Evaluation of Data-Driven Causality Discovery Approaches among Dominant Climate Modes

CyberTraining: Big Data + High-Performance Computing + Atmospheric Sciences

Steve Hussung[1], Suhail Mahmud[2], Akila Sampath[3], Mengxi Wu[4],
Research Assistant: Pei Guo[5] Faculty Mentor: Jianwu Wang[5]

[1]Department of Mathematics, Indiana University, Bloomington
[2]Computational Science Department, University of Texas at El Paso
[3]Department of Atmospheric Sciences, University of Alaska, Fairbanks
[4]Department of Earth, Environmental and Planetary Sciences, Brown University

[5]Department of Information Systems, UMBC

Technical Report HPCF–2019–12, hpcf.umbc.edu > Publications

Abstract

Identification of causal networks in atmospheric teleconnection patterns has applications in many climate studies. We evaluate and compare three data-driven causal discovery methods in locating and linking causation of well-known climatic oscillations. Four climate variables in the ERA-Interim reanalysis data (1979-2018) were examined in the study. We first employ dimension reduction to derive the the time-series for selected climate variables. Then time-series of dominant modes were processed using three different causal discovery methods: Granger causality discovery, Convergent cross-mapping (CCM), and PCMCI. Discovered causal links were different for different methods as well as for different variables. However, slightly similar causal links were observed between the Granger causality and CCM methods. Comparison of these three methods is discussed based on the El Niño-Southern Oscillation (ENSO) and its connection with other oscillations. Causal discovery methods were able to capture the linkage between the ENSO, North Atlantic Oscillation (NAO), and Pacific Decadal Oscillation (PDO), for some of the variables. Overall, this study identifies the usage of these statistical models in locating the direct and indirect causal links among the oscillations. Application of these data-driven causal discovery methods, both in terms of mediation and direct relationships between the observed teleconnection patterns, suggests that the data-driven statistical methods are efficient in locating the regimes of climate patterns and their 12 observed real connections to some extent. We present and provide our explanation of the evaluation results for each of the three causal discovery methods.

Contents

1 Introduction

The study of causality has recently emerged as a primary research tool in many areas of science [13]. The challenge is to reconstruct causal links along with their time lags using complex dynamical systems and connect them to the physical processes represented. Reconstructed causal graphs provide a new tool for identifying teleconnection patterns from spatio-temporal climate data [5]. In addition, many previous works suggest that temporal changes in the network's architecture and its graphical representations are widely used to identify signatures of global change due to the El Niño activity [8]. Here we introduce the basic ideas of causal connection, as well as some applications of statistical causality models to the climate system. We observed that the statistical causal discovery algorithms described in this paper are easy to implement.

Atmospheric teleconnection patterns influence the global climate through the atmospheric circulation [11]. Especially, the El Niño–Southern Oscillation (ENSO) has a significant effect on humanity's economic, social, and physical well-being. Many studies have been devoted to predict the extreme weather conditions of different regions of the globe in association with the ENSO [2, 6]. Unique and prominent circulation patterns occur in the atmosphere due to large scale ENSO variability [23], although mediation of other atmospheric teleconnection patterns and their causal connections are not well explained. Applying the graphical causal discovery model to the climate patterns could potentially discover the causal pathways that are responsible for extreme climate events of specific regions [21]. Dimension reduction is one important component in causal discovery in the case of large climate data sets [26].

In general, global climate model data reveals large spatial and temporal patterns that are relevant to multiple climate events. It is essential to isolate well-known climate patterns to study the specific climate events associated with them. The empirical orthogonal function decomposition (EOF), also called Principal Component Analysis, separates the data based on the orthogonal basis function computed from the climate data. The EOF method produces spatial modes in addition to the temporal patterns. Especially, the rotated empirical orthogonal function (REOF) approach is considered to be the most useful technique in capturing the spatial variabilities of the complex dynamical system that are more clear, well defined, and amenable to straightforward interpretation [11].

The purpose of this study is to explain how the climate system exhibits aspects of causal networks, with dominant modes corresponding to major teleconnection patterns. This study focused on identifying causal links between sea surface temperature (SST), 2-meter temperature (T2M), 10-meter wind speed (SI10), and mean sea level pressure (MSL). The time series of well-separated patterns associated with these variables were used as our primary data for the causal discovery algorithms. We will discuss the results from the perspective of causal discovery methods and their ability to capture the variabilities related to those variables. We mainly intend to explain the causal connection between the ENSO mode and other prominent climate variabilities using the Granger causality [9], Convergent cross Mapping (CCM) [27], and PCMCI [22] approaches. Reliability and limitation in the individual techniques should be addressed to some extent in the future, but are out of scope of this study.

2 Data and Methods

2.1 ERA-Interim Reanalysis

In this study, we use ERA-Interim global atmospheric reanalysis data [4] from 1979-2018. The 1.5 ERA-Interim modeling system includes 4-dimensional variational analysis (4D-Var) with a running Integrated Forecast System (IFS) model cycle of 31R2. The horizontal resolution of the data is T255 spectral (equivalent to about 80 km horizontal resolution) on 60 vertical levels. We obtained SST, MSL, SI10 and T2m from the monthly means of daily means fields from [10]. The analysis was focused on the monthly data spanning 40 years. For the causality study we use monthly time series that best represent dominant modes of larger variances, calculated for all years.

