Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., author comment AC2
https://doi.org/10.5194/hess-2021-445-AC2, 2021

Reply on RC3

Damien Delforge et al.

Author comment on "Detecting hydrological connectivity using causal inference from time series: synthetic and real karstic case studies" by Damien Delforge et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-445-AC2, 2021

1. General response

Dear reviewer, we would like to thank you for your attention to the preprint manuscript and the helpful comments you made. The major points you raised seem related to the lack of precision regarding the description of the statistical tests and concerns relative to the stability of the PCMCI-CMI method and, concurrently, its fair comparison to the other methods. Regarding the description of the tests, we briefly revised the methodological section to make their description more apparent. Regarding the stability issue of PCMCI-CMI and its comparison to other CIMs, we have chosen to clarify our point of view and use it as a basis for revising the discussion section of the manuscript.

Regarding the discussion, based on your remarks and those of the other reviewers, we will rewrite the section with improved references and comparisons to the literature on the following topics:

- a summary and appreciation of our results;
- a particular focus on the robust estimation of Conditional Mutual Information (CMI) concerning missing values, record length, dimensionality, the nature of the dependencies, or noise;
- and practical recommendations for the uses of causal inference methods and future research perspectives.

Concerning point 2 and the virtual experiment, you, as other reviewers, encourage us to extend the virtual experiment to study the effect of the sample size and/or the number of variables. Nevertheless, we have decided not to comply with this request for multiple reasons presented in the revised discussion, and in the responses below.

Thank you again for your contribution to this discussion,

On behalf of co-authors,

Damien Delforge

2. Major Comments

2.1 Major Comment 1

The lack of statistically significant links in the CMI method made more sense to me when I noticed the sparse dataset (e.g. fewer than 100 data points) that is available for the karst system analysis. In general, I think this method would not be expected to produce robust results given this amount of data, due to the high dimensional pdfs involved for information theory measures. With this, I think the interpretation should not be that CMI performed "worse" in some way, but that it has higher need for data length as a much higher dimensional approach, especially relative to the bivariate methods. Additionally, it seems like the bivariate methods actually do utilize the full amount of data available for any two variables, which means that the comparison is even less "fair" – it might be better to only consider the time window in which there is data available for all (or some, like the P1, P2, and P3 cases in Figures 5 and 6) variables. I think this aspect should be made more clear at the least, and possibly deserves a change in the time windows used for the comparison.

We thank the reviewer for his/her comment that helped clarify the paper. Since stability is a recurring point among other reviewers, we decided to write a section in the discussion. Our intention was not to convey that the PCMCI-CMI method is worse or bad in a general sense but in our real study case. Still, we fully agree that the message to be conveyed is not relative to a ranking of the method as better or worse but that the method has more important prerequisites in terms of sample size depending on the number of variables. Therefore, we will make sure that it is apparent in the revised discussion.

As you said, the bivariate methods use the full amount of data. We see the fair treatment of the time domain as an interesting discussion, which is also related to (1) the question that arises primarily in the presence of non-identically distributed missing data, "should we condition on all variables?", and (2) the problem of intermittent connectivity.

Our first point of view is to use the PCMCI without formulating any constraint on which variable is allowed to influence another or not. Hence, this attitude is very empirical as we expect the PCMCI and the data to speak for themselves. The second initial point of view is that we make the assumption of persistent connections in time between the variables, even if nonlinear, as opposed to intermittent connectivity. This assumption is supported by the fact that percolation is continuous, but of course not verified. Our research question is, therefore, straightforward and boils down to "Is there a connection between two variables", without any complement such as "in winter or in summer", or "when it is very wet".

With both viewpoints in mind, we consider two reasons why our approach may be unfair. First, with no regard to the hydrological problem, the methods are evaluated on different sample sizes, which is unfair if we aim at ranking them. Here, we may counterargue that the fact that bivariate methods allow for a more extensive cross-domain coverage is one of their advantages, which is not fair to dismiss. More generally, in the empirical generalization process of answering our question, "is there a connection between two variables?", it is unfair not to use all the available information to support the conclusion. Instead of harmonizing the temporal domain downwards, we prefer to discuss how to

increase the time-domain to avoid the stability and dimensionality problem. With missing values, the small size of the time domain is due to the fact that we test and condition each variable with respect to each variable. This is why focusing on P1, P2, P3 only has the effect of increasing the size of the cross-domain while removing the possibility that P1, P2, P3 influence each other. This brings us to the first aforementioned question (1) "should we condition on all variables?". Conceivably, when it comes to testing the connectivity between two points A or B within a system, conditioning on their past and the common driver, e.g.., evaporation and rainfall or effective precipitation alone, could be a dimensionally adequate and contained representation of the problem.

