Comment on "Estimating causal networks in biosphere–atmosphere interaction with the PCMCI approach"

Jarmo Mäkelä[1], Laila Melkas[1], Ivan Mammarella[2], Tuomo Nieminen[2,3], Suyog Chandramouli[1], Rafael Savvides[1], and Kai Puolamäki[1,2]

[1]Department of Computer Science, P.O. Box 68, FI-00014 University of Helsinki, Helsinki, Finland
[2]Institute for Atmospheric and Earth System Research / Physics, P.O. Box 64, FI-00014 University of Helsinki, Helsinki, Finland
[3]Institute for Atmospheric and Earth System Research / Forest Sciences, P.O. Box 27, FI-00014 University of Helsinki, Helsinki, Finland

Correspondence: Jarmo Mäkelä (jarmo.makela@helsinki.fi)

Abstract. This is a comment on "Estimating causal networks in biosphere–atmosphere interaction with the PCMCI approach" by Krich et al., Biogeosciences, 17, 1033–1061, 2020, which gives a good introduction to causal discovery, but confines the scope by investigating the outcome of a single algorithm. In this comment, we argue that the outputs of causal discovery algorithms should not usually be considered as end results but starting points and hypothesis for further study. We illustrate how not only different algorithms, but also different initial states and prior information of possible causal model structures, affect the outcome. We ~~demonstrate~~ provide a proof-of-concept demonstration of how to incorporate expert domain knowledge with causal structure discovery and remark on how to detect and take into account ~~overfitting~~ over-fitting and concept drift.

1 Main text

In a recent paper **?** tested and applied a newly developed PCMCI algorithm (**??**) in order to detect causal links in geophysical data. The algorithm is used on flux tower eddy covariance data and related meteorological measurements of six variables in order to detect which variables can be seen to steer the behaviour of others. The paper can be viewed as a proof-of-concept and is a good introduction to causality and underlying problems, given the novelty of applying these types of methods to better understand biosphere-atmosphere interactions. However, we feel that ~~contribution of Krich et al. covers only one part of~~ the approach in **?**—together with much of the other related work (**?**)—is limited in its contribution to the practical application of causal ~~discovery~~ structure discovery (CSD) algorithms. There were items that in our opinion are significant ~~for practical application of such causal discovery methods and which~~ that were only briefly mentioned or not at all addressed in ~~the paper by ?~~(**?**). These are:

- ~~The~~ Different CSD algorithms may produce distinct outcomes (models) ~~of causal structure discovery (CSD) algorithms are, in many cases, interchangeable: it is very~~ when operating on the same data. It is often difficult to identify the "correct" ~~model~~ among these models, purely based on data.

- The choice of initial state ~~affects the final model~~(known structures) affects the behaviour and output of CSD algorithms. Due to their setup, **?** employed an empty graph, but other choices are also possible.
- Utilising the knowledge of the domain experts and user interaction can be used to improve the models.
- ~~Overfitting~~ Over-fitting and concept drift ~~. Overfitting~~ were addressed in (**?**) via the use of Akaike information criteria (AIC) (**?**) but as these issues are central to any model selection or development we want to stress their importance. Over-fitting means that the analysis relies too much on the training data. Usually this happens when the amount of data is too small, resulting the causal model fitting to noise. Concept drift means that the underlying data distribution changes, rendering the causal model ~~irrelevant~~obsolete. An example of a concept drift is that a model trained on a certain location may not describe relations in another location; it is important to be able to take this ~~phenomena~~ phenomenon into account.

These comments are based on our recent workshop paper in the KDD 2021 conference (**?**). Since many experts in Earth system sciences are not likely to follow said conference, we wanted to convey the main findings via this reply to **?** as it also originally inspired us to explore the topic. In short, ~~our workshop paper presents a procedure on how to utilise prior knowledge of the domain experts in finding causal structure discovery (CSD) models and how a user might incorporate this knowledge with CSD algorithms . This knowledgecan be characterised by a prior distribution over all possible causal structures.We use both synthetic data as well as flux tower eddy covariance variables – same variables as in **?**) – measured at the SMEAR II station at Hyytiälä, Finland (**?**). We simulate the user 's choices~~ we try to find a model (directed acyclic graph) that best reflects the data, domain knowledge and user beliefs. Here we explore the behaviour of several CSD algorithms on both synthetic and real data and demonstrate how to incorporate prior knowledge and user interactions to this process. Before examining these topics in more detail, we present the underlying workflow in our approach:

1. Input domain knowledge (if any) as probabilities of known structures in the data.
2. Apply CSD algorithms to the data with the domain knowledge.
3. Choose a model from those provided by the algorithms, e.g. what the user regards as the best model in terms of their background knowledge and model score.
4. Apply user interactions to the chosen model. We have substituted an actual user with a greedy search ~~from the neighboring states~~ algorithm that examines the neighbouring models (one edit away) of the current ~~model. By "neighbourhood" we mean the models that can be reached from the current model by simple edits and "greedy" we mean that the user always chooses the best model from the neighbourhood of the current model,~~ one and chooses the best, in terms of model score.
5. Check the validity of the chosen model. We use cross-validation to detect over-fitting and concept drift due to its simplicity but other methods, e.g., AIC are possible as well.

