The Third International Workshop on Climate Informatics
National Center for Atmospheric Research, Boulder, Colorado, September 26-27, 2013
https://www2.image.ucar.edu/event/ci2013

Causal Discovery in Climate Science Using Graphical Models

I. Ebert-Uphoff
Electrical and Computer Engineering
Colorado State University

Y. Deng
Earth and Atmospheric Sciences
Georgia Institute of Technology

1 Introduction

Causal discovery is the process of *identifying cause-and-effect hypotheses from observational data*. The purpose of this document is to demonstrate the great potential of using causal discovery algorithms in climate science, by showing how they can be applied for two climate science applications.

2 Technical Approach

We use the framework of probabilistic graphical models developed by Pearl [1] and by Spirtes et al. [2]. Specifically, we use algorithms for constraint-based structure learning, such as the PC algorithm developed by Spirtes and Glymour [3] and modifications thereof that deal with temporal data. The PC algorithm generates one or more *graph representations* that describe the potential causal pathways in the system. At the core of this algorithm are *conditional* independence tests that allow us to distinguish between direct and indirect causal connections. Causal discovery of this type has already been applied with great success in disciplines ranging from the social sciences to computer science, engineering, medical diagnosis and bioinformatics.

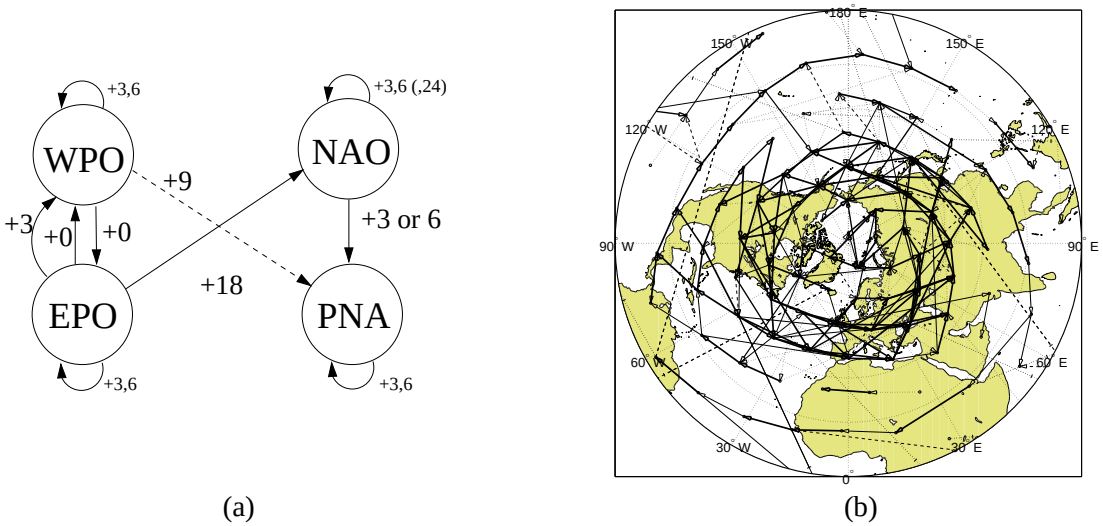
3 Limitations

There are certain limitations to the interpretation of the causal graphs [1,2,4], the most important one dealing with potential hidden common causes. Namely, we need to consider the possibility that any link detected by the PC algorithm may either present a direct causal connection, be due to a hidden common cause, or a combination of the two. Thus we call the results from the analysis *causal hypotheses*, and they must be tested one by one by a domain expert. The contribution of this causal discovery process is therefore to *reduce the number of causal hypotheses to a manageable set that can then be tested by a domain expert*.

4 Application 1: Four-mode example

We applied our method to derive hypotheses of causal relationships between four prominent modes of atmospheric low-frequency variability in boreal winter including the Western Pacific Oscillation (WPO), Eastern Pacific Oscillation (EPO), Pacific-North America (PNA) pattern, and North Atlantic Oscillation (NAO) [5]. Figure 1(a) shows a summary of the relationships we found, with numbers along links providing delays in days. To generate Figure 1(a) we used a temporal model with D=3 days between time slices, i.e. the only possible delays are 0, 3, 6, ..., 30 days. (For results using D=1 and D=2 days and further discussion, see [5].)

