CAUSAL DISCOVERY IN THE PRESENCE OF CONFOUNDING LATENT VARIABLES FOR CLIMATE SCIENCE

Savini Samarasinghe[1], Elizabeth A. Barnes[2] and Imme Ebert-Uphoff[1]

Abstract—**Identifying causal interactions in the presence of hidden common causes (latent confounders) is a well known challenge for causal discovery in climate science applications. In this exploratory investigation, we look into the suitability of a temporal extension of the Fast Causal Inference (FCI) algorithm, which is designed to identify latent confounders, for applications related to climate. The results from the initial experiments on synthetic and real world data show potential of this method being useful for climate applications. However, some apparent limitations of this method also indicate the necessity of a detailed suitability analysis prior to application.**

I. INTRODUCTION

The Earth is a complex system with many physical processes causally interacting with each other across space and time. Discovering these causal interactions can help to get a deeper understanding of the mechanisms governing the Earth's climate, which (for example) can be useful to create skillful climate/weather prediction models, to find the strongest causal pathways linking responses to climate change, and to anticipate the effects from altered atmospheric and oceanic flow patterns that may result in changing frequencies of extreme weather events. The most reliable approach to identify causal interactions is through an intervention study. However, as the ability to intervene in the climate system is extremely limited we often need to resort to observational studies. As causal interactions cannot be proved through an observational study, mainly due to the potential existence of latent variables, any interaction identified will only be a hypothesis of a *potential* causal interaction that needs to be further evaluated by climate scientists.

One popular approach to causal inference in climate science is based on *Granger causality* [1], [2]. A more recent approach, which is the one considered here, is based on *Pearl causality* [3], [4], which represents the causal structure among climate variables as probabilistic graphical models [5], [6], [7]. In these models, the climate variables are presented as vertices of a graph and the causal interactions are represented as graph edges. Constraint-based structure learning methods, such as the *PC* [4] and *PC stable* [8] algorithms, have been used to infer causal hypotheses in the climate [6], [7], using a temporal extension [5], [9] of the original method to be able to model delayed causal connections.

One of the biggest limitations of standard methods is that they assume causal sufficiency. That is, they assume that there are no hidden common causes (latent confounders), i.e. that every common cause of any two or more variables is already included in the model. Causal sufficiency is a very strong assumption that is often violated in climate applications. Firstly, we can only make an educated guess of the variables that need to be included in the model, resulting in a high risk of missing hidden common causes. Furthermore, it may be impossible to measure the exact variable of interest and we may need to resort to proxies of those unmeasured variables. The results from these methods must therefore always be interpreted with caution - namely, any connection found can be a true causal connection, be due to a latent variable, or both.

On the Granger causality side there have been some first efforts to infer causal structure in the presence of hidden variables in climate settings [10], [11]. On the Pearl causality side, algorithms for inferring hidden common causes exist, but to the best of our knowledge none of them have been tested for climate applications and their suitability is yet to be determined. Primary algorithms are the Fast Causal Inference (FCI) algorithm [12], [4], and tsFCI [13], which is an extension of

Corresponding author: Imme Ebert-Uphoff, iebert@colostate.edu. [1]Electrical & Computer Engineering, Colorado State University, Fort Collins, CO, USA. [2]Dept. of Atmospheric Science, Colorado State University, Fort Collins, CO, USA.

FCI for time-series data that enforces that the identified causal structure is time invariant. In this initial study, we choose to focus on a temporal extension of FCI, rather than tsFCI. The reason is that, as will be seen in Fig. 4, we sometimes encounter intermittent edges that are present in some time steps but not in others, and those cannot be easily identified as such using tsFCI.