The other aspect that could be unfair when comparing causal graphs over different time domains was pointed out by reviewer 1. This is to interpret dissimilarities in the represented connections as a lack of robustness when they could be different connectivity regimes associated with different environmental conditions. This is why the hypothesis of persistent connections over time, as opposed to intermittent connectivity (2) is essential to introduce and discuss. If the hypothesis proves to be true, we can expect similar results for a given method regardless of time domain or hydrological conditions, as long as we reach a sufficient sample size. On the other hand, if intermittent connectivity is expected, the suggestion would be to consider piecemeal applications of the causal inference methods for some temporal subdomains (e.g., summer vs winter) or hydrological conditions (wet vs dry). In our case, this would reinforce the problem (1) mentioned above.

In conclusion, we consider that changing the window as you suggest as a potential revision would be unfair in other respects mentioned above. Assuming we do so, we expect similar results in the case of CCF. For CCM, we may encounter robustness or convergence issues related to sample length similarly to PCMCI-CMI, without substantially enriching the discussion while requiring us to explain ourselves using different terminology related to nonlinear dynamical systems theory. In the end, the reader may also legitimately ask why not use longer records for CCM if we could.

2.2 Major Comment 2

I also have a question about the statistical significance. For example, in some information theory based studies, we use a shuffled surrogates method for statistical significance of a given link, which would differ from a p-value in a correlation analysis. If the method for identifying a statistically significant link varies at all between methods, this also needs to be apparent.

Currently, we rely on a Student’s t-test for CCF, CCM, and PCMCI-ParCorr to test the obtained dependencies or the mean correlation value in the case of CCM. PCMCI-ParCorr takes into account the larger degrees of freedom. PCMCI-CMI relies on a non-parametric test involving random local permutations (Runge, 2018). We agree on the importance of specifying the test as it controls the outcome of the causal inference methods. In the revised manuscript, we clarify the description of the methods to better evidence the

applied statistical tests.

As far as we would expect, the random shuffling surrogate test is comparable to a Student t-test for a bivariate normal distribution. Random shuffling is consistent with the principle that the data consists of independent draws from a fixed probability distribution. If the latter is normal, that is white noise, then, it is in phase with the null hypothesis behind a Student's t-test. Of course, we most likely do not compare two white noise signals, unless the causal inference method is multivariate and completely captures the deterministic signals. In the literature, it is common to rely on more constrained tests or randomization (e.g., Schreiber, 2000). We do not wish to expand on the appropriate test for causal inference as we believe it is a complex and open issue. We want to convey the main point that a test or p-value threshold is a way for the practitioner to control the number of links being output and limit the focus on the fewer, more substantial results.

2.3 Major Comment 3

For the synthetic case, I see you used a lot longer dataset than you have available for the real karst case study. This could lead to the better performance for the higher dimensional or multivariate methods – it might be useful to test the synthetic case for a much smaller dataset and observe or confirm whether these methods start to lose their detection of links. This could better show that the CMI/ParCorr types of methods do have better performance, but only when given a lot of data. I think the differences between the methods might make them inherently difficult to compare, but these are some things that could improve the attempt.

Indeed, the virtual experiment is a three-variable case that spans 365 days, while the real study case is an 11 (All) or 9 (P1, P2, P3) variable case over a restricted time domain of 48 (All), and 184 (P1), 62 (P2), and 218 (P3) timestamps. As also stated in response to your first comment, we fully agree that the sample size (together with the dimensionality) is an element impacting the robustness and performance of the method. In the revised discussion, the second point (see general response) aims at covering this point. Nevertheless, as stated in our general response, we decided not to pursue the suggestion to include sample size as a variable in our virtual experiment.

Our motivations are the following. First, our study remains a comparative study, and we think that such a focus on the PCMCI-CMI method and this complex problem would rather deserve a separate issue. Also, the conclusion of such an extended virtual experiment is reasonably known a priori, and as you have suggested: the results become more robust with increasing sample length (or decreasing number of variables). We see no reason why a non-trivial conclusion, such as recommendations of sample length as a function of the number of variables, would be transposable to a problem with different characteristics, such as different noise levels, model coupling patterns, signal behavior, or representative scales. In addition to the sample size and dimensionality, an adequate estimation of CMI also depends on the nature of the CMI dependencies, smooth or not smooth as it could be

expected in systems with highly dynamic connectivity, as well as the magnitude and the dynamic traits of noise. This CMI dependency and noise vary across spatial and time scales. The results also depend on the methods, for instance, kernel-based or nearest neighbors' estimators and their hyperparameters.

