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

## Team - 2

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

http://cybertraining.umbc.edu/

# Background: Testing of causal discovery Algorithms using the climate reanalysis

1. The study of causality has recently emerged as a primary research tool in many areas of science
2. ENSO is an important large-scale climate variability responsible for many extreme weather events (Donnelly & Woodruff, 2007)
3. 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
4. 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.

# Background: Testing of causal discovery Algorithms using the reanalysis climate data

5. 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).

6. We mainly intend to explain the causal connection between the ENSO mode and other prominent climate variabilities using the Granger causality (Granger, 1969), Convergent cross Mapping (CCM, Sugihara et al., 2012}, and PCMCI (Runge et al., 2018) approaches.

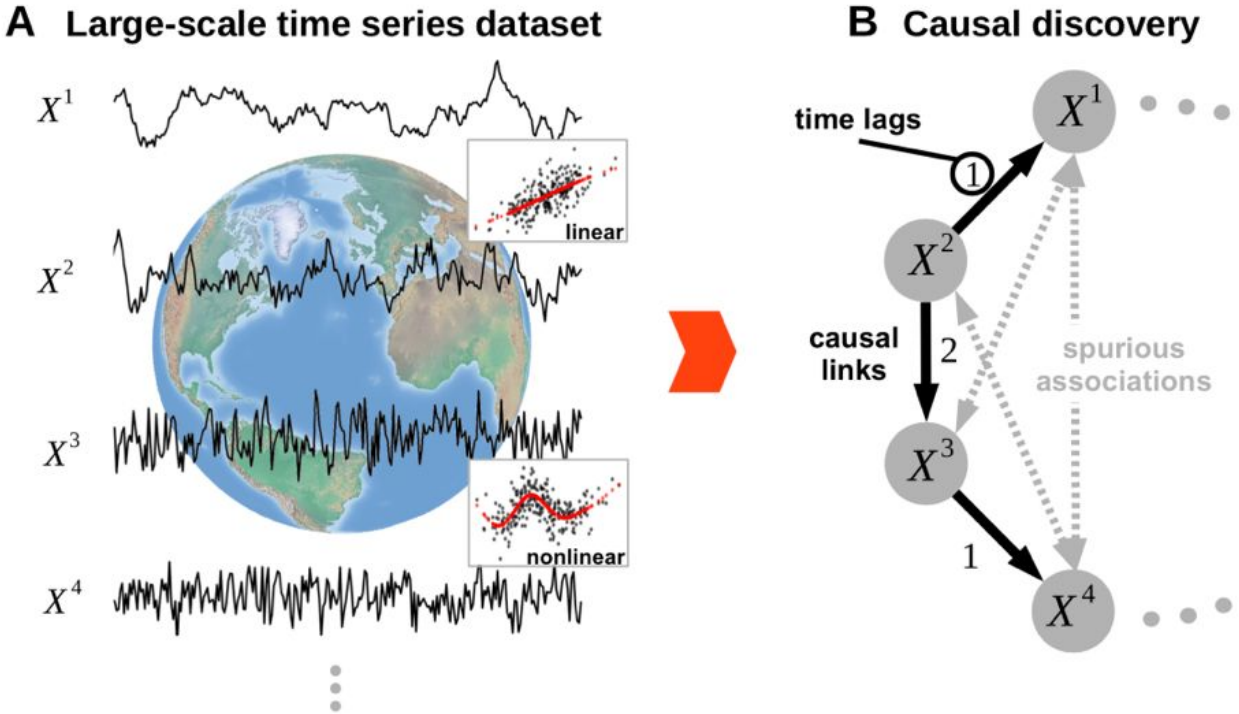
# Data:

1. ERA-Interim global reanalysis data obtained from European Centre for Medium-Range Weather Forecasts (ECMWF) - 1979-present

Weather parameters:

- 2 meter air temperature (T2M)
- Sea surface temperature (SST)
- Mean sea level pressure (MSL)
- 10 meter wind speed (SI10)