II. Fast Causal Inference (FCI) algorithm

FCI is a constraint-based structure-learning algorithm that is expected to give asymptotically correct information about the causal structure even in the presence of hidden common causes [4]. It builds on the classic PC algorithm, but it does not assume a causally sufficient model (unlike PC). Following similar steps to PC, FCI first represents all the observed variables as nodes in a fully connected undirected graph, where each variable is initially assumed to interact with every other variable. Then it uses conditional independence tests to remove as many edges as possible. An edge between two nodes X and Y is removed whenever the two nodes are independent conditioned on any subset of vertices adjacent to X or Y (aka a conditioning set), thus indicating that there is no direct connection between the two nodes. The conditioning sets that lead to the removal of edges are stored and used in the orientation phase of the algorithm. Once the initial skeleton is uncovered, the first step is to orient triples of variables adjacent in the format X – Y – Z (i.e., X is adjacent to Y and Y is adjacent to Z, but X is not adjacent to Z). These edges are oriented as X $\rightarrow$ Y $\leftarrow$ Z (a V-structure) if Y is not in the conditioning set of X and Z. In contrast to PC, the FCI algorithm goes on to further refine the initial skeleton to identify and replace some of the spurious direct interactions that are due to previously unmodeled latent variables. The refined final skeleton is then reoriented using V-structures and additional orientation rules [4], [14], [15]. The graphical model produced by FCI is called a partially oriented induced path graph or a partial ancestral graph [15]. In addition to the uni-directed edges used in PC, this graphical model uses new arrow symbols to express the presence of latent confounders. For example a bi-directed edge between two variables, X $\longleftrightarrow$ Y, indicates that there is a latent confounder of X and Y. Whereas, a circular symbol represents ambiguity. X o$\rightarrow$ Y indicates either X $\rightarrow$ Y or X $\longleftrightarrow$ Y (or both). See [4] for details.

III. Examples

As a first test for the suitability of FCI for climate applications, we carry out a few initial experiments. We use a temporal extension of FCI, as implemented in TETRAD [16], [17], a free software package.

A. Synthetic Example

We first simulate the time series of two observed variables, X and Z, in the presence of a latent confounder, Y. Fig. 1 shows the true causal structure of X, Y and Z.

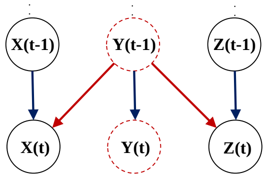

Fig. 1. The true causal structure of X, Y and Z, shows that X and Z are connected solely through the latent variable, Y. Throughout this paper a dashed circle indicates a latent (unobserved) variable, while a solid circle indicates an observed variable.

The time series for this example are generated based on Eq. (1)–(3), where e_x, e_y and e_z are standard normal errors:

$$\begin{aligned} \text{latent: } y(t) &= 0.4y(t–1) + 0.1e_y(t), && (1)\\ \text{obs: } x(t) &= 0.4x(t–1) + a_1 y(t–1) + 0.1e_x(t), && (2)\\ \text{obs: } z(t) &= 0.4z(t–1) + b_1 y(t–1) + 0.1e_z(t), && (3) \end{aligned}$$

Each auto-correlation term above has a parameter of 0.4, and noise of 0.1, while for the cross-connection parameters, (a_1, b_1), we consider three cases:

Case 1: $a_1 = b_1 = 0.3$ (medium strength)
Case 2: $a_1 = b_1 = 0.4$ (as strong as auto-corr.)
Case 3: $a_1 = b_1 = 0.1$ (as weak as noise)

For each case we now seek to recover the causal structure based on the simulated data for only X and Z. We use both the PC stable and FCI algorithms along with conditional independence tests based on Fisher's Z-test on partial correlation. We use a sample size of 5,000. (Unless stated otherwise, we use a statistical significance level $\alpha = 5 \times 10^{-5}$ throughout this paper.) The PC stable algorithm assumes that there are no confounding latent variables, which makes it impossible to infer the correct relationship between X and Z, while FCI has that capability.

In Case 1 (Fig. 2), FCI correctly identifies the hidden common cause between X and Z. These results are robust for a range of statistical significance values $\alpha \in [5 \times 10^{-5}, 0.01]$. In contrast, PC stable incorrectly identifies direct connections between X and Z.

