Referee 2

We appreciate the anonymous referee's valuable feedback, which helped identify weaknesses in the text. In response, we thoroughly revised the manuscript. We provide below a point-by-point response to each comment. Please note that the referee's comments are highlighted in **bold** font, while our answers are in regular font.

Major comments:

1) **Period. MA, AM, MAM? It is quite difficult to follow the various spring selections and to understand why one is chosen with respect to the other. As long as I understood, most results refer to AM, but the other seasons are cited here and there. I think this is quite confounding for the reader, and a single spring window would help, maybe just stating that changing it does not change the results.**

   We thank the reviewer for this comment. We choose to investigate both MA and MA spring windows to allow comparison with the study of Ossó et al. 2018 and 2020. In fact, we find it important to mention the different time means, since different causal links were detected in our analysis concerning AM or MA SST index (also, please see the reply concerning the specific comment to L166). This is nicely illustrated in Fig.3, where we find a causal link between MA SST index -> JA EA (Fig.3f, lag = 4) and between AM SST index -> JA SLP index (Fig.3c, lag= 3).

2) **The methods section 2.2 could be expanded to include a more detailed description of the bootstrap ensemble, which is currently just in the main text. It is unclear how the authors refer to the bootstrap ensemble throughout the text. Linked to this, the term "causal ensemble" or "causal timeseries" is not explained in the text. Also, since many indexes are considered in the text, a point-by-point list of all indexes considered with a unique and identifying name would help. For example the SST_Ridge just appears at the end of the results section, but would be useful to have it here for quick reference.**

   We thank the reviewer for these suggestions. We included two new sections and a table in the methodology (lines 169-188 in the revised manuscript):

   "2.4 Bootstrapping and ensemble subsampling

   We perform bootstrapping and ensemble subsampling (e.g. Dobrynin et al. (2018)) to investigate the sensitivity of the causal links to data sampling. When analysing 1958-2008 using observations in Sect.3.3, we randomly select 500 samples of 45-years long, i.e. excluding 6 random years at each iteration. Each of these time series are then analysed with CEN using the same hyperparameters (see Sect.2.3). We perform a similar bootstrapping using MR-30, but additionally include a second step of ensemble subsampling. That is, we first randomly exclude 6 years from the analysed period. Next, we randomly select 1 ensemble member from the 30-member set. Each time series is then analysed with CEN. This process is repeated 2000 times. It is important to note that reducing the length of the time series in this way increases the variability and hence lowers the significance of the obtained β-coefficients. However, this should not by itself lower the strength of the coefficients themselves.

2.5 Predictive skill assessment

In Sect.3.4, we perform a predictive skill assessment for SLP, T2m and Z500 at lead times of 3-4 months in MR-30 against ERA-20C. For this assessment we use point-wise detrended anomaly correlation coefficient (ACC, Collins (2002)). We are interested in assessing the predictive skill conditioned to the strength of significant β-coefficients (p-value < 0.1). Our hypothesis is that the predictive skill in summer is likely to increase in cases where MR-30 is able to capture the causal link between spring SST index and summer EA, as opposed to cases where the model fails to capture the observed causal link. We refer to these time series as MR-30 bootstrap ensemble. For example, we shall assume that we are interested in calculating the conditioned predictive skill of JA Z500. To accomplish this task, we first identify the specific years and ensemble members that correspond to significant β-coefficients for the spring SST and summer EA. With this information, we can then sample JA Z500 to create a time series of similar length. In case more than one ensemble member is randomly selected in a given year, we calculate an ensemble mean. We then determine the ACC between the MR-30 bootstrap and ERA-20C."

We additionally included a table (Table 1) containing a list of the investigated variables. We also added in Sect. 2.2 (lines 124-128) "(...) In a second step, we analyse the impact of NA-SST on summer T2m using two additional indices, i.e. T2m_CE and T2m_Ridge (Sects. 3.2, 3.4). The T2mCE index is calculated as JA T2m anomalies averaged over the region 45◦N-55◦N; 10◦E-35◦E (indicated by a red box in Fig2f), and the T2mRidge index is calculated over the region 40°N-55°N; 15°W-34°W (indicated by a black box in Fig7b)."

