Dear Editor,

We hope this message finds you well.

We have carefully revised the manuscript according to the helpful reviewer comments. You can find our responses as well as an updated manuscript attached, which we believe to be much improved due to your and their valuable work.

After further internal discussions regarding our manuscript, we recognised an opportunity to enhance the sensitivity analysis comparing the results from Causal Effect Networks (CEN) between observations and model outputs to address potential ambiguity present in some of the figures in our original manuscript. Specifically, in light of recent results published in Di Capua et al. (Weather and Climate Dynamics, 2023), we decided to improve our sensitivity analysis by fixing the causal parents based on those detected as significant for ERA-20C when calculating the causal graph for the late period in our MR-30 analysis. This adjustment allows us to provide a more equitable comparison between the strengths of the β-coefficients in the model and the observations and thereby further clarifies some of our original results. As a consequence, the corresponding figures have been enhanced accordingly. However, our central conclusions stand with the new, more thorough analysis.

We believe these revisions significantly enhance the clarity and robustness of our findings. Thank you for your consideration, and we look forward to your feedback.

Best regards,

Dr. Julianna Carvalho Oliveira

Responses to reviewer R1

I thank the authors for considering my previous comments. While I find the results of this manuscript very interesting, I believe the authors ought to further improve the description and explanation of their results.

We thank the reviewer for raising the need for further improving the explanation of our results. In the process of revising our manuscript accordingly, we have decided to enhance parts of the presented analysis by taking up a slightly different sensitivity analysis strategy as devised in a previous paper (Di Capua et al., Weather and Climate Dynamics, 2023) published after the original submission of our present manuscript (27 June 2023). The modified tests are described in detail in Sects. 2.4 and 3.5 of our revised manuscript and particularly address some results that could not be well explained in our original setting. We believe that these minor modifications of the manuscript content, though not having been explicitly requested by the reviewer, contribute substantially to further clarifying the reasons for some of our original findings on the sensitivity of causal links to the use of different datasets and time periods, respectively.

For example, while NA-SST stands for North Atlantic sea surface temperatures, the authors defined the variable called "SST index" as a diagnostic of the SST variability over the North Atlantic region, which includes the anomalization and detrending of subtracted SSTs over two defined regions. Yet, in the discussion of the results, the authors frequently use simply "SST" or "NASST" while implying SST index (by the way, in the text of the manuscript it is SST index, while in Figures it is SST_index). The same issue persists for SLP and SLP index. I would highly appreciate it if the authors revised this issue. Maybe it is worth revising the naming of the variables used for the causal graphs as listed in Table 1? Perhaps there is no need to define an additional SST index and instead define NA-SST as the anomalized detrended monthly data of subtracted SSTs over two defined regions?

We thank the reviewer for their thorough revision of our manuscript and appreciate the valuable feedback. We have carefully addressed all the specific points raised above. In particular, we have revised the manuscript to ensure consistent and accurate references to the indices and variables used when interpreting the results. Specifically, we have clarified the distinction between the SST and SST index (now consistently referred to as *SST index* throughout the text) and similarly for the SLP and SLP index. We also reviewed the variable names used in the figures to ensure consistency with the manuscript text. We are confident that these revisions resolve the issues highlighted.

Further issues.

Abstract.

L2. The authors state that the analysis is done for 1908-2008. The next sentence in L2-3 starts with "We find …" → this gives the impression that the results are related to the entire period. While in L6- 7 the authors clarify that this result is representative only for the 51-year period from 1958 to 2008. I recommend simplifying the sentence,

e.g. "We find that in ERA-20C reanalysis the causal link from … to.. was evident/robust for the period 1958-2008 with an estimated causal effect expressed by … However, this causal connection was not evident/detected from the analysis of the entire period".

We agreed and rephrased the sentence as: " *In the ERA-20C reanalysis, we find that the causal link from the meridional NA-SST gradient in spring (expressed by the SST index) to the summer EA is robust during the period from 1958 to 2008, with an estimated causal effect expressed by a beta-coefficient of about 0.2 (a 1 standard deviation change in the spring SST index causes a 0.2 standard deviation change in the EA 4 months later). However, this causal connection was not evident when analysing the entire period from 1908 to 2008.*"

L4. In my opinion, L4 reads very hard for an introduction; the reader is introduced to the NA-SST term, which further develops into the SST index. If the authors decide to keep both definitions, maybe the following makes it more clear: "NA-SST gradient in spring (expressed by SST index)…."

