Procedia Computer Science

Volume 51, 2015, Pages 404–412

ICCS 2015 International Conference On Computational Science

The construction of complex networks from linear and nonlinear measures – Climate Networks

J. Ignacio Deza[1] and Hisham Ihshaish[2]

[1] Department of Physics and Nuclear Engineering,
Polytechnic University of Catalonia, Barcelona, Spain
juan.ignacio.deza@upc.edu
[2] Department of Computer Science and Creative Technologies,
University of the West of England - Bristol, the UK
hisham.ihshaish@uwe.ac.uk

Abstract

During the last decade the techniques of complex network analysis have found application in climate research. The main idea consists in embedding the characteristics of climate variables, e.g., temperature, pressure or rainfall, into the topology of complex networks by appropriate linear and nonlinear measures. Applying such measures on climate time series leads to defining links between their corresponding locations on the studied region, whereas the locations are the network's nodes. The resulted networks, consequently, are analysed using the various network analysis tools present in literature in order to get a better insight on the processes, patterns and interactions occurring in climate system. In this regard we present ClimNet; a complete set of software tools to construct climate networks based on a wide range of linear (cross correlation) and nonlinear (Information theoretic) measures. The presented software will allow the construction of large networks' adjacency matrices from climate time series while supporting functions to tune relationships to different time-scales by means of symbolic ordinal analysis. The provided tools have been used in the production of various original contributions in climate research. This work presents an in-depth description of the implemented statistical functions widely used to construct climate networks. Additionally, a general overview of the architecture of the developed software is provided as well as a brief analysis of application examples.

Keywords: complex systems, climate networks, time series analysis, graph theory

Software availability

Availability: ClimNet - Software will be available shortly in a public repository, meanwhile researchers are recommended to contact the authors for the source code.
Programming languages: C++, Python
Software requirements: NetCDF, C++ compiler. Python (optional, for pre- and post-processing).

 Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2015

doi:10.1016/j.procs.2015.05.260

1 Introduction

Many techniques of complex network analysis have found application to complex systems. Examples of these include, but not limited to, applications in airline transportation networks modelling [**?**], social interactions [**?**] or the Internet [**?**]. The study of network properties have proven to be vital tools to analyse and predict the behaviour of such complex systems. These systems share the characteristic that they can be straightforwardly represented in terms of a well-defined set of nodes coupled via links that have a clear physical interpretation.

In a continuous system, however, like the atmosphere or the ocean, it is not always clear how to define the relevant network nodes, and there is usually no obvious interaction that can be used to define links between them. In order to tackle this problem, the so called interaction networks have been recently used. As such, the observed atmospheric or ocean locations serve as nodes, and the links are calculated based on statistical measures of similarity (e.g. correlation coefficient) between pairwise time series of observations at the studied locations.

In like manner, climate networks (CNs) represent a statistical similarity structure of spatio-temporal resolved climatological variables, which depend on this definition of nodes and links. Their spatial sampling results in a space-embedded topology, and thus, a careful interpretation of the inferred network is required.

CNs have been successfully employed to analyse climate features, for instance the analysis of significant correlations [**?**, **?**] and atmospheric teleconnections [**?**], the study of local phenomena as El Niño-Southern Oscillation (ENSO) [**?**]. Other CNs techniques have been applied to investigate the connections between sea surface temperature (SST) variability and the global mean temperature [**?**] and to study the possible collapse of the meridional overturning circulation (MOC) [**?**] on the north Atlantic.

On a related note, climate system is known for its multi-scale interactions and hence one would like to explore the interaction of processes over the different scales. Data are available through high-resolution ocean/atmosphere/climate observational and model simulations but they lead to networks with a big number of nodes (~ 10000) making the process of network construction computationally demanding. Moreover, when high-resolution model data are used, constructing such networks becomes very challenging using currently available software tools. For such scales, parallel climate network construction and analysis tools have been developed, e.g. Par@Graph [**?**]. However, these tools are not usually based on information theoretic techniques, but rather on linear correlation calculations.

In contrast, ClimNet enables the construction of *directed* and *non-directed* climate networks based on state of the art Information-theoretic tools, besides ordinary correlation coefficient measures. As well, ordinal analysis technique is implemented to analyse time scales directly from the data. Additionally, robust statistical analysis are in place to assure the significance of the results.

The provided features of ClimNet have been used in previous works [**?**, **?**, **?**] to construct and analyse climate networks from correlations in surface air temperature (SAT) time series. More precisely, the software have been used to process the interdependencies between the time series and use this information – together with a statistical significance test – to create an adjacency matrix suitable to be explored and analysed by conventional network analysis tools and interpreted from a climatological point of view (see section 5).

