Responses to reviewers's comments about the article
"A causality-based method for multi-model comparison: Application to relationships between atmospheric and marine biogeochemical variables"
by G. Bénard, M. Vrac, M. Gehlen

Review 1

Comment: **The manuscript by Benard and colleagues describes an approach to identify causal relationships across a subset of state or diagnostic variables in Earth system models. The method is based on the PCMCI+ algorithm, which quantifies how previous values of one variable (A) enhances the prediction of the others (B), while considering different time-lags. It differs from a typical correlation analysis in that this method includes intermediate 'explanatory variable' (C), which provides additional information (e.g., mechanistic understanding) linking variables A and B. As examples, the authors applied the method to analyse the relationships between surface nutrient concentrations and primary production in the eastern North Atlantic subpolar gyre as simulated in five ESMs. Using their preindustrial control simulations, they show that the simulated interannual spring bloom (primary production) intensity is predominantly modulated by the winter vertical mixing-induced nutrient variability. The role in lateral transport is less important.**

Answer: We would like to thank the reviewer for her/his constructive review. Below are our detailed responses in regular font, while the comments are in bold font.

Comment: **Not clear who will most benefit from using the PCMCI+. The method appears to be quite useful, especially for people with limited understanding of the non-linear and complex interactions between different ESM components. On the other hand, based on the example, the users ought to have a priori understanding of which variables are to be included in the analysis (performance goes down with increasing number of variables). If one fails to include key variables, the outcome could have been very different or misleading. I would like to see more information on how to select which variables to analyse and if there is a way to validate the results. For instance, can you show that mixing is actually driving the primary production variability? Do we need to run additional sensitivity experiments with the models to confirm the results?**

Answer: Runge et al.(2019) emphasized that a good application of PCMCI+ calls for a prior knowledge ofthe subject in order to select key variables.We will add the following sentence to the limitation section "it is important to note that PCMCI+ is not intended to be used as a data mining tool (Runge et al 2019)".
Our study specifically focuses on NPP with its variability as the primary scientific guiding the selection of relevant variables.

The method can, however, also be applied to explore the sensitivity of identified relationships to the variable selection by progressively incorporating or removing variables - examining, for instance, how the relationship between variables A and B might change when variable C is considered. We have partially pursued such an approach: our initial conceptual scheme was refined and variables such as temperature and Arctic Oscillation were removed without impacting the intensity of the other links. Stratification was also added creating a link not existing before. This led to our final, more focused, conceptual scheme. We propose to add

the following text to the discussion : "*The PCMCI+ method allowed us to explore multiple conceptual schemes through an iterative process of variable selection. By progressively adding or removing variables it is possible to assess the robustness of relationships between variables. Some relationships remain stable despite changes in the variable set, while others emerge or disappear, establishing the relevance of each variable."*

Comment: **Applying the method on different ESMs across the preindustrial simulations is useful to characterise the ESMs, but this comes out short of my expectations. It indicates that the models are different, which is not a surprise. When more models are involved, how can we synthesise the findings? A clustering approach could be an idea (Couespel et al., 2024).**

Answer: While clustering could indeed provide valuable insights into model behavior patterns, the scope of our study - with only five models, one scenario, and one simulation each - does not provide sufficient data points for a meaningful clustering analysis.

However, we propose to add the following text in the perspective section of the article: "*Future studies could extend this analysis to a larger ensemble of models and simulations, potentially enabling clustering approaches (e.g. Couespel et al., 2024) to identify systematic patterns in causal relationships across this larger range of ESMs.*"

Comment: **Model 'evaluations' were mentioned several times and appears to be doable, as also stated in the conclusions. So why not do this? I suggest the authors to consider applying the method on observations/reanalysis and ESMs historical simulations and thereby actually demonstrate its usefulness and added values beyond conventional model-data evaluation.**

Answer: We appreciate the suggestion about model evaluation using observations and reanalysis data. While this would indeed be valuable, we believe this warrants a separate dedicated study for several reasons.

Incorporating model evaluation would expand the scope of the current paper beyond its core focus. More importantly, applying this methodology to observational data raises several critical methodological questions that deserve thorough examination: the handling of false negatives in observational data, the phasing of climate events, and the development of appropriate evaluation metrics. These considerations would need exploration in a follow-up study.
We propose to add in the perspective section : "While this application is promising, its implementation would require an in-depth investigation of methodological aspects specific to observational data, particularly regarding the handling of false negatives and the phasing of climate events."

That said, we agree that comparing causal relationships between models and observations represents an important direction for future research, not just for our specific case but for any researcher applying this methodology to study Earth system processes. This approach could indeed enhance conventional model-data evaluation methods.