Please see the respective change in the response to L2-3.

L6. ''only evident during the late period from 1958 to 2008''. L7. How did the authors come up to a 45-year-long period? Why not randomly sampled 51 year-long period (like 1958-2008) to intercompare the strength of specific causal links based on timeseries of identical length?

Thank you for raising this point. In our analysis, we implement a leave-k-out cross-validation strategy, where we randomly exclude 6 years (approximately 12% of the dataset) at each iteration. This process generates 500 different samples of 45-year-long time series for the reanalysis data, allowing us to assess the robustness of the causal graph structure by testing how consistent the identified links are across various subsets of the data. Our method primarily serves as a sensitivity analysis of the causal graph. By excluding specific years while preserving the natural temporal order, we ensure the robustness of our findings.

L11-15. It is not clear which period of time the authors are talking about: full period (1908-2008) or a 51-year period (1958-2008)? Or all simulations have different analyzed period?

We performed a series of tests for the full period, 51-year periods (1908-1957 and 1958-2008) and the cross-validation of 45-year samples in the late period for all datasets. The only exception for this is shown in Fig.6c, where we show a causal graph for the pre-industrial simulation spanning the entire 1000 years. To account for this difference, we rewrote the sentence as: "*We then use different datasets with the MPI-ESM-MR to analyse the 1908-2008 period, focusing on a historical simulation and a 30-member initialised seasonal prediction ensemble. We specifically test the model's ability to reproduce the causal links detected in ERA-20C and evaluate its impact on the model's predictive skill for European summer climate. We find that MPI-ESM-MR generally fails to reproduce the causal link between the spring SST index and the summer EA across the datasets.*

However, the 30-member initialised ensemble occasionally reproduces the causal link, though it typically underestimates its strength."

L15. Do authors mean "SST index"? Same in L16.

This sentence has been modified and the new lines read as follows: *"We perform a predictive skill assessment conditioned on the spring SST index causal links for July-August sea level pressure, 500 hPa geopotential height and 2-metre air temperatures for predictions initialised in May. Our results suggest that while the overall impact may be limited, leveraging these causal links locally could help to constrain and modestly improve the seasonal prediction skill of European summer climate."*

Introduction

L28-30. I recommend that this sentence is moved to ~L58, in order to keep the continuity of the introduction.

We thank the reviewer for this suggestion. However, we believe that keeping the sentence about applying the PCMCI causal discovery algorithm at the end of the first paragraph of the introduction is important for immediately highlighting the core focus of our paper. Introducing the key method and its application early on helps readers quickly understand the primary aim of our study and its relevance.

L70: The sentence in L58 starts in a similar way. I recommend to unite the sentences from L28, L58 and L70 into one paragraph.

We thank the author for the suggestion. We altered the text as follows:
"Nevertheless, while the linear regression-based analysis provided in Ossó et al. (2018) suggests a contribution of spring NASST to summer SLP variability, this approach does not imply causation. Disentangling the complex causal-effect pathways underlying the mechanism proposed in Ossó et al. (2020) over a long observational record is a crucial step in evaluating EA predictability in dynamical climate models. Although dynamical seasonal forecasts of European summer climate typically show limited skill (e.g., Mishra et al. (2019)), recent studies suggest that improving the representation of teleconnections can increase forecast skill (Oliveira et al., 2020; Carvalho-Oliveira et al., 2022; Schuhen et al., 2022). The physical mechanism connecting NA-SST variability and jet stream dynamics proposed in Ossó et al. (2020) provides a framework for assessing the broader influence of NA-SST on seasonal predictability of the EA, which is the aim of the present study.