In this paper the software to construct complex climate networks is described, mainly from a functional (features) perspective, with special attention to the implemented statistical and Information-theoretic measures. The software functions have been computationally optimised and safe multithreading has been also introduced to enable parallel processing on multi-core

machines. However, since our major focus here is on the usability and application of the tools, their application to real climate data will be addressed here, whereas the scalability and performance of ClimNet will be discussed elseweher in a coming work.

On a different matter, It should be mentioned that knowledge in climate dynamics is necessary in order to understand the patterns calculated by the the presented software. However, the analysis of the networks are done by running common graph theory algorithms to find network metrics including degree distribution, closeness and betweenness centrality.

2 Measures for Climate Networks Construction

The software is equipped with functions to produce Information-theoretic measures - non linear correlation and information transfer between time series. All the measures (except entropy) are pairwise, i.e. based on the similarity found between each pair of time series. The list of these measures are described as follows:

2.1 Cross Correlation

It is implemented for benchmarking and for comparison with information theoretic techniques. The Pearson correlation coefficient [**?**] is defined as:

$$\mathrm{E}[(X-\mu_X)(Y-\mu_Y)] = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y) \tag{1}$$

where x_i and y_i are X and Y are two random variables with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y and E is the expected value.

2.2 Entropy

The Shannon's definition of entropy is generally defined in terms of the probability density function (PDF) p_i, of the system to be in state x out of a possible set of states $\mathcal{A}$:

$$H = -\sum_{i\in\mathcal{A}} p_i \log_b p_i. \tag{2}$$

Here, p_i is the probability of the value number i to appear in a sequence of characters of a given time series. This measure can be used to better understand the characteristics of the series. It is an univariate measure and the result of the calculations does not yield a network.

2.3 Mutual Information

Mutual information is computed from the probability density functions (PDFs) that characterise two time series in two nodes, p_i and p_j, as well as their joint probability function, p_{ij} [**?**]:

$$M_{ij} = \sum_{m,n} p_{ij}(m,n)\log\frac{p_{ij}(m,n)}{p_i(m)p_j(n)}. \tag{3}$$

M_{ij} is a symmetric measure

$$M_{ij} = M_{ji} \tag{4}$$

of the degree of statistical interdependence for the time series $i(t)$ and $j(t)$; if they are independent:

$$p_{ij}(m,n) = p_i(m)p_j(n) \tag{5}$$

and thus $M_{ij} = 0$.

2.4 Directionality Index

The directionality index (DI) is defined as:

$$\mathrm{DI}_{XY}(\tau) = \frac{I_{XY}(\tau) - I_{YX}(\tau)}{I_{XY}(\tau) + I_{YX}(\tau)}, \tag{6}$$

where the $I_{XY}(\tau)$ and $I_{YX}(\tau)$ are called conditional mutual information and are defined as:

$$\begin{aligned} I_{XY}(\tau) &= I(X;Y|X_\tau) \\ &= H(X|X_\tau) + H(Y|X_\tau) - H(X,Y|X_\tau); \qquad (7) \\ I_{YX}(\tau) &= I(Y;X|Y_\tau) \\ &= H(Y|Y_\tau) + H(X|Y_\tau) - H(Y,X|Y_\tau), \qquad (8) \end{aligned}$$

with $X_\tau = X(t-\tau)$, $Y_\tau = Y(t-\tau)$ and $H(X|Y)$ being the conditional entropy [**?**].

The directionality index, DI_{XY}, then quantifies the *net* information flow. From the definition of DI_{XY}, Eq.(6), it is clear that $\mathrm{DI}_{XY} = -\mathrm{DI}_{YX}$. Also, $-1 \leq \mathrm{DI}_{XY} \leq 1$: $\mathrm{DI}_{XY} = 1$ if and only if $I_{XY} \neq 0$, $I_{YX} = 0$ (*i.e.*, the information flow is $X \to Y$ and there is no back coupling $Y \to X$) and $\mathrm{DI}_{XY} = -1$ if and only if $I_{XY} = 0$, $I_{YX} \neq 0$ (*i.e.*, the information flow is $Y \to X$ and there is no back coupling $X \to Y$).

Naturally, $\tau > 0$ is a parameter that has to be tuned appropriately to the time-scales relevant to the particular dataset.

2.5 Symbolic Ordinal Patterns

For the information theoretic measures (entropy, MI and DI) the PDF of the time series has to be approximated. This can be done in several ways, the most usual is by means of histograms of the data.