In Case 2, FCI identifies the correct causal structure at a small α value, $\alpha = 5 \times 10^{-5}$, while PC stable fails to do so (results not shown). However, it appears that

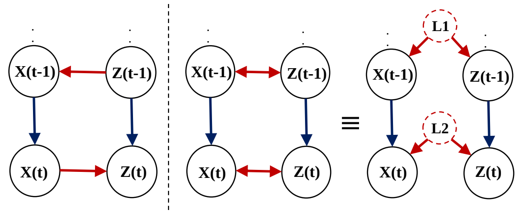

Fig. 2. Case 1 results for PC stable (left) and FCI (center). The bi-directed edges between X and Z in the FCI results indicate the presence of one or more latent confounders. A sample scenario is presented using latent confounders L1 and L2 (right).

the FCI results for Case 2 are quite sensitive to the α value. For example, at $\alpha = 0.01$ FCI also indicates incorrect direct interactions between X and Z. Therefore, identifying the best α values for different sample sizes can be a challenge in practical applications.

As shown in Fig. 3 (Case 3), both PC and FCI fail to identify the hidden common cause when the the signal-to-noise ratio of the hidden component is small. This is not too surprising, as weak signals are always difficult to pick up in causal discovery.

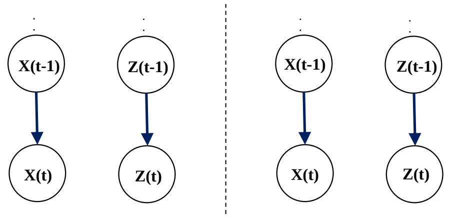

Fig. 3. Case 3 results for PC stable (left) and FCI (right).

In summary, FCI performs well in this example for the medium strength case, and it is neither surprising, nor overly concerning, that FCI performs poorly for the weak strength case. It is surprising, however, that FCI results lack robustness for the strong case and this case thus requires further study.

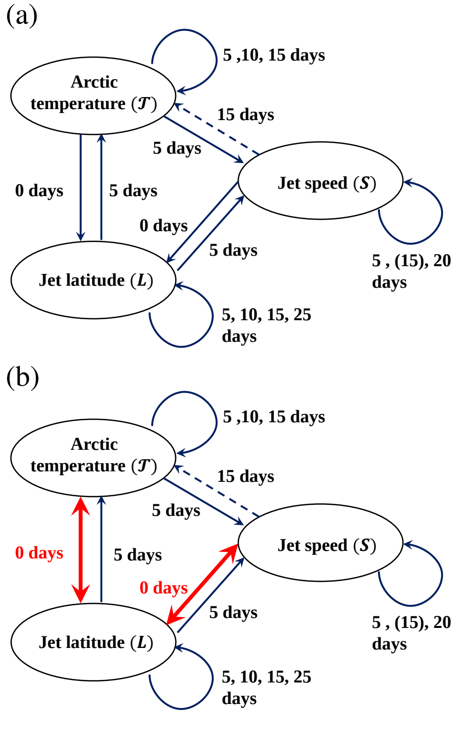

Fig. 4. Summarized potential causal structures identified by (a) PC stable and (b) FCI. The time lags of the interactions are shown along the edges. Dashed arrows here indicate interactions that are not completely consistent across time (either weak or intermittent). The red bi-directed edges in the FCI results indicate the presence of one or more latent confounders.