In this paper, we use a Causal Effect Network based on PCMCI (hereafter CEN, Kretschmer et al. (2016)) to test the hypothesis that spring NA-SST causally drives a response in summer SLP and temperature fields in the Euro-Atlantic sector during 65 the 20th century. CEN overcomes spurious correlations caused by autocorrelation, indirect effects, or common drivers (Runge et al., 2014, 2019). It has been successfully applied to hypothesis testing for other tropical and mid-latitude teleconnections in the Atlantic-Pacific region (e.g., Karmouche et al. (2023)), the Indian Ocean (e.g., Di Capua et al. (2020a)), and the Arctic region (e.g., Siew et al. (2020); Kretschmer et al. (2020)).

Specifically, we use CEN to investigate the circumstances under which spring extratropical North Atlantic SSTs causally influence summer EA conditions and their associated impact on surface climate. We also analyse pre-industrial, historical, and initialised simulations with the Max Planck Institute Earth System Model in its mixed-resolution setup (MPI-ESM-MR, Dobrynin et al. (2018)) to evaluate model performance in reproducing the observed NA-SST–EA link, aiming to identify how this performance may constrain the seasonal prediction skill of European summer climate."

Section 2.1. It is not clear from the text what the analysed periods for different MPI-ESM-MR simulations are (pre-industrial, historical, MR-30).

We see the issue and added the following paragraph at the end of Section 2.1: "*We focus our analysis on the 101-year period spanning 1908-2008, using data from both the historical simulation, the MR-30 hindcast ensemble, and ERA-20C. In addition, the piControl simulation, with its fixed external forcings, offers a unique opportunity to study long-term internal climate variability free from anthropogenic influences. To capture the full range of natural variability, we leverage the entire 1000-year period of the piControl in our analysis. This approach allows for benchmarking internal variability across observed, historical, and hindcast datasets.*"

Section 2.2.

L118. by subtracting the average NA-SSTs or NA-SST anomalies?

In the beginning of Section 2.2 we specify that we calculate anomalies for each variable. To make this clearer, we also rewrote the sentence as follows: " *As a second step, we calculate the SST index by subtracting the average NA-SST anomalies over the eastern box (35° W-20° W, 35° -42° N) from the average NA-SST anomalies over the western box (52° W-40° W, 42° -52° N), represented by green boxes in Fig.1b.*"

L120. Isnt it the same as in L 106-107? L125. Is it really just a second step? The first step was in L111 to define a reference EA index. Then SST index seems to be ignored, and no "step" was given for it. And the second step is the analysis of the impact of NA-SST on temperature.

We improved the text structure taking this into account. We removed L120, since this was redundant given L106-107. Please see the reply to L118, where we include a second step for the calculation of the SST index. Lastly, we also rewrote L125 as follows: *"Next, we analyse the impact of NA-SST on summer T2m using two additional indices (...)"*

L131. I recommend the authors to be cautious here: NA-SST is not the index/variable as well as SLP. According to the authors notation SLP stands for the general definition of the sea level pressure, while SLP_ind represents the analysed variable. Therefore, the use of NA-SST definition here might also be also confusing.

We changed SLP to SLP index to make it clear.

Section 2.3.

L136. This part of the sentence could also perfectly fit to Abstract or Introduction.

We agree that the original phrasing had elements that could fit within the abstract or introduction. To address this, we revised the sentence to focus more directly on the application of the Causal Effect Network (CEN) method and removed the broader context. We rewrote the sentence as follows: *"We employ the Causal Effect Network (CEN) method (Runge et al., 2015; Kretschmer et al., 2016) to analyse the causal influence of the spring SST index on summer EA and temperature variability in the Euro-Atlantic sector."*

L140. A note for the authors: they are partial correlations if the user chose partial correlation (ParCorr) as conditional independence test. I recommend to reformulate the sentence as follows: "This method/algorithm is based on iterative conditional independence testing among …"

We agree with the reviewer's suggestion and rewrote the sentence as follows: *"This algorithm is based on iterative conditional independence testing amongst a set of time series (actors) to assess whether a link between a potential precursor and a target variable at a certain time lag is (...)"*

L168. From the description it is not clear why did authors choose "parcorr"?
Section 2.4.

