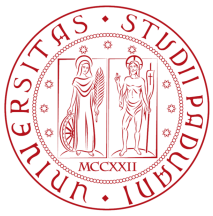

UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Corso di Laurea in Fisica

Tesi di Laurea

Ricostruzione delle dinamiche temporali in networks climatici: un approccio di ensemble

Reconstructing the temporal dynamics of climate networks: an ensemble approach

Relatore
Prof. Manlio De Domenico
Correlatore
Dr. Tomas Scagliarini

Laureando
Lorenzo Gamba

Anno Accademico 2022/2023

Abstract

Earth's climate is a compelling example of complex systems, with increasing interest in deciphering its underlying dynamics to improve weather forecasts and address climate change. This thesis represents a foundational step towards demystifying the intricate dynamics of Earth's climate system. It addresses the inverse problem of reconstructing a functional representation of this system through a rigorous and consistent Bayesian approach. The study analyzes network ensembles of daily temperature anomalies from 2664 global locations since 1970, uncovering significant changes in the network's structure, particularly after the early 2000s. These changes include a reduction in network connections and a rise in nodes with higher connectivity, notably in climatically important regions like the Antarctic and the Amazon Rainforest. This methodology not only illuminates the dynamic nature of the climate network but also deepens our understanding of its interconnectedness. The research introduces a novel and robust approach, proposing a method to increase our comprehension of the climate system's complexity.

Il clima terrestre, sistema complesso per eccellenza, è sempre più oggetto di studi per migliorare le previsioni meteorologiche e per comprendere e contrastare il cambiamento climatico. Questa tesi costituisce un passo essenziale verso la decifrazione delle complesse dinamiche del sistema climatico della Terra. Affronta il problema inverso di ricostruire una rappresentazione funzionale di questo sistema attraverso un metodo bayesiano rigoroso e sistematico. La ricerca analizza ensamble di reti basati su anomalie di temperature giornaliere registrate in 2664 località in tutto il mondo a partire dal 1970, evidenziando cambiamenti sostanziali nella struttura della rete, soprattutto dopo i primi anni 2000. Questi cambiamenti si manifestano con una diminuzione delle connessioni di rete e un incremento di nodi ad alta connettività, in particolare in aree climaticamente critiche come l'Antartide e la Foresta Amazzonica. Questo metodo non solo mette in luce la natura dinamica della rete climatica, ma arricchisce anche la nostra comprensione della sua interdipendenza. Il lavoro propone un approccio innovativo e solido, aprendo la strada a una maggiore comprensione della complessità del sistema climatico.

Contents

Introduction

Nowadays, the role of the analysis of the Earth's climate is widely recognized. Since the seminal work of Lorenz [12], which showed the intrinsic chaotic nature of the dynamics of the climate, today we assist to a dramatic improvement of the quality of the short-term weather forecasts. More recently, the already visible effects of climate change [15] raised the necessity of analyzing the long-term behavior of the climate variables, in an effort to understand and mitigate the impact of the human activities on the atmosphere.

In this thesis, we will analyze the temporal evolution of the correlation network of the daily temperature anomalies, extracted from a dataset starting from 1970 and collected over a world-grid with a size of 5×5 degrees. Using a novel Bayesian approach [16], our analysis seeks to determine the probability of existence of a link between two geographic locations, as opposed to the usual frequentist approach where the null hypothesis is accepted or rejected at a given confidence level. With this procedure, from a probabilistic network we are able to build an ensable of networks, which allows us to study the network descriptors with the appropriate confidence intervals.

The analysis of each calendar year from 1970 to 2022 shows a watershed in the temporal evolution of many topological descriptors between and after the year 2002. In particular, many indicators such as the global connectivity and the average path length, after remaining constant for the first decades, show a clear trend in the first 20 years of 2000. Our results suggest that the climate dynamics has reached a tipping point after a period of relative stationarity and that the changing is still ongoing.

Overall, this work aims to contribute to the growing field of climate network analysis by employing a Bayesian approach. Network science has become an increasingly influential perspective in climate studies, yielding insights into disaster prediction, climatic patterns, and tipping points [3, 11, 13]. Through this research, we hope to improve our understanding of complex climatic systems and ultimately, it seeks to propose a rigorous methodological approach that future studies can build upon, that could includes the analysis of further critical variables like precipitation or atmospheric pressure.

The thesis is structured as follow. In Chapter 1, we provide a brief methodological background for the principles of network theory and statistical inference. This will provide the necessary tools for understanding the subsequent chapters, as they provide the tools and frameworks necessary to analyze and interpret complex network structures and statistical models within our research.

In Chapter 2, we will describe the data set of climate data and the preprocessing steps. Subsequently, we will provide a detailed explanation of our Bayesian approach, which is employed to construct an annual ensemble of networks. This rigorous statistical method is designed to encode most of the information about the network connectivity.

Finally, in Chapter 3, we will present our findings from analyzing the climatic data. We will explore the temporal evolution of topological descriptors within each year's network. The goal is to uncover interaction patterns, discern trends, and deepen our understanding of the climate system's inter connectivity.

Chapter 1

Methods

In Section 1.1 we provide the basic tools of network theory, an ideal framework to gain insights on systems that exhibit complexity, an emerging feature that may occur when a huge number of units are interacting. In Section 1.2, we provide some elements of statistical inference, which involves extracting information from data.

1.1 Network Theory in Complex Systems

At the core of network theory is the concept of *network*, or *graph*, a discrete mathematical object that represents a collection of units, called *nodes* or *vertices*, and the interactions between them, also known as *edges*.

Formally, a graph is defined as $G = \{N, E\}$, where N is the number of nodes and E denotes the set of edges of the form $\{i, j\}$, representing the existence of a link between the nodes labelled i and j. A graph can be mathematically represented using the so called *adjacency matrix*, an $N \times N$ matrix whose elements indicate whether pairs of vertices are adjacent or not in the graph:

$$A_{i,j} = \begin{cases} 1 & \text{if } \{i, j\} \in E \\ 0 & \text{if } \{i, j\} \notin E. \end{cases} \tag{1.1}$$

If $A_{i,j} = A_{j,i}$ for all $\{i, j\} \in E$, the graph is said to be *undirected*, since links have no direction. On the other hand, a graph is *directed* if exists at least a $\{i, j\} \in E$ so that $A_{i,j} \neq A_{j,i}$. Directed graphs assign a direction to the link, implying that node A has a connection to node B, but not necessarily the other way around. An immediate example is the act of following on social networks. If the graph does not contain self-loops, then the diagonal elements of the matrix will be zero, as $A_{i,i} = 0$. Self-loops represent a connection of a node with itself. In a weighted graph, each edge is associated with a weight, which may represent the strength of the connection, the significance of the link or other characteristics. In this case, the adjacency matrix W is constructed using the weights $W_{i,j}$ corresponding to the edge i, j are used.

Graphs serve as versatile models for a wide array of systems and processes across various domains such as physical, biological, social, and informational. To describe these complex systems, numerous topological descriptors have been introduced to understand their structure and dynamics. Those mathematical descriptors have seen considerable development over the years, drawing inspiration from diverse fields of study including algebra, statistical mechanics, and quantum mechanics. In the following sections we describe the most important ones, that will be used for the analysis of the following chapters.

1.1.1 Degree

The degree of a node is the most fundamental and intuitive measure; it represents the number of connections or edges that the node has with other nodes. This measure has a different meaning

depending on whether the graph is directed or undirected. In the first case, it represents only the number of connections, while in the latter case we can differentiate between in-degree, the number of connections directed towards the node, and out-degree, the number of connections originating from the node. In the case of undirected graphs, the formula to calculate the degree of a node i from the adjacency matrix is

$$k_i = \sum_{j}^{N} A_{i,j}. \tag{1.2}$$

1.1.2 Clustering Coefficients

In many real world networks the connections between nodes has the property of *transitivity*, meaning that, if node A is connected with nodes B and C, then also nodes B and C are connected together. Intuitively, in social network terminology, we can say that "the friends of my friend are also my friends". The clustering coefficient quantifies how likely is that, given two nodes connected to a third node, they are also connected to each other. In other words, it could be said that if two nodes are neighbors to another, this measure captures the probability that they are also connected to each other. We can distinguish between the *local clustering* coefficient and the *global clustering* coefficient. The local coefficient assigns to each node its own coefficient through the formula

$$C_i = \frac{2e_i}{k_i(k_i - 1)}, \tag{1.3}$$

where k_i is the degree of node i and e_i is the number of edges between the neighbors of node i. Essentially, this is like dividing the number of present triangles that include that node by the number of possible triangles with the same node as a vertex. To obtain an indicative value of the entire network, it is possible to calculate the average local clustering, which is the mean of the local coefficients of all the nodes:

$$\langle C_{\text{local}} \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i. \tag{1.4}$$

The global coefficient, on the other hand, is another single value for an entire network and represents how interconnected the nodes are with each other. This measure is also known as the transitivity of the network and it reflects the probability that the adjacent vertices of a vertex are connected to each other.

$$C_{\text{global}} = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triplets of nodes}}. \tag{1.5}$$

The average local coefficient and the global coefficient do not always coincide and indicate slightly different things, the average local clustering coefficient can be high if many nodes have high local clustering, even if those clusters are not part of larger interconnected triads that contribute to the global clustering coefficient. Conversely, a high global clustering coefficient implies that not only do nodes tend to form clusters, but these clusters are interconnected throughout the network, forming a cohesive structure.

1.1.3 Assortative Mixing

In the context of network theory, mixing patterns are descriptors to describe connections types between nodes, in particular assortative mixing is the tendency of nodes to be connected to other nodes that have similar features. The most commonly used method to measure the assortative mixing is to calculate the degree-degree correlation coefficient r

$$r = \frac{\frac{1}{M}\sum_{e_{ij}} j_i k_i - \left(\frac{1}{M}\sum_{e_{ij}} \frac{1}{2}(j_i + k_i)\right)^2}{\sigma^2} \tag{1.6}$$

$$\sigma^2 = \frac{1}{M}\sum_{e_{ij}} \frac{1}{2}(j_i^2 + k_i^2) - \left(\frac{1}{M}\sum_{e_{ij}} \frac{1}{2}(j_i + k_i)\right)^2 \tag{1.7}$$

where the terms are defined as:

- M is the total number of edges in the network.
- e_{ij} is the edge that connect nodes i and j
- j_i is the degree of one endpoint of the e_{ij}-th edge.
- k_i is the degree of the other endpoint of the e_{ij}-th edge.

