviside function. Note that pragmatically, we calculate P_{ij} and $\Delta\Phi_{ij}$ simultaneously by considering the complex correlation coefficient $C^a_{ij} = \left\langle \hat{a}^a_i(t)^* \hat{a}^a_j(t) \right\rangle_t$ of the analytic signals $\hat{a}^a_i(t)$,$\hat{a}^a_j(t)$ (Brockwell and Davis (2002), Bergner et al. (2008)). Substantially, the analytic signal $\hat{a}^a_i(t)$ corresponds to the time series $\hat{a}_i(t)$ with its negative frequency contributions removed. It can be shown straightforwardly that $P_{ij} = |C^a_{ij}|$ and $\Delta\Phi_{ij} = \arg(C^a_{ij})$.

Now we can consider directed climate network measures, *e.g.*, the topologically local *in* -

1 Here we use the mean of angles, *i.e.*, $\langle \Phi_j(\omega) \rangle_{\omega>0} = \arg\left(2/\mathcal{T} \sum_{k,\omega_k>0} \exp(i\Phi_j(\omega_k))\right)$.

area weighted connectivity (in-AWC)

$$AWC_v^{in} = \frac{\sum_{i=1}^{N} A_{iv} \cos(\lambda_i)}{\sum_{i=1}^{N} \cos(\lambda_i)}, \tag{B.2}$$

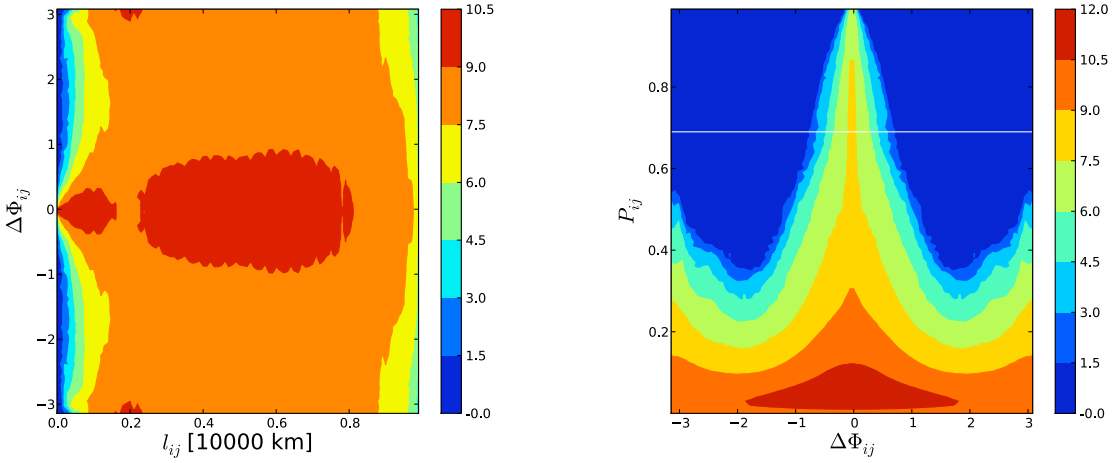
and the *out - area weighted connectivity* (out-AWC)

$$AWC_v^{out} = \frac{\sum_{i=1}^{N} A_{vi} \cos(\lambda_i)}{\sum_{i=1}^{N} \cos(\lambda_i)}. \tag{B.3}$$

If vertex v has a large AWC_v^{in} but a small AWC_v^{out}, its dynamics tends to lag behind that of its neighbors. If in contrast, v has small AWC_v^{in} but large AWC_v^{out} it tends to dynamically lead its neighbors.

Note that on the global topological scale we can also calculate *directed shortest path betweenness* by including only the contributions of directed paths in Eq. 2.14. Along a directed path $(i,v_1,v_2,\ldots,v_{d-1},j)$ of length d connecting vertices i and j, the cumulative average phase difference increases with every arc, *i.e.*, the directed path follows the direction of the mean phase flow in the spatially extended dynamical system[1].

The edges included in the directed HadCM3 surface air temperature climate network carry



(a) (b)

Figure B.1 Frequency plot of average phase shift $\Delta\Phi_{ij}$ vs. (a) edge distance l_{ij} and (b) Pearson correlation P_{ij} for HadCM3 SAT data set. The distributions are symmetric with respect to the $\Delta\Phi_{ij} = 0$ axis, since we compute the histogram for the complete antisymmetric matrix $\Delta\Phi_{ij}$. A histogram with 10^4 equally sized rectangular bins is used for this analysis. The white horizontal line indicates the threshold of $\tau = 0.69$ corresponding to an edge density of $\rho = 0.005$ for the associated climate network. The color bar gives the logarithm of frequency.

1 Specifically, this holds for any directed path, not only the shortest directed paths wielded for the calculation of directed betweenness centrality.