ClimNet provides two way to calculate the PDFs p_i, p_j and p_{ij}: by histograms of the original values (this case will be referred to as MIH) and by using a symbolic transformation, in terms of probabilities of ordinal patterns [**?**] (this case will be referred to as MIOP).

MIOP allows time scale selection, which can be used in any information theoretic measure. The procedure for this calculation is the following: Ordinal Patterns are calculated by pointing out the value of a data point relative to other values in the series. Using e.g. three symbols (letters) $3! = 6$ different patterns exist, for four symbols there will be $4! = 24$ patterns, and so forth – see Fig. 1 (a) and (b) respectively. The possibility of two equal adjacent values is not considered, partly due to the presence of noise in the time series.

As shown in Fig. 1 OPs of length 3 are formed by 3 symbols in the following way: if a value ($x_i(2)$) is higher than the previous one ($x_i(1)$) but lower than the next one ($x_i(3)$), it will yield the pattern '123' (Fig. 1 (1)), while the opposite case ($x_i(3) < x_i(2) < x_i(1)$) will give the pattern '321' (Fig. 1 (6)), etc. This symbolic transformation allows to detect correlations in the sequence of values which are not taken into account when using histograms of values as they do not consider the order in which the values appear in the time series. As a drawback, this

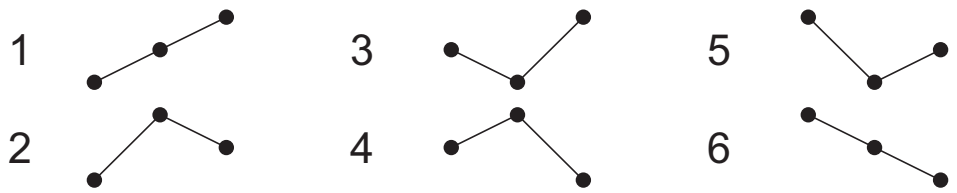

Figure 1: Ordinal patterns are computed from the comparison between (not necessarily adjacent) points on a time series. The number of possible patterns depend on the number of "letters" considered. In the general case they grow as $n!$. [**?**]

technique does not contain information about the relative magnitudes; this is usually useful as a natural robustness under low to moderate noise embedding.

After constructing time series of the OPs, histograms can be calculated from them. In this symbolic approach, the number of bins is naturally defined by the number of possible patterns, which in turn is determined by the number of symbols in the ordinal pattern. As explained above, if the OP word is of length n, there will be $n!$ possible patterns, and this will be the number of bins used for computing the probabilities associated with the symbolic sequences. This eliminates the binning problems which frequently appear when using histograms.

Ordinal patterns do not need to be constructed with immediately adjacent data points only. It is possible to construct them with data points that are separated in time, and in this way different time scales are considered. This symbolic transformation keeps the information about correlations present in a time series at the selected time scale, but does not keep information about the absolute values of the data points. Therefore, the e.g. Mutual Information computed from ordinal patterns (MIOP) can be expected to provide complementary information with respect to the standard method of computing the mutual information (MIH).

3 Statistical Significance

In any experiment or observation that involves drawing a sample from a population there is always the possibility that an observed effect would have occurred due to sampling error or chance alone. However if the probability of obtaining an extreme result – large difference between two or more sample means – given the null hypothesis is true, is less than a predetermined threshold (e.g. 5% chance), then it is possible to conclude that the observed effect is not due to chance[**?**].

To address the significance of the values, ClimNet uses bootstrap (BS) algorithm [**?**]. This algorithm randomly resamples with replacement from the original datasets using blocks of data of approximately the size of the autocorrelation time of the time series, and then computes the estimators (MI and DI) from the resampled data. Doing so, both the statistics (histogram) and the power spectrum of the original time series are approximately preserved. In this way, an empirical distribution can be obtained for each link, and significance thresholds can be extracted from them.

4 ClimNet Architecture

ClimNet is equipped with wide range of necessary functions to construct climate networks from raw time series. In line with the described measures in the previous section, the processing time series sequence carried out by ClimNet is shown in fig. 2. The core functionality of the software is to compute the adjacency matrix from time series. To do so, various options are possible: from the linear cross correlation, used mostly for benchmarking and for comparison to the other measures, through entropy (an univariate measure which estimates the amount of variability a certain region has) to Mutual information and the directionality index, which can be calculated both using histograms or using the more sophisticated method of ordinal analysis. Each link is repeated a number of times by using the bootstrap algorithm in order to assure the statistical significance of the result. Insignificant links are discarded, whereas significant ones are accepted and thus included in the resulted adjacency matrix.