Looking to the r coefficient it is possible to understand if nodes tend to connect with other nodes that have a similar number of connections (degree). The network is fully assortative mixing when $r = 1$, and fully dis-assortative mixing when $r = -1$. If r is approximately zero, the network is not showing statistically significant assortative or dis-assortative patterns, implying that its connectivity is similar to that of a random graph. Another way to understand the assortativity of a graph is to study the scaling hypothesis of $k^{(nn)}(k) \propto k^{\mu}$. If the scaling exponent is positive the network is assortative, if negative it is disassortative and if it is near to zero it is neutral.

1.1.4 Average Path Length

Every couple of nodes in the network can be linked through a path that can be direct or pass over a multiple nodes. We can define the shortest path length d_{ij} as the minimum number of edges that must be crossed to go from node i to node j. The average path length of a network is the average distance between any pair of nodes, where $N(N-1)$ is the number of all possible edges that a network can present.

$$L = \frac{1}{N(N-1)}\sum_{i \neq j} d(i,j) \tag{1.8}$$

This is the formula for an unweighted graph. In the event that the network is weighted, one must be careful in each situation by reasoning about what the weights represent. For example, if the weights are the geometric distance between two nodes, then the distance between two nodes will be the path that tends to minimize the sum of the weights on that path.

1.1.5 Molloy-Reed Coefficient

The Molloy-Reed coefficient, also known as the heterogeneity parameter, provides insights on the network structure as it is defined as the ratio of the average squared degree $\langle k^2 \rangle$ divided by the average degree $\langle k \rangle$:

$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle}. \tag{1.9}$$

Using this coefficient, we can gain a deeper understanding of the network's structure, as its value gives us information about the dimension of the Largest Connected Component (LCC). When κ is high, the LCC encompasses a substantial fraction of the network's nodes, creating a scenario where a significant number of nodes are interconnected. This interconnectivity allows for the possibility of reaching many other nodes from any given node through a series of steps. In fact, if we analyze a random network as for example an Erdős-Rényi, where each pair of nodes has a fixed probability p of being connected independently of other pairs, when $\kappa > 2$, when the number of nodes is large, the network undergoes a percolation phase transition and a macroscopic LCC emerges, indicating a high level of connectivity [1]. In a scale-free network, where the degree distribution follows a power law and a small number of nodes, known as hubs, have a high number of connections, the interpretation is quite different. The Molloy-Reed coefficient is also defined as the ratio between the second moment (the average of the squares of the degree) and the first moment (the average degree) of the degree distribution, k. This coefficient can be shown to be proportional to N^{-d}, as explained in [20], where N

is the number of nodes and d is a parameter inferred from the data to characterize the structure of the network. In our case, we will not examine the analysis for scale-free networks; instead, we will simply focus on observing how this coefficient changes over time as a measure of the network heterogeneity. Also, this analysis serves as an initial step towards applying percolation theory in order to understand the network's resilience and vulnerability, especially in the presence of hubs. [1].

1.2 Statistical Inference

Statistical inference is the process of analyzing data by employing statistic methods. The objective of inference is to enhance our understanding, enabling us to make predictions or decisions. This is particularly important in complex systems where we aim to discern patterns, rules, or properties. There are two school of thought in statistical inference: frequentist and Bayesian.

1.2.1 Frequentist Inference

The frequentist approach defines the definition of probability of an event as the limit of its relative frequency in a large number of trials. In this perspective, probabilities are considered objective and are calculated by considering long-run frequencies of events. As such, the frequentist methodology emphasizes the dimension of the sample and the accuracy of the results is proportioned within the number of test or analyzed data. A key concept in frequentist inference is hypothesis testing, which involves establishing a null hypothesis and an alternative hypothesis. Statistical tests are then used to determine whether the observed data significantly deviate from what is expected under the null hypothesis. Parameters are associate with confidence intervals, which, at a certain confidence level (typically 95%) gives a range where we expect the true value of a parameter to fall, with a certain degree of certainty. At a 95% confidence level, if we repeatedly sampled the population and calculated the interval each time, we would anticipate the true parameter to be within these intervals 95% of the time.

1.2.2 Bayesian inference

The Bayesian approach instead is based on a prior knowledge of the event we are indicating on. The main point of the Bayesian approach is to assign a probability, called prior, to an hypothesis (H), and to update this probability to a posterior probability, on the basis of a new evidence (E). This is done using the Bayes' theorem, which allows us to update the prior probability using data to obtain a corrected posterior probability. The mathematical form of the theorem is:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)} \tag{1.10}$$

where

- $P(H|E)$ is the probability of hypothesis H given the evidence E, known as the posterior probability.
- $P(H)$ is the probability of hypothesis H being true before the evidence is seen, known as the prior probability.
- $P(E|H)$ is the probability of observing evidence E given that H is true, known as the likelihood.
- $P(E)$ is the probability of observing the evidence.

The prior probability represents what is already know about our problem and about the hypothesis we are considering. It can be objective when it is based on previous research or data or subjective when reflect personal belief. The likelihood is a function applied to the analyzed data utilizing some statistical model. It is fundamental because it update the prior probability in a posterior one. Some criticize the Bayesian approach because it can be really depends on personal opinion represented y the prior probability but in some case like complex system it can be really helpful to direct the analysis

considering different aspect that with a frequentest approach can be difficult to incorporate. For example, in climate, distance is really a factor that can change the interpretation of inter connection of Earth locations and a prior probability based on distance could help distinguish between causality or casualty.

Chapter 2

Climate networks

In this chapter we give a comprehensive description of the data set used in the analysis and the various pre-processing steps. Then, we describe the procedure used to build the cross-correlation networks starting from the climatological time series. A pictorial representation of the analysis can be seen in fig. 2.1.

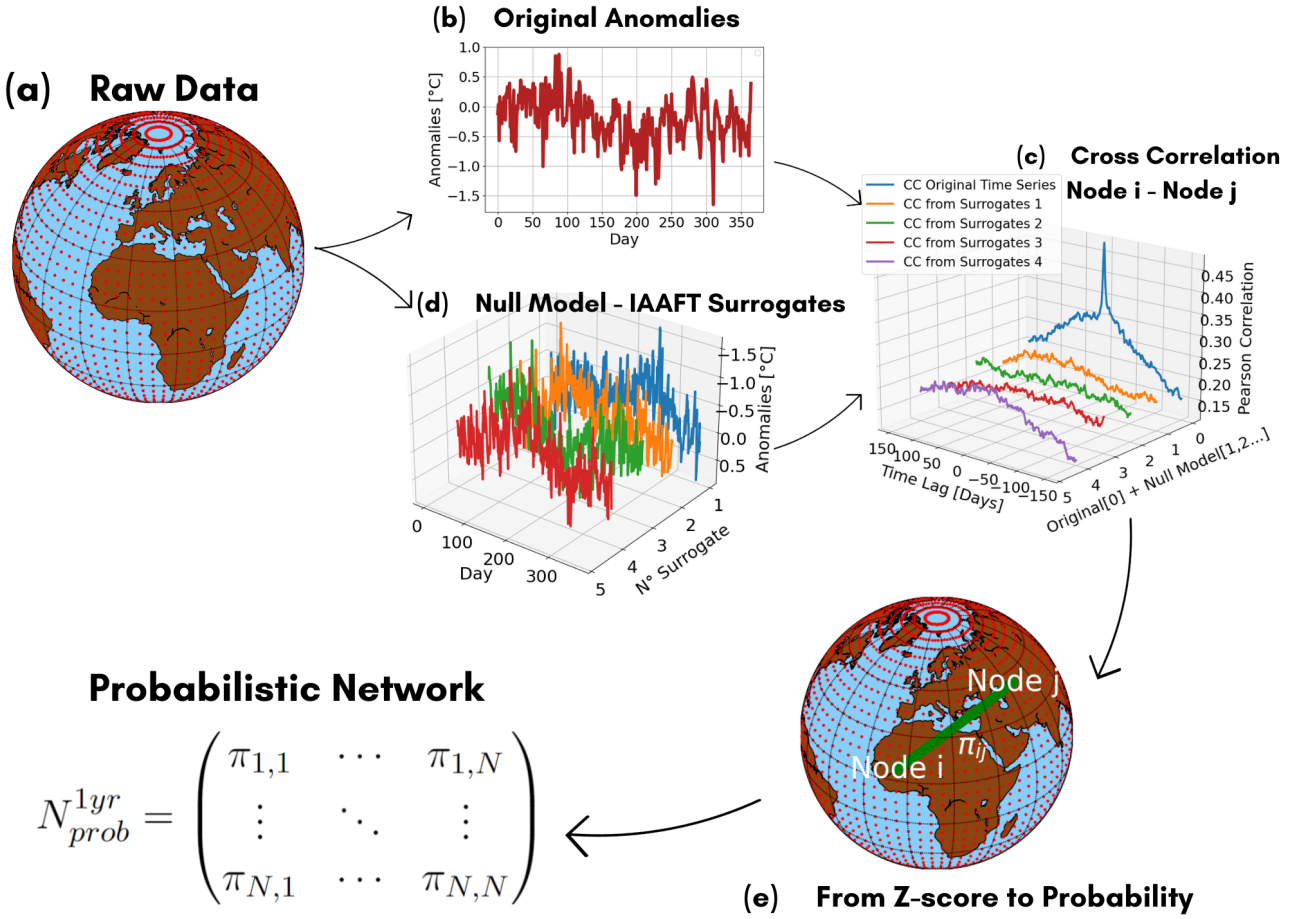

Figure 2.1: **The pipeline of the network construction .** Temperature data are collected on a world-grid (a) and then the anomalies are extracted (b) For each pair of links the cross-correlation is computed (c) and the relative null model (d). From this result, the probability matrix is computed using the Bayesian approach described in 1.2 (e).

