We would like to thank the anonymous reviewers for their assessments, and the Editor, Dr. Li, for managing our submission and providing guidance on how to proceed. Following their indications, we have grouped the main concerns shared by the reviewers by themes and addressed them together. We agree that substantial revisions are needed in our manuscript and in the responses below we explain in detail how we plan to do it.

The main themes, following the ones highlighted by Dr. Li, are:

- Ranking of the simulations
- Uncertainty and statistical significance
- Diagnostics and metrics
- Introduction and framing of the study
- Structure and content

As suggested by Dr. Li, at this stage we have not replied to the other more specific concerns, but we will gladly address them when preparing the new manuscript. We hope that the responses and proposed changes are adequate and we look forward to revising and improving our manuscript.

Please find below the reviewers' comments in black, followed by our replies in blue, with the main proposed changes highlighted in **bold**.

## Ranking of the simulations

(Rev. 1) I have doubt on the validity of the method used for these rankings, and that these rankings are robust. This is because of the same reason I give above. The relevant differences need to be checked for statistical signficance.

(Rev. 2) I would not associate the ranking with an assessment of the performance. Although it is considered the best representation of the reality, ERA5 is still a model output forced by sea surface temperatures as the ocean is not coupled. Moreover, this ranking is quite subjective as, even if a simulation is ranked high (e.g., between number 23 and 34), it does not mean that it is much "worse" than the HighResMIP simulations as the difference with the closest HighResMIP simulation may be very small. This difference value is not taken into account in the ranking. A metric using the actual range of HighResMIP may be more useful. As an example, in line 265, I would not say "they perform surprisingly well", I would say that they are much closer to ERA5 than the other simulations.

(Rev. 3) [...] I see no added value in ranking the results and comparing them systematically with HighResMIP and lower-resolution simulations – a rather tedious exercise that is both questionable and yields no additional insight. For starters, how does one assess the significance of such a ranking?

(Rev. 3) Discard or minimize the ranking: I think the authors should really reconsider devoting so much space to describing the specific absolute and relative ranking of each simulation, since differences may correspond to minute differences in the metric itself and the authors do not test for the significance of these differences (thus in the rankings).

Given that they are only dealing with three EERIE models and that all three simulated responses as well as that in ERA-5 can be shown in a single panel (e.g., as in Fig. 1 and 3), which allows for clear visual assessment of the differences, it seems odd to use the scatter plots, which are based on single subjective metrics, to assess the performance of the EERIE models. The "longitude" metric is particularly problematic since determining the maximum location of a pattern (especially precipitation) is often ambiguous. For instance, according to Fig. 2, HadGEM has a negative velocity potential maximum in the central Pacific "that is too strong and displaced eastward", but I have trouble seeing the displacement by eye in Fig. 1. And as I mention in the next comment, the discrepancies in the Maritime Continent seem a lot more concerning to me.

I would suggest the authors start by showing the first two figures and then skip to Figure 10 (but without the specific rank values, which seems to suggest they believe those numbers are meaningful). In this way they would be including the HighResMIP simulations in their comparison, for completeness, but they would not need all the scatter plots and the accompanying tedious descriptions of the rankings. They could then simply refer to that figure after analyzing each variable to illustrate whether the EERIE models perform better than HRMIP-HR or not. Skipping the comparison of the metrics based on latitudinal averages, subjective boxes, pairs of boxes, etc., would go a long ways towards making the manuscript more digestible, interesting and objective. It would also eliminate questions about how the confidence levels (error bars) were estimated.

We acknowledge the reviewers' concerns about the rankings presented in the manuscript. They were originally conceived as a simple way to provide an overview of the results, which had to be interpreted within the limitations imposed by our choice of metrics, models and all the other aspects detailed in the Discussion. However, we understand that they may be perceived as too subjective. We thank the reviewers for pointing out these issues and we will completely revise this part in the new manuscript.

**Specifically, we propose to remove all the parts concerning the numerical ranking, including Figure 10.** Instead, we plan to simply describe, for each metric, whether the EERIE models appear (in)consistent with the ERA5 reference, and compare them to the HighResMIP range, similarly to what is done already for the atmosphere-only simulations. As discussed in detail below, uncertainties and statistical significance will also be addressed.