Our point, with the synthetic studies, was to show the divergence of the methods on the same – simplistic - case study, not as an answer to the question "what should we do", but rather as an exploration of the behavior of the tested methodology in a case where we can give meaningful interpretations of the results. For each problem on which those methods are used, we consider that a good strategy to be highlighted in the perspectives and recommendation section would be to test the issues met and the insights gained by using fit-for-purpose models mimicking the property of signals they want to study.

2.4 Major Comment 4

Finally, there are several places with strange phrasing, or where a term is introduced before it is defined, so there is momentary confusion on whether a reference is missing or the sentence is relevant. I am highlighting some of these that I noticed in the minor line-by-line comments below.

We apologize for these writing problems and thank the reviewer for pointing out some of them. We will make sure to correct them and recheck the whole manuscript carefully.

3. Minor Comments

3.1 Minor Comment 1

Line 7: "appears unstable" relates to my comment on data length…I think the instability is at least partially due to a very small dataset. Either way, it is not very clear what this term means within the abstract.

We agree with your interpretation. We hope to have answered the questions related to instability in our responses to comments 1 and 3. The revised abstract mentions how unstable the results are, i.e., that they provide run-dependent variable results.

3.2 Minor Comment 2

Line 15: "between variables from variables" did not make sense to me

We modified "*…interactions between variables from time-series only*".

3.4 Minor Comment 4

Line 35-45: This paragraph seemed scattered, and I did not come out of it with a clear understanding of "effective connectivity" in particular. Suggest to revise

For more clarity, we propose a reordering of the paragraph and some edits: "[…] T*he functional one is dynamic and is retrieved from statistical time-dependencies between local hydrological variables. Functional connectivity is a matter of cross-predictability and reflects dynamic links between the variables. These dynamic links are potential connections subject to confounding factors, i.e., they may or may not be related to a flow process between variables. Effective connectivity precisely refers to actual connections linked through hydrological processes and flows. Since CIMs with a multivariate framework address confounding factors, they offer the promise of discriminating functional connectivity from the effective one. From the structural to the effective connectivity through the functional one, the search for hydrological connections can be seen as a progressive limitation of the possibilities, from the potential paths to the actual paths taken by water*".

3.5 Minor Comment 5

Line 41: "process-based water flows" – I'm not sure if there are non-process-based flows of water?

Corrected. See the above response to minor comment 4.

3.6 Minor Comment 6

Line 45: "progressive constraint" was not clear to me

Also edited in response to Minor comment 4.

3.7 Minor Comment 7

Line 50: "heterogeneity" instead of "hiddenness"?

We meant both. Edited to "hidden heterogeneity".

3.8 Minor Comment 8

Line 60: I'm not sure about the sentence "nonlinearity is imputed to nonlinear hydrological processes", seems redundant

L60: "*Nonlinearity is imputed to inherently nonlinear hydrological processes such as power laws or threshold effects triggering flows.*"

Indeed. To clarify what we meant, we propose: *"Nonlinearity could be the result of heterogeneities or could be imputed hydrological processes themselves, often mathematically described as nonlinear with power laws or threshold effects triggering flows".*

3.9 Minor Comment 9

Line 65: would be good to re-define CCF here

Corrected

3.10 Minor Comment 10

Line 74: What do you mean by "to appreciate the results" – to compare with the results, or validate them?

L74:76: "*To appreciate the results, previous dye tracing tests have revealed fast connected preferential flow between the surface and a particular spot in the cave (Poulain et al., 2018). This prior knowledge can be seen as a reality check on the blind CIMs.*"

We meant to say "compare and validate", but talking about validation is probably an exaggeration. It is more like an agreement with our (one of our) expectations. We propose to be more explicit, we propose: "*In terms of expected results, previous dye tracing […] can be seen as a partial reality check*".

3.11 Minor Comment 11

Line 85: "Being multivariate" – I'm not sure that a multivariate approach inherently deals with confounding effects. For example, a multiple linear regression is multivariate, but does not do any type of conditioning on confounding variables…

L85: *Being multivariate, those methods can cope with confounding variables.*

We are not sure to understand your point. What we mean here is that, as several causes can be entered in the test, it can distinguish - if the problem is well posed and fulfills the condition of causal sufficiency - between direct and indirect causation. A multilinear regression, such as Granger causality, can do the same. In that sense, we consider that it can cope with confounding variables.

3.12 Minor Comment 12

Line 93: PC and MCI are brought up, and then defined later – would be better to re-arrange such that we are not wondering what they are.

Accepted. PC and MCI can be introduced earlier at L85.

3.13 Minor Comment 13

Line 99: "not preselection" to "no preselection"?

Corrected.