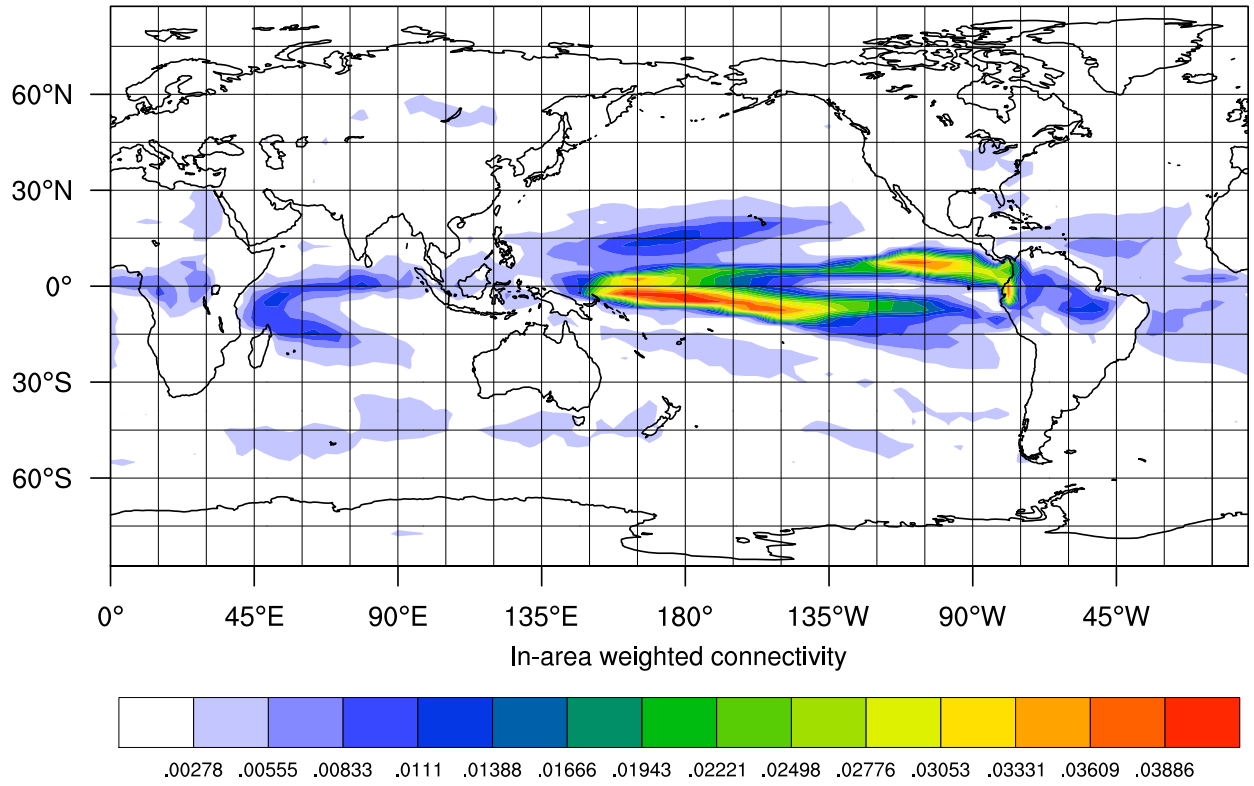
relatively small average phase shifts of $-1 < \Delta\Phi_{ij} < 1$ at an edge density of $\rho = 0.005$ (Fig. B.1(b)). The dependence of $\Delta\Phi_{ij}$ on edge length l_{ij} is relatively homogenous, except for very small, very large edge distances, two 'blobs' of high frequency close to the $\Delta\Phi_{ij} = 0$ axis and one smaller high frequency region at $l_{ij} \approx 0.2$ and $\Delta\Phi_{ij} = \pm\pi$ (Fig. B.1(a)).

Comparing the fields of in-AWC and out-AWC (Fig. B.2) we observe that the super-nodes found in both fields are complementary, *i.e.*, regions with large in-AWC tend to have a small out-AWC and vice versa. This suggests that some well defined centers of action on average have a leading phase relationship to their neighbors, *e.g.*, the out-AWC super-node in the tropical East Pacific at $\lambda \approx -5°N$ and $\phi \approx 270°E$ (Fig. B.2(b)), while others on average lag behind their neighbors, *e.g.*, the in-AWC super-node in the tropical West Pacific at $\lambda \approx -5°N$ and $\phi \approx 180°E$ (Fig. B.2(a)). There is also weaker evidence of complementary super-nodes over the Indian Ocean. We suggest that the complementary super-nodes in the tropical Pacific could be a footprint of ENSO in our directed SAT climate network. This is reasonable because ENSO varies on interannual time scales of three to eight years that is well captured by our monthly anomaly SAT data, while faster time scales are averaged out. More specifically, some of the complementary super-nodes resemble known centers of action of the ENSO phenomenon, *e.g.*, the pronounced out-AWC super-node off the west coast of South America corresponds to the warm water pool forming off the South American Pacific coast during El Niño episodes. Again, this is conceivable because of the strong heat flux coupling between sea surface and surface air temperature.

