4th International Workshop on Climate Informatics
September 25-26, 2014
Hosted by the National Center for Atmospheric Research in Boulder, CO

USING CAUSAL DISCOVERY TO LEARN ABOUT OUR PLANET'S CLIMATE - RECENT PROGRESS

Imme Ebert-Uphoff[1], Yi Deng[2]

***Abstract*—Causal discovery is the process of identifying cause-and-effect hypotheses from observational data. We use causal discovery to construct networks that track interactions around the globe based on time series data of atmospheric fields, such as daily geopotential height data. At last year's workshop we explained the basic concepts of using this approach to identify climate properties. Here we report recent progress, namely (1) an analysis of anticipated changes in the climate's network structure under enhanced greenhouse gases (GHGs) and (2) computational advances that allow us to move to spatial networks.**

I. INTRODUCTION

Causal discovery theory is based on *probabilistic graphical models* and provides algorithms to identify potential cause-and-effect relationships from observational data [1,2]. The output of such algorithms is a graph structure showing potential causal connections of the input variables. While used extensively in the social sciences and economics for decades, and more recently in bioinformatics with great success, it has only recently been applied in climate science.

The specific method used here is *constraint-based structure learning*, specifically the classic *PC* algorithm [2] and a new variation thereof, the *PC stable* algorithm [3]. Furthermore we extended those algorithms to learn temporal models, following the procedure first outlined in [4]. Since most climate scientists are unlikely to be familiar with this method, we provide a very short, intuitive explanation of the method. Constraint-based structure learning is based on two key facts:

1) We can distinguish between *direct* and *indirect* connections based on observed data, using *conditional* independence tests (CI tests). (Fisher's Z-test is most commonly used for the CI tests.)
2) We cannot *prove* causal connections (primarily due to potential hidden common causes), but we can *disprove* so many connections that only few potential causal relationships are left at the end.

[1]Dept. of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO, USA, iebert@engr.colostate.edu. [2]School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, GA, USA, yi.deng@eas.gatech.edu.

The method proceeds in the following five steps:

1) First we assume that every variable is a cause of every other variable (fully connected graph).
2) Then we perform CI tests to eliminate as many connections as possible (pruning).
3) Whatever is left at the end are the *potential* causal connections.
4) Arrow *directions* are determined (as far as possible) from additional CI tests and temporal constraints.
5) Evaluation step: In the final graph, every link (or group of links) must be checked by a domain expert. If we can find a physical mechanism that explains it (e.g. from literature), the causal connection is confirmed. Otherwise, the link presents a *new hypothesis* to be investigated.

To apply this method to analyze climate we define a grid around the globe and evaluate an atmospheric field at all grid points, which provides time series data at the grid points. We then use the method above to identify the strongest *pathways of interactions* around the globe based on the time series data [5].

Figure 1 shows a sample network plot. It turns out that the interactions captured by this particular network are *storm tracks*, thus satisfying the Evaluation step (Step 5 above). In general, which physical processes are tracked depends on the atmospheric field used and the time scale, e.g. daily data vs. monthly data.

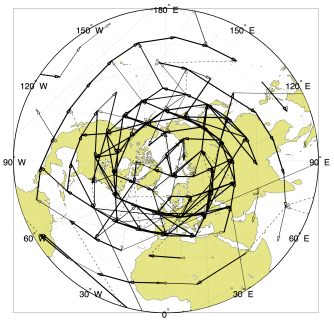
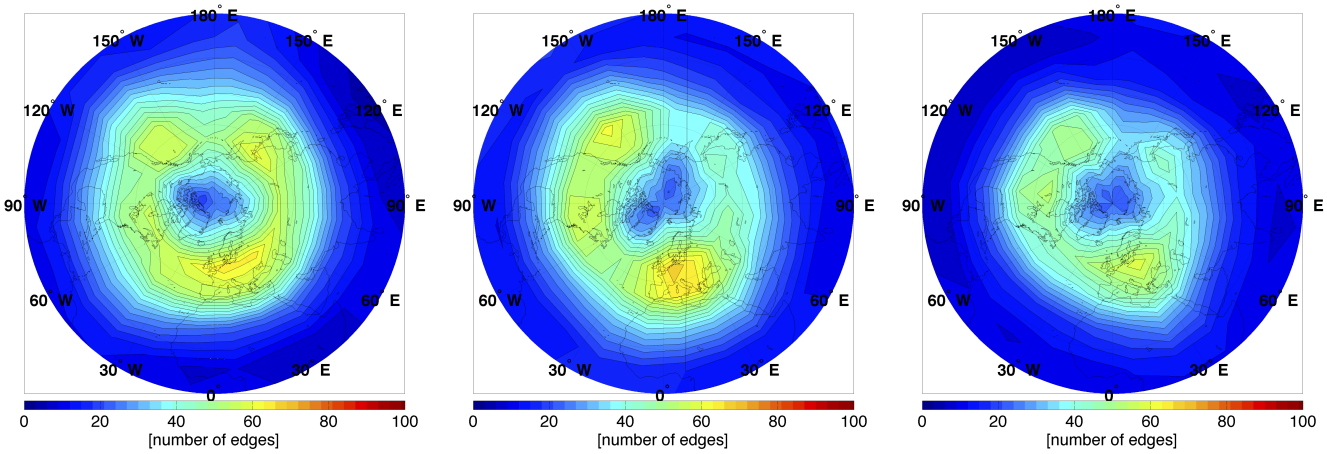


