Response to reviewers' comments.

We hereby provide a detailed response to all of the comments given by the two reviewers to the manuscript "Revisiting the Historical Drying of the Mediterranean in the LESFMIP Simulations".

The reviewer's original text is presented below in bold, and our responses are given inline.

Reviewer 1

Reviewer comments to: Avisar and Garfinkel, submitted for publication in Weather and Climate Dynamics

This manuscript was submitted to another journal and was unfortunately rejected from there. This is my second time reviewing this manuscript. As I saw earlier, the potential strength of this study is that it used new LESFMIP datasets and attempts to follow up on the physical mechanisms proposed in previous work.

However, I will have to be more critical this time, because the paper was resubmitted without addressing most of my comments, even simple ones (and those of other reviewers). My greatest concern is that there is a mismatch between the analyses and motivation for using the LESFMIP datasets. Major improvements can be made. My comments are in detail below.

Thank you for reviewing this manuscript, and we apologize for any misgivings. We have added figures, text, and analysis that directly address the reviewers' comments on the comparison of LESFMIP to observations, and also that better isolate the role of all of the forcings that are included as part of LESFMIP. The specific changes are described below in detail, in line with the reviewers' comments.

Major Comments

1.One of the advantages of using historical forcing experiments as LESFMIP has over using future forcing experiments is that model fidelity can be tested with observations. Are the LESMIP simulations reliable in simulating the historical climate? Authors do calculate model PSL biases in the supplementary materials, but the comparison is not done properly. The biases are calculated between (i) observations vs single-forcing experiments, and (ii) the comparison does not account for the role of internal variability.

If the authors do not have all-forcing experiments, can they show at least by adding all single-forcing experiments that LESFMIP well simulates the historical observed climate? If the LESMIP simulations do, it will make the analyses more powerful, and if they don't, it is an important result to be documented.

The reviewer points out the importance of comparing the all-forcings historical LESFMIP experiment with observations, and we fully accept the comment. The all-forcing historical experiments for the various models are available and we now directly compare the observational trends to these runs. We also compare biases in historical to observations. In addition, we discuss the cancellation of the aerosol induced response and GHG induced response throughout.

The internal variability was taken into account in the scatter plots of the original submission but not quantified. We now more fully quantify it. Furthermore, we have added a figure comparing the distribution of trends in Mediterranean PSL across all historical members to observational trends.

Specifically, along with the single-forcings multi-model responses maps, we include the responses obtained in the historical (all-forcings) experiment:

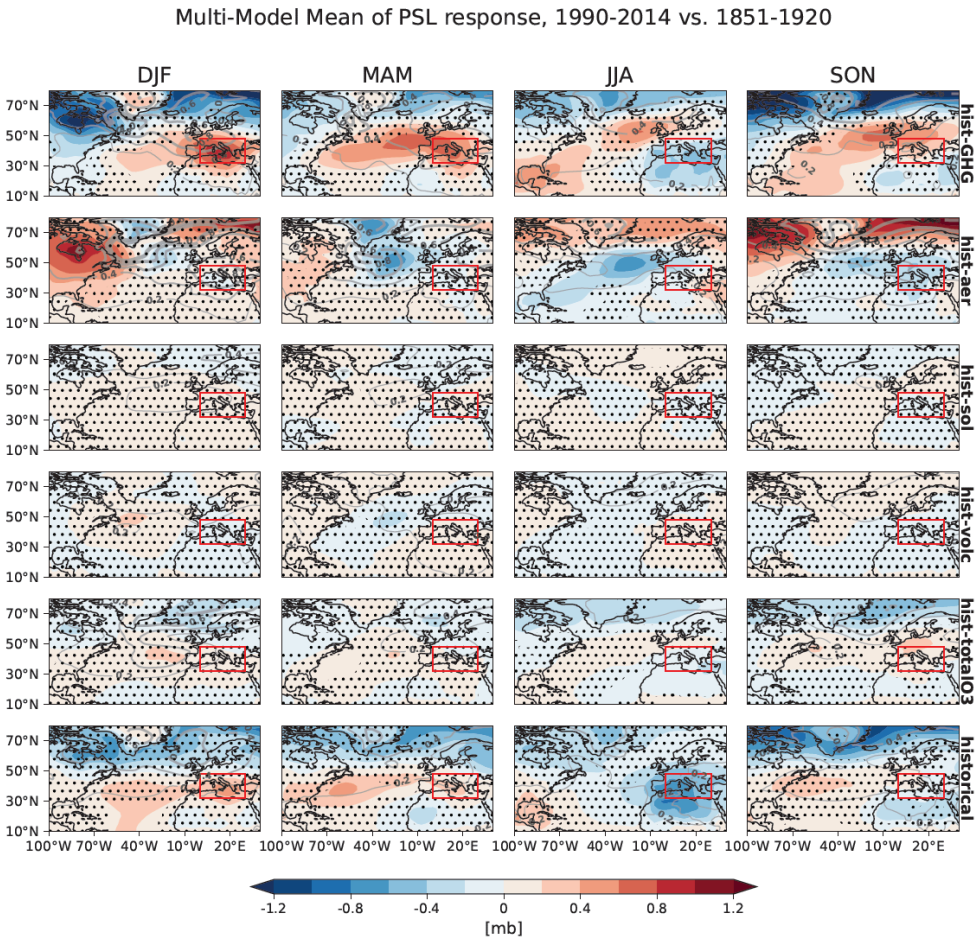


Figure 1. Multi-model means of the seasonal (column-wise) PSL responses (1990-2014 vs. 1851-1920) in the single-forcing experiments (row-wise). The red rectangle marks the Mediterranean region as defined in the main text. Regions without stippling indicates where at least 80% of the models agree on the sign of the response. A similar figures but for different averaging periods is shown in Supplemental Figure S1.

In addition, we added a corresponding figure, Figure 2, for the PSL response maps based on observations:

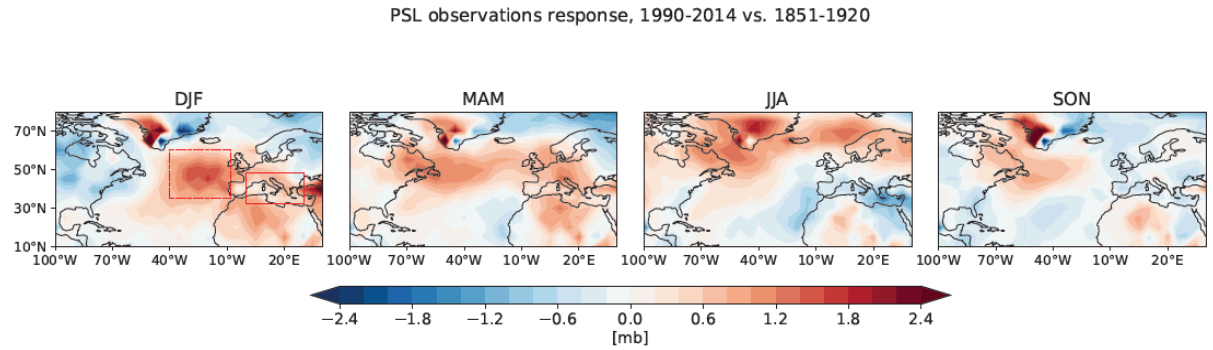


Figure 2. Seasonal (column-wise) PSL responses (1990-2014 vs. 1851-1920) in the HadSLP observations.

and a figure that compares the Mediterranean DJF distribution of the PSL responses in the all-forcings LESFMIP experiment with the observed response, Figure 3a (as well as compares the all-forcings DJF distribution of the PSL response with observations for the North Atlantic region where the observations show a more pronounced ridging signal, Figure 3b):

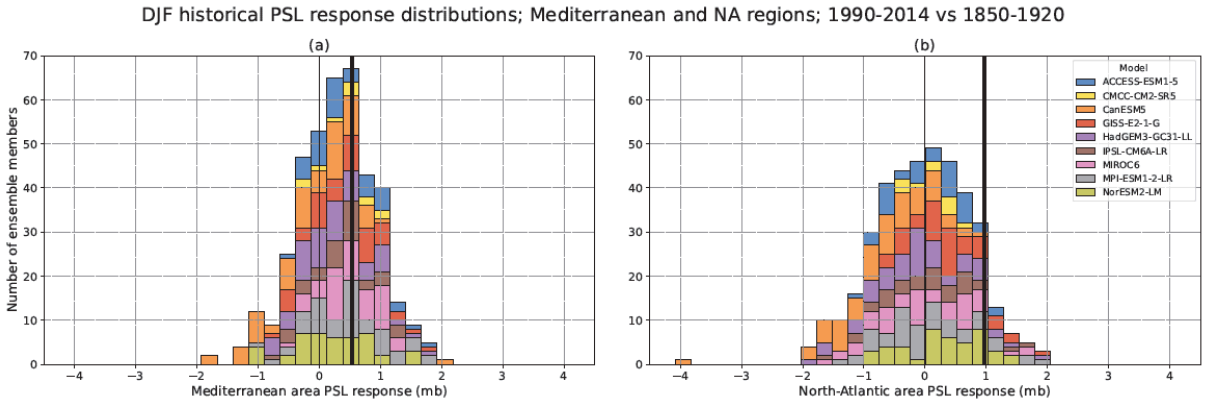


Figure 3. Distribution of models' (color-coded) ensemble members in the Mediterranean (left; solid red box on Figure 2) and North Atlantic (right; dashed red box on Figure 2) DJF PSL responses (1990-2014 vs. 1851-1920) for the historical experiment. The vertical thick black line refers to the observed response based on the HadSLP.