3.14 Minor Comment 14

Line 103: reference for causal sufficiency? In general, this is a good point for any analysis, you particularly reference it for CMI, could state that this hypothesis really underlies all your methods…

L100-104: "*While the CMI method is the most promising in that it does not assume linearity and accounts for confounding effects, as for the other CIMs, the reliability of the reported causal relationships nevertheless depends on underlying hypotheses (discussed in Runge, 2018a). Perhaps, the most important but the most challenging to verify and conceptualize in practice is the hypothesis of causal sufficiency. Causal sufficiency implies that the analysis should include all potential common causes.*"

The reference is the same as in the previous sentence. We repeat it and now also refer to Runge et al. (2019). We agree with your point and revise accordingly: "*Perhaps, the most important is the hypothesis of causal sufficiency because it underlies all CIMs (Runge 2018a, Runge et al. 2019). Causal sufficiency implies that the analysis should include all potential common causes. It is, however, challenging to verify and conceptualize in practice.*"

3.16 Minor Comment 16

Line 121: "overall good performance of this value" is vague, do you mean for the synthetic study, or the real study, or in general?

We now mention that the overall good performance of this value is relative to our dataset.

Still, from experience and based on the literature, the value of m equal to 2 is common. With m=2, CCM relies on actual trajectory segments rather than unique points (m=1). On average, the gain of performance can be thought of as the algorithm knowing in which direction the dynamic is going (e.g., lower discharge or higher discharge), hence a better state estimates (e.g., recession or rainy period), and mapping performance with the other variable. Usually, there is no significant gain of performance while passing from m=2 to m=3 (if gain there is), and one would hold with m=2 for parsimony.

3.17 Minor Comment 17

Line 136: I don't think you have introduced Granger Causality

The revision of the method section now correctly introduces Granger's pioneering work and these differences with the PCMCI-ParCorr method. For the preprint version, Granger was further explained in the supplementary materials.

3.18 Minor Comment 18

Line 153: Is two weeks of computation for a single processor? I figure it would take different amounts of time depending on whether you used a laptop or a server, etc, so could make this more clear.

Corrected. Two weeks parallelized on 6-cores intel i7 9th gen laptop.

3.19 Minor Comment 19

Line 176: It seems like for 2014-2017 time-series, there would be more than 465 time steps, implying the presence of gaps. This also comes into play in terms of your total data length for the multivariate methods. Basically, the counts in Table 2 make it seem like there is more data available than there actually is, when you start comparing multiple datasets (with 48 data points being the total overlap).

That is the point we are making in the last paragraph of the data section and the reason why we highlighted the restricted time-domain of the overlap in Figure 1.

3.20 Minor Comment 20

Line 204: "problematic case of the common cause" – after this, you define what this means, but as it is, the phrase seems a little mysterious, like a Sherlock Holmes story.

Rephrased into: "the common cause problem".

3.21 Minor Comment 21

Line 210: haven't defined Qb' yet

Qb is mentioned as the discharge of B earlier in L208. Yet, we modified to introduce Qb in L206 "In this case, two reservoirs, A and B, and their discharge Q_A, Q_B are subject to the same forcing …"

3.22 Minor Comment 22

Line 230: This is a "synthetic study" but the years make it seem like you are using actual data from your real study?

Indeed, see L214. We propose to remind it in the caption of Figure 2 to make it more apparent.

3.23 Minor Comment 23

Line 234: What is a differenced dataset? This comes up a few times and I'm not completely sure what it is…whether it is the increment or something done in the

modeling process.

The first-order difference is the time-series obtained by subtracting the values at the previous time step, i.e., Y_{t}-Y_{t-1}. This is now mentioned explicitly with a better motivation of the underlying reasons, as requested by other reviewers as well.

3.24 Minor Comment 24

Line 270: "If causality is hard to infer…" is not a great sentence, excusing a complicated figure and telling us it actually makes sense. You could just remove this.

We do as suggested.

3.25 Minor Comment 25

Line 297: "is be removed"

Corrected.

3.26 Minor Comment 26

Line 338: "evanescent singularity" is unclear

This sentence is removed from our revised discussion.

4. Cited references

Runge, J.: Causal network reconstruction from time series: From theoretical assumptions

to practical estimation, Chaos, 28, 075310, https://doi.org/10.1063/1.5025050, 2018.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., Nes, E. H. van, Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J.: Inferring causation from time series in Earth system sciences, 10, 2553, https://doi.org/10.1038/s41467-019-10105-3, 2019.

Schreiber, T. and Schmitz, A.: Surrogate time series, Physica D: Nonlinear Phenomena, 142, 346–382, https://doi.org/10.1016/S0167-2789(00)00043-9, 2000.