

We sincerely thank the reviewer for the time he or she spent editing the manuscript. We would also like to thank them for the numerous comments, some of which highlighted passages that were certainly still unclear or incomplete. We believe this has significantly improved the paper.

Our answers are integrated into the review below.

---

#### Review #1

“This manuscript presents ICON XPP, the first version of a new climate model specifically tailored for climate prediction applications at two alternative model resolutions.

Both the model documentation and the evaluation of the two resolution configurations are comprehensive, spanning seasonal to decadal scales and covering important drivers of predictability such as the ocean and atmospheric circulations, the stratosphere and key modes of internal climate variability. The authors employ a wide range of diagnostics and benchmark against well-established datasets, demonstrating the model’s strengths and identifying areas for improvement.

I find the paper to be of interest and suitable for publication in GMD, pending the authors’ response to a few clarifications and minor comments listed below. “

- **The manuscript is generally clear but presents frequent typos and grammatical errors. I recommend a thorough proofreading to improve its readability.**

Thanks a lot for pointing at this. We will have a thorough check, for example proof reading by a few co-authors. In relation to reviewer 2, we also will sharpen the abstract and restructure the introduction.

- **Sentence in lines 105-109: The final part of the sentence seems to be incomplete. Did you mean to say that MPI-ESM has been used for developing machine learning methodologies? If that’s the case, in which way?**

We will skip this part of the sentence, since machine learning is just applied for examination of heat extremes over Europe in historical experiments, but not yet for the assessment of predictions.

- **Sentence in lines 109-110: The phrasing is odd. I would simply say that MPI-ESM has been used to conduct operational decadal climate predictions.**

Thanks, we will change the sentence accordingly: “MPI-ESM has been also used for the assessment of decadal climate predictions and is used to conduct operational forecasts”

- **Sentence in lines 110-113: This sentence would benefit from some rephrasing too. I suggest simply saying that decadal prediction skill in the model has been shown to arise from near-term memory in the North Atlantic Ocean heat content and from the externally-forced long-term trends. Also, note that “prediction skill” should be written in singular.**

Thanks, will change the sentence accordingly.

- **Sentence in lines 113-116: Instead of indicating the processes for which the prediction skill has been assessed, it would be more interesting to state those for which the model shows actual skill, and those for which it doesn't.**

Agreed, we meant the actual skill. The sentence will be changed to: "In addition, actual predictions skill is found...". In addition, we will add a recent publication which demonstrates that summer heat extremes can be predicted by the model (Wallberg et al., 2025).

- **Lines 118-119: The carbon uptake by the ocean as not an Earth System component, it's an Earth system process.**

True! Will be changed accordingly.

- **Line 172: What is TERRA? You have not properly introduced it.**

Thanks! TERRA was adopted from the Latin for "earth". This note will be added to the sentence where TERRA first appears.

For clarity we will also explain the role of TERRA in ICON Land, which has not been stated yet:" ICON Land includes the JSBACH land-surface model developed for predecessors of ICON XPP such as MPI-ESM (Reick et al., 2013, 2021), and other land-surface model such as TERRA (from the Latin for "earth") which is implemented into the operational configuration of ICON NWP"

- **Lines 244 to 246: Node characteristics can largely vary across machines. Could you also indicate the throughput in terms of the number of processors per day (to allow a more direct comparison with other models)?**

For each node there are two CPUs (64 cores each). We will add a note in the text: "The experiments are run on the CPU-partition of the Levante High-Performance Computing system at the Deutsche Klimarechenzentrum (DKRZ), with each node consisting of 2 CPUs and 128 cores in total."

- **Section 2.3: The manuscript would benefit from additional detail on the tuning procedure. Specifically, it would be helpful to explain how the parameter choices summarized in Table 1 were determined, whether they were based on tailored experiments, and if so, what kind of experiments were conducted.**

Yes, tailored pre-industrial control experiments with a wider range of parameters were performed to get those optimal parameters and their values in Table 1. Those experiments include a series of coupled model runs to achieve the ideal convection, cloud cover and microphysics parameters for TOA radiation balance and GMT assessment, and then additionally sea-ice and ocean mixing parameters to get a balanced ocean circulation assessment (e.g. AMOC, Labrador Sea freezing and mixed-layer depth). We will include a statement accordingly in the main text:

" A series of tailored pre-industrial control experiments are employed to find the optimal parameterization values. First, a wider range of convection, microphysics and cloud cover parameters are examined to estimate their impacts on the TOA radiation balance and GMT. Then, with the resulting subset of atmospheric and oceanic parameters the ocean-circulation and sea-ice distributions are adjusted. With the optimized parameters a new spin-up is started. The values of the optimized parameter values are shown in Table 1."