We chose ParCorr for its suitability in detecting linear relationships, which aligns with our assumption regarding the SST index and EA. Its simplicity and interpretability, along with computational efficiency, make it ideal given our use of both observational and model data, as well the need to perform cross-validation. We added this sentence for clarification: *"ParCorr was selected due to its ability to detect linear relationships, its computational efficiency, and its established application in related studies (e.g., Siew et al. 2020), providing clear and interpretable insights into conditional relationships between variables."*

L172. Why is it necessary to exclude 6 random years from each iteration? Side note: Tigramite has also the methodology for the bootstrap aggregation, see tutorial here https://github.com/jakobrunge/tigramite/blob/master/tutorials/causal_discovery/tigram ite_tutorial _bootstrap_aggregation.ipynb

We appreciate the reviewer's comment regarding the exclusion of 6 random years, and thank the reviewer for sharing the tutorial. This decision was made to thoroughly assess the impact of sample size and interannual variability while preserving the chronological order of the remaining years. We opted against a bagging method, as it can lead to the repetition of the same years within a sample, which could complicate the interpretation of temporal dependencies essential to our analysis.

Section 3.1
L191. Large-scale cyclonic conditions where?

We rewrote the sentence for clarity as follows: *"The spatial pattern of the summer EA in its positive phase is characterised by large-scale cyclonic conditions across the Euro-Atlantic region, with the exception of the anticyclonic centre of action located south of Iceland and west of the British Isles."*

Section 3.2
L195-196. There is already the statement in Sect. 2.2 L117-118. Instead I suggest the following: "As explained in Sect.2.2, spring extratropical North Atlantic SSTs are represented via SST index, following the methodology of Osso et al.."

We rewrote the sentence as follows: *"As explained in Sect. 2.2, we evaluate spring extratropical North Atlantic SSTs via the SST index, following Ossó et al. 2018."*

L203 NA-SST or SST index? Here and below the authors use NA-SST. It is important to differentiate…

We rewrote the sentence as follows: *"This analysis suggests that the spring SST index - summer EA relationship is nonstationary."*

L206-207 Repetition from L105-107.

We deleted the sentence accordingly to avoid redundancy.

L207 "Correlation maps in Fig.2a-f"

Changed as suggested.

L208 summer SLP index?

The sentence correctly refers to SLP, as we investigate the point-by-point correlation of SST index with the SLP field, and not with the SLP index.

L213. If there is further interesting material, it could be added to the Supplements.

Thanks for this remark, we included the regression maps for early and late period as supplementary information.

L214. NA-SST anomalies or SST index?

We rewrote the sentence to improve readability as follows: *"Regression maps further suggest that the spring SST index is associated with summer NA-SST anomalies, which then influence atmospheric circulation (Fig.2g-i)."*

L217 Typo? No need of full stop.

We included the parenthesis missing.

L219 Figure 2d-f depicts significant correlation between the AM SST.... Given the order of the explanation, I would move panels d-f after regression maps (and regression maps move up as panels d-f).

We changed Fig. 2 as suggested.

L224-224. The definition of the T2mCE is already provided in Section 2.2 L126, however the coordinates of central Europe slightly differ. Please, correct this.

We corrected L126 and the coordinates are now the same in Table 1, L126 and L224: 46° N - 55° N; 11° E - 34° E.

L230. Is it correct to state here "spring extratropical SST" or rather "SST index"?

We rewrote the sentence to improve clarity as follows: *"To further test the robustness of the SST-EA relationship in ERA-20C, we evaluate whether spring SST index and summer EA are conditionally dependent. Specifically, we test the hypothesis that spring SST index is a causal driver for the summer EA, thus excluding autocorrelation effects or common drivers which could lead to spurious links."*

L233. '... in ERA-20C , as defined is Sect. 2.1?2.2?' Please, make sure that methodology section clearly states the periods that were analysed by different simulations and includes the definition of "early" and "late" periods.

Following the reviewer suggestion, we rewrote the sentence as follows: *"First, we build one CEN for each of the three investigated periods in ERA-20C, i.e. early, late and full periods, as defined in Sect.3.1."*

L237. Please, stay consistent: T2mCE (as in the text) or T2m_CE as in causal graph (Fig. 3a, Table 1)

As mentioned above, we made sure that all indices were written as subscripts in the manuscript to keep consistency, including those in Table 1.