These results for the historical LESFMIP response vs. the observations are described and analyzed in the text as follows:

> "*Figure 2 shows the observed PSL response between the same periods. A meridional dipole response is evident in DJF and MAM generally resembling the multi-model mean historical all-forcings and GHG-induced responses in the LESFMIP models. However, the signals differ in magnitude (note the different colorbars for Figure 1 and 2, which is not surprising since the observed anomalies include both the forced response and also internal variability. In order to assess whether the modeled response is discrepant with the observed PSL response, in Figure 3 we contrast, separately for two regions, the distribution of the PSL response as obtained in the historical all-forcings LESFMIP experiment for all of the models' ensemble members (color-coded bars) with the observed response (vertical thick black line). For the Mediterranean region (Figure 3a, solid red box on Figure 2) the observed response lies well within the LESFMIP historical distribution, and 36% (143/396) of the ensemble members exceed the observed PSL response. If we instead focus on the region to the west of France in which the observed response is stronger (Figure 3b, dashed red box on Figure 2), the LESFMIP multi-model mean shows a weak response, however 8% (32/396) of the ensemble members nonetheless exceed the observed PSL response. Hence, there is no evidence for a model vs. observations discrepancy over this period.*"

Furthermore, we also include the precipitation response in the all-forcings LESFMIP along with the single forcing responses:

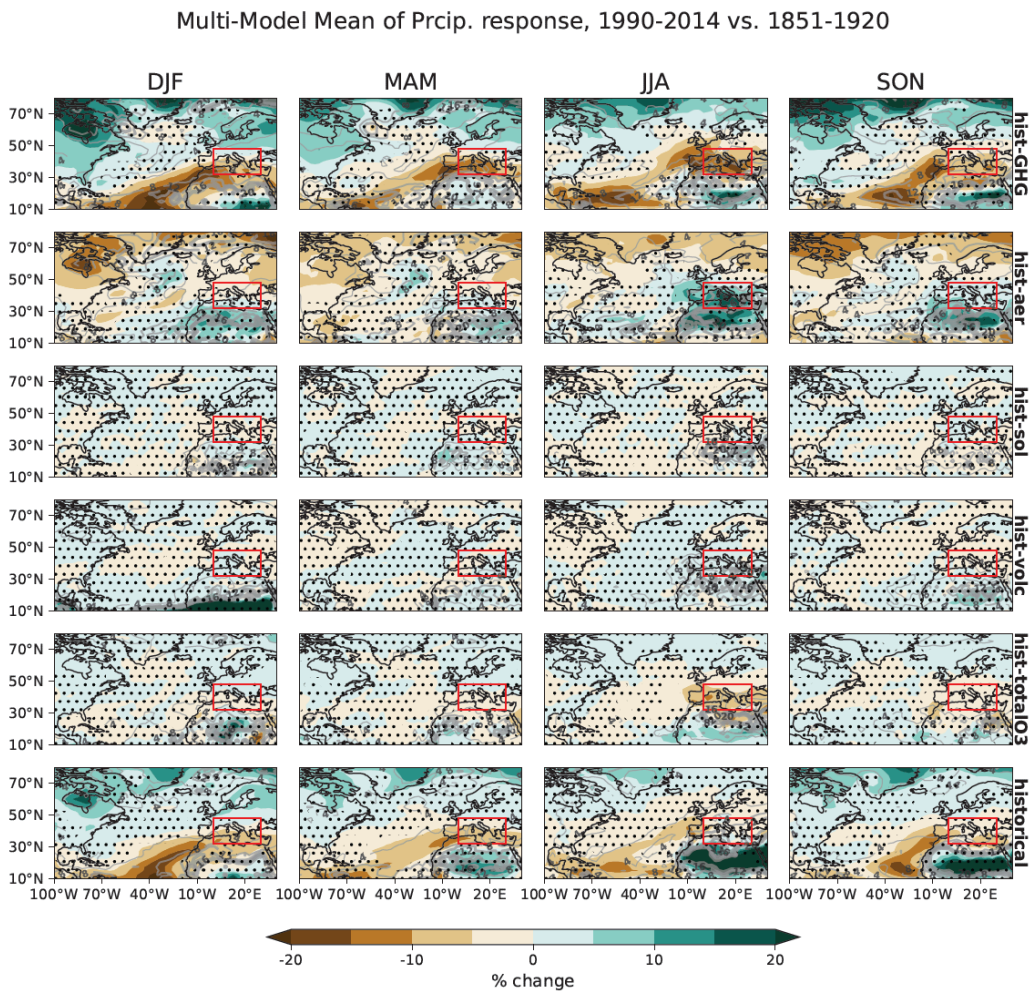


Figure 4. Multi-model means of the seasonal (column-wise) precipitation responses (1990-2014 vs. 1851-1920) in the single-forcing experiments (row-wise). The red rectangle marks the Mediterranean region as defined in the main text. Regions without stippling indicates where at least 80% of the models agree on the sign of the response. Gray contours indicate the intermodel spread in the forced response. Similar figures but for a different averaging period are shown in Supplemental Figure S2.