2.1 Climate data and anomalies extraction

The data set used in this analysis is freely available and is part of the ERA5 reanalysis produced by the European Centre for Medium-Range Weather Forecast (Appendix D). The climatic variables in the reanalysis are obtained indirectly by interpolating data between the earth's surface and the lowest atmospheric level, considering current weather conditions [10]. The hourly data of 2 meters surface temperature has been collected from 1st January 1970 to 31st December 2021. The data is arranged on a world-wide grid with a spatial resolution of 5°x5° (fig. 2.1.a) and represents the temperature of the air at 2 meters above the ground, sea, or inland waters. For our analysis, we resampled the data to daily frequency by averaging the hourly records, obtaining a total of 2664 time series, each of them having a length of 19,358 daily observations. To obtain the daily temperature *anomalies*, we subtract each observation from the usual temperature on that day (calculated over a 20 years baseline period, from 1970 to 1989)(fig. 2.1.b). This step is necessary to avoid spurious correlations due to effects of seasonality: if we used original data we would find that, on average, the temperatures are anti correlated when, for example, in one hemisphere is winter and in the other is summer. Also, for instance, the similarity between a temperature time series of a node in the Arctic and one in the Amazon rain forest would be low. In other words, an anomaly provides us information whether the temperature on that day was warmer or colder compared to the usual temperatures. Specifically, a positive anomaly indicates that the temperature for that day was warmer than the mean of all the days with the same calendar data and location during the baseline period, while a negative anomaly means it was colder.

In order to conduct a temporal analysis, we divided the whole time series into 52 time windows, from the 1st January to 31th December of each year, from 1970 to 2022.

2.2 Cross-correlation analysis

We used the linear Pearson cross-correlation to quantify the similarity between anomalies on two locations (fig. 2.1.d), as it is widely used in the study of the climate system [2, 3, 8, 13, 21] The cross-correlation ρ_{ij} between two time series $x_i(t)$ and $x_j(t)$ at lag τ reads:

$$\rho_{ij}(\tau) = \frac{\sum_{t=0}^{n-\tau} x_i(t)x_j(t+\tau)}{\sqrt{\sum_{t=0}^{n-\tau} x_i^2(t)x_j^2(t+\tau)}}, \tag{2.1}$$

where for our investigations we incorporate a maximum time delay τ of five months. As we aim to create an undirected edge when comparing two nodes, the maximum value of the cross-correlation may be associated with a positive or negative time delay. Identifying the implications of this distinction presents an opportunity for additional research.

Starting from the anomalies time series of temperature for each node, we build a null model (fig. 2.1.d) formed by 30 iterative amplitude adjusted Fourier transform (IAAFT) surrogates [18] (see Appendix B for details on how they are calculated). Specifically, the IAAFT method destroys the correlations between data by preserving the auto-correlation of the original series. It is necessary to emphasize that surrogates have been created for each year range, this way the mixed values of the original time series remain within the same time period, in order to keep the comparison consistent. Generating this null model it is possible to quantify, taken two nodes, how much the maximum correlation value ρ^{max} of the two original time series of the anomalies is significant compared to the mean of the maximum cross-correlation values of the IAAFT surrogates $< \rho^{max}_{\text{surr}} >$. We quantify this obtaining a Z-score that is calculated as follows:

$$Z_{\text{score}} = \frac{|\rho^{max} - < \rho^{max}_{\text{surr}} >|}{\sigma_{\rho^{max}_{\text{surr}}}}. \tag{2.2}$$

The Z-score itself is a powerful statistical tool that provides insight into the degree of deviation from a null hypothesis. But in this case, it measures how much the observed correlation differs from what

would be expected by chance. For each pair of nodes, a Z-score is calculated, yielding a total of 52 $N \times N$ matrices of Z-scores, one for each year:

$$Z_{\text{score}}^{1yr} = \begin{pmatrix} Z_{1,1} & \cdots & Z_{1,N} \\ \vdots & \ddots & \vdots \\ Z_{N,1} & \cdots & Z_{N,N} \end{pmatrix}.$$

Then, we map the Z-scores into the p-values p, that is the probability of observing the statistic if the null hypothesis is true, using the complementary error function which gives the area under the Gaussian curve from the Z-score to infinity:

$$p = \text{erfc}(Z) = 1 - \text{erf}(Z) = \frac{2}{\sqrt{\pi}} \int_Z^\infty e^{-t^2}\, dt. \tag{2.3}$$

This is particularly useful for computing the probability of observing a value in the tail of a normal distribution, beyond a certain number of standard deviations from the mean, and in our case this is exactly our Z-value. We expect a high z-score value for two nodes that are geographically close, and to investigate this aspect, we have depicted the distributions of the z-scores as a function of distance, dividing them into intervals of 500 km (fig. 2.2). After approximately 2000 km, the distributions tend to become similar.

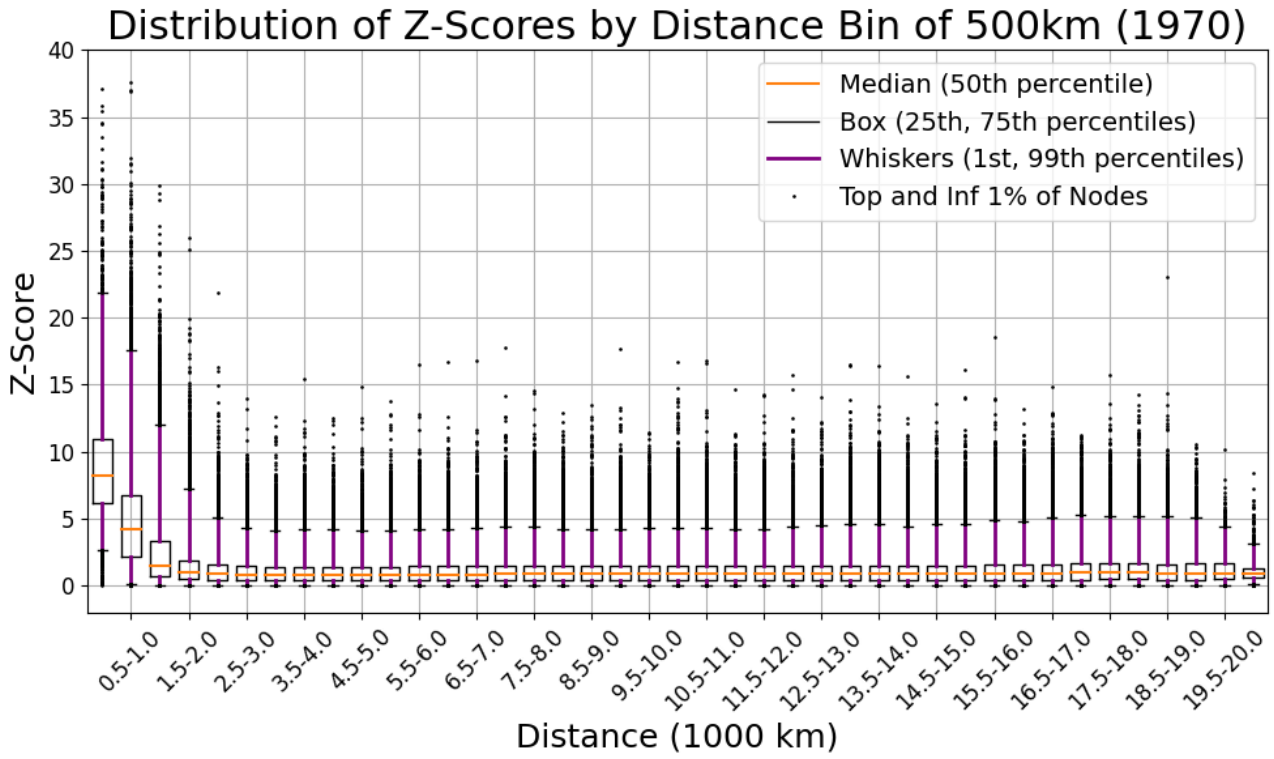

Figure 2.2: **The distribution of the Z-scores as a function of the distance.**T he z-scores of all possible node pairs have been categorized based on the distance in kilometers between the two nodes. In each boxplot, the z-score values whose distances fell within a specific range every 500km have been included. This way, we can get an idea of the intensity of the connections as a function of the distance. Each boxplot illustrates the median, lower (25th percentile) and upper (75th percentile) quartiles, with whiskers extending to the 1st and 99th percentiles to capture the full range of Z-score variability within each distance range. Outliers, which appear beyond these limits, highlight extreme variations.

2.3 Probability of a link with Bayesian approach

In our analysis, we adopt a Bayesian inference procedure described in [16] to transform the z-score obtained in the cross-correlation analysis into a probability of existence $\pi_{i,j}$ of a true interaction

between two locations labelled i and j. This method to create a climate network bypasses the standard procedure of validating results with a fixed significance threshold by updating a prior probability based on evidence or observations. In particular for our case, we choose a prior probability based on the spatial auto-correlation of the signals, that decays exponentially with distance. We use this equation as our prior probability:

$$P_{ij} = \exp\left(-\frac{d_{ij}}{K}\right), \tag{2.4}$$

where d_{ij} is the distance in kilometers between the points i and j and K a positive parameter. We choose $K = 2000$ in the case of temperature anomalies, as it was observed that this value is the typical spatial scale beyond which the spatial correlations become negligible [2]. Also the z-score values exhibit a very similar distribution after 2000 km (fig. 2.2), which reinforces the justification for the choice of this prior. The distance d_{ij} is calculated using the haversine formula (Appendix C). Upon obtaining the p-value, we apply the Bayesian method to calculate the probability of a link which includes calculating the Bayes factor:

$$B_{\text{factor}} = \begin{cases} -ep\log p & \text{if } p < e^{-1} \\ 1 & \text{if } p > e^{-1} \end{cases} \tag{2.5}$$

A Bayes factor is needed because p-values are frequently misinterpreted and this can lead to incorrect conclusions in statistical analysis. For this reason calibrating p-values is crucial to correct their common misinterpretation as error probabilities or as a measure of the hypothesis's truth [9, 19]. Finally, using the definition, the Bayes factor can be mapped into the minimum posterior probability for the null hypothesis given the p-value:

$$\pi_{ij} = 1 - \left[(1 + \left(\frac{B_{\text{factor}} \cdot (1 - P_{ij})}{P_{ij}}\right)^{-1}\right]^{-1}. \tag{2.6}$$

π is the probability of the existence of the edge between two nodes (fig. 2.1.e).
In (fig. 2.3) you can have a geographical representation of the results of our method to calculate the probabilities of links. The zscores heatmap presents the values obtained in the comparison between the cross-correlation of local temperature anomalies with those of the node indicated by the black cross at the coordinates (-60,-25). Applying the Bayesian method described above the zscores become probabilities represented in the second heatmap.