Comment: **There are some inconsistencies of messages in the manuscript. For instance, statement such as in L169: NAO affects MLD and again in L289: "MLD partially driven by NAO" but the models indicate insignificant relationship (Fig. S7). Does this mean the models are wrong?**

Answer: The absence of a statistically significant relationship in our analysis does not imply that no relationship exists between these variables. When a relationship is relatively weak, its strength may fall below the significance threshold.

We propose adding the following clarification to the section method of the manuscript:

"*It is important to note that the absence of a statistically significant causal link in our analysis should not be interpreted as definitive evidence for the absence of any relationship between variables. Rather, it indicates that any potential relationship is not strong enough to pass the significance threshold.*"

Comment: ***As is, the manuscript presents a proof of concept for understanding model behaviour and I wonder if it is more appropriate for journals such as the Geoscientific Model Development. For ESD, I would expect new understanding of Earth system processes, i.e., beyond that MLD affect PP intensity. I would recommend expanding the analysis to include future scenarios, which will considerably enhance the manuscript value for the ESMs community. As the authors hinted (Sect. 6), this additional analysis could reveal new understanding on how the non-linear mechanism of primary production will evolve in the future and therefore potentially explains the spread in the projections, which is a key outstanding research question today.***

Answer: We appreciate your suggestion of applying this method to the analysis future scenarios. The analysis of future scenarios is indeed a critical research direction, and we can share that this work is currently underway with a study in preparation.

The present study presents an original methodological framework for understanding Earth system processes, demonstrating how causal discovery methods can reveal and quantify relationships within ESMs. While GMD would be appropriate for new model developments, our work does not propose innovations in model developpement. Rather, it provides the Earth system dynamics community with a new way of understanding ESM behaviour.

Comment: **There are numerous 'hand-waving' statements, without any clear demonstrations, that should be avoided such as: L407-408, L409-410, L447-448, L462-463.**

Answer: These statements are part of the conclusion and perspective sections. We do not consider them problematic as they serve to outline potential future research, directions or applications of our study.

L1: … a novel causality-based approach to compare Earth System model outputs.

Modified "We introduce a novel approach to compare Earth System Model output using a causality-based approach." to "We introduce a novel causality-based approach to compare Earth System Model outputs."

Repetition: L9-10 and L11-12

L9-10 is modified: “The analysis of the causal links suggests a dominant contribution of winter vertical mixing to nutrient concentration compared to transport.”

L11-12 stays the same: Stratification is identified as an important factor controlling spring bloom NPP in some, but not all, models.

L12-13: “Most of the links … contributing to inter-model spread.” This is not specifically shown. Unless you have evidence, remove it.

Removed

L25: “leading to projections with still scattered outcomes” is redundant, as you already mention differences in future climate states and dynamics.

Modified “these disparities can lead to differences in future climate states and dynamics leading to projections with still scattered outcomes” to “these disparities can lead to differences in future climate states and dynamics with scattered projections.”

Paragraph L27-36: This is not the full picture. Increasing studies are now applying emergent constraint to link biogeochemical variables in causal relationships to understand the reasons behind projection spread, such as Fu et al. (2016), Kwiatkowski et al. (2017), Goris et al. (2023), and many others.

While we acknowledge the increasing use of emergent constraints to reduce the inter-model spread, it is not a model intercomparison method.

We propose a new version for this paragraph: “However, none of these methods characterise the interactions between variables or quantify their differences between models. Recent advances have explored two approaches to address these limitations. First, the emergent constraint framework has been increasingly applied to biogeochemical variables to reduce projection spread (e.g., Fu et al., 2016; Kwiatkowski et al., 2017; Goris et al., 2023). This approach identifies relationships between observable present-day variables and future projections across models, helping to constrain uncertainties in climate predictions. Second, correlation-based approaches have been developed to intercompare models and analyse the interactions between variables (e.g., Charakopoulos et al., 2018; Anagnostopoulos et al., 2010; Gleckler et al., 2008). However, while both these methods provide valuable insights, they cannot fully establish causality between variables”

L42-45: previous studies on should be cited here, e.g., Thomas et al. (2008), Tjiputra et al. (2012), Keller et al. (2012)

Citation added at the end of the sentence: (Thomas et al. 2008; Tjiputra et al. (2012) ;Yamamoto et al., 2020; Feucher et al., 2022; Keller et al., 2012),

L57: … Pelegri et al. (1996) and Williams et al. (2011) highlight ….

We removed the parenthesis : “Moreover, Pelegrí et al. (1996) and Williams et al. (2011) highlighted the..”

L62-64: what graphs? These would also fit better in Sect. 2. Perhaps the authors can also add a figure illustrating this so-called graph.

An example of a causal graph will be given in the supplementary.

L66: 'model spread' is mentioned again here, but it is never shown and analysed in the paper.

We deleted: "as well as the consequences on inter-model spread.". We maintain our hypothesis that differences in causal structures can be at the origin of inter-model spread.

L67: remove "The next"

We removed "The text", now the sentence starts with: "The section 2,...