In addition, since precipitation observations are limited spatially and temporally, in the Supplemental (Figure S3) we present model ensemble-means maps for precipitation trends, based on the all-forcings (historical) LESFMIP experiments, that correspond to the same periods analyzed in [Vicente-Serrano, Sergio M., et al. "High temporal variability not trend dominates Mediterranean precipitation." Nature 639.8055 (2025): 658-666] based on a comprehensive precipitation data from Mediterranean stations

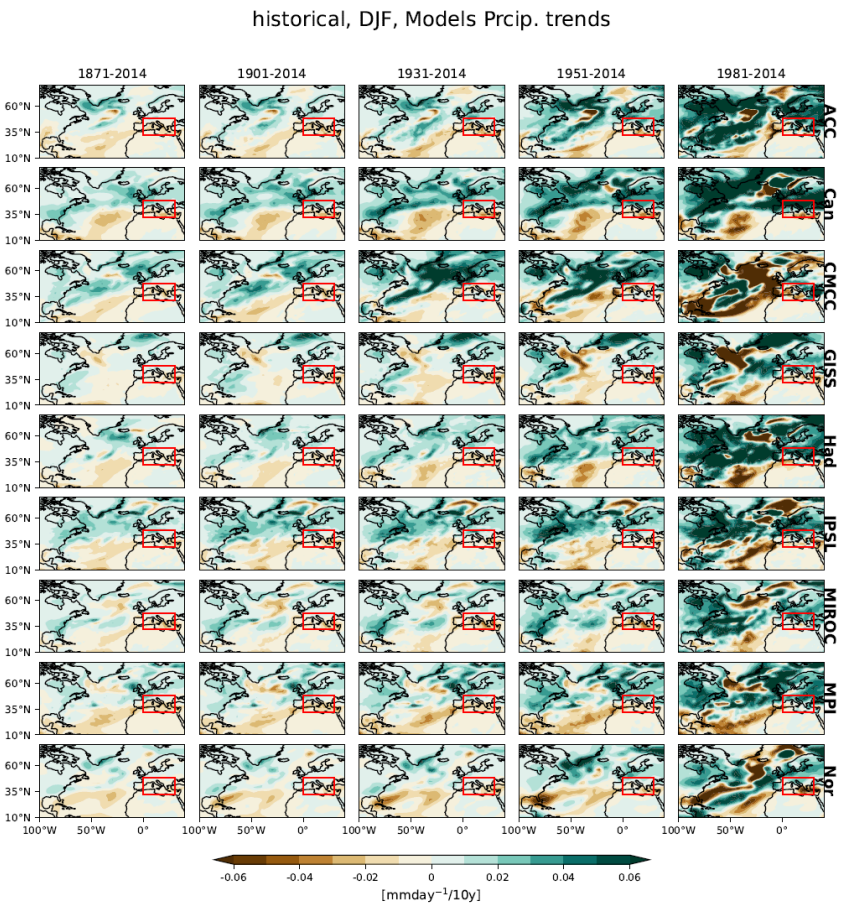


Figure S3. Ensemble means of the DJF precipitation trends [mm/day/10y] for the different models (row-wise) during different time periods (column-wise) based on the historical LESFMIP experiment.

and we note on the text:

> *"In agreement with [serrano_25], recent trends in precipitation (i.e., 1981-2014) indicate wettening in most of the Mediterranean (and not only there) with the majority of the models indicating a positive precipitation trend. Drying trends in the most of Mediterranean in the historical simulations only emerge for start-dates of 1931 or earlier, however many individual models simulate wettening trends over Southern Europe (especially Southern France, Italy, and the Balkans) even for a start date in 1871."*

In addition to comparison with observations, we also relate to the responses seen in the all-forcings experiment in light of the stronger responses in the single-forcings experiments:

> *"The responses in hist-GHG and hist-aer are larger than in any of the other single forcing experiments for both precipitation and PSL (though note a robust response in JJA in hist-totalO3 during summer for precipitation that mimics the GHG induced response). In most regions aerosols and GHGs have opposite signed responses for both*

PSL and precipitation, and hence the historical all-forcings changes are relatively muted and are less robust across models than for GHG alone."

We then summarize the analysis:

> "*Based on the above comparison of the LESFMIP PSL response and the observed one, it may be suggested that the ridging (and drying) signal across the Mediterranean has not yet emerged out of the system's internal variability, in agreement with [serrano_25] and [seager_24], presumably because aerosols have canceled out a substantial portion of the GHGs-induced drying. That is, GHGs and aerosols have had a strong but opposite impact on historical wintertime climate in the Mediterranean."*

My impression is that the LESFMIP simulations are utilized in this study as if they were future forcing experiments, and the analyses are focused on the inter-model spread. My major confusion is that the analyses, or the explanation of the results, don't relate well to the strength of the LESFMIP simulations mentioned in the introduction (that it could be used to study intermodel spread as well as internal variability).

As we noted above in detail, we now show figures for the multi-model means of each individual forcing simulation together with the historical run, and directly analyze it in the context of available observations that are also included now in the text. By that, we utilize the LESFMIP simulations as they were originally intended.

