Response to Anonymous Referee 1

1. General response

Dear reviewer, on behalf of the co-authors and myself, I would like to thank you for your attention to the preprint manuscript and the helpful comments you made. Below are detailed responses to each of your major and minor comments.

Many points you raised seem related to the lack of precision in the methodology section and, for the discussion section, a lack of clear insights and missing references to existing literature. Regarding the methods, although several of your questions are answered in the supplementary materials (SM), we agree that the manuscript should be self-contained. Accordingly, the revised paper is improved by briefly revising the methodological section while keeping the section succinct. Regarding the discussion, based on your remarks and those of the other reviewers, we propose a rewriting of the section with improved references and comparisons to the literature on the following topics:

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

Concerning point 2 and the virtual experiment, the other reviewers strongly encourage us to extend the virtual experiment to study the effect of the sample size and/or the number of variables. This was not explicitly your request, although you expressed concern about the instability of the PCMCI-CMI method and its relation to the record length (Major comment 4). Nevertheless, we have decided not to comply with this request for multiple reasons that are reflected in the revised discussion, as any reader may have the same concerns.

These motivations are numerous and quite related to your major comment 4 about the sample size. First, our study remains a comparative study. Such a focus on the PCMCI-CMI method would rather deserve a separate issue. Also, the conclusion of such an extended virtual experiment is reasonably known *a priori*: the results become more robust with increasing sample length or decreasing number of variables (including delayed ones up to d_{max}). 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, the 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 characteristic 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 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.

Thank you again for your contribution to this discussion,

Damien Delforge

2. Major comments

2.1 Major Comment 1

> In the application of multivariate causal inference methods (Partial Correlation and Conditional Mutual Information), the authors didn't illustrate whether the history of variables is used in the conditioning set of variables or not. For instance, if one wants to test for the hypothesis that Y causes X using multivariate methods, then ideally the following conditional test should be implemented: $I(Y_t, X_t | Y_{t-1}: Y_{t-\tau}, X_{t-1}: X_{t-\tau}, Z)$. This conditional test means that one is testing for the statistical relationship between Y and X at time t while conditioning in the history of both variables up to a lag time of τ as well as the set of variables Z which includes all other confounders. This is a crucial point in the implementation of multivariate methods because it removes the effect of autocorrelation. There are many ways of conditioning in the history of variables; for example, the classical Granger causality (Granger, 1969) conditions in the history of X but not Y. Similarly, Transfer Entropy (Schreiber, 2000) does the same, whereas methods such as momentary information transfer (Pompe & Runge, 2011) conditions on both variables. I think that the authors need to mention explicitly what is the conditioning test for both multivariate methods. Please consider adding this information with clear mathematical expressions. Also, the parameter τ is often one of the most important hyperparameters in causal inference methods, so a discussion on the value of this parameter needs to be included. Please note that τ in this context is slightly different than $dmax$ which the authors use to set the maximum lag time for testing interactions.

The PCMCI algorithm tests for Momentary Conditional Independence (MCI) between variables, which implies conditioning on the Parents of both variables, that are subsets of the historical variables selected using the PC algorithm. This is further explained in the Supplementary materials following the description given in Runge et al. (2019).

Based on your remark, we agreed that the main manuscript could give more details about the algorithm and the tests. As suggested, the revised methodological section now includes the mathematical description for the independence test with a short description in section 2.1. A longer description will still be available in the supplementary materials. Also, we agree that the method section lacks clarity and that the references to supplementary material are not visible enough. Therefore, we propose a few edits in section 2.1.3 to remind that the PCMCI algorithm and the independence tests are further described in the SM. To facilitate a cross-sectional reading of the paper, we also refer to the test equation mentioned above of the MCI in section 2.1, in sections 2.1.3 and 2..1.4 related to ParCorr and MCI. Finally, as you said, the maximum delay is an important parameter that affects the outcome. Some remarks about this parameter are scattered in the preprint article. In particular, in line L190-195, we mention that it should be large enough in virtue of the causal sufficiency but cannot exceed five days due to missing values. In practice, this delay is often set based on the analysis of bivariate dependencies. We propose to stress the importance of the delay and mention these points earlier in the more appropriate section 2.1. Also, we can refer to figure SM.3 and 4 showing the CCF and CCM bivariate dependencies to motivate a maximum delay of 5 days. Both figures show that most of the dependencies do not sustain beyond four days.