We believe that the comparison with ERA5 is a reasonable way to assess whether the models can produce a realistic response. We stress that the large-scale circulation in ERA5 is constrained by data assimilation, including millions of satellite retrievals, which indirectly account for atmosphere-ocean interactions despite the lack of a properly coupled ocean. It is customarily considered as the observational reference in most studies and we think that, bearing in mind its limitations, it can be used here to evaluate the models.

Similarly, we consider the comparison with HighResMIP a valuable part of our study, since it provides a benchmark for the EERIE models. Our objective is to assess whether these novel eddy-rich models are qualitatively and/or quantitatively different to previous-generation models, and we thus believe that, even without the "ranking", the scatter plots and figures with the full comparisons constitute an important outcome of this work. To address Rev. 3' concern on the length of the analysis, we propose to remove some metrics from the revised version (see below).

We emphasize that removing the rankings will change the structure and partly the tone of our manuscript, but it will not affect our main results, namely that we don't find evidence for large qualitative differences in the behavior of eddy-rich versus non-eddy-rich models.

## Uncertainties and statistical significance

(Rev. 1) Lines 270 - 272, 294 - 296, 298 - 300, 301 - 302, 321 - 323, 326, etc: My criticism here is about the comparisons with ERA5 (or between model experiments). I assume the qualitative (better, best, worse, etc) comments from the comparisons are based on the mean values. The confidence levels, although shown on the figures, are not used in the evaluation nor discussed. There are substantial overlaps for these intervals.

I think at the very least there should be suitable statistical tests on the difference of the means between data with different variances. Statistical tests for climate teleconnections research is a standard thing to do. In the current form of the analyses presented, I am unable to form an opinion for the validity of these evaluations.

(Rev. 1) Figures 2, 7: Are the longitudinal locations indicated the mean locations? What are ranges of variations in the longitudinal locations? Are the differences locations statistically significant?

(Rev. 2) Could the authors explain in more details how the error bars are calculated? It is very unclear to me.

(Rev. 2) For IFS-LR-AMIP, I would have expected a use of the members in the calculation of the uncertainty, but it does not seem to be the case. Also, since there are no error bars on

the HighResMIP simulations (coupled and AMIP), does it make sense to include the error bars when comparing with the ensemble range (see my comment below for line 296)?

We understand the reviewers' concerns about the lack of targeted discussion about some aspects of statistical significance, and we apologize for the confusion around the error bars. **In the new manuscript, we intend to revise the confidence intervals and assessment of statistical significance in our figures and modify the text accordingly.**

We will apply a bootstrapping technique to all the metrics in Figs. 2, 4, 6, and 7 to update the error bars for the amplitudes, which are currently based on t-tests. In Figs. 2 and 7 (scatter plots) we will add error bars for the longitudes, which are currently missing. Additionally, the differences between the models and ERA5 will also be tested for significance with the same method.

Figure R1 is an example of what these updated figures would look like. It combines updated versions of Figs. 7a,b (top) and Figs. 6c,d (bottom), merged into a new figure (see details below). In this figure, the estimates from the linear regressions (same as the original figures) are complemented with confidence intervals calculated from the 2.5th and 97.5th percentiles of the empirical distributions obtained from 1000 bootstrapped regression estimates. For the amplitude, these confidence intervals replace the previous ones estimated with a t-test; for the longitude, they constitute a new addition. Furthermore, we indicate whether the difference with respect to ERA5 is statistically significant (95% level), based on the same method: outlines indicate significance for the difference in amplitude and small dots inside the markers indicate statistical significance for the difference in longitude. Simple markers, with no outline and no dot, mean that the difference with ERA5 in both amplitude and longitude is not significant.

