Dear Reviewers, dear Editor,

We are grateful for the substantial feedback and the assessments and comments of the reviewers. This has allowed us to revise technical clarity across the manuscript and to conduct several major improvements:

1. We substantially expanded the sensitivity experiments for the methodical experiments to address the parameters of data generation more closely. Specifically, we added three experiments in Appendix G, which analyze the impact of the noise scale of the stochastic differential equations, the impact of the sampling rate of the time discretization (or time resolution of data for observational experiments), and the robustness to randomized changes to the underlying network structure at identical network density.
2. For the applied experiment on the interactions between the AMOC and Arctic summer sea ice, we also added two further robustness checks: We show that the involvement of Arctic temperatures is not required to detect the analyzed interactions, and that the sea ice aggregation area at the 66th percentile of variance of sea ice concentration comes up as a natural middle ground between larger areas that show the slow destabilizing signal from Arctic sea ice to the AMOC, and smaller aggregations that only show the fast stabilizing / negative interaction.
3. We acknowledge that the reconciliation of our applied approach focussing on PCMCI with the methodical results showing worse performance of PCMCI than LKIF+GCSS required more careful consideration and made an effort in Section 5 to explain more clearly the reasons for the exclusion of GCSS (and of LKIF in detailed analysis), for our higher confidence in PCMCI results, and for the lack of an expected causal effect from Arctic temperatures to Arctic sea ice.

We also respond to the reviewer and community comments point-by-point in the following. We attach the revised version of our manuscript with additions marked in blue, removals in red.

Sincerely yours,
Niki Lohmann, David Strahl, Annika Högner, Willem Huiskamp, Matthias Boehm, and Nico Wunderling

Referee 1

Referee Comment 1

General Comments:

In this work, the authors conduct a quantitative investigation into the reliability and robustness of three multivariate causal inference methods:

1. Liang–Kleeman Information Flow (LKIF)
2. Peter–Clark Momentary Conditional Independence (PCMCI)
3. Granger Causality for State Space Models (GCSS)

This is done in the context of studying the interactions of climate tipping elements in various facets of the Earth system which pose specific operational challenges. Through the quantitative metric of choice (Matthews Correlation Coefficient; MCC), the authors showcase unique advantages for each method, while also identifying three general principles for addressing nonlinear responses, delayed effects, and confounders during the application of these causal methods to climate tipping points. The use of MCC is natural and justified, as it considers balanced ratios of the confusion matrix in binary classification.

Following a preliminary study on synthetic data generated by a network of differential equations, they apply LKIF and PCMCI, based on their recommendations, on reanalysis data to detect tipping point interactions between Atlantic Meridional Overturning Circulation (AMOC) and Arctic summer sea ice (ASSI), confirming established physical mechanisms (bidirectional stabilizing interactions) beyond confounding influences (Arctic temperatures).

This study is a welcome addition to both the climate tipping literature and to the causal inference community. The structure of the paper is sound, transitioning from a synthetic-data investigation to a realistic application to climate tipping point interactions between AMOC and ASSI. Physically consistent results are derived in the latter study, both in terms of state-space causality and temporal causal influence regions, by applying the recommendations derived from the former experiment.

My general assessment is that this is a well-written paper overall, with the authors presenting their methodology and results succinctly and clearly, which should be of interest to the relevant researchers. The results are put in context, well interpreted, and presented without drawing strong conclusions. This work fits into the scientific scope of NPG. My recommendation is that it can be published to NPG following some major revisions and clarifications, as well as some minor corrections and adjustments.

We thank the reviewer for their encouraging feedback on the clarity and relevance of our study, especially with regard to the contribution to both the causal inference and climate tipping fields. We would also like to thank the reviewer for the constructive feedback to which we respond in the following.