It is found that WPO and EPO are nearly indistinguishable from the cause-effect perspective as strong simultaneous coupling is identified between the two. In addition, changes in the state of EPO (NAO) may cause changes in the state of NAO (PNA) approximately 18 (3-6) days later. These results are consistent with previous findings on dynamical processes connecting different low-frequency modes (e.g., interaction between synoptic and low-frequency eddies), and provide the basis for formulating new hypotheses regarding the time scale and temporal sequencing of dynamical processes responsible for these connections. Thus the results are consistent with effects reported in current literature, but some of the time scales obtained are new and have not yet been validated.



(a) (b)

Figure 1: Summary graph for four-mode problem (a) and snapshot of causal-discovery climate network (b).

5 Application 2: A New Type of Climate Network

The basic idea of *climate networks* is to use atmospheric fields - or other physical quantities - to define a network of nodes, where each node represents a point on a global grid. In a traditional correlation-based climate network two nodes are connected if the cross-correlation of the data associated with those two nodes is beyond a threshold. We recently used causal discovery methods to define a new type of climate network. The key idea is to interpret large-scale atmospheric dynamical processes as *information flow* around the globe, and to identify the pathways of this information flow using causal discovery. While *correlation-based* climate networks focus on *similarity* between nodes, this new method provides an alternative viewpoint by focusing on *information flow* within the network over time [6].

Figure 1(b) shows a sample network result based on daily values of 500 mb geopotential height over the entire globe for boreal winter during the period 1950 to 2000 using NCEP-NCAR reanalysis data. We use *Fekete points* [7] as grid points to obtain an equally spaced grid around the globe. Fig. 1(b) shows the strongest pathways of information flow happening within a single day - and is just one of a series of figures we would look at for such a network. Results suggest that synoptic-scale, sub-weekly disturbances act as the main information carrier in this network. We also define a variety of network measures, e.g. we can measure for how many days information of the initial geopotential height value at a grid point can still be significantly felt at the same grid point (local memory), or for how long it can be felt at other grid points (remote impact). This new approach serves as a tool to better understand certain dynamic processes of the earth's climate. For example, comparing boreal summer and winter we found a poleward retreat of synoptic-scale disturbances in boreal summer, which is largely responsible for a corresponding poleward shift of local maxima in local memory and remote impact, most evident in the North Pacific sector. For the NH as a whole, both local memory and remote impact strengthen from winter to summer leading to intensified information flow and more tightly-coupled network nodes during the latter period [6]. We are currently exploring the changing characteristics of atmospheric information flow in a warming climate, by applying climate networks to the output of GCM models for current and future climate projection.

Conclusions Causal reasoning has tremendous potential to generate causal hypotheses from data that domain experts can then investigate further. We hope to have stimulated more interest in this exciting area.

Acknowledgments The NCEP-NCAR reanalysis data was provided through the NOAA Climate Diagnostics Center. This research was in part supported by the DOE Office of Science RGCM program (grant DE-SC0005596) and NASA Energy and Water Cycle Study (NEWS) program (grant NNX09AJ36G).

6 References

[1] Pearl, J., 1988: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. 2nd ed. Morgan Kaufman Publishers, 552 pp.

[2] Spirtes, P., C. Glymour, and R. Scheines, 1993: *Causation, Prediction, and Search*. Springer Lecture Notes in Statistics. 1st ed. Springer Verlag, 1993, 526 pp.

[3] Spirtes, P., and C. Glymour, 1991: *An algorithm for fast recovery of sparse causal graphs.* Social Science Computer Review, 9(1):6772, 1991.

[4] Koller, D., and N. Friedman, 2009: *Probabilistic Graphical Models - Principles and Techniques.* 1st ed. MIT Press, 1280 pp.

[5] Ebert-Uphoff, I. and Y. Deng, 2012: *Causal Discovery for Climate Research Using Graphical Models*, Journal of Climate, Vol. 25, No. 17, doi:10.1175/JCLI-D-11-00387.1, Sept 2012, pp. 5648-5665.

[6] Ebert-Uphoff, I. and Y. Deng, 2012: *A New Type of Climate Network based on Probabilistic Graphical Models: Results of Boreal Winter versus Summer*, Geophysical Research Letters, vol. 39, L19701, 7 pages, doi:10.1029/2012GL053269, 2012.

[7] Bendito, E., A. Carmona, A. M. Encinas, and J. M. Gesto, 2007: *Estimation of Fekete points*, J. Comput. Phys., 225, 23542376, doi:10.1016/j.jcp.2007.03.017.