Advection-diffusion-networks

on the relationship of flow dynamics and climate network topology

DISSERTATION

zur Erlangung des akademischen Grades
doctor rerum naturalium
Dr. rer. nat.
im Fach Physik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von
Nora Molkenthin MASt. (Cantab)

Präsident der der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:
Prof. Dr. Elmar Kulke

Gutachter:
1. Prof. Dr.Dr.h.c. Jürgen Kurths
2. Prof. Dr. Emilio Hernández-García
3. Dr. Sc. Serhiy Yanchuk

Tag der mündlichen Prüfung: 6.11.2014

Abstract

The earth's climate is an extraordinarily complex, highly non-linear system with a multitude of influences and interactions between a very large number of variables and parameters. Complementary to the description of the system using global climate models, in recent years, a description based on the system's interaction structure has been developed. Rather than modelling the system in as much detail as possible, here time series data is used to identify underlying large scale structures. The challenge then lies in the interpretation of these structures.

In this thesis I approach the question of the interpretation of network measures from a general perspective, in order to derive a correspondence between properties of the network topology and properties of the underlying physical system. To this end I develop two methods of network construction from a velocity field, using the advection-diffusion-equation (ADE) for temperature-dissipation in the system.

For the first method, the ADE is solved for δ-peak-shaped initial and open boundary conditions. The resulting local temperature profiles are used to define a correlation function and thereby a network. Those networks are analysed and compared to climate networks from data. Despite the simplicity of the model, it captures some of the most salient features of climate networks. As the method allows the free choice of node locations, this model is used to study the influence of the node locations on the topology of the network.

The second network construction method relies on a discretisation of the ADE with a stochastic term. The resulting linear stochastic recursive equation can be used to define a correlation matrix, based only on the transition matrix and the variance of the uniformly random noise term. This allows the construction of larger networks, due to its relative computational simplicity. It is straightforward to generalize this method to systems, with more general transition matrices and time-dependent velocity fields. I construct weighted and unweighted networks for four different cases and suggest network measures, that can be used to distinguish between the different systems, based on the topology of the network and the node locations.

For better applicability, another related model is described, that represents aspects of the monsoon system of the past 1000 years, demonstrating that it is possible to capture climate changes in a simple model, that generates correlation structures.

All these methods provide networks of well-understood physical systems with parameters, that can be varied freely. To track and quantify the topological changes, resulting from parameter changes in the underlying system, a method is developed to quantify topological changes in linearly ordered sets of networks. This method is used to study changes in various types of networks.

The reconstruction methods presented in this thesis successfully model many features, found in climate networks from well-understood physical mechanisms. This can be regarded as a justification of the use of climate networks, as well as a tool for their interpretation.

Zusammenfassung

Das globale Klimasystem ist ein ausgesprochen komplexes und hochgradig nichtlineares System mit einer Vielzahl von Einflüssen und Interaktionen zwischen sehr vielen Variablen und Parametern. Komplementär zu der Beschreibung des Systems mit globalen Klimamodellen, wurde in letzter Zeit eine Beschreibung entwickelt, die auf der Interaktionsstruktur des Gesamtsystems beruht. Statt möglichst viele Details so genau wie möglich zu modellieren, werden hier Zeitreihendaten verwendet um zugrundeliegende Strukturen zu finden. Die Herausforderung liegt dann in der Interpretation dieser Strukturen.

Um mich der Frage nach der Interpretation von Netzwerkmaßen zu nähern, suche ich nach einem allgemeinen Zusammenhang zwischen Eigenschaften der Netzwerktopologie und Eigenschaften des zugrundeliegenden physikalischen Systems. Dafür werden im Wesentlichen zwei Methoden entwickelt, die auf der Analyse von Temperaturentwicklungen gemäß der Advektions-Diffusions-Gleichung (ADE) basieren.

Für die erste Methode wird die ADE mit offenen Randbedingungen und δ-peak Anfangsbedingungen gelöst. Die resultierenden lokalen Temperaturprofile werden verwendet um eine Korrelationsfunktion und damit ein Netzwerk zu definieren. Diese Netzwerke werden analysiert und mit Klimanetzen aus Daten verglichen. Trotz der Einfachheit des Modells, können einige der characteristischen Merkmale von Klimanetzen reproduziert werden. Da in dieser Methode die Lage der Knoten frei gewählt werden kann, eignet sie sich gut um den Einfluss der räumlichen Knotenverteilung auf die Netzwerktopologie zu untersuchen.

Die zweite Methode basiert auf der Diskretisierung der stochastischen ADE. Die resultierende lineare, stochastische Rekursionsgleichung wird verwendet um eine Korrelationsmatrix zu definieren, die nur von der Übergangsmatrix und der Varianz des stochastischen Störungsterms abhängt. Aufgrund des vergleichsweise geringen Rechenaufwands erlaubt dies die Konstruktion größerer Netzwerke. Die Methode kann auf allgemeinere Übergangsmatritzen ebenso einfach verallgemeinert werden wie auf zeitabhängige Prozesse. Ich konstruiere gewichtete und ungewichtete Netzwerke für vier verschiedene Fälle und schlage Netzwerkmaße vor, die zwischen diesen Systemen zu unterscheiden helfen, wenn nur das Netzwerk und die Knotenpositionen gegeben sind.

Ein weiteres verwandtes Modell beschreibt Aspekte des Monsun-Systems über die vergangenen 1000 Jahre und zeigt die Möglichkeit auf, die topologischen Auswirkungen von Klimaveränderungen in einem einfachen Modell zu untersuchen.

Die entwickelten Methoden erlauben die Konstruktion von Netzwerken in Abhängigkeit von freien Parametern. Um die Änderungen in der Topologie mit Parametervariationen zu untersuchen und quantifizieren wird eine Methode entwickelt, die topologische Veränderungen in einer geordneten Menge von Netzwerken beschreibt. Diese Methode wird auf verschiedene Netzwerktypen angewendet.

List of Publications

This thesis builds upon the following publications of research conducted during my PhD.

- Molkenthin, N.; Rehfeld, K.; Marwan, N.; Kurths, J.: Networks from Flows-From Dynamics to Topology. *Scientific reports* (2014), doi:10.1038/srep04119.
- Molkenthin, N.; Rehfeld, K.; Stolbova, V.; Tupikina, L.; Kurths, J.: On the influence of spatial sampling on climate networks. *Nonlinear Processes in Geophysics* 21 (2014), p. 651-657.
- Rehfeld, K.; Molkenthin, N.; and Kurths, J.: Testing the detectability of spatio–temporal climate transitions from paleoclimate networks with the START model, *Nonlinear Processes in Geophysics* 21 (2014), p. 691-703.
- Tupikina, L.; Rehfeld, K.; Molkenthin, N.; Stolbova, V.; Marwan, N.; Kurths, J.: Characterizing the evolution of climate networks. *Nonlinear Processes in Geophysics* 21 (2014), p. 705-711.
- Goppelsröder, F.; Molkenthin, N.: Mathematik/Geometrie.
  In: Günzel, S.; Mersch, D. [Eds.] : *Bild. Ein interdisziplinäres Handbuch.* J.B. Metzler, to appear in 2014.
- Molkenthin, N.; Tupikina, L.; Kurths, J.: Flow networks from the discretized advection-diffusion-equation. *Submitted to Physical Review E* (2014)

Earlier publications:

- Molkenthin, N.; Hu, S.; Niemi, A.J.: Discrete nonlinear Schrödinger equation and polygonal solitons with applications to collapsed proteins. *Physical Review Letters* 106 (2011), p. 078102.
- Eatough, R.P.; Molkenthin, N.; Kramer, M.; Noutsos, A.; Keith, M.J.; Stappers, B.W.; Lyne, A.G.: Selection of radio pulsar candidates using artificial neural networks. *Monthly Notices of the Royal Astronomical Society* 407 (2010), p. 2443–2450.

Contents

1 Introduction

Chaos is the score upon which reality is written
Henry Miller

Physics is an attempt to understand the world, and by understand of course we mean formulate equations and calculate. From the smallest to the largest scale, from elementary particles to galaxies. The trouble is that said world is incredibly complex. Any attempt to describe it requires simplifying assumptions and the singling out of one particular aspect of it. Be it an intricate experimental set-up to study a single atom in vacuum or the blunt assumption, that a thrown object essentially behaves like a point mass.
In most fundamental theories, technically, almost everything interacts with almost everything else all the time. Still, most interactions, even the behaviour of very large numbers of particles, like a gas, are described as a series of two-body interactions at most. The reason for the success of such a radical reductionism lies in the typically fast decrease of the interaction strength between the constituents with their relative distance. Theoretically of course the sun's movement is affected by the motion of humans on earth, but the effect is far too small to be measurable.
It should also be noted, that the abstraction typically depends on the problem. An apple can be described as a point mass if we want to compute where it lands, when we throw it, but if we are more interested in what it tastes like, a description of its molecular composition is probably more helpful. Knowing about the physical laws that govern the behaviour of an apple, the molecules the apple is made of, the atoms the molecules are made of and the elementary particles the atom is made of, helps us to understand a variety of aspects of apple behaviour on

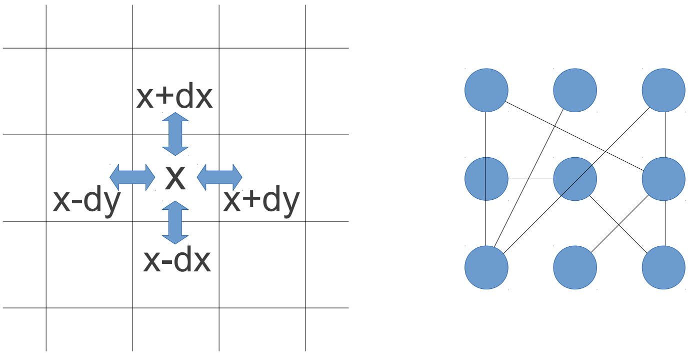


Figure 1.1: **Left:** The description of a system with differential equations is based on local values of a function and its derivatives. **Right:** The description of a system as a network is based on the interdependence of a property between all pairs of locations.

many scales.
With constituents and simple interactions a surprising number of phenomena can already be explained. But there are many other phenomena, that arise from more complex interactions of the constituents and those phenomena are much less well understood than the individual parts. While systems in physics have always been too complex to map them mathematically without approximation, usual simplifying assumptions include the assumption, that the medium is regular and can be represented as a grid or continuum, in which interactions happen only between the nearest neighbours (Fig. 1.1). Examples for this range from crystal structures to descriptions by differential equations. This provides a very accurate description for systems with a simple interaction structure, but the problem is, that partial differential equations are rarely solvable, especially when the interactions are non-linear and numerical evaluation shows chaotic behaviour.
One way to resolve this issue is to work with ensembles of trajectories, rather than individual ones and analyze the correlations of a variable, rather than its exact value. Hence, shifting the description from a description of local interactions to one, that relies on non-local correlations. This is the idea behind the study of correlation networks in systems, that can otherwise be approximated well by a continuous description.
The earth's climate system is one of the most complex systems. Influences include the sun, tectonics, vegetation and of course, and most controversially, humankind's activities on this planet

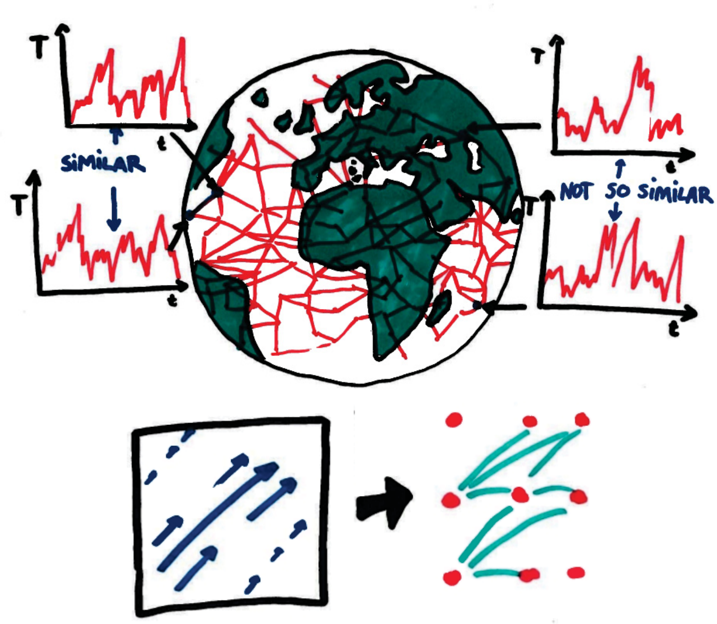


Figure 1.2: **Top:** Climate networks connect locations with similar climate data time series. **Bottom:** Network of a simple flow system.

[20]. While climate models get more and more complex as computing power grows, including more and more influences and parameters, the model output has more and more characteristics

of a “numerical experiment”, in which the influence of the parameters is observed rather than a known feature of the model, leaving the problem of interpretation and identification of underlying mechanisms unsolved.

The introduction of climate networks [45] is an attempt to approach the problem from a different perspective. The idea is to understand the large scale structure of the system by looking at the interactions between all of its parts, with the aim of finding mechanisms, by which simple constituents work together to produce complex phenomena.

In other fields, graph theory has already been applied with great success [44, 31]. Typical examples include social networks [25], the internet or transport networks [51]. In all of these cases, the nodes are clearly defined and distinct from each other. The nodes of a social network are people, their links are defined by the interaction, subject to the current study. There is no ambiguity in the definition of the nodes, because a person is a well-defined entity and the question whether person A’s arm might be part of person B would never occur. This is not the case in the definition of networks from a continuous system, as Fig. 1.3 illustrates.

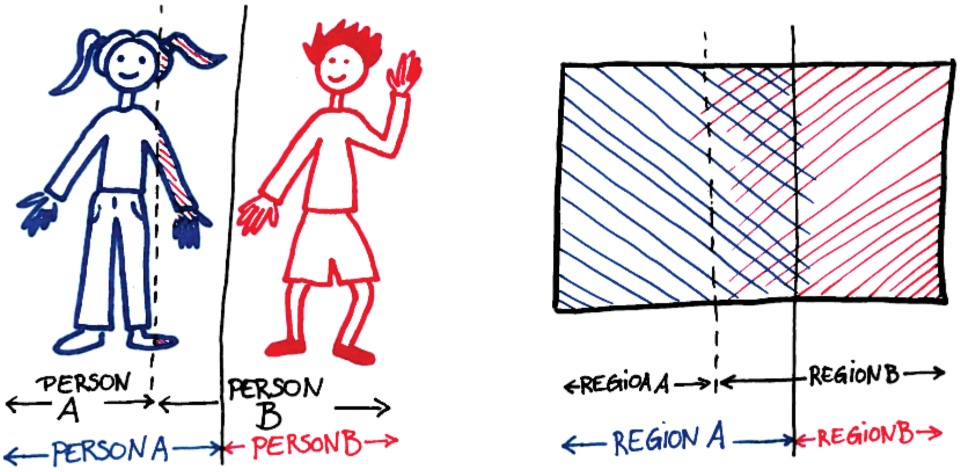


Figure 1.3: There is no ambiguity in the definition of nodes in a network of clearly separated discrete constituents, like a social network (a), in a network of a continuous system however, many different choices might be equally valid (b)

My thesis is concerned with some conceptual issues of networks of continuous systems and strives to look at the interpretation of climate networks from a new angle:

1. The description of continuous systems as networks, that are discrete by definition, is relatively new. Such a description reduces the system’s complexity in a fundamentally different way to the description with differential equations. Rather than only looking at local interactions, it takes all possible interactions into account, but reduces the spatial resolution to the node locations. This provides a fundamentally non-linear description and has recently given many important insights into the climate system. What can be said about such networks and their relation to differential equations in general?

2. Climate networks can be constructed in various ways from climate data time series, such as temperature, pressure or precipitation. Nodes in a climate network are typically associated with locations on the earth. Links are formed between two locations x_1 and x_2 if there is a high statistical interdependence between time series $T(x_1, t)$ and $T(x_2, t)$ (Fig. 1.2). The interdependencies can be measured with Pearson correlation, mutual information or event synchronization. Can we find a simple analytical model system, to produce networks, that are structurally and conceptually similar to climate networks?

3. An important issue of such climate networks is their interpretation. It is interesting to see the topological structures of the system, but as physicists and climatologists, we want to understand the underlying mechanisms. Many of the most interesting discoveries of climate network approaches aim to find patterns in network measures. Can we find a relationship between topological features and properties of the underlying system?

4. Given a method of network construction from a manageable physical system, how do construction parameters like spatial node distribution or influences other than the flow field itself impact network topology?

To answer these questions I developed two related but distinctly different methods. The first method starts from a continuous description of the flow system by the advection-diffusion-equation. A tracer peak (of temperature or concentration) is introduced and the equation is solved for its decay, giving a temperature profile over time for each location. Similar to the time series in traditional network construction, we define a continuous Pearson correlation and compute this for a set of node locations.

The second method starts out from a discretized advection-diffusion-equation with added random noise. This can be solved for random initial conditions and noise matrices with a mean of 0 and a variance of 1. In the Pearson correlation we also average over all realizations and obtain a correlation matrix, that is independent of the initial conditions, and which we interpret as a weighted network.

These models are used to develop more realistic models for climate applications, study the influence of inhomogeneous node distributions and verify a method for the characterization of changes in a network.

The outline of my thesis is the following:
Chapter 1 formulates questions and goals and provides an overview over the topic.
Chapter 2 introduces network concepts and gives definitions used throughout the thesis.
Chapter 3 derives network construction methods from flows for the continuous and discrete case.
Chapter 4 applies the new methods to study the influence of spatial sampling, temporal evolution and climate networks of the Asian monsoon region.
Chapter 5 summarizes the findings and further questions are formulated.

The main findings from this thesis have already been published. Specifically Chap. 3 is based on [28, 30] and Chap. 4 is based on [39, 29, 48].

1.1 Descriptions of complex continuous systems

A natural question, that arises when discussing the possibility of describing a continuous system as a network is the question why this is necessary and useful. What additional insights can be gained by reducing the continuous system to a discrete one and seemingly disregarding the principle of locality by allowing long-range links?

The principle of locality states that a local variable can only be influenced by its direct surroundings. Therefore, the typical approach to the description of most systems in physics uses differential equations:

$$F(x_i, U, U_{x_i}, U_{x_i x_j}, U_{x_i x_j x_k}, ...) = 0, \quad \text{where} \quad U_x = \frac{\partial U}{\partial x}. \tag{1.1}$$

The function $U(x_1, x_2, ..., x_n)$ of n variables $\{x_1, x_2, ..., x_n\}$ is described by a function F of the x_i and the derivatives of U with respect to these variables. A differential equation describes a process by its infinitesimal changes, or derivatives, thereby incorporating the principle of locality. This has been a very fruitful intuition for several centuries.

However, while the description of a system as a differential equation is as close to an exact description as we can hope to get, it yields many problems. Differential equations can rarely be solved exactly and while they are in principle deterministic, even smallest changes in the initial conditions can lead to very large differences in the (typically numerical) solution. This means that we have a description, that is in principle exact and should allow the prediction of the entire evolution of the system, but since small differences in the initial conditions have such a large impact on the systems state at later times, and measurements are only finitely accurate, predictions can only be made on relatively short time scales.

A natural solution of this issue is to work with averages. This can be averages over different realizations of the initial conditions or added noise. Instead of individual trajectories, correlations between different locations are computed and averaged over all realizations:

$$C(U(\vec{x}_i), U(\vec{x}_j)). \tag{1.2}$$

The correlation function describes persistent features of the system in a way, that is independent of the noise realization and can be interpreted as a complex network (by evaluating it at discrete node positions), giving insights into the topology of the interaction structure.

So instead of studying the evolution of the state at long time scales, we study the correlations induced by short term dynamics. In this thesis I will show approximate methods for solvable differential equations and compare them to networks obtained from data, that require no solution of differential equations. A general, continuous solution is beyond the scope of this work.

1.2 Networks and climate

The climate system is an extraordinarily complex continuous system [9]. It is a very large system with a vast number of parameters and influences, some of which, like advection and diffusion, can easily be modelled as differential equations [33]. The influences of human activity, however, are more difficult to model. Consequently, there are many climate models on all levels of complexity [42], depending on the required level of detail.

Following the reasoning from the section above, another possibility is to describe the climate as a correlation network. Since data is available, this would not even require the development or choice of a climate model, since the network can be computed directly from the available climate data time series.

Climate networks were first suggested in [45], and have been found fruitful in many applications [50, 47, 38, 41]. In climate networks, nodes are locations and connections are established if an observable, measured over time in two locations, is found to be correlated. In this way networks can be constructed heuristically from data or models, that allow the definition of a similarity measure, even without further knowledge of the exact form of the interaction.

They have provided important insights regarding various questions in climate sciences, ranging from the impact of the El Niño Southern Oscillation on global climate [46, 16, 2], to the dynamics of the Asian monsoons [26, 38], ocean [10, 17] and atmospheric dynamics [4, 13].

The construction of spatially embedded climate networks uses nodes corresponding to (geographical) locations, and links corresponding to statistical interdependence between climate time series observed at the locations of the nodes [45]. The strength of statistical dependence can be quantified in several different ways, depending on the type of data.

Probably the simplest and most common one is the Pearson correlation. Given two time series T_1 and T_2:

$$\rho(T_1, T_2) = \left\langle \frac{T_1 - \langle T_1 \rangle}{||T_1 - \langle T_1 \rangle\,||}, \frac{T_2 - \langle T_2 \rangle}{||T_2 - \langle T_2 \rangle\,||} \right\rangle . \tag{1.3}$$

Angular brackets $\langle . \rangle$ denote the expectation value or $\langle\, . \,,\, . \,\rangle$ the scalar product, the norm is defined as $||.|| = \langle (\,.\, - \langle\, . \,\rangle)^2 \rangle$.

This is also the correlation measure, that will be used predominantly in this thesis, as it can be used for deterministic functions as well as for noisy data, on top of that it has been shown, that most of the structure of the climate system is captured by the linear approach [19].

Other similarity measures include mutual information [11, 13], an entropy based measure, that is useful when dealing with non-linear correlations and event synchronization [36, 26, 43], a measure which compares event series.