- **Sentence in lines 270-271: I agree that reported TOA values are within the acceptable range and compare well with residual imbalances documented in other CMIP6 models. However, in the last 500 years of the spin-up simulations both the 160/40 and 80/20 configurations exhibit consistently positive and negative TOA imbalances, respectively, without oscillating around zero. This suggests a persistent net energy gain in one case and loss in the other, which may have implications for long-term climate stability and should be acknowledged and discussed in the manuscript.**

We don't think that a further discussion is needed since such inconsistencies are well-known in coupled models, and – as the reviewer pointed out - the TOA imbalances in the presented configurations are relatively small. Further it is unclear what their net effects on the long-term climate would be, since there are yet unresolved issues such as energy leakages due to missing parametrization, or incomplete atmosphere-ocean coupling, and small but long-term trends in the ocean interior due to the fact that the coupled system would require much longer runtime to reach equilibrium. We think a discussion at this point, would be rather speculative than scientific founded, and would require a proper analysis.

- **Line 304: Could you indicate here and in Table 2 how many members you have run per ensemble?**

The number of ensemble members are already given in Table 2 in the description of HIST. This information appears perhaps a bit hidden. We will now explain the number of members in the table caption: "For HIST, three ensemble members are performed for the period 1850-2014."

- **Figure 8: In this figure the red color represents the 160/40 configuration and the blue color the 80/20 one, but in Figure 1 is the other way around. I suggest using the same color convention to ease the comparability of the figures.**

Thanks, we will change the colors accordingly.

- **Line 335: It would be fair to comment that the reference period for PIOMAS corresponds to a much warmer climate than for the preindustrial simulations, which implies that the preindustrial sea ice thickness is expected to be larger.**

Yes, agreed. We will add a note by the end of this paragraph: "In fact, since the PIOMAS reanalysis depicts the current state of the climate, the preindustrial sea-ice thickness is expected to be larger."

- **Line 344: A peak magnitude should be a flat number, not a range.**

True. This sentence will be changed to: "For the last 500 years of simulation, the overturning circulations in the Atlantic show values between 14-17 Sv for 80/20 and 16-19 Sv for 160/40 at 26° N at 1000 m depth, which is comparable to the RAPID array (~17 +/-4 Sv, Frajka-Williams et al., 2019)."

- **Line 350: I would explicitly say that you refer to the transport through ocean passages that are important for the climate system.**

We will add a note where the transports are first mentioned: "... and transport through various ocean passages that are important for various climate sub-systems (table 3)."

- **Figure 5: A key process for the AMOC and decadal variability (and predictability) in the North Atlantic is deep water mixing in the Labrador and Irminger Seas, which is controlled**

**by density stratification. It would be extremely useful to show how they are represented in the two model configurations, given the goal of using them to perform decadal climate predictions.**

We will include a new figure showing the mixed-layer depth for the two configurations and will further add the following text:

” The state of the ocean circulation in the North Atlantic is closely related to the deep-water mixing in the Labrador Sea and Irminger Sea, and at higher latitudes in the Norwegian and Greenland Seas. The deep convection of the Labrador Sea and Irminger Sea can drive the deep-water formation, and are suggested to impact on the AMOC. The mixing in the Norwegian and Greenland Seas contribute to the Arctic overflows and Atlantic bottom water. The mixed-layer depth in March is used here as a proxy for deep-water mixing (Fig. 6). It shows that the 80/20 configuration provides deep mixed layers in the Labrador Sea with maximum values of up to 3000 m. In the Irminger Sea, the mixed-layer depth reaches values of up to 2500 m. The maximum of the deep mixed layers in the 160/40 configuration is shifted to the Irminger Sea and reaches values of about 2500-3000 m. The shift of the maximum values of the mixed-layer depth is closely related to the production of sea ice, which is larger in this configuration compared to the 80/20 configuration (see Fig 1d). The values of mixed-layer depths are generally higher compared to recent climate estimates for which maximum values of ~1000 m in the Labrador Sea and Irminger Sea are suggested (e.g. Königk et al., 2021). Finally, the mixed-layer depths in the Norwegian Sea are similar in both configurations and reach values of up to 3000 m.”