2.2 Major Comment 2

> Related to the previous point, the reason why conditioning in the history of variables is important is because hydrologic timeseries of variables are often highly autocorrelated which leads to spurious causal links, and this testing removes autocorrelation. This context is important in interpreting the results obtained from causal inference methods either in the synthetic or real-

> world case studies. The authors used first-order difference of the original time series to remove the effect of seasonality and autocorrelation; however, this is not needed if the implementation of multivariate methods already conditions on the history of variables. Please revise the interpretation of results both in sections 3.2 and 4.2 to highlight this issue of autocorrelation.

We agree with your point on auto-correlation. Therefore, we will revise sections 3.2 and 4.2 as suggested. Even if the first-order difference is not needed, we consider it worth illustrating it as we believe that potential causal inference method users are unaware of this issue and its implication. This change will also meet the concerns of reviewer #4, who also requested some clarification and explanations about it.

2.3 Major Comment 3

> I found the discussion section to be lacking and it does not report any insightful comparisons to previous work of causal inference in hydrology. For instance, Ombadi et al. 2020 used four causal inference methods on a synthetic and real-world case studies with formal investigation on the impact of sample length, observational and process noise. Some of the methods used in this paper (e.g., CCM and CMI) were also used in that study. It would be important to compare the findings of both studies on the performance of different causal inference methods as this will allow us to build a consensus on the suitability of causal inference methods for hydrologic applications. Although Ombadi et al. 2020 is perhaps the most relevant to this study, there are other studies that used specifically information-theoretic approaches for hydrologic systems characterization such as (Jiang & Kumar, 2019). Please enrich the discussion section by linking the findings of this study to previous work.

We thank the reviewer for sharing these additional references, particularly the recent work of Ombadi et al., 2020. In addition, we propose to rewrite this section as proposed in our general response to be, as the commentary suggests, more in line with the existing literature and focused on specific issues as the impact of missing data or sample size, or the number of variables (including lagged ones up to d_{max}).

At first glance, the study of Ombadi et al. does not challenge our conclusions but complements them. We were particularly interested regarding the results of the bivariate Transfer Entropy (TE) method as we didn't explore it. The asymptotic behavior for True Positive Rate was moderately comparable to the PC method, and the reduction of the False Positive Rate is quite impressive. TE, we believe, could be less demanding in terms of record length as no conditioning is performed for the other variables and could be a viable potential alternative when the sample size is problematic. The behavior of CCM with respect to noise also allowed us to respond to some concerns of reviewer 3. We will make sure to report and discuss these interesting findings properly. The work Jiang and Kumar (2019) is also very relevant as it uses the PCMCI framework implemented in Tigramite to characterize hydrological systems.

2.4 Major Comment 4

> The record length of the timeseries used in this study is relatively short. This is one of the main challenges that face the application of causal inference methods in hydrology. In my experience, I found that methods based on information theory (e.g., CMI) often needs a long record (~ 2000 – 3000 data points) to provide reasonable performance. The results shown in this study either for the synthetic or real-world case studies are perhaps significantly impacted by the record length yet no discussion was included on the effect of record length. Please enrich the discussion section by highlighting the potential impacts of sample length.

Our beginning discussion mentions the problem of missing values (L316-320). We consider missing values to be problematic in the same sense as the record length problem. Missing values reduce the overlapping time domain over which conditioning can be performed. This is equivalent to a reduction in available samples, triggering issues similar to those of a short record length. Concerning Conditional Mutual Information (CMI), the adequate record length would depend on the dimensionality of the problem (the number of necessary variables to characterize the system) and the complexity of conditional dependencies (e.g., highly non-smooth dependencies, which could be the very case of intermittent or highly dynamic connectivity), which is not scale-invariant in time and space, and may vary geographically from study site to study site.

To the extent that hydrology deals with problems that are empirically complex to characterize beyond an overly simplistic lumped view, we agree that record length should be relatively large, but to what extent, we believe, is highly variable and case-dependent. It also depends on the method chosen to characterize the CMI. In particular, this is the reason why we chose the nearest-neighbor estimator and the shuffle test of Runge (2018) that is better suited than kernel-based approaches for short record length (< 1000), based on numerical experiments covering sample sizes from 50 to 2,000 and dimensions up to 10. Yet, despite the use of method recommended for small records, the real study case of the manuscript is concerned by the pitfall of estimating CMI with short record length resulting in non-robust test results, as you also interpret (minor comment #1). We showed that the robustness could be increased by performing an ensemble of tests. Of course, the robustness of a numerical result is one issue; its reliability in terms of connectivity is another.