2.2 Empirical Orthogonal Function (EOF)

We preprocessed climate data variables to retain the patterns that are inherent to the large scale variabilities and remove the patterns corresponding to regional climate change. The first step in preprocessing is to remove the annual cycles from the monthly data. The second step is to detrend the data at every grid cell. The empirical orthogonal functions have been calculated for preprocessed monthly anomalies. In this study, we perform a varimax rotation of calculated EOFs using the first 30 modes [18]. The conventional EOF analysis produces spatial patterns (EOFs) and time series (PCs) that are both orthogonal. The Kaiser varimax rotation criterion [12] is being applied on each spatial mode to derive a simple structure, which contains a localized center with maximized variability. The rotated EOFs are orthonormal [19]. We chose six EOFs out of the 30 rotated EOFs based on their percentage of variance as well as their meteorological significance. These six selected EOFs and their associated principal components were used in further causal discovery analyses.

3 Granger Causality Method

3.1 Description of Method

Granger causality [9] was proposed by Nobel Laureate Clive W. Granger in 1969 and developed as a predictive model in economics. Granger causality is defined as follows: given two time series variables x and y, we say that x Granger causes y if the regression for y based on past values of both x and y is a statistically significant improvement on the regression of y only based on past values of y. Let the set of lagged variables of x be written x_t^P for t from 1 to P, the maximum time lag. Similarly, let the lagged variables of y be notated y_t^P. To evaluate Granger causality we generate the following regressions:

$$y_t = A_1 \cdot y^P + \varepsilon_1 \tag{3.1}$$

$$y_t = A_2 \cdot y^P + B_2 \cdot x^P + \varepsilon_2 \tag{3.2}$$

and consider whether or not the regression function in Equation 3.1 is better than Equation 3.2 in predicting y. A statistical hypothesis test method such as F-test or Chi-squared (χ^2) test is used to decide which regression is better by getting a p-value of the test and checking to see if this value is low enough to indicate statistical significance.

The F-test is applied in regression problems to decide whether a model fits the data significantly better than a naive model. In the case of Granger causality, model 3.1 is the naive model, model 1, and model 3.2 is model 2. We write the residual sums of squares RSS_1 and RSS_2 and the number of parameters p_1 and p_2, respectively. Let n be the number of total data samples. The null hypothesis is that model 2 does not fit data better than model 1. We calculate the F statistics of these two regression models via

$$F = \frac{\frac{RSS_1 - RSS_2}{p_2 - p_1}}{\frac{RSS_2}{n - p_2}} \tag{3.3}$$

Then we look up the F statistics in F-distribution with its corresponding degree of freedom to get the p-value of the null hypothesis and reject the null hypothesis if appropriate: The null hypothesis is rejected only if the p-value is smaller than the given significance level. For instance, if we set the level of significance of p-value as 0.05, and get a p-value from F-test as 0.01, it means that the probability of the statement "model 2 does not fit data better than model 1" being true is only 1%. Since 1% is less than 5%, we believe that model 2 is better, which indicates that x Granger causes y.

In conclusion, the Granger causality test includes two important parts: the regression models and the hypothesis test. The first regression model is a naive model that assumes the time series is only predicted by its own lagged variables, and the second regression model considers lagged variables of other time series as well. Hypothesis testing is used to compare the two models to decide which is better. To support causality discovery among multiple variables, the original model was extended to a graph model so that it could measure whether and how one variable is caused by multiple other variables [28, 1].

3.2 Implementation

The *statsmodels* module [15] is a Python [17] module that provides classes and functions for the estimation of many different statistical models. The Vector Autoregressions *tsa.vector_ar* package [16] from *statsmodels 0.9.0* was our primary software for our Granger causality approach.

The input data of the Granger causality model is the EOF modes of our time-series data. The most significant modes are chosen to generate the causality plots. Based on the team's understanding of atmospheric science, the maximum time lag in the Granger causality model is set to be 3 months.

The pairwise regression models are fitted and their RSSs are used to generate the F-test. We use a p-value significance level of 0.05 for the F-test. The p-value threshold of 0.05 ensures that our tests reject the null hypothesis at 95% probability. By applying this p-value threshold, the causality relationship pairs number is still large, making human interpretation

difficult or impossible. So we set another threshold that only includes the 20 causal links with lowest p-values, since these are the most significant ones.

4 Convergent Cross Mapping Method

4.1 Description of Method

Convergent cross mapping (CCM) is a causality analysis method designed for weakly or moderately coupled dynamic systems. It was introduced in [24] and extended in [27]. The fundamental theory for this method is phase space reconstruction for a dynamic system and Takens' Theorem [25]. The evolution of a dynamic system can be represented by a certain phase orbit in the phase space, and any two nearby phase points will remain close to each other for at least a short amount of time before diverging. The definition of causality in CCM and the complete algorithm is described below.

The phase space of the dynamic system can be reconstructed from discrete observations using the time delay embedding method. Assume there is a regular time series for the variable y: y_{t0}, y_{t0+1}, y_{t0+2}, ... If this system can be described by a set of n-th order ordinary differential equations, then a vector $Y = (y, \frac{dy}{dt}, \frac{d^2y}{dt^2}, ..., \frac{d^ny}{dt^n})$ will contain all the information of the system at each moment following $\frac{dY}{dt} = f(Y)$, where f is a given function. We can approximate derivatives using consecutive observations, i.e., $(y_t, \frac{dy}{dt}|_t, \frac{d^2y}{dt^2}|_t, ..., \frac{d^ny}{dt^n}|_t) \approx A(y_t, y_{t-\tau}, y_{t-2\tau}, ..., y_{t-n\tau})$, where A is a constant matrix. Thus, the phase space can be reconstructed as the vector space for $Y_t = (y_t, y_{t-\tau}, y_{t-2\tau}, ..., y_{t-n\tau})$.