Furthermore, we have added a paragraph to the discussion that directly relates to the reviewer's excellent point that our approach has relevance to storyline approaches to explaining intermodel spread in future climate projections [Shepherd, Theodore G et al., Storylines: an alternative approach to representing uncertainty in physical aspects of climate change, Climatic change 151, 555--571, (2017)]:

> *"Our focus on understanding intermodel spread in historical climate has implications for attempts to build storylines of future climate change that explicitly consider intermodel spread in underlying processes [shepherd2018storylines]. In Europe and the Mediterranean, these storyline approaches couch future projections in terms of how each model simulates, e.g., changes in the polar vortex or changes in polar amplification [Giuseppe Zappa and Theodore G. Shepherd, Journal of Climate, 30, 6561 - 6577, (2017)]. We find no evidence that a stronger global mean warming leads to more pronounced drying in this region, which implies that dynamical effects overwhelm thermodynamic mechanisms [seager_19,elbaum_22]. Furthermore, our results imply that the rate of warming of SSTs in the North Atlantic warming hole region has a stronger impact than either the polar vortex or Arctic amplification on Mediterranean drying, and should be included when building storylines of European and Mediterranean climate. Finally, the regression coefficients derived from hist-aer and hist-GHG are not, in general, quantitatively similar. That is, the correlations and regression coefficients of the*

majority of the large-scale metrics with the Mediterranean PSL response are higher under the anthropogenic aerosols forcing than under the GHGs forcing. This could arise if European sector aerosols have an impact on the local atmospheric circulation. Confirming this requires a more detailed analysis of the historical and hist-aer simulations, including analyzing the potential for reversals of trends since the 1980s when European aerosols peaked. Ongoing work is aimed at clarifying this effect."

2.The title of section 3.3. is 'Towards Understanding the Intermodel Spread'. However, the correlation coefficients reported in Figures 5–7 are across all ensemble members. These quantities include 'intramodel' spread (i.e., correlation between ensemble members of one model), and the number of ensemble members used per model varies. The correlation coefficients in Figures 5–7 conflates spread due to structural uncertainty and internal variability. This section has to be revised, or the section title has to be reframed to match the analyses. For intermodel spread, Table 3 is more relevant. Also, the 'relative' correlation in Table 3 is still hard to understand. What is a meaningful number? In general, structural uncertainty in the forced response and internal variability are not well separated in the analyses, while the motivation for using the LESFMIP in the analyses is based on it. Lastly, the statistical significance of the correlations is not documented.

In order to address the reviewers' comment, we made several changes in Section 3.3. First, in the original Figures 5-7 (Figures 7-9 in the revised submission) we now calculate the correlation coefficients using each model's ensemble mean response (in addition to the correlations calculated based on all available members of all the models). The correlations based on the models' ensemble means are similar, and in fact higher in all cases, relative to the correlations that are based on all of the individual ensemble members of all models. See, for example, Figure 7 below:

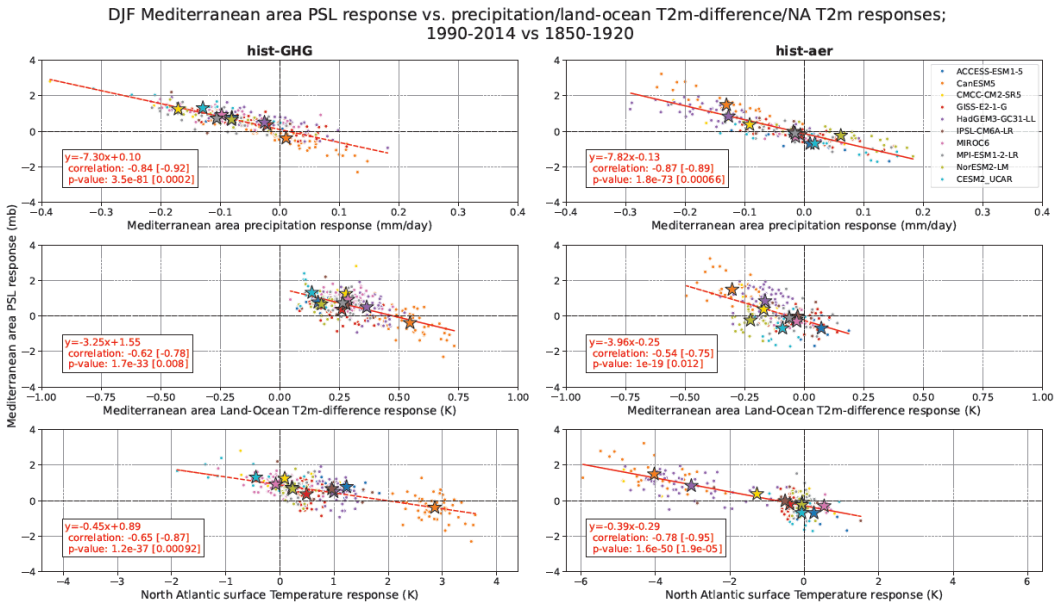


Figure 7. PSL response vs. responses of precipitation (top panels), land-ocean near-surface temperature difference (middle panels), and near-surface temperature within the NA warming-hole region (blue rectangle in Figures 5 and 6; bottom panels) obtained for each ensemble member of the LESFMIP models (color coded dots) for the hist-GHG and hist-aer experiments (left and right panels, respectively). Stars refer to the individual models' ensemble-means. The correlation coefficient and its p-value when using the individual ensemble-members (-means) are indicated outside (inside) the square brackets.