3) **The end of the results section could possibly make for a separate subsection: see comment at lines 256-281. I think this is one of the most interesting results of the paper, but is currently difficult to grasp and could be easily lost in the main text.**

Following the reviewer's suggestion, we included a new section "3.5 Forecasts of opportunity: could causality help?" concerning these results.

4) **Physical pathway. The CEN framework is a powerful tool but should be used with caution. In particular two things are necessary, following Kretschmer et al. (2021):**
    a) **the CEN should only be used to "measure" a causal link for which there is already an hypothesized physical pathway;**
    b) **the causality of the link is always conditional to the choice of the variables included in the model, meaning that if a relevant variable is missing, the CEN result may be wrong.**

   I think the authors respect both "requirements", but I suggest to:
    c) **recall the possible physical mechanism behind the link when presenting the CEN. It is currently only cited in the introduction.**
    d) **put more emphasis and discuss on the possible impact of a missing process in the CEN.**

   Thanks for your input. Concerning a), b) and d), we included a paragraph in Sect. 2.3 to provide a more in-depth discussion on the challenges of using CEN (please see below, under the reply to the specific comment for L131). Concerning c), we rewrote the beginning of Sect. 2.3 (L134-136) as "We use Causal Effect Network analysis (CEN, Runge et al. (2015); Kretschmer et al. (2016)) to test whether spring NA-SST anomalies causally influences the variability of summer SLP and temperature fields in the Euro-Atlantic sector during the 20th century, investigating the mechanism proposed in Ossó et al. (2018, 2020)."

Specific comments

L94. The ocean state is derived ~~in~~ from?

Text modified as suggested (L100).

L94-97. It is not completely clear to me how the assimilation experiment is performed for the ocean. If is said at line 94 that an ocean-only simulation with MPI-OM forced with ERA-20C is performed. Is this ocean state then used for the nudging of the 30 ensemble members? I think this choice should be (quickly) motivated in the text.

We thank the reviewer for raising this question. In the assimilation experiment for the ocean, an ocean-only simulation is conducted using the MPI-OM model, which is forced with ERA-20C atmospheric reanalysis data. This choice is made because it allows for a longer observational record compared to using an ocean reanalysis. Additionally, using ERA-20C for atmospheric forcing enables the production of century-scale assimilation runs without the need to create a separate ocean reanalysis. The ocean state obtained from this ocean-only simulation is then used for nudging in the ensemble members. This means that the ocean model within the hindcasts is nudged towards the ocean state obtained from the ocean-only simulation forced with ERA-20C data. By using this approach, the initialization shock is reduced because the same ocean model is used in both the hindcasts and the assimilation run, ensuring consistency in model physics. A similar approach using MPI-ESM has been followed in Borchert et al. 2018.

To draw attention to this point, we added in Sect. 2.2 (L100-101): “To help reduce initialisation shock, the ocean state is derived from an ocean-only simulation performed with MPI-OM forced with the atmospheric variables from ERA-20C, thus maintaining consistency in model physics.”

L102. ..at lead times of 3-4 months..

Text modified as suggested (L108).

L106. I would say: "... the second principal component (PC) of the leading empirical orthogonal function (EOF) decomposition of ..."

Text modified as suggested (L112).

L131. Since the technique is pretty novel and can easily generate misunderstandings, I suggest the authors to add some further disclaimer for the reader. In particular, as implicitly stated a few lines above, a limitation of the Causal effect network analysis is that the choice of the variables to be considered is crucial for determining the causality of the link. In this sense, the possibility of a spurious correlation can never be completely excluded. I think this should be made clear in the text, expanding the sentence at L131.

The reviewer has a very good point. We added the text below to convey this information (L144-154):

“We emphasise that the term "causal" should be interpreted cautiously within the context of this study. When we refer to causality, we mean causality relative to the set of investigated variables and under the specific assumptions considered in the PCMCI algorithm (such as the stationarity of time-series data). As a consequence, the possibility of spurious correlations cannot be entirely ruled out. The choice of variables included in the analysis is another crucial aspect for determining the causality of the identified links. Yet, this poses a challenge as including more variables enhances the credibility of causal discoveries but introduces complexities. For instance, accommodating numerous variables and significant time lags to address physical delays, like identifying atmospheric teleconnections, leads to high dimensionality. This, in turn, can significantly affect the reliability of statistical outcomes. Hence, a successful application of CEN requires (such as for any data-driven method), expert knowledge of the underlying physical processes, including relevant variables, time-scales and temporal resolution. For a more detailed understanding of the CEN analysis and the PCMCI algorithm, we refer the reader to Runge (2018), which provides a comprehensive description of these techniques."