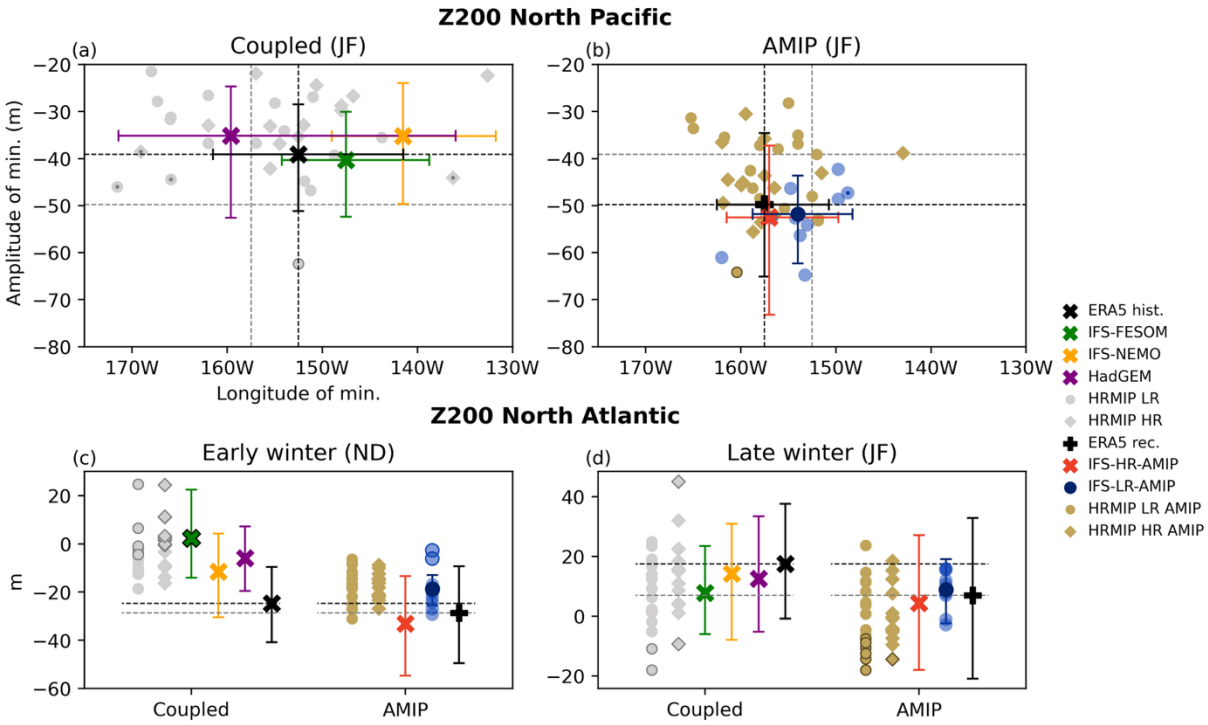


**Figure R1.** Top: Longitude and amplitude of the minimum in the 200-hPa geopotential height anomalies averaged between 30°N-50°N and regressed on the N3.4 index, in Jan-Feb. Left: Coupled runs. Right: AMIP runs. Bottom: linear regression on the N3.4 index of the 200-hPa geopotential height dipole in the North Atlantic, as described in the main text, in Nov-Dec (left) and Jan-Feb (right). ERA5 hist. =1950-2014, ERA5 rec.= 1980-2023. The light blue circles indicate the single members from the EERIE IFS-LR ensemble, while the dark blue circle represents the ensemble mean. The error bars represent confidence intervals at the 95% level based on bootstrapping. Outlines indicate statistical significance for the difference in amplitude between the models and ERA5; small dots inside the markers indicate statistical significance for the difference in longitude.

It is evident that the statistical significance is often limited, as it was already suggested by the large error bars shown in the original manuscript. Only values relatively far from ERA5 are significantly different, while closer values are inevitably not significant, which constitutes a limitation when trying to assess the model's fidelity. Indeed, we were very cautious in drawing conclusions in the final section of our manuscript, where we discussed extensively several limitations of our assessment, including large uncertainties that are intrinsic to the ENSO atmospheric response. This limitation was also highlighted in our abstract. In the revised manuscript, we will make sure to comment systematically on the significance (or lack thereof) and add a dedicated point in the discussion. We believe that this will improve our analysis, though our main results and conclusions will remain unaffected.

Concerning error bars for the single atmosphere-only IFS runs, the members are implicitly used in the sense that the average of the ensemble mean is more certain than the average over a single member; we propose to emphasize this in the new manuscript. Additionally, the statistical significance of the single members will be indicated in the new figures.

## Diagnostics and metrics

(Rev. 2) Although the manuscript is of interest, the methodology used to compare the simulations between each other and with ERA5 is not very convincing. I am not sure that averaging the response fields in boxes, sometimes even taking the difference between two boxes, is the best way to compare the simulations.

