Submitted to *Entropy*. Pages 1 - 11.

OPEN ACCESS

Article

Reliability of inference of directed climate networks using conditional mutual information

Jaroslav Hlinka [1,*], David Hartman [1], Martin Vejmelka [1], Jakob Runge [2,3], Norbert Marwan [2], Jürgen Kurths [2,3,4] and Milan Paluš [1]

[1] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodarenskou vezi 2, 182 07, Prague 8, Czech Republic

[2] Potsdam Institute for Climate Impact Research (PIK), 14473 Potsdam, Germany

[3] Department of Physics, Humboldt University, 12489 Berlin, Germany

[4] Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen AB24 3UE, United Kingdom

* Author to whom correspondence should be addressed; hlinka@cs.cas.cz, (+420) 266053808

Version January 21, 2013 submitted to Entropy. *Typeset by LaTeX using class file mdpi.cls*

Abstract: Across geosciences, many investigated phenomena relate to specific complex systems consisting of intricately intertwined interacting subsystems. Dynamical complex systems can be represented by a directed graph, where each link denotes an existence of a causal relation, or information exchange between the nodes. For geophysical systems such as global climate, these relations are commonly not known theoretically but estimated from recorded data using causality analysis methods. These include nonlinear methods based on information theory, as well as their linear counterparts. Importantly, the choice of causality analysis methods affects the reliability of the constructed networks and any further inference regarding existence of significant features or climate variability. We compare a range of methods and parameter settings with respect to the reliability of directed climate networks, using surface air temperature data from reanalysis of 60-year global climate records preprocessed by VARIMAX-rotated principal components analysis. Overall, causality methods provided reproducible estimates of climate causality networks, with the linear approximation outperforming the nonlinear methods. Interestingly, optimizing the nonlinear methods with respect to reliability has lead to improved similarity of the detected networks to those discovered by the linear approach, in line with the hypothesis of near-linearity of climate reanalysis data.

Keywords: causality; climate; nonlinearity, transfer entropy, network, stability

1. Introduction

Across geosciences, many investigated phenomena relate to specific complex systems consisting of intricately intertwined interacting subsystems.

These can be suitably represented as networks, an approach that is gaining increasing attention in complex systems community [1,2]. The meaning of the existence of a link between nodes of a network depends on the area of application, but in many cases it is related to some form of information exchange between the nodes.

This approach has already been adopted for the analysis of various phenomena in the global climate system [3**? ?** –5]. Typically, a graph is constructed by considering two locations linked by a connection, if there is an instantaneous dependence between the localized values of a variable of interest.

This dependence can be conveniently quantified by mutual information - an entropy-based general measure of statistical dependence that takes into account nonlinear contributions to the coupling. In practice, for reasons of theoretical and numerical simplicity, linear Pearson's correlation coefficient might be sufficient. In particular, while initial works by Donges et al. stressed the role of mutual information in detecting important features of global climate networks [6,7], more detailed recent work has shown that the differences between correlation and mutual information graphs are mostly spurious, such as due to trivial or erroneous-data-related nonstationarities of the data [8].

However, these methods do not allow to assess the directionality of the links and of the underlying information flow. This motivates the use of more sophisticated measures, known also as causality analysis methods.

The family of causality methods include linear approaches such as the Granger causality analysis [9] as well as more general nonlinear methods. A prominent representative of nonlinear causality assessment method is the conditional mutual information [10] known also as transfer entropy [11].

Arguably, the nonlinear methods, due to their model-free nature, have the theoretical advantage of being sensitive to forms of interactions that linear methods may detect only partially or not at all. On the other side, this advantage might be more than outweighed by a potentially lower precision. Depending on specific circumstances, this may adversely affect the reliability of detection of network patterns.

Apart from uncertainty about the general network pattern, reliability is important when the interest is in detecting changes in time, with the need to distinguish them from random variance of the estimates among different sections of time series under investigation - a task that is relevant in many areas of geoscience including climate research.