If Y is only influenced by one external variable x, i.e. $\frac{dY}{dt} = f(x, Y)$, then neighboring phase points of Y are also associated with similar x. Therefore, we can estimate the historic values of x by averaging x over a group of known neighboring phase points of Y. The causal effect from x to y is defined and measured by the capability of estimating historic x from Y using this cross mapping method.

Here are the steps of extended CCM [24, 27]:

(1) Reconstruct the phase space for Y with specified parameters n and τ: $Y_t = (y_t, y_{t-\tau}, y_{t-2\tau}, ..., y_{t-n\tau})$.

(2) For each time t, we will find $n+1$ nearest neighbors of Y_t in the reconstructed phase space at time t_1, t_2, ..., t_{n+1}, and their corresponding x at the same time or earlier.

(3) The historic x at time t is estimated as a weighted average of x at time t_1, t_2, ..., t_{n+1}. The weight is an exponentially decaying function of the Euclidean distance in the Y space.

(4) The skill of cross mapping is measured by the correlation coefficient between true x at time t and the estimates. For a better comparison with the other two causality methods, we also compute the p-values for the correlation coefficient assuming the residuals from a linear model follow independent and identical normal distributions.

(5) Because x does not have to be concurrent with Y, the steps above can be repeated for different lags between x and Y to determine the optimal time lag of the causal relationship. Particularly, a causality is unreal if the optimal time lag requires x to occur later than the time t. If the optimal time lag is 0, then it means x and Y are synchronized and it is hard

to tell the true causal direction between these two variables [24]. We choose to dispose these uncertain causal relations in our study, but we are aware that this conservative approach may miss some quick processes and cause a decrease in the number of identified causal relations.

4.2 Implementation

Following the detailed description of the algorithm in [24, 27], we independently developed a Python [17] program to compute the CCM causality for pairs of time series. In our analysis, we set $n = 3$ and $\tau = 1$, consistent with the other two methods. In order to make most use of the relatively short data set, the neighboring points of Y_t are selected from all time steps except t. In addition to the causal relations with a non-positive optimal time lag, those relations with a p-value above 0.05 are also disposed.

5 PCMCI Causality Method

5.1 Description of Method

PCMCI is a causal discovery method described in [22] which identifies relevant variables for conditioning and estimates causal networks from time series data. The method makes use of a "time series graph" made of nodes representing the state variables at different time-lags. If the time lag is denoted by τ, a causal link is notated $X^i_{t-\tau} \to X^j_t$, and this link exists if $X^i_{t-\tau}$ is not conditionally independent of X^j_t given the past of all variables. Assuming the causal structure does not change over time, the same links are present at each time step.

The parents $\mathcal{P}(X)$ of a variable X are defined as the set of all nodes with a link towards X. For example, in Figure 5.1, the parents of the variables X^1 at $t-1$ and variable X^3 at time t are shown with the red boxes and blue boxes respectively. However, estimating these parents directly by testing for conditional independence on the whole past is problematic due to high-dimensionality and because conditioning on irrelevant variables leads to biases.

PCMCI estimates causal links by a two-step procedure [22]:

1. **Condition-selection**: For each variable X^j, estimate a superset of parents $\tilde{\mathcal{P}}(X^j_t)$ with an iterative Markov discovery algorithm [7] such as PC_1 algorithm.

2. **Momentary conditional independence**: To test whether $X^i_{t-\tau} \to X^j_t$ with MCI, we evaluate

$$X^i_{t-\tau} \perp X^j_t \mid \tilde{\mathcal{P}}(X^j_t), \tilde{\mathcal{P}}(X^i_{t-\tau}) \tag{5.1}$$

The condition-selection step reduces the dimensionality and avoids conditioning on irrelevant variables. The second step checks momentary conditional independence (MCI) conditions between $X^i_{t-\tau}$ and X^j_t, and checks whether or not $X^i_{t-\tau}$ and X^j_t are not conditionally independent given $\tilde{\mathcal{P}}(X^j_t)$ and $\tilde{\mathcal{P}}(X^i_{t-\tau})$.

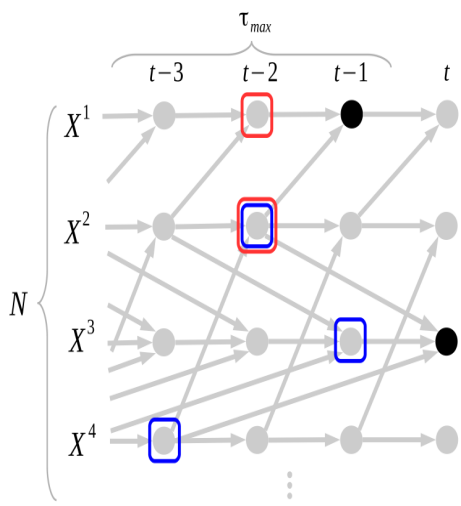

Figure 5.1: Time series graph,representing the time-lagged causal dependency structure underlying the data [20].

5.2 Implementation

Tigramite [20] is a Python [17] package implementing the above causal time series analysis. It helps to efficiently reconstruct causal graphs from high-dimensional time series data sets, the meteorological variables in our case, and model the obtained causal dependencies for causal mediation and prediction analyses. Causal discovery is based on linear as well as non-parametric conditional independence tests applicable to discrete or continuously-valued time series [22].