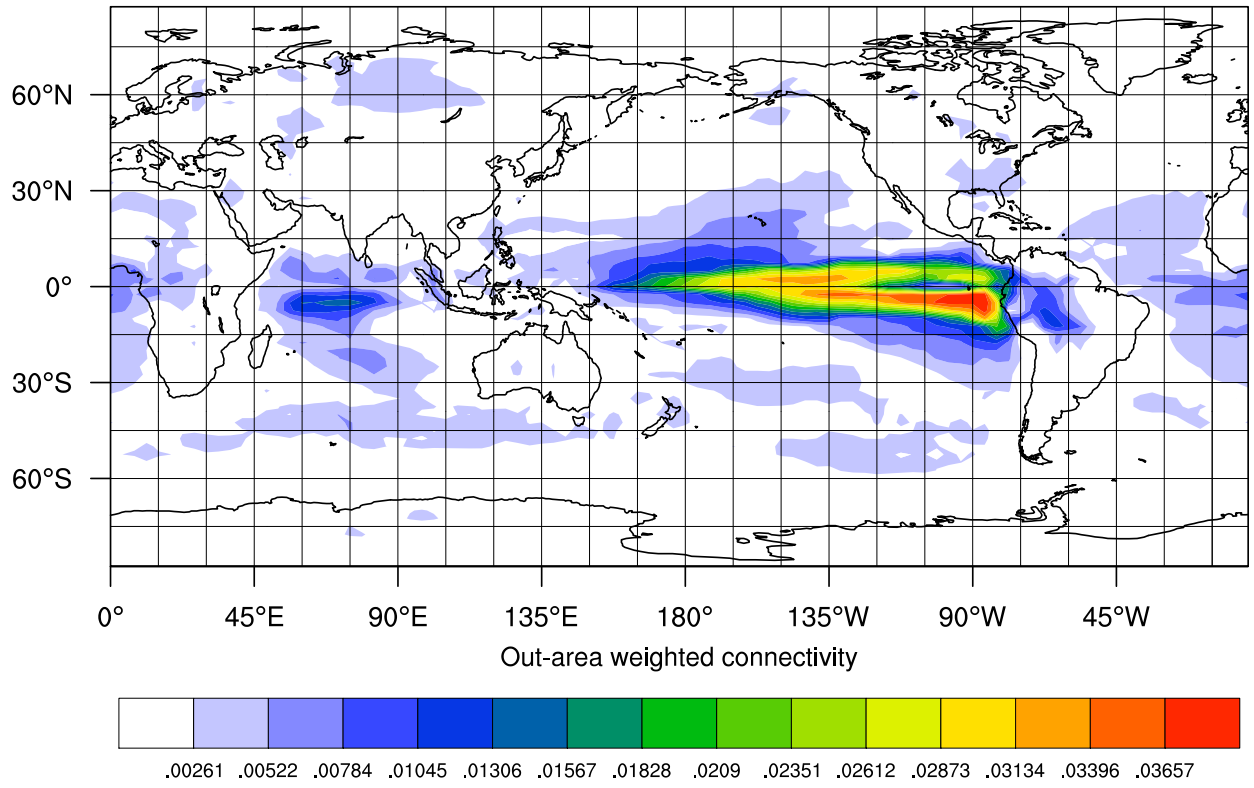
The directed betweenness field for the same directed climate network (Fig. B.3) possesses backbone structures resembling those found in the undirected Pearson correlation climate network constructed from the same data set (Fig. 3.8(a)). Given the hypothesis that perturbations in the dynamics of the underlying SAT field are advected preferentially by strong surface ocean currents we expect that there should on average also be a monotonous phase evolution in the dynamics of vertices along the path of surface ocean currents. This in turn should result in a comparable contribution of directed and undirected shortest paths to the betweenness of these vertices, finally leading to similar backbone structures in the fields of directed and undirected betweenness centrality. Ergo, following this reasoning the similarity of the directed and undirected betweenness fields calculated for the HadCM3 SAT network gives supporting evidence that shortest path betweenness can be interpreted as a measure of advective dynamical information flow in climate networks.

One has to be aware that the purely linear analysis performed above presents just a preliminary step *towards directed climate networks*, since (i) the average phase difference $\Delta\Phi_{ij}$ is not suitable to study causal relationships and (ii) it is only interpretable for time series with small banded power spectra which is not true in a strong sense for the climatological anomaly time series studied here. Inspite of these deficiencies our simple method is able to detect pronounced structures in directed network measure fields, *i.e.*, in- and out-AWC and directed betweenness. Furthermore note that we obtained similar results for the NCEP/NCAR SAT data set substantiating their physical relevance. To construct directed climate networks

from time series where directionality is related to causality, more sophisticated measures of information transport and causality such as transfer entropy (Schreiber (2000)) and granger causality (Granger and Hatanaka (1964)) or recurrence based methods (Romano et al. (2007)) could be used in the course of future research.



(a)



(b)

Figure B.2 (a) In - and (b) out - AWC fields for a directed climate network at $\rho = 0.005$ constructed from the HadCM3 SAT data set using the complex correlation coefficient method.

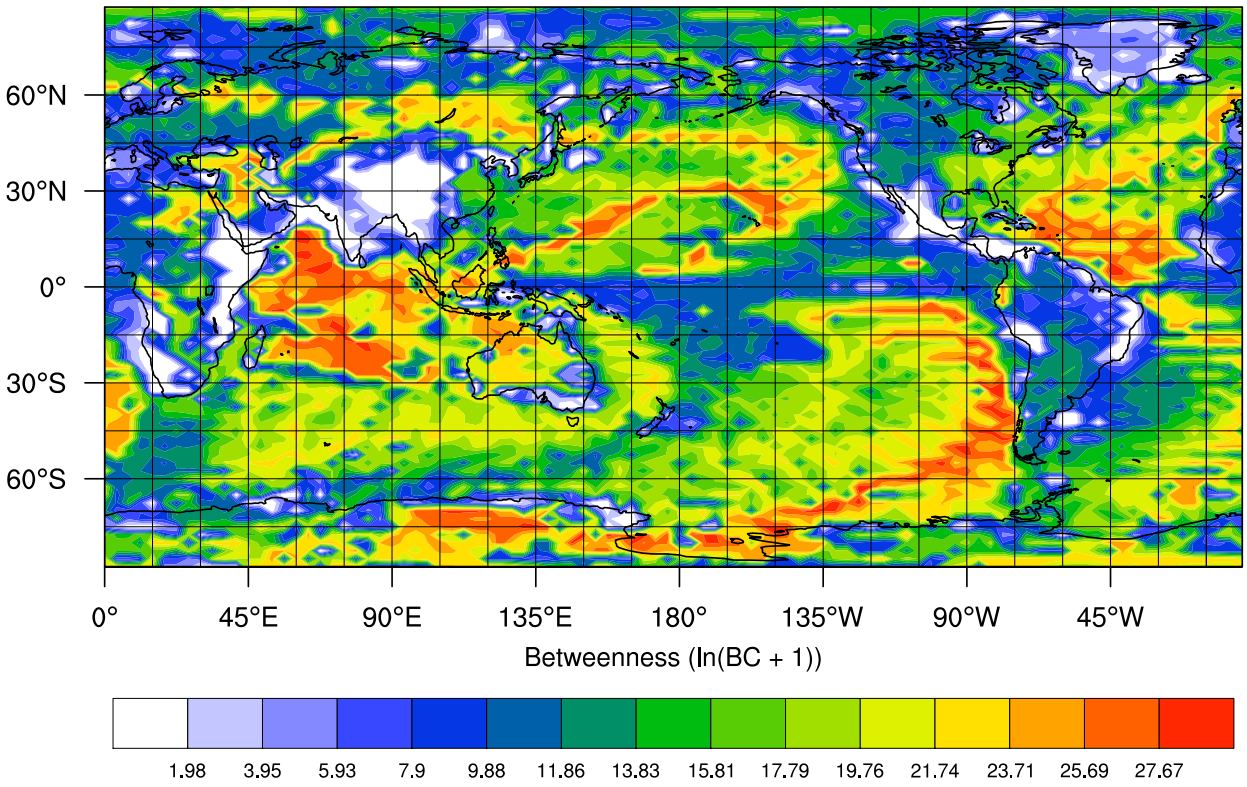


Figure B.3 Directed betweenness field for a directed climate network at $\rho = 0.005$ constructed from the HadCM3 SAT data set using the complex correlation coefficient method.