In other words, before analyzing the a complex dynamical system using network theory, a key initial question is that of the reliability of the network construction, and of its dependence on the causality method choice and settings.

We study this question for a selection of standard causality methods, using a timely application in the study of climate network and its variability. In particular, surface air temperature data from the NCEP/NCAR reanalysis dataset [12,13] are used. The original data contain more than 10,000 time

series - a relatively dense grid covering the whole globe. For efficient computation and visualization of the results, it is convenient to reduce the dimensionality of the data. We use principal component analysis and select only components that have significantly high explained variance compared to corresponding colored random noise.

As the causality network construction reliability may crucially depend on the specific choice of the causality estimator, we quantitatively assess the effect of choice of different causality measures and their parametrization.

We assess the network construction reliability by quantifying the similarity of causality matrices reconstructed from independent realization of a stationary model of data. These realizations are either independently generated, or they represent individual non-overlapping temporal windows in a single stationary realization. Optimal parameter choice of the applied nonlinear methods is detected, and the reliability of networks constructed using linear and nonlinear methods compared.

The latter method, i.e. comparing networks reconstructed from temporal windows, allows to assess the network variability on real data and compare it with variability on the stationary model.

2. Data and Methods

2.1. Causality assessment methods

2.1.1. Granger causality analysis

A prominent method for assessing causality is so-called Granger causality analysis, named after Sir Clive Granger, who proposed this approach to time series analysis in a classical paper [9]. However, the basic idea can be traced back to Wiener [14], who proposed that if the prediction of one time series can be improved by incorporating the knowledge of a second time series, then the latter can be said to have a causal influence on the former. This idea was formalized by Granger in the context of linear regression models. In the following, we outline the methods of assessment of Granger causality, following the description given in [15] and [16,17].

Consider two stochastic processes X_t and Y_t and assume they are jointly stationary. Let further the autoregressive representations of each process be:

$$X_t = \sum_{j=1}^{\infty} a_{1j} X_{t-j} + \epsilon_{1t}, \qquad \mathrm{var}(\epsilon_{1t}) = \Sigma_1, \tag{1}$$

$$Y_t = \sum_{j=1}^{\infty} d_{1j} Y_{t-j} + \eta_{1t}, \qquad \mathrm{var}(\eta_{1t}) = \Gamma_1, \tag{2}$$

and the joint autoregressive representation be:

$$X_t = \sum_{j=1}^{\infty} a_{2j} X_{t-j} + \sum_{j=1}^{\infty} b_{2j} Y_{t-j} + \epsilon_{2t}, \tag{3}$$

$$Y_t = \sum_{j=1}^{\infty} c_{2j} X_{t-j} + \sum_{j=1}^{\infty} d_{2j} Y_{t-j} + \eta_{2t}, \tag{4}$$

where the covariance matrix of the noise terms is:

$$\Sigma = \mathrm{Cov} \begin{pmatrix} \epsilon_{2t} \\ \eta_{2t} \end{pmatrix} = \begin{pmatrix} \Sigma_2 & \Lambda_2 \\ \Lambda_2 & \Gamma_2 \end{pmatrix}. \tag{5}$$

The causal influence from Y to X is then quantified based on the decrease in the residual model variance when we include the past of Y in the model of X, i.e. when we move from the independent model given by Equation 1 to the joint model given by Equation 3:

$$F_{Y \to X} = \ln \frac{\Sigma_1}{\Sigma_2}. \tag{6}$$

Similarly, the causal influence from X to Y is defined as:

$$F_{X \to Y} = \ln \frac{\Gamma_1}{\Gamma_2}. \tag{7}$$

Clearly, the causal influence defined in this way is always nonnegative.

2.2. Estimation of GC

Practical estimation of the Granger causality involves fitting the full and depleted models described above to experimental data. While the theoretical framework outlined above is formulated in terms of infinite sums, the fitting procedure requires selection of the model order p for the models. For our report, we have selected $p = 1$ to allow direct comparability of the Granger causality analysis to the nonlinear methods considered later. This choice is the most common choice for Granger causality in literature and amounts to looking for links with lag 1 time unit.