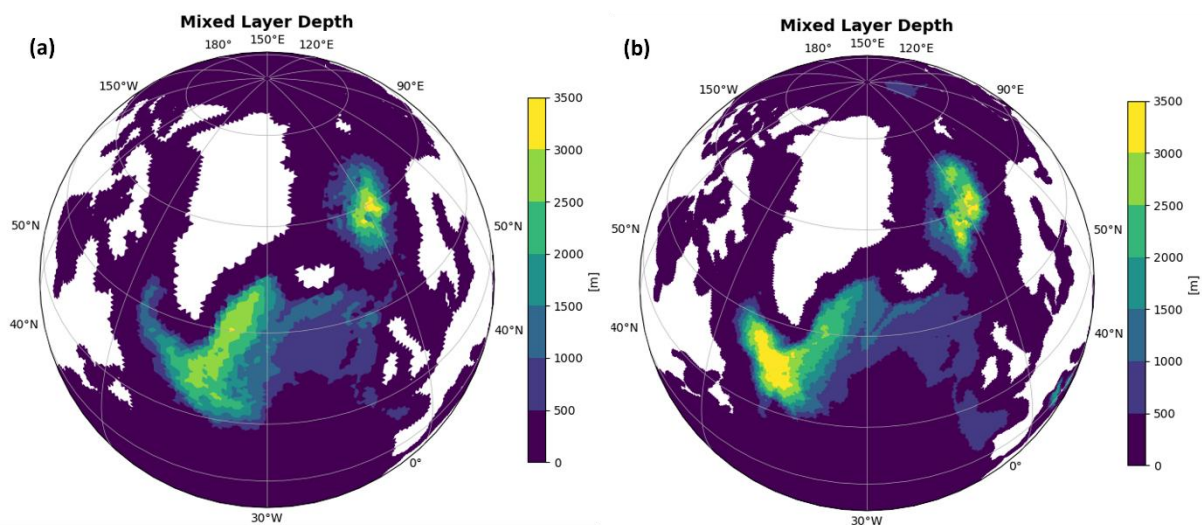


Figure 1: The mixed-layer depth in March of CTRL for (a) 160/40 and (b) 80/20. Units are in meters [m].

- **Table 3: The title of the first column is incorrect. It's not an experiment, but an ocean passage that you are listing in the column.**

Thanks. Typo! Will be changed accordingly.

- **Line 489: I would change “key indicator for” with “critical parameter that determines the”.**

Will be changed accordingly.

- **Sentences in Lines 502-505: The assumption of linearity doesn't seem to hold in the last 130 years of the 80/20 configuration, which has an R square of 0.1 that is most likely not statistically significant. Can you discuss which implications this has for your estimate?**

The low R-value of the 80/20 experiment is dominated by a group of points near TOA=0. Upon closer examination, we found that this cluster of points corresponds to the model time steps at which technically necessary restarts of the experiment were performed. The configuration for this experiment was extremely unstable, so on-the-fly adjustments were made (e.g., by changing the time steps in the ocean and atmosphere, or the Rayleigh coefficient). These parameter changes were somewhat too strong. This resulted in the model climate adjustment that reflected those changes rather than the climate sensitivity. We repeated the 4xCO<sub>2</sub> experiment for the 80/20 configuration with only minor parameter changes. The figure below shows the ECS of this new experiment for all years (150) and the last 130 years. R-value is now 0.53 and ECS=2.49. We will change the figure in the main text accordingly.