Our implementation of the PCMCI causality method mirrors that found in the basic tutorial documentation for the Tigramite Package in [20]. We have modified the code to read data from our data files, rather than generating a random sequence, to suppress visual output when desired, to store the images separately, and to store the connection information in output files for postprocessing. We can then run separate scripts on these data files to prepare them for visualization, sorting and filtering by connection strength.

We note that connection strength is reported in two ways, by p-values and by MCI partial correlations. We must be careful in using these values, since MCI partial correlations is more faithful to the PCMCI method, but is not easily compared to the other methods described in this document. However, p-values are quite ubiquitous and can be easily considered in light of the CCM and Grange Causality methods, although they do not contain as much information about the results of the PCMCI algorithm.

6 Identification of Causal Links between Atmospheric Oscillations using Statistical Models

6.1 Granger Causality

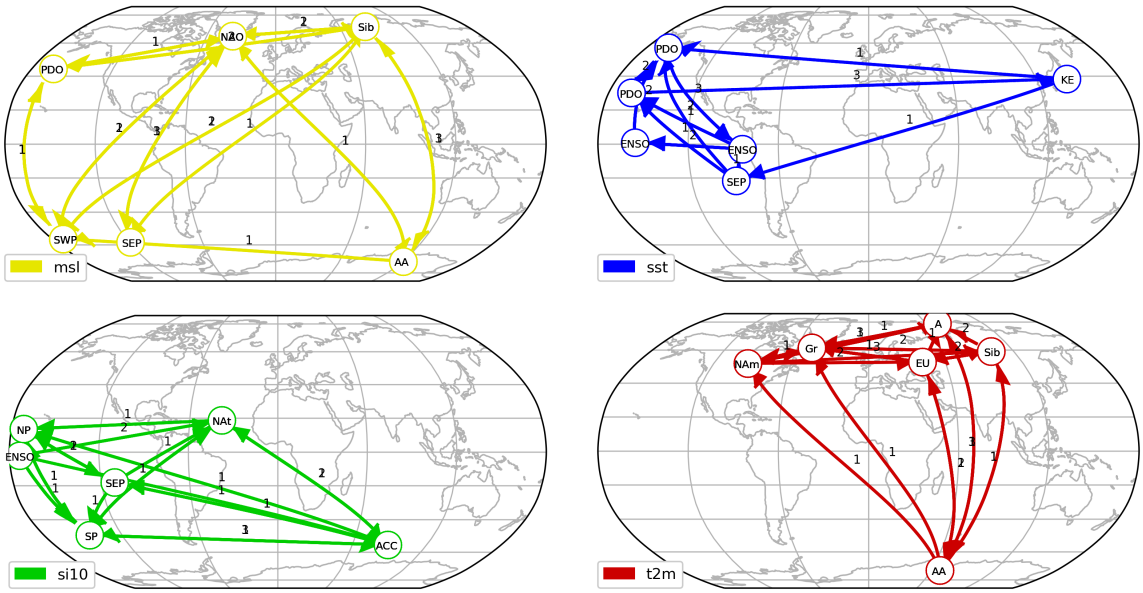

Figure 6.1: Causal links discovered between dominant REOF modes of mean sea level pressure (top left), sea surface temperature (top right), 10m wind speed (bottom left), and 2m-temperature (bottom right) using the Granger causality method. Some of the well-known modes shown here are North Atlantic Oscillation (NAO), Pacific Decadal Oscillation (PDO), El Niño-Southern Oscillation (ENSO). The directionality of causal modes are shown using the arrow directions, and the corresponding time lags are printed near the middle of each arrow.

Fig. 6.1 shows the identified causal links from the Granger causality approach, based on dimensionally reduced time series of mean sea level pressure, sea surface temperature, 10m wind speed, and 2m temperature. In mean sea level pressure, NAO mediates all the other observed teleconnection patterns. However, the linkage of ENSO has not been identified. The effect of ENSO on other atmospheric teleconnection patterns is explained via the SST variable. The observed change in the eastern Pacific (ENSO) with one or two months lag time is most likely to affect the PDO oscillation. Likewise, the 10m wind speed captured the ENSO variability; however, the identification of the Northern Hemispheric wind patterns and their relation to ENSO has not been found. The 2 meter temperature could not capture the ENSO, PDO or NAO but did discover the causal connection between the Arctic and Antarctic oscillations, both directly and indirectly through Siberia, North America, and Greenland. The calculated time lag for any variables suggests that one-month effective change may be enough to drive the change in other linked patterns.