L79-85: The graph is mentioned again, but I find it difficult to visualise this.

An example of a causal graph will be given in the supplementary.

L92: … irrelevance of event B to explain event A …

New sentence: "The test checks if event B explains or not event A when conditioned by C"

L103-105: could you clarify what happens if variables Y and Z are not independent, e.g., surface wind speed and MLD?

Answer: The partial correlation test remains relevant even when Y and Z are not independent. The test will decide which variable is the most important for X.

The dependence of two variables often stems from common driving factors. The MCI step refines this conditioning by including Z and Y parents in the conditioning set, providing a more precise information on the influence from potential dependencies.

L111-115: this is not very clear to me. An example with actual model variables would be useful.

Answer: Sentence added at the end of subsection "PCMCI+": *"For example, let's assume that NAO is identified as a parent of MLD and that transport and MLD are parents of nutrients. Then, when testing the link between nutrient and transport we will not condition by MLD but by NAO, the parent of MLD."*

L154: remove first period and correct references format.

Correction : "*...between-model spread of projected NPP (Whitt and Jansen, 2020; Kwiatkowski et al., 2020; Johnson et al., 2013)."*

Fig.1: increase axes label sizes. Increase fonts size of NAC/IC/EGC

The size of the fonts will be increased.

L164: remove first period

Removed

L1784: the surface nutrient stock

"Nutrient stock" modified to "surface nutrient stock".

Fig. 2: why there's an arrow from NAO to nutrients? The only arrow to primary productivity should be from nutrient, as PP in models is usually driven by phytoplankton growth rate (temperature, light) and nutrient availability.

There is an arrow connecting NAO to nutrients based on correlations found in various studies (Oschlies et al. 2001, Patara et al. 2011). In reality, this is an indirect relationship mediated by other physical variables. We included this arrow in our diagram as part of the conceptual framework guiding our study.

Arrows pointing to PP are from nutrients, but also from nutrient transport, and stratification. We acknowledge that the relationships between nutrient transport, as well as stratification and PP are indirect, with nutrients serving as an intermediate variable.

L189: extended

Answer: Modified ‘’extented” to “extended”

L202: Brody et al. (2013)

Answer: Removed the parenthesis in (Brody et al. (2013)).

Table 2: what is T in the slp formulation?

Answer: T is the number of timesteps, t takes value between 0 and T.

Modified: “with (i, j) indicating a grid point, P 90 (t) the 90th percentile of slp at time t, and T the number of timesteps”

L227, L235: Switch Figure S1 and S2 in supplementary. Use model names instead of modelling centres in figs. S1 and S2.

Answer: Correction of the names and order of the figures.

L245: via horizontal advection

Answer: Modified: “The inflow of nutrients via horizontal advection is..”

L247: beginning

Answer: Correction: “beggining” to “beginning”

L252: which graph?

Answer: The graph corresponds to the causal structure obtained from PCMCI. We propose to call it a “causal graph” for a better comprehension.

Modified: “We obtain one distinct graph for each nutrient.” to “We obtain one distinct causal graph for each nutrient.”

L266: …most and least similar links displayed in Fig. 4. The results …

Answer: Modified: “... with the most similar links displayed on 4(a) and the most different on 4(b)” to “with the most and least similar links displayed in Fig. 4.”

L267: Figs. S3-S6.

Answer: Correction of the figure numbers.

L268: … not necessarily true … (?)

Answer: Modified: “the converse is not true” to “the converse is not necessarily true”

L269: IPSL-CM6A-LR

Answer: Correction of the name.

L271: Figs. S3-S6

Answer: Correction of the figure numbers.

Fig. 4: increase fontsize and color readability

Answer: The figure has been modified.

Fig. 4 caption: …. introduced in Eq. 4.

Answer: Modified: “introduced in 4” to “introduced in Eq.4”

L288: Patara et al. (2011)

Answer: Removed the parenthesis in “(Patara et al. (2011))”

L308-309: not true for IPSL-CM6A-LR (nitrate)

Answer: Modified : “with values below 0.19” to “with values around 0.2”

L324-326: this statement is true only if the relationship holds into the future, i.e. there is no emerging new limiting factor for PP. I really recommend the authors to consider analyzing the scenario simulations in order to back up such statements.

Answer: Modified: “Specifically, if stratification continues to increase, the intensity of the bloom in the eastern part of the subpolar gyre may decrease more significantly in the UKESM1-0-LL and IPSL-CM6A-LR models compared to the others.”

to

“Specifically, if stratification continues to increase and assuming that the relationship will remain unchanged into the future, the intensity of the bloom in the eastern part of the subpolar gyre may decrease more significantly in the UKESM1-0-LL and IPSL-CM6A-LR models”

L343: +0.3 and +0.2

L346: +0.33

L350: -0.24

L358: +0.33

Answer: Added all the signs for a clarification.