Specific Comments:

(Format: p.##, l.## - Page number, line number | Section/Appendix/Figure/Table ##)

p.2, l.32—44 – In terms of references, the authors appropriately cite most relevant works in the associated fields throughout the manuscript. But, while the authors succinctly explain climate tipping points and provide constructive and relevant examples here, I would recommend that they note, either implicitly or explicitly, how they essentially describe bifurcation-generated tippings here (with a hint towards rate-induced tips when referring to effects across time scales), with other regime-switching driving mechanisms (internal variability and rate-limited tipping) also being possible **[1]**.

[1] https://arxiv.org/abs/1103.0169

Answer: We agree with the reviewer that the study focuses on bifurcation-induced tipping in the main experiments, and specified this in contrast to noise and rate-induced tipping on p.2, l. 35-36. Additionally, we conducted a new experiment on the noise scale, which confirms the possibility of purely noise-induced tipping in our model system, as specified in Appendix G, p. 30, l. 616-620.

p.7, l.178—179 – I would recommend elaborating a bit more on the details of how the time lag analysis is calculated in LKIF ("Time lag analysis is implemented by shifting any single input time series by a given number of time steps.") and whether the adopted approach is operationally consistent with the approaches in PCMCI and GCSS. These details can be added in the appendix if preferred.

Answer: We agree with the reviewer that the methodical description of LKIF can be improved both for mathematical rigor and readability. We state the approach by LKIF more clearly on p. 8, l. 217-220, and specify that all three methods analyze time lags by extending the dimensionality of the problem on p. 7, l. 166-167.

Section 2.3 – I would recommend adding a very brief paragraph here providing an interpretation of MCC and its values for people not familiar with the metric (e.g., maximum values and zero values, relation to the chi-squared statistic or other scores for intuition, etc.), which would also help with the self-containedness of the work.

Answer: Absolutely, we now specify the min/max values and their meaning (p. 9, l. 250-252) as well as the relationship to the phi coefficient and chi-squared statistic on p. 9, l. 243-244.

p. 9, l. 223—227 (and p.21, l.481—483 by extension) – As a quick clarification, how are the results affected by a different heuristic choice of the time step Δt to account for causal delays? Specifically, how are the results in Panel (b), Figure 2 affected by implicitly changing the signal-to-noise ratio of each relation? A quick note here would also help with empirically elaborating on Recommendation 1 in Section 4.

Answer: We agree with the reviewer that the timestep Δt is a crucial parameter to be analyzed in the study as it poses an additional time step constraint related to the internal dynamics of the involved systems. We conducted an additional experiment varying the orders of magnitude of Δt, and find that the default $\Delta t = 0.1$ indeed lies in the optimal order of magnitude for the coupling strength of 1. Larger Δt comes with a larger signal-to-noise-ratio on paper, but the time discretization poses a more crude approximation in that case, with a

good tradeoff in the order of magnitude of 0.1. Figure G1b (p.30) shows the results of this experiment with corresponding interpretation in l. 621-625. We additionally specify the constraint of sampling timescales in recommendation 1 on page 15, l. 353-355.

p.10, l. 239—240 – Indeed, since GCSS assesses causal relationships by projecting the latent state process (cause) onto the space spanned by the infinite past of the observation variables with and without the effect, it provides higher explanatory power with the autoregressive part resolving the presence of time lags. Elaborating a bit more on this argument will make the justification slightly more rigorous.

Answer: We thank the reviewer for this feedback on the explanation for the advantages of GCSS. We developed our argument more clearly, stating that the advantages of the latent state lie (in contrast to the other methods) in noise separation and time lag resolution, as to keep the explanation concise and accessible, see p. 11, l. 294-295.

p. 14, l. 325 – A quick note on why GCSS cannot be implemented here in a straightforward manner due to implementation details would be welcome (the small number of samples available for this experiment is also a valid limitation, based on the results of Fig. 2a). I do note the additional clarification in p. 17, l. 412—413, which is more than enough, but it does come towards the end of the case study.

