Dear Reviewers, dear Editor,

We are grateful for the positive feedback on the previous revisions and additional experiments. We also thank the reviewers for drawing our attention to the remaining technical errors. Besides these fixes, we clarified some methodological details and made minor changes to Figures 3 and 4 to improve clarity.

We also respond to the reviewer and community comments point-by-point in the following, with page and line references. 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

The authors did a good job of taking into consideration and addressing my and the other reviewer's comments and questions with detailed responses and satisfactory revisions. I like and appreciate the now better flow of Section 1, the newly added exposition and streamlining of Section 2, the clarifications and grounding of their results and recommendations in Sections 3-5, and the revisions to the styling/visuals and new sensitivity and robustness experiments in the Appendices.

We thank the reviewer for their positive feedback on the clarifications and additions conducted in our last revision.

My only additional (and final) suggestions are:

(i) p.11, l.263-272: Clarifying (and verifying) that the chosen growth duration of T=50 for the linear deterministic drift c(t) is slow enough so as the climate tipping mechanisms studied in Figure 4 are indeed a consequence of bifurcation-generated tipping (or of noise-dependent tipping in the critical orange region), controlled by the forcing strength f, and not due to rate-induced tipping processes.

Answer: We agree that rate-induced tipping needs to be excluded as a cause from this experiment. We therefore tested the same network configuration under a deterministic setting for growth durations of T = 50 (as in the paper) and T = 5000, and found identical tipping thresholds. This strongly suggests there is no rate induced tipping at T = 50. This finding is also in agreement with the fast dynamics of the system itself. We now specify this on p. 11, l. 271.

(ii) p.7, l.170: Strictly speaking, W is the Wiener vector process and dW are its independent-Gaussian/white-noise increments.

Answer: We thank the reviewer for drawing our attention to this issue and corrected the wording on p. 7, l. 171.

(iii) p.26, Figure F2: I would recommend adding a brief conclusive exposition or comment on the results of this figure, like done for Figure F1 and the newly added Figure F3, both for completeness and consistency.

Answer: We agree with the reviewer and state in its caption that the Figure F2 shows identical results to the Fig. 5 it is compared to.

After addressing these very minor suggestions, I am satisfied that the authors addressed all of my concerns and I therefore recommend that the revised manuscript is accepted for publication to Nonlinear Processes in Geophysics.

Referee 2

I thank the authors for addressing my comments, especially running the additional experiments. I believe the paper has substantially improved. I have one last comment, a thought which you may include in the discussion and some minor specific points.

We thank the reviewer for their positive feedback on the additional experiments conducted in our last revision.

1. GCSS – I still do not fully understand how you implement this. Is this again fitting the model to the data and then testing it, as for LKIF? Then please say so. This only becomes clearer when reading the appendix.

Answer: We agree with the reviewer that the fitting procedure wasn't made clear enough. We specify this now at the introduction of GCSS on p. 7, l. 186, as to summarize the complexity of multiple state space model transformation steps required for the Granger causality approach taken by the original paper.

Thought for the discussion – What does a monthly feedback timescale tell us on tipping dynamics on the order of years to decades? Physically, it takes time for water to be transported from the sea ice region, sink and contribute to the AMOC. Even 5 months is short for the deeper water masses.

Answer: Indeed the (causal) dynamics on the order of years or decades are a substantial area that could be explored using more long-term model experiments. We discuss the implications and limitations of monthly timescales in lines 447-452 of the manuscript and now also add a note on the role of developing causal structure under AMOC weakening or shutdown on p. 18, l. 458-459. The observed causal effects on timescales of months from ASSI to AMOC most likely indicate direct changes in surface temperature, though a detection of specific underlying causal dynamics (e.g. sea surface heat loss, meltwater influx) would most likely require spatially resolved model data. The more directly tipping-relevant link here is surely the AMOC -> ASSI link, for which monthly timescales are physically on the same order as actual potential non-linearities in ASSI.