Once the correlation matrix C_{ij} is found, the adjacency matrix of the unweighted network (Eq. 2.1) can be constructed by employing a threshold value α, using the the Heaviside theta function $\theta(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$ and the Kronecker delta $\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$:

$$A_{ij} = \theta(C_{ij} - \alpha) - \delta_{ij} \ . \tag{1.4}$$

Alternatively, the network can be weighted with the correlation values (Eq. 2.2):

$$w_{ij} = C_{ij}\theta(C_{ij} - \alpha) - \delta_{ij} \ . \tag{1.5}$$

In the weighted case, the thresholding is not strictly necessary, it does however reduce computational costs for some network measures, such as betweenness to work with sparse networks, while leaving the result almost unchanged.

2 Complex networks background

Handle nur nach derjenigen Maxime, von der du zugleich wollen kannst, dass sie allgemeines Gesetz werde, Silvio Berlusconi

Marc-Uwe Kling, *Das Känguru-Manifest*

A graph is a mathematical object, that describes a relational structure. The branch of mathematics, that concerns itself with such structures is called *graph theory.* For many years, graph theory has been found to be a useful tool for the description of various, very different types of systems in physics and other areas. Most of these systems, like the internet, the dynamics of disease spreading or power grids, have a natural network structure, in which nodes and links have a clear physical interpretation [44, 12, 35] and in many cases, the spatial embedding of the networks is not of interest.

Recently, however, the spatial embedding has attracted increased attention, as many systems rely on geometry as well as the topology. In power grids the length of the links is an important cost factor and in our area of application, the climate system, the physical distance of nodes has a large influence on their connection probability. Such networks will be referred to as *spatial networks.* To describe spatial aspects, some of the network measures, discussed here will take the geometry into account as well as the topology. In this section I introduce the network concepts and measures necessary for this thesis, for a more complete overview I refer the reader to [3, 1].

2.1 Definition

A *graph* $G = (V, E)$ is defined as an ordered pair of a set of nodes or vertices V and links or edges E. A link is a pair of nodes, that are connected.

Complex networks are graphs with a non-trivial topology. That means that they are neither grids nor random graphs. In a grid every node is structurally identical, this can be achieved by arranging the nodes on a regular spatial grid and connecting every node to its k nearest neighbours. In a random network [15, 14] every pair of nodes has the same probability to be connected, which also creates a kind of homogeneity of the nodes.

A *subgraph* G' of a graph G, is a graph with the vertex set $V' \in V$ and an edge set $E' \in E$.

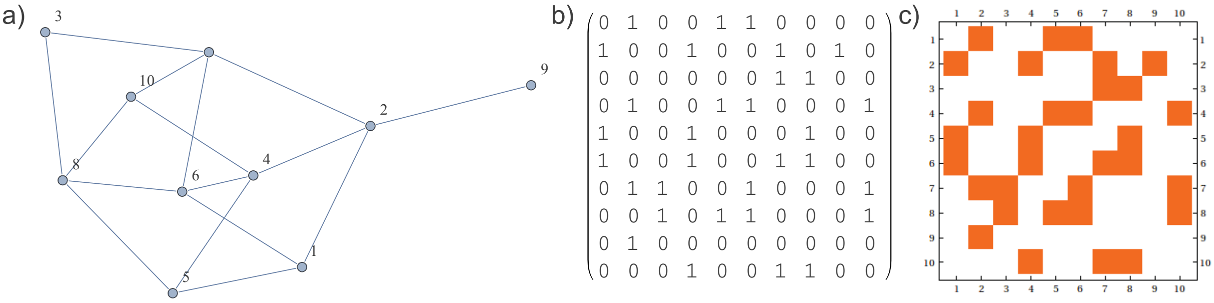
A graph is called *undirected*, if its links have no orientation.

2.2 Adjacency matrix

A graph of N nodes can be characterized by an $N \times N$-matrix, where each node corresponds to a row and column. The entry (i, j) is 1 if i and j are connected and 0 if they are not. Such a matrix is called an *adjacency matrix*. Fig. 2.1 b shows an adjacency matrix of a graph Fig. 2.1 a.

$$A_{i,j} = \begin{cases} 1 & \text{if nodes } i \text{ and } j \text{ are connected} \\ 0 & \text{if nodes } i \text{ and } j \text{ are not connected} \end{cases} \tag{2.1}$$

Note, that the adjacency matrix of an undirected graph is always symmetric.
Adjacency matrices are an intuitive representation of the graph and many network measures are defined using them directly. The matrix of all paths of length n, for example is the nth power of $\mathbf{A}$, while the degree is given by it's row sum.



$$\begin{pmatrix} 0&1&0&0&1&1&0&0&0&0 \\ 1&0&0&1&0&0&1&0&1&0 \\ 0&0&0&0&0&0&1&1&0&0 \\ 0&1&0&0&1&1&0&0&0&1 \\ 1&0&0&1&0&0&0&1&0&0 \\ 1&0&0&1&0&0&1&1&0&0 \\ 0&1&1&0&0&1&0&0&0&1 \\ 0&0&1&0&1&1&0&0&0&1 \\ 0&1&0&0&0&0&0&0&0&0 \\ 0&0&0&1&0&0&1&1&0&0 \end{pmatrix}$$

Figure 2.1: A graph can be displayed as: a) Graph, b) Adjacency matrix, c) Matrix plot

In terms of storage space, adjacency matrices are inefficient, particularly for sparse matrices, as they contain redundant information. In this respect, edge lists are more efficient, instead of storing zeros and ones, they only store the ones. In this thesis however, all graphs are relatively small, so the representation as adjacency matrix is sufficient.
The notion of an adjacency matrix can be generalized to describe weighted graphs, incorporating the connection strength. For a network with weights $w_{i,j}$ this notation is:

$$w_{i,j} = \begin{cases} c_{i,j} & \text{if nodes } i \text{ and } j \text{ are connected with connection strength } c_{i,j} \\ 0 & \text{if nodes } i \text{ and } j \text{ are not connected} \end{cases} \tag{2.2}$$

In this thesis, all networks are correlation networks. The correlation matrix can be interpreted directly as a weighted adjacency matrix. For reasons of computational efficiency, however, a lower cut-off is used, setting all entries below a certain threshold to zero, so that sparse algorithms can be employed.

2.3 Weighted vs. unweighted network measures

To analyse and describe the topological features of a network, a variety of network measures have been developed. Here I will present those used in this thesis. Network measures can describe properties of the network as a whole or single nodes or links. As the main motivation and application here is the climate system, we are particularly interested in measures, that can be interpreted as regional properties. Thus the emphasis lies on node properties. Since our networks are spatially embedded and their spatial embedding is an integral aspect of their structure, we also present network measures, that combine network topology and embedding geometry.

2.3.1 Degree or node strength

The *degree* k_i of node i of a network with N nodes is given by the number of nodes it is connected to, that is the row sum over row i of the adjacency matrix, defined in Eq. 2.1:

$$k_i = \frac{\sum_{j=1}^{N} A_{ij}}{N-1}. \tag{2.3}$$

The weighted version of the degree is called *node strength* s_i of node i and is given by the sum of the weights of node i's links, or the row sum over row i of the weighted adjacency matrix, defined in Eq. 2.2:

$$s_i = \frac{\sum_{j=1}^{N} w_{ij}}{N-1}. \tag{2.4}$$

Both degree and node strength are linear network measures, that only rely on the immediate surrounding of a node.

2.3.2 Betweenness

The *shortest path betweenness* b_i of a node i is defined as the number of all shortest paths that go through it,

$$b_i = \frac{\sum_{j,k\in 1,\ldots,N, j\neq k} \frac{n_{jk}(i)}{n_{jk}}}{(N-1)(N-2)}. \tag{2.5}$$

Where n_{jk} is the number of shortest paths connecting k and j and $n_{jk}(i)$ is the number of those paths, that go through i. This means that the entire network topology is needed to compute the betweenness of each individual node.

2.3.3 Connected component or community size

A *connected component* is a subgraph of a graph G, such that each pair of nodes within the subgraph is connected by a path, while no node of V' is connected to a node outside of V'. The size of a subgraph is the total number of its nodes.

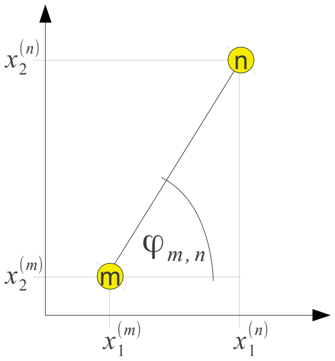
Since complete separation rarely occurs in climate networks a more subtle measure is used to analyse the community structure. Similar to the notion of connected components, we would like to find subgraphs of G, that are highly connected within themselves and have few connections to the other subgraphs. First let us define a division $d = d_1, ..., d_n$ as a set of n non-intersecting sets of vertices d_i. *Modularity* is a measure of how well a given division d divides the network into subnetworks, that have more links within themselves than interconnecting them. The modularity of a network with the adjacency matrix $\mathbf{A}$ and the division d, is defined as [8]:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(d_i, d_j), \tag{2.6}$$

where m is the number of edges in the network. The modularity is high if there are many links within the d_i and few links connecting two nodes from different sets. The goal is to find a good division into communities. For this there are optimization algorithms, like the one presented in [8], which is used throughout this work.

2.3.4 Anisotropy

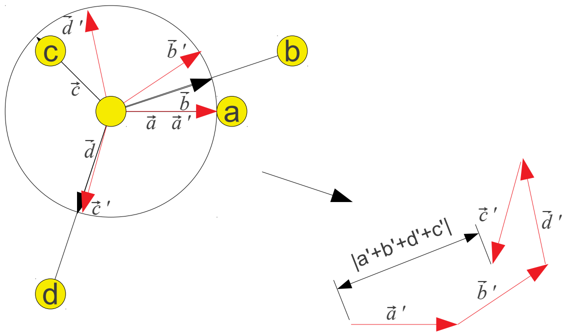
Unlike the previous network measure, the *local anisotropy*, as defined in [23], takes the spatial embedding into account as well as the topology. It is a measure for how aligned the links of a node are. For this the angles between all links between point m and its neighbours n_i and the reference axis are compared for all existing links. If the locations of the nodes m and n are given in Cartesian coordinates, $(x_1^{(m)}, x_2^{(m)})$ and $(x_1^{(n)}, x_2^{(n)})$, the angle is given by the equation in Fig. 2.2.



$$\varphi_{mn} = \arctan\left(\frac{x_2^{(n)} - x_2^{(m)}}{x_1^{(n)} - x_1^{(m)}}\right)$$

Figure 2.2: The angle φ_{mn} is the angle between the vector from vertex m to vertex n and the x-axis.

The anisotropy is the vector sum of unit vectors along all connections of node m. The vectors are described by their length A_{mn}, which is 1 for existing links and 0 otherwise (see Fig. 2.3).



$$R_m = \frac{1}{k_m}\left|\sum_{n=1}^{N} A_{mn} e^{i2\varphi_{mn}}\right|. \qquad (2.7)$$

Figure 2.3: Each link corresponds to a vector of the same angle φ_{mn} (black) of length 1. The red vectors have the phase $2\varphi_{mn}$ and are added to find R_m.

The angles φ_{mn} are doubled. This ensures that perpendicular vectors cancel, while an angle of π adds up. We define the *weighted local anisotropy* as:

$$R_m = \frac{1}{s_m} \left| \sum_{n=1}^{N} w_{mn} e^{i2\varphi_{mn}} \right| . \tag{2.8}$$

For weighted networks we also have to take the connection strength into account. It comes in as the length of the vectors.

2.3.5 Link length distribution

The *link length distribution* is a global network measure, that, like anisotropy, combines topology of the network and the geometry of the embedding. It is a histogram of frequencies of link lengths in the network.

The *complete link length distribution* is the link length distribution of the complete graph of a set of spatially embedded nodes. The complete link length distribution can be used to compare to the actual link length distribution. If, for example, the spatial distance is unrelated to the connection probability, the actual link length distribution should differ from the complete link length distribution only by a factor $c < 1$, equal to the link density. On the other hand, if the network is a unit disk graph, as described in [7], of a given node distribution with disk size r, the link length distribution is the same as the complete link length distribution, cut off at link length r.

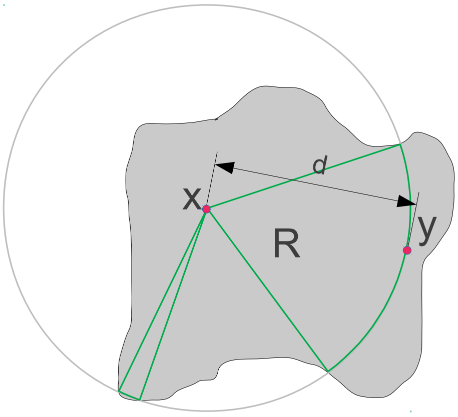


Figure 2.4: The frequency of distance d in area R is the integral over all line segments, that lie inside R (green), integrated over all positions of the origin x.

The continuous equivalent of the complete link length distribution is the *distance distribution*, which is defined as the integral over the distances between all pairs of points x and y in a region R. This is equivalent to integrating over the area of the region and all circle segments, that lie within, of circles of radius d as illustrated in Fig. 2.4.

2.3.6 Voronoi tessellation analysis

A Voronoi tessellation is a division of a space into non-overlapping regions, such that every point of the space belongs to one region. The regions are defined around a set of nodes, such that every point in the cell is closer to that node than to any other, as shown in Fig. 2.5. If X is a metric space with a distance function d and there are N nodes at the locations $p_1, ..., p_n$, then the cell R_i around point i is:

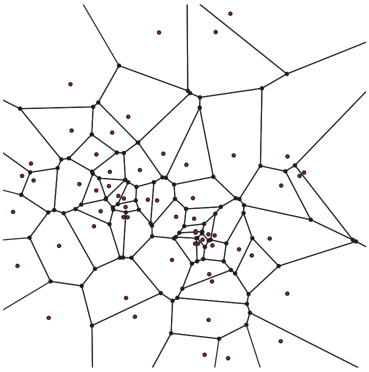


Figure 2.5: The Voronoi cell of a node n includes all points, that are closer to n than to any other node. Here the nodes are the red dots, the tessellation is depicted as black lines.

$$R_i = \{x \in X \mid d(x, p_i) \leq d(x, p_j) \, \forall \, j \neq i\} \tag{2.9}$$

Voronoi tessellation analysis is a simple diagnostic method to estimate the homogeneity of a node distribution.

The *Voronoi cell size distribution* is a histogram, computed from the sizes of the areas of all Voronoi cells. In a grid all cells are identical, thus the cell size distribution consists of a single line at a^2, where a is the grid constant. If the nodes are uniformly distributed in space the resulting cell sizes will have a clear peak around the mean cell size A_{sam}/N_{nod}, where A_{sam} is the size of the total sampled area and N_{nod} is the number of nodes. If the nodes are scattered around a centre with small distances in the centre and large ones in the outer regions, the peak will be shifted to smaller cell sizes.

3 Construction of networks from flows

From the maelstrom of the knowledge
Into the labyrinth of doubt
Frozen underground ocean
Melting - nuking on my mind
Gogol Bordello *Supertheory Of Supereverything*

In this chapter I first present the general idea for the bottom-up construction of networks from a well-understood physical system. This is done for a continuous system first, which has the advantage of allowing to freely distribute the nodes in space, but is computationally expensive. Therefore an alternative method is introduced, that is based on a discretisation of the advection-diffusion-equation (ADE). This model makes it possible to generalize the method to time dependent flows, allow external heating, interaction with another layer of flow or obstructions for the tracer.
Sect. 3.2 is based on my paper [28] and largely follows its presentation. Sect. 3.3 closely follows my publication [30].

3.1 Question and approach

Simply put, the starting question is: *What is the climate network of a bucket of water at rest?* Which later transforms into the advanced question: *What is the climate network of an inhomogeneous velocity field with external heating or interaction with a secondary velocity field?*
Here, I approach the problem of the interpretation of climate networks from a new angle. Instead of starting from climate data, constructing a network and interpreting the network based on climatological knowledge, I apply a bottom-up approach. The idea is to construct complex networks from a simple model, which is physically well-understood and shares crucial properties with the climate system. This allows me to draw a connection between network topology and properties of the underlying system.
As outlined before, climate networks are constructed by choosing node locations with climate data time series and evaluating pairwise similarity measures between all pairs of nodes. The resulting correlation matrix is either thresholded or used as a weighted adjacency matrix. The reason for a high correlation can be that there is a *direct connection*, meaning that fluctuations are transported from one node to the other within the system in question, *indirect connection*, meaning that fluctuations are transported from node to node via another system, and *common driver*, which is an external forcing, that acts on both nodes [37].
It appears reasonable to assume, that a lot of climate's direct interactions happen via simple advection and diffusion in the ocean or the atmosphere. So the ADE provides a good starting point. Given a two dimensional flow field and an initial distribution of a tracer quantity

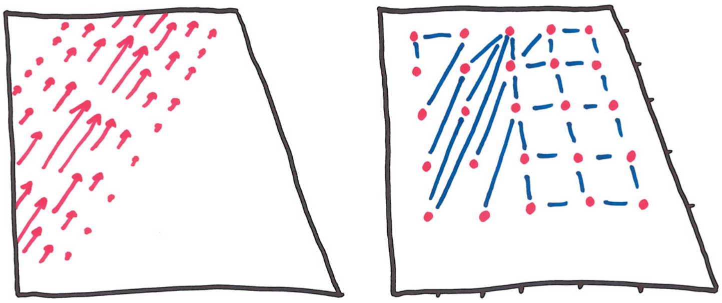


Figure 3.1: We construct a complex network (right) from a physical flow (left).

(which will be called "temperature", throughout this thesis), the ADE describes how the tracer distribution in the flow field changes over time. The resulting tracer changes over time in each location can be viewed as the theoretical equivalent to the measured time series, used to construct climate networks. Thus, they are used to define a correlation and construct a theoretical network.
This is done here in two different ways, allowing to investigate different aspects of the system. The two approaches are presented in Sect. 3.2 (continuous model) and Sect. 3.3 (discrete model).

3.2 Continuous method

The starting point for this model is a continuous two dimensional fluid. To model temperature fluctuations, that decay too fast to interact with each other, a δ-peak is chosen as initial conditions. Temperature transport in the fluid obeys the deterministic advection-diffusion-equation (ADE). The location of the initial peak should be irrelevant, because the system to be analysed is the velocity field. We therefore average over all realizations of the initial condition, which is all locations of the peak.
By also integrating over a time window ranging from shortly after the initial conditions until a time long after everything has decayed, we define a continuous correlation measure, which is then evaluated between all pairs of node positions to construct a correlation matrix and finally a network.

3.2.1 The advection-diffusion-equation

The ADE is a parabolic partial differential equation, that states how the physical quantity of interest (here temperature), is transported in a liquid. Local temperature changes over time are governed by the spatial temperature change, the velocity and the diffusion coefficient in the fluid. For simplicity, the system is restricted to the sourceless, incompressible ADE:

$$\frac{\partial T}{\partial t} = \chi \Delta T - \nabla \cdot (\vec{v}(\vec{x})T), \tag{3.1}$$

where χ is the diffusivity of the fluid and $\vec{v}(\vec{x})$ is its velocity at location $\vec{x}$. The ADE is obtained by inserting the advective and diffusive flux

$$\vec{j} = \vec{j}_{diff} + \vec{j}_{adv} = -\chi \nabla T + \vec{v} T \tag{3.2}$$

into the sourceless continuity equation for temperature

$$\frac{\partial T}{\partial t} = -\nabla \cdot \vec{j} \ \ . \tag{3.3}$$

Here, $T(\vec{x}, t)$ is the value of the temperature in position $\vec{x}$ at time t. For a homogeneous velocity field $\vec{v}(x) = v$ in one dimension, this can easily be solved by making the coordinate transformation:

$$\begin{aligned} t' &= t \\ x' &= x - vt \end{aligned}$$

Leading to the simple heat equation:

$$\frac{\partial T}{\partial t} = \chi \frac{\partial^2 T}{\partial x^2}, \tag{3.4}$$

Which can be solved using Fourier transformations, as explained in [24]. To solve a partial differential equation, it is necessary to fix initial and boundary conditions: The temperature is taken to be 0 at $\pm\infty$. Furthermore, the interactions are restricted to a flat 2D Euclidean space. The δ-peak shaped initial conditions, located at $\vec{x}_0$, serve as a simplified model for the local temperature fluctuations. So finally the Cauchy problem of Eq. (3.1) is solved with the initial conditions:

$$T(\vec{x}, 0; \vec{x}_0) = \delta(\vec{x} - \vec{x}_0) \ . \tag{3.5}$$

For homogeneous velocity fields, the ADE can then be solved analytically, giving:

$$T(x,t) = \frac{e^{-\frac{(x-x_0-vt)^2}{4\chi t}}}{\sqrt{4\pi\chi t}} + T_0 \tag{3.6}$$

for the temperature development of a δ-peak, located at x_0, as initial temperature distribution. More complex flows $\vec{v}(\vec{x})$ require an approximate solution to the ADE, as a direct analytic derivation becomes impossible. The stationary flows are assumed to vary slowly over space, $|\nabla \vec{v}(\vec{x})| \ll \chi$ (Unless it is stated otherwise, we set $\chi = 1$). As a consequence, all derivatives of the velocity field can be ignored, which is necessary to ensure the applicability of the approximation. In d dimensions this leads to the approximated temperature field

$$T_{appr}(\vec{x}, t; \vec{x}_0) = \frac{e^{-\frac{|\vec{x}-\vec{x}_0-\vec{v}(\vec{x})t|^2}{4\chi t}}}{\sqrt{4\pi\chi t}^d}. \tag{3.7}$$

To evaluate the validity of this assumption for a given velocity field $\vec{v}(\vec{x})$, we compute a diagnostic residual R from:

$$R(\vec{x},t) = \frac{\partial T_{appr}}{\partial t} - \chi \Delta T_{appr} + \nabla \cdot (\vec{v}(\vec{x}) T_{appr}) \ . \tag{3.8}$$

$R(\vec{x},t)$ is zero for a perfect solution, which is the case if $|\nabla \vec{v}(\vec{x})| = 0$. If the maximum of this function is small compared to the other terms in equation 3.8, $R(\vec{x},t) \ll \frac{\partial T_{appr}}{\partial t}$, the approximation is considered to be good. This condition was checked numerically for all examples in this thesis.