Answer: We agree with the reviewer and now specify earlier that the model fitting procedure of GCSS is the reason, where the solution of the discrete time algebraic Ricatti equation is based on a single stationary noise covariance matrix, which can only be derived from a continuous time series at once, see p. 16, l. 392-394.

p.14, l. 341—343 – Is the choice of including cells above the 66th percentile based on a specific heuristic in the associated literature? How sensitive are the results to this choice?

Answer: We agree with the reviewer that the specific choice of the 66th percentile requires further justification. The percentile was originally chosen to show a clearer signal than aggregating all regions, while also retaining many regions considered important for freshwater influx, e.g. the Denmark strait (see Weijer et al., 2022). We added an experiment for aggregation at the 0th, 50th, 75th and 90th percentile to Appendix F, Fig. F3, which shows that links signs and strengths are consistently detected, even though the significance of links varies.

Figure 5 and p.15—16, l. 369—389 – I would recommend a clarification of the results in Figure 5 and the associated text: The colour of the causal arrow from ASSI to AMOC indicates a stabilizing/negative effect at one month delay but at five months delay there's a destabilizing/positive link instead. While that is the weaker link, as the text notes, adding two arrows that are independently coloured instead of single one that is coloured with respect to the stronger link would remove any ambiguity or confusion. I would recommend the same for the causal effect from ASSI to the Arctic temperatures (adding three arrows). That way, the coefficient of each link can also be superimposed next to the corresponding arrow for clarity. If the authors choose to implement these changes, they can also apply them to Figure F1 for consistency.

Answer: We thank the reviewer for the feedback on the readability of our figures. We edited all corresponding figures (Fig. 5 and Appendix figures F1-3) to display multiple arrows for different time lags for improved readability. We decided to keep the concrete edge weights outside of the plot, both for readability and to avoid potential confusion: The effect strength estimation (Linear Mediation) that comes with the PCMCI module provides interpretability (e.g., to calculate the impact of 1Sv AMOC weakening etc.), but may disagree with the edge significance calculated by PCMCI itself, which is indicated by the colors in the plot. We specify the meaning of arrow shading in the caption of Fig. 5.

Table C1 – For the synthetic data experiments, has a larger (or non-uniform) noise level been tested (but not too large as to break the required assumption of stationarity through the linear couplings)? Also, I would recommend adding the variable next to the parameter name in the first column, which would also clarify the use of uniform coupling strengths (not considering the choice of sign). E.g., “Noise scale (σ)”.

Answer: We agree with the reviewer that the role of the noise scale is worthwhile to explore. We conducted an experiment on the sensitivity to the noise scale and found no significant impact, as long as no noise-induced tipping occurs, see Appendix G, Fig. G1a. We thank the reviewer for the technical correction on parameter names and updated the parameter table to contain the variable names of Table C1 as in Eq. (1).

Appendices D, E, and F – Just wanted to note that these are a very nice addition to the text, further illuminating the implementation intricacies behind the causal methods utilized in terms of different variants, operational complexity, and robustness to observational noise and different reanalysis approaches, respectively.

We thank the reviewer for this positive feedback, and would like to note further additions in Appendices F and G that explore the robustness and sensitivity in more detail.

Technical Corrections:

(Format: p.#, l.# - Page number, line number)

1. p. 1, l.9 – The LKIF abbreviation is used in l.14 of the abstract but not defined here.
2. p. 3, l.50—53 – If possible, the authors can slightly revise this sentence for clarity and readability.
3. p. 3, l.59 – The PCMCI acronym is used here but first defined in l.77.
4. p. 4, l.104 – Small typo: “…not known to the causal method **and** introduces common…”.
   Answer: We thank the reviewer for drawing our attention to these technical corrections, all of which we implemented.
5. p. 5, l.116 – As a small note, since the Wiener processes for each state variable are mutually independent (as also noted in the text), please consider adding an appropriate subscript to W to indicate this (I assume the diffusion feedback σ is held constant across x_i).
   Answer: We agree that the subscript is necessary to indicate separate random processes, and added this subscript to all occurrences (in Eq. 1 and Appendix B).