Overall, the problem of estimating the CMI in case of missing values, short record length, or high dimensionality was the concern of all reviewers. Therefore, as mentioned in our general response, we propose revising the discussion section to discuss this topic properly.

2.5 Major Comment 5

> 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.

2.6 Major Comment 6

> Some of the basic information on causal inference methods that was mentioned in section 2 is not accurate and incorrect. For instance, the description of partial correlation in lines 136-138 is not accurate. The authors mention that "partial correlation is like Granger causality". This is quite vague. What the word "like" means specifically here? It is true that partial correlation shares similarities with Granger causality in the sense that both use linear regression to assess interactions while conditioning on potential confounders. However, there are crucial difference too. For instance, Granger causality is technically implemented in a different way than partial correlation with setting both restrictive and unrestrictive regression models, and testing for statistically significant differences using t-test and Ftest. These are very crucial differences. Also, on a higher level, the concept of Granger causality is based on what is known as predictive causality and it take into account time precedence. The authors should be careful in introducing the different methods and use precise information. I suggest that the authors revise section 2.1 by writing clear mathematical expressions for each causal inference method, and also be more precise in their description.

As mentioned in response to Major Comment 1 and the general response, the revised methodology section aims at being more specific, but not too much longer, while recalling the additional description in the SM. In this case, you will notice that the questions about the differences between the PCMCI-Parcorr and Granger causality are clearly mentioned in the SM1.3. How crucial the differences between the PCMCI-ParCorr and Granger approaches are is a matter for a debate that, in our opinion, does not belong to the paper's scope. We agree that changing the PC routine, or the test to an F-test or a Wald test are alternatives that will probably change the outcomes of causal analysis without altering the general philosophy. Of course, we may understand that if a choice alters the outcome, it is somehow crucial when it comes to causality. By being more specific in the main manuscript, we hope to provide information that may be of interest to the reader in relation to this question they may have or for a simple question of reproducibility. However, we do not have the arguments for preferring one implementation over another and, therefore, we do not wish to elaborate on the different variants of Granger causality.

3. Minor Comments

3.1 Minor Comment 1

> Lines 315-320 and elsewhere: the instability of CMI here is attributed to missing data. This might be one of the reasons, but I suspect that the main reason is the short record length (see my major comment #4). Also, a possible but unlikely reason is that the instability is the result of changes in the dynamic connectivity. This might be true if the timeseries used in Figure 6 (a, b, c and d) correspond to different hydrologic conditions (wet vs dry). If this latter case is possible, then it is worth of highlight and discussion.

We agree that the record length is the main reason for the instability and is related to the missing value problem (see the general response and major comment 4). The instability to which we first refer is related to the variable results of the test obtained on the same dataset and on the same time domain given by the time distribution of missing values, the record length, and the maximum delay. Since the time domain on which the conditioning is performed does not vary, the same holds for the hydrological conditions. It is, however, confirmed that for the graphs a, b, c, d of Figure 6, the temporal domain for which the variables are conditioned varies since the datasets vary. Hence, the analysis covers more or less hydrological states according to the size of this domain.

Although not stated clearly in the current preprint, we hypothesized that hydrological connections are perennial, i.e., not intermittent, even if sensitive to the hydrological conditions, i.e., nonlinear. We stress this assumption in the revised document. With this in mind, we might expect the hydrologic connections revealed by the MCI to be more robust as the sample size increases. However, this assumption may not be correct if the connectivity is intermittent or highly dynamic. In this case, we believe that the revealed connections are representative of averaged connectivity and may differ if the time domain of the analysis varies for small sample sizes, as you have suggested. It is worth being notified and discussed together with our hypothesis of constant connectivity over time as opposed to intermittent connectivity. We encourage further studies about intermittency in the perspectives at the end of the discussion.

3.2 Minor Comment 2

> Figures 4 & 5: there are several causal links with arrows pointing toward RF (Rainfall)?? Apparently, this is physically incorrect, but I was not able to find any discussion on this in the paper. Are these arrows drawn in the wrong way? Or these are the real results obtained from causal inference methods?

If it is the latter case, then this needs to be discussed. In general, this raises a red flag on the accuracy of causal links obtained from the different methods.

