Biogeosciences Discussions

Biogeosciences Discuss., referee comment RC2
https://doi.org/10.5194/bg-2021-231-RC2, 2021

Comment on bg-2021-231

Michele Ronco (Referee)

Referee comment on "Technical note: Incorporating expert domain knowledge into causal structure discovery workflows" by Jarmo Mäkelä et al., Biogeosciences Discuss., https://doi.org/10.5194/bg-2021-231-RC2, 2021

The paper raises a warning with regard to the blind usage of causal structure discovery (CSD) algorithms and correctly suggests that CSD should be seen more as a useful guidance in the understanding and modeling of multivariate systems rather than as the final outcome of analysis. In particular, the authors identify four main weaknesses of CSD: different causality methods produce different results, the outcome depends on the initial graph, domain knowledge is not taken into account, overfitting and performance drops due to distribution shifts are usually ignored. All these concerns are of great interest and yet have not been sufficiently explored in the literature. For this reason the paper could have an impact on how CSD are deployed for scientific discovery. However, the authors do not satisfactorily develop the ideas presented in the introduction or at least do not provide enough details. Even for a first contribution towards the interesting interplay between domain knowledge and CSD I think that the authors should make an extra effort and expand further the content of the paper. Find below more detailed comments per section.

1. Main text: what do you mean by "outcomes (models) of causal structure discovery (CSD) algorithms are, in many cases, interchangeable"? are you referring to the fact that different CSD may produce distinct causal graph over the same data? Perhaps explain or rephrase this sentence. How exactly does a greeady search over models help injecting previous knowledge into CSD?

2. Differences in CSD algorithms: the authors should provide a concrete example of a case in which different CSD produce different results (maybe taken from the literature). In its current form, section 2 is too generic and does not really add much information besides what is already stated in the introduction.

3. The choice of initial state: this is one of the main parts of the paper where the autorhs describe their first experiment on how the initial graph affects the result. Interestingly, they show that increasing the level of prior knowledge the outcomes gradually converge. However, more details on how the experiment is performed should be given. What are the syntetic data used? Which are the four CSD methods? How is the parameter "k" (encoding the prior) used to generate initial states?

4. Utilising expert knowledge and user interactions: here the authors suggest that not only user knowledge should be used for defining an optimal starting point (or graph) but also within a sort of interaction scheme between CSD and the user. What is the kind of procedure the authors have in mind and how is it used in the example? Is it a Bayesian-inspired approach with prior/CSD/posterior? More specifically, how is the "expert model" in (d) produced? Which CSD have been applied to generate (a) and to end up in (c)?

5. Overfitting and concept drift: none of these two issues are developed in the section. I would recommend to either strengthen Section 5 with some experiment or simply move the observations in Sec5 to the conclusions.

In summary, the paper addresses some crucial aspects regarding the application of CSD to scientific discovery. However, the interesing ideas presented are not developed in enough details. If the authors succeed in adequately expand their work then I would certainly recommend publication since the topic is of interest and can have potential impact in the forthcoming literature.