6.2 CCM method

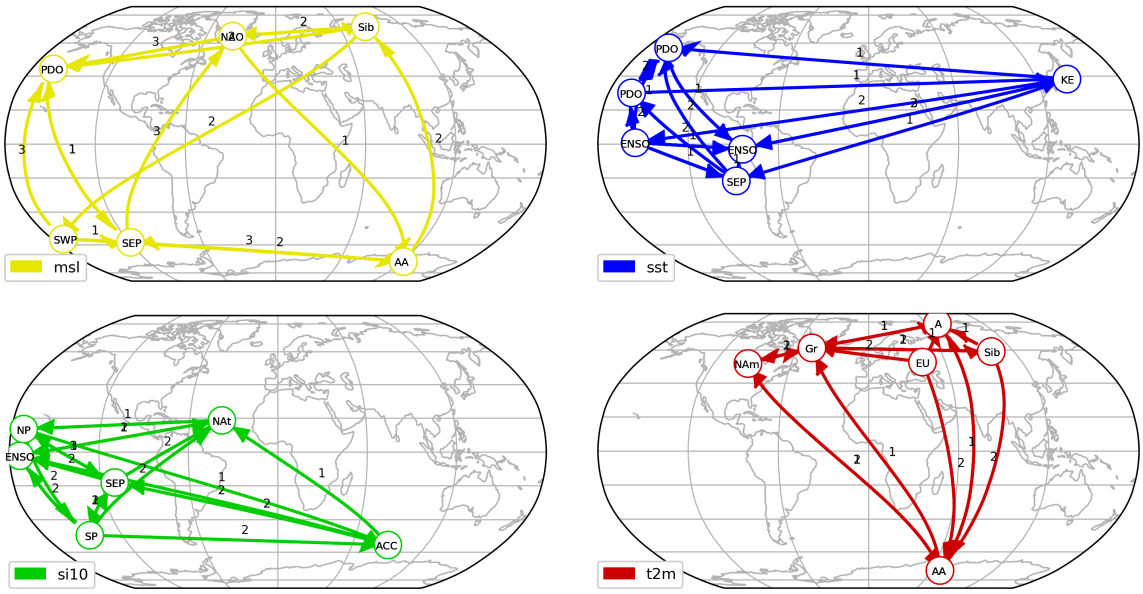

Figure 6.2: Causal links discovered between dominant REOF modes of mean sea level pressure (top left), sea surface temperature (top right), 10m wind speed (bottom left), and 2m-temperature (bottom right) using the CCM causality method. Some of the well-known modes shown here are North Atlantic Oscillation (NAO), Pacific Decadal Oscillation (PDO), El Niño-Southern Oscillation (ENSO). The directionality of causal modes are shown using the arrow directions, and the corresponding time lags are printed near the middle of each arrow.

The observed causal links in CCM (Fig. 6.2) are similar to those found using the Granger causality method, but some of the causal relationships have changed in direction. The CCM method was able to capture the mediation of PDO in casual effect on the NAO oscillation and other oscillations in mean sea level pressure. The intervention of ENSO and their causal links observed for both SST and wind speed. However, ENSO didn't have any causal effect on the 2 meter temperature variable.

6.3 PCMCI method

Fewer causal links were discovered for all variables in PCMCI (Fig 6.3). Also, PCMCI has more causal links for sea level pressure than 2m temperature and 10m wind speed. In particular, the SST variable could not capture any of the observed causal discoveries of other methods. The PCMCI method may require a long time-series data to obtain any well-defined causal links between the oscillations.

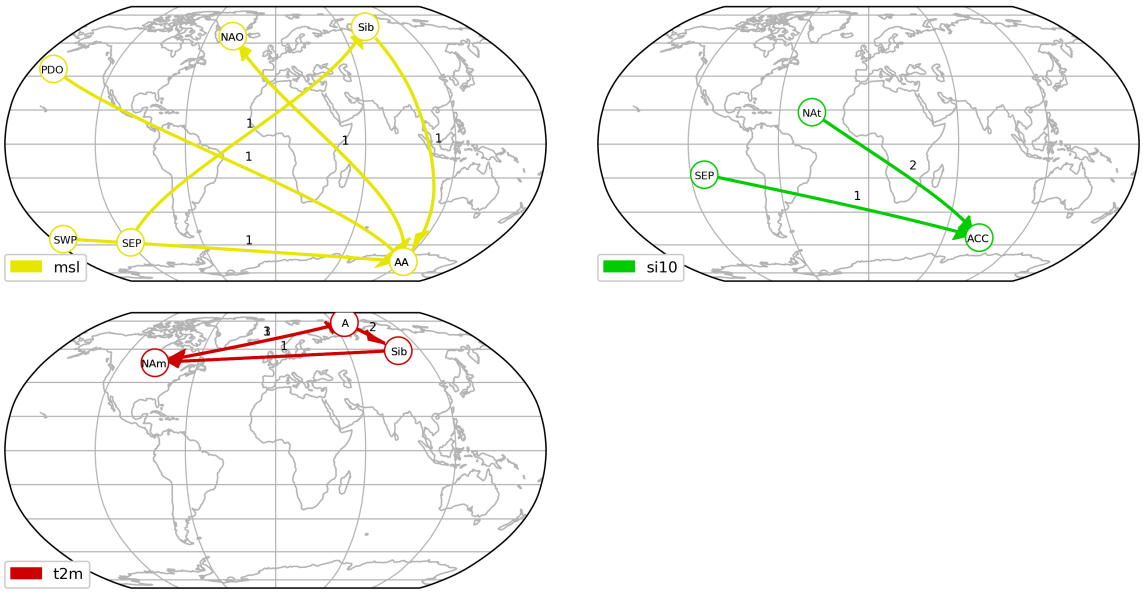

Figure 6.3: Causal links discovered between dominant REOF modes of mean sea level pressure (top left), 10m wind speed (top right), and 2m-temperature (bottom left) using the PCMCI method. No significant causal links are identified for sea surface temperature. Some of the well-known modes shown here are North Atlantic Oscillation (NAO), Pacific Decadal Oscillation (PDO), El Niño-Southern Oscillation (ENSO). The directionality of causal modes are shown using the arrow directions, and the corresponding time lags are printed near the middle of each arrow.

7 Identification of causal discovery across the variables

7.1 Comparison of cross-variable causality between the statistical models

Figures 7.1-7.3 show the causal graphs by the three causality discovery menthods for all four variables. Similarly to how we plotted single variables, we only draw the 20 causal links with the lowest p-values. In this section, we will focus on discussing cross-variable causal links identified by different methods.