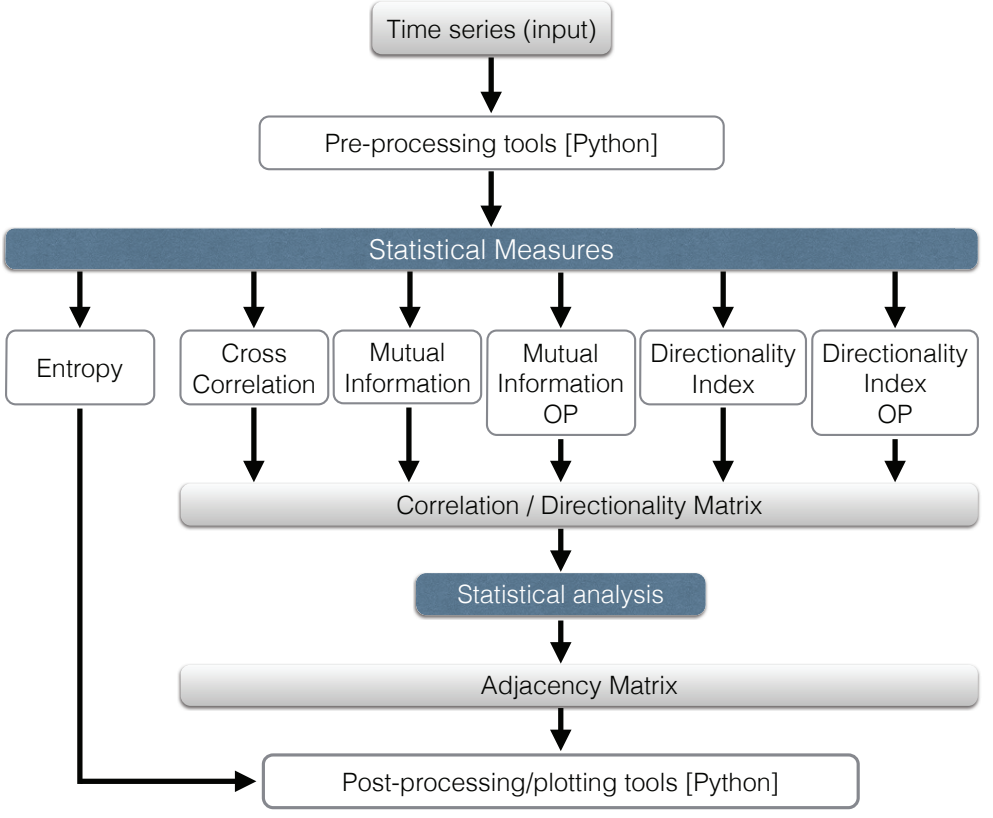

Figure 2: Software description. ClimNet provides an interface to read and pre-process climate time series. Afterwards, the user is given the choice to select the type of network based on a statistical measure(s). The resulting network (adjacency matrix) is subject to further analysis by graph analysis by common graph analysing algorithms/libraries.

Having produced the adjacency matrix (network), it can be processed consecutively by any graph analysing algorithm(s) in order to obtain more advanced measures as betweenness centrality and community detection. Accordingly, the general architecture of ClimNet is gown in fig. 3.

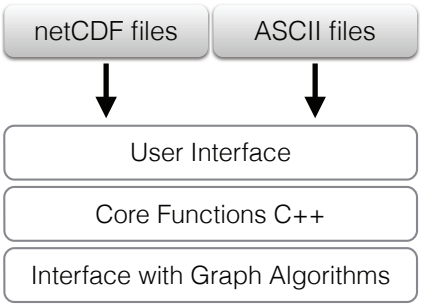

Figure 3: ClimNet architecture. The software core is implemented in C++ with python tools for pre-processing (removal of the seasonal cycle, or trends from the time series, normalisation, among others) and post-processing (calculating some basic network properties and graphing the network as a map in order to be interpreted by climatologists). ClimNet's interface is provided with routines to read netCDF files (self-describing, machine-independent data formats - very popular in climate science) and simple ASCII files.

5 Applications - Climate Networks' Examples

The software presented in this paper has been used in various climatological studies through its development. Some of the results obtained with these techniques is shown below. All the maps shown are statistically significant.

In fig. 4 Mutual information using ordinal patterns is shown for the SAT field. This technique allows to select the time scales of the phenomena directly, something harder to achieve using traditional methods. The figures (left) show networks for shorter time scales while in the right longer time scales are displayed. The top maps are of connectivity while the bottom maps are of connections to a given point (in the central Pacific ocean). Note that for longer time scales the connectivity is higher and there are longer connections than for shorter timescales. This is compatible with the current understanding of climatological patterns, dominated by inter annual time scales like ENSO.