3.2.2 Definition of the correlation

In analogy to the wide-spread Pearson correlation [6] (Eq. 1.3), the continuous cross-correlation analogue (CCA) is defined as the normed scalar product of solutions of the Cauchy problem of the ADE at two points $\vec{x}_1$ and $\vec{x}_2$

$$C(\vec{x}_1, \vec{x}_2) = \left\langle \frac{T(\vec{x}_1, t; \vec{x}_0)}{||T(\vec{x}_1, t; \vec{x}_0)||}, \frac{T(\vec{x}_2, t + t_l; \vec{x}_0)}{||T(\vec{x}_2, t + t_l; \vec{x}_0)||} \right\rangle . \tag{3.9}$$

The scalar product is then defined as the integral over time and peak position $\vec{x}_0$. The time integration is analogous to the sum over time steps, the integration over the peak position serves as an average over realizations of the peak, corresponding to stochastics in the time series, where peaks appear at random in arbitrary places. So we define the CCA in this context as:

$$C(\vec{x}_1, \vec{x}_2) = \frac{\frac{1}{(t_1 - t_0)} \frac{1}{V} \int_{t_0}^{t_1} \int_{\mathbb{R}^2} T(\vec{x}_1, t; \vec{x}_0) T(\vec{x}_2, t + t_l; \vec{x}_0) d\vec{x}_0 dt}{||T(\vec{x}_1, t; \vec{x}_0)|| || T(\vec{x}_2, t + t_l; \vec{x}_0)||}, \tag{3.10}$$

where $\vec{x}_0$ is the position of the peak. The time lag t_l is the difference in travel time of the peak from $\vec{x}_0$ to $\vec{x}_1$ and $\vec{x}_2$. Note that, due to the much finer time resolution in the model, this time lag is different from those commonly used in data analysis, where the time span between one data point and the next is often much longer than the time lag would be.

$$t_l = t_{max}(\vec{x}_1, \vec{x}_0) - t_{max}(\vec{x}_2, \vec{x}_0), \tag{3.11}$$

where $t_{max}(\vec{x}, \vec{x}_0)$ is the time when the temperature at $\vec{x}$ reaches its maximum, with the initial peak starting at $\vec{x}_0$. The norm is defined as:

$$||T(\vec{x}, t; \vec{x}_0)|| = \sqrt{\frac{1}{(t_1 - t_0)} \frac{1}{V} \int_{t_0}^{t_1} \int_{\mathbb{R}^2} T(\vec{x}, t; \vec{x}_0)^2 d\vec{x}_0 dt.} \tag{3.12}$$

The factors $\frac{1}{(t_1 - t_0)} \frac{1}{V}$ cancel out due to the division by the norm in Eq. 3.10. Applying the

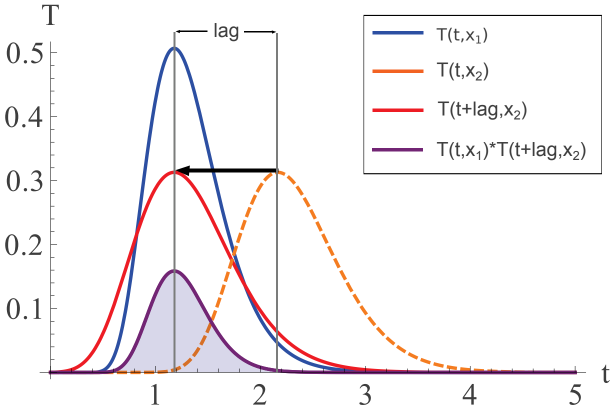


Figure 3.2: The continuous analogue of the Pearson correlation is defined using the temperature profiles at $\vec{x}_1$ (blue) and $\vec{x}_2$ (dashed orange). They are aligned such that their peaks (dashed orange → red) coincide and multiplied (purple). The CCA is given by the integral over time and initial peak location.

solution of Eq. 3.7 in Eq. 3.10, the expression simplifies due to a Gaussian integration:

$$\begin{aligned} &||T(\vec{x}_1,t;\vec{x}_0)||||T(\vec{x}_2,t+t_l;\vec{x}_0)|| \\ &\quad= \sqrt{\int_{t_0}^{t_1}\int_{\mathbb{R}^2} T(\vec{x}_1,t;\vec{x}_0)^2 dt d\vec{x}_0}\sqrt{\int_{t_0}^{t_1}\int_{\mathbb{R}^2} T(\vec{x}_2,t;\vec{x}_0)^2 dt d\vec{x}_0} \\ &\quad= \sqrt{\int_{t_0}^{t_1}\int_{\mathbb{R}^2} \frac{e^{-2\frac{|\vec{x}_1-\vec{x}_0-\vec{v}(\vec{x}_1)t|^2}{4\chi t}}}{\sqrt{4\pi\chi t}} dt d\vec{x}_0}\sqrt{\int_{t_0}^{t_1}\int_{\mathbb{R}^2} \frac{e^{-2\frac{|\vec{x}_2-\vec{x}_0-\vec{v}(\vec{x}_2)t|^2}{4\chi t}}}{\sqrt{4\pi\chi t}} dt d\vec{x}_0} \\ &\quad= \int_{t_0}^{t_1} \frac{1}{8\pi\chi t} dt = \frac{1}{8\pi\chi t}(\log(t_1)-\log(t_0)) \end{aligned} \tag{3.13}$$

The lower limit of the integration, t_0 is chosen small but non-zero (here $t_0 < 10^{-2}$) as the correlation function is not defined for $t = 0$. The upper limit is chosen such that all temperature profiles have decayed to a value very close to zero (here: $t_1 = 5000$).

The CCA is evaluated between all node pairs (i,j) from a given set of nodes. This provides the correlation matrix C_{ij} from which the adjacency matrix A is constructed by choosing a fixed significance threshold α. This can be expressed with the Heaviside θ function and Kronecker δ as

$$A_{ij} = \theta(C_{ij}-\alpha) - \delta_{ij}\ . \tag{3.14}$$

α is chosen such that the adjacency matrix is sparse, yet almost all nodes are connected by a path. In all cases studied, the main features of the network topology were robust with changes of the threshold (compare Fig. 3.11). A link density of 10-20 % was found to satisfy these requirements best.

3.2.3 Application to different velocity fields

This method can be applied for many different velocity fields. The first step is always to solve the ADE Eq. 3.1, the result is then used to compute the correlation matrix using Eq. 3.10. The ADE can be solved analytically for $\vec{v} = 0$, $\vec{v} = const.$ (translation) and $\vec{v} = \begin{pmatrix} y \\ -x \end{pmatrix}$ (rotation). The second step is to compute the correlation function. This can be done exactly only for zero velocity. For constant velocities it can still be done approximately, giving a closed solution. In the case of more general, slowly varying velocity fields, the time integral has to be done numerically. Here I present the results for the different cases.

i) For $\vec{v} = 0$

The movement of the peak's maximum is dominated by the advection, which is zero here. Therefore the time lag is approximately zero, $t_l \approx 0$. This simplifies the problem sufficiently so that the correlation function (see Eq. 3.10) can be computed analytically, giving:

$$C(\vec{x}_1, \vec{x}_2) = \frac{\Gamma(0, \frac{|\vec{x}_1 - \vec{x}_2|^2}{8\chi t_1}) - \Gamma(0, \frac{|\vec{x}_1 - \vec{x}_2|^2}{8\chi t_0})}{\log(\frac{t_1}{t_0})} \tag{3.15}$$

Where t_0 and t_1 are the limits of the time integration and $\Gamma(a, x)$ is the upper incomplete gamma function. Fig. 3.3 shows how the correlation behaves with the distance of points $\vec{x}_1$ and $\vec{x}_2$. To obtain a network from this, node locations $\vec{x}_i$ are chosen and Eq. 3.15 is applied to find their pairwise correlations. The correlation strength in this case only depends on the distance of the nodes. If the node locations form a square lattice, the resulting network is a regular grid.

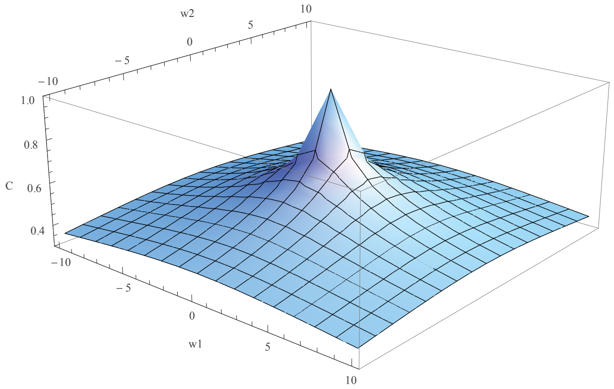


Figure 3.3: Correlation measure for $\vec{v} = 0$, plotted over distance vector $\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$ between the points. For zero distance, the correlation is maximal and drops off symmetrically with distance.

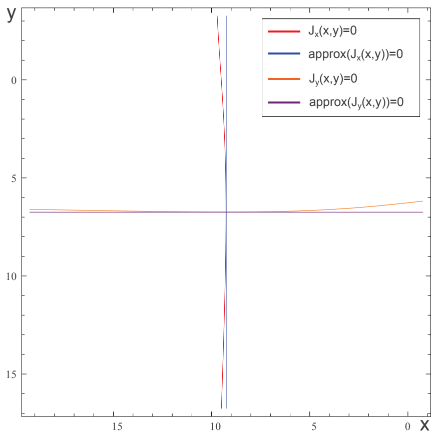


Figure 3.4: The peak is located at the values for $\vec{x}$, for which both components of the gradient are zero, which happens at the intersection of the zero-lines of J_x (red) and J_y (orange). To find the peak location, Newton's method is applied (blue and purple lines).

ii) For $\vec{v} = const.$

For homogeneous velocity fields the correlation can still be computed analytically with a few approximations. The added challenge lies in the time lag, which, instead of zero, is now given by Eq. 3.11, where the following expression for the maximum can be obtained from the roots of the numerator of the exponent, $|\vec{x} - \vec{x}_0 - \vec{v}t|^2$, from Eq. 3.7:

$$t_{max}(\vec{x}, \vec{x}_0) = \frac{\sqrt{\chi^2 + |\vec{v}|^2 |\vec{x} - \vec{x}_0|^2}}{|\vec{v}|^2}. \tag{3.16}$$

In this case, Eq. 3.7 is the exact solution of the ADE. Using the definition of the correlation Eq. 3.10, gives the following expression:

$$C(\vec{x}_1, \vec{x}_2) = \frac{\int_{t_0}^{t_1} \int_{\mathbb{R}^2} \frac{1}{(4\pi\chi t)^2} \frac{t}{t+t_l} e^{-\frac{|\vec{x}_1 - \vec{x}_0 - \vec{v}(\vec{x}_1)t|^2 + \frac{t}{t+t_l}|\vec{x}_2 - \vec{x}_0 - \vec{v}(\vec{x}_2)(t+t_l)|^2}{4\chi t}} dt d\vec{x}_0}{\frac{1}{8\pi\chi t}(\log(t_1) - \log(t_0))} \tag{3.17}$$

To approximately evaluate it, Laplace's method [5] is applied. Laplace's method is a technique for the estimation of integrals of the form: $\int_a^b h(y) e^{-g(y)} dy$, where $g(y)$ is a smooth function with a unique global minimum. As the exponential decays much faster than the non-exponential prefactor $h(y)$ varies, all significant contributions to the integral come from a narrow neighbourhood of the location of the minimum of $g(y)$. So $g(y)$ is replaced by its Taylor expansion to second order around the location y_{max} of the minimum of $g(y)$:

$$\begin{aligned}\int_a^b h(y)e^{-g(y)}dy &\approx h(y_{max})e^{-g(y_{max})}\int_{-\infty}^{\infty} e^{g''(y_{max})(y-y_{max})^2/2}dy \\ &\approx h(y_{max})e^{-g(y_{max})}\sqrt{\frac{2\pi}{g''(y_{max})}}.\end{aligned} \tag{3.18}$$

Where y_{max} is the value of y, for which $g(y)$ takes on the global minimum and therefore the peak reaches its maximum.
For the evaluation of Eq. 3.17, the functions $h(\vec{y})$ and $g(\vec{y})$ are chosen to be:

$$\begin{aligned}h(\vec{y}) &= \frac{1}{(4\pi\chi t)^2}\frac{t}{t+t_l(\vec{y})} \\ g(\vec{y}) &= \frac{|\vec{x}_1-\vec{y}-\vec{v}(\vec{x}_1)t|^2+\frac{t}{t+t_l(\vec{y})}|\vec{x}_2-\vec{y}-\vec{v}(\vec{x}_2)(t+t_l(\vec{y}))|^2}{4\chi t}\end{aligned} \tag{3.19}$$

The two dimensional equivalent of g'' is the Hessian matrix, which transforms Eq. 3.18 as follows:

$$\int_{\mathbb{R}^2} h(\vec{y})e^{-g(\vec{y})}d\vec{y} \approx h(\vec{y}_{max})e^{-g(\vec{y}_{max})}\frac{2\pi}{\sqrt{\det H(g(\vec{y}_{max}))}}, \tag{3.20}$$

where H denotes a Hessian matrix. The Gaussian integral is then evaluated. The main challenge left, is locating the maximum of the peak. By definition it is in the location, where both components of J are zero simultaneously (Intersection of zero lines for J_x and J_y in Fig. 3.4), but this system of equations can not be solved analytically for the peak location. The peak position has to be approximated using Newton's method (here one iteration) to find the root of the gradient of g.

$$\vec{x}_{m,n+1} = \vec{x}_{m,n} - H^{-1}(\vec{x}_{m,n}) * J(\vec{x}_{m,n}) \tag{3.21}$$

The initial guess is a peak, that moves with the speed $\vec{v}$, starting from $\vec{x}_1$:

$$\vec{x}_{m,0} = \vec{x}_1 - \vec{v}(\vec{x}_1)t \tag{3.22}$$

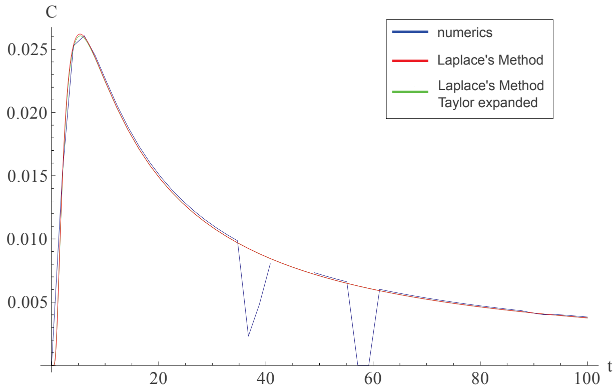


Figure 3.5: The spatial integration results in the integrand of Eq. 3.23, here it is plotted over time: numerically (blue line), using Laplace's method (red line) and using Laplace's method, but additionally Taylor expanding $\frac{2\pi}{\sqrt{\det H(\vec{x_{max}})}}$ to first order (green line).

After all these approximations the correlation function can now be written out as:

$$C(\vec{x}_1, \vec{x}_2) \approx \int_{t_0}^{t_1} \frac{2\pi}{\sqrt{\det H(\vec{x}_{max})}} \frac{1}{(4\pi\chi t)^2} \frac{t}{t+t_l} e^{-\frac{|\vec{x}_1 - \vec{x}_{max} - \vec{v}t|^2 + \frac{t}{t+t_l}|\vec{x}_2 - \vec{x}_{max} - \vec{v}(t+t_l)|^2}{4\chi t}} dt \qquad (3.23)$$

Fig. 3.5 shows how the approximation of the spatial integral relates to its numerical evaluation. The integral poses a difficult numerical problem, leading to errors in the numerical evaluation.

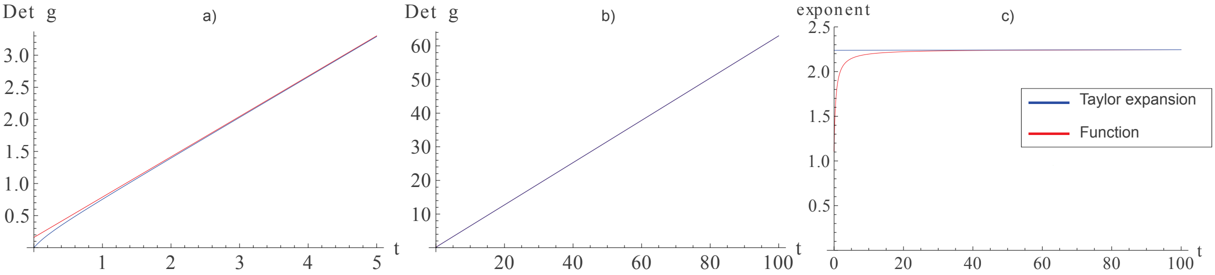


Figure 3.6: Taylor expansions for prefactor and exponent together with the full function for the case of a homogeneous velocity field: a) $2\pi/\sqrt{\det H(\vec{x}_{max})}$ with it's first-order Taylor expansion plotted from 0 to 5, to magnify the discrepancy, b) over the full interval t_0 to t_1, c) numerator of the exponent $|\vec{x}_1 - \vec{x}_{max} - \vec{v}t|^2 + \frac{t}{t+t_l}|\vec{x}_2 - \vec{x}_{max} - \vec{v}(t+t_l)|^2$ with its first-order Taylor expansion.

To evaluate the time integration, the gauss integral factor $2\pi/\sqrt{\det H(\vec{x}_{max})}$ and the numerator of the exponent $|\vec{x}_1 - \vec{x}_{max} - \vec{v}t|^2 + \frac{t}{t+t_l}|\vec{x}_2 - \vec{x}_{max} - \vec{v}(t+t_l)|^2$, are expanded in a Taylor series to first order around t_1 (see Fig. 3.6). The result is of a form, that can be integrated analytically over time:

$$\int \frac{e^{a_0 - \frac{a_1}{t}}(c_0 + c_1 t)}{b_1 t + b_2 t^2}\, dt = -\frac{e^{a_0}\left(e^{\frac{a_1 b_2}{b_1}}(b_2 c_0 - b_1 c_1)\mathrm{Ei}\left(-\frac{a_1(b_1+b_2 t)}{b_1 t}\right) + b_1 c_1 \mathrm{Ei}\left(-\frac{a_1}{t}\right)\right)}{b_1 b_2} \tag{3.24}$$

where the a_i, b_i and c_i are given by functions of the Taylor coefficients of the expansions. The exponential integral function $\mathrm{Ei}(x)$ is defined as:

$$\mathrm{Ei}(x) = -\int_{-x}^{\infty} \frac{e^{-t}}{t} dt \tag{3.25}$$

Finally, a set of nodes is chosen and the correlation is evaluated for every pair of nodes.
As the velocity field is homogeneous, the resulting network is also homogeneous, apart from boundary effects. This implies that an analysis of network measures, provides no interesting information. However, in contrast to the $\vec{v} = \vec{0}$ case, the network is not isotropic. Links are longer in flow direction than perpendicular to it (see Fig. 3.7).

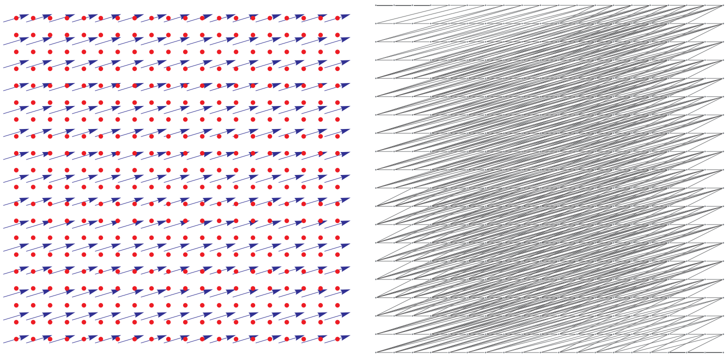


Figure 3.7: Network for the velocity field: $v_x(x,y) = 1$, $v_y(x,y) = 0.3$ with 20×20 nodes arranged in a regular grid.

iii) The circular flow

Like the homogeneous flow, the circular flow field $\vec{v}(\vec{x}) = \omega \begin{pmatrix} y \\ -x \end{pmatrix}$ can be described as a coordinate transformation of the resting case. Using the replacements: $\vec{x}' = \begin{pmatrix} \cos\omega t & -\sin\omega t \\ \sin\omega t & \cos\omega t \end{pmatrix} \vec{x}$ leads to an analytical expression for the temperature evolution that is (with $\omega = 1$):

$$T(\vec{x}, t; \vec{x}_0) = \frac{1}{4\chi t} e^{-\frac{(x\cos t + y\sin t - x_0)^2 + (-x\sin t + x\cos t - y_0)^2}{4\chi t}} \tag{3.26}$$

The spatial integration was done analogously to the homogeneous field.
While $2\pi/\sqrt{\det H(\vec{x}_{max})}$ can still be expanded to first order in a Taylor series, in a reasonable approximation (see Fig. 3.8 a and b), the exponent now has to be expanded to the second order in t (Fig. 3.8 c). With the second order expansion, however, the integration can not be done

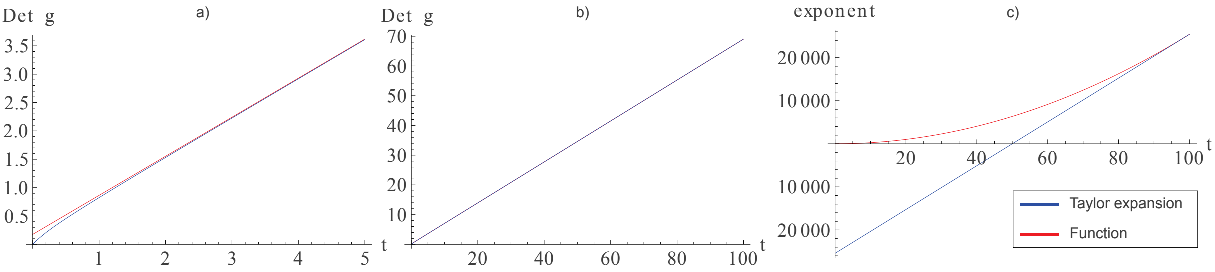


Figure 3.8: Taylor expansions for prefactor and exponent together with the full function for the case of the circular flow field: a) $2\pi/\sqrt{\det H(\vec{x}_{max})}$ with it's first-order Taylor expansion plotted from 0 to 5, to magnify the discrepancy, b) over the full interval t_0 to t_1, c) numerator of the exponent $|\vec{x}_1 - \vec{x}_{max} - \vec{v}t|^2 + \frac{t}{t+t_l}|\vec{x}_2 - \vec{x}_{max} - \vec{v}(t+t_l)|^2$ with its first-order Taylor expansion.

analytically and a numerical evaluation has to be performed.
A grid of 20 × 20 nodes was chosen and the correlation was evaluated for every pair of nodes. The resulting network is shown in Fig. 3.9. As expected, there are more longer links in the faster flowing outer regions of the flow. As in the homogeneous case, links are longer in flow direction, than perpendicular to it. The result is less regular than the network of the homogeneous flow

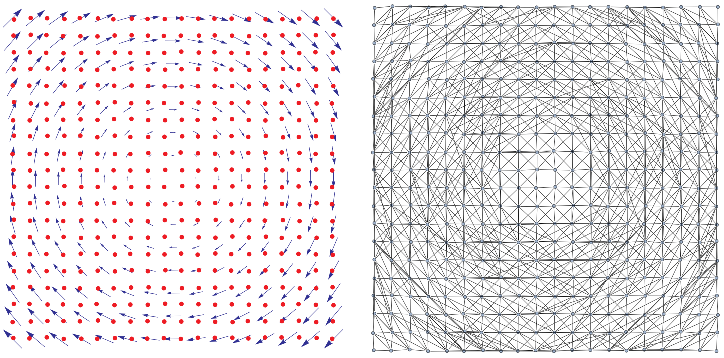


Figure 3.9: Network for the velocity field: $v_x(x,y) = 1$, $v_y(x,y) = 0.3$ with 20 × 20 nodes arranged in a regular grid. Figure from [28], supplementary material.

shown in Fig. 3.7, but the symmetry of the flow is well visible. The network measures k_i (see Eq. 2.3) and b_i (see Eq. 2.5) are dominated by boundary effects.

iv) For inhomogeneous velocities

Similarly to the circular velocity field, slowly varying (low-gradient) inhomogeneous velocity fields can be treated using a combination of analytical approximations and numerical evaluation. This was done for the two paradigmatic flows shown in Fig. 3.10. One is composed of three narrow parallel flows, with alternating directions, and the other flow is made up of two narrow flows intersecting in the middle. Using the approximate solution of the ADE, given in Eq. 3.7, the correlations in a 20×20-grid are computed for all pairs of nodes. Any pair of sites with a correlation larger than α is connected with a link in the network. The threshold α is determined such that the link density $\rho = \frac{L_{net}}{L_{full}}$ is constant, where L_{net} denotes the number of links in the flow network and L_{full} denotes the number of links in the fully connected graph with the same nodes. The value for ρ is chosen such that the network has almost no isolated nodes and is sufficiently far from being fully connected.