L380: sidered. The results …. the blue curve illustrates …

Answer: Modified: “The result is displayed in Figure 8, where the blue curve in Figure 8 illustrates that precision is not critical,” to

“The results are displayed in Figure 8, the blue curve illustrates that precision is not critical,”

L391: … lower it. … number of variables

Answer: Removed “the” at the end of sentence: “does not lower the it.”

L404: previous study

Answer: Modified “previous studies” to “previous study”

L408: great if the authors can actually show this

Answer: This is a topic for another study.

L410: again, this is hand waving, and need to be shown or reformulate the sentence

Answer: The point on hand-waving statement has already been answered.

L440: explore the interannual variability

Answer: Modified: “explore the variability” to “explore the interannual variability”

L445: than horizontal transport … Also elsewhere, it is better to explicitly state ‘horizontal’ when referring to advection or transport.

Answer: Modified: “than transport.” to “than horizontal transport.”

Review2

Comment: **The authors present an approach to compare outputs from Earth System Models (ESMs) by identifying causal relationships between atmospheric and marine biogeochemical variables. The goal of this approach is to improve the understanding of how model differences affect projections, particularly for marine primary productivity in climate-sensitive regions like the North Atlantic subpolar gyre. They leverage a causality technique known PCMCI+, which discovers causal graphs with strengths and lagged interactions, allowing comparisons across models. The study introduces a custom dissimilarity metric to quantify differences between models' causal graphs, incorporating both the strength and lag of interactions.**

Overall I found the work interesting. However, there are many segments lacking precise writing making the work needlessly difficult to follow. I believe the paper could be greatly improved by providing explicit details and addressing a few other key concerns.

Answer: We would like to thank the reviewer for her/his constructive remarks. We appreciate the detailed comments which allowed us to improve the clarity of the manuscript. Below are our detailed responses in regular font, while the comments are in bold font.

Comment: **I found the description of PCMCI+ very confusing and had to turn to source material to learn what exactly it was doing. As written, there are a lot of terms that seemingly pop up out of nowhere (e.g. what is α_Yτ=ti, ϵXτ=0|Zτ=tj,...) that are not defined. I believe the paper could be strengthened if the language in this subsection was tightened, and provided more explicit details. If the authors think the details are too distracting, moving them to the appendix would also be acceptable.**

Answer: We acknowledge that the detailed mathematical formulation of PCMCI+ could indeed be confusing and distracting for readers primarily interested in the application and results of such a study. As we did not develop the method ourselves, we agree that interested readers can refer to the original work in Runge et al. (2019) for a comprehensive mathematical treatment. We have therefore revised this subsection to be more concise, focusing only on the essential information needed to understand our study. Below is our proposed revision of the PCMCI+ subsection:

We are now replacing the subsection of PCMCI+ by: "*To investigate the causal links between the variables in our study, we use the PCMCI+ method (Runge et al. 2019 , Runge et al. 2020). This approach is based on Granger Causality (Granger et al. 1969), which examines whether past values of variable A enhance the prediction of B beyond what B's own past values allow.*

PCMCI+ generates a causal graph where physical and biogeochemical variables form the nodes, while edges represent contemporaneous or lagged relationships between these variables. Each causal link is characterized by a time lag (0 for contemporaneous relationships) and a strength ranging from -1 to 1, with the sign indicating whether the relationship is positive or negative, and the value indicating the intensity of the relationship.

To detect these causal relationships, the method relies on the concept of conditional independence. Two variables A and B are conditionally independent given a variable C if knowledge of B provides no additional information about A when C is already known. In other words:

$A \perp B \mid C \Leftrightarrow P(A|B,C) = P(A|C)$

In our analysis, this concept is applied through partial correlation tests which evaluate whether the relationship between two variables persists when the influence of one or more conditioning variables is removed. The partial correlation is computed by first performing regressions of both variables on the conditioning set, then calculating the correlation between their residuals.

The algorithm proceeds in two main steps: PC ("Peter & Clark") and MCI ("Momentary Conditional Independence"). The PC step first identifies potential relationships between variables through a series of conditional independence tests. The most significant relationships are then validated during the MCI step, which examines indirect influences between variables.

For a detailed description of the PCMCI+ method, we refer to (Runge et al. 2020).

The variables of the conceptual scheme discussed in Section scheme will be given to PCMCI+ and the resulting causal graphs for different climate models will be compared to each other."