6. p. 5, l.117 – As a minor note, I would recommend first noting here the role of c as the common confounder in the induced causal diagrams, akin to p.10, l.255—262, which would also clarify Panel (a) in Figure 1.
   Answer: We agree with the reviewer that the role of c was not specified clearly enough and clarify the confounder role on p. 6, l. 150.
7. p. 5, l.121—123 – It could be that I'm missing something, but shouldn't the cited negative critical value correspond to the transition threshold in the absence of the additive noise and linear coupling terms? If yes, maybe consider slightly rephrasing this sentence to avoid ambiguity.
   Answer: We thank the reviewer for noticing the ambiguity. Since linear coupling terms and noise are additive, we previously described them here as "other forcing sources" which may contribute to shifting the system beyond the threshold. We rephrased this to refer to coupling and noise on p. 6, l. 152.
8. p. 6, l. 140 – Small typo: "…are established in **the** literature.".
9. p. 7, l. 163 – Small typo: "…can be found in **Appendix** D.".
10. p. 10, l. 239 – Small typo: "…for time lag**. We** consider…".
11. p. 10, l. 258 – For clarification: "…(without interactions **and noise**).".
    Answer: We thank the reviewer for the typo corrections and implemented each accordingly.
12. p. 10, l. 260—279 – "exclusion" and "inclusion" are used here for the confounder term, but "hidden" and "known" are used in the legend of Figure 4. I would recommend sticking to the former throughout for consistency.
    Answer: We changed the legend in Figure 4 accordingly.
13. p. 10, l. 263—266 – I could be misinterpreting Figure 4, but I think this excerpt should read as "…**for the LKIF algorithm in the absence of forcing**.", "…**the forcing strength does not have an influence on the true positive rate of LKIF**.", and "…GCSS drops for an unknown confounder, but **the false positive rate** remains unchanged if the confounder is included in the causal analysis.".
    Answer: We agree with the reviewer that the previous phrasing was unclear and rephrased the corresponding paragraph on p. 13, l. 321-326. Our descriptions are mostly describing the trend of single metrics (e.g. true positive rate of GCSS) upon varying the forcing strength, i.e., reading the graph left to right rather than comparisons between inclusion and exclusion. This stresses the finding that a system with a confounder (and therefore lots of causal connections) can easily be handled as long as it's included.
14. p. 12, Figure 4 Caption – For clarification: "The LKIF method **instead** sees a large decline in false positives **and** the PCMCI method does not show a clear effect".
    Answer: We agree with the reviewer that clarification was required. Following further reviews, in order to not discuss concrete results in figure captions, the above sentence was removed among others from the caption of Fig 4.
15. p. 13, Section 4 – Referencing the relevant panels from Figures 2—4 here would help for fast lookup.
    Answer: We now reference the figures 2-4 across Section 4, e.g. p. 14 l. 343
16. p. 13, l. 293 – Small typo: "…as described in **Appendix** B.".
17. p. 15, l. 346 – Small typo: "…(Carvalho and Wang, 2020)**. A** reduction…".
    Answer: We thank the reviewer for drawing our attention to these typos and technical corrections and implemented all of them accordingly.

18. Panels (e) & (f), Figure A1 – Some connections (e.g., 9→11) might be construed as being moderated by an intermediate variable (10 in this case). While this shouldn't be an issue considering the structure of the linear and explicit couplings in Eqs. (1) and (2), having the arrows circumvent the extraneous nodes in the diagram would leave no room for misinterpretation. Finally, I would recommend noting Panel (c) as the default model network for clarity, just like the last column of Table C1.
    Answer: We agree with the reviewer that the dense network structures in Fig. A1 lacked readability and improved the figure to be more clearly readable, and specified c as the default option.
19. p. 21, l.470 – I would recommend writing "$\Delta x_i = x_i - 1$" here for simplicity.
20. p. 21, l.478—483 – Small typo: $\Delta x_{1,t}$ should read as $\Delta x_{i,t}$ in Eq. (B2). I would also recommend including the noise term (using Euler—Maruyama), making (B2) a coupled VAR(1) process that is consistent with the preceding exposition, as the simplification without the stochastic term does not really simplify things that much. The approximation statement in l.479—480 is still true then, both weakly and strongly (under appropriate convergence orders).
    Answer: Following the reviewer comments, we implemented the technical corrections accordingly on p. 23, l. 556, 565. We additionally introduced the Euler-Maruyama derivation on p. 23, l. 564-567.