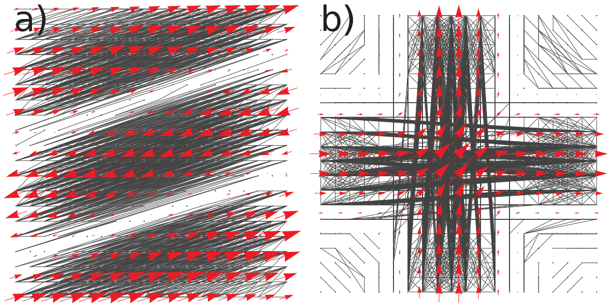


Figure 3.10: The correlation network (black) computed from the given velocity field (red arrows) for two flow fields for: a) Counter-currents, b) Crossing currents, for better visibility, a low link density of 2 percent was chosen. The networks display longer links in flow direction and a higher link density in regions with higher velocity.

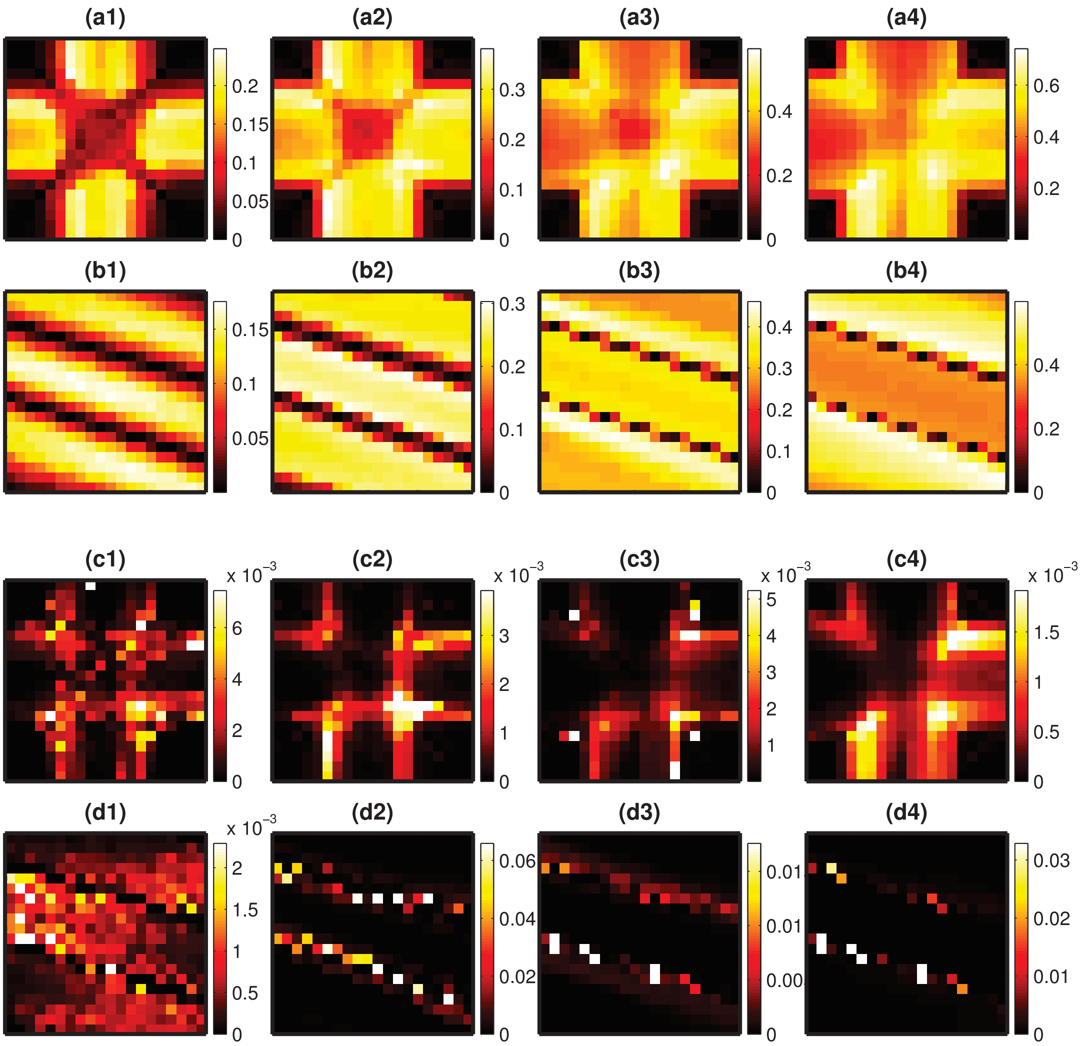


Figure 3.11: Degree of crossing flows (row a) and counter currents (row b) for a link density of 10 % (1), 20 % (2), 30 % (3), 40 % (4). Betweenness of crossing flows (row c) and counter currents (row d) for a link density of 10 % (1), 20 % (2), 30 % (3), 40 % (4). Figure from [28], supplementary material.

Fig. 3.11 shows degree (see Eq. 2.3) and betweenness (see Eq. 2.5) for link densities of 10 to 40 % to illustrate the robustness of the general features of degree and betweenness. This implies that there is no need to use a more sophisticated method of threshold selection, throughout this chapter the threshold was set at $\rho = 0.2$.

We observe that, while intensity and width of the patterns vary with the threshold, the degree is consistently higher in faster flowing regions and the betweenness marks a certain transition region for most values of the link density. In Fig. 3.11 (d1) the node betweenness is not highest in the transition region. This implies that the network of the counter currents is disconnected at a link density of 10 %, the rare and thereby high betweenness connections between the flows do not exist in this case.

The resulting networks and underlying flows are illustrated in Fig. 3.10. In areas of the flow with a higher velocity, the networks show a higher density and length of links than in slower regions. High absolute velocity coincides with high node degree (see Fig. 3.12 and Fig. 3.13, top row).

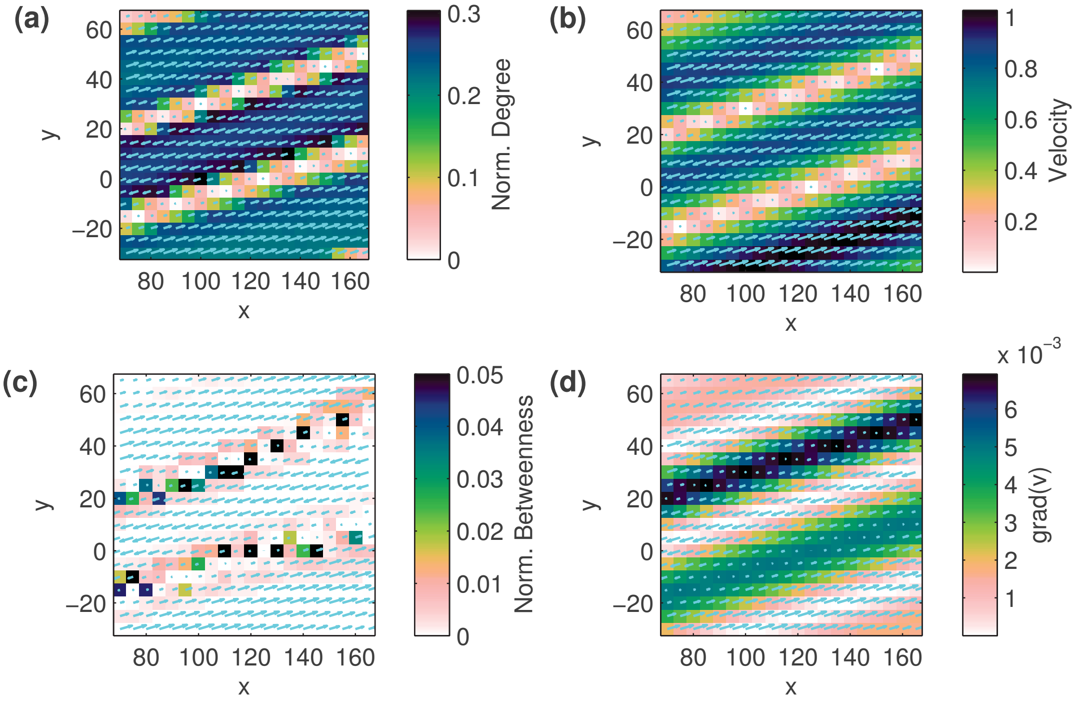


Figure 3.12: Flow field and network measures for the counter-currents in Fig.(3.10a): a) The normed degree, relates to b) the absolute value of the flow's local velocity; c) The maxima of the normed betweenness are co-located with d) the maxima of the absolute value of the gradient gradient of the absolute current velocity. See equations (2.3) and (2.5) for definitions of the network measures. Figure from [28].

For low velocities, degree and flow speed are approximately proportional, for higher speeds a saturation occurs due to the finite size of the grid. In Fig. 3.13a, the region, where both flows are highest (around (115,15)), however, has a lower degree, than the surrounding. The reason for this is most likely, that the approximate solution of the ADE works less well in this region than everywhere else.

High values for shortest path betweenness occur in the transition zones between opposing flow directions (Fig. 3.12, bottom row), or regions of distinctly different flow velocities (Fig. 3.13, bottom row). In both cases, the regions of highest betweenness outline the underlying velocity field.

Other network measures such as local clustering coefficient or local assortativity [1] yield structures similar to those of the node degree (results not shown).

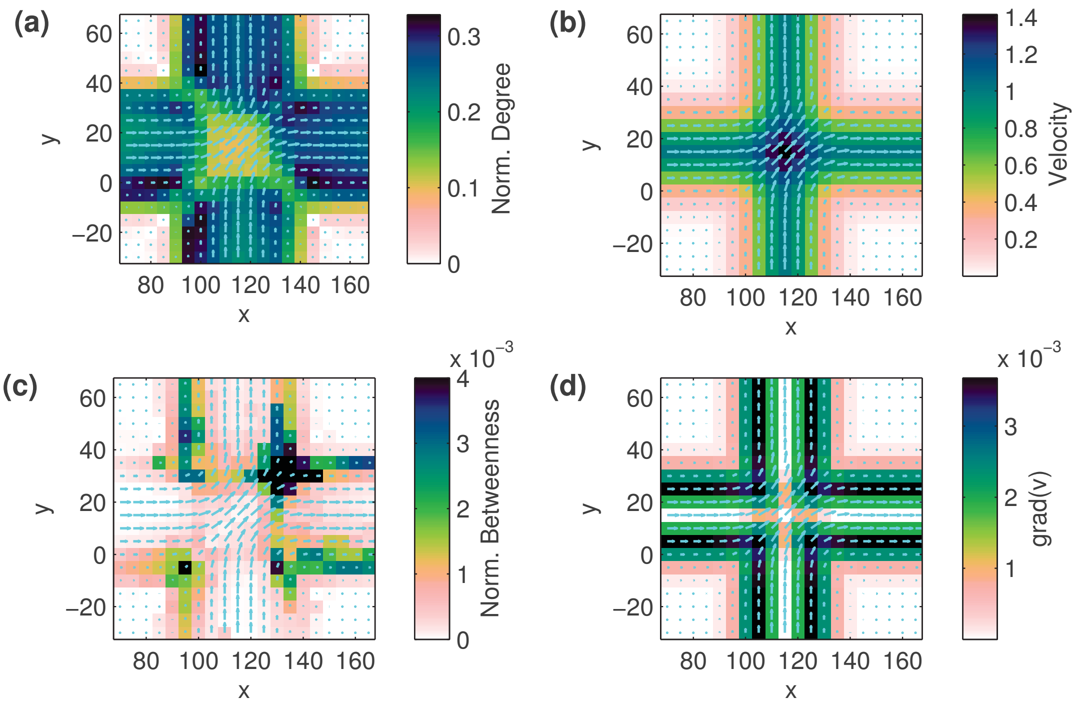


Figure 3.13: Flows and network measures for the crossing currents, see caption of Fig. 3.12. Figure from [28].

3.2.4 Application to data

In the model networks, it was found, that the degree is highest, where the absolute velocity is highest and high betweenness marks transition zones. In the actual climate system, many other influences act on the local temperature, yet these tendencies can be observed in correlation networks derived from sea surface temperatures (SST) in the tropical Pacific.

The daily anomaly SST data is based on the optimum interpolation data (OI.v2) as provided by NOAA/NCDC [40, 34] and the averaged monthly current's velocity data was provided by the OSCAR Project Office (Earth and Space Research, Seattle). The data stems from the region 120° − 160°W, 15°S−15°N and spans the time period August 1996 to August 1997. The chosen year is neither an El Niño nor a La Niña-year, and the results are largely robust against the choice of the particular year. The network is calculated by standard cross-correlation with a time lag of up to one day.

Flow velocity, gradient, and the obtained network measures are given in Fig. 3.14. To suppress turbulent effects, only the longitudinal component of the gradient was used. As in the paradigmatic flows investigated above, there is a reasonable agreement between the absolute values of the velocity field and the degree in the correlation network (Fig. 3.14 (a) and (b)). Again, the

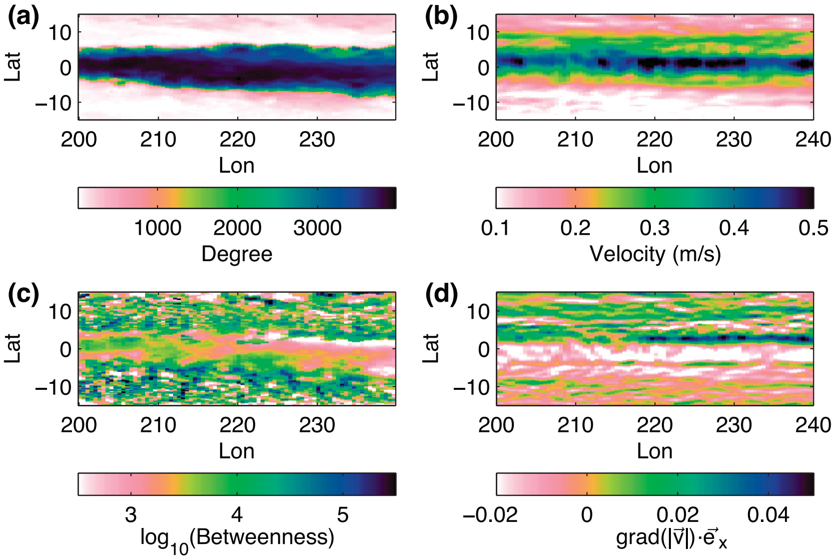


Figure 3.14: Network measures of the correlation network of the equatorial counter-currents from 1997 daily anomaly SST data in comparison with flow velocity and gradient. The region of highest degrees coincides with the region of highest flow velocity, while the regions of highest betweenness coincide with the highest velocity gradient. Figure from [28].

degree is maximal where the ocean current's velocity is, and the betweenness shows large values in regions with large values of the longitudinal velocity field's absolute gradient (Fig. 3.14 (c) and (d)), hereby confirming the results obtained for the paradigmatic flows.

3.2.5 Summary

The method can easily be generalized beyond 2D static flows and to flow systems outside of climate science, as temperature can be replaced by any quantity described by the heat equation such as density or chemical concentrations. In multivariate settings, reaction, advection and diffusion processes could be studied simultaneously. Given sufficient computing resources, non-stationary flows $\vec{v}(\vec{x}, t)$ could be treated similarly, using a time offset for integration range and peak appearance, as the ADE can still be solved analytically for time dependent velocity fields. This could give new insights in the dynamics of evolving flows, highly valuable not only in the analysis of changing climates.

The line-like structures in the betweenness fields of global climate networks [10] were previously attributed to "information flow" in underlying ocean currents. Here it was found that regions of high betweenness outline the flow rather than tracing it. The results therefore suggest some corrections concerning the former interpretation and suggest that a high betweenness occurs in transition zones between regions of different magnitude or direction of the underlying velocity. This qualitative observation can be seen when comparing the betweenness with the absolute

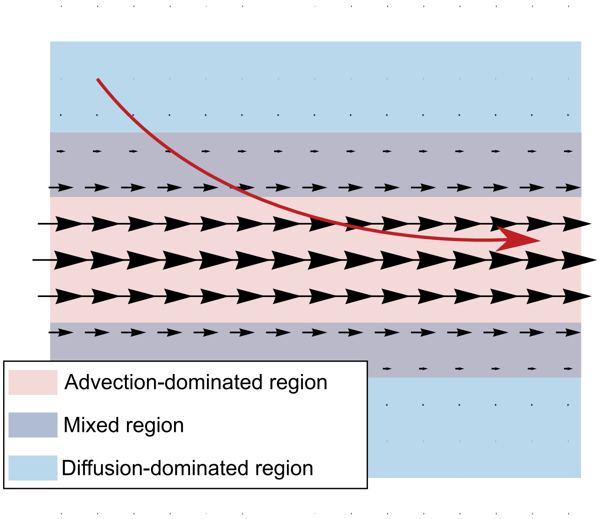


Figure 3.15: Schematic illustration of flow properties, that result in distinctive network properties: While advection dominates the transport of temperature fluctuations in regions of fast propagation, localized diffusion dominates in stagnant regions. Signals that leave the stagnant area by diffusion through the mixed region are subsequently transmitted along the flow. This leads to the asymmetry seen in the betweenness, where the betweenness values rise in flow direction. Figure from [28].

gradient. Physically, this could be due to the fact that advection dominates in fast flowing regions, which results in a higher parallel but lower perpendicular link density compared to the stagnant case. Additionally, there is a correlation between regions with a high node degree and high average current velocity. Considering the advective-diffusive nature of these surface currents, a physical explanation could be that a fast flow transports the signal farther.

Both, the degree and the betweenness increase marginally along the flow direction. This indicates that the signals from the slow flowing region first travel through diffusion, once they hit the fast region their main peak will travel downstream (the trajectory approximately follows the red arrow in Fig. 3.15). This leads to a network, in which points downstream in the fast flowing area have connections even to points in the slow region upstream from them, while points upstream in the fast region mainly get connected to other points in the fast region downstream from them, because diffusion plays a smaller role there. This leads to an increased degree and betweenness in the upstream slower regions close to the flow's centre.

In future research, such idealized case studies may be highly useful to study the influence of spatial embedding, and to test hypotheses concerning the dynamics of observed correlation networks. Given sufficiently low-gradient flow data, this method can be used to construct correlation networks from observed oceanic or atmospheric flows.

It was shown how correlation networks can be constructed directly from flow fields and given an example of how to use these networks to interpret network measures. We thereby provide a foundation for climate network analysis and bridge the gap between the dynamics of underlying flows and climate network interpretation.

3.3 Discrete method

A second, complementary, approach to network construction from flows is based on stochastic recursive equations. For this, the ADE is discretized and a stochastic term is added to the resulting linear recursive equation. The correlation matrix is averaged over all realizations of the noise, which can be interpreted as a weighted adjacency matrix and analysed using network measures.
Additionally to systems, that are fully determined by the velocity field, it is then also possible to study systems with external heating at a subset of the nodes, systems where some interactions are artificially lowered and systems in which two velocity fields interact by exchanging temperature. I propose a combination of network measures, which can identify the underlying system if only the network is known.