Figure 1 caption. Specify the period considered for the SST anomaly. From the text it is spring SSTs and summer EA (L151), but this is not clear for the caption. Also, it is not clear what period panels c, d, e are referring to. In general, since different periods are considered for different variables, a suggestion would be to put the period as a subscript: SSTMA.

Thank you for this comment. We rewrote the caption to highlight which period is being investigated, and included which months are included in the calculation of each index.

"Variability and linear relationships of EA in ERA-20C. a) Positive phase of the EA teleconnection, defined as the second EOF of July-August (JA) SLP. b) Regions used to calculate the NA-SST and SLP indices proposed in Ossó et al. (2018). c) Pointwise correlation of EA index with concurrent JA anomalies of 2-metre air temperatures in the full period (1908-2008). d) Same as c), for JA anomalies of total precipitation. e) Time series of April-May (AM) SST (blue) and JA EA (grey) indices in ERA-20C for 1908-2008, smoothed by a 3-year running mean. f) Running-correlation between AM SST and JA EA indices for a 20-year window. Coloured markers indicate significant correlations at the 95% confidence interval, illustrated by dashed lines."

L151. Linked to the above. Is the spring SST referred to MAM, MA or AM?

We rewrote the sentence as "A Pearson correlation analysis reveals a time-dependent relationship between the AM SST index and the EA in summer (Fig.1e)." (L195-196)

L166. Why not use MAM? Is there some process changing significantly between early and late spring?

We thank the reviewer for this question. First, we chose to analyse bimonthly means of SST to allow a direct comparison with the studies of Ossó (2018 and 2020). Second, we decided to include both March-April and April-May SST indices in the analysis because we found that in ERA-20C both indices can causally influence SLP in summer (Fig.3c,f).

L200. The T2mCE looks like potentially correlated with EA with no lag (from the composite in Fig. 1).. isn't this correlation appearing in the CEN?

We thank the reviewer for raising this point. The contemporaneous correlation does not appear in the CEN. The reason for that is two-fold. Firstly, we aim to identify precursor signals in spring, specifically before the initialisation of prediction systems typically in May, that could forecast the summer EA. Thus, we concentrated on uncovering interseasonal causal links, i.e. those between spring and summer. In other words, we used tau_min=3 and not tau_min=0. Secondly, we chose to perform our analysis using PCMCI (version 4.2 from Tigramite), which cannot identify causal links at lag 0.

L207-8. I appreciate that the author acknowledge that something could be missing from the CEN. Also, could this mean that the observed causal link may be to some extent spurious, since some key process is missing from the CEN?

We thank the reviewer for this question. Indeed, we acknowledge the concept of "causal sufficiency" and the possibility that some key processes may be absent from the CEN, potentially leading to spurious causal links. Adding more variables could potentially alter the network structure, highlighting the dynamic nature of causal inference within complex systems. For this reason, it's crucial to interpret the term "causal" cautiously within the context of our study. When we refer to causality, we do so relative to the set of investigated variables and the specific assumptions considered in the PCMCI algorithm. We recognise that the possibility of spurious correlations cannot be entirely ruled out, given the inherent challenges in variable selection and the complexities introduced by including additional variables.

However, we have good reasons to believe that the spring SST index -> summer EA link is unlikely to be spurious. A follow-up analysis including a tropical SST index (suggested as another predictor for summer EA, e.g. Wulff et al. 2017) in the network showed that the link is stable (see figure below).

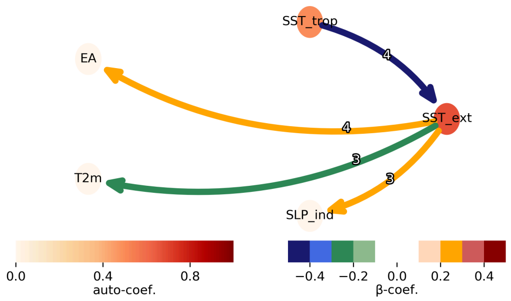

Fig 1: Follow-up analysis showing a causal graph including the additional variable “SST_trop”, following Wulff et al. 2017.