Referee 2

Referee Comment 2

The authors compare multiple causal inference methods for their applicability to climate tipping points. They do this by applying all methods to different datasets generated from a network of nodes with tipping dynamics and comparing the dependence of the results on certain data parameters. Next, they apply two of the three methods to climate data to study the interaction between the AMOC and Arctic summer sea ice.

This is an important paper to establish which methods are most suitable for which types of data. However, since the results may depend on the methodological choice made, these need to be discussed more clearly and argued for thoroughly. My main comments are given below, and I attach a pdf with my specific comments.

Answer: We thank the reviewer for their positive feedback on the relevance of our study from a methodological perspective. We conducted several additional experiments to analyze the sensitivity of our findings both with respect to experiments on synthetic data and the applied experiment. In the following, we answer the concrete points of feedback individually.

Main comments:

1. Please mention the data parameters you vary more clearly and discuss as to why these are relevant.
   Answer: We agree with the reviewer that the description of varied parameters could be improved and extended on the varied parameters and their importance on p.10, l. 257-260. Basic data parameters like number of samples, coupling strength and delays can be highly variable across application examples, and can often be estimated before, and thus form the first experiment. Network structure like size, density, and confounding may complicate prediction further and is not known before conducting an experiment, forming experiments 3 and 4.
   Also, the description of the generation of the data can be improved (e.g. we link variables through a network, where at each node variability is modelled by SDE), where I suggest including the delay and explicit time dependence directly.
   Answer: We agree and cleared up the connection of the network and SDEs on p. 6, l. 156-157, stressing that the network structure serves as a conceptualization of the underlying coupled dynamic equations. We further introduce the time dependence and delays immediately in Eq. 1 and p. 6 l. 140-145 rather than in a second step as before.
   The networks given in the appendix need to be justified up to some level, and some discussion as to how sensitive the results are to the choices made would be valuable.
   Answer: We agree with the reviewer that the choice of networks require a further justification. We conducted an additional experiment, seen in Appendix Fig. G1c, that explores the role of randomized network structures (while keeping the same number of nodes, edges, cycles and negative/positive signs of edges). We find that only PCMCI sees some significant variation of performance across networks for high numbers of samples. In background analysis, we also found that our manually

designed systems were more resilient to noise than randomly generated ones, which confirmed the advantage of manual feedback loop design to ensure higher resilience to noise, in turn showing a clear transition from a non-tipping to a tipping space in Fig. 4 and the corresponding experiment. Overall, we added further sensitivity tests in a new chapter that we appended to the supplementary material of the manuscript (see Section G).

Furthermore, how sensitive are your results to parameters like noise size?

Answer: We thank the reviewer for this feedback on sensitivity investigation, and conducted an additional experiment seen in Appendix G, Fig. G1a. The role of noise is negligible for all three methods at default parameter settings, as long as no noise-induced tipping events occur (at the highest tested noise rate of 0.5).

It would be valuable to discuss the limitation of the chosen parameter settings and take those into account for the recommendations.

Answer: We agree with the reviewer that the usage of default parameter values may pose a limitation of this study, especially with regard to experiments on observational data. Results may appear optimistic in comparison to real world examples that might not reach the magnitude of any default parameter (e.g., 1000 samples). The MCC scores of the experiments should not be considered guaranteed scores for applied experiments, but reveal trends and comparability across methods, which then form our recommendations. We specify this on p. 14-15, l. 346-348.