APPENDIX C

Betweenness in El Niño and La Niña climate networks

Inspired by the work of Tsonis and Swanson (2008), we would like to check for qualitative changes in the betweenness backbone structure of SAT climate networks (Chap. 5) with respect to different states of the El Niño Southern Oscillation (ENSO). We first calculate the southern oscillation index (SOI) from monthly averaged NCEP/NCAR surface pressure (SP) data (Kistler et al. (2001)), that is provided on the same grid as the NCEP/NCAR surface air temperature (SAT) data used to construct El Niño and La Niña climate networks from (Table 3.1). The SOI index is defined as the normalized surface pressure difference $\Delta P(t)$ between Tahiti and Darwin (Trenberth et al. (1997)), *i.e.*,

$$SOI(t) = \frac{\Delta P(t) - \langle \Delta P(t) \rangle_t}{STD(\Delta P(t))}, \tag{C.1}$$

where $STD(\Delta P(t))$ denotes the standard deviation of pressure difference. The coordinates

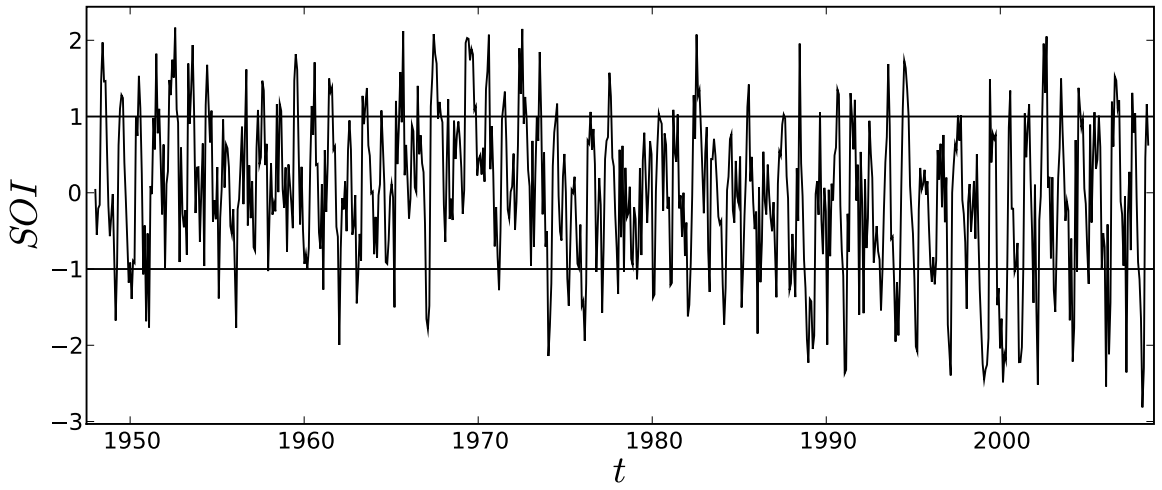


Figure C.1 . Time evolution of the SOI index calculated from NCEP/NCAR reanalysis surface pressure data.

of Tahiti (T) and Darwin (D) are approximately given by ($\lambda_T = -17.4°, \phi_T = 210.5°$) and ($\lambda_D = -12.3°, \phi_D = 130.5°$).

We now define the El Niño state as comprising all time series indices t satisfying $SOI(t) < -1$, likewise all indices t' with $SOI(t') > 1$ are considered to belong to the La Niña state (Fig. C.1). Using this criterion, we construct concatenated anomaly time series from the NCEP/NCAR SAT data set, that contain only samples from the El Niño and La Niña states, respectively. The El Niño time series contain 115 samples while the La Niña time series consist of 124 data points. From the resulting El Niño and La Niña data sets we generate El Niño and La Niña climate networks using the method presented in Chap. 3. We choose Pearson correlation as the correlation measure, because the small number of samples contained in the concatenated El Niño and La Niña time series limits our ability to calculate a meaningful mutual information between pairs of time series.

First note, that handling climate networks with nearly four times as many vertices as contained in the networks analyzed by Tsonis and Swanson (2008), we can confirm their result that the La Niña network possesses a greater number of long range connections than the El Niño network (Fig. C.2). Comparing the El Niño and La Niña betweenness fields, we first observe that backbone structures appear in both states of ENSO (Fig. C.3). During La Niña the backbone structures appear to be more delicate and well defined, whereas they are broader and distorted during El Niño. This effect could be related to the more regular (*i.e.*, more predictable) climate during La Niña as compared to the greater irregularity (*i.e.*, lower predictability) during El Niño episodes (Tsonis and Swanson (2008)). The major backbone structures described in Chap. 5 are seen during both ENSO states when allowing for translations, deformations and differences in relative strength[1]. Some more pronounced deviations in backbone structure appear over the Indian, South Pacific and North Atlantic Oceans.

Further work is required to consolidate these findings and relate them to known features of ENSO dynamics. Backbone structures in the betweenness fields of El Niño and La Niña climate networks promise to provide novel insights into changes of the information flow structure with the SAT field during ENSO (Sect. 5.2).