2.3. Transfer entropy

To provide a framework for discussion of the related issues, it is useful to consider that for a general bivariate stochastic process the Granger causality concept, can be captured in information-theoretic terms. In particular, we can define that X causes Y if the knowledge of past of X decreases the uncertainty about Y (above what the knowledge of past of Y and potentially all other relevant confounding variables already informs). This simple concept is captured in the definition of *transfer entropy* (TE, [11]). TE as can be defined in terms of *conditional mutual information* as shown below, following closely [10].

For two discrete random variables X, Y with sets of values Ξ and Υ and probability distribution functions (PDFs) $p(x), p(y)$ and joint PDF $p(x, y)$, the Shannon entropy $H(X)$ is defined as

$$H(X) = -\sum_{x \in \Xi} p(x) \log p(x), \tag{8}$$

and the joint entropy $H(X, Y)$ of X and Y as

$$H(X, Y) = -\sum_{x \in \Xi} \sum_{y \in \Upsilon} p(x, y) \log p(x, y). \tag{9}$$

The conditional entropy $H(X|Y)$ of X given Y is

$$H(X|Y) = -\sum_{x \in \Xi} \sum_{y \in \Upsilon} p(x, y) \log p(x|y). \tag{10}$$

The amount of common information contained in the variables X and Y is quantified by the mutual information $I(X; Y)$ defined as

$$I(X; Y) = H(X) + H(Y) - H(X, Y). \tag{11}$$

The conditional mutual information $I(X; Y|Z)$ of the variables X, Y given the variable Z is given as

$$I(X; Y|Z) = H(X|Z) + H(Y|Z) - H(X, Y|Z). \tag{12}$$

Entropy and mutual information are measured in bits if the base of the logarithms in their definitions is 2. It is straightforward to extend these definitions to more variables, and to continuous rather than discrete variables.

Transfer entropy from process X_t to process Y_t then corresponds to the conditional mutual information between X_t and Y_{t+1} conditional on Y_t:

$$T_{X \to Y} = I(X_t, Y_{t+1}|Y_t). \tag{13}$$

While the definition of these information-theoretic functionals describing dependence structure between variables is very general and elegant, the practical estimation faces challenges related to the problem of efficient estimation of the PDF of the studied variables from samples of finite size. For the further considerations, it is important to bear in mind the distinction between the true quantities of the underlying stochastic process, and their finite-sample estimators.

2.4. Potential causes of observed difference

Interestingly, it can be shown that for linear Gaussian processes, transfer entropy is equivalent to linear Granger causality, up to a multiplicative factor [18]:

$$\mathcal{T}_{X \to Y} = \frac{1}{2} \mathcal{F}_{X \to Y}. \tag{14}$$

However, in practice, the estimates of transfer entropy and linear Granger causality may differ. There are principally two main reasons for this divergence between the results. Firstly, when the

underlying process is not linear Gaussian, the true transfer entropy may differ from the true linear Granger causality corresponding to the linear approximation of the process. However, there is a second reason for divergence between sample estimates of transfer entropy and linear Granger causality, valid even for linear Gaussian processes. This is the difference in the properties of the estimators of these two quantities. Typically, the estimators of the transfer entropy pay for their generality by some bias and higher variance of the estimates.

2.5. TE estimation

As mentioned in the Introduction, it is important to bear in mind the distinction between the true underlying characteristics of the multivariate process, and their finite sample estimates. There are many algorithms for estimation of information-theoretical functionals, that can be adapted to compute transfer entropy estimates. We focus on two of them which have been tested and applied on real-world data. The first is an algorithm based on discretization of studied variables into Q equiquantal bins (EQQ, [19]) and the second is a k-nearest neighbor (kNN, [10]) algorithm.