# Methods: Dimension reduction of Atmospheric variables




Runge et al., 2017

# Step 1:  EOF and PC analyses

1. Data decomposition of T2M, SST, SI10, and MSL using the rotated Empirical Orthogonal Function (EOF)
2. Calculated 30 EOF modes and PC modes for each variable
3. Then, 6 PCs were selected for each variable based on the amplitude and spread of the associated spatial variability

# Step 2:  Causal discovery

4. These selected 6 modes were then subjected to causal discovery processes using the statistical models
5. Three models tested in this study are Granger causality, CCM, and PCMCI

# Causal discovery Algorithms

## Granger Causality- Granger (1969)

- The Granger causality test is a statistical hypothesis test for determining whether one time series is useful in forecasting another, first proposed in 1969.
- A variable X that evolves over time Granger-causes another evolving variable Y if predictions of the value of Y based on its own past values and on the past values of X are better than predictions of Y based only on its own past values.
- If (1) predicts better y by involving x than (2), x granger causes y

$$y_t \approx A \cdot \vec{y_{t-1}} + B \cdot \vec{x_{t-1}} \quad (1)$$

$$y_t \approx A \cdot \vec{y_{t-1}} \quad (2)$$

# Granger Causality Implementation

- Granger causality implementation utilized the Vector Autoregressions tsa.vector_ar package from statsmodels 0.9.0 in Python
- Parameters:
  - 30 EOF modes
  - Maximum lag: 3 months
  - P-value threshold: 0.05
- Selected 20 edges with lowest p-values for visual inspection

# Convergent Cross mapping (CCM) - Sugihara et al. (2012)

1. Nonlinear state space reconstruction method to distinguish causality from correlation
2. It separates the weak signals of dynamical system
3. This approach is based on the assumption that the causation is unilateral
4. $X \Rightarrow Y$ ( estimation of X is possible using Y)

# Mathematical mechanism of CCM

$$\frac{d\boldsymbol{X}}{dt} = f(\boldsymbol{X}, \boldsymbol{U}) \tag{1}$$

Here, $\boldsymbol{X}$ and $\boldsymbol{U}$ are vectors of the variables and parameters: $\boldsymbol{X} = [x_1, x_2, \ldots, x_n]^{\mathrm{T}}$ and $\boldsymbol{U} = [u_1, u_2, \ldots, u_l]^{\mathrm{T}}$. For example, taking the two variables $x$ and $y$ of $\boldsymbol{X}$, each element of the corresponding shadow manifolds of $x$ and $y$ are constructed according to

$$\boldsymbol{M}_{x,t} = \left[x_t, x_{t-\tau_x}, x_{t-2\tau_x}, \ldots, x_{t-(E-1)\tau_x}\right] \tag{2}$$

$$\boldsymbol{M}_{y,t} = \left[y_t, y_{t-\tau_y}, y_{t-2\tau_y}, \ldots, y_{t-(E-1)\tau_y}\right] \tag{3}$$

https://www.sciencedirect.com/science/article/pii/S0098135417301448#bib0070

# Conditional Independence

This is found through the notion of Conditional Independence.

This takes three variables, Ex: **Height** is conditionally dependent on **Vocabulary** (or vice versa) given **Age**.

This means that fixing Age, we see that height and vocabulary have no relationship. I.e., a tall 12 year old will not be any more likely to have a larger vocabulary than a short 12 year old.

# PCMCI Method - Runge et al. (2018)

The PCMCI method works to discover causal links by a two step method. We begin with a list of variables (X, Y, Z), and a maximum time lag (2).

The PC method trims irrelevant parents for each variable

The MCI method does a more stringent causality test for each possible parent.

Each possible parent is a value pair: Not just a variable, but **a variable and a time delay**.

# Example PC run

We follow possible parents of the variable X, but this process will be repeated for Y and Z. (And there is a parallel opportunity here.)

We call the possible parents P(X) = {(X,1), (X,2), (Y,1), (Y,2), (Z,1), (Z,2)}, and the chosen parents P’(X) = { }, empty at first.

**1st Iteration**. We compute simple correlation for each variable and delay in P(X) with X. Any uncorrelated pairs are removed from P(X), the **strongest** pair is moved to P’(X)

# Example PC run

Say now that P(X) = {(X,2), (Y,1), (Y,2), (Z,1)}, and P’(X) = {(Z,2)}.