Fig. 1. Sample network showing the strongest direct interactions in the Northern hemisphere that take about 1 day from cause to effect. Based on 500mb daily geopotential height data for boreal winter (DJF months) from NCEP/NCAR Reanalysis data (1948-2011).

4th International Workshop on Climate Informatics
September 25-26, 2014
Hosted by the National Center for Atmospheric Research in Boulder, CO



(a) NCEP/NCAR 1950-2000 (b) CCSM4 1950-2000 (c) CCSM4 2050-2100

Fig. 2. Contour plots showing number of outgoing edges for boreal winter for three different data sets.

II. Results for a Warming Climate Based on CCSM4.0 Model

To study the effect of a warming climate we applied our analysis to daily geopotential height data at 500mb for boreal winter (DJF) from three different data sets: (1) NCEP/NCAR reanalysis (observation) for 1950-2000; (2) NCAR CCSM4.0 model for 1950-2000; (3) NCAR CCSM4.0 model's future climate projection under RCP8.5 scenario for 2050-2100.

Figure 2 shows one set of properties obtained using the three different data sets, namely the number of outgoing edges at each location. Light colors indicate locations that have a strong impact on other locations. Key observations from these and other plots include: (1) Prominent midlatitude interaction pathways in midtroposphere weaken and shift poleward. (2) Major tropical interaction pathways start diminishing. (3) Averaged over the entire Northern Hemisphere, the atmospheric interactions weaken. (4) This weakening, especially in the tropics, leads to reduced interconnectivity among different geographical locations and thus a more chaotic atmosphere in the future. These findings are consistent with the literature, since midlatitude storm tracks are known to move poleward in a warming climate. Using our methods we can now localize some of the effects, based on the output of climate models [6].

III. Moving toward 3D models

We started out using available implementations of the *PC* and *PC stable* algorithms (in *Java*, *Matlab* and *R*). However, those severely limited how many grid points we were able to use in our models. Thus we created our own implementation in *C*, which increased speed by a factor of 300, and introduced multi-threading - another factor of 4 on a standard laptop. This allowed us to increase the number of grid points per layer and to move to spatial models that include *several* height layers. The first such spatial plots can be found in [7] and in our poster at this workshop. These plots track *large-scale atmospheric waves* in three dimensions and thus can now identify interactions also between several different height layers. New information that can be gained from these plots include (1) location of the maximum wave source and (2) preferred pathways of wave propagation.

Acknowledgments

This work was supported in part by NSF Climate and Large-Scale Dynamics (CLD) program Grant AGS-1147601 awarded to Yi Deng.

References

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

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

[3] Colombo, D., and M.H. Maathuis, *Order-independent constraint-based causal structure learning,* (arXiv:1211.3295v2), 2013.

[4] Chu, T. D. Danks and C. Glymour, *Data Driven Methods for Nonlinear Granger Causality: Climate Teleconnection Mechanisms*, Tech. Rep. CMU-PHIL-171, Dep. of Philos., Carnegie Mellon Univ., Pittsburgh, Pa., 2005.

[5] Ebert-Uphoff, I. and Y. Deng, *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.

[6] Deng, Y., and Ebert-Uphoff, I., *Weakening of Atmospheric Information Flow in a Warming Climate in the Community Climate System Model*, Geophysical Research Letters, 7 pages, doi: 10.1002/2013GL058646, 2014.

[7] Ebert-Uphoff, I. and Y. Deng, *High Efficiency Implementation of PC and PC stable Algorithms Yields Three-Dimensional Graphs of Information Flow for the Earth' Atmosphere*, Tech. Rep. CSU-ECE-2014-1, Dept. of Electrical and Comp. Engineering, Colorado State University, Sept 3, 2014. Available at http://hdl.handle.net/10217/83709.