The presented approach is Bayesian in nature and can be formulated as building a probabilistic model of the data. The aim is to find (locally) optimal model and ~~this~~ as such, we assume that the domain knowledge can be characterised by a prior distribution over all possible causal structures (known features in the graph and confidence in that knowledge). Similarly, in our

simulation, the user will have confidence, represented by parameter k, in certain structures between any pair of variables (A $\rightarrow$ B, B $\rightarrow$ A or no link). The user (in our case greedy search) is presented with options for simple edits and how these edits would affect the model score. This process is iterated, until the current model is at least as good as any of the neighbours – see **?** for details. The outcomes are also compared to a model produced by actual domain experts (IM and TN). The takeaway message is that instead of using ~~expert~~ domain knowledge to merely quality check the final model produced by a CSD algorithm, the prior knowledge should be incorporated into the ~~process~~causal structure discovery process. The CSD methods we have used are PC-Stable with two significance levels 0.1 and 0.01, GES, and ICA-based LiNGAM. We use both synthetic data as well as flux tower eddy covariance variables – same variables as in **?** – measured at the SMEAR II station at Hyytiälä, Finland (**?**).

All presented numerical analyses use synthetic data, which enables us to know the "true model". This data is created by generating a random (directed acyclic) graph and sampling it with random edge weights to produce data sets of varying size. Each graph is generated with a sparsity of 0.3, which means that each pair of variables has an edge between them with a probability of 0.3. All edges are oriented away from the first variable and in the same order the variables are defined, which ensures acyclicity. Noise from either uniform distribution (-0.01, 0.01) or Gaussian with a standard deviation of 0.01 was added for each variable (for each variable the choice of the distribution was random). The reason for including both types of noise distributions is to create data sets which almost follow assumptions made by the algorithms while still breaking some of them. All of the algorithms we use in the experiments assume linearity but, additionally, PC-Stable and GES assume Gaussianity of noise and LiNGAM assumes non-Gaussianity.

2 Differences in CSD algorithms

While **?** have focused on PCMCI, it is worthwhile to note that different CSD algorithms have varied outputs (models) for the same input data ~~(?)~~ (**??**) since each algorithm operates differently and makes different assumptions about the underlying data (Fig. 1). Additionally, even if the modelling assumptions in the causal discovery process are correct, insufficient or biased data may result in skewed results. Therefore, the model gained from any one of these algorithms should not be viewed as the end result, but rather a starting point for further analysis. Often it is not clear, which among the discovered models is the "best", although we can argue that some of them are more plausible (**?**), given the expert's knowledge. In some algorithms, inputting this prior knowledge (e.g., probabilities of certain structures) is possible, but the ability to iteratively refine this background knowledge during the data analysis process nor the possibility to express uncertainty in the prior information have not been built in. These caveats hinder the usability of many CSD algorithms.

3 The choice of initial state

As different algorithms produce different models, so does the choice of initial state affects the outcome. These states can be, for example, empty graphs (as in **?**), states produced by sampling methods, or states that reflect certain ~~expert~~ domain knowledge. Depending on the choice of initial state and on how uncertain the prior information is, different locally optimal models that