Comment: **I personally like the idea of the dissimilarity measure between causal structures. However, the current stated distance feels a bit ad hoc, and could use some further explanation.**

1. **As I understand, the authors use a convolution across time lags to try and dull the otherwise potentially sharp dissimilarity that might arise between models due to small time lag differences. Is this an accurate summary of why convolution is being applied? If so, further elaboration for why this is actually an 'issue'?**

2. **How is padding for the convolution being applied? It looks like the authors are applying full padding based on their example. They should explicitly state this, and perhaps show the full padding in their example. If they are not doing full padding, please explain what you are doing for your convolution at the edge.**

3. **What was the justification for width 3 Gaussian? Did that just seem to cover the expected minor differences in time lags between models? How should one choose the width of the kernel?**

Answer: We acknowledge that additional clarification would strengthen the manuscript and propose the following revisions:

1. We will clarify the reason for using convolution by adding the following explanation to the subsection "Dissimilarity" of the manuscript:

   "*Consider three causal graphs that are identical except for a single causal link with different time lags k1, k2, and k3, where k1 < k2 < k3. The convolution ensures that the dissimilarity measure reflects the temporal proximity of these lags, such that Dis(M1,M2) < Dis(M1,M3). However, when an expert considers that two time lags k1 and k2 are sufficiently distant from each other, they can adjust the convolution kernel size (set to 3 in our study) to reflect this assessment. Specifically, by choosing a kernel size smaller than the difference between k2 and k1, the dissimilarity between models M1 and M2 becomes maximal. In this case, the expert effectively indicates that the temporal distance between these lags is so significant that both M2 and M3 should be considered equally dissimilar from M1, resulting in Dis(M1,M2) =*

Dis(M1,M3)."

2. Regarding the convolution padding: While the equation describes the convolution application, we will explicitly state in the text that: "*For the convolution operation, we use full padding, which extends the input vector by a total of 2 elements (size of kernel minus one), adding one element at each end.*"

3. Concerning the Gaussian kernel width: We will clarify in the subsection "Dissimilarity" that the convolution width is a user-defined parameter by adding: "*Here we selected a width of 3 years because we consider that lag differences beyond three years reach maximum dissimilarity. The width of the kernel is a user-defined parameter that can be changed for specific studies.*" This choice reflects our knowledge and the temporal scales relevant to our analysis.

Comment: **My primary concern with the dissimilarity measure however has to do with its lack of resilience to the number of variables. By definition, the number of entries in the tensor M[i,j,t] depends on the partial correlation between all the processes and the number of time lags. This can be extremely large, or moderately small. If I add more processes to consider, the size of M can grow rather quickly (in a worst case scenario, if you have L variables, I believe PCMCI+ can return N=L(L-1)^2 connections, so M has L^2(L-1)^4T entries). If you want to measure the distance between M and M', then the Euclidean distance has the tendency to crush "near" and "far" points in high dimensions. So the metric of model comparison is heavily dependent on the number of variables involved in the problem. In other words, two models might look very distant if you only choose a small number of variables, but actually look quite close when you choose a larger number of variables. How do the authors control for this to ensure that this is a consistent and reliable measure of similarity between models?**

Answer: We appreciate this insightful observation about the challenges of measuring dissimilarity in high-dimensional spaces. It is correct that we should acknowledge certain limitations of our approach.

While analyzing dissimilarity across a large causal graph could indeed mask important local differences, our methodology allows for focused analyses on specific subgraphs or connections of interest, as demonstrated in our study. By examining selected submatrices that correspond to particular causal relationships, we can focus the dissimilarity analysis on specific links of interest. This targeted approach helps maintain interpretability and meaningful comparisons, avoiding the challenges that would arise from computing dissimilarity across all possible connections in a large network.

We propose to add a sentence in section "Similar models" : "*Looking at the dissimilarity of the entire causal graph could mask important local differences.*" followed by the rest of the section : "*Focusing on sub-matrices provides…*"

Comment: **Line 191: "Interactions between variables should not evolve during the 500 years of simulation." I don't believe this to be true. The causal interaction should have bursts of strength due to the nonlinearity inherent in the system. For example, if you took the wavelet power spectrum of any of your time series, I would strongly bet that strength of the signal at different frequencies is heavily time dependent. The choice of starting date is known to be important for causal interactions, even within**

preindustrial control (see Brandstator, Teng "Is AMOC More Predictable than North Atlantic heat Content")

Answer: While individual time series may indeed show time-dependent variations in signal strength, our focus is on quantifying the relationships between those signals, which operate on seasonal to interannual timescales. These underlying mechanisms are expected to remain consistent over the pre-industrial simulation, even if their relative strength may vary. This is different from predictability studies where initial conditions are crucial, and the non-linearity of each signal needs to be reproduced - our aim is to characterize the causal relationships between variables that are mechanistically linked through well-understood physical processes.

We therefore suggest rephrasing the statement to: "The primary mechanisms driving interactions between spring productivity and its controlling factors are expected to maintain consistent relationships over the 500-year simulation period, though their relative strength may show some temporal variations."