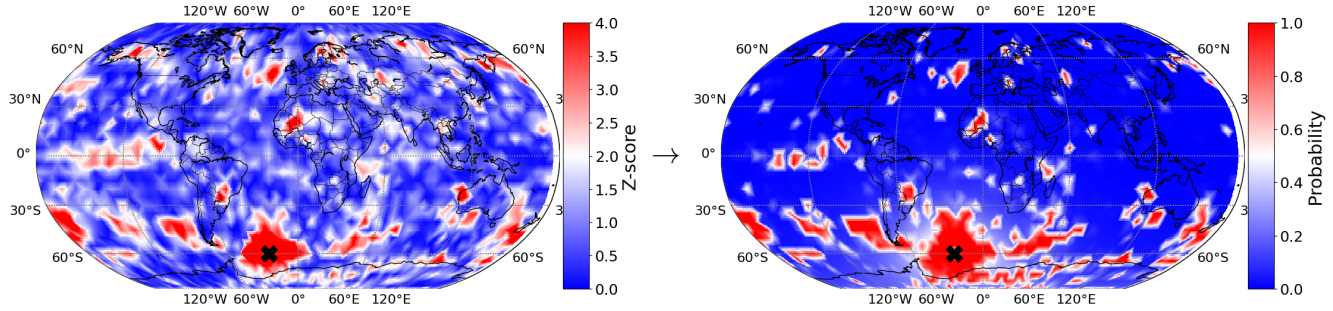

Figure 2.3: **From Z-score to Probability Heatmaps.** Transformation of Z-scores into posterior probabilities for the node at coordinates (-60, -25) in 2016. This particular node is the one with the highest connectivity.

2.3.1 Reconstruction via Probabilistic Networks

All those link probabilities lead to the creation of a matrix N_{prob}^{1yr} which represents a probabilistic network, also known as a fuzzy network:

$$N_{prob}^{1yr} = \begin{pmatrix} \pi_{1,1} & \cdots & \pi_{1,N} \\ \vdots & \ddots & \vdots \\ \pi_{N,1} & \cdots & \pi_{N,N} \end{pmatrix}.$$

We will compile 52 of these probabilistic matrices to represent the climate network for each year. After that, we generate matrices of the same size with random numbers between 0 and 1, defining an edge if the value is less than the respective probability. With this process we create one hundred adjacency matrices. These annual sets represent an ensemble of climate networks (fig. 2.4), which serve as the basis for our analyses to reconstruct the temporal dynamics. In this way, we also incorporate the possibility of measurement error, improving the analysis even in the absence of data uncertainty information. This allows us to find the descriptors most likely to represent the true underlying structure [14].

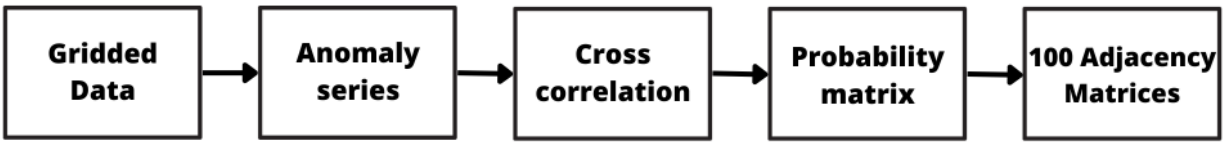

Figure 2.4: **Construction scheme of annual climate-based ensamble.**

Chapter 3

Results and Discussion

To characterize the evolution of networks over time, we adopt the following approach. For each year, we take the calculated fuzzy networks and generate an ensemble of 100 networks. This method enables us to compute both the average and standard deviation of descriptors to underline the annual evolution of the climate structure based on temperatures. Our approach proves to be effective in accurately monitoring the dynamics even in situations characterized by uncertainty.

3.1 Number of Connections

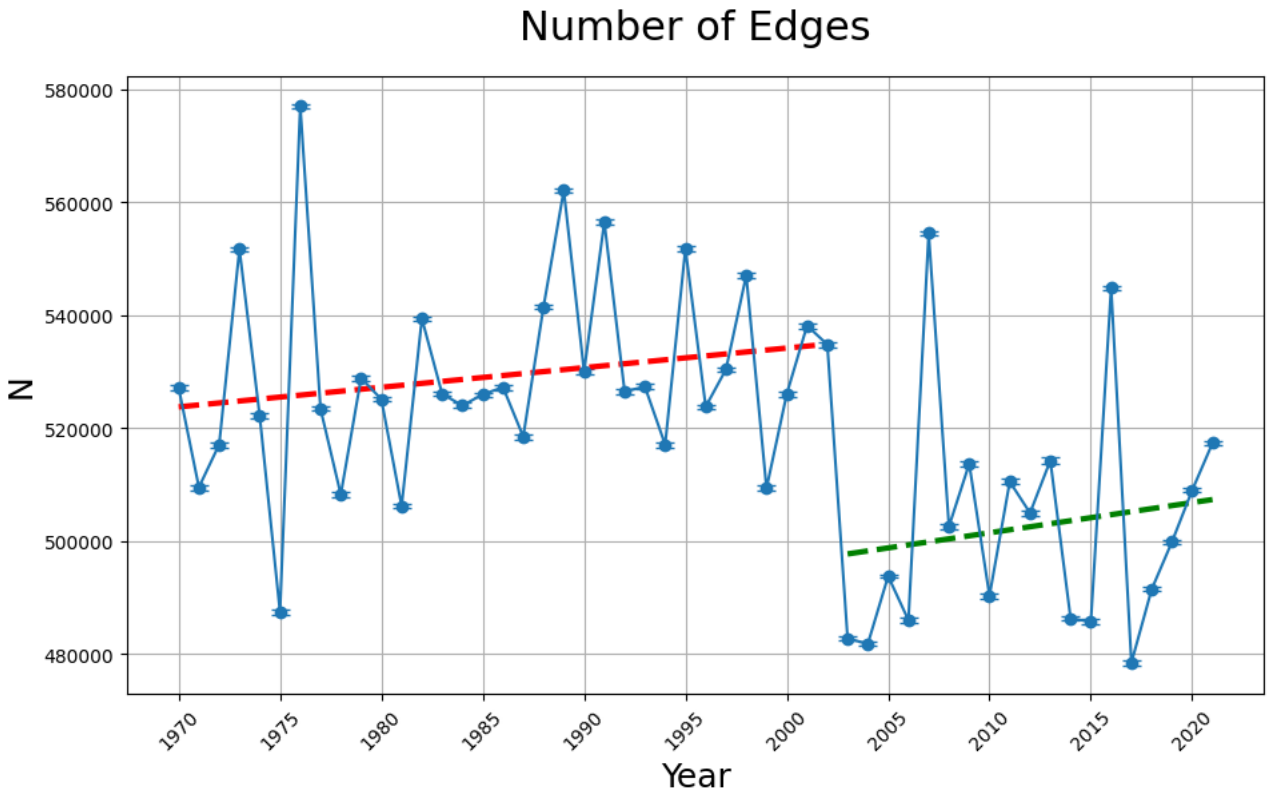

Figure 3.1: **The total number of connections over time.** Mean and Standard Deviations of the number of edges extracted from the annual ensemble of networks generated for each year. The two lines interpolate the data before and after the year 2002.

Starting with a basic measure of the network, the first metric we want to examine is the annual number of links. As previously mentioned, the graph displays the mean and standard deviations derived from the ensemble of generated networks. Notably, across a 52-year period, the most striking

observation is a clear shift after the year 2002, where the number of edges significantly decreases, with the exception of two years (2007 and 2016). This shift is visibly significant even the trend keeps positive increasing. The year 2003 is remembered in the history of climate to have shown an abnormal heatwave that hit Europe with the subsequent assignment of year of breakthrough for Europe [6]. The standard deviations are quite small relative to the mean values, indicating a high similarity among the generated networks. This consistent probability distribution may be attributed to the large number of potential links $N(N-1)/2 = 3.547.116$. Therefore, in light of the Central Limit Theorem and the Law of Large Numbers, it's reasonable to observe such small standard deviations [Appendix A].

3.2 Connectivity as Degree for Climate Networks

The nodes do not cover equal surface areas, in fact those nearer to the Earth's poles encompass smaller areas. Given the closer proximity of these nodes, it's likely that their time series of temperature anomalies will be similar, leading to a higher probability of strong correlation and, consequently, a greater likelihood of forming links. This behavior predictably distorts the degree of the nodes. In fact, a node at the pole is expected to have a higher number of connections. To address this, we introduce a customized weighted version of the degree, as described in [21]. This descriptor for a Climate Network structure represents the *surface area of the globe to which the node is connected*, relative to the total surface covered by the network's nodes, as defined by the following equation:

$$C_i = \frac{\sum_{j=1}^{N} \cos(\theta_j)}{\sum_{\forall\theta,\phi} \cos(\theta)}, \tag{3.1}$$

where C_i represents the connectivity of node i, θ is the latitude, and ϕ is the longitude.

The degree of each node was determined by calculating the average degree across the ensemble of 100 networks generated from the probabilistic matrix. For each of the 52 years, we represented the average weighted degree for each node by creating a heat map covering the Earth (3.2). In those heat map the red part represent points with high connectivity in therm of percentage of surface reached.

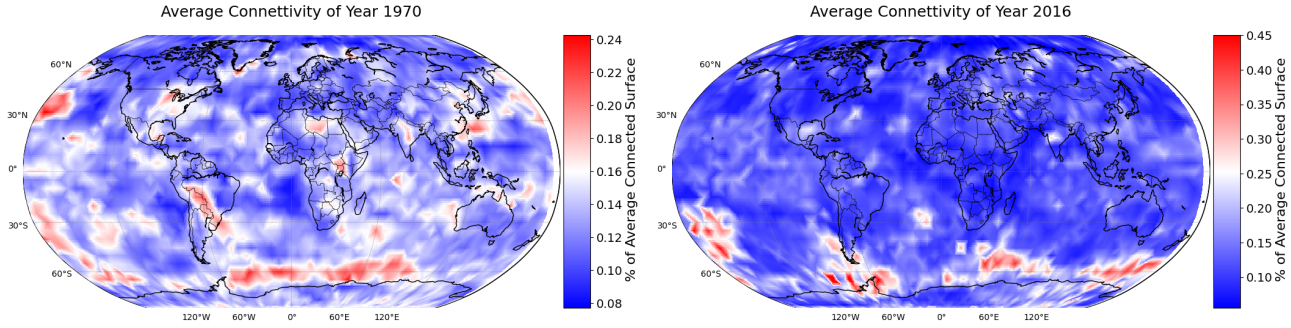

Figure 3.2: **Heatmaps depicting the connectivity of each individual node.** The years 1970 and 2016 are highlighted in particular, showing very different maximum values, as can be seen from the colorbar scale.

