When I first finished reading this manuscript, my initial inclination was to recommend rejection. However, after further consideration, I decided to give the authors an opportunity to revise the manuscript. That said, **extensive revision and substantial improvement will be necessary** before the manuscript can be considered for acceptance. See my attached document "Reviewer_comments.docx".

We want to thank the reviewer for the critical but thorough and constructive comments which helped improve the quality of our manuscript. Following the received comments we plan to address these in the revised manuscript in the following fashion:

Major comments:

1. Previous studies addressing future changes in stationary waves, clarifying the mechanisms responsible for stationary-wave variability, and examining the drivers of European heatwaves often use the eddy streamfunction (Psi*), calculated by removing the zonal-mean component from the total streamfunction, to depict the spatial structure of subtropical/midlatitude stationary waves. Could you please do the similar analyses in the main Fig. 1 and Fig. S2, but with 200-hPa Psi*?
Thank you for the suggestion. We have computed the trend patterns of the Psi* (please see figure below).Results show consistent findings and similar patterns among Z200_az and Psi*. For this reason we will include such Psi* analyses in the Supplementary Material to complement the findings in Fig. 1 and Fig. S2.

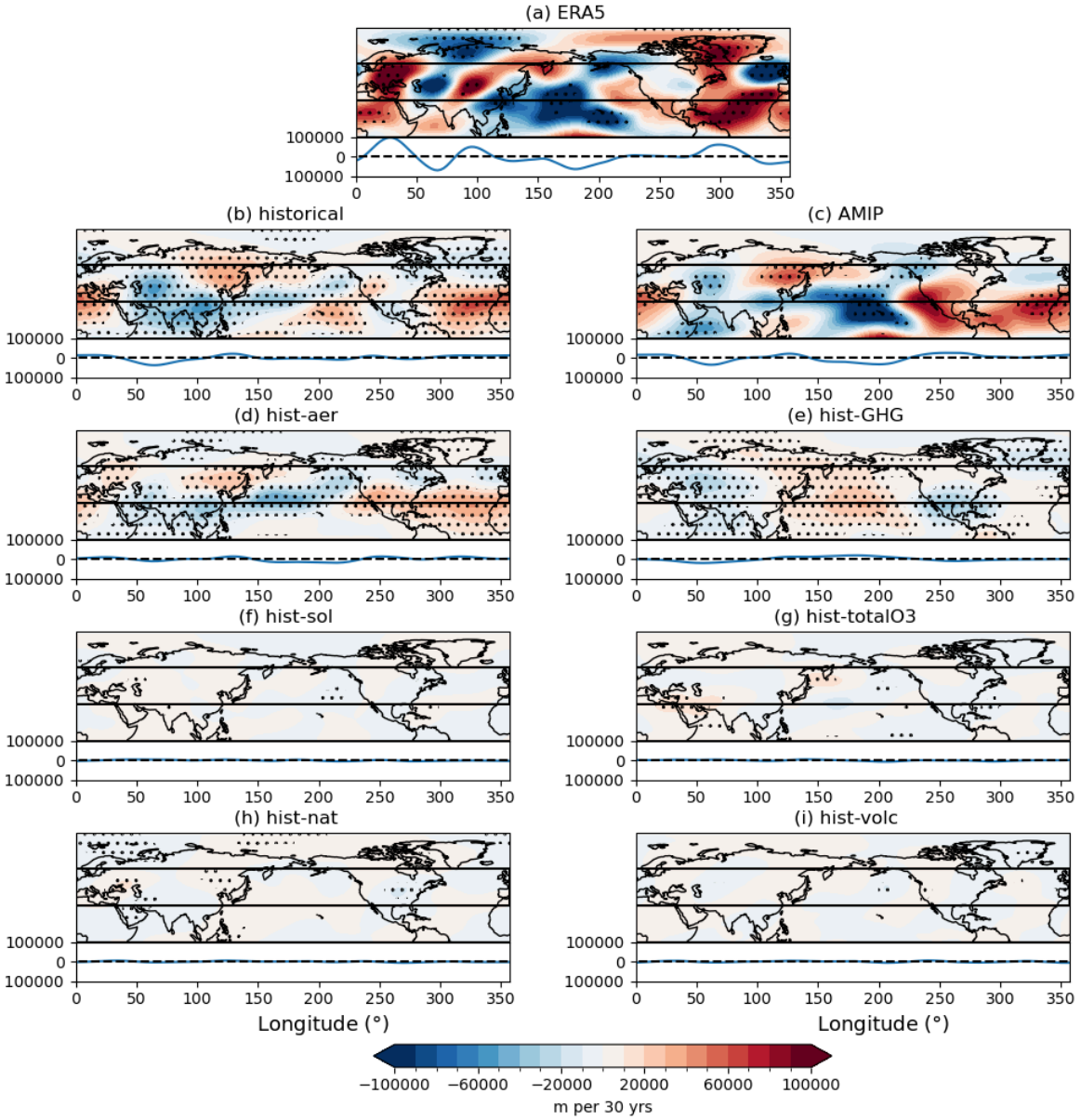

Figure R1. Same as Figure 1 but for the eddy streamfunction (Psi*).

2. Is it possible for the authors to switch from pattern correlation to the skill score (S; Taylor 2001)? See the equation for S in Wang and Wu (2018). I ask for this change because the hist-volc experiment shows significant pattern correlation with the observed trend during 1979–2014, and hist-sol also exhibits a somewhat high (R ~ 0.5) pattern correlation with the ERA trend during 1943–1978. However, in the spatial maps (Fig. 1i and Fig. S2e), the magnitudes of the trends in these two experiments are extremely weak compared to those in the historical and hist-aer experiments. Could you at least show the skill score for the main Fig. 3 and Fig. S3 in the supporting information? It would be even better if the authors could replace all pattern-correlation analyses with the skill score, as S additionally considers spatial standard deviation.

Thank you for the suggestion. We agree that skill score together with pattern correlation can provide added context to the results. For this reason, we will show the skill-score equivalent of figures containing pattern correlation metrics (Fig. 3-7 and Fig. S3) and add them as supplementary figures. However, given the overall small magnitude of changes in model

ensemble means, S score is mostly low. Following the suggestions by Reviewer #3 we are extending the analysis to check whether Signal-to-Noise errors exist, and correct them if possible.