Indeed, some links that are obtained from the causal inference methods are physically incorrect. We report some of them, e.g., in L296-297 in the result section: "*We also denote two upward links to R4 and ET. These links seem physically unrealistic and potentially problematic since the effect of P2 is be removed from these variables, which may alter the whole causal graph.*". We further remind the how problematic it is in the discussion L353-356: "*For the multivariate methods, we have chosen to let the causal graphs be formed from the data. We have not prescribed any constraint on the conditioning of variables. This means that variables can be conditioned on potentially aberrant links, negatively impacting the whole causal graph*".

For more clarity, we propose to discuss this in the first part of the discussion about the summary and general appreciation of the results (see general response).

3.3 Minor Comment 3

Lines 228-229: the standard deviation of the noise added to the precipitation signal is unrealistically large!! Even the smallest value used here which is 0.05 of the standard deviation of precipitation is still large. A proportion of 0.001 of the standard deviation of precip is often sufficient to satisfy the condition of causal sufficiency. If process noise is very large, this will impact the results. See Ombadi et al. 2020 for the impact of process noise on the performance of different causal inference methods.

We understand that our process noise may be unrealistic because we did not attempt to mimic a realistic environmental rainfall noise. The noisy rainfall series, $P_{eff,A}$ and $P_{eff,B}$ are intermediate variables that are used in the causal analysis. Ultimately, what matters is the resulting noise in the reservoir discharge series Q_A and Q_b. In practice, we adjusted the parameters to our liking by graphically interpreting the discharge differences between consecutive runs of the toy model. The noise parameter may seem large because the unit hydrograph and reservoir act as low-pass filters, removing a significant portion of the injected noise.

Following your remark, we wanted to have a better representation of the impact of the noise level parameter ε_{lvl}. The table below shows the 100-runs average correlation $\bar{\rho}$ between two discharge series of 365 days generated with the same unit hydrograph $H = [0.7, 0.2, 0.1]$ and linear storage discharge equation $Q = 0.1S$ per noise level parameter. In the manuscript, this is equivalent to the average correlation between two Q_A obtained with model configuration 1. These values, although roughly evaluated, seem to us to be a reasonable noise panel to explore, especially given the wide range of scales of possible hydrological applications and where we could imagine any kind of dissimilarities.

$\boldsymbol{\varepsilon_{lvl}}$	0.05	0.1	0.15	0.2	0.25	…	0.7
$\boldsymbol{\bar{\rho}}$	0.96	0.87	0.76	0.66	0.57		0.24

3.4 Minor Comment 4

Lines 230-231: why only the last year was used? Is it a spin-off period to eliminate the impact of initial conditions or for computational reasons?

Yes, the early period was considered as a warming-up period. It will be mentioned. Of course, it also impacts computational time.

3.5 Minor Comment 5

Lines 252-253: this is perhaps related to the conditioning on the history of variables (see my major comments #1 and #2)

253-253: "*This finding supports our theoretical assertion: the multivariate nonlinear method is the best suited to address effective hydrological connectivity. Furthermore, the method appears to perform better if seasonality is left present in the time-series".*

As we did with the major comments, we agree here as well. We revised to make sure that this is explicit for the reader.

3.6 Minor Comment 6

Lines 260-262: This is a well-known issue with causal inference methods. You can refer to some studies that pointed to the same issue in evaluating causal inference methods either in hydrology or other fields.

260-262: *This is particularly contrasting with other methods and provides a valuable piece of information. However, the high precision comes at the cost of a low recall: CMI misses about half of the actual causal links. On the contrary, ParCorr misses none but has a bad precision, i.e., many false positives.*

It is indeed in phase with Runge's reports on the methods. This also could be related to the results of Ombadi et al, 2020. We will look for other references in hydrology.

3.7 Minor Comment 7

Table 4: please replace the abbreviations with the full name (e.g., TP: True positives) or alternatively add this info to the caption of the table so that it can be a standalone component.

We will choose one of the two options to improve the readability of the paper.

3.8 Minor Comment 8

Figures 4, 5 and 6: I suppose that the numbers in the arrows denote the lag time of interaction in days; however, this was never introduced or mentioned in the captions. Please revise.

Please, note that all captions mention it: "*An undirected line represents contemporaneous dependencies. Delayed dependencies are shown using directed curved arrows. All corresponding delays d are displayed in the middle of its corresponding arrow*"

We will, however, notify the reader that the delay is expressed in days.

3.9 Minor Comment 9

Lines 27-29: some applications of causal inference in hydrology are missing here. For instance, soil moisture-rainfall feedback (Wang et al., 2018) or differential impact of environmental drivers of evapotranspiration (Ombadi et al., 2020). There are others too if you look in the literature.