We are interested in studying how the distributions change over the years. To achieve this, we created a boxplot depicting the average connectivity of all nodes. The averages were calculated from the data of all 100 networks generated (fig. 3.3). In this graph, we emphasize the medians, along with the 1st, 25th, 75th, and 99th percentiles, as well as the minimum and maximum values. To better highlight the changes in these distributions, we plotted the widths of the percentiles and the min-max range. As the final indicator, we introduced the Coefficient of Variation, defined in the following manner:

$$CV = \frac{\sigma}{\bar{x}} \tag{3.2}$$

The CV is consistent with the standard deviation, but it also visibly follows the trends of other measures. Observing the scale of widths, we notice a strong upward trend, with values doubling the

percentage of the globe's surface covered. This trend is also mirrored by the IQR. The growth begins more abruptly around the year 2000, where it takes on a visibly positive trajectory. However, seeing that the medians remain relatively constant while the width increases reflects a shift in the more extreme connections, particularly in the higher values of connectivity.

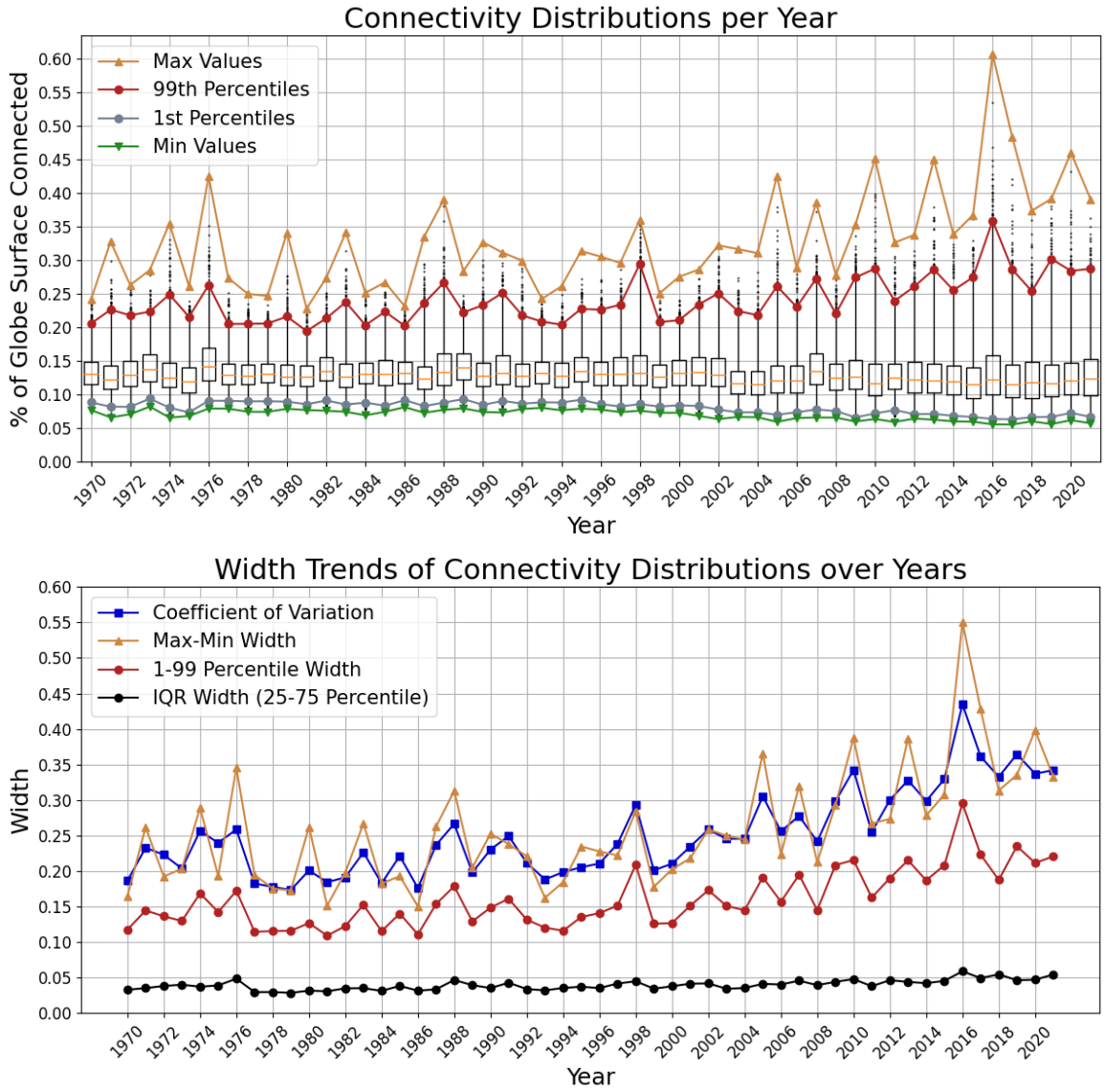

Figure 3.3: **Distributions of connectivity for each year.** The first image depicts the boxplots with the medians and the interquartile range. The maximum and minimum values, as well as the 1st and 99th percentiles, are then highlighted with lines. The second figure, on the other hand, aims to emphasize the widening of these distributions by depicting the thickness of the interquartile ranges, the 1-99 percentiles, and the minimum-maximum thickness. The coefficient of variation is included to reflect the standard deviation, which is sensitive to extreme values, so it provides a global parameter of the distribution.

To identify regions with the highest connectivity over a 52-year period, we generated a global heatmap by averaging annual connectivity values (fig. 3.4). This heatmap shows that areas with the most connectivity are primarily located in the oceans near Antarctica, with remarkable areas also present in South America. In these regions there is the Antarctic Circumpolar Current (ACC) and the Amazon Rainforest. Both of them are well known as potential tipping points [5,11,23] in the Earth's ecological system and, as our analysis demonstrates, they are highly connected to the rest of the Earth's surface. Certainly noteworthy is the presence of a distinctly dark blue zone adjacent to the Amazon Rainforest's red zone, which sharply contrasts with other global areas. This contrast may signal a region with a unique connectivity profile, suggesting the necessity for more in-depth studies to fully comprehend

this phenomenon.

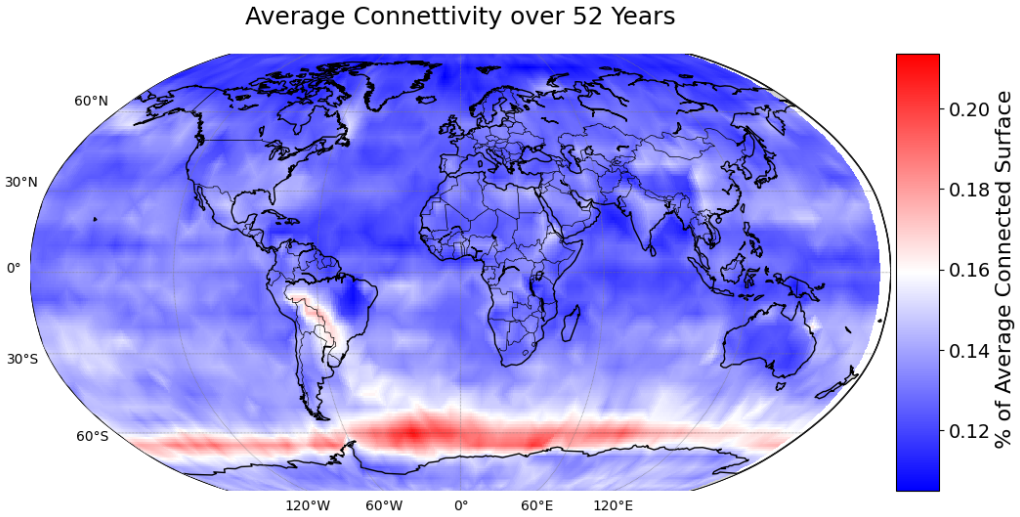

Figure 3.4: **Heatmap depicting the average connectivity of every single node over a span of 52 years.** This heatmap highlights the average percentage of surface that each node was connected to over the analyzed 52-year period. We can observe that nodes with highest values are located in the ocean near Antarctica and in the Amazon rainforest.

In addition to our findings, the Amazon Rainforest, as mentioned in [5], shows a significant loss of resilience since the early 2000s and this is observed even in areas where the broad leaf fraction has not decreased significantly. The Amazon Rainforest is crucial for global climate stability and biodiversity and it usually acts as a carbon sink but has exhibited signs of declining ecosystem productivity, and during major droughts, it temporarily becomes a carbon source due to increased tree mortality. Always in this article the authors report that the Amazon's resilience loss during periods of significant droughts is linked to sea surface temperature anomalies in the northern tropical Atlantic Ocean and shifts in the Atlantic Multidecadal Oscillation. In our analysis the fact that the nodes with highest connectivity lay in the ocean, as indicated by our findings, could suggest the ocean's strong role as a regulator of the climate structure.

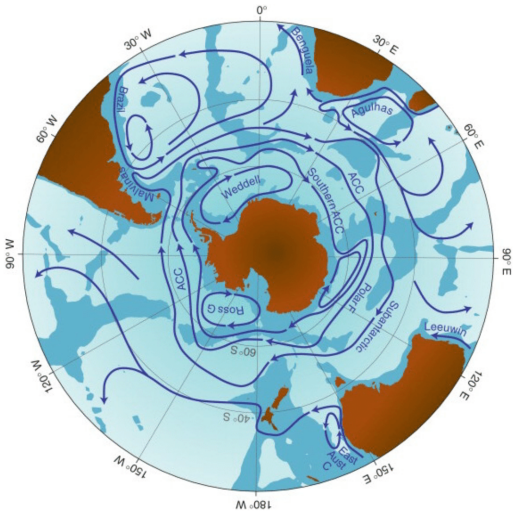

Figure 3.5: **The Antarctic Circumpolar Current (ACC)** This image showcases the path of the Antarctic Circumpolar Current (ACC), the largest and most powerful ocean current system that encircles Antarctica. The relentless flow of the ACC plays a pivotal role in global climate regulation and in linking the world's oceans, impacting marine biodiversity and climate cycles. The image is taken from [17]

The nodes that have a high connection value within our climate network are significantly influenced by the Antarctic Circumpolar Current (ACC) (3.5), the largest oceanic current system on Earth. The ACC plays a pivotal role in the global distribution of water masses, as extensively discussed in [23]. The current's strength is chiefly determined by the Southern Westerly Winds and surface buoyancy forces. A key geographical feature that affects the ACC is the Drake Passage, located under South America, where we find nodes with maximum values of connectivity over a span of 52 years (fig. 3.4). This passage serves as a crucial bottleneck, shaping the ACC's eastward trajectory around Antarctica. Furthermore in those locations, the ACC's path is integral to the formation of the northward-returning southward flow of the Circumpolar Deep Water, a component of the Atlantic Meridional Overturning Circulation (AMOC), both part of the global Thermohaline circulation (fig. 3.6). Recent observations, as detailed in [7], have highlighted notable changes in the AMOC, including a discernible decline in its strength between 2004 and 2012. These shifts in the ACC and AMOC underscore the dynamic nature of oceanic currents and their profound impact on the global climate system.