3.3.1 A discrete description

As in the previous section, the starting point is the ADE for an incompressible fluid, given in Eq. 3.1. Instead of solving it directly, it is discretized to a regular $N \times N$-lattice, using the central step method and $\Delta x = \Delta t = 1$. For the discretisation to appropriately represent the differential equation, the following requirements have to be met:

$$\begin{aligned} \kappa \Delta t &\ll \Delta x^2 \\ v_{max} \Delta t &\ll \Delta x. \end{aligned} \tag{3.27}$$

These requirements are met by choosing κ and v_{max} appropriately. The partial differential equation turns into a set of coupled equations for the temperature T_{ij} depending on the two components of the velocity v^i_{ij} and v^j_{ij}, that is $\vec{v}(\vec{x})$ evaluated at the grid points x_{ij} for $i, j = 1, ..., N$:

$$\begin{aligned} T_{ij}(t+1) = T_{ij}(t)(1-4\kappa) &+ (\kappa - v^i_{ij}/2)T_{i+1j}(t) \\ &+ (\kappa - v^j_{ij}/2)T_{ij+1}(t) + (\kappa + v^i_{ij}/2)T_{i-1j}(t) \\ &+ (\kappa + v^j_{ij}/2)T_{ij-1}(t). \end{aligned} \tag{3.28}$$

This notation can be simplified to a linear matrix equation:

$$\mathbf{T}(t+1) = \mathbf{A}\mathbf{T}(t). \tag{3.29}$$

$\mathbf{T}(t)$ is a list of the temperatures at time t in every location. The boundary conditions in the discrete case are implemented by defining the temperature outside the sampled region. Here the temperature is set to zero in all sites outside of the region of interest. The entries of the matrix $\mathbf{A}$ are the coefficients in Eq. 3.28. Fig. 3.16 shows a colour-coded example of $\mathbf{A}$. Time series can be derived from an initial temperature field $\mathbf{T}(0)$ using Eq. 3.29. It is deterministic and the initial conditions provide the only source. For a more realistic model, a noise term $\boldsymbol{\epsilon}(t)$

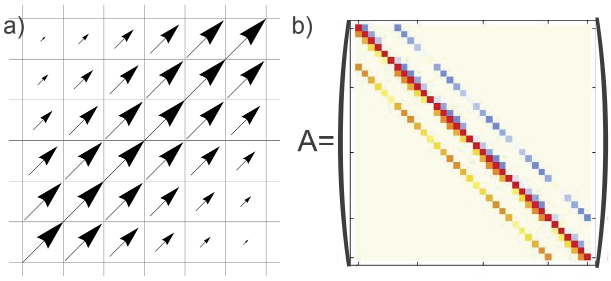


Figure 3.16: Example velocity field (a) with its transformation matrix **A** (b). Figure from [30].

is introduced in the recursive equation:

$$\mathbf{T}(t+1) = \mathbf{A}\mathbf{T}(t) + \boldsymbol{\epsilon}(t). \tag{3.30}$$

Where $\boldsymbol{\epsilon}(t)$ is a vector of uniformly random, statistically independent noise with zero mean variance one in each grid point at each time step:

$$\begin{aligned}\langle \epsilon(t)_x \rangle &= 0,\\ \left\langle \epsilon(t)_x^2 \right\rangle &= 1,\\ \langle \epsilon(t)_x \epsilon(t')_y \rangle &= \delta_{xy}\delta_{tt'}.\end{aligned} \tag{3.31}$$

Where the expectation value of all noise realizations is denoted by $\langle . \rangle$. It seems natural to choose the initial conditions: $\mathbf{T}(0) = \boldsymbol{\epsilon}(0)$. Finally, Eq. 3.30 can be integrated to yield:

$$\mathbf{T}(t) = \sum_{k=0}^{t-1} \mathbf{A}^{t-1-k}\boldsymbol{\epsilon}(k). \tag{3.32}$$

Due to the noise, each individual component of the temperature can not be predicted, but their correlations can be determined. Such correlation matrices can be analysed as networks and contain topological information about the system.

The covariance of two time series with zero mean is defined as the average over n time steps of coeval values of the time series, or in short their scalar product, divided by n. The covariance matrix of n time steps, $\mathbf{C}_n$, is therefore defined as the average over n time steps over tensor products of coeval temperature realizations.

$$\begin{aligned}\mathbf{C}_n &= \frac{1}{n}\left\langle \sum_{t=0}^{n} \mathbf{T}(t) \otimes \mathbf{T}(t) \right\rangle \qquad (3.33)\\ &= \frac{1}{n}\sum_{t=0}^{n}\sum_{k=0}^{t-1}\sum_{k'=1}^{t-1} \left\langle \mathbf{A}^{t-1-k}\boldsymbol{\epsilon}(k) \otimes \mathbf{A}^{t-1-k'}\boldsymbol{\epsilon}(k') \right\rangle .\end{aligned}$$

Writing Eq. 3.33 out in components and applying Eq. 3.31 leads to:

$$\begin{aligned}
&\sum_{k=0}^{t-1}\sum_{k'=1}^{t-1}\left\langle A^{t-1-k}_{ij,i''j''}\epsilon(k)_{i''j''}A^{t-1-k'}_{i'j',i'''j'''}\epsilon(k')_{i''j''}\right\rangle \\
&=\sum_{k=0}^{t-1}\sum_{k'=1}^{t-1} A^{t-1-k}_{ij,i''j''}A^{t-1-k'}_{i'j',i'''j'''}\delta_{i''j'',i'''j'''}\delta_{kk'} \\
&=\sum_{k=0}^{t-1}(AA^T)^{t-1-k}_{ij,i'j'}, \qquad (3.34)
\end{aligned}$$

Where $\mathbf{A}^T$ denotes the transposed of matrix $\mathbf{A}$.
In an idealized static case we can additionally take the number of time-steps n to infinity. So we can evaluate Eq. 3.33 to:

$$\begin{aligned}
\mathbf{C}_\infty &= \lim_{n\to\infty} n^{-1}\sum_{t=0}^{n}\sum_{k=0}^{t-1}(\mathbf{AA}^T)^{t-1-k} \\
&= \lim_{n\to\infty} n^{-1}(nI - n\mathbf{AA}^T + \mathbf{AA}^T((\mathbf{AA}^T)^n - I))(I-\mathbf{AA}^T)^{-2} \\
&= (I-\mathbf{AA}^T)^{-1} \qquad (3.35)
\end{aligned}$$

This expression only depends on the transformation matrix $\mathbf{A}$. It describes how the temperature evolution in a point is related to the temperature evolution in other points. Note, however, that the largest eigenvalue of matrix $\mathbf{AA}^T$ is required to be smaller than one to ensure stability and allow the use of the geometric series for the computation.

The time-dependent case

The discrete approach can easily be generalized to incorporate time-dependence. The time dependence of the system is encoded in a time dependent transformation matrix $\mathbf{A}(t)$, which transforms Eq. 3.30 into:

$$\mathbf{T}(t+1) = \mathbf{A}(t)\mathbf{T}(t) + \boldsymbol{\epsilon}(t). \qquad (3.36)$$

When subsequent application of the transformation matrix lead to its exponentiation before, now it results in the multiplication of the matrix at different times:

$$\mathbf{T}(t) = \sum_{k=0}^{t-1}\prod_{j=0}^{t-1-k}\mathbf{A}(t_j)\boldsymbol{\epsilon}(k). \qquad (3.37)$$

And thereby the correlation function:

$$\mathbf{C}_n = \sum_{t=0}^{n}\sum_{k=0}^{t-1}\prod_{j=0}^{t-1-k}\mathbf{A}(t_j)\mathbf{A}(t_j)^T \qquad (3.38)$$

The resulting correlation matrix can be thresholded and analysed as a weighted network.
The description of the system with a transformation matrix makes it possible to include other influences beyond advection and diffusion, which is a step towards a more realistic model. The goal is to construct networks from different transformation matrices and find network measures, that can distinguish between them. In subsequent sections three different cases will be introduced.

Secluded regions

In the first variation, a part of the region is disconnected from the rest. This can be useful for modelling regions with large differences in altitude, or separation by a geographical barrier, that does not influence the atmospheric velocity field.
For this, the system is divided into regions and all components of $\mathbf{A}$, that connect a node from one region to a node from another region are multiplied by small factor.
The new transformation matrix is used to compute a correlation matrix according to Eq. 3.35 and analyse it as a network.

External heating

Another form of a common driver is a region, that is affected differently by the decay of the signal of the last step. It could be a region with a higher heat capacity, such that the ground reheats the air. This is realized by adding a heat term to the diagonal elements of $\mathbf{A}$, that are part of the affected region:

$$\mathbf{A} \to \mathbf{A} + \mathbf{H}, \tag{3.39}$$

where $\mathbf{H}$ is a diagonal matrix, with $h_{i,i} > 0$ in regions, that are heated externally and $h_{i,j} = 0$ otherwise.
The network is constructed from the modified $\mathbf{A}$ using Eq. 3.35.

Interacting flows

To model temperature exchange between two separate velocity fields, i.e. ocean flows and atmospheric flows, the following coupled recursive equations are introduced:

$$\begin{aligned} \mathbf{T_a}(t+1) &= \mathbf{A}_a \mathbf{T_a}(t) + \mathbf{B_a}\, \mathbf{T_b}(\mathbf{t}) + \boldsymbol{\epsilon}_a(t) \\ \mathbf{T_b}(t+1) &= \mathbf{A}_b \mathbf{T_b}(t) + \mathbf{B_b}\, \mathbf{T_a}(\mathbf{t}) + \boldsymbol{\epsilon}_b(t) \end{aligned} \tag{3.40}$$

Here T_a is the primary temperature field, the network is constructed from. T_b is the secondary temperature field, that interacts with T_a and thereby influences network a. A_a and A_b are the corresponding transformation matrices, depending on the velocity field in the respective medium and B_a and B_b are inter-medium diffusion matrices. The effect of T_a on T_b is ignored, because the impact of a fluctuation to be transported from system a to b and back again, are small. The resulting equations are:

$$\begin{aligned}\mathbf{T_a}(t+1) &= \mathbf{A}_a\mathbf{T_a}(t) + \mathbf{B_a}\ \mathbf{T_b}(\mathbf{t}) + \boldsymbol{\epsilon}_a(t)\\ \mathbf{T_b}(t+1) &= \mathbf{A}_b\mathbf{T_b}(t) + \boldsymbol{\epsilon}_b(t)\end{aligned} \tag{3.41}$$

Solving Eq. 3.41, the following expression for the temperature evolution is found:

$$\mathbf{T_a}(t) = \mathbf{A}_a^t\mathbf{T_a}(0) + \sum_{k=0}^{t-1}\mathbf{A}_a^{t-1-k}(\boldsymbol{\epsilon}_a(k) + \mathbf{B_a}(\mathbf{A}_a^k\mathbf{T_b}(0) + \sum_{k'=0}^{k-1}\mathbf{A}_b^{k-1-k'}\boldsymbol{\epsilon}_b(k'))). \tag{3.42}$$

The correlation matrix for flow A is:

$$\mathbf{C}_n = \sum_{t=0}^{n}\sum_{k=0}^{t-1}((\mathbf{A}_a\mathbf{A}_a^T)^{t-k} + \mathbf{B_a}\mathbf{B_a}^T\sum_{k'=0}^{k-1}(\mathbf{A}_b\mathbf{A}_b^T)^{k-k'}) \tag{3.43}$$

In principle the inter-medium interaction matrix $\mathbf{B_a}$ could have any shape but for simplicity I assume that the layers only exchange temperature with the same node in the other medium and do so with a fixed rate c, i.e. $\mathbf{B_a} = c\mathbf{I}$.

Constructing the network

In principle, the resulting correlation matrix can directly be interpreted as a weighted network, where the link weight between x_1 and x_2 is equal to their correlation $C_n(x_1, x_2)$. However, as most correlations are very small, the introduction of a threshold, below which the correlations are set to zero, does not significantly change the result, while largely simplifying the computations. Unweighted networks can be constructed by setting all non-zero weights to one. This however results in a larger dependence of the network on the threshold.

In the following I show the results for weighted and unweighted networks. All networks were computed with a threshold keeping the strongest ten percent of links, but the resulting patterns in the network measures have been found to be robust with respect to changes in the threshold.

3.3.2 Distinguishing different cases by network measures

Networks are constructed from the four cases using Eq. 3.35. While the other influences (secluded region, external heating and interaction with a secondary flow field) change, the velocity field is always kept the same for comparability. It is defined by the stream function $\Phi(x, y)$ as:

$$\vec{v}(\vec{x}) = u\begin{pmatrix} -\frac{\partial\Phi}{\partial y} \\ \frac{\partial\Phi}{\partial x}\end{pmatrix}. \tag{3.44}$$

where u is the velocity parameter and $\Phi(x, y)$ is defined as:

$$\Phi(x, y) = 1 - \tan((y - \sin(2(x))) \cos(\arctan(\cos(2(x))))) \tag{3.45}$$

This kind of flow is called *meandering flow*, together with the other influences it is depicted in Fig. 3.17. It constitutes a simple but non-trivial physical example, and can be found approximately in ocean currents such as the gulf stream.

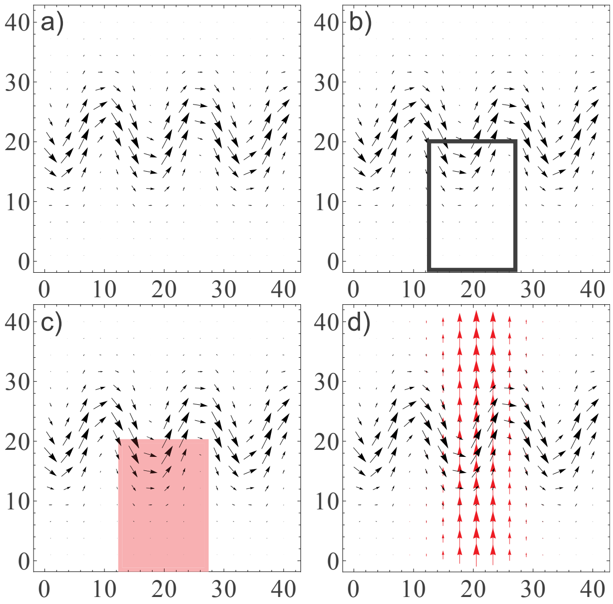


Figure 3.17: Schematic visualization of the underlying system with a) plain flow, b) secluded region, c) external heating, d) interacting flows. Figure from [30].

Here, the parameters for the velocity $u = 0.2$ and a diffusion constant of $\kappa = 0.01$ were used. For this flow field, networks are computed with all of the variations mentioned above. The effects of secluded regions, external heating and interacting flows are discussed and compared to networks generated from the velocity field only. The grid is regular and has 40 × 40 nodes.

For reasons of computational complexity, the sums are restricted to $n = 20$. In the following sections the resulting network measures are presented in their weighted form, applied to the meandering flow: plain, with a secluded region, with external heating and with interaction with a secondary flow.

To analyse the networks and distinguish between the different cases, I suggest the use of three network measures: Node degree or node strength, community detection and edge angle anisotropy [23].

Degree or node strength

Fig. 3.18 shows the degree and node strengths (Subsec. 2.3.1) of the networks of the four cases. In agreement with previous studies [28], in the case of the meandering flow without other influences (Fig. 3.18 a and e), reaches the node strength in regions with a high absolute velocity. The flow transports random temperature fluctuations, thus strengthening the correlation between two nodes, that are connected by a fast flow. While this can be seen both in the unweighted (top row) and weighted (bottom row) network measure, the weighted version (e) provides a far clearer picture, with less border effects and almost no dependence on the threshold.

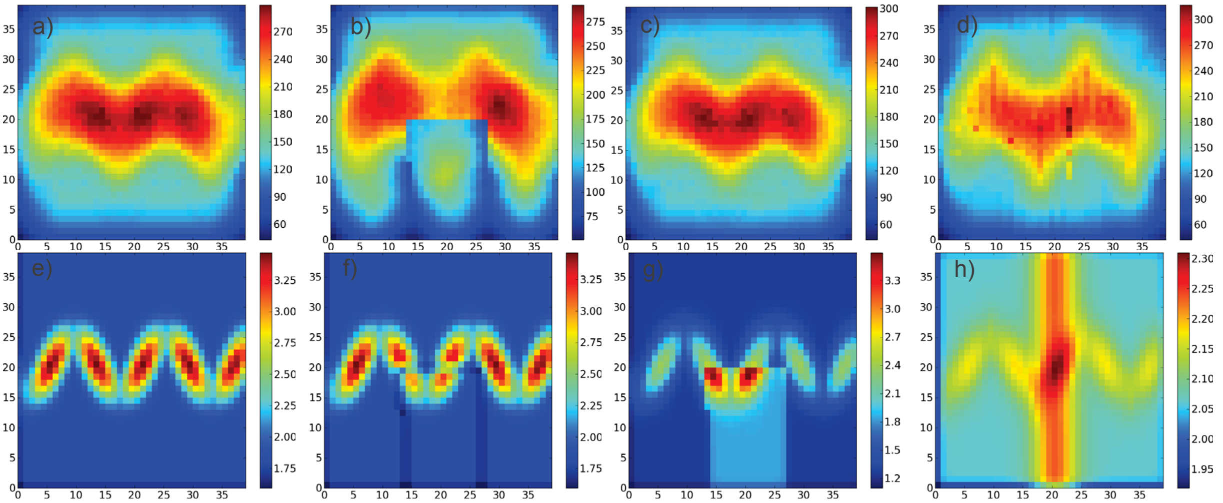


Figure 3.18: Node degree of the unweighted network resulting from a) plain flow, b) secluded region, c) external heating, d) interacting flows.
Node strength of the weighted network resulting from e) plain flow, f) secluded region, g) external heating, h) interacting flows. Figure from [30] and supplementary material.

In the case of a secluded region (see Fig. 3.18 b and f), this effect can still be observed. In the area, where links were cut, the degree or node strength is lowered. The restriction of the region also lowered the node strengths in the fast flowing regions around the cutting line. So the node strength can be lowered by restricting the size of the influencing region. In the unweighted case (b) this effect is far more prominent.

If there is an externally heated region however, this also leads to an increase in the node strength, as can be seen in Fig. 3.18 g, in the unweighted case (c), however, this effect can not be seen at this threshold. Another mechanism for an increased node strength is the indirect transport of fluctuations via an interaction with a secondary velocity field in the interacting case in Fig. 3.18 h. Again, the unweighted network measure merely shows a mild distortion of the pure flow, instead of resolving the secondary flow.

As we found multiple possible causes for high node strengths, degree and node strength alone are insufficient to distinguish uniquely between the different cases. Furthermore, this shows that, the unweighted degree does not resolve many of the phenomena, going beyond the plain flow.

Edge angle anisotropy

The anisotropy (Subsec. 2.3.4) is a measure of how aligned the links of a node are. As such, it should allow to distinguish between a mechanism, that creates links with a preferred direction (i.e. advection), and a mechanism, that creates links in all directions with equal probability (i.e external heating).

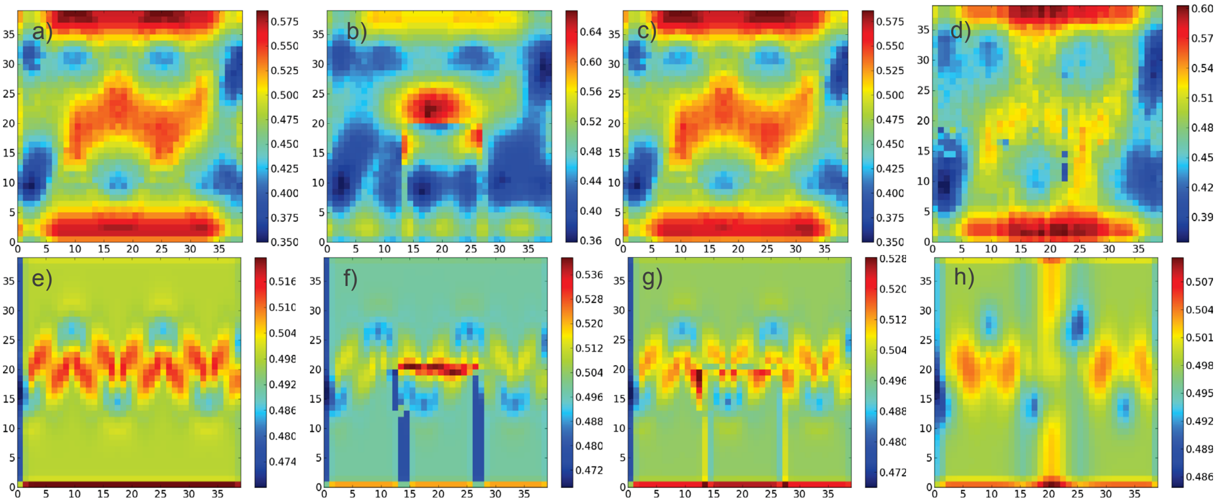


Figure 3.19: Local anisotropy of the network resulting from a) plain flow, b) secluded region, c) external heating, d) interacting flows.
Weighted local anisotropy of the weighted network resulting from e) plain flow, f) secluded region, g) external heating, h) interacting flows. Figure from [30] and supplementary material.

The anisotropy for all cases is displayed in Fig. 3.19. We observe, that the weighted anisotropy in the network of the meandering flow (Fig. 3.19 e) is highest, at the sides of the fastest straight flowing areas and lowest, at the turning points. In the middle of the straight flow links go to both sides, whereas at the sides of it, links only go to the middle, leading to smaller differences in angles. The turning points have links to two of the straight stretches, leading to large angles between them. At first sight, the unweighted anisotropy (Fig. 3.19 a) appears to trace the flow, just like the degree, but upon closer examination, it is actually out of phase with the degree. This is due to the fact, that the anisotropy is lowest at the turning points, that lie within the flow. The pattern in the unweighted anisotropy can again be seen as a blurred version of the weighted pattern.

This pattern persists in the other cases, but is distorted. In the case of the secluded region (Fig. 3.19 f) the anisotropy is lowered where the cut goes against the flow direction and increased where the cut goes with the general flow direction. In the first case the links that defined a direction were cut, where in the latter case all those links are cut, that destroyed the alignment. In the unweighted version in (Fig. 3.19 b), none of these structures are clearly visible.

In the case of external heating (Fig. 3.19 g) we find the same pattern as in the meandering flow, with some distortion at the borders of the heated region. Together with the node strength, this can help to distinguish between a heated and a fast flowing region. If the node strength is high and the anisotropy follows its shape, being large in straight and low in bent regions, the network structure can best be explained by a flow. If the node strength is high and the anisotropy is largely unaffected, this hints towards external heating in the high node strength region.

Interaction with a straight flow perpendicular to the average direction of the meandering flow as depicted in Fig. 3.19 h, leads to a blurring of the patterns of the meandering flow and adds high anisotropy along the secondary flow. In the overlapping region however, the anisotropy is lowered, which distinguishes separate layers of interacting flows from a simply additive flow pattern. In the unweighted version, Fig. 3.19 c and d are distorted patterns from the plain flow and show no additional information.

Community detection

The communities (Subsec. 2.3.3) are shown in Fig. 3.20. It appears that networks of the meandering flow do not form independent clusters (Fig. 3.20a), adding external heating (Fig. 3.20c) or interaction with perpendicular flow (Fig. 3.20d) does not change this. Separating a region from the rest like in Fig. 3.20b leads to this region forming a separate cluster.