1 By the strength of a backbone structure at vertex v we refer to the value of the betweenness field BC_v.

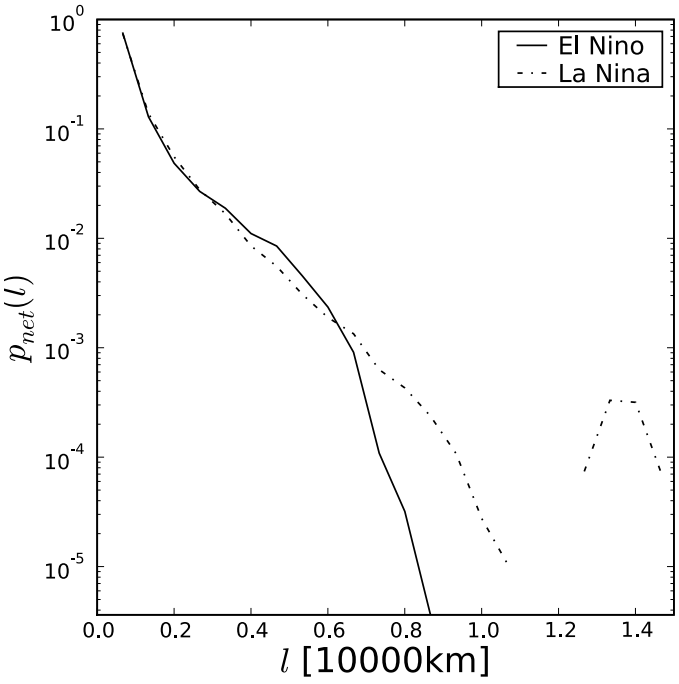
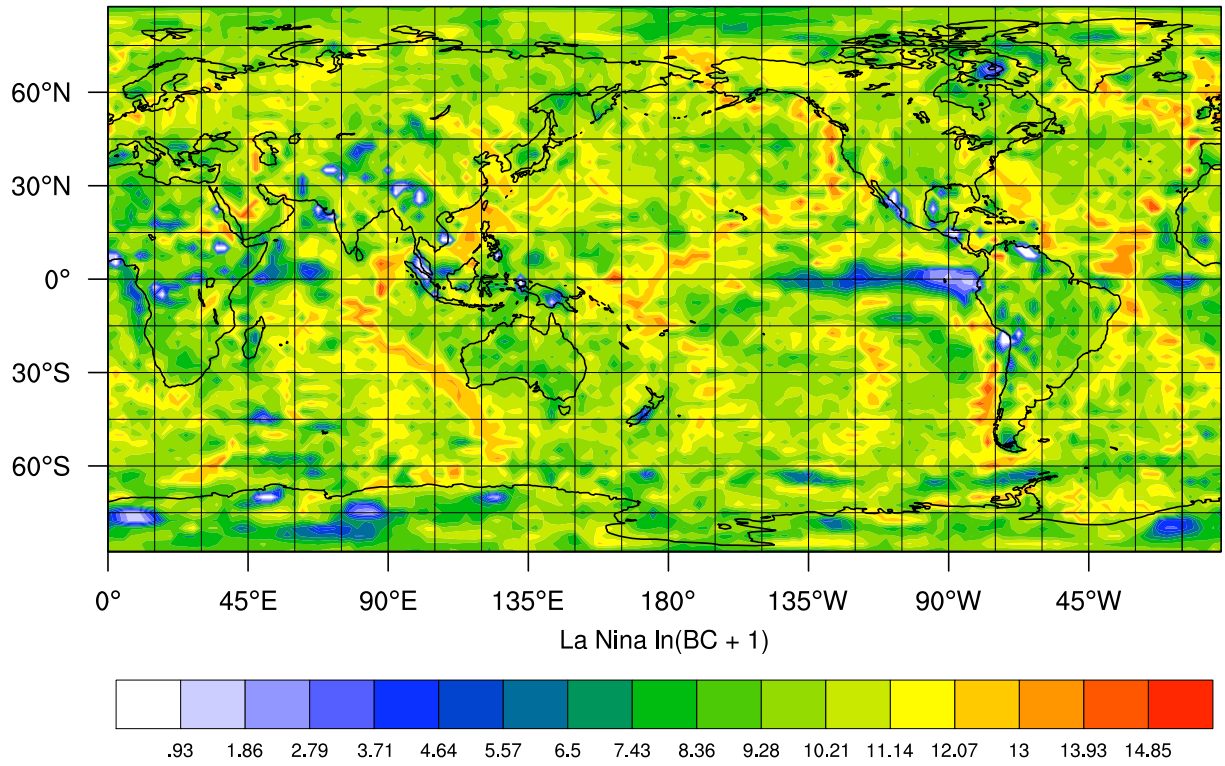
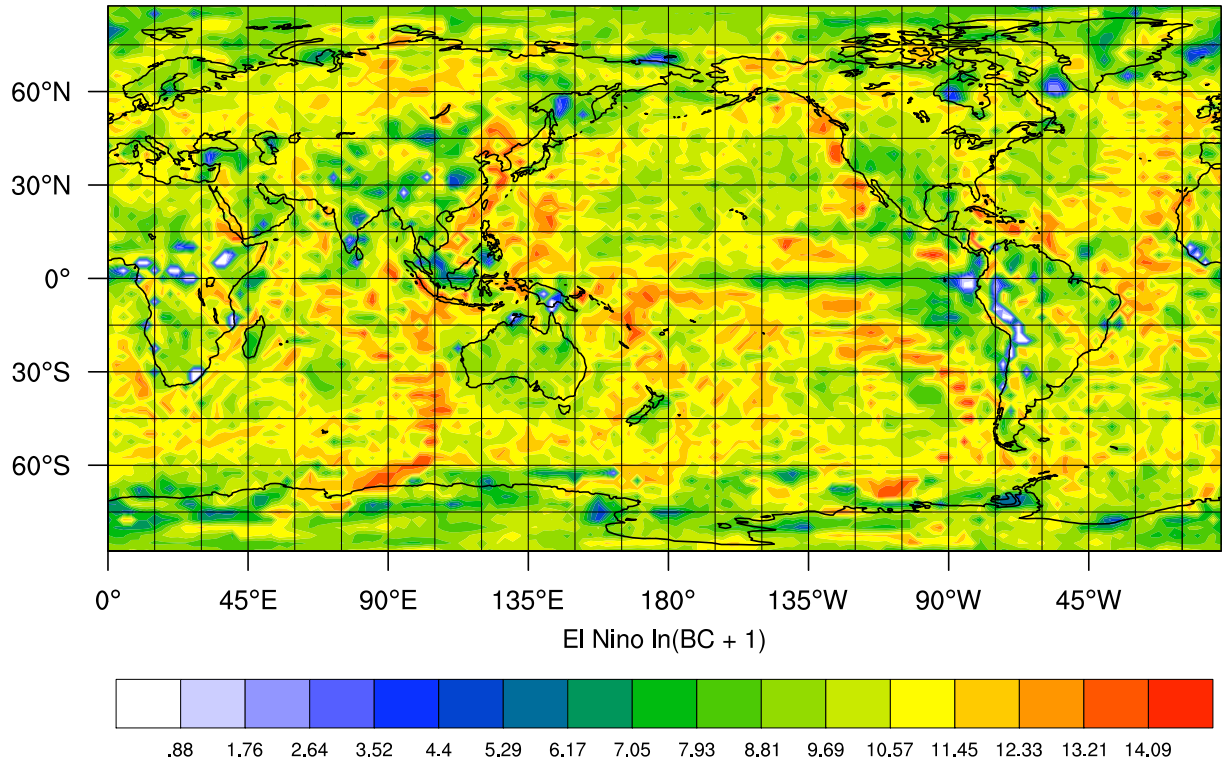