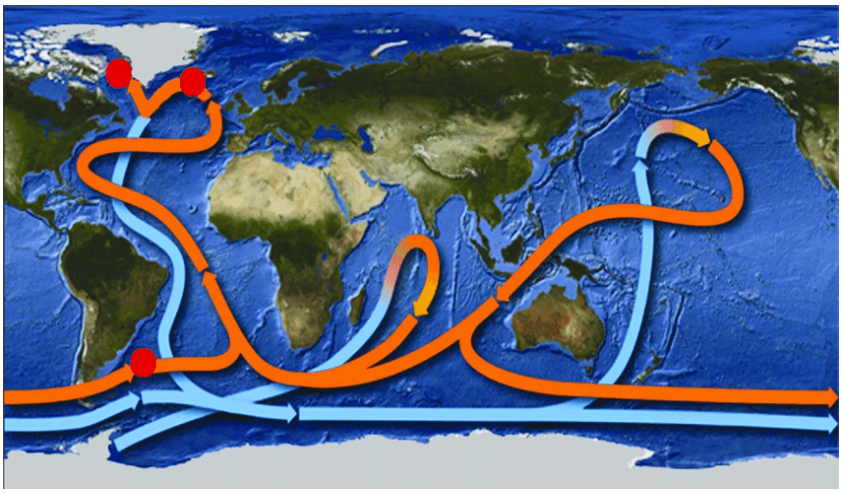

Figure 3.6: **Diagram of the global thermohaline circulation .** *Thermo* refers to the influence of temperature gradients, and *Haline* pertains to the influence of salinity (salt concentration) gradients. The red circles indicate the locations where heat exchange occurs between the atmosphere and the water masses in thermohaline circulation. ACC is the current flowing around Antarctica while AMOC is the current that characterizes the Atlantic Ocean. The figure is reproduced from [4].

3.3 Clustering Coefficients

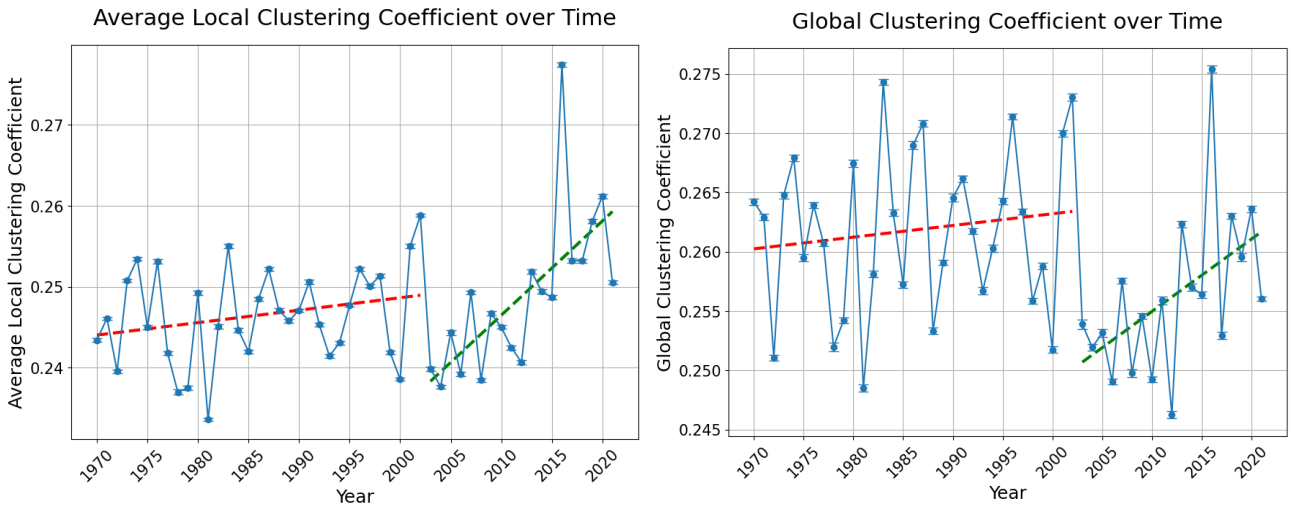

Figure 3.7: **Average Local and Global Clustering Coefficients over Time.** Mean and Standard Deviations of the Average Local and Global Clustering Coefficients, derived from the yearly ensemble of networks generated annually. The two lines interpolate the data before and after the year 2002.

As explained in Section 1.1.2, the clustering coefficient measures the likelihood that nodes connected to a given vertex are also interconnected. We depicted both the average local coefficient and the global coefficient (fig. 3.7) as the mean values from 100 networks generated by the probabilistic network model. It's worth noting that, in this case as well, there is a change in the trend after the beginning of the years 2000.

3.4 Assortative Mixing

In the context of assortative mixing (Sect. 1.1.3), we evaluate whether nodes have a propensity to connect with others of similar connectivity. We calculate the degree-degree correlation coefficient for the customized degree representation and for the topological degree (fig. 3.8). The customized degree reflects the likelihood of a node being linked to others with a comparable percentage of surface area connected. The 'topological' term refers to using the standard degree, which may be skewed due to the varying proximity of nodes.

Interestingly, these measurements reveal opposing trends. While the measure related to the connected surface area increases over time, the topological one shows a decrease. This suggests that the network's topological structure is moving towards neutral assortativity. Additionally, while nodes increase their extreme values of connectivity as seen above, similar ones tend to be more linked.

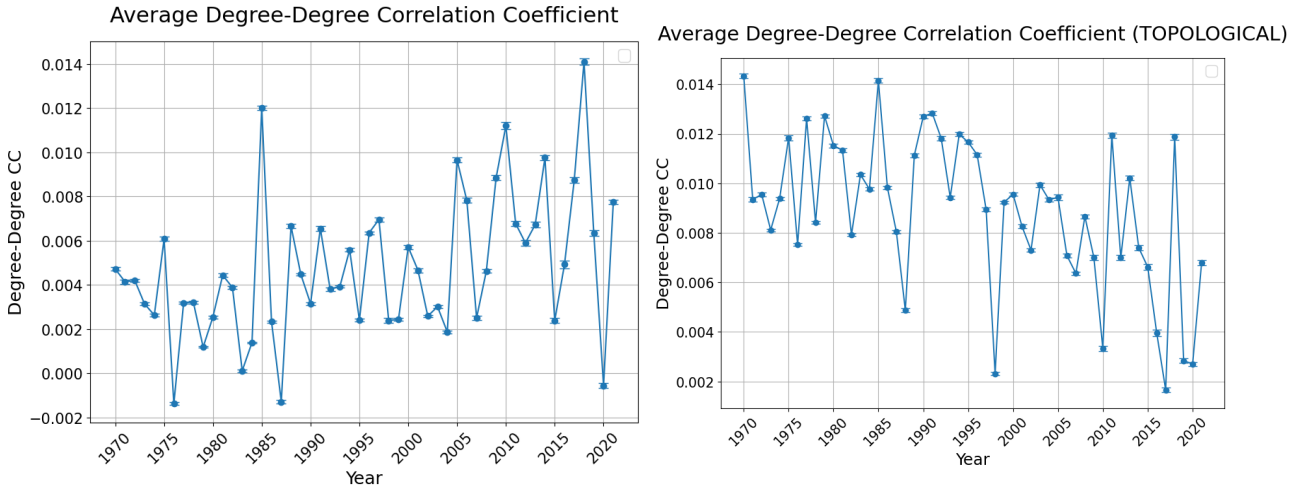

Figure 3.8: **Degree-Degree Correlation Coefficients over Time.** Mean and Standard Deviations of the degree-degree coefficient, extracted from the annual ensemble of networks generated each year. In the first image, the degree is defined as connectivity (percentage of the surface connected to that node), while in the second one, it refers to the topological degree.

For the scaling exponent (fig. 3.9), we can adopt the same considerations, as the trend of the function is the same as that of the degree-degree correlation.

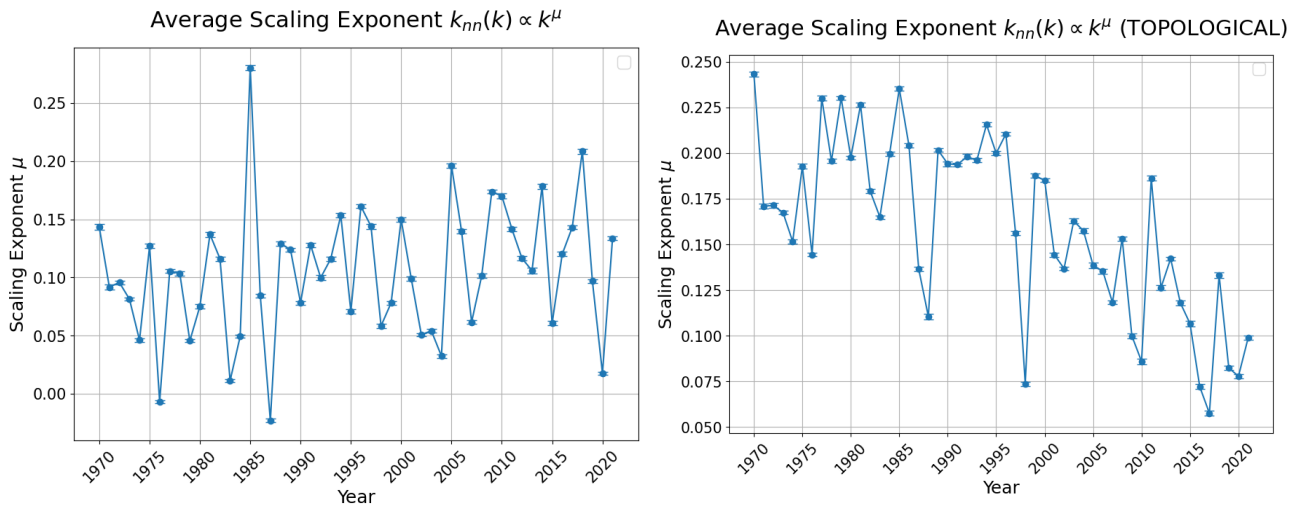

Figure 3.9: **Scaling Exponent over Time.** Mean and Standard Deviations of the Scaling Exponent, extracted from the annual ensemble of networks generated each year. In the first image, the degree is defined as connectivity (percentage of the surface connected to that node), while in the second one, it refers to the topological degree.