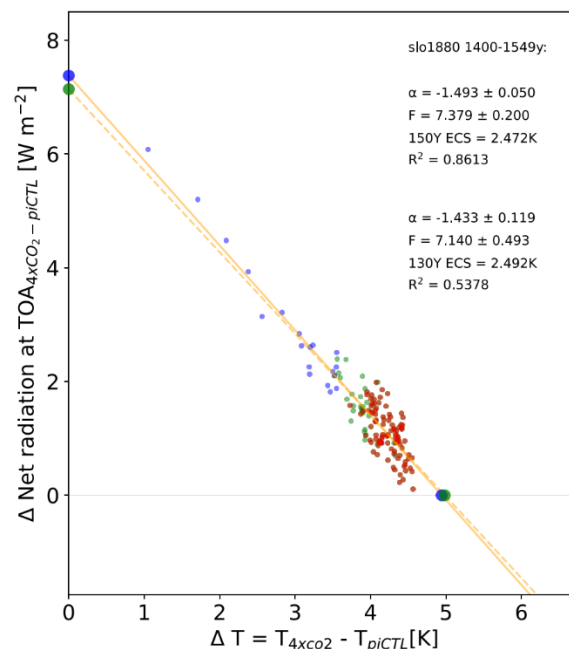


Figure 2: The Equilibrium Climate Sensitivity (ECS) as diagnosed from the scatterplot between TOA net radiance and global mean surface temperature anomaly, including a linear regression for 80/20. ECS is estimated from 150 years of the 4xCO<sub>2</sub> experiments (yellow line) and for the last 130 years (yellow dotted line). R-squared values are included.

- **Line 520: Did you mean to say “principal modes of climate variability”?**

Yes, will be changed here accordingly and throughout the text.

- **Sentence in lines 528-529: You should specify that this statement refers to the atmosphere.**

Yes, will be changed accordingly.

- **Sentence in lines 539-540: The phrasing could be improved. I suggest simply saying that you use OLR as it is generally assumed that is a reasonable proxy for deep tropical convection and precipitation.**

Thanks, will be changed accordingly.

- **Paragraphs in lines 541-560 and Figure 12: You discuss (and cite) first the symmetric component, but you show first the antisymmetric one. I would swap the two rows in the figure to follow the order of the discussion. Also, I suggest acknowledging in the text that your comparisons are just visual and do not consider any statistical significance. Indeed, it is unclear to what extent some of the highlighted improvements for the higher resolution happen by chance.**

Thanks, there has been an error when including the figure. In fact, the two rows should be swapped. In addition, in the text the following sentence appeared twice and will be removed: "One exception is the lack of the strong signal of the  $n=1$  WIG waves (cmp. Fig. 12a and Fig. 12b,c)."

Regarding significance we agree and will remove the subordinate clauses in which "improvements" with respect to the configurations appear.

- **Sentence in lines 555-556: You refer to Figure 12d, but should it be to Figure 12a-c, that are the ones corresponding to the antisymmetric component.**

True. We will swap the rows in figure 12, Then 12d becomes 12d-f.

- **Sentence in lines 589-590: I would be more clear if you specify that Tropflux is your reference for evaluation. I didn't notice until I read the caption of the next figure. Also, could you provide some details on that dataset and provide the corresponding reference? It is not a well-known product.**

We will add some details of TropFlux in the figure 13 caption: "As reference in (c-e) TropFlux is used (Praveen Kumar et al., 2012). TropFlux consists of daily and monthly fluxes, SST and wind stress for the tropical region for  $30^{\circ}$  S to  $30^{\circ}$  N, and combines ERA-Interim and ISCCP corrected using Global Tropical Moored Buoy Array data from 1979 to present."

- **Line 601: Change "for all" to "for all metrics".**

Will be changed accordingly.

- **Figure 14: It misses a legend in one of the spaghetti plots to explain what each line represents.**

True. We will include a legend in figure 14b.