B. Application to Observed Climate Data

Next we revisit an application we previously studied using the PC stable algorithm [18]. Namely, we look at the causal links identified by PC stable and FCI on data representing the Arctic temperature and mid-latitude jet stream. While there is ample evidence of connections between these two variables [18], [19], it is also possible that hidden common causes are present, for example, due to stratospheric processes [20], [21]. Here we use (1) 850hPa Arctic temperature averaged over 70^{o}N-90^{o}N ($\mathcal{T}$), (2) jet latitude ($\mathcal{L}$) and (3) jet speed ($\mathcal{S}$) in the North Pacific. We analyze daily data of the boreal winter months (Dec.-Feb.) from the Community Earth System Model–Large Ensemble (CESM-LE, [22]) for years 402 to 2200. The seasonal cycle is removed from the data to weaken the impact of time acting as a common cause of the variables. For more information of the data and preprocessing, see [18]. The results for PC stable and FCI are shown in Figures 4(a) and (b), respectively. These results show **summarized** inferred causal structures identified by the algorithms. The results show that the auto-correlations of $\mathcal{L}$, $\mathcal{S}$ and $\mathcal{T}$, the causal links between $\mathcal{T}$ and $\mathcal{S}$, and the causal links from $\mathcal{L}$ to $\mathcal{T}$ and $\mathcal{L}$ to $\mathcal{S}$, occur in both PC stable and FCI. However, the results also show the direct instantaneous edges - instantaneous here means fewer than 5 days - from $\mathcal{T}$ to $\mathcal{L}$ and $\mathcal{S}$ to $\mathcal{L}$ identified in PC stable being replaced with bi-directed edges in FCI, indicating that those instantaneous connections are due to a hidden common cause, and that there is no direct instantaneous causal interaction between the two variables. We do not attempt to interpret these results at this initial stage, as we need to further evaluate the reliability and limitations of FCI. However, the fact that FCI identifies *just a few bi-directed edges*, and otherwise confirms the direct connections identified by PC (rather than indicating *lots* of bi-directed edges), indicates that this type of method could be of use to the climate science community, if reliability can be shown.

IV. CONCLUSIONS AND FUTURE WORK

Standard tools to infer causality in climate science assume that the modeled systems are causally sufficient. This assumption is often violated. We make an initial effort to study the suitability of the FCI algorithm, which is systematically designed to handle hidden common causes, for climate applications. The case studies suggest that FCI has potential to be useful. However, limitations such as lack of robustness and reliability are also apparent. As future work we propose to test FCI, tsFCI, and Granger-based methods, for more synthetic case studies, as well as for geophysical data for which the causal signatures are already known. A detailed analysis will help determine the usefulness of these algorithms for the the climate science domain. In addition, we plan to explore using different conditional independence tests for FCI, such as KCI [23] or CCI [24], to account for the non-linearities of relationships and non-Gaussianity of data.

ACKNOWLEDGMENTS

We thank Marie C. McGraw at Colorado State University for providing climate data. We thank Dr. Clark Glymour and Dr. Joseph Ramsey at Carnegie Mellon University (CMU) for their valuable insights, and the Center of Causal Discovery at CMU for making the TETRAD software available to the public. This work was supported by NSF grants AGS-1445978 (Ebert-Uphoff) and AGS-1545675 (Barnes).

REFERENCES

[1] C. Granger, "Testing for causality: A personal viewpoint," *Journal of Economic Dynamics and Control*, vol. 2, no. 1, pp. 329–352, 1980.

[2] A. Arnold, Y. Liu, and N. Abe, "Temporal causal modeling with graphical granger methods," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 66–75, ACM, 2007.

[3] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufman Publishers, revised second printing ed., 1988.

[4] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search.* MIT Press, 2nd ed., 2000.

[5] T. Chu, D. Danks, and C. Glymour, "Data driven methods for nonlinear granger causality: Climate teleconnection mechanisms," tech. rep., Carnegie Mellon University, Department of Philosophy, 2005.

[6] I. Ebert-Uphoff and Y. Deng, "Causal discovery for climate research using graphical models," *Journal of Climate*, vol. 25, no. 17, pp. 5648–5665, 2012.