3.5 Average Path Length

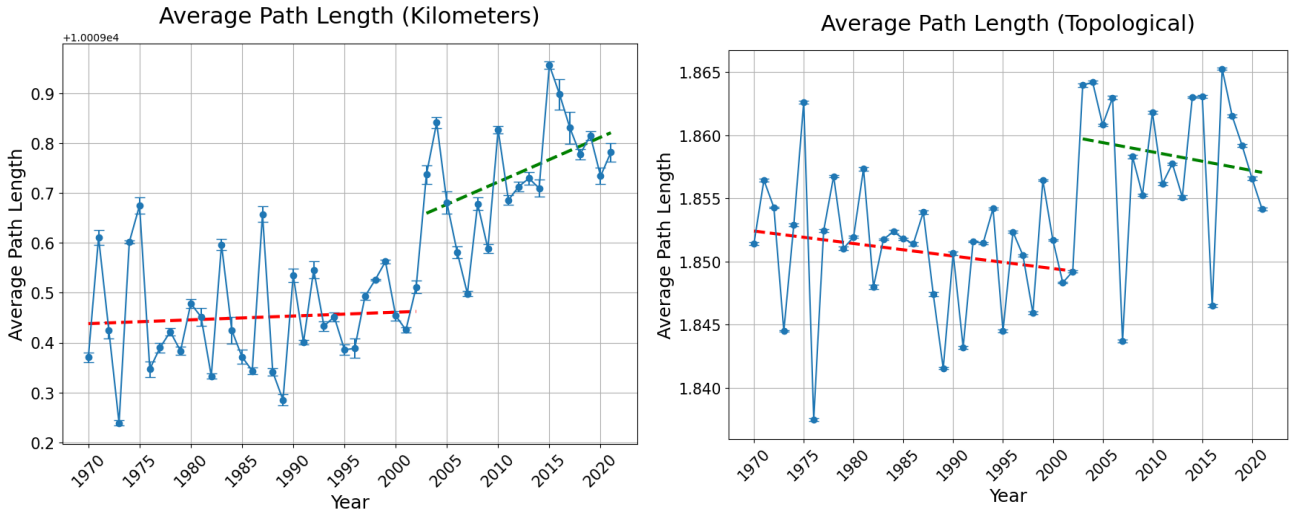

Figure 3.10: **Average Path Length over Time.** Mean and Standard Deviations of the Average Path Length, derived from the annual ensemble of networks generated each year. In the first image, the edges are weighted by the distance in kilometers between connected nodes. In contrast, the second image focuses on the topological path, where the y-values represent the average number of edges needed to connect any pair of two nodes. The two lines interpolate the data before and after the year 2002.

We calculated the Average Path Length (Sect. 1.1.4) for both the version based on the actual distances between nodes and its topological counterpart. The distance-weighted version of the Average Path Length represents the average of the shortest paths to connect all possible pairs of nodes, with edges measured in kilometers. A value of 10,010 km is acceptable as it is approximately a quarter of the Earth's circumference. This indicates an increase in the average, which can be associated with a distancing of the connected nodes. On the other hand, the topological version, while showing the usual trend change after the beginnings of 2000s, remains relatively constant. The low value of this measure indicates a high overall connectivity, likely attributable to a small-world structure. In such a structure, despite the nodes appearing numerous and thus more dispersed, they are in reality easily connected by a path consisting of a few edges (fig. 3.11).

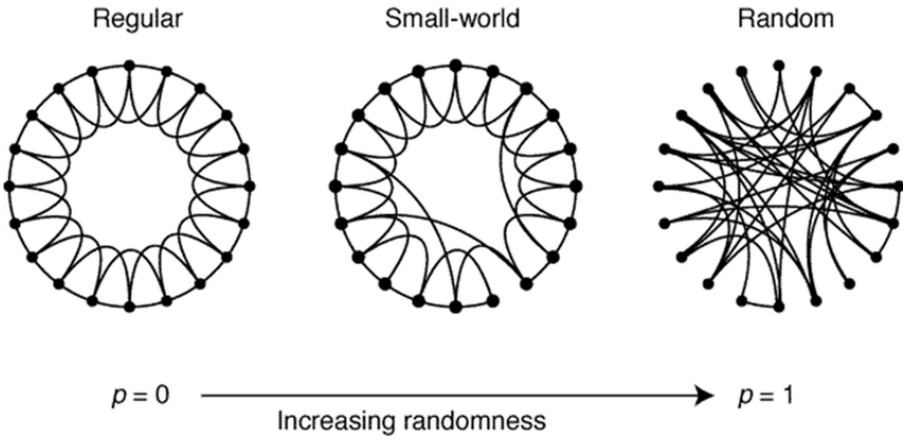

Figure 3.11: **Different structures of a network increasing randomness.** A "small-world network" is a type of mathematical graph in which most nodes are not neighbors of one another, yet most nodes can be reached from every other by a small number of steps. Small-world networks represent a balance between order and randomness. The figure is reproduced from [22].

3.6 Molloy-Reed Coefficient

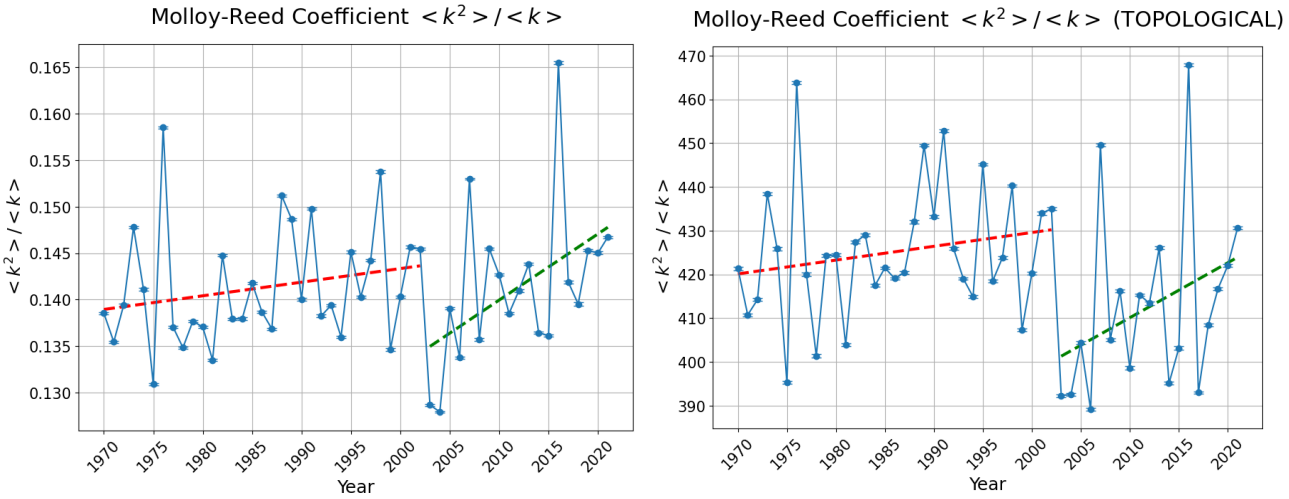

Figure 3.12: **Molloy-Reed Coefficient over Time.** Mean and Standard Deviations of the Molloy-Reed Coefficient, extracted from the annual ensemble of networks generated each year. In the first image, the degree is defined as connectivity (percentage of the surface connected to that node), while in the second one, it refers to the topological degree. The two lines interpolate the data before and after the year 2002.

The Molloy-Reed coefficient provides clear information when we consider its topological version (fig. 3.12). Its value, significantly greater than 2, indicates the strong presence of a Largest Connected Component (LCC) which includes a big fraction of the nodes. This can also be explained by the Bayesian prior used in the construction of the probabilistic network, which accounts for the proximity between two nodes. The presence of a LCC and the facility to form connections with nearby nodes, along with the results of the average path length, confirm the small-world structure of the climate networks.

3.7 Considerations about our Method

This rigorous Bayesian approach to studying climate reveals that the structure of the climate network has been changing over the years. Starting with the number of edges, there is a noticeable reduction in their numbers after the year 2002. Additionally, many descriptors change trends after the early

2000s, with some doubling their value. In fact, as we can see in Fig. 3.2, there is an emergence of nodes with higher connectivity. This is evident in the mapped representation of connectivity in two different years, 1970 and 2016, which exhibit distinct distributions. The climate networks over the years have always maintained a Largest Connected Component that includes a large fraction of the network's nodes and a small-world structure, but the topological and geographical connections are changing.

Additionally the described method could be also used as an optimal visual instrument to identify long distance connection also know as Teleconnections [3]. Just to give an example on how this could be made let's have a look to figure 3.13. Here we can see a strong correlation between temperature anomalies from a node in the middle of the Sahara Desert and the Drake Passage.

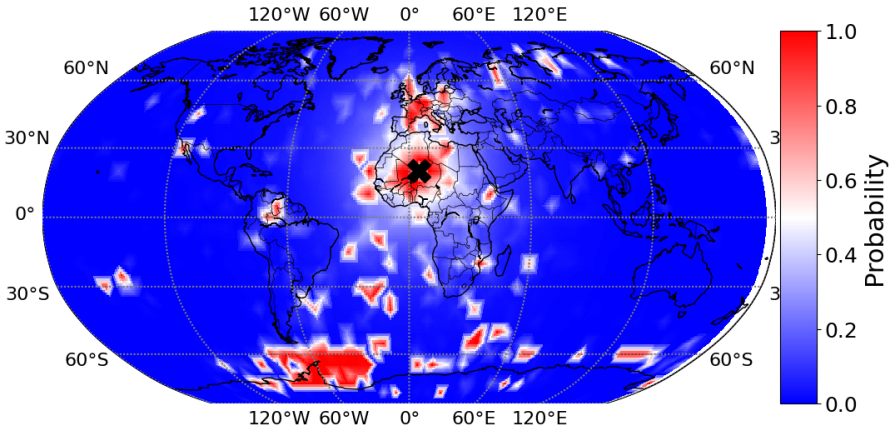

Figure 3.13: **Probability Node at coordinates (20° latitude, 5° longitude) in Year 2016.** The heatmap highlights for each node the value of the probability of a link with the node marked by a black cross during the year 2016.