We thank the reviewer for sharing these additional references, as well as those from Jiang and Kumar 2019. We will update our literature review and ensure that we further compare our results with existing studies.

3.10 Minor Comment 10

Lines 35-44: I liked the distinction between structural, functional and effective connectivity. However, from the text, it was not clear what is meant by the effective connectivity and how it differs from the functional one.

35-44: *We refer to the terminology of 35 Rinderer et al. (2018), which is inspired by and borrowed from the field of neurological and brain connectivity (Friston, 2011). There are three types of connectivity: (i) structural, (ii) functional, and (iii) effective connectivity. The structural connectivity is derived from the medium and highlights the potential, static, and time-invariant water flow paths from the geological environment's topography, spatial adjacency, or contiguity. The functional one is dynamic and is retrieved from statistical time-dependencies between local hydrological variables. A statistical association may result from confounding factors, e.g., rainfall acting on two disconnected reservoirs or a shared seasonal pattern. Therefore, dependencies do not necessarily imply factual causation, such as process-based water flows. Then, the functional connectivity is a matter of cross-predictability and still reflects potential rather than actual flow paths for water. CIMs with a multivariate framework address confounding factors. They offer the promises of discriminating functional connectivity from the effective one, which reveals actual flow paths and processes within the system. From the structural to the effective connectivity through the functional one, the search for hydrological connections can be seen as a progressive constraint from the potential paths to the actual paths taken by water.*

For more clarity, we propose a reordering of the paragraph and some edits: "[...] *The 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.*

3.11 Minor Comment 11

11- Line 71: remove "obtained from". Typo.

Corrected. Thank you for pointing the issue.

3.12 Minor Comment 12

12- Line 144: the correct name is transfer entropy not "entropy transfer"

Corrected. Thank you for pointing the issue.

3.13 Minor Comment 13

13- Line 150: “computationally expensive and quickly require …”. The sentence is not logically correct. Please revise.

Corrected. Thank you for pointing the issue.

3.14 Minor Comment 14

14- Line 152: grammatical error in “at section 3”. It should be “in section 3”

Corrected. Thank you for pointing the issue.

3.15 Minor Comment 15

15- Figure 1 caption: when referring to (a) and (b), please remove the parentheses because it is confusing. Only use the parentheses the first time you introduce them.

Corrected. We will avoid parenthesis or parenthesis with a single letter when referring to the subplots. Thank you for pointing the issue.

3.16 Minor Comment 16

16- Line 291: replace “As for CCM” with “Similar to CCM”. The sentence does not read well currently.

Corrected. Thank you for pointing the issue.

3.17 Minor Comment 17

17- Line 15: this sentence does not read well at all. I understand what you want to convey, but it needs to be rephrased. Something like “…interactions between variables from timeseries only…etc.”

Corrected. We now use “from timeseries only” as suggested. Thank you for pointing the issue.

3.18 Minor Comment 18

18- Lines 98-99: the description of the parameter alpha_pc is not very clear and intuitive to me. Could you elaborate?

98-99: *The PC procedure has a tuning hyperparameter named αP C , which controls the number of potential causes. αP C varies from 0 to 1, where 1 is the less restrictive case which implies not pre-selection.*

The alpha_PC is a significance level used to select the parents in the PC stage of the PCMCI algorithm. We agree that “tuning hyperparameter” is too vague when used alone and will use a succinct description stating explicitly that it is a significant level, closer to the one of Runge et al. (2019): “The main free parameter of PCMCI is the significance level α_{PC} in PC_1, which should be regarded as a hyperparameter ...”

More details are available in the SM.

4. Cited references

Runge, J.: Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information, in: International Conference on Artificial Intelligence and Statistics, International Conference on Artificial Intelligence and Statistics, 938–947, 2018.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D.: Detecting and quantifying causal associations in large nonlinear time series datasets, 5, eaau4996, https://doi.org/10.1126/sciadv.aau4996, 2019.

Jiang, P. and Kumar, P.: Using Information Flow for Whole System Understanding From Component Dynamics, 55, 8305–8329, https://doi.org/10.1029/2019WR025820, 2019.

Ombadi, M., Nguyen, P., Sorooshian, S., and Hsu, K.: Evaluation of Methods for Causal Discovery in Hydrometeorological Systems, 56, e2020WR027251, https://doi.org/10.1029/2020WR027251, 2020.