(Rev. 2) Concerning the regional averages:

The boxes do not always fit the location of the anomalies. For example, in Fig. 5a, the southern box in the North Atlantic sector overlaps with both positive and negative anomalies. Isn't it a problem? The values in this box would be closer to 0 than if a more southward box was chosen. Moreover, the authors add again uncertainty by taking the differences between the averages in the two boxes over the North Atlantic domain. Personally, I would find more interesting, and maybe more robust, to look at pattern correlations within basins (North Pacific, North America, and North Atlantic) to evaluate the representation of the teleconnection in comparison with ERA5.

(Rev. 3) Second, the boxes used to define the metrics appear highly subjective. Comparing spatial correlations would have made more sense to me, since at least the differences can be tested for statistical significance and are less sensitive to the choice of domain.

We thank the reviewers for expressing their concerns. We acknowledge that the exact values estimated by our metrics can be sensitive to the details of the regions, whose definitions are based on the observed patterns. However, the fact that same metric is used for all the models and for the observations provides a fair comparison.

Our rationale to build the diagnostics is that they should encapsulate the main aspects that we are interested in, based on the common understanding of the key processes involved in the ENSO teleconnections. For example, we have decided to focus on the anomalous divergence in the central Tropical Pacific because this is the main trigger of the Rossby wave train, while the accompanying anomalies over the Maritime Continent and South America are considered as modulating factors. Similarly, the current knowledge of the ENSO response in the North Atlantic indicates that a dipole in geopotential height/sea level pressure dominates in both early and late winter, but with opposite sign. We thus selected the boxes to compute the related metrics so that we could encapsulate the response in both

seasons, considering that the centers of actions show slightly different locations in ERA5 among the two seasons.

To address the concern about the choice of boxes in the North Atlantic and provide an example of the sensitivity of our metrics to the exact boundaries, we have repeated this analysis using different regions. Figure R2 shows the alternative boxes, which have been adjusted to better match the location of the dipole lobes in ERA5 and avoid the overlap between positive and negative values. In contrast to the analysis presented in the manuscript, different coordinates are used in early and late winter (cf. Fig. 5a,e).

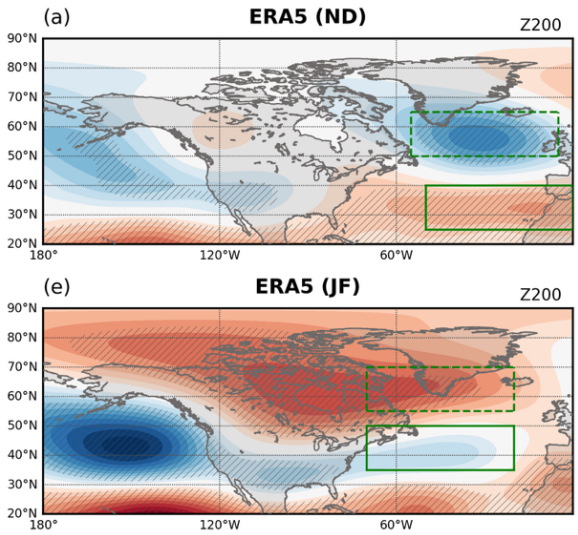


**Figure R2.** Linear regression on the N3.4 index of 200-hPa geopotential height anomalies in Nov-Dec (top) and Jan-Feb (bottom) for ERA5. The green boxes show the regions used to compute the metric shown in Fig. R3.

Figure R3 shows the values of the corresponding metrics for ERA5 and the EERIE coupled models, based on the difference between the two boxes. The observed values have greater amplitude (cf. Fig. R1c,d), which is consistent with the choice of more targeted boxes. However, the qualitative conclusions for the EERIE models are similar, namely the fact that their early-winter response is inconsistent with ERA5, with significant differences for IFS-FESOM and HadGEM, while it is more consistent in late winter, though with large uncertainty.