And please specify earlier whether or not some variables undergo tipping in the time series you consider.

Answer: We specify c to be 0 by default in the subsection on data generation, p. 6, l. 150-151, and the lack of tipping in the default experiments on p. 10, l. 264.

2. The discussion and explanation of the three methods need to be improved. Currently particularly LKIF and GCSS are not clear to me, with e.g. the equations given not explained (what are X, A, …?). It would be valuable to discuss the choices that can be made when implementing this model and how much the results depend on this.

   Answer: We thank the reviewer for noting that the explanations of LKIF and GCSS could be made more mathematically rigorous and clear, we added descriptions of the variables of LKIF and its potential extension to other functional shapes of the underlying system on p. 8, l. 214-216. The time-lag approach of LKIF is also explained more clearly, see p. 8, l. 217-220. For the GCSS method, we explain the involved variables more clearly on p. 9, l. 232-233.

3. There is great value in including a case study, however I have two concerns relating the discussed example.
   1. Firstly, temperature is included as a potential common driver, but then the causal interaction is found to go the other direction with multiple lags. Have you checked the results when excluding temperature and the sensitivity of the results to method parameters?

      And can you discuss the physical realism of the network, as it states that AMOC is the main drivers of sea ice with temperature only following? Also, please clarify in the main text that you have considered other temperature datasets.

      Answer: Following the reviewer's comments, we checked the results when excluding temperature and found that the interactions between AMOC and ASSI are detected identically, see Fig. F2. We acknowledge that the lack of an expected link from Arctic temperatures to Arctic summer sea ice is a

limitation of this study. We consider it likely that the fast bidirectional connection (i.e., both influence each other on small timescales) prohibits detection by PCMCI, which is also indicated by the contemporaneous link between the two variables found for larger sea ice aggregations in Fig. F3a,b. We added a reference to Appendix F to the main text on p. 17, l. 443-444.

2. And secondly, it would be valuable to include an example where GCSS can be used, e.g. using climate model data. This allows for a more robust comparison and verification whether GCSS does as well for real data. Furthermore, in your recommendations PCMCI comes out worst, while here you deem it more reliable. Can you reconcile these conclusions?
   Answer: We agree with the reviewer that further explanation is needed why we recommend PCMCI over GCSS in our setting. The following reasons are part of this explanation.

   First, we selected the application example for its relevance to the climate tipping research area rather than as a pure example of method application, and are thus convinced that data masking would be required for this experiment, which GCSS doesn't provide. An additional application of causal inference methods or GCSS specifically to one or multiple datasets of Earth system model results goes beyond the scope of the paper in our view.
   The reasons to focus on PCMCI results rather than LKIF ones are detailed on p. 18, l. 450-453 as the number of samples is lower due to LKIF sampling, and the existence of significant delays is to be assumed following robust PCMCI results.
   A further reason for confidence in our PCMCI-results lies in the underlying dynamics. An important difference between the synthetic experiments and the applied experiment, specifically Arctic summer sea ice, is the importance of nonlinearity in variable dynamics. PCMCI performs significantly better in linear settings, which may also help explain why the results are robust and in line with explanations offered by the literature. We specify this reconciliation on p. 19, l. 489-492.

In addition to the above comments, it would be valuable to check the structure of the paper, specifically:

- In the introduction you go causal – tipping elements – causal, while I think tipping – causal would be more insightful.
  Answer: We thank the reviewer for this structural feedback and adapted the introduction for a clearer flow as recommended, see p. 2-3.
- Throughout the paper there is a lot of repetition. Up to some level this is ok, but some parts are mentioned over four times. I suggest reducing repetition throughout. E.g. in the introduction you already discuss a lot of results, which I suggest leaving for later and instead focus reasoning why this is a valuable exercise.
  Answer: We agree with the reviewer that there are some spots where repetition could be reduced, while still keeping both technical and applied sections accessible to a broad readership. We removed the direct results from the introduction and instead clarified the contribution to the remaining research gap on AMOC-ASSI interactions, and why the experiment is a valuable demonstration of important application steps

and constraints (variable identification, preprocessing, method selection), see p. 4, l. 101-116. As recommended, we also shortened the method details in the section head of the Methods and Data section, to leave details for the corresponding sections, see p. 5, l. 124-128.