We visualize similar multivariate results for the CCM method in Figure 7.1. When using the CCM method, the top six strongest cross-variable causal relations identified all have clear physical bases. First, the causal link from SST-4/SST-7/SST-0 to SI10-5 indicates that the trade wind anomaly in the central Pacific responds to SST anomaly in the equatorial eastern/central Pacific and along the western coast of South America where ocean upwelling frequently occurs, consistent with the Bjerknes feedback (e.g., [14]). Second, the causal link from SI10-5 to SST-3 suggests that the 10m wind anomaly in the equatorial central Pacific can cause an SST anomaly there. This relationship may reflect the zonal advection feedback of ENSO, where a weaker easterly trade wind induces eastward advection of surface water

from the western Pacific warm pool (e.g., [3, 14]). Third, the causal link from MSL-1 to SI10-0 shows the connection between Antarctic sea level pressure and 10m wind anomalies in the Antarctic Circumpolar Current (ACC) region. As is expected from a geostrophic balance, wind speed strongly depends on the local pressure gradient. Fourth, a similar connection between pressure and wind anomalies can be found in the southern Pacific, indicated by the causal link from MSL-3 to SI10-1. However, some of the weaker cross-variable causations are more difficult to interpret, such as the link from wind anomaly in the southern Pacific (SI10-1) to the SST anomaly near Alaska (SST-4).

Similar to CCM, Granger causality results in Figure 7.2 also suggest causal links from SST-4 to SI10-5 (the Bjerknes feedback in the equatorial Paciifc), from MSL-1 to SI10-0 (pressure and wind near the ACC) and from MSL-3 to SI10-1 (pressure and wind in the southern Pacific). The cross-variable causality detected by Granger causality around Antarctica is more difficult to interpret, perhaps due to confounding and indirect factors. For example, Granger causality identifies a strong causal link from MSL-1 to T2M-2, which is the influence of sea level pressure on surface temperature in Antarctica. However, we usually observe the influence of temperature on pressure in local circulations, such as a sea-breeze. Furthermore, Granger causality suggests a causal link from surface winds (SI10-0) to temperature (T2M-2), too. Therefore, it is possible that both of these causalities capture the importance of atmospheric circulation, especially low-level heat transport, in Antarctica.

We also include a multivariate causal graph for the PCMCI method in Figure 7.3. PCMCI tends to identify more long-range cross-variable causalities, such as the link from 10 m wind anomalies in the southeast Pacific (SI10-4) to Antarctic sea level pressure (MSL-1), from Greenland temperature (T2M-8) to Siberian pressure (MSL-7), from temperature/wind anomalies in the southeast and north Pacific (SST-7, SST-4 and SI10-7) to Siberian temperature (T2M-6), or from temperature anomalies in the central Pacific (SST-0) to surface winds in the north Atlantic (SI10-6). All of these causalities are more difficult to interpret.

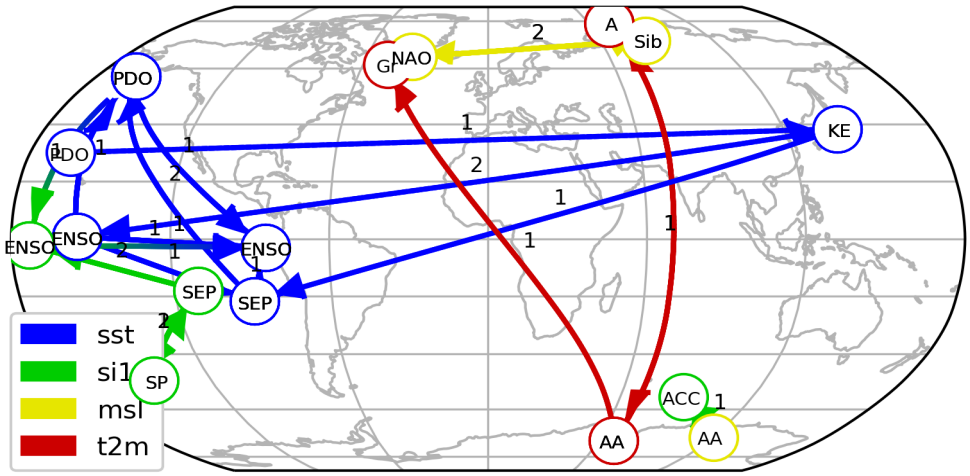

Figure 7.1: Causal graph of SST, SI10, MSL and T2M using the Convergent Cross Mapping method.

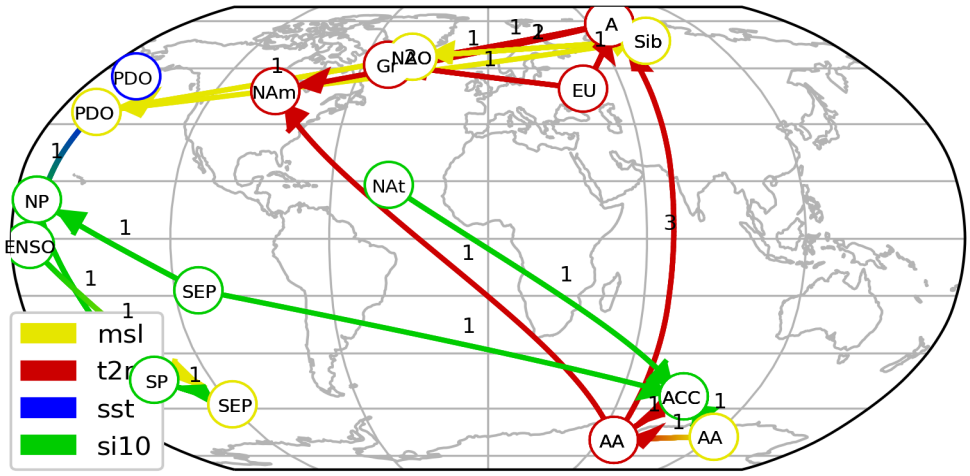

Figure 7.2: Causal graph of SST, SI10, MSL and T2M using the Granger Causality method.