As previously explained, these probabilities are derived from temperature anomalies. However, by incorporating additional variables such as precipitation, pressure, and humidity, and reapplying the Bayesian approach, we can construct a multi-layer probabilistic network. Analyzing the geographical positions of these connections and integrating data across all layers enables a more comprehensive understanding of teleconnections and possibly the identification of tipping points using percolation theory. This approach could allow for a deeper exploration of the complex interactions within the climate system.

Conclusions

This thesis has provided a comprehensive analysis of the climate network by employing a Bayesian approach, revealing the changing dynamics of this complex system since the 1970s. Through the analysis of daily temperature anomalies across global locations, it has been shown that the network's structure has undergone significant transformations, especially after the early 2000s. The methodology developed and utilized in this thesis not only illuminates the dynamic nature of the climate network but also proposes a new approach to increase our understanding of the system's complexity. The study's findings suggest that the climate network is changing and perhaps it is the dynamics of ocean currents play a important role. This research represents a significant step in the field of climate network analysis, providing insights that could aid in the development of more accurate predictive models for climate change and its impacts.

The implications of this research emphasize the need for continued study and adaptation in our approaches to understanding the Earth's climate system. It also highlights the necessity of integrating various climatic variables and data sources to build a more comprehensive picture of the climate network. The work done in this thesis lays the foundation for the way for future explorations in this field, contributing to the global effort to better predict and mitigate the effects of climate change.

Appendix A

Behavior of Standard Deviations with Increasing Number of Nodes

In this appendix, we demonstrate how the standard deviations in Erdős-Rényi graphs become smaller as the number of nodes N increases, applying the Central Limit Theorem (CLT) and the Law of Large Numbers (LLN). In Erdős-Rényi graphs, each edge between a pair of nodes exists independently with a fixed probability p. As the number of nodes increases, the number of potential edges, considered as independent random variables, also increases.

According to the CLT, the distribution of the sum or average of a large number of these independent variables will approximate a normal distribution. Consequently, in larger graphs, the distribution of the number of edges becomes more concentrated around the mean, leading to smaller standard deviations.

Additionally, the LLN states that as the number of observations (here, the potential edges) increases, the average of these observations approaches the expected value. For Erdős-Rényi graphs, this expected value is p times the total number of possible edges, $N(N-1)/2$. Therefore, as N grows, the actual number of edges in the graphs tends to converge to this expected number, further reducing the variability and standard deviation. To validate our assumptions, we generated 100 Erdős-Rényi graphs for each specified number of nodes: 10, 100, 500, 1000, 2500, and 5000. While plotting the mean and standard deviation offers some insights, these graphs alone may not effectively convey the nuances of the resulting distributions. Therefore, we chose to plot the coefficient of variation 3.2, which provides a clearer understanding of the distribution's skewness and variability relative to the mean (fig. A.1). This metric helps in better interpreting how 'peaked' or dispersed the distributions are, offering a more meaningful analysis of the above conclusion.

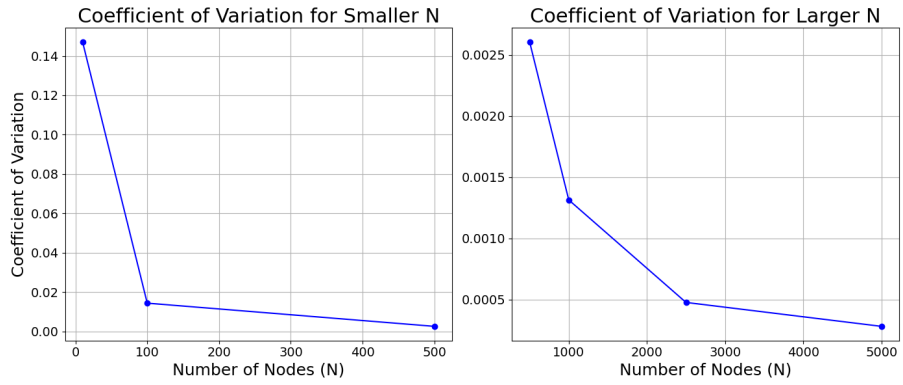

Figure A.1: **Coefficient of variation as the number of nodes changes.** Variation coefficients for an ensemble of 100 Erdős-Rényi networks with a link probability $p = 0.5$, varying the number of nodes N. A low coefficient of variation value indicates that the standard deviation is small in terms of magnitude compared to the absolute value of the mean.

Appendix B

IAAFT surrogates as null models

Surrogate data testing is used to build confidence intervals for the corresponding null hypothesis. A null hypothesis for a wide class of stochastic processes can be formulated by stating that all the structures in a time series are encoded in the mean, the variance and the auto-covariance function. For a Gaussian linear process x_t, these quantities are specified from the power spectrum

$$|S_k|^2 = |\frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} x_t \exp\Big(\frac{i2\pi kt}{N}\Big)|^2.$$

In this case, surrogate time series x_t^* are readily created by multiplying the Fourier transform of the data by random phases and then transforming back to the time domain:

$$x_t^* = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} e^{i\alpha_k} |S_k| \exp\Big(-\frac{i2\pi kt}{N}\Big),$$

where $0 \leq \alpha_k \leq 2\pi$ are independent uniform random numbers.

Appendix C

Haversine Formula

The haversine formula is a function of spherical trigonometry that calculates the distance between two points on a spherical surface, given their longitudes and latitudes.

$$d_{ij} = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_j - \phi_i}{2}\right) + \cos(\phi_i)\cos(\phi_j)\sin^2\left(\frac{\lambda_j - \lambda_i}{2}\right)}\right) \tag{C.1}$$

Where

- r is the authalic radius of Earth that is 6371 km.
- Latitude and longitude of the first node (in radians): (ϕ_1, λ_1)
- Latitude and longitude of the second node (in radians): (ϕ_2, λ_2)

Appendix D

Data Availability

The ERA5 reanalysis data used are publicly available at
https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels.

Bibliography

[1] A.-L. Barabási and M. Pósfai. *Network science*. Cambridge University Press, Cambridge, 2016.

[2] Y. Berezin, A. Gozolchiani, O. Guez, et al. Stability of climate networks with time. *Sci Rep*, 2:666, 2012.

[3] N. Boers et al. Complex networks reveal global pattern of extreme-rainfall teleconnections. *Nature*, 566(7744):373–377, 2019.

[4] R. Bojariu, M.-V. Birsan, R. Cică, L. Velea, S. Burcea, A. Dumitrescu, S. Dascălu, M. Gothard, A. Dima, F. Cărbunaru, and L. Marin. *Schimbările climatice – de la bazele fizice la riscuri și adaptare*. 01 2015.

[5] C. A. Boulton, T. M. Lenton, and N. Boers. Pronounced loss of amazon rainforest resilience since the early 2000s. *Nature Climate Change*, 12(3):271–278, 2022.

[6] N. Christidis, G. S. Jones, and P. A. Stott. Dramatically increasing chance of extremely hot summers since the 2003 european heatwave. *Nature Climate Change*, 5(1):46–50, 2015.

[7] P. Ditlevsen and S. Ditlevsen. Warning of a forthcoming collapse of the atlantic meridional overturning circulation. *Nature Communications*, 14(1):4254, 2023.

[8] J. Fan, J. Meng, J. Ludescher, X. Chen, Y. Ashkenazy, J. Kurths, S. Havlin, and H. J. Schellnhuber. *Statistical physics approaches to the complex Earth system*, volume 896. 2021.

[9] Held. A nomogram for p values. *BMC Medical, Research Methodology*, 10:21, 2010.

[10] H. Hersbach et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[11] T. Liu, D. Chen, L. Yang, et al. Teleconnections among tipping elements in the earth system. *Nat. Clim. Chang.*, 13:67–74, 2023.

[12] E. N. Lorenz. Deterministic nonperiodic flow. *Journal of atmospheric sciences*, 20(2):130–141, 1963.

[13] J. Ludescher, M. Martin, N. Boers, A. Bunde, C. Ciemer, J. Fan, S. Havlin, M. Kretschmer, J. Kurths, J. Runge, V. Stolbova, E. Surovyatkina, and H. Schellnhuber. Network-based forecasting of climate phenomena. *Proc Natl Acad Sci U S A*, 118(47):e1922872118, 2021.

[14] L. Peel, T. Peixoto, and M. De Domenico. Statistical inference links data and theory in network science. *Nat Commun*, 13:6794, 2022.

[15] S. Rahmstorf and D. Coumou. Increase of extreme events in a warming world. *Proceedings of the National Academy of Sciences*, 108(44):17905–17909, 2011.

[16] S. Raimondo and M. De Domenico. Measuring topological descriptors of complex networks under uncertainty. *Physical Review E*, 103(2):022311, 2021.

[17] S. R. Rintoul. Antarctic circumpolar current. In J. H. Steele, editor, *Encyclopedia of Ocean Sciences (Second Edition)*, pages 178–190. Academic Press, Oxford, 2009.

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

[19] T. Sellke, M. Bayarri, and J. Berger. Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71, 2001.

[20] M. Serafino, G. Cimini, A. Maritan, A. Rinaldo, S. Suweis, J. R. Banavar, and G. Caldarelli. True scale-free networks hidden by finite size effects. *Proceedings of the National Academy of Sciences*, 118(2), Dec. 2020.

[21] A. Tsonis and K. Swanson. Topology and predictability of el niño and la niña networks. *Physical review letters*, 100:228502, 2008.

[22] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[23] S. Wu, L. Lembke-Jene, F. Lamy, H. W. Arz, N. Nowaczyk, W. Xiao, X. Zhang, H. C. Hass, J. Titschack, X. Zheng, J. Liu, L. Dumm, B. Diekmann, D. Nürnberg, R. Tiedemann, and G. Kuhn. Orbital- and millennial-scale antarctic circumpolar current variability in drake passage over the past 140,000 years. *Nature Communications*, 12(1):3948, 2021.