- **Sentence in lines 606-609: The link between SST and wind stress biases in the western and central Pacific is not entirely clear. Aside from the edges ( $150^{\circ}$ E- $160^{\circ}$ E and  $240^{\circ}$ E- $270^{\circ}$ E), the SST slope is quite similar across both simulations and the reference dataset, which is not the case for wind stress.**

Thanks, we will disentangle this link in the text: "The west-east SST gradient is about  $4^{\circ}$  C and the SST slope is close to what is shown in the TropFlux reference. In the western Pacific edge ( $150^{\circ}$ E- $160^{\circ}$ E), the SST gradients are relatively steep in both configurations. In the eastern Pacific edge ( $240^{\circ}$ E- $270^{\circ}$ E), the SST gradient reverses in both configurations. The ENSO-related zonal wind stress substantially improves in the higher-resolved configuration

compared to the 160/40 resolution (Fig 14c). In 80/20 the magnitudes are much closer to the reference, and the minimum is shifted eastward closer to what is observed. “

- **Line 618: How do you define this ENSO amplitude?**

Thanks, the ENSO amplitude is defined as the standard deviation of SST anomalies, we will add this definition.

- **Line 622: Can you explain why it is important to evaluate ENSO skewness?**

Evaluating skewness is crucial because many ENSO-related climate impacts (precipitation extremes, teleconnections, drought risk) depend not only on ENSO amplitude but also on whether warm or cold events dominate. Models that simulate ENSO amplitude realistically may still misrepresent skewness, which limits confidence in projections of future ENSO behavior (An & Jin, 2004; Timmermann et al., 2018; Cai et al., 2021). We would prefer not to explain the skewness in such details in the text, since the ENSO paragraph still is extensive and we think that the skewness should be a standard diagnostic for ENSO.

An, S. I., & Jin, F. F. (2004). Nonlinearity and asymmetry of ENSO. *Journal of Climate*, 17(12), 2399–2412.

Timmermann, A., et al. (2018). El Niño–Southern Oscillation complexity. *Nature*, 559, 535–545.

Cai, W., et al. (2021). Changing El Niño–Southern Oscillation in a warming climate. *Nature Reviews Earth & Environment*, 2(9), 628–644.

- **Line 649: Please avoid using the term “significant” in this context as you have not really assessed statistical significance.**

“significant” will be skipped here.

- **Lines 658-660: This sentence could be rephrased for clarity. I interpret that you mean to say that with the cheaper configuration you can more easily explore the space of hyperparameters in your model to identify potential tuning improvements for ENSO representation.**

Thanks! The sentence will be changed to: “Thus, the much faster and cheaper configuration can be used to more easily explore the space of hyperparameters to identify potential tuning improvements for ENSO representation.”

- **Lines 695-697: I don’t think this statement is correct as it is written. The way I understand it, in the extra-tropics, changes in the zonal and meridional jets are closely linked to changes in major modes of climate variability like the NAO, a link that needs to be well represented for the predictability of these modes and their climate impacts.**

Will change the sentence accordingly.

- **Sentence in lines 731-732: I don’t understand what you mean to say here.**

These two sentence will be rephrased to: “ERA5 reveals divergence of the E-vector downstream of the maximum zonal wind, which indicates that momentum fluxes are able to force the jets towards the north-eastward direction.”

- **Sentence in lines 769-770: For me the most important advantage of showing the wind anomalies is that they better show the downward propagation.**

Thanks for bringing this up. For example, Bushell et al (2020) show that most of the climate models seem to have an eastward time mean wind bias throughout the depth of the QBO, which means that the easterly winds are too weak in the time mean of their multi-model ensemble. This is not the case for these configurations of ICON XPP, especially for the lower to mid equatorial stratosphere. In ICON the easterlies are too strong, hence, ICON XPP is overestimating the easterlies. To account for this bias and focusing stratospheric oscillation itself, the time mean of the zonal mean zonal wind is removed. We will rephrase this paragraph:

“The observed QBO is characterized by descending alternating easterly and westerly jets in the tropical stratosphere and their downward propagation into the troposphere, as shown by the zonally averaged zonal wind (Fig. 17a). In the ICON XPP 160/40 configuration, the descending winds are weakly easterly with a high periodicity of roughly 12 months at 32 km ( $\sim 10$  hPa), compared to roughly 28 months in observations (Fig. 17b). For the higher resolution 80/20 a QBO is present and the period increases to 17 months, although the amplitudes still appear smaller than observations (Fig. 17c). A quasi-permanent easterly wind in the lower-to-middle stratosphere is prominent in both resolutions (Fig. 17 b, c). In order to assess the QBO independently from the climatological state, the long-term mean is removed from the QBO time series (Fig. 17e, f). The zonal wind anomalies emphasize that ICON XPP is capable of developing spontaneous QBO phases and their downward propagation (Fig. 17e,f). However, in 160/40 with 90 vertical levels only, the QBO appears disruptive and the downward propagation is not well established (Fig. 17e). The long-term mean equatorial zonal mean wind in the model configurations further exhibit strong easterly winds at an altitude of about 20 km height...”