Both these algorithms require setting an additional parameter. While some heuristic suggestions have been published in the literature, the suitable values of the parameters may depend on specific aspects of the application including the character of the time series. For the purpose of this study, we use a range of parameter values and subsequently select the parameter values providing the most stable results for further comparison with linear methods, see below.

2.6. Data

2.6.1. Dataset

Data from the NCEP/NCAR reanalysis dataset [12] have been used. In particular, we utilize the time series $x_i(t)$ of the monthly mean SAT from January 1948 to December 2007 ($T = 720$ time points), sampled at latitudes λ_i and longitude ϕ_i forming a regular grid with a step of $\Delta\lambda = \Delta\phi = 2.5°$. The points located at the globe poles have been removed, giving a total of $N = 10224$ spatial sampling points.

2.6.2. Preprocessing

To minimize the bias introduced by periodic changes in the solar input, the mean annual cycle is removed from the data to produce so-called anomaly time series. The data were further standardized so that the time series of each point has unit variance. The time series are then scaled by the cosine of the latitude to account for grid points closer to the poles representing smaller areas and being closer together (thus biasing the correlation with respect to grid points farther apart). The poles are thus omitted entirely by effectively removing data for latitude ± 90.

2.6.3. Computing the components

The covariance matrix of the scaled time series obtained by preprocessing is computed. Note that this covariance matrix is equal to the correlation matrix, where each correlation is scaled by the inverse of the product of the cosines of the latitudes of the time series entering the correlation.

Next, the eigendecomposition of the covariance matrix is computed. The eigenvectors corresponding to genuine components are extracted (estimation of the number of components is explained in the next paragraph). The eigenvectors are then rotated using the VARIMAX method [20].

The rotated eigenvectors are the resulting components.

2.6.4. Estimating the dimensionality of the data

To reduce the dimensionality, only a subset of the components is selected for further analysis. The main idea rests in determining significant components by comparing the eigenvalues computed from the original data to eigenvalues computed from a surrogate model. The surrogate model consists of autoregressive (AR) models fit to each time series independently and its dimension is estimated using the Bayesian Information Criterion. See [**?**] for more detail. This approach has lead to the selection of 67 components.

2.6.5. Model

To provide a stationary model of the potentially non-stationary data, a so-called surrogate data was constructed.

Technically, linear surrogate data are conveniently constructed as multivariate Fourier transform (FT) surrogates [21,22]; i.e. obtained by computing the Fourier transform of the series, keeping unchanged the magnitudes of the Fourier coefficients (the amplitude spectrum), but adding the same random number to the phases of coefficients of the same frequency bin; the inverse FT into the time domain is then performed.

The surrogate data represent a realization of a linear stationary process conserving the linear structure (covariance and autocovariance) of the original data, and hence also the linear component of causality.