Comment: **Line 194: How are the variables normalized? Does each variable receive the same, say, standard normalization?**

Answer: All variables receive standard normalization. This will be clarified in the revised version: "Each processed variable is normalized prior to running PCMCI+" becomes "*Each processed variable is centered (mean at 0) and reduced (standard deviation at 1) prior to running PCMCI+.*"

Comment: **Section 4 needs some clarification, particularly with regards to the PCMCI+ algorithm and the metric chosen. In Section 4.1, the authors write "Focusing on sub-matrices provides insight into differences between models with respect to specific causal relationships. For instance, computing the dissimilarity on the sub-matrix corresponding to the drivers of one variable allows to explore specific dynamics shared by two models (similar impacts, similar lags)." Are you doing the following: 1. For a single variable, you get a causal structure for each ESM (which can look quite different). 2. You can then get a measure of distance between the causal graphs. This measure provides the "average" difference between the causal graph structures. If that is the case, how are you getting the causal graph structure from a single variable? Doesn't PCMCI+ give the structure of the whole system? Or, are you computing some sub-part of the network using a PCMCI+ type algorithm? For your dfe example, does this mean that you have looked at parents starting at dfe for all models. This results in a causal graph for each ESM (with dfe as base node), and then you computed distances between these causal graphs? This requires further clarification.**

Answer: Thank you for seeking clarification about our methodology. We will revise the manuscript to clarify how we analyze specific causal relationships within the full PCMCI+ output.

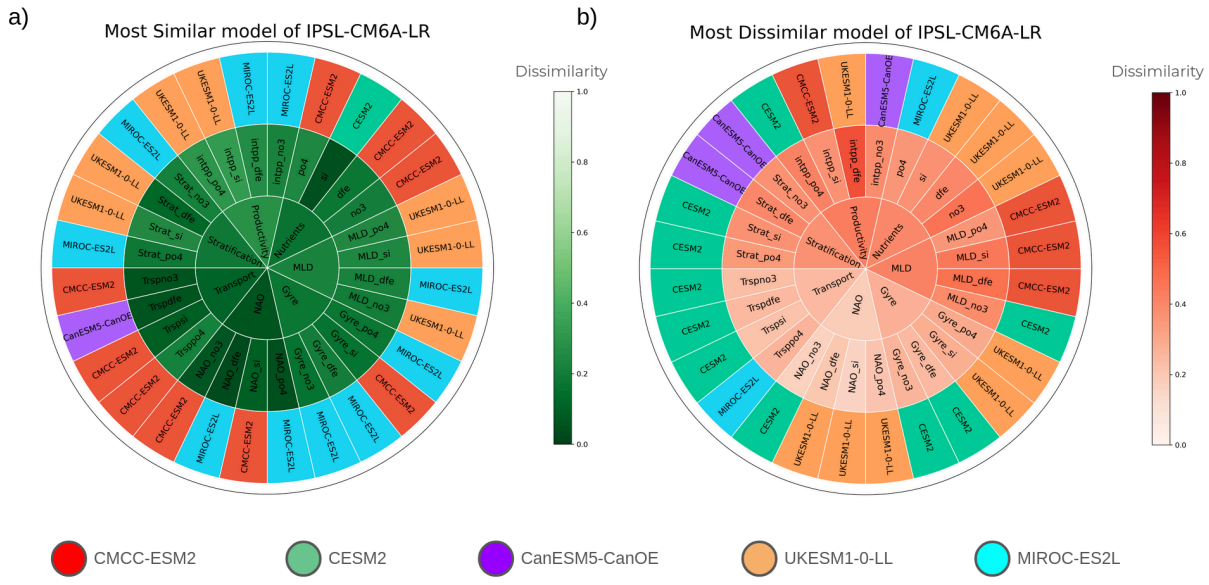
While PCMCI+ indeed generates a complete causal graph structure, our methodology allows for targeted analysis of specific relationships by examining selected parts of the full matrix. We propose modifying our current text at the beginning of the section 4.1 to read:

"*Focusing on sub-matrices provides insight into differences between models with respect to specific relationships. For instance, when analyzing dynamics around a specific variable (the i^{th} one), we can extract its corresponding column (the i^{th} column) from the complete PCMCI+ matrix. This column represents all causal influences from the ith variable on the other variables in the system. By applying our dissimilarity measure to these extracted columns from different models, we can systematically compare how various models represent the causal influences linked to this specific variable.*"

Comment: **Figure 4: This information feels too compressed. How am I to evaluate the difference between "close" and "far" here? How much further is the far model from the close model? Can I compare the results within a figure? How does MLD_no3 compare to Trsp_dfe? What is the point of the outermost circle, since you have a color bar? What does "no model assigned" mean? What do the _si, _no3, _dfe mean on the variables gyre, transport, MLD, ect? Does MLD_no3 mean the effect of MLD on no3 (or other way around)? Overall, I think this point is rather confusing. You pick ISPL model. Then for each variable, you build the causal graph (somehow, again not clear from my previous comment). You then measure distances between other models. The closest model is then determined based off your metric. Is that what is happening?**