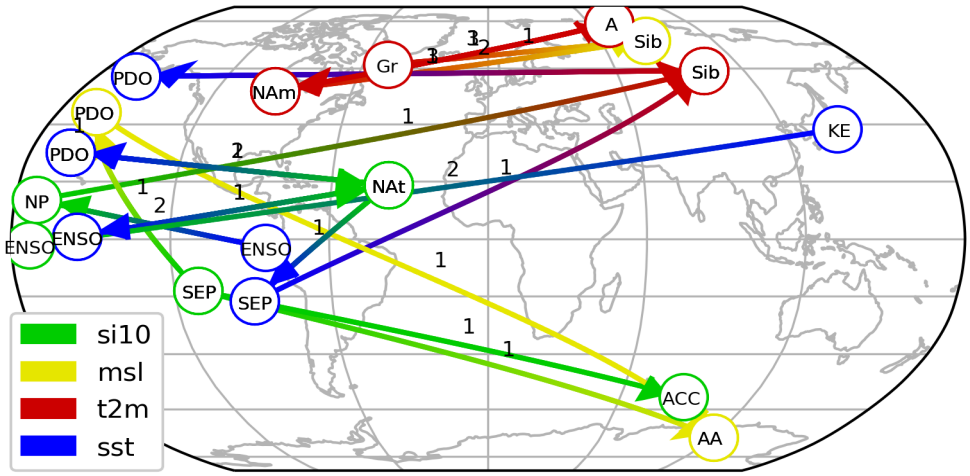

Figure 7.3: Causal graph of SST, SI10, MSL and T2M using PCMCI method.

7.2 Comparison of similarity measures between the statistical models

Causal connections were different for different methods. It is important to quantify the causality methods that have the most similar results. To quantify the similarity coefficients, we used the well-known matrix distance calculation to measure Jaccard coefficients for a different combination of the model results.

The calculated Jaccard distance between the statisticals model is shown in Table 7.1. The lower value corresponds to larger similarity measures between the methods. The Granger causality and CCM pair has a lowest value which corresponds to a relatively large similarity between their findings. However, the identified causal connections had few similarity between CCM and PCMCI due to their high Jaccard distance.

Table 7.1: Matrix distance between the statistical models

Matrix Distance	Granger Causality	CCM	PCMCI
Granger Causality	0	0.6926	0.8683
CCM	0.6926	0	0.9173
PCMCI	0.8683	0.9173	0

8 Conclusions

We tested the functionality of the causal discovery methods based on their causality and direction using the atmospheric variables (e.g., 2m temperature, sea surface temperature, sea ice fraction, and mean sea level pressure). The Granger causality and CCM methods

were most likely to produce similar causal connections for most of the variables. This study was identified the cross variable causal connections that can be related to previous findings of Bjerknes feedback of ENSO, the geostrophic winds around Antarctica and the southern Pacific [14]. However, this study is not intended to explain the important dynamical processes and their effects on causal connections among the variables. The causal connection results were suggested that significance of each selected PC components may vary the results of the statistical model reasonably depending on the variables. The Jaccard coefficients were calculated to test the similarity between the statistical models. More significant Jaccard coefficient was found between PCMCI and CCM methods. Further work may be needed to explain the physical processes involved in the causal connections and their application in regional climate prediction.

9 Appendix

9.1 Code and Results

The source code and causality discovery results of this work can be found at https://github.com/big-data-lab-umbc/cybertraining/tree/master/year-2-projects/team-2.

Acknowledgments

This work is supported by the grant CyberTraining: DSE: Cross-Training of Researchers in Computing, Applied Mathematics and Atmospheric Sciences using Advanced Cyberinfrastructure Resources from the National Science Foundation (grant no. OAC–1730250).

The hardware in the UMBC High Performance Computing Facility (HPCF) is supported by the U.S. National Science Foundation through the MRI program (grant nos. CNS–0821258, CNS–1228778, and OAC–1726023) and the SCREMS program (grant no. DMS–0821311), with additional substantial support from the University of Maryland, Baltimore County (UMBC). See hpcf.umbc.edu for more information on HPCF and the projects using its resources.

References

[1] Andrew Arnold, Yan Liu, and Naoki Abe. "Temporal Causal Modeling with Graphical Granger Methods". In: *In Proceedings of the 13th Int. Conference on Knowledge Discovery and Data Mining, 66 – 75: Association for Computing Machinery*. 2007.

[2] S. Brönnimann et al. "Extreme climate of the global troposphere and stratosphere in 1940-42 related to El Niño". In: *Nature* 431.7011 (2004), pp. 971–974. ISSN: 1476-4687. DOI: 10.1038/nature02982. URL: https://doi.org/10.1038/nature02982.

[3] Antonietta Capotondi et al. "Understanding ENSO diversity". In: *Bulletin of the American Meteorological Society* 96.6 (2015), pp. 921–938.

[4] D. P. Dee et al. "The ERA-Interim reanalysis: configuration and performance of the data assimilation system". In: *Quarterly Journal of the Royal Meteorological Society* 137.656 (2011), pp. 553–597. DOI: 10.1002/qj.828. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.828. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.828.