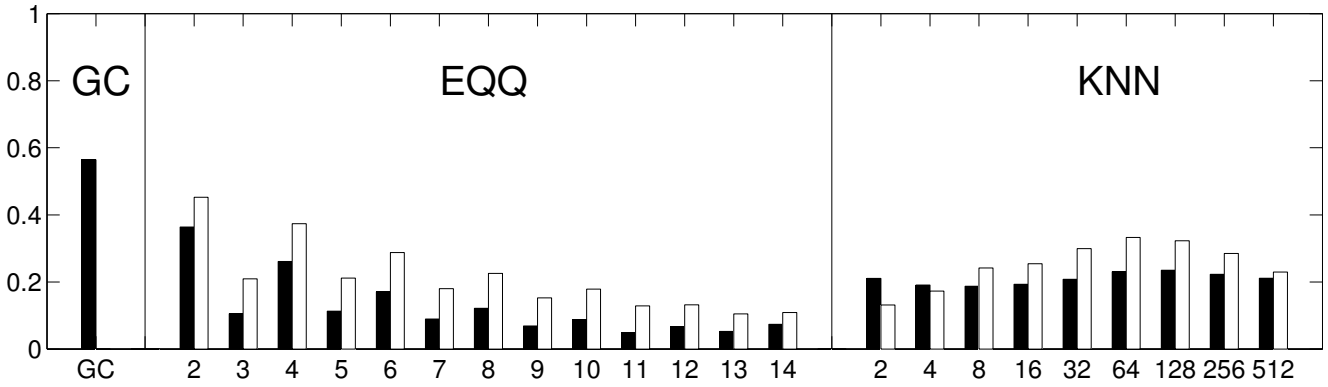
2.7. Network analysis

Both the stationary model and real data time series were split into 6 windows (one for each decade, i.e. with approximately 3650 time points). For each of the windows, causality matrix has been computed with several causality methods.

In particular, we have used pairwise Granger causality as a representative linear method, and conditional mutual information (transfer entropy) computed by two standard algorithms, using a range of critical parameter values. The first is an algorithm based on discretization of studied variables into Q equiquantal bins (EQQ, [19], $Q \in \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14\}$) and the second is a k-nearest neighbor algorithm (kNN, [10], $k \in \{2, 4, 8, 16, 32, 64, 128, 256, 512\}$).

Each of these algorithms provides a matrix of causality estimates among the 67 climate components within the respective decade. We further assess the similarity of these matrices both across time and

Figure 1. Reliability of causality network detection using different causality estimators, and the similarity to linear causality network estimates. For each estimator, six causality networks are estimated, one for each decade of modeled stationary data. Black: the height of the bar corresponds to the average Spearman correlation across all 15 pairs of decades. White: the height of the bar corresponds to the average Spearman correlation of nonlinear causality network and linear causality network across 6 decades.



methods; first in stationary data (where temporal variability is attributable to method instability only) and subsequently in real data.

Apart from direct visualization, the similarity of constructed causality matrices is quantified by the Spearman's rank correlation coefficient of on off-diagonal entries. The reliability is then estimated as the average Spearman's rank correlation coefficient across all $(6*5)/2 = 15$ pairs of temporal windows.

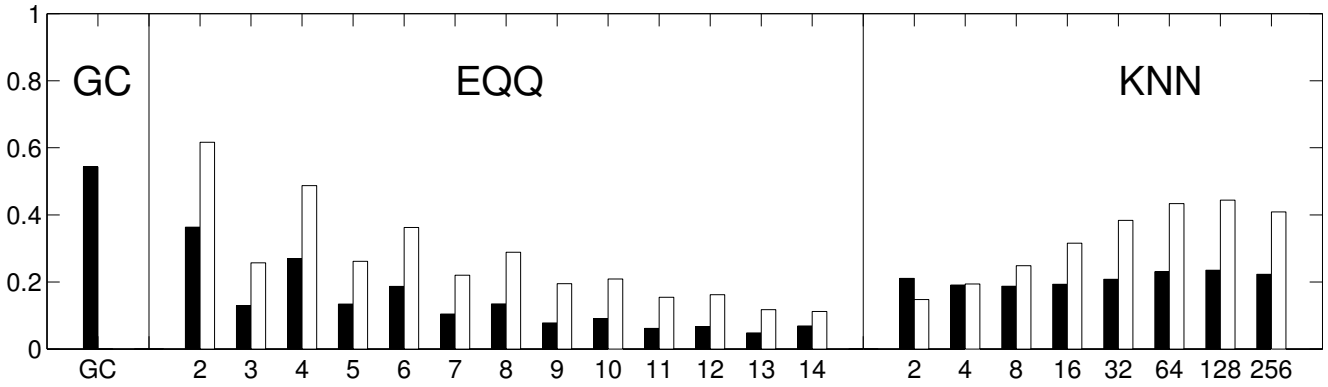
3. Results and Discussion

The reliability of causality networks for all methods and settings is shown in Figure 1. The linear Granger causality shows the highest reliability, with the average Spearman's rank coefficient ~ 0.6. The equiquantal binning method provided most reliable network estimates for $Q = 2$ ($\bar{r} \sim 0.36$), with reliability generally decreasing for increasing Q. The k-nearest neighbors algorithm provided even less reliable network estimates, with only weak dependence on the values of the k-parameter and optimum reliability of $\bar{r} \sim 0.33$ for $k = 64$.