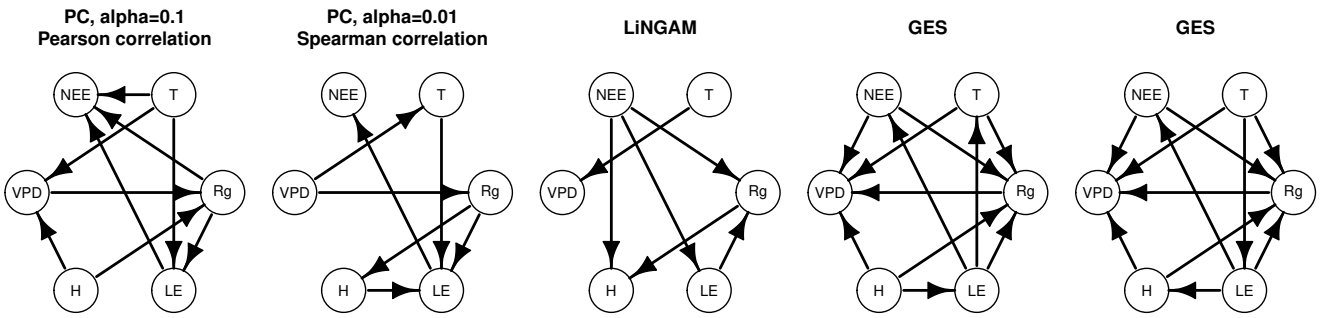


Figure 1. Different algorithms produce different causal graphs for the same data. PC algorithm is started from a full graph, LiNGAM has no defined initial graph and we started GES from an empty graph. GES produces (multiple) graphs with indistinguishable conditional dependency relationships.

fit the data may be found. Intuitively, it would be interesting to have a set of initial states that would cover all local optima, which could give rise to a global maximum-a-posteriori (MAP) solution. The underlying problem here would be to find a representative set of starting points for the exploration.

We demonstrate the combined effect of utilising multiple initial states and different levels of prior knowledge (k) with synthetic data (Fig. 2). The initial states are generated by four different CSD algorithms and are complemented by an empty graph and the correct model, which we know as the data is synthetic. ~~The user knowledge is reflected by parameter~~ There are three possible causal states for a pair of model variables A and B: $A \rightarrow B$, $A \leftarrow B$, and no causal connection between the two. The user knows the correct state between each pair of variables with the probability of k, ~~where $k = 1$ indicates that user has full knowledge of the causal structure and~~ e.g. $k = 1/3$ ~~means that the user has no prior information (see ? for details; values of $k > 1/2$ lead to near constant results).~~ means flat prior and $k = 1/2$ means that user knows the true states of the pairs with a probability of $1/2$. In these simulations, the level of prior knowledge $k \in [1/3, 1/2]$ We do not take into account wrong information ($k < 1/3$), and values above $1/2$ do not produce interesting results as such high certainty leads to near-constant results.

The structural Hamming distance (SHD) ~~indicates how many modifications to a model have to be made in order to end up with another model~~between two models indicates the minimum number of edge modifications (simple edits) required to transform either of the models into the other one. Even with a small amount of prior information, the end result after user interactions (greedy search) becomes much more stable – the spread of SHD diminishes as k increases (Fig. 2).

4 Utilising ~~expert~~ domain knowledge and user interactions

The knowledge of the domain experts is classically used to provide suitable initial states for the CSD algorithms or to quality check the outcomes, but this knowledge should also be used to steer the CSD processes via user interactions and to allow reassessment of both user's own prior knowledge and related uncertainty as well as the algorithm process. When this knowledge

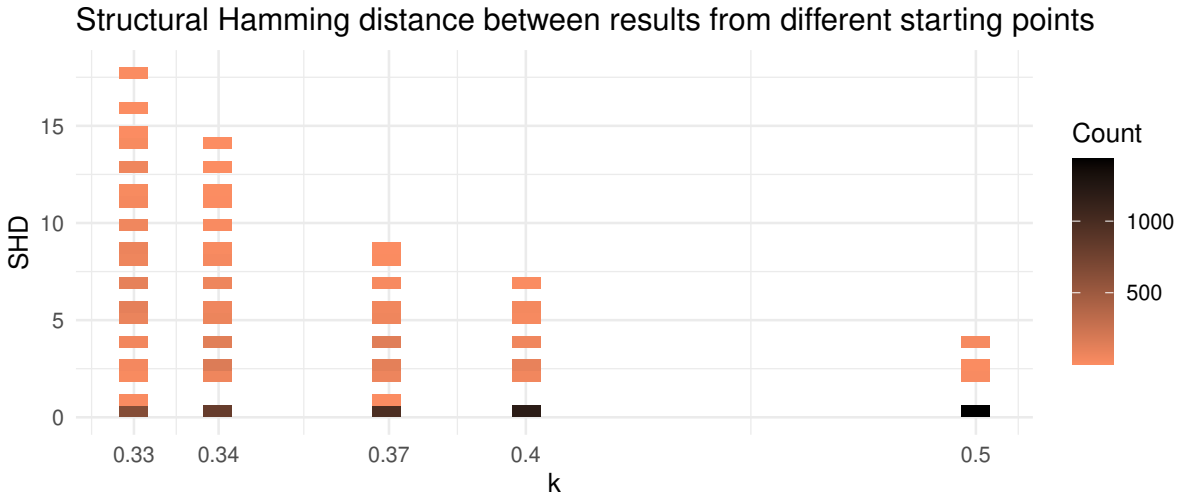


Figure 2. Pairwise structural Hamming distances when running analysis on the same data starting from different initial models. Variance in the distances show that the final model is affected by choice of initial model. Additionally, the spread of distances decreases rapidly with increasing prior knowledge.