- Lots of paragraphs consist of only one sentence, which I found distracting while reading. Please check for coherent paragraphs.
  Answer: We improved paragraph structure across many spots in the manuscript, including the spots indicated by the reviewer in the technical correction list.
- The order in which you explain the methods is opposite to what you show in the results figures. I suggest making this consistent.
  Answer: We thank the reviewer for catching this inconsistency and we corrected the legends and orders in all figures.

Technical corrections:

- p1 l. 9 LKIF abbreviation missing
- p2 l 35 Comma missing
- p 7 l 163: “Appendix” missing
- p 10 l 263: Mention Fig. 4
  Answer: We thank the reviewer for catching these technical errors and corrected them accordingly.
- p2 l 24: Shift paragraphs
- p3 l 72: Clarifications to confounding in physical systems
  Answer: We agree and clarified in more natural language that global warming introduces common trends (which may include nonlinearity and noise), see p. 3-4, l. 84-87.
- p3 l 74: Missing parameter names
  Answer: We agree and added all parameters explicitly.
- p3 l 79 and next paragraph: Combined into one paragraph
- p4 l 85: Removing conclusions in the introduction
  Answer: We agree and removed the conclusions on p. 4, l. 110-115.
- p4 102: Method section head too detailed
  Answer: We agree and removed the detailed introduction, see p. 4 l. 124-127.
- p4 104: Does a confounder contribute noise or forcing?
  Answer: We thank the reviewer for this specification. We considered this part of the section too detailed and shifted it to the corresponding confounder experiment 3.3, where we specify that the confounder results in forcing effects, which may lead to false causal attributions, see p. 13, l. 310-312.
- p 5 l123-125: Shift the unique role of cubic differential equations
- p5: Make c time dependent
- p5: Combine Eq. 1 and 2
  Answer: We thank the reviewer for these suggestions and agree with all of them. We adapted section 2.1 accordingly (p. 6), and slightly adapted Appendix B to specify tau=0 (p. 23, l. 554).
- p6 l 133: Name the methods here already
- p 7 l 175: Parameters for the LKIF equation
- p 7 l 188-189: Parameters for GCSS equations
- p 8 l 202: Add MCC meaning