Bushell, A. C., Anstey, J. A., Butchart, N., Kawatani, Y., Osprey, S., Richter, J. H., Serva, F., Braesicke, P., Cagnazzo, C., Chen, C.-C., Chun, H.-Y., Garcia, R. R., Gray, L. J., Hamilton, K., Kerzenmacher, T., Kim, Y.-H., Lott, F., Mclandress, C., Naoe, H., Scinocca, J., Stockdale, T. N., Watanabe, S., Yoshida, K., & Yukimoto, S. (2020). Evaluation of the Quasi-Biennial Oscillation in global climate models for the SPARC QBO-initiative. *Quarterly Journal of the Royal Meteorological Society*, 1–31. <https://doi.org/10.1002/qj.3765>

- **Sentence in lines 771-772: For a robust assessment on the simulated QBO periodicity you could perform a spectral analysis of the QBO index for the models and ERA5.**

We thank the reviewer for bringing this up. We agree that showing the equatorial zonal mean wind or the anomaly is perhaps only an exploratory method to assess the quality of the modeled QBO. We further estimate explicitly the QBO amplitude and the period (see figure below). Both quantities are only estimated for the equatorial wind, not the temperature. The amplitude is determined as  $\sqrt{2} \sigma$ , with  $\sigma$  is the standard deviation of the de-seasonalized monthly mean zonal wind at each level (Dunkerton and Delisi, 1985). There are some methods for determining the QBO period: using the fast Fourier transform (FFT) or determining the transition times. The FFT method is used here to determine the QBO period.

The figures show the improvements we can get from the higher-resolving configuration. However, we are a bit uncertain to include this figure in the main text, as it would just confirm what can be extracted from the equatorial mean wind cross-sections, which are a common diagnostic though. Instead, we will include a few markers in the text making a statement about the periodicity of each configuration. See revised paragraph in the previous comment.



Dunkerton, T. J. and Delisi, D. P., "Climatology of the Equatorial Lower Stratosphere." Journal of the Atmospheric Sciences, vol. 42, no. 4, pp. 376–396, 1985. doi:10.1175/1520-0469(1985)042<0376:COTELS>2.0.CO;2.