Figure C.2 Intrinsic edge distance distribution $p_{net}(l)$ of El Niño and La Niña SAT climate networks at $\rho = 0.005$ constructed from NCEP/NCAR reanalysis surface pressure data using Pearson correlation. Compared to the El Niño network, the La Niña network possesses a considerably larger number of long range teleconnections.



(a)



(b)

Figure C.3 Betweenness fields of (a) La Niña and (b) El Niño SAT climate networks calculated from NCEP/NCAR SAT data at $\rho = 0.005$ using Pearson correlation. Note that the color scale is logarithmic.

APPENDIX D

Supplementary results from additional AOGCM runs

For reference, we provide additional betweenness fields calculated for monthly surface air temperature (SAT) Pearson correlation climate networks generated from a representative subset of model output from the World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model data set (Meehl et al. (2007)). As in Sect. 3.1.1, we choose 20th century reference runs for optimal comparability with reanalysis data (Table D.1).

The betweenness fields of CCCma (Fig. D.1(a)), NCAR PCM1 (Fig. D.1(b)), CNRM (Fig. D.2(a)) and GFDL CM2.0 (Fig. D.2(b)) all show backbone structures qualitatively resembling those described and discussed for NCEP/NCAR reanalysis and HadCM3 SAT climate networks in Chap. 5. Note particularly the effect of spatial resolution or equivalently network size N on the shape of backbone structures. The lower N, the more the major backbone structures tend to be blurred while weaker and narrower features disappear completely. This is consistent with the hypothesis raised in Chap. 5 that western boundary currents (WBCs) are not clearly identifiable in the betweenness fields of climate networks studied until this point, because they are too narrow to be resolved by the coarse grids available.

In contrast, the betweenness fields of high resolution SAT climate networks studied here, *i.e.*, those constructed from GFDL CM2.0 and ECHAM5 data sets, do indeed include backbone structures along the east coasts of continental land masses resembling WBCs (Fig. D.2(b), **??**). For example, note the Gulf Stream and Brazil current along the east coast of North and South America, the Agulhas along the east coast of Africa, the Kuroshio along the western rim of the North Pacific and finally the East Australia current (Fig. 5.2).