In this way, we account for the intermodel spread.

Second, the "relative" correlations in Table 3 were intended to quantify the role of intramodel spread, but we agree this was hard to understand and confusing. Therefore, to make the analysis and results more clear, we have removed Table 3 and the accompanying discussion. We have also removed understanding intramodel spread as a core aim of the paper. The third change we made in this section accounts for the statistical significance of the correlations, which was not indicated, as the reviewer noted. For that, we now include within the correlations figures significance values (p-values) for the correlations (based on the entire ensemble members and on the ensemble means). Accordingly, all correlations we focused upon in the original submission are significant.

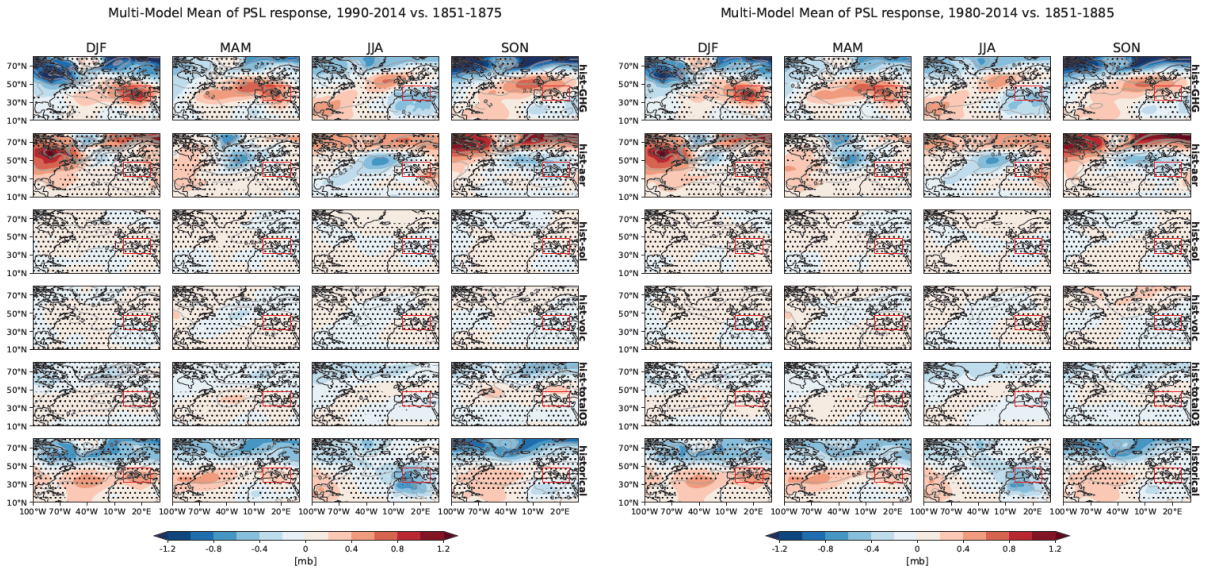
3.Various climate indices from previous work are explored in the manuscript. How the indices are calculated is important, but it would be nice to provide the readers with which

underlying physical hypothesis is being tested here by examining the indices. This will suit the paper better for the scope of Weather and Climate Dynamics.

We agree with the reviewer's comment. Text describing the motivation for each index, including the underlying physical hypothesis and a suitable reference for more details, has been added in the methods section for each index.

4.Further details are required for the time periods chosen here. Why did the authors choose a 70-year window (1851-1920) and a 25-year window (1990-2014) for those particular years? This is important for understanding the signals from the hist-aer simulations. In particular, 1990-2014 is after the European aerosol emissions peaked in 1980. I suggest that the authors better explain why the specific time periods were chosen to quantify the response and clarify the spatial pattern of the aerosol forcing over those periods. This is something I pointed out earlier, but hasn't been addressed.

The choice of long windows was made to maximize the forced signal while averaging out the noise as much as possible. We have added this motivation to the "Data and Methodology" section. In order to further address the reviewer comment and concern, we now include in the supplemental figures for the sea-level pressure and precipitation multi-model mean responses, for the single forcings and the historical, that are calculated based on similar lengths of epochs and where the epochs ranges vary. Accordingly, the results are insensitive to changes in the averaging window, as we also report now in the text. For example, see Figure S1 below:



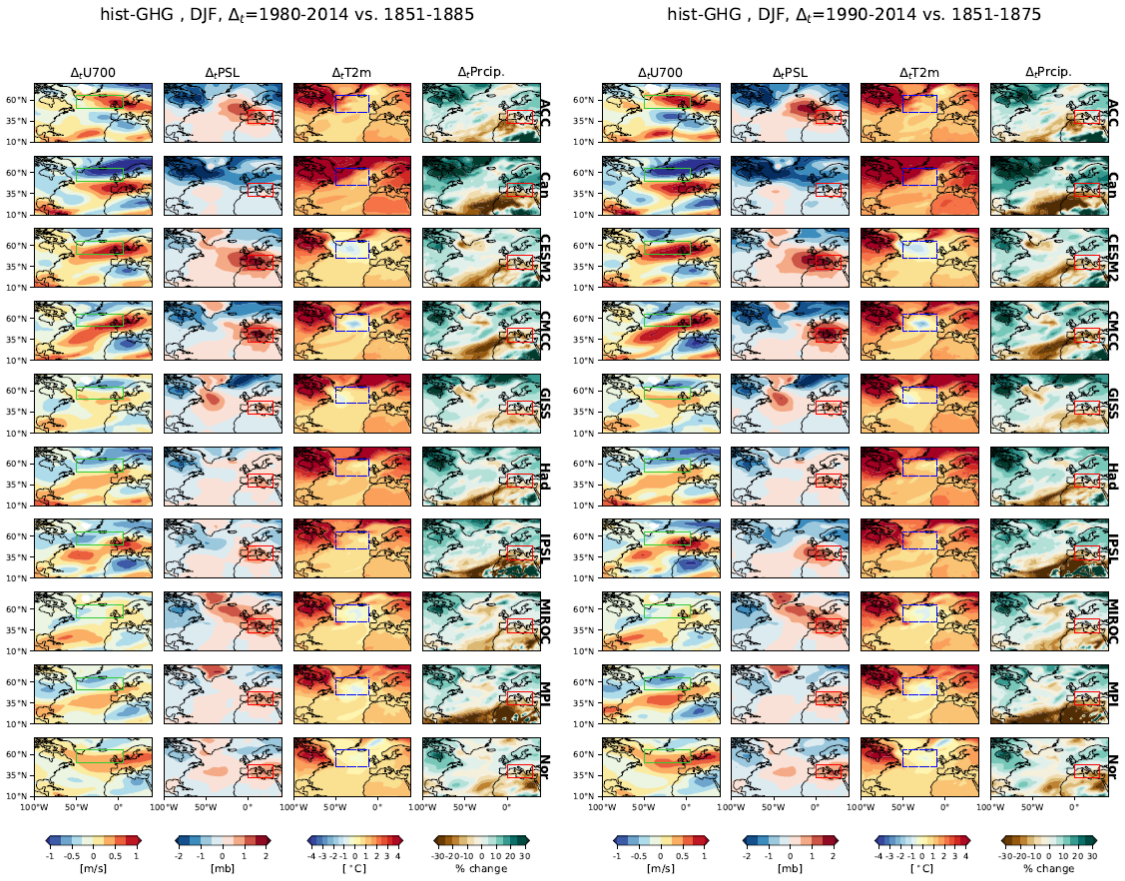
(a) 1990-2014 vs 1851-1875

(b) 1980-2014 vs 1851-1885

Figure S1. Sea-level pressure multi-model means single-forcings and historical responses for different epochs.

Similarly, we tested the results sensitivity to changing the epochs lengths and ranges for the various responses calculated for the individual models in the hist-GHG and hist-aer

experiments. These figures are also included now in the supplemental and show no significant sensitivity. See for example:



(a) 1 (b) 3

Figure S5. Ensemble means of the DJF responses for the U700, PSL, T2m, and Prcip. (column-wise) fields, obtained for each model (row-wise) in the hist-GHG experiment, for different averaging ranges than in the main text.

5. Why are the CESM2 outputs treated so differently here? I believe the CESM2 model level outputs can be easily interpolated onto standard pressure levels. This is also something I pointed out earlier, but hasn't been addressed.

As we note in the manuscript, the pressure levels provided by CESM2 are nearly identical to the standard pressure level in the LESFMIP simulations, and we prefer to use the data as provided instead of using any interpolation methods. This is motivated by the fact that we are taking temperature at 992hPa as a proxy for surface temperature, and there is no practical way to interpolate from the fields available to get the true surface temperature. Since we are considering anomalies in any event, it seems farfetched to imagine that this choice would in any way affect the results.

6. When PSL response is related to global mean temperature response, would it mean that dynamical sensitivity is related to climate sensitivity?

We aren't sure that we understand the reviewers' question well, but we will try to explain the issue more fully. We are adopting the terminology of Grise and Polvani 2016 (already cited) who examined the relationship between a model's equilibrium climate sensitivity (ECS) and its atmospheric circulation response to increased CO2. Their key finding is that the magnitude of the midlatitude atmospheric circulation response in the NH winter is not well correlated with the global-mean surface temperature response (ECS). They focus on three key measures of the circulation response, and we are focusing on a fourth one: the PSL response over the Mediterranean. We are using different simulations that they did as well. Nonetheless, our conclusions agree with theirs: there is no relationship between dynamical sensitivity in the Mediterranean and climate sensitivity.