Figure 5 show the comparison between the mutual information and the directionality index. Both calculated without ordinal patterns. MI show that the Pacific ocean is a highly clustered area for the SAT field while DI show that information is flowing from east to west in the equator, probably following the trade winds. All this information is inferred directly from the data and can be of big help for climatologists in order to analyse non linear relationships and the flow of the information hinting (but not proving) causal effects.

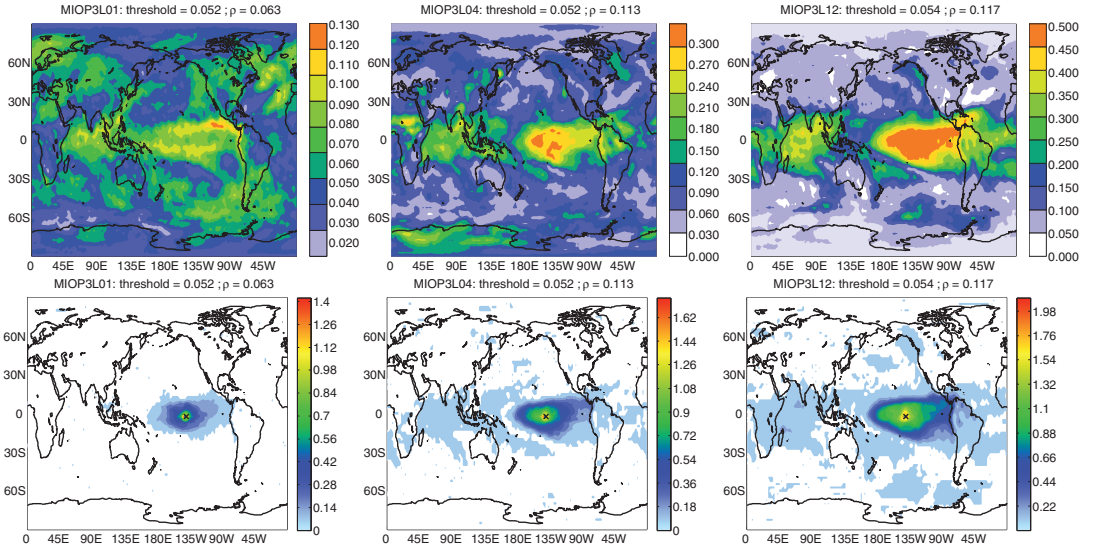

Figure 4: Example of Mutual information using ordinal patterns (MIOP). The Area weighted connectivity (equivalent to the degree of the network over a sphere) is shown in the top maps. Bottom maps show one column of the matrix showing the connections to the point shown with an **x**. First row: mostly time scales, second row, seasonal time scales, third row, yearly time scales [**?**].

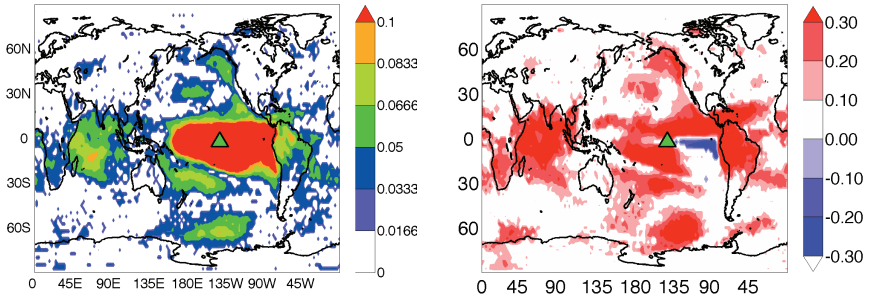

Figure 5: Columns of the Adjacency matrix, showing connections for a point in the central Pacific ocean (marked with a green triangle), for (left) Mutual information and (right) Directionality Index. Notice the similarities in the shape of the patterns and the new structure unveiled by DI showing the direction of the flow of information, not present in the figure on the left. [**?**]

6 Conclusions

The presented software tools in ClimNet allow the construction of relatively large climate networks from linear and non-linear statistical measures. ClimNet is equipped with easy-to-use user interface to facilitate the reading and the pre-processing of climate time series. The developed techniques to evaluate linear and the nonlinear relationships between variables at different timescales has been addressed in detail. Additionally, we have discussed examples on its application to real climate data, resulting in recent findings and original contributions in climate research. We believe ClimNet will help researchers in various complex sciences domains, and it is hoped that it will contribute to further findings in the investigation of complex systems as well as in climate research

References