[5] Imme Ebert-Uphoff and Yi Deng. "Causal Discovery for Climate Research Using Graphical Models". In: *Journal of Climate* 25.17 (2012), pp. 5648–5665. DOI: 10.1175/JCLI-D-11-00387.1. URL: https://doi.org/10.1175/JCLI-D-11-00387.1.

[6] Klaus Fraedrich and Klaus Müller. "Climate anomalies in Europe associated with ENSO extremes". In: *International Journal of Climatology* 12.1 (1992), pp. 25–31. DOI: 10.1002/joc.3370120104. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.3370120104. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.3370120104.

[7] Maria L. Rizzo Gábor J. Székely and Nail K. Bakirov. "Measuring and testing dependence by correlation of distances". In: *The Annals of Statistics* 35.6 (Oct. 2007).

[8] A. Gozolchiani et al. "Pattern of climate network blinking links follows El Niño events". In: *EPL (Europhysics Letters)* 83.2 (July 2008), p. 28005. DOI: 10.1209/0295-5075/83/28005. URL: https://doi.org/10.1209%5C%2F0295-5075%5C%2F83%5C%2F28005.

[9] C. W. J. Granger. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". In: *Econometrica* 37.3 (1969), pp. 424–438. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1912791.

[10] *Homepage — ERA Interim, Monthly Means of Daily Means*. https://apps.ecmwf.int/datasets/data/interim-full-moda/levtype=sfc/. Accessed: 2019-6-27.

[11] John D. Horel and John M. Wallace. "Planetary-Scale Atmospheric Phenomena Associated with the Southern Oscillation". In: *Monthly Weather Review* 109.4 (1981), pp. 813–829. DOI: 10.1175/1520-0493(1981)109<0813:PSAPAW>2.0.CO;2. eprint: https://doi.org/10.1175/1520-0493(1981)109<0813:PSAPAW>2.0.CO;. URL: https://doi.org/10.1175/1520-0493(1981)109%3C0813:PSAPAW%3E2.0.CO;2.

[12] Henry F Kaiser. "The varimax criterion for analytic rotation in factor analysis". In: *Psychometrika* 23.3 (1958), pp. 187–200.

[13] Anna Krakovská and Filip Hanzely. "Testing for causality in reconstructed state spaces by an optimized mixed prediction method". In: *Phys. Rev. E* 94 (5 Nov. 2016), p. 052203. DOI: 10.1103/PhysRevE.94.052203. URL: https://link.aps.org/doi/10.1103/PhysRevE.94.052203.

[14] J David Neelin et al. "ENSO theory". In: *Journal of Geophysical Research: Oceans* 103.C7 (1998), pp. 14261–14290.

[15] Josef Perktold et al. *Statsmodels, Statistics in Python*. https://www.statsmodels.org/. Accessed: 2019-06-06.

[16] Josef Perktold et al. *Vector Autoregressions*. https://www.statsmodels.org/devel/vector_ar.html/. Accessed: 2019-06-06.

[17] Python Software Foundation. *Python*. Python Software Foundation. 2019. URL: https://docs.python.org/3/.

[18] Michael Richman. "Review article, rotation of principal components". In: *J. Climatol.* 6 (Jan. 1986), pp. 293–355.

[19] Michael B. Richman. "Rotation of principal components". In: *Journal of Climatology* 6.3 (1986), pp. 293–335. DOI: 10.1002/joc.3370060305. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.3370060305. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.3370060305.

[20] J. Runge. https://github.com/jakobrunge/tigramite. software downloaded April 18th, 2019.

[21] Jakob Runge et al. "Identifying causal gateways and mediators in complex spatio-temporal systems". In: *Nature Communications* 6 (Oct. 2015). Article, 8502 EP. URL: https://doi.org/10.1038/ncomms9502.

[22] J. Runge et al. *Detecting causal associations in large nonlinear time series datasets*. https://arxiv.org/abs/1702.07007v2. Accessed: 2018-06-28. 2018.

[23] David M. Straus and J. Shukla. "Does ENSO Force the PNA?" In: *Journal of Climate* 15.17 (2002), pp. 2340–2358. DOI: 10.1175/1520-0442(2002)015<2340:DEFTP>2.0.CO;2. eprint: https://doi.org/10.1175/1520-0442(2002)015<2340:DEFTP>2.0.CO;2. URL: https://doi.org/10.1175/1520-0442(2002)015%3C2340:DEFTP%3E2.0.CO;2.

[24] George Sugihara et al. "Detecting causality in complex ecosystems". In: *science* 338.6106 (2012), pp. 496–500.

[25] Floris Takens. "Detecting strange attractors in turbulence". In: *Dynamical systems and turbulence, Warwick 1980*. Springer, 1981, pp. 366–381.

[26] Martin Vejmelka et al. "Non-random correlation structures and dimensionality reduction in multivariate climate data". In: *Climate Dynamics* 44.9 (May 2015), pp. 2663–2682. ISSN: 1432-0894. DOI: 10.1007/s00382-014-2244-z. URL: https://doi.org/10.1007/s00382-014-2244-z.

[27] Hao Ye et al. "Distinguishing time-delayed causal interactions using convergent cross mapping". In: *Scientific reports* 5 (2015), p. 14750.

[28] Cunlu Zou, Katherine J. Denby, and Jianfeng Feng. "Granger causality vs. dynamic Bayesian network inference: a comparative study". In: *BMC Bioinformatics* 10 (Dec. 2009). PMC2795767[pmcid], pp. 401–401. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-401. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2795767/.