L239-240. Shouldn't this information be stated in Sect. 2.3 instead of Results section?

Following the reviewer's suggestion, we moved this sentence to Sect. 2.3 and the last paragraph of Sect. 2.3 now reads: *"Lastly, the PCMCI parameters are chosen as follows: pc alpha = 0.2, alpha level to print results = 0.1, independence test = ParCorr, significance = 'analytic', masking type 'y'. ParCorr was chosen for its effectiveness in detecting linear relationships, computational efficiency, and established use in related studies (e.g. Siew et al. (2020)), offering clear insights into conditional relationships. Our CEN analysis focuses on τmin = 3 months and τmax = 4 months, which for simplicity we refer to as 3 and 4-month lags."*

Fig 3. Is there a typo in panel (b) with node "EAP"? It would be very useful if the authors would plot the nodes in panels (a) and (b) at the same position (implying that EA, SST_ind, and SLP_ind are plotted at the same position on both panels). This is easily done in Tigramite in plot_graph function using keyword node_pos.

We corrected the typo and now both nodes are correctly labeled as "EA".

Caption Fig 3. Circle → node? Typo "time tag" → "time lag"

Thanks for spotting the typo – changed as suggested.

L243. $\beta SST \rightarrow$ EA--> $\beta SST\_ind \rightarrow$ EA, same L245, L246, L253, L275.

Changed as suggested.

L246. Can you really expect that the link SST $\rightarrow$ T2mCE is mediated via EA? It would be a mediator if the authors observe the link EA $\rightarrow$ T2mCE. Contrary, there is no causal link from EA to T2mCE , so how does EA mediate it?

In the CEN analysis, we focus only on the 3-4 month lag, which is why no causal link is found from EA to $T2m_{CE}$, since the relationship between EA and $T2m_{CE}$ is contemporaneous. To answer the reviewer's question, we performed a mediation analysis using the path tracing rule. We found that while the SST index has a direct effect on $T2m_{CE}$ (-0.2) after accounting for EA, the indirect effect via EA is weaker and equal to -0.15. The direct effect of the SST index on $T2m_{CE}$, without conditioning on EA, is −0.31. Hence, the total effect equals −0.46, meaning that about 32.6% of the total effect of the SST index on $T2m_{CE}$ is mediated through EA. We rewrote the sentence as follows: *"We find a causal link of similar strength at a 3-month lag, $\beta_{SSTind \rightarrow SLPind} \approx 0.21$, between AM SST index and JA SLP index, as well as $\beta_{SSTind \rightarrow T2mCE} \approx -0.2$ between AM SST index and JA $T2m_{CE}$. Using the path-tracing rule (e.g. Kretschmer et al. (2021)) we find that about a third of the influence from AM SST index on JA $T2m_{CE}$ is mediated via the EA."*

L248. From the manuscript it follows that in order to avoid non-stationarity, the authors split the analysis to early and late periods. Yet in this sentence, the authors claim that since the full timeseries is nonstationary, they do not find a significant causal link in early period. It is not clear why they then found significant links in the late period.

We agree that the sentence was misleading. We therefore rewrote it as follows: *"We find no significant causal links when using the early or full periods."*

L250. SLP index? Also in L256.

We rewrote the sentence as follows: *"Next, we test the sensitivity of the detected causal links between spring SST index and summer SLP to slight differences in the analysed years. We assess summer SLP using both EA and SLP indices."*

L261. Remove redundant full stop.

Changed accordingly.

L262. How did authors find it out? Higher correlations?

In the analysis of Fig. 4c, we compare the EA index calculated from MR-30 with the ERA-20C EA index. The fact that the observed EA lies within the range of MR-30 values, depicted as light grey dots, leads us to conclude that MR-30 mostly encompasses the observed variability.

L263-264. Do authors refer to just historical simulations? Figure 5. There is no panels g, h, i as explained in the caption.