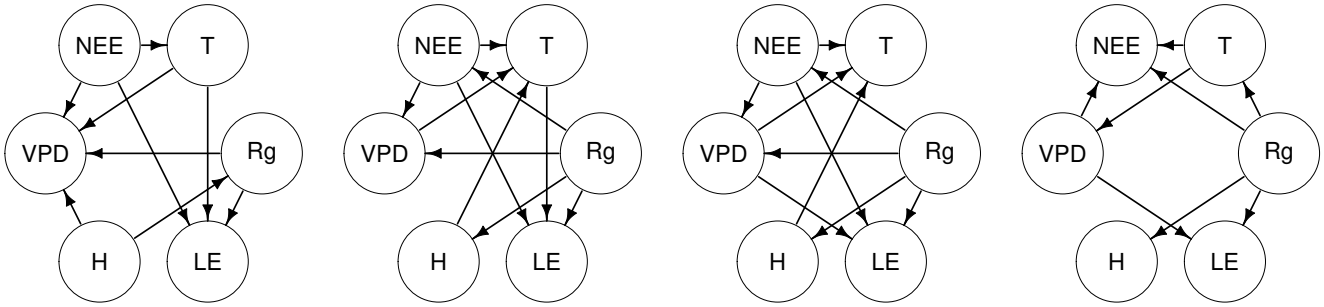
is disregarded and the data is blindly trusted, any CSD algorithm or user (e.g., our greedy search) can uncover erroneous connections and miss relevant ones (Fig. 3). For example, the expert model (d) identifies four direct and well-established causal links from downwelling shortwave radiation (Rg) to latent and sensible heat fluxes (LE,H), temperature (T) and net ecosystem exchange (NEE). Two of these links (T and NEE) are missing from the best scoring model among the CSD algorithms (a), which also erroneously asserts that H is a driving force behind Rg. Both user models (b,c) find a new unrealistic link from Rg to vapour pressure deficit (VPD) and indicate that Rg is affecting T only indirectly through NEE.

5 ~~Overfitting and concept drift~~

~~Overfitting the model to the data is a common problem in statistical modelling, but to the best of our knowledge this problem has not been addressed in the context of CSD. In **?** we demonstrate how to detect overfitting using k-fold blocked cross-validation (**?**). The same method is also applicable in detecting concept drift, which we induced by including a set of measurements taken in August 2015 to calibration data containing measurements taken in April 2013–2015 – this violates causal stationarity stemming from seasonality.~~

5 Concluding remarks

Novel CSD algorithms, and more generally many machine learning methods, offer new insights in Earth system sciences. We argue that combining these methods with already abundant knowledge of the domain experts ~~will~~ may yield more robust results



(a) Initial model from algorithms, produced by PC-stable.

(b) Final model starting from (a) with greedy search.

(c) Final model starting from an empty graph with greedy search.

(d) Expert model, produced by IM and TN.

Figure 3. The user (greedy search) finds slightly different models (b,c) whether we start the search from the best scoring model among our CSD algorithms (a) or an empty graph. The underlying causal structures were given a uniform prior ($k = 1/3$). Also shown is the expert model, produced before these experiments. The SHD from the expert model to (a),(b) and (c) are ten, seven and five.

and provide promising questions for future research. We also argue that while there are plethora of CSD algorithms that has been applied in earth sciences the question of how to use them in practice is still open. We have briefly presented here ~~one fairly simple~~ a fairly simple proof-of-concept approach as how to achieve this, demonstrated its effectiveness and highlighted some pitfalls – we direct anyone interested in a more detailed presentation to see **?**, where we have also demonstrated how to detect over-fitting and concept drift, two common problems in statistical modelling, using k-fold blocked cross-validation (**?**) . Hopefully, ~~this~~ the work presented here will encourage developers to implement and study further interactive workflows. ~~We direct anyone interested in a more detailed presentation to see **?**.~~

Author contributions. JM prepared the comment, while LM ran the simulations and prepared the KDD manuscript under supervision of KP. IM and TN provided the domain expert knowledge and the expert model and together with SC and RS participated in producing this paper.

Competing interests. The authors declare that they have no competing of interests.

Acknowledgements. We thank Helsinki Institute for Information Technology, Future Makers Funding Program, and Finnish Center for Artificial Intelligence for support.