- p 8 l 208: Add parameter descriptions at start of results
  Answer: We agree with the reviewer that the methodical and metric sections required some additional details and added variable descriptions for LKIF and GCSS (p. 8, l. 214-216 and p. 9 l. 232-233), ranges for the MCC score (p. 9, l. 250-252), and list and describe the parameters varied in the experiments more closely (p. 10, l. 257-260), also linking to further parameters analyzed in the appendices.
- Fig 2 caption: Lacking default parametrizations, captions without results.
  Answer: We now visualize and state default parameterizations in Fig. 2. We also adapted figure captions to not describe results in detail across all figures.
- p 9 l 223: Coupling strength and signal-to-noise ratio unclear
  Answer: We agree that the previous phrasing was not clear enough and thus specified that the coupling strength is difficult to interpret and offer the signal-to-noise ratio as a possible way to think about the size of coupling strengths, see. p. 11, l. 276-277.
- p 10 l 257: Errors in c_lim notation
  Answer: We thank the reviewer for catching this error and clarified the background of c_lim and f and corrected the error of notating f_s twice, see p. 13-14, l. 313-317.
- p13 l 279: Difficult reliability comparison between PCMCI and LKIF.
  Answer: We agree that this statement was too broad, the baseline performance of PCMCI in either confounder inclusion or exclusion just isn't great, there may be a niche for PCMCI in tipping scenarios, which we clarified on p. 14, l. 339-340.
- p 13 l 282: Timescale dependence of data availability
  Answer: We conducted an additional experiment on the timescale of sampling (at a constant inherent timescale of the dynamical system and at constant number of samples) and show the analysis in Appendix G, Fig. G1b now, showing a rather strong dependence of detection power on the timestep used for sampling the data. Our chosen default timestep of 0.1 indeed performs better than larger or smaller orders of magnitude.
- p 13 l 299-300: Confounder may be unknown in many experiment settings
  Answer: We agree with the reviewer that confounders may not be known in advance of analysis. To differentiate the recommendation accordingly, we added a sentence that GCSS and PCMCI depend a little less on the inclusion (i.e., this may impact method selection), see p. 15, l. 368-369.
- p 14 l 324-325: Role of third recommendation unclear.
  Answer: We removed the phrase "following our third recommendation" on p. 16, l. 391. The contrast of weak performance of PCMCI on synthetic data is reconciled with the application results later on p. 19, l. 489-492.
- p 20 app. A: Networks with larger sizes?
  Answer: We specified that we only derive networks in between these sizes (since 12 is the maximum size used in the experiments, both due to typical applications to climate variables and computational performance reasons) in Appendix A, p. 22, l. 543-545.
- p 25 app. F: LKIF results for the additional AMOC datasets
  Answer: We thank the reviewer for this question regarding consistent method results for all experiments. We looked at the LKIF results and found that they are qualitatively identical to those in Fig. F1a for HadISST4, while no links are detected for COBE-SST2 and ERSSTv5. HadCRUT5 shows an additional weak impact on AMOC on Arctic temperatures of 2% information transfer. Since the LKIF results are

not in the focus of analysis, we decided to keep the Figure F1 with a focus on PCMCI results.

Community Comment 1

Community Comment 1

I was notified by someone of this nice piece of work. It was a good read. I have some comments but here I don't have enough time writing down all of them. Here I just want to bring to the authors' attention that the IF-based causality analyses with external forcing and with time-delayed systems are handled a little bit differently; they should be treated with systems with augmented dimension(s). Here I am attaching two files, causality_ext_frc.pdf & delayed_causality.pptx, for your reference. The pdf regards the effect of external forcing with a system similar to your Eq. (1). Comparing Fig. 2 with Fig. 3 in the pdf, you will see how the external forcing issue is resolved. The second contains two PPT slides, regarding a VAR with time delays. In your case, if the delay tau=0.1, while you choose a time step dt=0.01 to generate the series, then there should be ten extra variables, represented by delayed series, to be added to augment the dimension of the system.

References:

X.S. Liang, 2016: Information flow and causality as rigorous notions ab initio. PRE 94, 052201.

X.S. Liang, 2021: Normalized multivariate time series causality analysis and causal graph reconstruction. Entropy, 23, 679.

Answer:
Thank you for your detailed feedback on the application of the LKIF algorithm. You are right to point out that the description of time lag analysis with the algorithm was too brief. The python package by Yineng Rong implementing the LKIF method provides the dimension augmentation for time lags larger than one, exactly as you described. It returns an n x n x t_lag matrix, where an entry (i,j,k) gives the information flow from variable j shifted by k time steps to the present state of variable i. We conduct the significance test against the standard error on this matrix to find significant edges at different time lags. If there is a significant interaction between two nodes at any time lag, this implies the existence of an edge in the causal graph as seen in Fig. 1.
We specified the approach more clearly on p.8, l. 217-220.

Regarding your comment on the external forcing experiment, we agree that the inclusion of the confounder variable is very important for LKIF. Fig. 4 shows exactly the difference between an application of LKIF to the set of variables without integrating the forcing variable (confounder exclusion), and with integration of that variable (confounder inclusion). As you expected, the algorithm indeed performs significantly better when the confounder is known,

which is notably a difference to the other algorithms which only see small improvements in true and false positive rates.