L212. I wouldn't call this "skill in reproducing the summer EA" since Fig. 4a is evaluating the pdf of the EA index, so just from a statistical/climatological point of view.

Indeed. We rewrote the sentence following the reviewer's suggestion as "We find that MPI-ESM generally captures the range of variability, although its performance in replicating the summer EA varies across different simulation sets." (L258-260).

L216. In what sense is Fig. 4b showing "MR-30 capturing the temporal variability of the relationship in the early period" ? I think the only information is about the spread of the relationship, but I do not see a tendency for a negative correlation in the early period as observed for ERA20C.

Thanks. We rewrote the sentence as "We find that the model shows limited skill, particularly in the late period." (L262-263)

L221. I agree MR-30 looks slightly better for the early period, but still is quite far from ERA20C (the positive correlation in the southern North Atlantic is not significant and does not extend so much North). This is true for the ensemble mean, but have you checked whether some individual member is getting a response closer to the observed relation? This could possibly inform on the "missing" process in the chain.

We thank the reviewer for raising this question. As you can see in Fig. 2 below, there is a great variability amongst ensemble members for correlations between AM SST index and JA SLP, with the majority showing too weak positive correlations, mostly displaced to the west, in comparison to Fig.4.c. This is not surprising for MR-30, which has shown limited performance in reproducing teleconnections in the North Atlantic European region, particularly in summer (e.g. Carvalho-Oliveira et al. 2022).

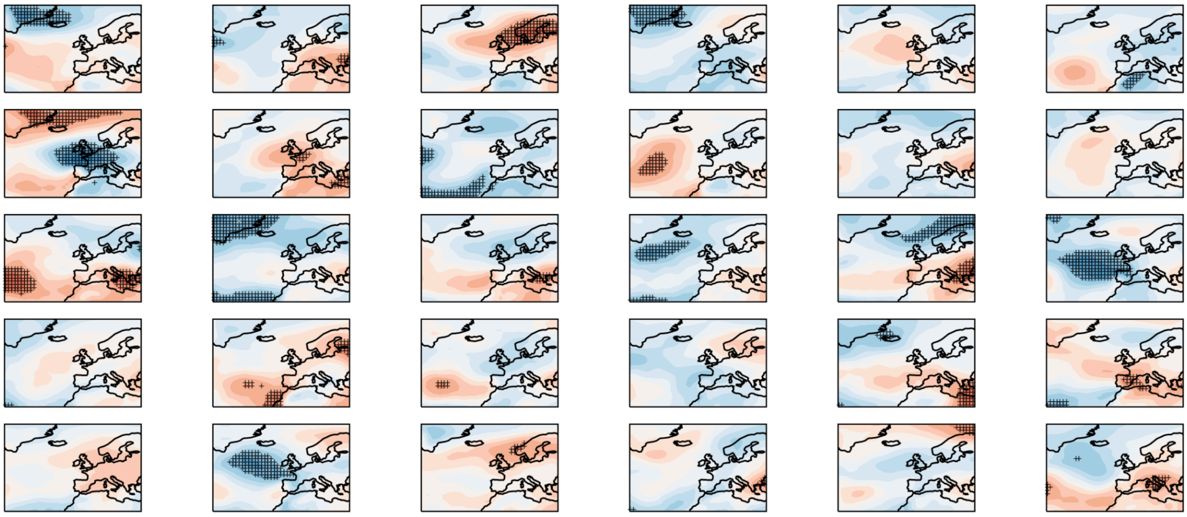

Fig. 2 Correlations between AM SST index and JA SLP for each ensemble member in MR-30 in the full period (1908-2008).

L222. I do not understand "first" in the sentence.

Please note that this section has been rewritten, while addressing comments of the other reviewer. The new sentence is “The variables analysed in the CEN sets are SST, EA and SLP indices and the time lag of interest is spring - summer (3 and 4 months lag).” (L271-273).

L228. 0.03 seems very small. Is it significant?