3. I am disappointed that the authors did not make any effort to understand why the hist-aer simulations show similarity to the observed wavenumber-5 pattern during 1979–2014. If the exploration of mechanisms is left entirely to future work, then this manuscript should also be accepted only in the future, when the authors clarify the underlying physical processes. I fully understand that a complete investigation and understanding of why the observed trends show a wavenumber-5 pattern over the Northern Hemisphere during the past 40–50 years may be too ambitious for a single study. Teng et al. (2022) did a much better job than this study in attempting to clarify the underlying mechanisms responsible for this observed pattern. Nevertheless, even they could not fully address the underlying dynamical processes.

I hope that this study can at least try to understand why the hist-aer experiment simulates a zonal dipole pattern of 200-hPa eddy geopotential height over Northeast Asia and the North Pacific.

This wavenumber-1 eddy geopotential height trend is consistent with that seen in the ensemble mean of the historical simulations and, to some extent, is also analogous to the pattern revealed by ERA5, albeit with weaker magnitude. What are the long-term trends in SST and precipitation in the hist-aer experiment? Can you identify significant changes in the Rossby wave source (Sardeshmukh and Hoskins 1988) and the stationary wavenumber (Ks;Hoskins and Ambrizzi 1993; Karoly 1983)? I would greatly appreciate it if the authors could attempt to address this using simplified models, such as a stationary wave model (SWM; Ting and Yu 1998) or a linear baroclinic model (LBM; Watanabe and Kimoto 2000). Teng et al. (2022) also used a linear planetary wave model to better understand the mechanisms responsible for the observed wavenumber-5 pattern.
We want to thank the reviewer for the interest and thorough suggestion. We agree that an in-depth analysis of the mechanisms involved in the observed changes would be of great interest. However, we also believe that the current manuscript together with the proposed changes by Reviewer #3 adding a formal signal-to-noise analysis would provide a meaningful contribution to the community, providing a comprehensive analysis of the effect of each forcing in driving the upper tropospheric circulation in Northern Hemisphere summer together with a quantification/ possible correction of S/N issues. Therefore, we prefer not to add the suggested analysis to the revised manuscript as it would carry the study in a different direction than the signal-to-noise analysis suggested by Reviewer #3, and beyond the scope of one journal publication. We believe that our conclusions of aerosol emissions (and in the revised version also volcanoes) being important drivers of the observed wave-like circulation changes, and that the LESFMIP simulations show significant signal-to-noise errors in the northern hemisphere summer circulation responses to these forcings, represent substantial new knowledge of relevance to the scientific community, and worthy a publication in Weather and Climate Dynamics.