We refer to both historical and MR-30. We rewrote the sentence to make it clearer. *"Next, we evaluate the model skill in reproducing the spring SST index - summer EA relationship. We find that both historical and MR-30 simulations show limited skill, particularly in the late period (Fig.4b, Fig5). A comparison between correlation maps computed for the evaluated periods shows that while historical simulations do not show agreement in the spatial pattern of the spring SST - summer EA relationship against observations (Fig.2a-c), the MR-30 ensemble mean shows an improvement in reproducing the mechanism (Fig.5d-f)."*
Additionally, we corrected the caption as indicated.

L272. From the text/Methodology it is not clear if pi-control, historical, and MR-30 are the same length as analysed ERA-20C data. Are the EA phases in agreement between the model and ERA-20C? If not, that all could contribute to the reasons why the direction of the link is different in comparison to the reanalysis.

We thank the reviewer for highlighting this point. Please refer to our response to the specific comment in Section 2.1 for details on the analysed periods. To ensure comparability, the EA indices for the model were calculated as projections onto the EA index derived from ERA-20C. This method allows for a consistent comparison of the EA phases between the model and ERA-20C.

Section 3.5
L313. I believe it should be βSST_ind as well as in the caption of Fig.7 and L321, L322.

Changed accordingly.

L319. This already explained in Sect 2.2. I recommend avoid duplication and rather refer the reader to Sect.2.2.

Changed as suggested.

L321. ".. that are able to reproduce …"?

Changed as suggested. The sentence now reads: *"Targeting the two causal regions CE and Ridge, we test the hypothesis that the predictive skill of summer surface (...)".*

L377. 1902→ 1908?

Thank you for spotting the typo, we corrected it accordingly.

L387. "..detected in ERA-20C during the late period".

Changed as suggested.

Responses to reviewer R2

The authors carefully revised the manuscript, rewriting significant parts, and satisfactorily addressed all issues raised in the first round. In particular, I find the new methods sections valuable for the reader's understanding and the whole paper easier to follow, despite the complex methodology. Also, section 3.5 is much clearer in this version and adds value to the overall story.

We thank the anonymous reviewer for the positive feedback. We have carefully revised the manuscript, addressing all specific comments raised in the second round, as outlined below. During the revision, we decided to enhance parts of the analysis by adopting a slightly different sensitivity analysis strategy, inspired by a previous paper (Di Capua et al., Weather and Climate Dynamics, 2023), published after the original submission of our manuscript (27 June 2023). This updated analysis, detailed in Sections 2.4 and 3.5, significantly improves the clarity of our findings on the sensitivity of causal links across different datasets and time periods.

I only have a couple of minor suggestions that could further improve the presentation:

- Major point 1: seasons. I understand now that MA and AM are considered to build on Ossò et al. (2018, 2020) and that this choice is relevant to the results (e.g. Fig. 3). I would state this clearly in the text at lines 106-107 (it is clear from your answer, I would put at least part of that in the main text).

We appreciate the reviewer's comment. We included this information in the beginning of Sect. 2.2: *"We analyse bimonthly means in March-April (MA) and April-May (AM) for spring NA-SST and July-August (JA) SLP and T2m. We choose to investigate both MA and MA spring windows to allow comparison with previous studies (e.g. Ossó et al. 2018)."*

- Robustness of the CEN. I find it interesting that the links identified by the CEN are robust to the introduction of a new driver (tropical SST following Wulff et al., 2017; Fig. 1 in your response), I would also state this in the text when commenting the possibility of other missing processes.

Besides mentioning this link in the discussion, we now additionally wrote at the end of Sect. 3.2: *"This sensitivity in the causal link strength due to sampling suggests that the relationship between the spring SST index and summer SLP may be influenced by an external physical mechanism. Specifically, this could involve an additional variable not included in this CEN, such as the mechanism linking tropical SSTs described in Wulff et al. 2017."*

- L303 (was 249). Please add the information on how rare are the positive causal links in MR-30 here, I think this might be relevant.

Thank you for this remark. We included the percentage information and the new sentence reads as follows: *"Our sensitivity results suggest that the model predominantly fails to*

reproduce the observed links between SST index and EA or SLP indices (Fig6b), showing only in about 5% of the cases beta-coefficients in the positive range as in ERA-20C (Fig.3)."