The causality networks constructed by each nonlinear method have been compared to the causality network obtained using the linear Granger causality analysis, see white bars Figure 1. In general, the nonlinear causality networks have shown higher similarity to linear estimates than to nonlinear estimates for different section of the stationary model time series. Interestingly, the parameter settings that optimized the reliability also provided the (almost) closest results to the linear methods.

Figure 2 shows the results of an analogous analysis on original data rather than the stationary model. Note that here the computed causality network similarities reflect a combination of (lack of) reliability of the methods and real variability in the dynamical properties of the time series across time (i.e. true changes in the causality pattern). The results are both qualitatively and quantitatively similar to those shown in Figure 1, suggesting that the true variability of the causal networks on this time scale is likely rather small compared to the coarseness of the causality assessment methods.

Figure 2. Variability of causality network detection using different causality estimators, and the similarity to linear causality network estimates. For each estimator, six causality networks are estimated, one for each decade of the data. Black: the height of the bar corresponds to the average Spearman correlation across all 15 pairs of decades. White: the height of the bar corresponds to the average Spearman correlation of nonlinear causality network and linear causality network across 6 decades.



4. Discussion

The outlined pairwise causality estimators suffer from inherent limitations. To give an example, a system consisting of three processes X, Y, Z, where Z drives both X and Y, but with different temporal lags, may erroneously show causal influence between X and Y even if these were not directly coupled. To deal with such situations, the concepts can be generalized to allow to take into account the variance explained by third variable(s).

In practice, estimation of these generalized causality patterns from relatively short time series is technically challenging, particularly in the context of nonlinear, information-theory based causality measures, due to the exponentially increasing dimension of probability distributions to be estimated. However, recent work has provided promising approaches to tackle this problem by sidestepping the curse of dimensionality by heuristically selecting a subset of candidate variables to be taken into account [23].

For completeness we mention that apart from the time-domain treatment of causality, the whole problem can also be reformulated in the spectral domain, leading to frequency-resolved causality indices such as partial directed coherence (PDC, [24]) or Directed Transfer Function (DTC, [25]).

5. Conclusions

Meaningful interpretation of climate networks and their observed temporal variability requires knowledge and minimization of the methodological limitations of the methods of their construction. In the presented work, we discussed the problem of reliability of network construction from time series of finite length, quantitatively assessing the reliability for a selection of standard causality methods. These included two major algorithms for estimating transfer entropy with a wide range of parameter choices, as well as the linear Granger causality analysis, which can be understood as linear approximation to transfer entropy. Overall, causality methods provided reproducible estimates of climate causality networks, with

the linear approximations outperforming the nonlinear methods. Interestingly, optimizing the nonlinear methods with respect to reliability has lead to improved similarity of the detected networks to those discovered by linear methods, in line with the hypothesis of near-linearity of climate reanalysis data.

Acknowledgements

This study is supported by the Czech Science Foundation, Project No. P103/11/J068 and by the DFG grant No. KU34-1.