Comments between minor and major:

1. “The 1943–1978 period was chosen to complement the analysis of recent changes with a different period, and understand the evolution of the trend in time and identify possible differences in the mechanisms involved.” → The spatiotemporal evolution of aerosol forcing is different between these two periods.
Thank you for bringing up this point. We agree that this choice can appear arbitrary in the manuscript. We will add a more thorough explanation behind the decision and the differences between the two periods. “*We chose specifically the 1943-1978 period as it is the period immediately before 1979-2014 of equal length, so the timescales of the analysis are consistent. It is relevant to note that there have been important differences regarding the forcings between the two periods. Most notably, sulfate aerosol emissions have increased broadly in the Northern Hemisphere during the 1943-1978 period, especially over Europe and North America. While, in the 1979-2014 period, sulfate aerosol emissions decreased over Europe and North America but strongly increased over China and India (R. M. Hoesly et al (2018)).*”

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., Seibert, J. J., Vu, L., Andres, R. J., Bolt, R. M., Bond, T. C., Dawidowski, L., Kholod, N., Kurokawa, J., Li, M., Liu, L., Lu, Z., Moura, M. C. P., O’Rourke, P. R., & Zhang, Q. (2018). Historical (1750–2014) anthropogenic emissions of reactive gases and aerosols from the Community Emissions Data System (CEDS). Geoscientific Model Development, 11(1), 369–408. https://doi.org/10.5194/gmd-11-369-2018

2. “Accordingly, we find no evidence for oceanic variability contributing to the wave-5-like trend pattern in Z200_az.” → How did you reach this conclusion? There are similarities between Fig. 1a and Fig. 1c. Could you explain this statement in more detail and show the analyses that support it?
Our reasoning behind this sentence comes from the results from Fig. 1 and Fig. 3. Despite there being similarities between fig. 1a and fig. 1c, these are not qualitatively different from fig. 1b except in the Pacific region and stronger magnitudes. Given that AMIP experiments also contain the forcings from historical simulations, we would expect a degree of similarity between AMIP and historical simulations (pertaining to the forced contribution) while the differences among the two (given sufficiently large ensembles) would be the result of ocean variability. Fig.3 shows that the pattern correlation in midlatitudes of historical simulations and AMIP have similar values and spreads. From this we conclude that in model simulations, correct (observed) representation of the modes of variability in the ocean does not yield changes in atmospheric circulation that are structurally closer to reanalysis than simulations containing free-running oceans. Therefore, results suggest that to obtain the spatial structure of changes observed in AMIP simulations, SST forcings are not required. Given that historical simulations with freerunning oceans are able to match it. Thus, in CMIP6 models, we find no evidence of oceanic variability being the cause of these changes.
To further clarify our statement, we revised the wording to make clear that this is a finding based on (and specific to) the model simulations analysed in this study: “*Accordingly, based on the climate simulations with CMIP6-class models analysed here, we find no evidence for….*”

Minor comments:

“More specifically, we used simulations isolating anthropogenic aerosol concentration, well-mixed GHG concentrations, volcanic emissions, solar activity changes, total ozone concentrations, as well as historical forcing simulations combining all of the above.” → It could be helpful if you could indicate the number of ensemble members used for each single-forcing experiment here.
Thank you for the remark. We will add the specific number of members used in each forcing experiment.