[7] J. Runge, V. Petoukhov, J. F. Donges, J. Hlinka, N. Jajcay, M. Vejmelka, D. Hartman, N. Marwan, M. Paluš, and J. Kurths, "Identifying causal gateways and mediators in complex spatio-temporal systems," *Nature communications*, vol. 6, 2015.

[8] D. Colombo and M. H. Maathuis, "Order-independent constraint-based causal structure learning," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3741–3782, 2014.

[9] T. Chu and C. Glymour, "Search for additive nonlinear time series causal models," *Journal of Machine Learning Research*, vol. 9, pp. 967–991, 2009.

[10] M. T. Bahadori and Y. Liu, "Granger causality analysis with hidden variables in climate science applications," in *Climate Informatics workshop (CI 2011)*, 2011.

[11] M. T. Bahadori and Y. Liu, "An examination of practical Granger causality inference," *SIAM International Conference on Data Mining*, pp. 467–475, 2013.

[12] P. Spirtes, C. Meek, and T. Richardson, "Causal inference in the presence of latent variables and selection bias," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, (San Francisco, CA, USA), pp. 499–506, Morgan Kaufmann Publishers Inc., 1995.

[13] D. Entner and P. O. Hoyer, "On causal discovery from time series data using fci," 2010.

[14] J. Zhang, "On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias," *Artificial Intelligence*, vol. 172, no. 16, pp. 1873 – 1896, 2008.

[15] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson, "Learning high-dimensional directed acyclic graphs with latent and selection variables," *Ann. Statist.*, vol. 40, pp. 294–321, 02 2012.

[16] C. Glymour, R. Scheines, P. Spirtes, and J. Ramsey, "The Tetrad project - graphical causal models." http://www.phil.cmu.edu/tetrad/index.html, as of June 2018.

[17] J. D. Ramsey and D. Malinsky, "Comparing the performance of graphical structure learning algorithms with tetrad," *arXiv preprint arXiv:1607.08110*, 2016.

[18] S. Samarasinghe, M. C. McGraw, E. A. Barnes, and I. Ebert-Uphoff, "A study of links between the Arctic and the midlatitude jet-stream using Granger and Pearl causality," *Environmetrics*, 2018 (in press).

[19] E. A. Barnes and I. R. Simpson, "Seasonal Sensitivity of the Northern Hemisphere Jet Streams to Arctic Temperatures on Subseasonal Time Scales," *Journal of climate*, vol. 30, pp. 10117–10137, Sept. 2017.

[20] J. Cohen, J. A. Screen, J. C. Furtado, M. Barlow, D. Whittleston, D. Coumou, J. Francis, K. Dethloff, D. Entekhabi, J. Overland, and J. Jones, "Recent Arctic amplification and extreme mid-latitude weather," *Nature geoscience*, vol. 7, p. 627, Aug. 2014.

[21] L. Sun, C. Deser, and R. A. Tomas, "Mechanisms of Stratospheric and Tropospheric Circulation Response to Projected Arctic Sea Ice Loss," *Journal of climate*, vol. 28, pp. 7824–7845, Aug. 2015.

[22] J. Kay, C. Deser, A. Phillips, A. Mai, C. Hannay, G. Strand, J. Arblaster, S. Bates, G. Danabasoglu, J. Edwards, M. Holland, P. Kushner, J.-F. Lamarque, D. Lawrence, K. Lindsay, A. Middleton, E. Munoz, R. Neale, K. Oleson, L. Polvani, and M. Vertenstein, "The Community Earth System Model (CESM) Large Ensemble project: A community resource for studying climate change in the presence of internal climate variability," *Bull. Amer. Meteorol. Soc.*, vol. 96, pp. 1333–1349, 2015.

[23] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf, "Kernel-based conditional independence test and application in causal discovery," in *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, (Arlington, Virginia, United States), pp. 804–813, AUAI Press, 2011.

[24] J. D. Ramsey, "A scalable conditional independence test for nonlinear, non-gaussian data," *CoRR*, vol. abs/1401.5031, 2014.