References

1. Newman, M.E.J. The structure and function of complex networks. *SIAM Review* **2003**, *45*, 167–256.
2. Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M.; Hwang, D.U. Complex networks: Structure and dynamics. *Physics Reports* **2006**, *424*, 175–308.
3. Tsonis, A.; Roebber, P. The architecture of the climate network. *Physica A* **2004**, *333*, 497–504.
4. Tsonis, A.A.; Swanson, K.L.; Roebber, P.J. What do networks have to do with climate? *Bulletin of the American Meteorological Society* **2006**, *87*, 585+.
5. Yamasaki, K.; Gozolchiani, A.; Havlin, S. Climate networks around the globe are significantly affected by El Nino. *Physical Review Letters* **2008**, *100*.
6. Donges, J.F.; Zou, Y.; Marwan, N.; Kurths, J. The backbone of the climate network. *EPL* **2009**, *87*, 48007.
7. Donges, J.F.; Zou, Y.; Marwan, N.; Kurths, J. Complex networks in climate dynamics. *European Physical Journal* **2009**, *174*, 157–179.
8. Hlinka, J.; Hartman, D.; Vejmelka, M.; Novotna, D.; Paluš, M. Non-linear dependence and teleconnections in climate data: sources, relevance, nonstationarity. *submitted*.
9. Granger, C.W. Investigating causal relations by econometric model and cross-spectral methods. *Econometrica* **1969**, *37*, 414–&.
10. Vejmelka, M.; Palus, M. Inferring the directionality of coupling with conditional mutual information. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)* **2008**, *77*.
11. Schreiber, T. Measuring information transfer. *Physical Review Letters* **2000**, *85*, 461–464.
12. Kistler, R.; Kalnay, E.; Collins, W.; Saha, S.; White, G.; Woollen, J.; Chelliah, M.; Ebisuzaki, W.; Kanamitsu, M.; Kousky, V.; van den Dool, H.; Jenne, R.; Fiorino, M. The NCEP-NCAR 50-year reanalysis: Monthly means CD-ROM and documentation. *Bulletin of the American Meteorological Society* **2001**, *82*, 247–267.
13. Kalnay, E.; Kanamitsu, M.; Kistler, R.; Collins, W.; Deaven, D.; Gandin, L.; Iredell, M.; Saha, S.; White, G.; Woollen, J.; Zhu, Y.; Chelliah, M.; Ebisuzaki, W.; Higgins, W.; Janowiak, J.; Mo, K.; Ropelewski, C.; Wang, J.; Leetmaa, A.; Reynolds, R.; Jenne, R.; Joseph, D. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society* **1996**, *77*, 437–471.
14. Wiener, N., Modern Mathematics for Engineers; McGraw-Hill, New York, 1956; chapter The theory of prediction, pp. 165 – 190.

15. Ding, M.; Chen, Y.; Bressler, S.L., Granger Causality: Basic Theory and Application to Neuroscience. In *Handbook of Time Series Analysis*; Wiley-VCH Verlag GmbH & Co. KGaA, 2006; pp. 437–460.
16. Geweke, J.F. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association* **1982**, *77*, 77.
17. Geweke, J.F. Measures of Conditional Linear Dependence and Feedback Between Time Series. *Journal of the American Statistical Association* **1984**, *79*, 907–915.
18. Barnett, L.; Barrett, A.B.; Seth, A.K. Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables. *Physical Review Letters* **2009**, *103*.
19. Palus, M.; Albrecht, V.; Dvorak, I. Information theoretic test for nonlinearity in time series. *Physics Letters A* **1993**, *175*, 203–209.
20. Kaiser, H. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **1958**, *23*, 187–200.
21. Prichard, D.; Theiler, J. Generating surrogate data for time series with several simultaneously measured variables. *Physical Review Letters* **1994**, *73*, 951.
22. Palus, M. Detecting phase synchronization in noisy systems. *Physics Letters A* **1997**, *235*, 341–351.
23. Runge, J.; Heitzig, J.; Petoukhov, V.; Kurths, J. Escaping the Curse of Dimensionality in Estimating Multivariate Transfer Entropy. *Physical Review Letters* **2012**, *108*.
24. Baccala, L.A.; Sameshima, K. Partial directed coherence: a new concept in neural structure determination. *Biological Cybernetics* **2001**, *84*, 463–474.
25. Kaminski, M., D.M.T.W.A..B.S.L. Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biol Cybern* **2001**, *85*, 145–157.

© January 21, 2013 by the authors; submitted to *Entropy* for possible open access publication under the terms and conditions of the Creative Commons Attribution license http://creativecommons.org/licenses/by/3.0/.