Specific feedback:

L18 – "Recent decades" suggests at the latest after 2000, while the AMOC as having a tipping point is older than that.

Answer: We agree with the reviewer. We removed the part "In recent decades, " to focus on the core of the introduction, see p. 2,l .18

L26 – was -> has been

L45 – Remove "statistically"

L63 – points -> elements

L92-98 – Inconsistent use of Section v Sect.

Answer: We thank the reviewer for these technical corrections and changed them accordingly. The long form of “Section” only remains at the start of sentences now.

Eqn. (1) – Say x_i is a variable, t is time, either before or after.

L113-115 – x -> x_i

Answer: We clarify x_i and t before the equation on p. 5, l. 103-104 and corrected the technical errors in naming.

L170 – dW -> sigma dW

Answer: We thank the reviewer for catching this error and corrected it.

Eqn. (3) – Say what x_t is.

Answer: We now clarify that x and y are here vectors of variables and refer to discrete time steps t, see p. 7, l. 188-189.

L199-200 – I might have missed this, but what are “all four basic metrics”? Do you mean the TP/FP/TN/FN? Please clarify.

Answer: Indeed, this was unclear. We now specify TP, FP, TN, FN in brackets, see p. 8, l. 203.

L246-247 – Why does GCSS having a state space model means it would have a higher explanatory power?

Answer: The state space model abstracts from the observation variables (i.e., the actual states of the system are underlying and do not have to match exactly one observation variable). We state more clearly now (see p. 10, l. 250) that precisely this abstraction provides noise filtering (i.e., some noise is considered observational and does not affect causal structure), and flexible time lag handling. More formally, instead of the assumption of an underlying VAR process as the other two models, the state space can describe VARMA (vector autoregressive moving average) processes.

L250 – each variable ->? each method

Answer: We thank the reviewer for catching this error and corrected it to “each method”.

Fig 3 – (a) <-> (c) in caption and throughout the text. Also point out the different scales for the different methods, this is slightly confusing and makes it hard to compare.

Answer: We corrected the errors of labels, since the panels were reordered in the last revision. We updated the figure 3 to have identical scales for each method, and added gridlines to improve the visibility of the trends in the performance of each method.

Fig 4 – Can you put in gridlines to allow for better comparison between the methods? It would help comparing the values for e.g. GCSS and PCMCI, which appear comparible for the TPR.

Answer: We added gridlines to Fig. 4, which indeed improve visibility quite noticeably. As suggested, the GCSS and PCMCI performances under tipping are quite similar in terms of FPR and TPR (unless the confounder is excluded, then PCMCI is clearly advantageous).

L289-290 – Can you comment a bit on PCMCI earlier on? And also, it's TPR appears comparible to GCSS, and better after tipping, when LKIF is very unreliable because of the high FPR. So then the conclusion would be that neither really works well-enough.

Answer: We agree and added an earlier comment on PCMCI in the setting without tipping events. We revised the paragraph overall to be more clear that none of the methods works reliably when tipping events are present, see p. 11-13, l. 280 - 296.

Sec 4. – Recommendations 1 and 2 are on how and what data to consider, while 3 is on the suitability of the methods. I would suggest naming then 1a and 1b (and 3 becomes 2) and specify they are on data, not on method.

Answer: We agree with the reviewer and changed to the suggested labeling. We also indicate the different affected areas (data selection and method choice) in the preceding paragraph, see p. 13 l. 305ff.

L378 – decided for -> decide on

Answer: We corrected this error.

L451 – Larger timescales can also be considered using piControl instead of the scenario runs.

Answer: We agree with the reviewer that the preindustrial control runs offer a good way of gathering longer time series, increasing the reliability of our causal analyses. The previous comment on long experiment runs was directed at exploring the causal dynamics under an AMOC weakening, which would most likely require longer runs. We specify this more clearly on p. 18, l. 458-459.