Please note that this section has been rewritten, while addressing a specific comment of the other reviewer (concerning P9 L222-223). We unfortunately had used the wrong causal graph in Fig.6a by mistake, since over the course of our analysis we also tested the effect for tau_min = 2. The correct causal graph containing the same set of variables but focusing on tau_min = 3 and tau_max = 4 is now in Fig. 6a. As a consequence, the causal link in question has a beta-coefficient of 0.04, and not 0.03. In both cases, the beta-coefficient is significant at the 0.1 level (also see L166 for the chosen CEN parameters).

L245. What do you mean by "MR-30 causal timeseries"? If referred to the "MR-30 causal ensemble" in Fig. 6 caption, I would change the wording to something different.. e.g. MR-30 bootstrap ensemble

We thank the reviewer for the suggestion. A new section in the methodology (2.5: Predictive skill assessment) seeks to clarify this point. Please note that this issue has been already addressed in the reply to “major comments” #2.

L245. How do you perform the "predictive skill assessment"? By selecting random members with beta close to beta_1 and checking the skill only for those?

We acknowledge that this section needed further explanation to improve readability. To conduct this predictive skill assessment, we first identify ensemble members with beta coefficients equal to beta_1, indicating significant relationships between the spring SST index and summer EA. We then assess the predictive skill of our target variable (in the case of Fig.6d, JA SLP) exclusively for these selected years and ensemble members. We hope that including Sect. 2.4 and 2.5 in the manuscript will clarify this step for the reader. We added, in particular “To accomplish this task, we first identify the specific years and ensemble members that correspond to significant β-coefficients for the spring SST and summer EA. With this information, we can then sample JA Z500 to create a time series of similar length. In case more than one ensemble member is randomly selected in a given year, we calculate an ensemble mean. We then determine the ACC between the MR-30 bootstrap and ERA-20C.” (L179-188)

L249. How rare is this? The information might be relevant.

About 1% of the times.

L257. I can't easily see the contours in Fig. 7b. Would be better a separate figure, or black contours.

Following the reviewer’s suggestion, we splitted Fig.7 into two figures, so that the causal map in Fig. 7b can be seen more clearly. Please see Fig. 7 and Fig. 8 in the revised manuscript. We modified the beginning of the caption in Fig 7. as “Spatial features of the causal influence of spring SST index on summer climate. a) (...)” and in Fig. 8 “Does the spring SST index influence summer predictive skill in MR-30? a) (...)”.

L256-L281. I would separate this last part, since it is focussing on a different topic: how does the existence of a causal link between SST spring and T2m Ridge (a different predictand from the rest of the paper) influence the forecast skill on the Ridge region in the random MR-30 ensemble? Also, I had a hard time following the section, which I think should be separated from the rest, better framed and possibly expanded. I say this because the result looks interesting but is quite difficult to grasp from the current text. Also, the choice of the new predictand might look like a cherry pick, but I think it could be better motivated with the fact that it is the only causal link reproduced in the random MR-30 ensemble. The question "what happens to the skill when a causal link is reproduced?" seems relevant.

Following the reviewer’s suggestion, we included a new section “3.5 Forecasts of opportunity: could causality help?” to address these results. Also relevant for this section, we explained in more detail how the predictive skill assessment is achieved in the new Sect. “2.5 Predictive skill assessment”. We also rewrote parts of the text to increase readability.

“We aim to identify a robust fingerprint of spring NA-SST on summer predictive skill, which could potentially enhance targeted forecasting opportunities (Mariotti et al., 2020). Our correlation analysis, as depicted in Fig. 2, indicates the potential influence of spring NA-SST on summer T2m variability across the Euro-Atlantic region during the late period. Thus, we conduct an additional causality analysis in ERA-20C to pinpoint the regions within the T2m field where a causal relationship with spring NA-SST is anticipated. We also explore whether this causal relationship might impact the predictive skill of MR-30.” (L306-310)

L312-316. This part looks a bit technical for the discussion, I suggest to remove it.

Sentence removed as suggested.

References

Borchert, L.F., Müller, W.A. and Baehr, J., 2018. Atlantic ocean heat transport influences interannual-to-decadal surface temperature predictability in the North Atlantic region. *Journal of Climate*, *31*(17), pp.6763-6782.

Carvalho-Oliveira, J., Borchert, L.F., Zorita, E. and Baehr, J., 2022. Self-organizing maps identify windows of opportunity for seasonal European summer predictions. *Frontiers in Climate*, *4*, p.844634.