“These are idealized atmosphere-only experiments that use SSTs and sea ice concentration from reanalysis data as boundary conditions as well as prescribed historical forcings.” → Please indicate which SST/sea ice dataset is used to drive the CMIP6 AMIP simulations (e.g., HadISST).
We will add explicitly the AMIP SST dataset used. Lines 77-78: These are idealized atmosphere-only experiments that use SSTs and sea ice concentration from reanalysis *following the PCMDI-AMIP procedure (Hurrell et al., 2008)* as boundary conditions as well as prescribed historical forcings.

Hurrell, J. W., Hack, J. J., Shea, D., Caron, J. M., & Rosinski, J. (2008). A New Sea Surface Temperature and Sea Ice Boundary Dataset for the Community Atmosphere Model. Journal of Climate, 21(19), 5145–5153. https://doi.org/10.1175/2008JCLI2292.1

“Similarity of the spatial patterns of changes between models and reanalysis was assessed using the area-weighted Pearson pattern correlation at midlatitudes (30°N– 60°N) for the whole circumference and the two longitude ranges defined in Happ. et al. (2025): Eurasia (15°E–110°E) and North America–Atlantic (NA-Atl) (100°W–0°E).” → I am just curious: How did you calculate the area-weighted Pearson pattern correlation? Could you provide a bit more detail here?
We will add the explicit method we used to weight the areas “*The area-weighting was done by applying weights proportional to the cosine of each latitude at each grid-point.*”

“Our results indicate that only simulations forced with anthropogenic aerosol or wellmixed greenhouse gas emissions show significant trends in NH midlatitudes Z200_az during summer (Fig. 1b–i).” → Should this be Figs. 1b–1e?
Thank you for pointing out this typo. We will change it accordingly.

“Regarding the spatial structure of the trends, the aerosol, AMIP and historical ensembles present very similar patterns (Fig. 1b–d).” → It is safer to state that simulated trends in the historical simulations are similar to those in the hist-aer simulations; therefore, the forced response in the historical simulations is largely driven by evolving anthropogenic aerosols. The overall trend in AMIP is more pronounced and exhibits differences from those in the historical and hist-aer simulations. For example, AMIP simulates negative trends over the subtropical central North Pacific (Fig. 1c), which could be related to the observed La Niña- (or negative PDO-) like SST trend. The negative trends over the subtropical central North Pacific seen in AMIP are very weak in both the historical and hist-aer simulations.

Thank you for the valuable insight. We will restructure the paragraph separating the historical and aerosol from AMIP analysis for added clarity, which now reads as follows:

"*Regarding the spatial structure of the trends, the aerosol and historical ensembles present very similar patterns (Fig. 1b,d). This suggests that the forced response in the historical simulations is largely driven by evolving anthropogenic aerosols. Notable features are the decrease in Z200_az in the North Atlantic, as well as the wave train over Eurasia with two centres of increased Z200_az over central Europe and Asia combined with a decrease in Z200_az between them over western Asia. This structure is present and significant in reanalysis (Fig. 1a). The decrease in Z200_az over west Asia is displaced southward in the aerosol only experiments (Fig. 1d). The overall trend in AMIP is more pronounced and exhibits differences from those in the historical and hist-aer simulations. For example, AMIP simulates negative trends over the subtropical central North Pacific (Fig. \ref{fig:trends}c), which could be related to the observed La Niña- (or negative PDO-) like SST trend. The negative trends over the subtropical central North Pacific seen in AMIP are very weak in both the historical and hist-aer simulations. However, AMIP fails to capture the GPH increase over western Europe. Simulations forced with only well-mixed greenhouse gases (Fig. 1e), while showing some regionally significant trends, do not show a spatial structure of changes similar to reanalysis.*"

"As we can see in Fig. 3 for 1979–2014, AMIP simulations and experiments using aerosol and historical forcings show similar values of positive pattern correlation across all models." → I suggest changing "historical" to "historical all-forcings" for clarity.
Thank you for this suggestion, we will change "*historical*" to "*historical all-forcings*"

"Therefore, these results suggest that GHG emissions alone, and the associated global warming, do not play a role in driving the observed trends in Z200_az." → Climate models inherently suffer from biases, and it is still possible that CMIP6 models do not perfectly capture the forced response to increasing GHG emissions. I therefore suggest changing this to "likely do not play a dominant role in driving the observed trends in Z200_az."
Thank you for the remark. We will rephrase the sentence accordingly.