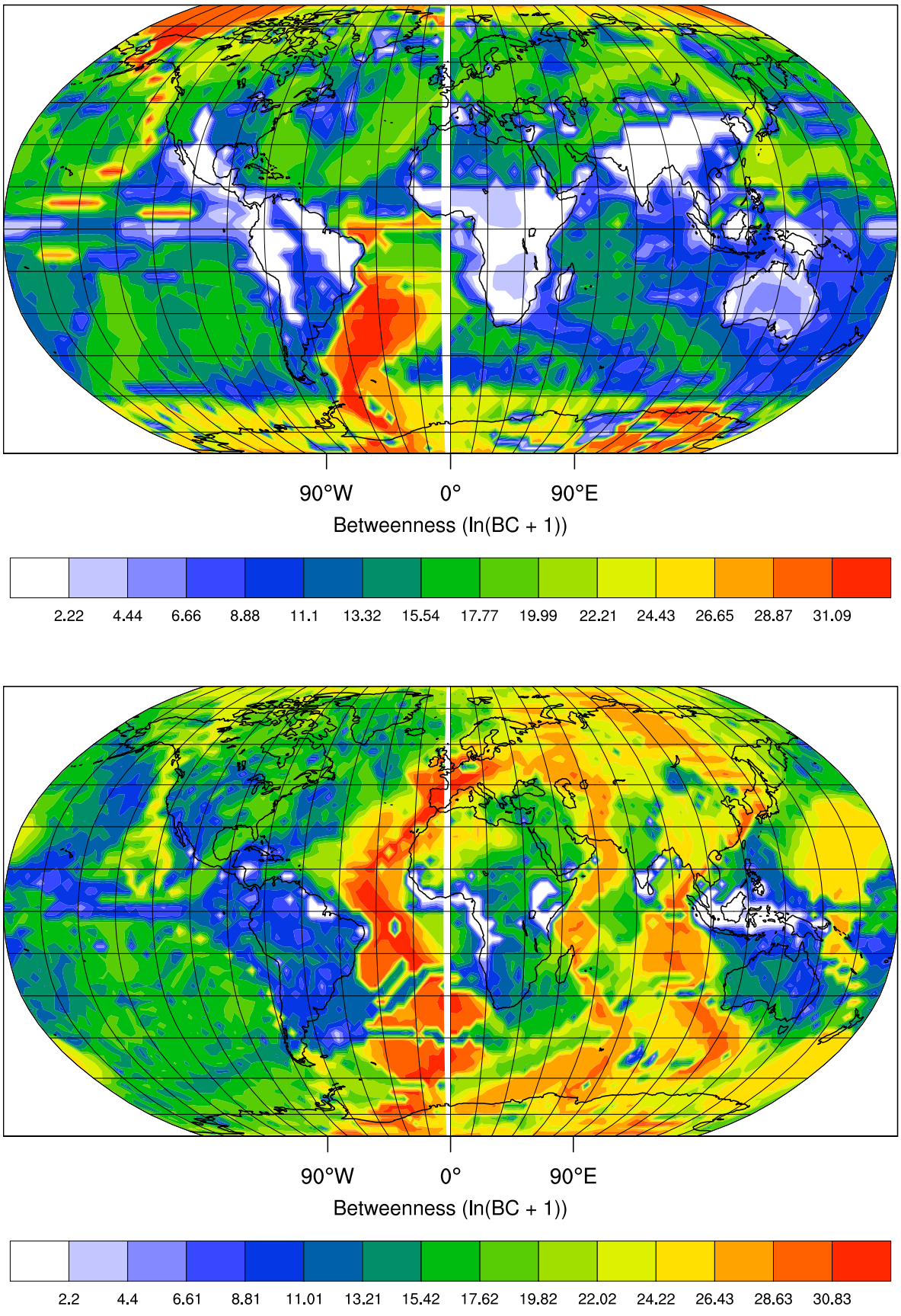


Figure D.1 Betweenness fields for (a) CCCma and (b) NCAR PCM1 SAT climate networks at $\rho = 0.005$ constructed using Pearson correlation.

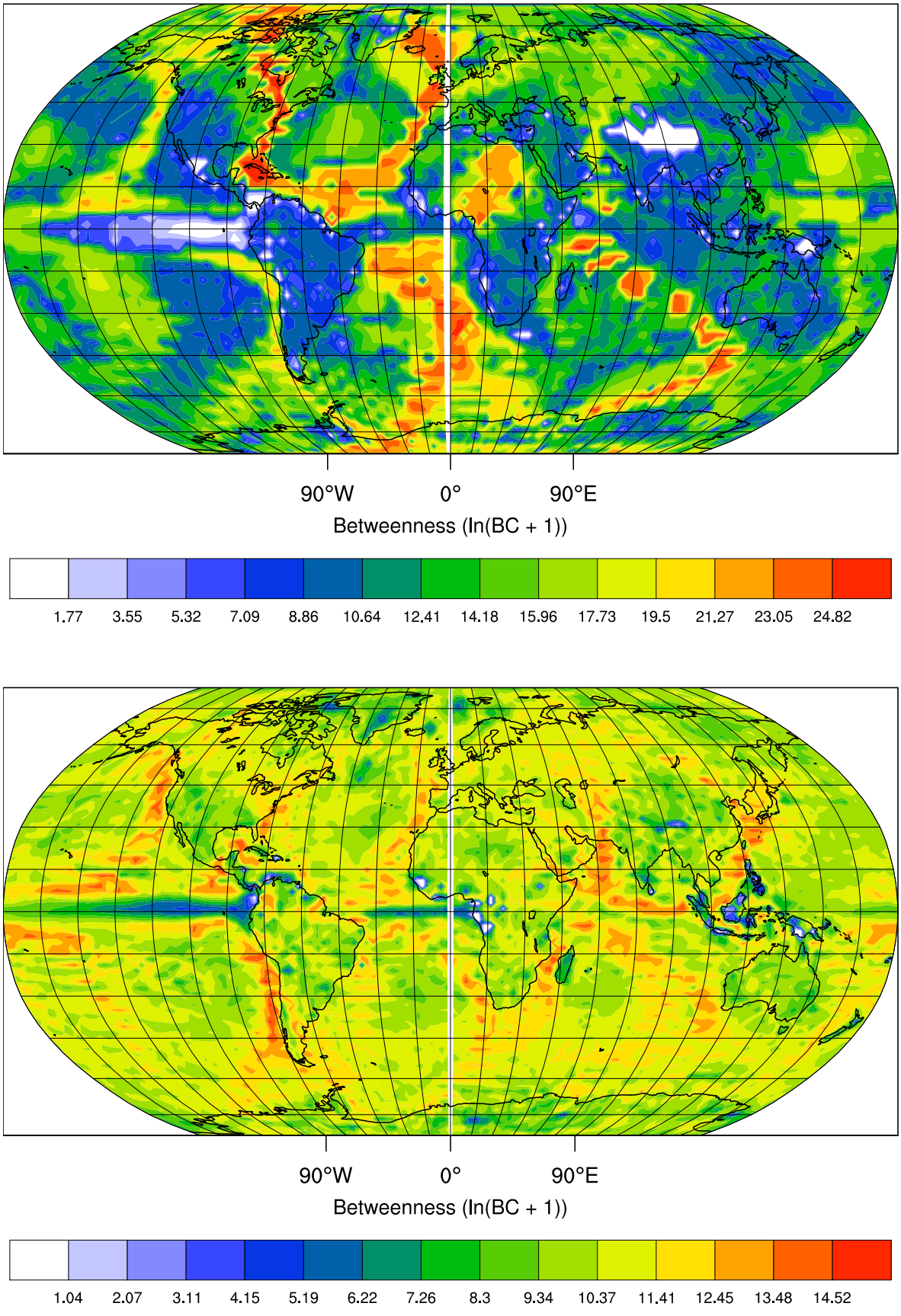


Figure D.2 Betweenness fields for (a) CNRM and (b) GFDL CM2.0 SAT climate networks at $\rho = 0.005$ constructed using Pearson correlation.

