

Review of "Benchmarking Conditioners in Liang–Kleeman Information Flow: Application to Land–Atmosphere Interactions"

The revisions of the manuscript improved its clarity and limitations are now stated more explicitly. In particular, the modifications make it easier to understand the concepts used in the manuscript. However, some changes do not fully resolve the initial criticisms of the reviewers, and some new questions arise from the clarifications of the definition / interpretation of the concepts. Therefore, I recommend further revisions to the manuscript to make it suitable for publication in ESD.

General comments:

- Sect. 2.4 / Order of adding conditioners for computing the indices: It would be more convincing to report the mean and range (as uncertainty measure) over all permutations in the main manuscript instead of using one specific permutation (for example in Fig. 5).
- Definition and interpretation of Moderation Gain and Confounding Pressure: The authors provide some new interpretation of the indices in the revised discussion, but due to the importance of the indices, the interpretation should already be improved in the methods and results section. In particular, I find the interpretations in the methods section (and at times in the results) counterintuitive. ΔIF is positive when $T^{bi} > T^{mv}$ (Equ. 7-9). This is consistent with the figures. However, if this is the intended formulation, why do "High MG_m values identify conditioners that dampen the apparent driver influence in the bivariate signal."? Isn't it rather the other way around, that the bivariate signal overestimates the coupling for high MG values because it does not account for a conditioners? Similarly for the interpretation of confounding pressure: CP is positive when $T^{mv} > T^{bi}$, suggesting that the true IF (after accounting for conditioners) is higher than the estimate by T^{bi} and not that the "conditioner (m) suppresses inflated bivariate influence"? Related, I find the interpretation "Large CP_m values indicate strong confounding by the conditioner." confusing, because (a) it is stated before that the method cannot separate between confounding and mediation, and (b) I'm more used to using confounding for cases where ignoring a third variable leads to an overestimation of the true coupling, rather than an underestimation. Please also check the interpretations in l. 469-479 regarding the direction of the differences between bivariate and multivariate formulations. For example, Fig. 1c suggests to me that the coupling (IF) of $SM \rightarrow GPP$ is too low (dampened) when not conditioning on VPD (and T, SSR), but l. 479 states that the bivariate representation overstates the true conditioning.
- Sect. 3.4 / Interpretation of the ANOVA: Strictly speaking, the ANOVA doesn't identify specific values of drivers / conditioners that lead to large $|\Delta IF|$ values. It only identifies which (combination of) variable(s) can explain the variability of $|\Delta IF|$. While the interpretations in the section are ecologically plausible, a different model is required to quantify the direction and effect sizes of these influences on $|\Delta IF|$, for example a multiple linear regression model. Please also comment on the very strong influence of the three-way interactions on the explained variability, as according to Fig. S4 the direct influences and two-way interactions seem to explain essentially no $|\Delta IF|$ variability.
- Mediation / confounding / conditioning: conditioning is still not used consistently across the manuscript instead of either confounding or mediation, since according to the authors the LKIF can not separate distinguish between the two (this includes the index name "confounding pressure")

Specific comments

- l. 63-65: This sentence is still unnecessary vague and difficult to understand. It would be more informative to describe the main ideas / axioms behind the formalism. I also don't understand what the "contrast" is between the first and the second part of the sentence. While some other methods are mentioned in the revised manuscript, it would be helpful for readers to state more explicitly what makes the LKIF approach different from other causal inference methods and what limitations it overcomes.
- l. 124-135: Is this the same specification of the Kalman Filter as in Z24? What is the rationale for the window lengths (24 months and 90 months)? Since the optimal specification is likely depending on the length of the timestep, it would be helpful here to rather state the lengths in months and not in timesteps. In this context I also wonder how IF is estimated for the years 2004 - 2010 if date is used from 2004-2018 but the Kalman filter uses a 90 months look-back window. Are shorter windows used for the first years? If yes, could that explain why IF estimates are mostly near 0 at the start of the analysis period?
- l. 145-152: This is another example where more work to understand the interpretation of the methods would be very helpful because the description of the meaning of the IF stays very vague and hard to grasp. In particular "the likelihood of control" is rarely mentioned afterwards, but the IF is rather interpreted as a strength of control and not a likelihood. It would also be helpful to state more explicitly what is meant by increased vulnerability.
- l. 178-180: Does that preserve the spatial (and/or temporal) autocorrelation structure of the data? This would be necessary for the bootstrap to provide meaningfully superior results to the t-test which assumes independence of the samples.
- l. 249: How is $\varepsilon = 0.02$ selected? What percentile across all couplings is that (since this is the proposed threshold selection criterion in the brackets above)?
- l. 291: I find the reformulation of this sentence still unnecessarily difficult to understand. Doesn't the loss of significance simply indicate an overestimation of the directional coupling when not accounting for conditioners.
- l. 332-333: Does that indicate that conditioning improves the IF estimation in this situation and identifies a bias of the bivariate framework? It would be nice to state this more explicitly if this is meant.
- l. 350: What is meant by "feedback" here and is there a difference between coupling and feedback?
- l. 448-451: This difference seems to be only because of the larger overall variability of $|\Delta IF|$ for GPP than LAI. The relative explained variances of the factors are very similar (see Fig. S4).
- l. 469: I don't think "amplify" is a good wording here, because MG and CP are measuring opposing directions by which bivariate and multivariate estimates differ, but not opposing amplitudes. Please choose a more careful wording. If you indeed want to measure amplification / dampening, I would rather define MG and CP as multiplicative measures and not additive measures.
- Fig. 5c: CP is defined in the methods for individual conditioners, but no individual conditioner is mentioned here. Why?
- l. 595: Please rephrase to "Synthetic data experiments" instead of "Theoretical experiments" because this is just one example with synthetic data and not a theoretical proof.