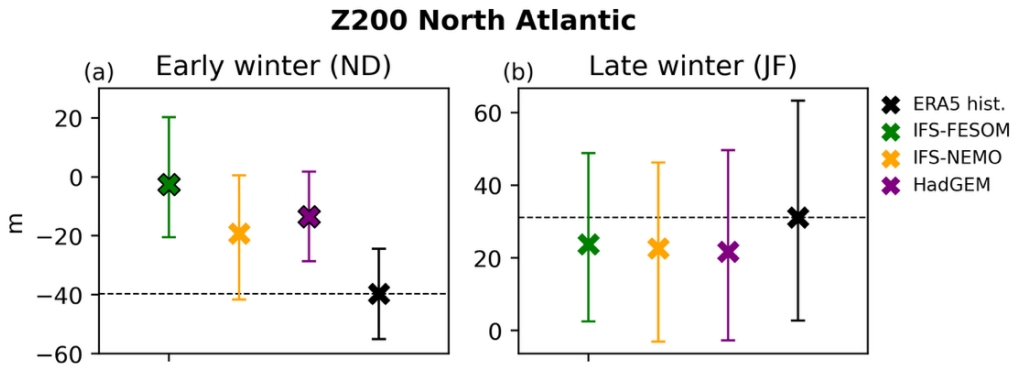


**Figure R3.** Linear regression on the N3.4 index of the 200-hPa geopotential height dipole in the North Atlantic, in Nov-Dec (left) and Jan-Feb (right), using the boxes shown in Fig. R2. The error bars represent confidence intervals at the 95% level based on bootstrapping. Outlines indicate statistical significance for the difference in amplitude between the models and ERA5.

We also acknowledge the suggestion to look at pattern correlations, which we believe is a complementary approach that would provide different insights: by design, area-averages miss the spatial structure, but correlations miss the area-averaged contribution. For some metrics we already show spatial correlations in our manuscript, specifically in the analysis of the North Atlantic sea-level pressure (SLP) signal. The correlations are shown as angles in the Taylor diagrams in Fig. 9. As discussed in the text, this surface response is tightly linked to the patterns of 200-hPa geopotential height, for which we use one of our tailored metrics. Both the SLP correlation patterns and our metric suggest similar conclusions for the EERIE models, namely that they share the same issues as the HighResMIP models in early winter but struggle less in late winter.

To further address the reviewers' concerns, we have performed a similar analysis of the spatial patterns of the 200-hPa velocity potential shown in Fig. 1 of the manuscript (contours). Figure R4 depicts the spatial correlations (angle) and standard deviation ratio (distance from the origin) between the patterns simulated by the EERIE coupled models (Fig. 1b-d) and ERA5 (Fig. 1a). The results are similar for IFS-FESOM and IFS-NEMO, which have correlations of 0.87 and 0.81, respectively, and partly underestimate the standard deviation. In contrast, a lower correlation (0.78) and high standard deviation ratio place HadGEM further away from the observational reference. These results agree with and complement what we had found using our metric based on the minimum of the velocity potential (Fig. 2 in the original manuscript; Fig. R5 showing the EERIE models only for easier comparison). This diagnostic confirms the similar behavior of IFS-FESOM and IFS-NEMO, with underestimated amplitudes, while HadGEM shows the opposite.

We emphasize that our metric focuses on one specific part of the response, while the pattern correlation is an assessment of the similarity over the entire region shown in Fig. 1. However, both methods suggest similar conclusions regarding how consistent with the observations is the response in the tropical Pacific in the EERIE models. We propose to include this assessment in the revised manuscript and expand the text describing Figure 1 accordingly.

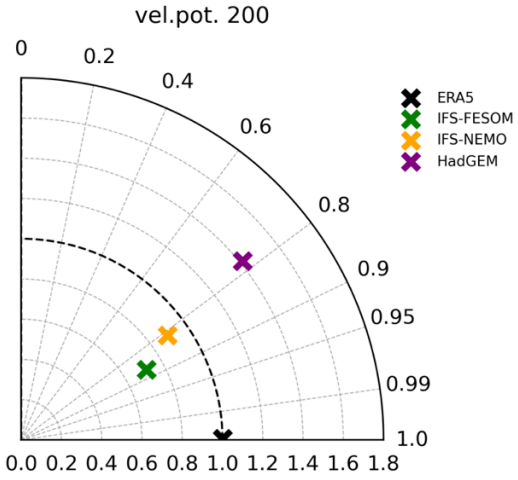


**Figure R4.** Taylor plots for the anomalous patterns of 200-hPa velocity potential over the region shown in Fig. 1 (60-360°E; 45°N-45°S). EERIE coupled models compared to ERA5 over 1950-2014.