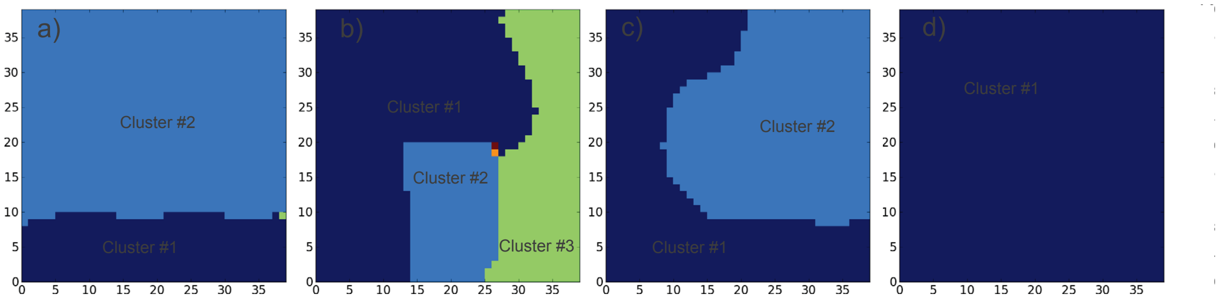


Figure 3.20: Communities of the network resulting from a) plain flow, b) secluded region, c) external heating, d) interacting flows. Figure from [30].

Therefore a region, that forms its own community can be identified as a secluded region if the structure also appears in the node strength (as a line of lowered node strength) and the anisotropy (depending on the relative direction of the edge of the separate cluster to the flow direction).

3.3.3 Time dependent network construction

The time dependent network construction is done using a shifting meandering flow. In 10 time steps the flow is shifted over one whole period. Examples of the time slices are depicted in Fig. 3.21. The networks are constructed according to Eq. 3.38 and node strength and weighted betweenness are computed for it and shown in Fig. 3.22d and 3.22e. Looking only at the network of the time dependent velocity field gives a similar result as one would expect from the average velocity field (Fig. 3.22f), in which the amplitude of the meandering structure is reduced significantly compared to the time slices (Fig. 3.21). The node strength is highest where the average velocity field is fastest and the betweenness reaches its maxima in the transition region between

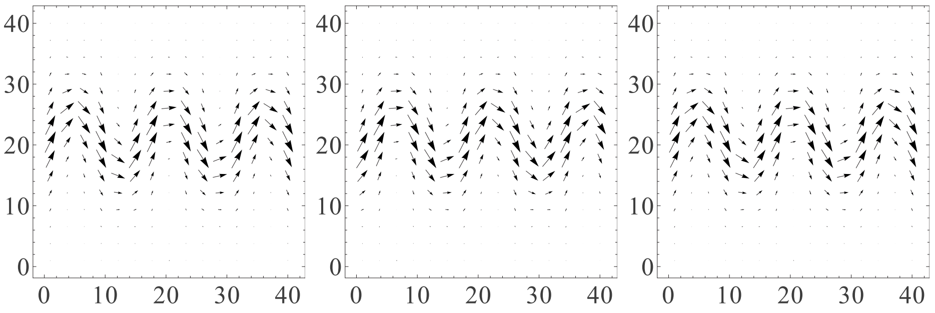


Figure 3.21: Velocity field for a) The first, b) The fifth and c) The 9th time step.

fast and slow areas of the average flow, showing an outlining pattern.
As the time dependence is not visible in the network measures, it might be necessary to analyse single time slice networks (Fig. 3.22a-c). If a trend is visible between the time slice networks and between the single slice networks and the time dependent one, this indicates a time dependent flow. In networks from climate data, this is hard to accomplish. The closest to realizing this is a sliding window analysis.

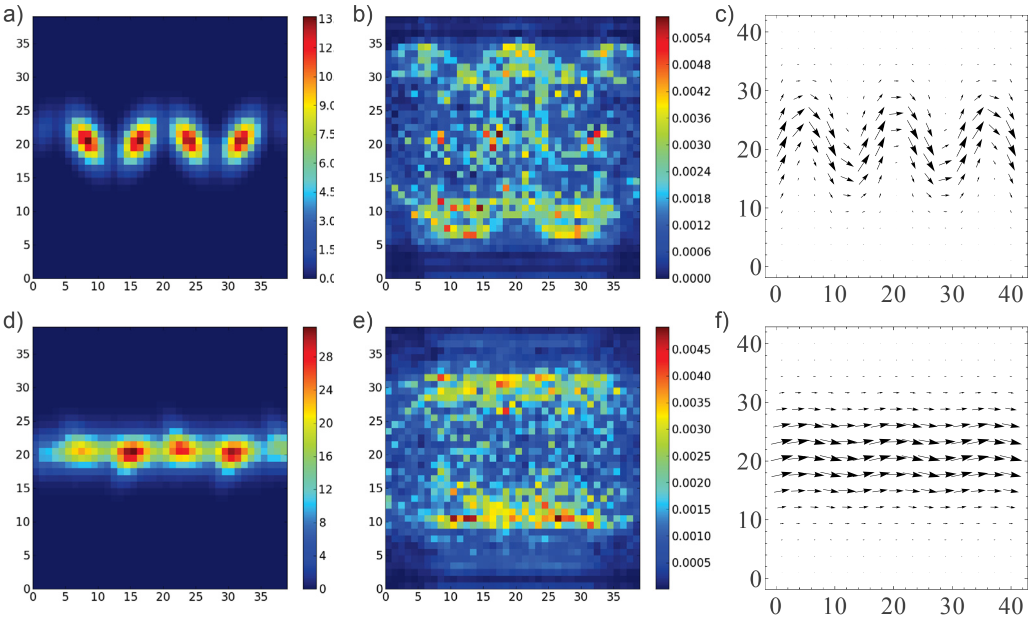


Figure 3.22: a) Node strength of time slice 0, b) Weighted betweenness of time slice 0, c) Velocity field of time slice 0, d) Node strength of the whole period, e) Weighted betweenness of the whole period, f) Average velocity field.

3.3.4 Summary

A new method to construct networks from flows was introduced, that also allows to modify the set-up from purely relying on the velocity field, to also including secluded regions and external heating, as well as interacting flows.

Example networks were constructed for all these cases and network measures were found to distinguish between different effects.

A purely advection driven system is characterized by high node strengths in combination with high anisotropy, where the flow velocity is high and straight and high node strength and low anisotropy, where the region of high velocity has a corner.

An externally heated region shows a high node strength but no changes in the anisotropy, apart from border effects.

A secluded region can be characterized by a separate community. Additionally the typical distortions in the anisotropy with the characteristically high values for borders along the flow direction and low values for borders against flow direction can be used to verify that result.

The node strength in interacting flows adds up while the anisotropy can cancel out if the flows go in different directions.

Furthermore, it was found, that the unweighted measures suffer far more from problems, such as threshold selection and border effects and provide far less clear pictures. I therefore recommend the use of weighted networks where possible, as it is not that much more computationally expensive and provides clearer and more robust statements.

As a next step, this method could be generalized to more realistic models and used to establish a set of tools for the reconstruction of a physical system from a correlation network. Using techniques presented in [21], it would be interesting to further investigate the inverse problem of reconstructing an underlying physical system from correlation data.

The approach presented in this work could also be applied to other systems, that can be described by an evolution equation.

4 Spatial issues of flow networks

If you point your questions
The fog will surely chew you up
But if you want the answers
You better get ready for the fire
System of a down, *Suggestions*

In this chapter the methods, developed in the previous section, are used to analyse and interpret some aspects of climate networks. First a simplified model of the monsoon region is introduced, that is based on the continuous method. Then the influence of the spatial distributions of the nodes on the network topology is analysed, using common network measures. And finally an evolution detection measure is applied to track the changes in the network due to variation of a flow parameter.
Sect. 4.1 is based on [39]. Sect. 4.2 follows my publication [29] closely and Sect. 4.3 is relies on [48].

4.1 Stream Transported Auto Regressive Temperature (START) model

In palaeoclimate data analysis, past climate conditions are reconstructed from proxy records. These can be be found i.e. in ice cores, tree ring data or stalagmites. One of the challenges of the use of such data is their spatial heterogeneity. Here, a semi-empirical model is introduced, that was developed to illuminate the impact of this inhomogeneity on the Asian monsoon system, which is subject to a time dependent forcing.
Observations and analysis of palaeoclimate data [38] have shown that the region in question is influenced by the Indian summer monsoon (ISM) and the East Asian summer monsoon (EASM). The idea behind this model is similar to the methods presented in Chap. 3: The system is modelled as a velocity field, in which temperature fluctuations are transported via advection and diffusion. In contrast to the previous methods however, palaeoclimate data is approximated at the level of the time series and not only the network.
The flow system consists of three narrow flows with Gaussian envelopes (see Fig. 4.1). Two of the flows come from the ocean in the south and go northwards at 70° E (representing the ISM branch) and 115° E (representing the EASM branch). The other one models the longitudinal influence of the ISM, and goes from west to east at 30° N. In contrast to the methods in Chap. 3, however, there are now four main sources of variability: One at the centre of each of the flows and a local noise term.
Over the centuries the respective strengths of the different influences (X, Y and Z) have varied

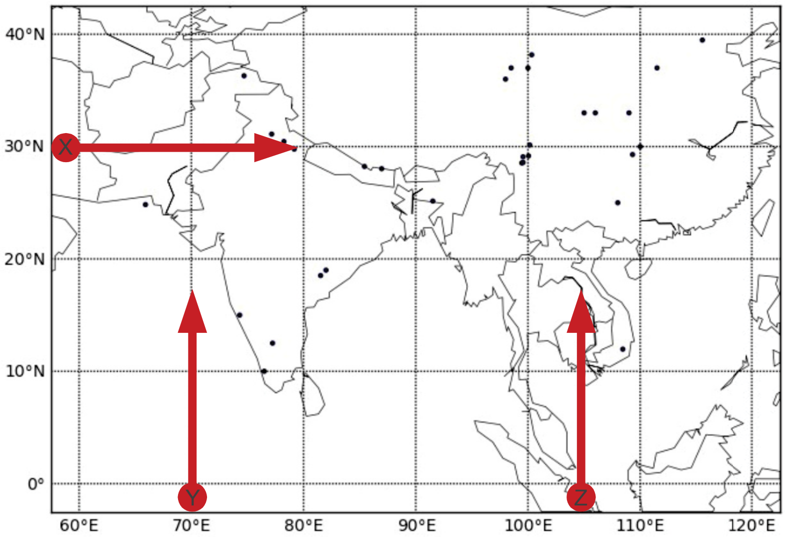


Figure 4.1: The monsoon system is modelled as a system of three flows, each with a fluctuation source at its origin. The black dots mark the locations of the measurements.

significantly. In the model, this is described by the forcing, which controls the relative strength of the three branches.

At each point in the region, the time series is computed as the sum of the random fluctuations generated by each of the four sources, scaled with a factor, that quantifies the transport from that source to the point:

$$R_i = f_X(i,F)R_X + f_Y(i,F)R_Y + f_Z(i,F)R_Z + R_{noise}. \tag{4.1}$$

Where R_X is the signal at point X, R_Y is the signal of the vertical Indian source, R_Z is the signal of the vertical East Asian source, R_Z is the signal of the Chinese source, R_{noise} is the noise and $f_X(i,F)$, $f_Y(i,F)$ and $f_Z(i,F)$ are the scale factors at point i and forcing F. The forcing controls the respective strengths of the different velocity fields.

To compute these factors the ADE (3.1) was solved approximately, starting from the source's position. $f_X(i,F)$ is defined as the maximum height of the temperature front from the vertical Indian source, when it reaches point i.

The initial condition used for the solution of the ADE is a Gaussian shaped temperature front of unit height (in longitudinal- or lateral-direction, depending on the source). It is of the form:

$$e^{-\frac{(x-x_0)^2}{s}}. \tag{4.2}$$

Where s is the width and x_0 the position of the temperature fronts. One front starts at the western border of the region and the other starts at the southern border. Both fronts are transported by the three flows and deformed in the process. The interaction with the flow is of the form:

$$T(x,t) = \sqrt{\frac{s}{s+4\chi t}}\, e^{-\frac{(x-x_0-vt)^2}{s+4\chi t}}. \tag{4.3}$$

Snapshots of the temperature profile in the flow are given in Fig. 4.2. Eq. 4.3 describes how the temperature front starting from x_0 develops over time. At $t = 0$, it is near zero in most positions, rises to a maximum and decreases again.

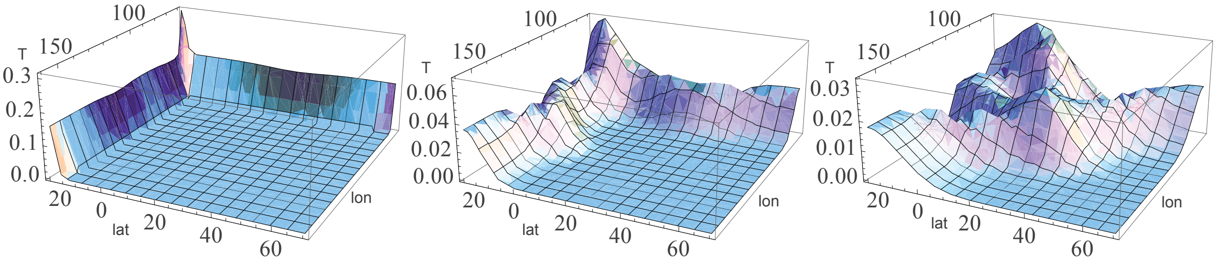


Figure 4.2: The temperature profile for the computation of the scale factors, at t=1, t=20 and t=80 in the chosen region.

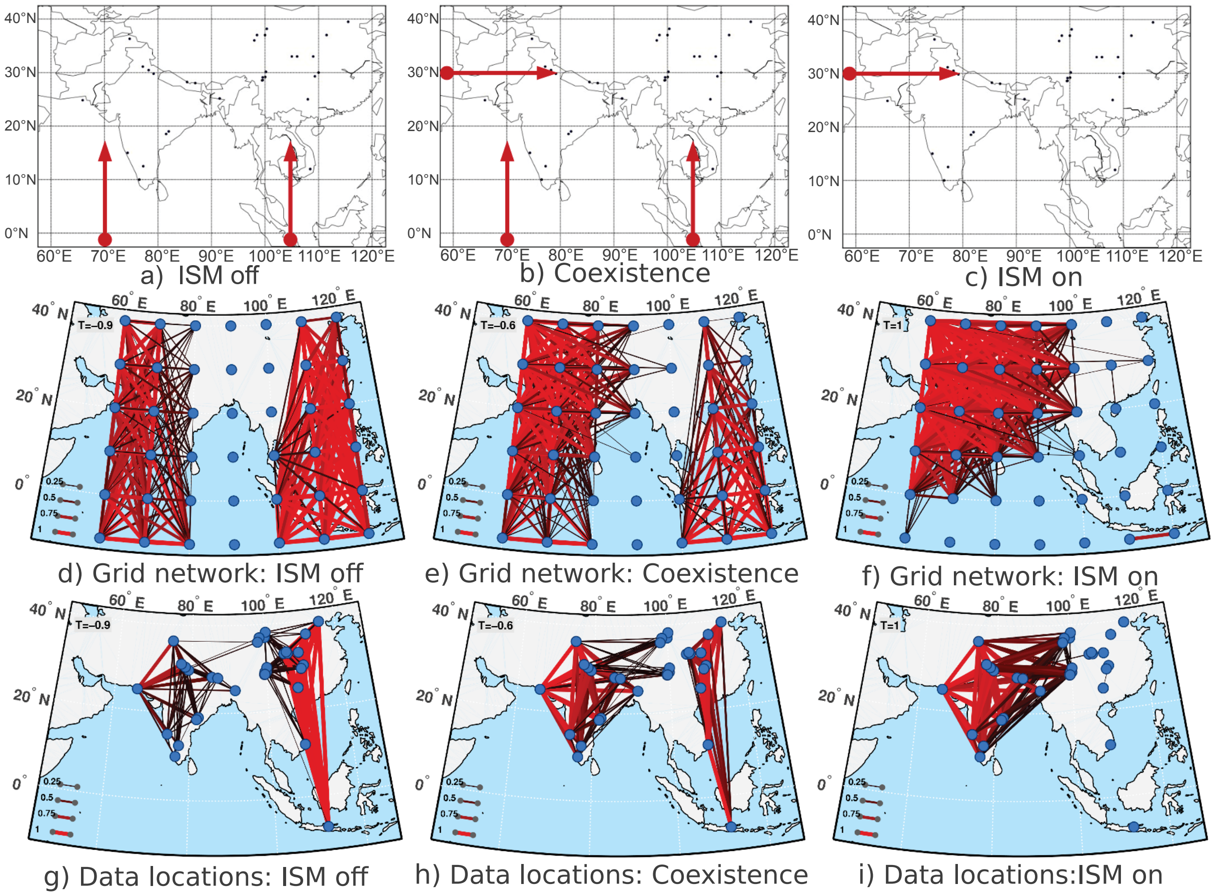
The scale factors $f_X(i,F)$, $f_Y(i,F)$ and $f_Z(i,F)$ are defined as the as the maximum temperature in location i, caused by the front from X, Y and Z.

Artificial time series R_X, R_Y, R_Z and R_{noise} are generated at each point i and Eq. 4.1 is used to combine those time series into a local time series. Using common similarity measures, such as Pearson correlation and mutual information, these local time series are used to construct a network.

This method was applied to the monsoon region. The forcing was varied from $F = -1$, resulting in only the south-north streams being present ("ISM off", only Y and Z in Fig. 4.1), over $F = 0$, leading to an equal coexistence of longitudinal and latitudinal branches ("Coexistence", X, Y and Z coexist), to $F = 1$, featuring only the longitudinal branch ("ISM on", only X). The results for this are shown in Fig. 4.3.

Each of the origins of the red arrows in Fig. 4.3a)-c) is a source of variability. For each source a random time series is generated and at each point these three sources, together with a local source are combined according to Eq. 4.7, to generate a local time series. Several different similarity measures (cross correlation, mutual information and event synchronization) are used to determine the amount of similarity between time series in a robust way. The threshold is set at the 20 % strongest links. This is done for nodes arranged on a grid (Fig. 4.3d)-f)) and for nodes at their original measuring locations (Fig. 4.3g)-i)).

As one might expect, in the "ISM off"-case, depicted in a), both in the grid and in the data locations (Fig. 4.3d) and 4.3g)), the network is almost disconnected into two large clusters, one for the latitudinal ISM branch and one for the EASM branch. The addition of the longitudinal ISM branch extends the ISM cluster towards the east, but the clusters remain separate. If the latitudinal ISM and EASM branch are switched off, the EASM cluster disappears and the ISM cluster is extended even further east, connecting the formerly distinct ISM and EASM regions. These features are robust under changes in the spatial node distribution and can be seen both, in the gridded version and with the nodes at the data locations.



a) ISM off

b) Coexistence

c) ISM on

d) Grid network: ISM off

e) Grid network: Coexistence

f) Grid network: ISM on

g) Data locations: ISM off

h) Data locations: Coexistence

i) Data locations:ISM on

Figure 4.3: The three branches at 70° E, 115° E and 30° N. model different influences in the monsoon system. The forcing term varies their respective strength. a)-c) Illustrate the effect of the forcing, d)-f) Show the resulting networks, when the nodes are arranged on a grid, g)-i) Show the resulting networks, when the node locations are the data locations. Figure from [39].

It is found, that the results from the START model are consistent with the results from palaeoclimate records found in [38]. There networks were constructed and analysed for the medieval warm period (700 to 1100 years BP), the little ice age (100 to 400 years BP) and the recent warm period (-30 to 100 years BP). The results suggested, that, during the little ice age, ISM and EASM region were less connected than during the warm periods. In START this can be modelled as different values for the forcing.
START has thereby been shown to reproduce features present in the palaeoclimate data, thereby strengthening the hypothesis, that the cooler climate during the little ice age may have lead to a weaker influence of the ISM on the EASM region. Furthermore, the model suggests, that the main features of the network topology are robust against the spatial inhomogeneity.