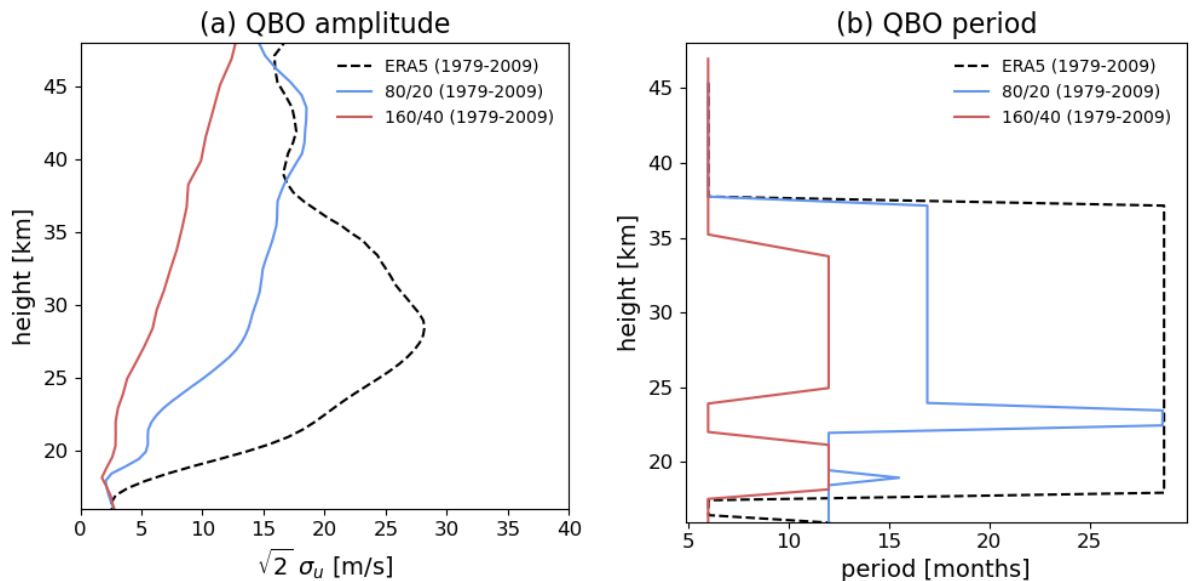


Figure 3: The QBO amplitude (a) and period (b) for ERA5 (black), and the 160/40 configuration (red) and the 80/20 configuration (blue).

- **Sentence in lines 777-779: You cannot really tell if the improvement comes just from the enhanced vertical resolution or from other factors, as there are other things that also change between the two configurations.**

This sentence will be deleted. However, we note the role of the vertical resolution for the QBO in the next paragraph. In AMIP-like experiments we have tested the number of levels and their vertical grid distances. It turned out that with an L130 vertical grid the model is able to generate internal generated QBO.

- **Sentence in lines 785-786: Could you explain what is different in those experiments with respect to your experiments besides being performed only with the atmospheric component?**

There are a few differences between the simulations of Niemeier et al. (2024) and those presented in this paper. Niemeier et al used an atmospheric model with a horizontal resolution of 160 km and 130 vertical levels. Therefore, their simulations either share the horizontal resolution with the 160/40 configuration, but not their number of vertical levels (L90), or the vertical resolution with the 80/20 configuration, but not their horizontal resolution. The lower vertical resolution of 90 levels is found too coarse to generate an internally generated QBO. It does, but not so well as presented in Niemeier et al (2024).

Another important difference is the time step. Niemeier et al used a time step of 360 s for the horizontal grid resolution of 160 km. Simulations for the 160/40 and the 80/20 configuration were performed with a time step of 450 s, which is considerably larger. Coupled tests experiments with the 160/40 configuration revealed a dependency of the representation of the QBO on the timestep. The reason for this is still not fully understood.

The paragraph will be changed accordingly:” In atmosphere-only experiments (160 km, 130 levels), the frequency of the QBO phases has been examined (Niemeier et al., 2024). In these experiments the QBO is well established and benefit from the increase of number of vertical levels. The lower vertical resolution of 90 levels is found too coarse to generate an internally generated QBO. Further, in the atmosphere-only experiments a much smaller time step was used, which seems to further improve the QBO (360 seconds in atmosphere-only experiments compared to 450 seconds in the coupled configurations). However, the reasons for such impact are yet not fully understood. In addition to the QBO, the atmosphere-only experiments reveal a well-represented stratospheric transport. As an example, the transport of the water vapor cloud after the Honga Tonga eruption is found very close to observations (Niemeier et al., 2024).”

- **Sentence in lines 807-811: This comparison is subject to considerable uncertainty, as it is based on counts from a single decade. This is particularly relevant given that the historical simulations do not capture the observed internal variability. To assess whether the model produces a reasonable number of MSSWs, it would be more informative to construct a histogram of event counts across multiple decades, both in the simulations and in ERA5. It would also be useful to examine the seasonal distribution of MSSWs, i.e., the months in which they tend to occur. In your current plot, based on the selected decade, it appears that both models simulate MSSWs earlier in the season compared to ERA5, and the events also seem to be shorter in duration.**

We agree with the reviewer’s concerns. A solid classification of the SSWs requires a larger sample size and a more detailed study of both configurations. Both would currently require considerable effort, especially an in-depth analysis of the SSWs of both configurations would increase the scope of this chapter. However, we accept the reviewer's suggestion and plan a separate publication on the Northern Hemisphere stratosphere in ICON XPP. For the presented paper, we will therefore remove the two sections related to the polar vortex and SSW, including Figure 18.