Answer: The reviewer correctly understands our methodology. We acknowledge that Figure 4 could benefit from several clarifications to improve its interpretability. We propose the following improvements:

We will enhance the figure by incorporating a color scale to represent the dissimilarity intensity, while maintaining the outer circle to provide an additional visual reference for model classification. This will allow readers to quantitatively assess the differences between 'close' and 'far' models, and facilitate comparisons across different variables. Here is the new figure:



Regarding the nomenclature with suffixes (_no3, _dfe, _si), this stems from our methodology as explained in the manuscript: 'Causal links differ slightly between nutrients. For example, for a given link between two variables, the nutrient may not be directly involved but indirectly influence the relationships through changes in conditioning variables, resulting in slight variations in the calculated strength. Nutrients are identified in the following by adding ``\_no3" (nitrate), ``_dfe" (dissolved iron) or ``_si" (silicate).'

We propose to revise the figure and the caption to include more clarification:

"*Most similar (a) and dissimilar (b) models compared to IPSL-CM6A-LR based on the dissimilarity metric introduced in section 4. The inner circle indicates the variable, the middle circle shows each variable's variant depending on which nutrient is considered(_no3 for nitrate, _dfe for dissolved iron, _si for silicate). The outer circle indicates the most similar or dissimilar model. The color scale in the inner and middle circles represents the dissimilarity intensity, while the outer circle's color identifies the most similar or dissimilar model*"

Comment: **Figure 5: How is model agreement measured? Is "model agreement" the same as model strength? Is that the same as the metric you proposed earlier? There is a "significance" thrown around. What is significance and how is that determined?**

Answer: We acknowledge the need for clarification regarding model agreement and significance. The significance refers specifically to the statistical significance determined by PCMCI+ and the agreement is a binary indicator based on consensus among models about link significance. We propose revising the caption of figure 5 to be more explicit:

"*Each marker represents a different model, with the marker color indicating the statistical significance of the causal link as determined by PCMCI+ (green for significant links, $p < \alpha$; red for non-significant links, $p > \alpha$). We define model agreement in two ways: when all models consistently identify a link as significant (highlighted in green), or when all models consistently identify a link as non-significant (highlighted in red).*"

Comment: **Figure 7: Am I to read this as NAO has a larger impact on stratification than Gyre does on transport? Are these boxes comparable? If so, is this just a consequence of the time lag you chose?**

Answer: You are correct in your reading of the figure the boxes are comparable as the value in the y-axis is the strength of the link (intensity of the the relationship between two variables). We acknowledge that we should be more explicit about the temporal aspect of these relationships.

We propose adding the following clarification to the manuscript at the beginning of section 4.2:

"*In the following subsections, we focus exclusively on contemporaneous lags (lag-0 relationships). While lagged relationships were also identified in our analysis, they were much fewer and much weaker and therefore not robust enough to be considered*."

As this is a contemporaneous link, it does not result from the choice of time lag.

Comment: **Line 387: This needs clarification. Do you mean the ground truth is the values of $\alpha_{i,j}$? Or, the non-null values of $\alpha_{i,j}$?**

Answer: The ground truth corresponds to the non null values of alpha. We check if PCMCI is able to retrieve the existence or absence of a link. We do not evaluate how close the strength is to the ground truth in this article. We propose to modify this sentence to "The non-null values of $\alpha_{i,j}$ in these auto-regressive vectors serve as our ground truth, allowing us

to evaluate the ability of PCMCI+ to correctly identify the presence or absence of causal links."

Minor comments:

1. **Line 183 typo: wall->well**

   Answer: Typo removed

2. **Table 1: A very brief appendix which discusses the primary differences between the Ocean, atmosphere and BGC models would be nice.**

   Answer: A table summarizing the differences of biogeochemical models will be added in the revised version of the supplementary. For the oceanic and atmospheric components, they are not our primary concern in this study.

3. **Line 255: Might be helpful to have a plot here of a causal graph structure.**

   Answer: A conceptual scheme is already provided earlier to illustrate the structure of a causal graph. In the results, we deliberately focus on specific links from this structure, presenting another complete causal graph would detract from our targeted analysis.

4. **Line 382: Typo: ".A"**

   Answer: Typo removed

5. **Line 390: Typo: ".."**

   Answer: Typo removed

6. **"Thus, this experiment clearly illustrates that it is important to give a relatively low number of variables to the algorithm." I suppose this comment is sort of in alignment with my previous on choice for number of variables.**

   Answer: These are two distinct aspects: your previous comment addressed the computation of dissimilarity with many variables, while this sentence specifically discusses PCMCI+'s performance and computational efficiency. The recommendation for a limited number of variables here stems from the algorithm's ability to handle conditioning sets.