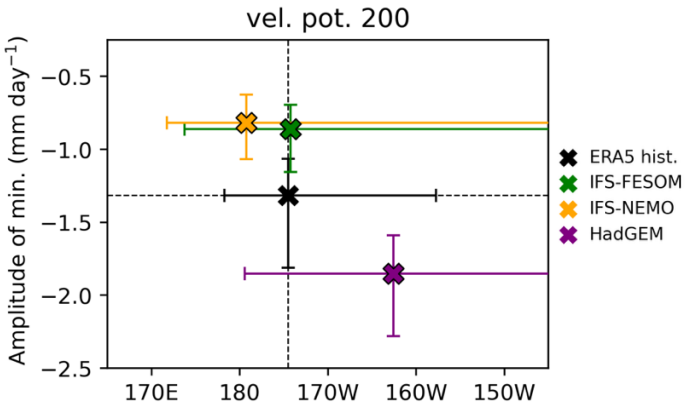


**Figure R5.** Longitude and amplitude of the minimum in the 200-hPa velocity potential anomalies averaged between 15°N and 15°S and regressed on the N3.4 index, in Nov-Dec. As in Fig. 1 of the manuscript but showing the EERIE coupled models only and with a corrected value for IFS-FESOM.

## Introduction and framing of the study

(Rev. 3) A more compelling introduction is needed. Since I am recommending a major overhaul of the manuscript, I will not point out every instance of wording that could be improved, but I will do so for the Introduction, which I believe requires a thorough rewrite and restructuring to make it clearer and more engaging. Here are some specific issues:

- The opening sentences are very vague. What is meant by "different climate processes"? What is meant by a "variety of ocean regimes"? It does not make for a very compelling opening.
- There is some confusion as to whether the text is referring to oceanic or atmospheric resolution. The first introductory sentence alludes to "horizontal resolution of Earth system models", but the rest of the first paragraph and most of the second one refer to ocean resolution only. Atmospheric resolution is suddenly mentioned in line 45, with no previous reference to its importance or benefits.
- It takes two paragraphs to get to ENSO, which is the focus of the paper, but even so we are told nothing about the specific oceanic processes that may be better represented in eddy-rich models and result in more realistic ENSO SST patterns (only that there are "serval (sic) ways through which a better resolved ocean could affect ENSO teleconnections"). Maybe you could cite the recent Siqueira and Kirtman (2026) paper and discuss improvements in the basic state, cold tongue, SST-wind feedbacks.
- Likewise, the reader would appreciate some discussion of what better resolved atmospheric processes and aspects of the atmospheric circulation are expected to contribute to a more realistic ENSO teleconnection (sharper vorticity gradients? more accurate waveguides? synoptic eddy feedbacks? )
- Lines 58-64: Stating that the improvements in the high-resolution simulations were attributed to the increased ocean resolution is underwhelming – unless the authors are trying to say that the increased atmospheric resolution itself did not play a role (if so, it is unclear)
- Lines 58-64: Should there not first be a discussion of how the simulated ENSO teleconnection is flawed in current models so that we may understand the importance of "a more accurate representation of the position of the North Pacific response" while "no improvement in the strength of the teleconnection" ?

(Rev. 2)

- Line 57: "a more realistic extra-tropical response": this is a bit vague. What is the response the authors are talking about? In which basin? And how is it "more realistic"?

- Line 60: “the position of the North Pacific response”: do the authors mean the centre ofthe mean sea level pressure anomaly or something else? Please be more precise.

- Line 65: “more accurate late-winter ENSO teleconnections”: again this statement deserves more details. How is the accuracy measured and which field is used to assess the teleconnection (sea level pressure, precipitation,...)?

- Line 68: “the extra-tropical teleconnection”: what field do the authors refer to? And in which basin? North Atlantic?

(Rev. 3) Finally, while I understand that the paper is intended as an assessment of these eddy-rich simulations with regards to ENSO, the manuscript as it stands is quite dry. It would benefit from some dynamical explanations both for context and as possible explanations for the lack of improvement in the ENSO simulation. For instance, a brief explanation of the mechanisms that affect the polar stratospheric vortex during ENSO would be welcome, as would attempting to connect the deficiencies in the TRWS and in the ensuing teleconnections to biases in the wave-guide or basic-state vorticity gradients.