**Continued iteration.** Now we do numerical tests for **conditional independence** between (X,0) and each pair in P(X) given all pairs in P’(X), removing any independent pairs. The most dependent is added to P’(X).

This way, if (Z,2) was driving other variables, they will be removed.

Now suppose P(X) = {(Y,1), (Y,2)}, P’(X) = {(X,2), (Z,2)}.

**We continue until P(X) is empty.**

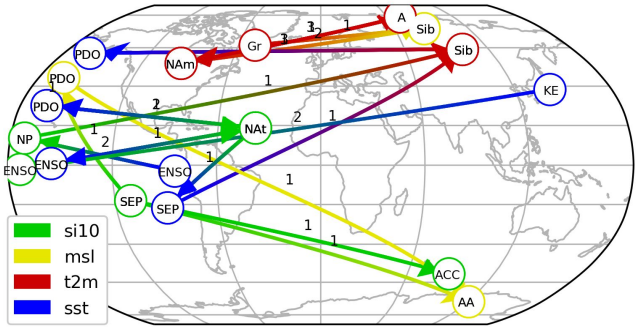
# MCI Method

Now each variable X has a set P’(X) of possible parents. For each variable and each pair in P’(X), we numerically test for conditional independence. We say the pair (A, i) has a causal link to X if

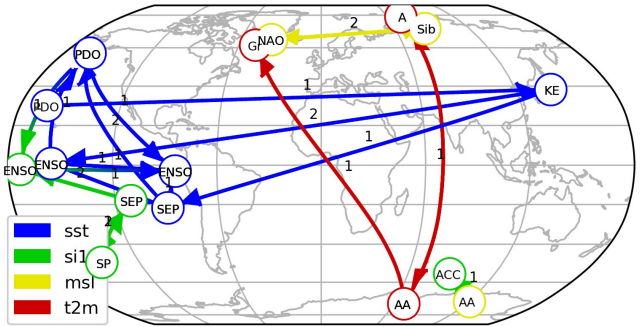
**X is conditionally dependent on (A,i) given P’(X) \ (A,i) and P’(A,i)**

This way we remove any pairs conditionally dependent given the other possible parents of X, or the possible “grandparents” of X through the pair (A,i).

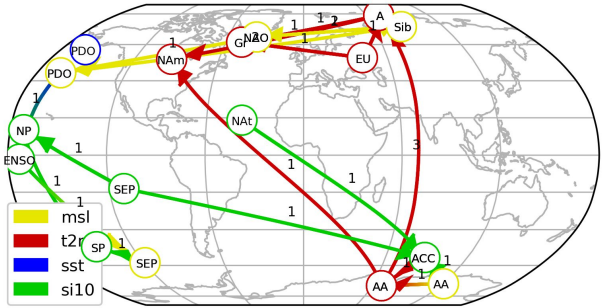
# Results and Discussion




PCMCI




CCM




GC

# Results and Discussion

1. Both CCM and GC detect the the causal links the Bjerknes feedback of ENSO and geostrophic balance around the Antarctica and in the Southern Pacific.

2. Some of the remote cross-variable causal links are also detected, especially by PCMCI.

3. GC identifies some causal links in Antarctica, which may involve confounding and indirect factors.

# Similarity transformation

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 |

- Calculated Jaccard distance between the statisticals model
- 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.
- The identified causal connections had few similarity between CCM and PCMCI due to their high Jaccard distance.

# Conclusion

1. 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, 10m wind speed, and mean sea level pressure).
2. 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
3. The Granger causality and CCM methods were most likely to produce similar causal connections for most of the variables.
4. The Jaccard coefficients were calculated to test the similarity between the statistical models.
5. More significant Jaccard coefficient was found between PCMCI and CCM methods.

# References

Christensen, J. H., et al. (2013) Climate Phenomena and their Relevance for Future Regional Climate Change. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Stocker, T. F., et al. (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Runge, J., et al. (2017) Detecting Causal Associations in Large Nonlinear Time Series Datasets. *arXiv [stat.ME]*. arXiv. http://arxiv.org/abs/1702.07007