Reviewer 2

This manuscript investigates the sources of inter- and intramodel spread in historical Mediterranean precipitation trends using LESFMIP simulations. The key analysis involves taking correlations of wintertime precipitation trends with those of various atmospheric dynamics metrics under greenhouse gas and aerosol forcings across all LESFMIP ensemble runs.

Understanding inter- and intramodel spread is important and I think this manuscript presents valuable results on what dynamics matter for understanding the sources of Mediterranean precipitation spread. However, I share the same concerns raised by reviewer 1, especially regarding how the current manuscript doesn't fully leverage the usefulness of LESFMIP simulations. It seems to me that separating the role of various forcings on the precipitation trend is not the main focus of this paper. There are brief discussions in passing about how the correlations between precipitation trend and various dynamics metrics differ among GHG and aerosol forcings but many of the differences remain unexplained. Thus the key results found here (sources of inter- and intramodel spread) could largely be done using regular SMILE ensembles (see e.g., Deser et al. 2020, Maher et al. 2023), which offers the advantage of having even more ensemble members than in LESFMIP.

We now show figures for the multi-model mean of each individual forcing experiment as well as that for the historical (all-forcings) experiment, and so utilize the LESFMIP simulations as they were originally intended. We also discuss cancellation of the aerosol induced response and GHG induced response throughout, which cannot be done using regular SMILE runs. In the context of single-model vs. multi-model ensembles, we also discussed the fact that for some of the climatic metrics we consider in the work, the individual models do not show the responses' relations that are revealed by the multi-model. Meaning, single-model ensembles would actually obscure these relations. Also, we want to stress, as reported in the manuscript, that GHG and aerosols do share similar correlations for the dynamical metrics responses vs. the PSL response. Indeed, in some cases the magnitude of the correlation differs in the two experiments, but that is outside the scope of the current work. The reviewer shares the same concerns raised by reviewer 1 and we kindly refer the reviewer to our detailed response to reviewer 1 on the various issues.

As reviewer 1 points out, I agree the usefulness of the LESFMIP ensemble could be better leveraged by putting the model trends in the context of observations. Vicente-Serrano et al. (2025) argue observed Mediterranean precipitation trends are dominated by internal variability and are not a response to forcing. LESFMIP could provide more clarity behind the sensitivity of the trend to the choice of time period (e.g., opposing responses to GHG and aerosol forcing may lead to weaker total forced response during some time periods).

As we present in detail in our response to reviewer 1, we now directly compare the observational trends to the historical all-forcing run. Specifically, we do that for the historical all-forcings multi-model mean, as well as vs. the individual ensemble members distribution, and show that the model response is not discrepant with the observed response. We also cite and discuss the paper of Serrano et al. In particular, we analyze the LESFMIP precipitation trends in similar periods analyzed by Serrano and show that the two sets agree in a recent-decades trend. Full details are in our response to reviewer 1 and we kindly refer the reviewer to them.

Regarding taking into account internal variability, this was shown in the original submission in our scatter plots, but not quantified. We now more fully quantify it in the updated plots as we specify in detail in the response to reviewer 1; see, e.g., Figure 7 presented in the detailed response to reviewer 1.

Furthermore, in the updated manuscript we now refer to the choice of the response time periods and to the response sensitivity in changing it. Specifically, as we present above in detail in our response to reviewer 1, we have added figures to the supplemental material regarding the choice of time period, and in general find nearly identical results. We show this both for the multi-model means and for each of the models' ensemble mean. We refer the reviewer to our detailed response to reviewer 1. Nonetheless, the reviewer is correct that part of the GHG induced response is cancelled by aerosols. This has been noted explicitly in the text, as we cite above in the response to reviewer 1.

As a first step I suggest the authors either 1) align the choice of dataset with the current focus on understanding inter- and intramodel spread or 2) reframe their research question to better align with the usefulness of the LESFMIP dataset.

Based on the comments made by both reviewers we now include a more thorough comparison of the LESFMIP experiments with observations. Both based on the historical multi-model mean and the single ensemble members distribution. We also utilize the LESFMIP experiments to demonstrate that aerosols have had an impact on historical climate in this region, a finding that is not possible using conventional SMILES. Last, we better quantify now intermodel spread in the responses correlations we analyze.

Deser, C., Lehner, F., Rodgers, K.B. et al. Insights from Earth system model initial-condition large ensembles and future prospects. Nat. Clim. Chang. 10, 277–286 (2020). https://doi.org/10.1038/s41558-020-0731-2

Maher, N., Wills, R. C. J., DiNezio, P., Klavans, J., Milinski, S., Sanchez, S. C., Stevenson, S., Stuecker, M. F., and Wu, X.: The future of the El Niño–Southern Oscillation: using large ensembles to illuminate time-varying responses and inter-model differences, Earth Syst. Dynam., 14, 413–431, https://doi.org/10.5194/esd-14-413-2023, 2023.

Vicente-Serrano, Sergio M., et al. "High temporal variability not trend dominates Mediterranean precipitation." Nature 639.8055 (2025): 658-666.