4.2 The influence of node distribution

The construction of climate networks from time series data relies on locally measured or extrapolated observations. As reanalysis has its own pitfalls [49, 27] or regularly sampled data may not be available at all, it can be desirable to construct the networks directly from spatially irregular measurements. Since the method, introduced in Sect. 3.2 allows free choice of node locations, it provides a good testing ground for the influence of the spatial distribution of nodes. In this section we want to study how network topology, velocity field and node distribution are intertwined.
We construct networks for different node distributions and analyse their degree, betweenness and link length distribution to compare the influence of spatial sampling on the topology of the resulting network.
Flow networks for a diagonal flow pattern, are constructed for the node distributions shown in row A of Fig. 4.4, on a regular 20×20 grid, a jittered 20×20 grid and for two versions of Gaussian clustering. The jittered sampling pattern was generated by uniformly drawing a point from each grid cell. For the clustered sampling, two Gaussian distributions of 100 nodes each, were generated around two central points on opposite sides of the flow (Ac) and along the flow (Ad).
Degree: The degree (shown in row B in Fig. 4.4) takes its maximum in the middle stripe, where the flow velocity is highest. It is not very sensitive to the spatial sampling and all four plots clearly show the expected pattern. Clustered sampling can lead to a skewed shape of the stripe (Bc), or underestimate its width if the outer region is poorly sampled (Bd).
Betweenness: The betweenness is shown in row C in Fig. 4.4. Grid and grid plus jitter show similar patterns: the betweenness is highest in the transition zone between the fast and slow flowing areas. This structure is barely visible in the clustered sampling plots. In the clusters on opposite sides, two stripes are visible, but the lower one of them is located in the centre of the flow, rather than at its side, in the middle between the two clusters.
This indicates that rather than showing a transition zone of the flow, it emphasizes a region, that is poorly sampled but central in the flow. In the plot of the two clusters in the flow, the outer regions are poorly sampled, therefore the transition is not visible.
Voronoi tessellation analysis: A Voronoi tessellation assigns a cell to each node, such that every point in the cell is closer to that node than to any other. If the nodes are uniformly

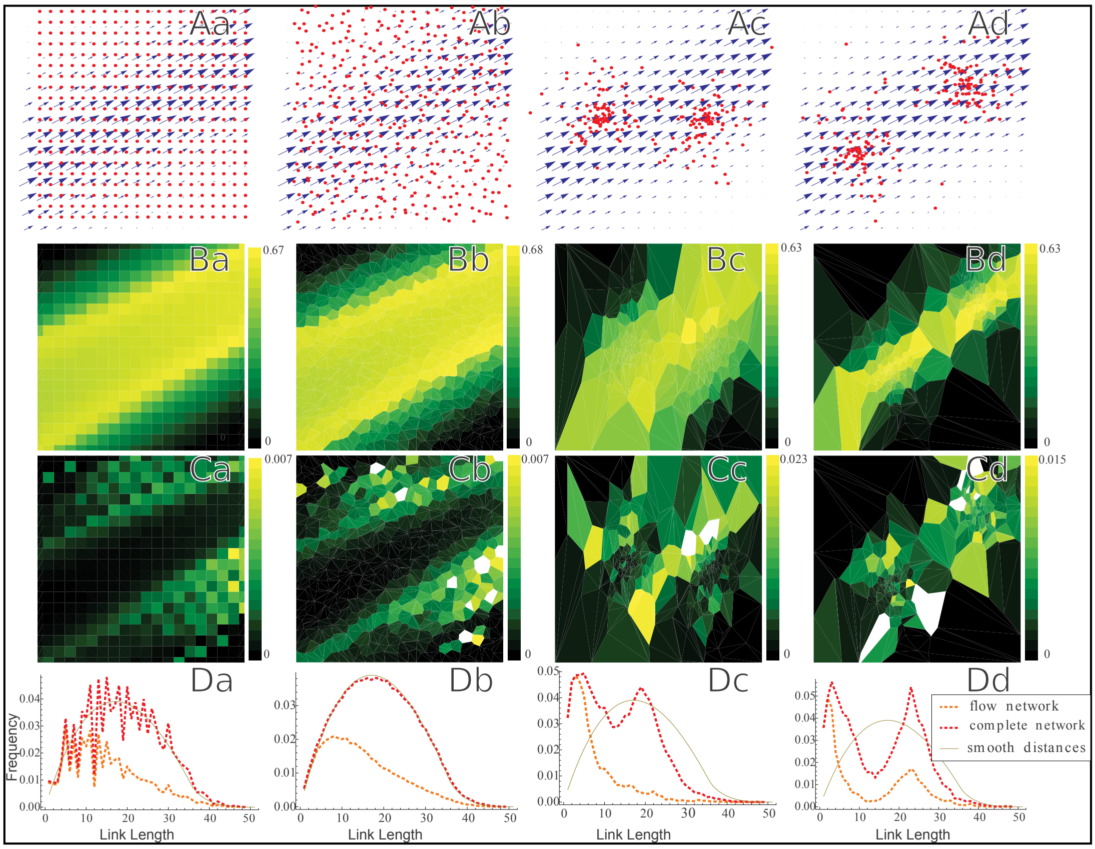


Figure 4.4: A diagonal flow is sampled with Aa) a grid, Ab) a jittered grid, Ac) two clusters on opposite sides of the flow, Ad) two clusters inside the flow. The networks are constructed with a link density of 40 percent. Row B) shows the degree, row C) the betweenness and row D) the link length distribution. Figure from [29].

distributed in space the resulting cell sizes will have a clear peak around A_{sam}/N_{nod}, where A_{sam} is the size of the total sampled area and N_{nod} is the number of nodes. The differences in the sampling can be quantified using the area size distributions of their Voronoi tessellations [1] as shown in Fig. 4.5.

In the grid all tiles are exactly the same size, so the histogram would be one very sharp peak. In the case of grid with jitter the peak is broadened. The two clustered distributions peak around much smaller values.

Link length distribution: The link length distribution (see Sect. 2.3.5), shown in row D in Fig. 4.4 is spiky for the grid and considerably smoother for the jittered node distribution. In the clustered sampling, the link length distributions show two peaks. The beige line shows the distance distribution (see Sect. 2.3.5) of pairs of points in a continuous square for comparison.

The regular arrangement of points in the grid leads to an over-representation of some specific link lengths i.e. multiples of the grid constant, while other node-node distances are excluded by the node distribution. The distance distribution of pairs of points in a continuous square is

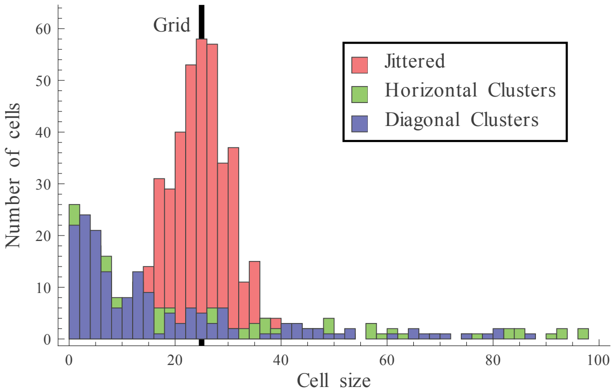


Figure 4.5: Tile size distribution for jittered and clustered sampling. The jittered sampling is peaked around the grid cell size, clustered sampling results in many small and few very large tiles. Note that grid and jittered grid have twice as many nodes as the clustered sampling. Figure from [29].

the continuous analogue to the link length distribution of the fully connected graph of nodes in the discretized network description. Since climate is a continuous phenomenon, an appropriate discretisation should well approximate the distance distribution. This is true for the link length distribution of the fully connected nodes of the jittered sampling. For the regular grid, it is more spiky but varies around the same curve, while the clustered node sampling leads to a biased link length distribution with one maximum for links within each cluster and another maximum of links connecting the clusters.

Compared to the link length distributions of the fully connected networks, the actual flow networks only express a subset of links, that is dominated by short ranged links. In the clustered case c) this eliminates the inter-cluster peak, because the clusters are not connected by a flow. In the clustered case d) however, this second peak persists, suggesting that links of length 30 are preferred, even though this effect is entirely a sampling effect rather than a flow property.

4.2.1 Spatial sampling in networks from the START-model

Using the START (Stream Transported Auto Regressive Temperature) model [39] described in Sect. 4.1 gridded, jittered and clustered networks are constructed to compare their topology.

It is used to simulate time series at each node and construct the network by computing the correlations for each pair. The 20% strongest correlations are used to set the links in the network. The START-model assumes three sources, R_X, R_Y and R_Z of random noise, that is transported by three independent flows. The ADE is used to compute variance factors $f_X(\mathbf{p}, F)$, that approximate the influence of source X on position $\mathbf{p}$ at forcing F. At each point, the signal is computed as the sum of the contributions from the three sources, scaled with the corresponding variance factor. A local noise contribution is also added:

$$R_i = f_X(i, F)R_X + f_Y(i, F)R_Y + f_Z(i, F)R_Z + R_{noise}. \tag{4.4}$$

Here, two of the factors were set to zero to model a diagonal flow like the one used above. Despite also being constructed according to advection and diffusion, START is significantly different from the other flow networks. Firstly, instead of averaging over all possible initial peak positions, there is a particular source location. This results in a decay of signal strength with distance from the source. Fig. 4.7 shows the node degrees for the resulting networks, which drop off with distance to the source, unlike the flow networks, where the degree only depended on the velocity. While the jittered node distribution shows the same structure as the grid, using clustered node locations again leads to a distortion of the pattern.

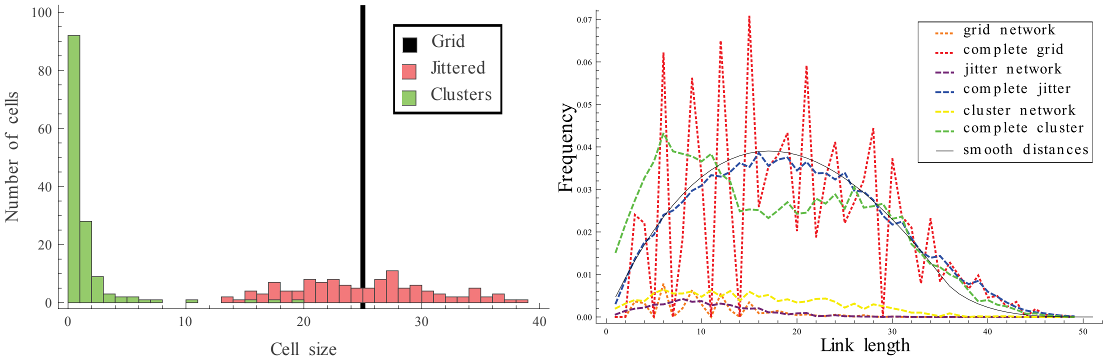


Figure 4.6: **Right:** Tile size distribution for the three node distributions in START. **Left:** Link length distribution for the three node distributions in START. Figure from [29].

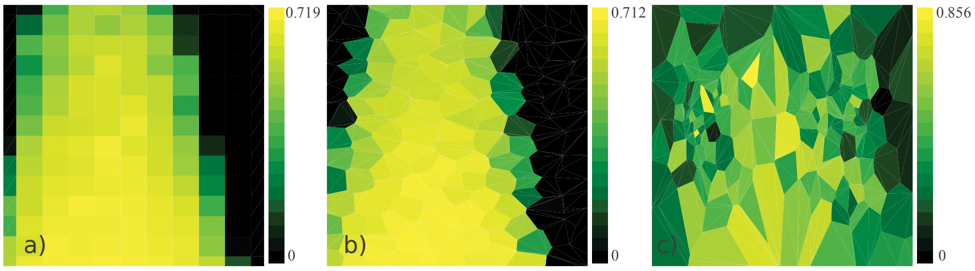


Figure 4.7: Degree of the network of the flow generated with START for grid (a), grid with jitter (b) and two clusters (c). Figure from [29].

Due to the addition of local random fluctuations, that are themselves not transported by advection and diffusion, the degree in the stagnant and outer regions goes to zero.

The Link length distribution is similar to the link length distributions for the flow networks, see Fig. 4.5.

The betweenness (results not shown) shows the same tongue like structure as the degree. This can be explained with the random fluctuations reducing the dynamic areas to only one, which is the region inside the flow, while everything else is essentially unconnected.

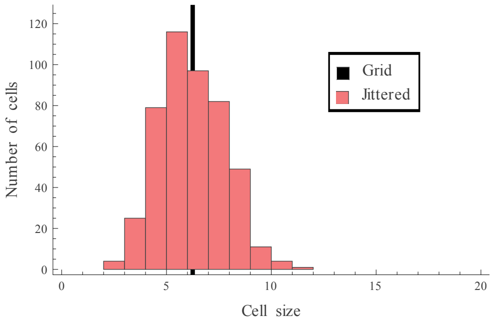


Figure 4.8: Tile size distribution for the jittered tile size distribution of the reanalysis data together with the grid line. Figure from [29].

4.2.2 Spatial sampling effects in data from the Asian monsoon domain

The data analysis was done on NCEP/NCAR [22, 32] reanalysis daily surface temperature anomalies for the summer monsoon months (June to September) of the years 1971-2010 with a resolution of 2.5×2.5 degrees, which results in a grid constant $a = 280$ km. The network was constructed from the anomaly time series using Pearson correlation. The significance threshold was set at the 5 percent strongest correlations. Degree and betweenness are analysed as shown in Fig. 4.9. The geographical coordinates of each node are marked by a point at the centre of each cell. To investigate sampling effects, a jitter was added in longitude and latitude at each grid point with a uniformly random number between $\pm a/2$. The tile size distributions are shown in Fig. 4.8. The new time series are constructed as Gaussian weighted averages of all of the original reanalysis time series on the grid. The weights depend on the euclidean distance between the grid points and the new location.

Fig. 4.9 shows a) the annual average of the surface wind speeds, where the arrow displays the direction and the colour shows the absolute value, that is computed by averaging over longitudinal and latitudinal components separately and taking the absolute value of the resulting vectors. Subfigure b) shows the link length distribution. The degree is presented for grid and jittered sampling in the second row (c,d). The bottom row shows the betweenness for grid and jittered sampling (e,f).

The node degree (Fig. 4.9c and d) is highest over the ocean and lowest over the Indian subcontinent. This coincides with the absolute wind velocities, which are highest over the ocean, Bangladesh and the Himalayas. This supports the theory, that connections in the climate network are often caused by an atmospheric or ocean flow between the node locations. On a daily scale, temperature fluctuations are transported along the flow, leading to a high correlation of nodes connected by a high velocity. Despite the high wind speeds over the Himalayas, however, the degrees there are not elevated. This may be because the plateau is relatively secluded from the rest of the network due to its high altitude. Consequently, even if the link density inside the secluded area would be very high, the degree would be low, as connections to other regions are less likely.

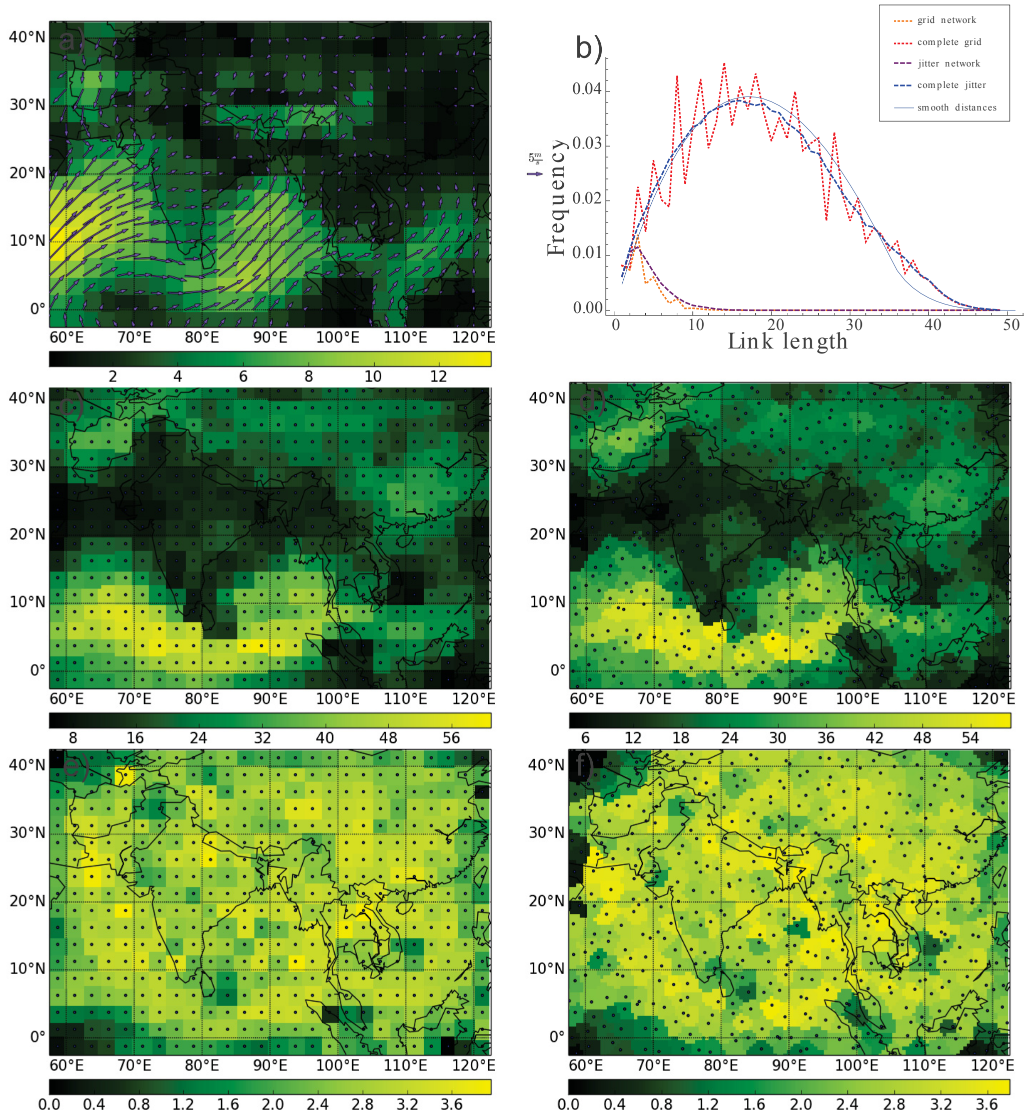


Figure 4.9: The correlation network of temperature data from NCEP/NCAR on a grid and a jittered grid. a) absolute mean wind velocity, b) link length distributions, c) degree of the grid network, d) degree of the jittered network, e) betweenness of the grid network, f) betweenness of the jittered network. Figure from [29].

We find that the effect of spatial sampling on the degree patterns is small and the overall patterns are not affected by the node distribution. This coincides well with our findings from the theoretical approaches in the previous subsections, as the jittered node distribution is still homogeneous as shown in the Voronoi tile size distribution in Figure 4.8.
The shortest path betweenness (e,f) is more sensitive to the spatial distribution of the nodes and the spatial resolution. Therefore, few clear patterns can be seen, most regions have a

betweenness between 2.5 and 3.5. The jittered network shows higher boundary effects in the west. There are three regions with a particularly low betweenness in both the gridded and the jittered version of the network at (80-85 E, 5-10 N), (100 E, 0-5 N) and (110 E, 10-15 N), these coincide with small regions with a low degree.

4.2.3 Summary

Different node distributions were examined in Flow networks, the START model and surface temperature correlation networks. The degree is robust, while the betweenness reacts more sensitively to the spatial sampling. All examples given, are related to climate, yet the results can be generalized to all networks coming from discretisation of a continuous system.

For a smooth link length distribution the node to node separation has to cover all possible distances. A grid does not fulfil that as all distances are of the form: $d(x_1, x_2) = a\sqrt{n^2 + m^2}$, where a is the grid constant and n and m are integer numbers. Distances, that can not be represented in this way, are not present in a grid.

While the effects of the area sizes themselves are discussed in detail in Heitzig et al. [18] and can be removed using their proposed consistently weighted network measures, this study shows that the problem goes further, when the underlying physical system is taken into account. While the node size distribution of the two clustered sampling versions in the flow network approach, is very similar, their relative position to the underlying flow is not, resulting in very different distorting effects on the network structure.

In future work it might be interesting to compare different jitter realizations quantitatively, for example using the common component evolution function (CCEF) as introduced in [48].

So in summary, it was found that, as long as the node distribution in space is sufficiently homogeneous, the exact spatial sampling chosen has little impact on the topology of the network in all cases studied. Only in cases of significantly inhomogeneous sampling, distortion and misleading structures arose. It is therefore important to discuss the sampling and its impact when analysing spatially sampled data. As a simple test of the spatial sampling the Voronoi size distribution can be employed. In a homogeneous sampling this will have a clear peak around A_{sam}/N_{nod}, where A_{sam} is the size of the total sampled area and N_{nod} is the number of nodes. If the peak is shifted or very spread out, that suggests poor spatial sampling.

4.3 Common component analysis

Since the methods described in Chap. 3 provide a means of constructing networks with controlled parameters, it is now possible to evaluate the evolution of flow networks with the parameters of the underlying flow.

In climate, there are processes like the daily weather, from which climate networks can be inferred by evaluating correlations. But there are also long-term trends and fluctuations on much larger time scales, that could be interesting phenomena. If the correlations are evaluated on intermediate time scales, by either segmenting the original data into smaller portions or using a sliding window approach, where the segments can overlap, multiple networks can be generated from the data. It is then interesting to compare these networks and examine their differences systematically. One method for this was presented in [2], where the correlation between weighted adjacency matrices was computed.

Here we want to generalize a related but different measure, which was introduced in [48]: The common component function (CCF). The common component of T unweighted networks $[N_1, ..., N_T]$ of the same number of nodes, is the network of those nodes and only the links, that occur in all T networks. It is defined as the sum of the product of each entry of all T adjacency matrices $[A^1_{l,m}, ..., A^T_{l,m}]$:

$$CCF(N_1, ..., N_T) = \sum_{l,m} A^1_{l,m} A^2_{l,m} ... A^T_{l,m} \tag{4.5}$$

For $T = 2$ this is identical to the measure in [2], and for unweighted matrices it is also identical to the definition in [48], but immediately generalizes to weighted networks.

The common component evolution function (CCEF) tracks the overlap of two networks with a certain distance in the ordering. The CCEF is defined for a linearly ordered set of T $n \times n$ networks $[N_1, ..., N_T]$, that can be described by their adjacency matrices $[A^1_{l,m}, ..., A^T_{l,m}]$. Two networks N_i and N_j in this set have a distance or *lag* δ in this order if $|i - j| = \delta$. Then, the ratio of the number of common links of two networks with the distance δ and the average number of links in a network, is the $CCEF(\delta)$.

$$CCEF(\delta) = \frac{\frac{1}{T-\delta} \sum_{i=1}^{T-\delta} \sum_{l,m} A^i_{l,m} A^{i+\delta}_{l,m}}{\frac{1}{T} \sum_{i=1}^{T} \sum_{l,m} (A^i_{l,m})^2} \tag{4.6}$$

By definition, $CCEF(0) = 1$ for all sets of networks, and $CCEF(\delta)$ can only take values between zero and one. For an artificially linearly ordered set of Erdős-Réyni-Graphs (ER-Graphs) with a fixed number of n nodes and a fixed, constant connection probability p, The CCEF can be computed analytically.

As the networks are completely random, the size of the common component should not depend

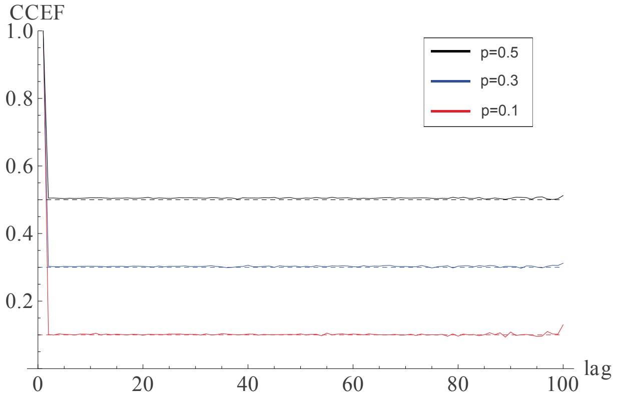


Figure 4.10: CCEF of a set of 100 linearly ordered ER-networks of 100 nodes for linking probabilities $p = 0.1$, $p = 0.3$ and $p = 0.5$.

on the lag, but be a function that is 1 for $\delta = 0$ and constant otherwise. The total number of possible links in such a network is $n(n-1)/2$, at a linking probability p, the expectation value for the number of realized links is $pn(n-1)/2$. For the numerator in Eq. 4.6 we need to find the expectation value of the number of links present in two networks. As the probability of each of the edges in one network to also appear in the other network is p, the total number of common links is $p^2 n(n-1)/2$, which with the normalization leads to:

$$CCEF(\delta) = \begin{cases} 1 & : \delta = 0 \\ p & : \delta > 0 \end{cases} \tag{4.7}$$

In Fig. 4.10, the theoretical prediction is compared to a set of 100 randomly generated networks.