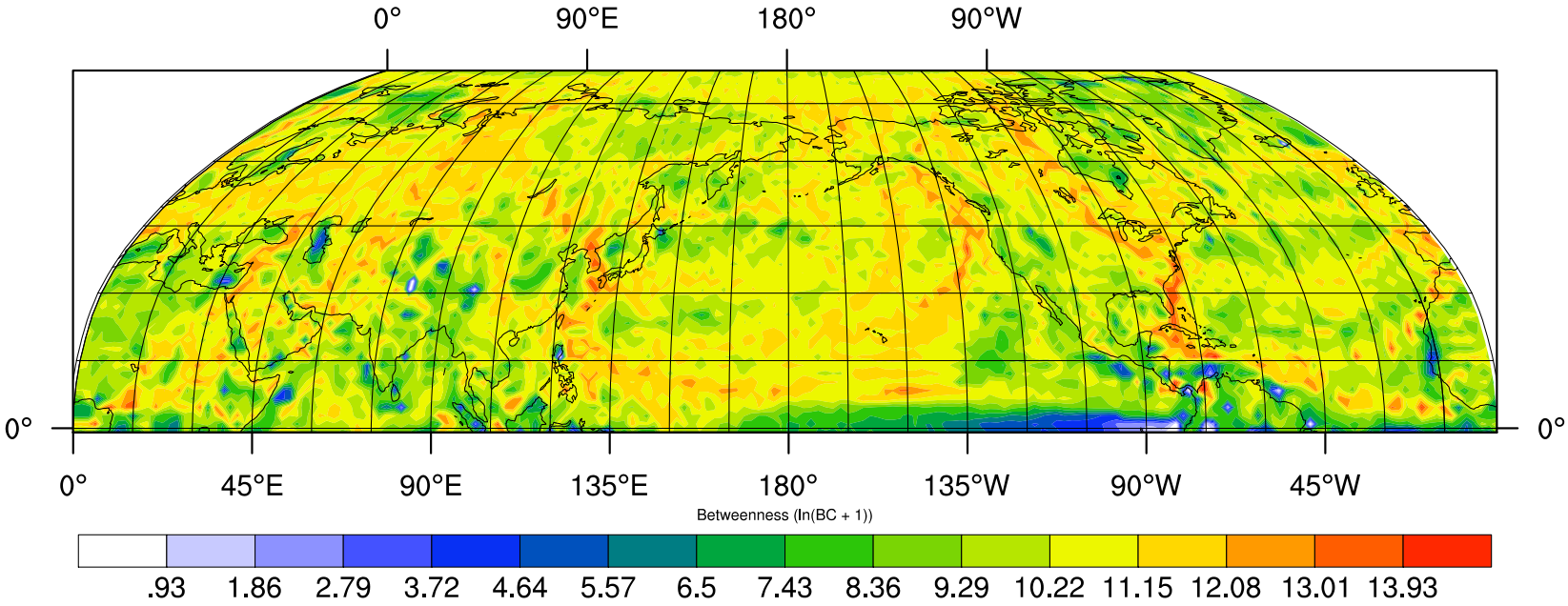


Figure D.3 Betweenness field for ECHAM5 SAT climate network at $\rho = 0.006$ constructed using Pearson correlation.

Table D.1 Properties of additional WCRP CMIP3 model surface air temperature data sets. The data sets are sorted with spatial resolution increasing from left to right.

	CCCma	NCAR PCM1	CNRM	GFDL CM2.0	ECHAM5
Spatial extent	global	global	global	global	northern hemisphere
Temporal coverage	01/1850 - 12/2000	01/1890 - 12/1999	01/1860 - 12/1999	01/1861 - 12/2000	01/1860 - 12/2100
$\mathcal{T}$ [months]	1320	1812	1680	1680	2892
$\Delta\lambda$ [°]	3.75	2.81	2.81	2.00	1.88
$\Delta\phi$ [°]	3.75	2.81	2.81	2.50	1.88
N	4608	8192	8192	12672	9216

APPENDIX E

Implementation

For the numerical calculations performed in this work we have created the object oriented library "pyClimateNetworks" using the open source scripting language Python (van Rossum et al. (1991–2009)). "pyClimateNetworks" encapsulates methods for time series analysis, climate network construction and complex network analysis and at the same time enables a high reusability of code together with the flexibility to generate and analyze climate networks interactively from the Python shell. Another great advantage of using the interpreted Python language is platform independence, we can use "pyClimateNetworks" on a great variety of computer architectures with high efficiency without having to adjust or recompile the source code. This will particularly allow a straightforward parallelization of our library in the future.

Computationally expensive algorithms have been implemented by utilizing the high performance open source libraries Numpy (Oliphant (2006)), SciPy (Jones et al. (2001–2009)) and embedded C++ code segments within SciPy.Weave. For most of the graph theoretical computations we relied on the fast open source library iGraph and its Python interface (Csárdi and Nepusz (2006)). For plotting on maps we employed PyNGL (Computational & Information Systems Laboratory at the National Center for Atmospheric Research (NCAR) (2004–2008)), the Python interface to the NCAR Command Language (NCL). We used the open source package GraphViz for drawing graphs (AT&T Research and Bell Labs (2004–2009)). All other figures were created using the open source Python library matplotlib (Barrett et al. (2004)). All calculations were performed on workstations running Mac OS 10.5 and SUSE Linux.

Selbstständigkeitserklärung

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit mit dem Titel „Complex Networks in the Climate System“ selbstständig und ausschließlich unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Jonathan F. Donges
Potsdam, 31. März 2009