Review 3

Comment: **The manuscript entitled "A Causality-Based Method for Multi-Model Comparison: Application to Relationships Between Atmospheric and Marine Biogeochemical Variables" examines the relationships between different atmospheric, oceanic, and biogeochemical variables in the North Atlantic Ocean subpolar gyre. The authors propose a new approach for model intercomparison, using the PCMCI+ algorithm, and demonstrate its application in comparing CMIP6 GCMs. The authors**

find that vertical mixing has the greatest impact on controlling the spring bloom in subset of CMIP6 GCMs.

Answer: We thank the reviewer for her/his thorough review and constructive suggestions to improve our manuscript. We appreciate the recognition of the value of our approach for investigating causal links between different components of the Earth system..

Comment: **The study proposes a valuable tool for examining causality links and comparing different GCMs, making the approach and results of significant interest to the community. However, my major concern is that the authors miss the opportunity to investigate additional variables that could provide a more complete picture of biogeochemical dynamics in the subpolar North Atlantic region.**

First, the subset of nutrients is limited to NO_3, Si, and Fe, but the authors do not include PO_4. For example, Laufkötter et al. (2016) show that the North Atlantic is phosphate-limited in some GCMs (their Fig. 7). Furthermore, the analysis could benefit from studying not only NPP but also different phytoplankton size classes. Most GCMs have small and large phytoplankton classes and including this in the causality analysis could provide an important perspective on marine biogeochemistry in relation to the large and small phytoplankton dynamics in the North Atlantic. Also, the authors could look into Myksvoll et al., 2023 (their Fig.1) for possible additional variables which might be interesting to include into analysis.

Answer: Regarding your suggestions for additional variables, we acknowledge that including phosphate would indeed enrich our analysis, especially given the potential for phosphate limitation in some GCMs in the North Atlantic, as shown by Laufkötter et al. (2016). We will extend our analysis to PO4 in our revisions.

While we acknowledge the value of explicitly considering phytoplankton size classes in future studies, it is outside of the scope of the present one. Here we choose to focus on Net Primary Production (NPP) as the primary metric of biological activity, rather than investigating the subsequent cascade of biological processes.

Comment: **Additionally, I believe that increasing the number of GCMs could improve the robustness of the statistical results. The authors could consider including a larger number of GCMs and include more biogeochemical models (GFDL-ESM4 with COBALTv2, GISS-E2-1-G with NOBM, etc.)**

Although the results of the study are interesting, I believe the authors are missing an opportunity to investigate additional causal links in their analysis. I strongly encourage the authors to explore the data further, to make what will be a nice contribution to our understanding of subpolar North Atlantic biogeochemistry. For this reason, and given that it would require substantial additional work, I have to recommend a major revision of the paper.

Answer: We acknowledge that expanding the size of the model ensemble would strengthen our statistical analysis. Hence, we will include one additionnal model, MIROC-ES2L. Please note that GISS-E2-1-G will not be added to our analysis, as the 500 years long pre-industrial simulation is not available through ESG. Also, not all needed variables were available for GDFL-ESM4.

We believe that these proposed changes and the additional analyses will improve our analysis while maintaining the focus and methodological rigor of the study. We will incorporate these changes in our revised manuscript.

Minor comments:

Title could also include oceanic variable, e.g. "… relationships between atmospheric, oceanic and marine ..."

Answer: New title: "A causality-based method for multi-model comparison: Application to relationships between atmospheric, oceanic and marine biogeochemical variables"

LN57: "Moreover …" -reference format

Answer: Removed the parenthesis

LN154: check parentheses placement

Answer: Correction : "*...between-model spread of projected NPP (Whitt and Jansen, 2020; Kwiatkowski et al., 2020; Johnson et al., 2013)."*

LN188: with constant natural external forcings

Answer: Addition of the word "natural"

Table 1 – check biogeochemical model names (e.g. should be MEDUSA-2.0)

Answer: Typo corrected

Figure 5 – x labels with var names not consistent with the text. Also would be easier to follow if table 2 had all the abbreviations used in Figures.

Answer: Modification of x labels to be consistent with the text (e.g. TRSPNO3 to TRSP_no3, NO3 to no3,...)

LN328-329: "transport plays a role … "maybe cite fig. 5?

Answer: Modification from : "*However, as discussed in the previous subsection, transport plays a role in determining pre-bloom nutrient concentrations for some models."*

to

"*However, as discussed in the previous subsection, transport plays a role in determining pre-bloom nutrient concentrations for some models (Figure 5)."*

LN359: "In the previous analysis" ?

Answer: Modification to "*In the previous results*"

Figure 7 NAO and Transport – why only 3 GCMs shown for NAO/TRSPSI?

Answer: Thank you for your observation. Upon checking the figure, we found that two points are overlapping, with the 4th GCM hidden behind another model. We will increase the marker size to make this hidden model visible.