The evolution of a flow network can be analysed for the gradual change of any flow parameter using the CCEF. The set of flows, whose networks will be considered here is a set of ten diagonal flows of increasing widths given by the following velocity function:

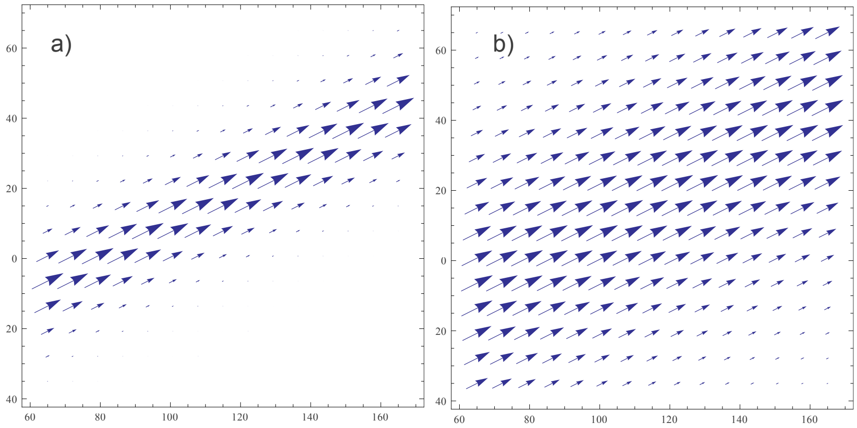


Figure 4.11: The diagonal flow changes from width parameter 200 (a) to width parameter 2000 (b) in ten steps of 200. Figure from [48].

$$\vec{v}(x,y) = \begin{pmatrix} e^{\frac{-(y-0.5x)^2}{c}} \\ 0.5e^{\frac{-(y-0.5x)^2}{c}} \end{pmatrix}. \tag{4.8}$$

The parameter c for the flow-width varies from 200 to 2000 in 10 steps, thus gradually changing the flow-network. Node position and node number are kept constant. For each value of c, correlation matrices $C_1, ..., C_{10}$ are obtained according to the method presented in Sect. 3.2. The resulting matrices are thresholded to obtain unweighted networks. Using Eq. 4.6, the CCEF is computed and shown in Fig. 4.12.

As expected, the resulting CCEF, shown in Fig. 4.12 decreases monotonously with the index lag, indicating that there is a systematic evolution happening in the system.

Finally, the CCEF was also computed for climate networks of 1971-2010 in [48], using daily NCEP/NCAR reanalysis temperature anomaly data [34, 22] for the Asian monsoon domain.

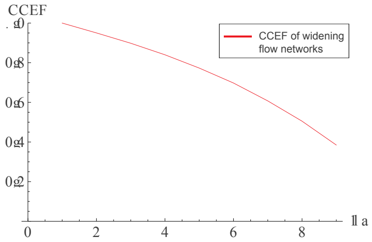


Figure 4.12: CCEF of ten networks of a widening diagonal flow for a threshold of 0.2. Here the CCEF depends on the lag δ.

The area between 2.5 °S to 42.5 °N and 57.5 °E to 122.5 °E was sampled with a spatial resolution of 2.5 °, resulting in time series for 468 nodes. Pearson correlation was used in windows of full years from January till December and the link density was chosen to be 5%.
At first sight, the resulting CCEF shows the same overall structure as the CCEF of random networks: A constant level of overlap for all lags except $\delta = 0$. But the levels (see Fig. 4.13) are significantly higher than those of random networks of the same link density. This suggests a persistent core network with fluctuations on top of it. The slight decline between 25 and 30 years time lag is statistically not significant, but suggests that it might be interesting to look at longer time spans.

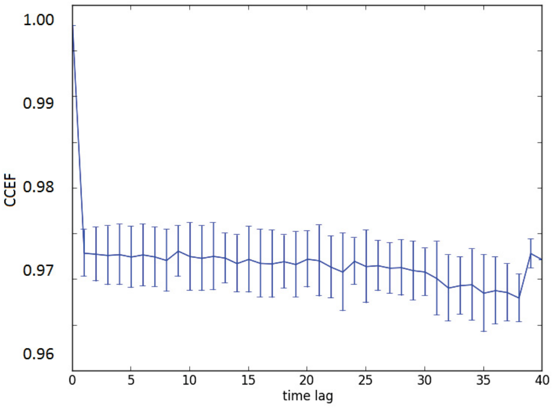


Figure 4.13: CCEF Here the CCEF depends on the lag δ. Figure from [48].

The CCEF has been introduced as a tool for the analysis of linearly ordered networks, or the evolution of networks with a parameter. It was applied to three distinct cases, namely random networks, showing an overlap of p, independently of the lag, flow networks of a widening flow, displaying a monotonous decline with the parameter distance and climate networks from data, showing a persistent core with additional fluctuations. The method thereby allows the distinction between deterministic and random behaviour and can be used to find periodicities. The definition of the CCF in Eq. 4.5 furthermore allows the computation of the persistent core by evaluating the total overlap between all networks.

5 Summary and outlook

I seldom end up where I wanted to go, but almost always end up where I need to be.

Douglas Adams

In the previous chapters, two complementary methods were developed for the construction of networks from a flow system. This helped to get an insight into the physical meaning of climate networks and their network measures. These methods were then applied to climate and palaeoclimate questions, as well as to address issues of the spatial distribution of nodes. Here I will summarize and discuss and the results from the previous chapters and draw my conclusions from them, as well as point out open ends.

5.1 Network construction from flow systems

In Chap. 3 of this thesis, two complementary methods were established to construct networks from the ADE, analogously to network construction from time series in climate networks.

The first method allowed the construction of networks of stationary flow systems with a small gradient using a combination of approximations, analytical calculations and, in the more complex cases, numerical evaluation. Thereby enabling the comparison of climate networks to the dynamics of advective-diffusive systems, for the first time giving insights into the relationship between dynamics and topology of such systems.

The second method used a discretisation of the ADE with added white noise. The resulting stochastic linear recursive equation was solved and the correlation matrix could be computed analytically, depending only on the variance of the white noise and the transformation matrix. This method was found to be significantly faster and therefore applicable to larger networks. The derivation is valid for any linear evolution system with noise and thus, by altering the transformation matrix, one can describe effects, that are not related to the ADE, use temporally changing transformation matrices or analyse coupled flow systems. It was shown how network measures can be employed to distinguish between different cases.

So far these methods have only been applied in the context of climate networks, but they could be applied to any linear evolution equation without much adjustment. The discrete method may also lead the way to a generalization to linear stochastic differential evolution equations.

5.2 Analysis of spatial issues with networks

In Chap. 4, the methods from the previous chapter were applied to related problems.

Spatio-Temporally AutocoRrelated Time series (START) is a semi-empirical model of the Asian summer monsoon region, designed to track the effects of transitions in the velocity field. Unlike in the methods presented in Chap. 3, random time series are generated at the source locations and combined according the the ADE. The resulting networks can explain structures seen in data networks.

To test the influence of spatial sampling, the continuous network construction method was used to construct networks of the same flow for different node distributions. It was found that inhomogeneous spatial sampling can lead to distorted or misleading patterns in the network measures. The Voronoi tile size distribution was found to be a measure for the sampling quality.

Common component analysis was introduced as a method of studying the evolution of networks. The method was applied to random matrices, matrices generated from flow networks and climate networks of the Asian summer monsoon.

Despite the simplicity of the model, the above examples showed, that it can shed light on a number of complex problems, related to the role of spatial proximity in climate networks.

5.3 Conclusion

Coming back to the questions from the introduction, the following answers were found:

- *Can we find a simple analytical model system, to produce networks, that are structurally and conceptually similar to climate networks?*
  We have implemented such models in Chap. 3 and Chap. 4.1. To evaluate the structural similarity of the created networks with climate networks, network measures, such as degree, betweenness and anisotropy were used.

- *What can be said about such networks and their relation to differential equations in general?*
  While we are not yet in a position to answer this question in general, multiple ways of constructing networks from flows were introduced and while the inverse problem is hard, some publications like [21] suggest that the converse might be possible in special cases at least.

- *Can we find a relationship between topological features and properties of the underlying system?*
  Topological features can be quantified using network measures. Those can be related to local parameters of the underlying flow system such as absolute velocity, velocity gradient and external heating. Examples for this were given in Chap. 3.

- *How do construction parameters like spatial node distribution or influences other than the flow field itself impact network topology?*
  The spatial distribution of nodes in the flow plays a large role in the network topology, as analysed in Chap. 4.2, any interpretation of the topology should consider this. To quantify the impact of flow and construction parameters, the CCEF was introduced in Chap. 4.3 to track topological changes in networks.

5.4 Outlook

Answering questions in science is like beheading a Hydra, whenever a question is answered, two new ones open.

- The two methods from Chap. 3 are a first step on the way to understanding the correspondence between stochastic partial differential equations and complex networks. Using methods from the analysis of stochastic differential equations it might be possible to obtain a more general expression for the correlation function between solutions at different locations, which would allow the definition of a network. From there, the relationship between dynamical properties and topology could be studied in even more generality, revealing for example possible connections between non-linearity and non-locality. Such a correspondence would be widely applicable in physics and beyond.
- How can the theory of time dependent correlations be used to improve the analysis of climate networks?
- Can a consistent weighting similar to that presented in [18] be developed based on the additional information provided by the velocity field, to obtain better results from inhomogeneously sampled flow networks?
- The CCEF method can be generalized in various interesting ways, including monitoring the evolution of node degrees, eigenvector centrality and other network measures and using a known periodicity to determine the persistent core of all networks at a distance of the period.
- So far only the evolution of scalar fields has been used for the network construction. It would be interesting to extend the concepts to vector fields or a multivariate description of the system.

Bibliography

[1] Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1-3):1–101, February 2011. ISSN 03701573. doi: 10.1016/j.physrep.2010.11.002.

[2] Y. Berezin, A. Gozolchiani, O. Guez, and S. Havlin. Stability of climate networks with time. *Scientific Reports*, 2:666, January 2012. ISSN 2045-2322. doi: 10.1038/srep00666.

[3] S Boccaletti, V Latora, Y Moreno, M Chavez, and D Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, February 2006. ISSN 03701573. doi: 10.1016/j.physrep.2005.10.009.

[4] Niklas Boers, Bodo Bookhagen, Norbert Marwan, Jürgen Kurths, and José Marengo. Complex networks identify spatial patterns of extreme rainfall events of the South American Monsoon System. *Geophysical Research Letters*, 40(16):4386–4392, August 2013. ISSN 00948276. doi: 10.1002/grl.50681.

[5] R. Butler. *Saddlepoint Approximations with Applications*. Cambridge University Press, 2007. ISBN 1139466518.

[6] C. Chatfield. *The analysis of time series: an introduction*. CRC Press, Florida, USA, 6 edition, 2003. ISBN 978-0203491683.

[7] B. N. Clark, C. J. Colbourn, and D. S. Johnson. Unit disk graphs. *Annals of Discrete Mathematics*, 86(1-3), 1991. doi: 10.1016/0012-365X(90)90358-O.

[8] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004. doi: 10.1103/PhysRevE.70.066111âĂŐ.

[9] H.A. Dijkstra. *Nonlinear Climate Dynamics*. Cambridge University Press, 2013. ISBN 9780521879170.

[10] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. The backbone of the climate network. *EPL (Europhysics Letters)*, 87(4):48007, August 2009. ISSN 0295-5075. doi: 10.1209/0295-5075/87/48007.

[11] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. Complex networks in climate dynamics. *The European Physical Journal Special Topics*, 174(1):157–179, July 2009. ISSN 1951-6355. doi: 10.1140/epjst/e2009-01098-2.

[12] SN Dorogovtsev, AV Goltsev, and JFF Mendes. Critical phenomena in complex networks. *Reviews of Modern Physics*, (80):1–79, 2008. doi: http://dx.doi.org/10.1103/RevModPhys.80.1275.

[13] Imme Ebert-Uphoff and Yi Deng. Causal Discovery for Climate Research Using Graphical Models. *Journal of Climate*, 25(17):5648–5665, September 2012. ISSN 0894-8755. doi: 10.1175/JCLI-D-11-00387.1.

[14] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[15] E.F. Gilbert. Random graphs. *Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[16] A. Gozolchiani, S. Havlin, and K. Yamasaki. Emergence of El Niño as an Autonomous Component in the Climate Network. *Physical Review Letters*, 107(14):1–5, September 2011. ISSN 0031-9007. doi: 10.1103/PhysRevLett.107.148501.

[17] O. Guez, A. Gozolchiani, Y. Berezin, S. Brenner, and S. Havlin. Climate network structure evolves with North Atlantic Oscillation phases. *EPL (Europhysics Letters)*, 98(3):38006, May 2012. ISSN 0295-5075. doi: 10.1209/0295-5075/98/38006.

[18] Jobst Heitzig, JF Donges, and Yong Zou. Node-weighted measures for complex networks with spatially embedded, sampled, or differently sized nodes. *The European Physical ...*, pages 1–21, 2012. doi: 10.1140/epjb/e2011-20678-7.

[19] J Hlinka, D Hartman, M Vejmelka, D Novotná, and M Paluš. Non-linear dependence and teleconnections in climate data: sources, relevance, nonstationarity. *Climate Dynamics*, 2013. doi: 10.1007/s00382-013-1780-2.

[20] IPCC-AR5. Climate Change 2013 The Physical Science Basis. Technical report, 2013.

[21] S. A. Jeffress and T. W. N. Haine. Correlated signals and causal transport in ocean circulation. *Quarterly Journal of the Royal Meteorological Society*, (October), March 2014. ISSN 00359009. doi: 10.1002/qj.2313.

[22] E Kalnay and M Kanamitsu. The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, 1996. doi: 10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

[23] H. Kutza. *Pattern recognition in complex networks, based on spatially embedded time series.* PhD thesis, Humboldt University of Berlin, 2012.

[24] G Lebon and D Jou. *Understanding non-equilibrium thermodynamics: foundations, applications, frontiers.* Springer, 2007. ISBN 9783540742517.

[25] Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási. Controllability of complex networks. *Nature*, 473(7346):167–73, May 2011. ISSN 1476-4687. doi: 10.1038/nature10011.

[26] Nishant Malik, Bodo Bookhagen, Norbert Marwan, and Jürgen Kurths. Analysis of spatial and temporal extreme monsoonal rainfall over South Asia using complex networks. *Climate Dynamics*, 39(3-4):971–987, August 2011. ISSN 0930-7575. doi: 10.1007/s00382-011-1156-4.

[27] Gareth J Marshall and A Harangozo. An appraisal of NCEP/NCAR reanalysis MSLP data viability for climate studies in the South Pacific. 27(19):3057–3060, 2000. doi: 10.1029/2000GL011363.

[28] N. Molkenthin, K. Rehfeld, N. Marwan, and J. Kurths. Networks from Flows-From Dynamics to Topology. *Scientific reports*, 2014. doi: 10.1038/srep04119.

[29] N. Molkenthin, K. Rehfeld, V. Stolbova, L. Tupikina, and J. Kurths. On the influence of spatial sampling on climate networks. *Nonlinear Processes in Geophysics*, 21:651–657, 2014. doi: 10.5194/npg-21-651-2014.

[30] N. Molkenthin, L. Tupikina, and J. Kurths. Flow networks from the discretized advection-diffusion-equation. *Submitted to Physical Review E*, 2014.

[31] Yamir Moreno, R Pastor-Satorras, and Alessandro Vespignani. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 26(4):521–529, 2002.

[32] NCEP/NCAR. *NCEP/NCAR*. URL `http://www.erls.noaa.gov/psd`.

[33] Zoltán Neufeld and Emilio Hernández-García. *Chemical and Biological Processes in Fluid Flows, a Dynamical Systems Approach*. 2001. ISBN 9780080439242. doi: 10.1016/B978-008043924-2/50055-9.

[34] NOAA/NCDC. NOAA Optimum Interpolation 1/4 Degree Daily Sea Surface Temperature Analysis (08.08.2013). URL `http://www.ncdc.noaa.gov/oa/climate/research/sst/oi-daily-information.php`.

[35] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic Spreading in Scale-Free Networks. *Physical Review Letters*, 86(14):3200–3203, April 2001. ISSN 0031-9007. doi: 10.1103/PhysRevLett.86.3200.

[36] R. Quian Quiroga, T. Kreuz, and P. Grassberger. Event synchronization: A simple and fast method to measure synchronicity and time delay patterns. *Physical Review E*, 66(4): 041904, October 2002. ISSN 1063-651X. doi: 10.1103/PhysRevE.66.041904.

[37] K. Rehfeld and J. Kurths. Similarity estimators for irregular and age-uncertain time series. *Climate of the Past*, 10(1):107–122, January 2014. ISSN 1814-9332. doi: 10.5194/cp-10-107-2014.

[38] K. Rehfeld, N. Marwan, S. F. M. Breitenbach, and J. Kurths. Late Holocene Asian summer monsoon dynamics from small but complex networks of paleoclimate data. *Climate Dynamics*, 41(1):3–19, September 2012. ISSN 0930-7575. doi: 10.1007/s00382-012-1448-3.

[39] K. Rehfeld, N. Molkenthin, and J. Kurths. Testing the detectability of spatio–temporal climate transitions from paleoclimate networks with the START model. *Nonlin. Processes Geophys.*, 21:691–703, 2014. doi: 10.5194/npg-21-691-2014.

[40] Richard W. Reynolds, Thomas M. Smith, Chunying Liu, Dudley B. Chelton, Kenneth S. Casey, and Michael G. Schlax. Daily High-Resolution-Blended Analyses for Sea Surface Temperature. *Journal of Climate*, 20(22):5473–5496, November 2007. ISSN 0894-8755. doi: 10.1175/2007JCLI1824.1.

[41] Karsten Steinhaeuser, Nitesh V. Chawla, and Auroop R. Ganguly. An exploration of climate data using complex networks. *ACM SIGKDD Explorations Newsletter*, 12(1):25, November 2010. ISSN 19310145. doi: 10.1145/1882471.1882476.

[42] Thomas Stocker. *Introduction to climate modeling*. Springer, 2009. ISBN 9783642007729.

[43] Veronika Stolbova, Paige Martin, Bodo Bookhagen, Norbert Marwan, and J. Kurths. Topology and seasonal evolution of the network of extreme precipitation over the Indian subcontinent and Sri Lanka. *Submitted to Nonlin. Proc. Geophys.*, (March), 2014.

[44] S H Strogatz. Exploring complex networks. *Nature*, 410(6825):268–76, March 2001. ISSN 0028-0836. doi: 10.1038/35065725.

[45] A. A. Tsonis and P. J. Roebber. The architecture of the climate network. *Physica A: Statistical Mechanics and its Applications*, 333, 2004. doi: 10.1016/j.physa.2003.10.045.

[46] Anastasios A. Tsonis, Kyle L. Swanson, and Paul J. Roebber. What Do Networks Have to Do with Climate? *Bulletin of the American Meteorological Society*, 87(5):585–595, May 2006. ISSN 0003-0007. doi: 10.1175/BAMS-87-5-585.

[47] Anastasios A. Tsonis, Geli Wang, Kyle L. Swanson, Francisco A. Rodrigues, and Luciano Da Fontura Costa. Community structure and dynamics in climate networks. *Climate Dynamics*, 37(5-6):933–940, July 2010. ISSN 0930-7575. doi: 10.1007/s00382-010-0874-3.

[48] L. Tupikina, K. Rehfeld, N. Molkenthin, V. Stolbova, N. Marwan, and J. Kurths. Detecting evolution of networks using spatial-temporal autocorrelation function. *Nonlin. Proc. Geophys.*, 21:705–711, 2014. doi: 10.5194/npg-21-705-2014.

[49] Renguang Wu, J. L. Kinter, and B. P. Kirtman. Discrepancy of Interdecadal Changes in the Asian Region among the NCEP–NCAR Reanalysis, Objective Analyses, and Observations. *Journal of Climate*, 18(15):3048–3067, August 2005. ISSN 0894-8755. doi: 10.1175/JCLI3465.1.

[50] K. Yamasaki, A. Gozolchiani, and S. Havlin. Climate Networks around the Globe are Significantly Affected by El Niño. *Physical Review Letters*, 100(22):228501, June 2008. ISSN 0031-9007. doi: 10.1103/PhysRevLett.100.228501.

[51] Liang Zhao, Ying-Cheng Lai, Kwangho Park, and Nong Ye. Onset of traffic congestion in complex networks. *Physical Review E*, 71(2):026125, February 2005. ISSN 1539-3755. doi: 10.1103/PhysRevE.71.026125.

List of Figures

Acknowledgements

I would like to thank Prof. Kurths for the atmosphere of freedom, trust and support, that provided a fertile ground for ideas to grow. Another big thanks goes to Kira Rehfeld for a very fruitful and productive office-companionship, I have yet much to learn from her about climate, data analysis and organisational skills. I want to thank Liubov Tupikina for her open, communicative and yet rigorously mathematical style of science. And of course I would like to thank Veronika Stolbova, for her data skills and appreciation of the importance of chocolate. Furthermore, I want to thank Norbert Marwan, Hannes Kutza, Jobst Heitzig, Reik Donner, Jonathan Donges and Jakob Runge for helpful and inspiring discussions, that were particularly valuable in the beginning of my PhD for finding the question.

I would like to thank the DFG Graduate School “Sichtbarkeit und Sichtbarmachung” for financial support and for an inspiring environment of interdisciplinary exchange of ideas. In particular I would like to thank Prof. Mersch for his unique combination of philosophy and maths, Fabian, for the interesting interdisciplinary collaboration, Mira for being a wonderful mentor and Lenore for her support with organizational matters.

I thank Frank Hellmann for many discussions, which clarified the conceptual foundations of this work, and for general awesomeness. I also thank Lisa Torlina for making our WG the best one I ever knew, for her support and the discussions about everything. Also big thanks to my awesome friends Ralf, Toni, Ewa, Burk, Benni, Thomas and Timo and my amazing Family: thanks Hedi and Merle and Thomas and Livia. And last but not least, thanks to Erik Rudolf for upside-down maths discussions.

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Berlin, den 02.07.2014 Nora Molkenthin