We thank the reviewers for highlighting these issues in our Introduction and for the valuable suggestions on how to improve it. **We plan to fully revise this section to improve its engagement and clarity, with special attention in contextualizing our study within the existing literature. Similarly, we will include throughout the entire text more detailed discussions of the dynamical mechanisms and implications.**

We would like to emphasize that we had already carried out a careful review of the existing literature on the topic, and, to the best of our knowledge, there is no consensus on specific oceanic and atmospheric processes that may be better represented in eddy-rich models that consequently impact ENSO teleconnections. In the revision, we will try to list some of the speculated mechanisms, including those discussed in the Siqueira and Kirtman study recommended by the reviewer, which was only published after our initial submission.

Specifically, we will clarify the dynamical mechanisms of ENSO teleconnections in the atmosphere (tropospheric Rossby waves and stratospheric pathway) and how changes in model ocean/atmosphere resolution could potentially influence both the ENSO forcing and its teleconnections, including (but not limited to): changes in eddy activity, which modifies tropical/extratropical SST biases and sub-surface ocean biases; changes in atmospheric mean state, which impacts tropospheric/stratospheric atmospheric waves guides; changes in extratropical air-sea interactions and interactions with position/intensity of storm tracks; increased atmospheric resolution and synoptic transient eddy feedbacks.

## Structure and content

(Rev. 1) Supplementary figures: There are ten supplementary figures and they are cited extensively in description of the results and discussion in the main text. I don't think this is a correct use of supplementary figures. This affects the readability of the paper and processing of the information on the part of the reader. More attention is needed on what are the key results and focusing on presenting them nicely in the main body of the manuscript.

(Rev. 3) Another issue, in my opinion, concerns the order of presentation of the results, although I am somewhat ambivalent on this point. On the one hand, one could argue that, following common practice, results should be presented in order of increasing complexity, beginning with the simpler AMIP simulations. On the other hand, such an approach would make it somewhat awkward to subsequently motivate the analysis of the coupled simulations, where most of the effort clearly lies, given their poor performance. In that sense, the current structure is understandable, and it is not obvious to me that a better alternative exists. It does raise the question of why one would have expected the presence of mesoscale activity in the ocean to improve the atmospheric response in the first place, although this seems to stem more from the design of the overall project than from the presentation choices made in the paper itself.

(Rev. 3) At the very least, substantial shortening is needed, and perhaps some re-ordering as indicated above, if not of the paper, then of the motivation for the project.

(Rev 3) Stratospheric polar vortex: This section cannot rely only on supplementary figures alone. Since I suggest you eliminate the scatter plots, Fig. S6 could be moved to the main text. I would suggest, however, that you elaborate on the "expected deceleration of the polar vortex", so the reader can follow. Also, what is happening in HadGEM? Is this model LOWTOP? Why is the signal discontinuous?

We thank the reviewers for expressing their concerns about the structure of the manuscript and the figures. We detail here how our plan for the revisions considering their insightful suggestions:

- Removing all the parts related to the ranking, as explained above, will substantially shorten the paper in terms of text, figures (Fig. 10) and supplementary material (Fig. S7).
- We propose to reduce the number of metrics by removing some panels from Fig. 6 and 7. Specifically, we plan to remove Fig. 6a, which shows the metric based on the amplitude of the early-winter response of the 200-hPa geopotential height in the North Pacific. In contrast to the North Atlantic, where the early- and late-winter responses are opposite in sign, the early-winter signal in the North Pacific is simply a

weaker version of the late-winter one, since the Rossby wave train is still developing. We find it sufficient to comment briefly on the spatial patterns of the EERIE models (Fig. 5) without the need for a dedicated metric. Similarly, we also find that the metric based on the high latitudes (Fig. 7b,c) does not add insightful information. Since this response is also contributing to the metrics in the North Atlantic (Fig. 6c,d), we propose to remove these two panels. The text would be shortened accordingly.

- Consequently, we propose to merge Figs. 6c,d and 7a,b in a new figure (see Fig. R1), and to restructure the Results by creating two sub-sections on the North Pacific and North Atlantic instead of the current ones focusing on early winter and late winter.
- We propose to move Fig. S6 to the main text and merge it with Fig. 6a, thus creating a new figure fully dedicated to the response of the stratospheric polar vortex. We would expand the text and provide more details on the response.
- We propose to further reduce the number of supplementary figures by removing Figs. S8-11, which show some of the key metrics as differences with respect to ERA5 and thus do not really provide additional information compared to the corresponding figures in the main text.

Concerning the question about the HadGEM model: the top is at 85 km, but we did not manage to output all variables in order to plot the figure up to this level. We will indicate this in the revised manuscript.