Dear Editor,

The documents containing our point-by-point responses have already been attached to our replies to each reviewer separately. For convenience, I have copied them below.

Kind regards,

Martin Andrews

Contents

Response to reviewer 1 1
Response to reviewer 2 8
Response to reviewer 3 18
Response to reviewer 4 27

Response to reviewer 1

Response to Reviewer #1 (Prof. Garfinkel)

This study presents first results from the QBOi phase 2 experiments in which the QBO is nudged to observations while the rest of the atmosphere is allowed to freely evolve, with the hopes of understanding how teleconnections improve with a "perfect" QBO. While there is a modest improvement in teleconnections both to the polar night jet and to the subtropical Pacific, the signals are still weaker than those observed. The remaining difference could be either to model deficiencies in the teleconnections processes themselves, or to the fact that part of the observed signal is a fortuitous alignment of internal variability.

This study is clearly on its way to being publishable in WCD, however moderate revisions are needed as detailed below. My main critique is that I'm not fully convinced that the problem is with the teleconnection process itself, while the authors appear to conclude otherwise; see details below. I'm going to sign my review, as the papers I'm referencing in my review probably give my identity away in any event.

Chaim Garfinkel

The authors thank Prof. Garfinkel for kindly reviewing the manuscript. Specific points raised are addressed below.

general comments:

1. The paper doesn't directly show that the nudging leads to a reasonable QBO in the lowermost stratosphere and instead cites Anstey et al in prep (line 225). I realize how this can happen with community efforts; however, I think it is important to demonstrate that the lowermost stratospheric QBO (both amplitude and meridional width) are reasonable in the nudged runs, in

order to confirm that the nudging is acting as intended and that we should expect teleconnections to be better. (For the purposes of this paper on teleconnections, only the lowermost stratosphere is what is relevant.). This concern is motivated by previous work which suggests that on the rare occasion that free-running models simulate a reasonable QBO in the lowermost stratosphere, the PNJ modulation is actually realistic in the extended winter mean (Rao et al 2020) even if the seasonality within winter is still biased.

Response: The protocol paper describing the experiments and how effective the nudging is, is now available in the public domain (https://egusphere.copernicus.org/preprints/2026/egusphere-2026-1165/egusphere-2026-1165.pdf). Results in that paper address the issue raised here. In addition, a further paper analysing the QBO per se in the phase 2 experiments has been submitted to WCD . (https://egusphere.copernicus.org/preprints/2026/egusphere-2026-2856/)

One way to bypass this concern (and also the lack of any figure documenting the QBO itself in this paper) is to focus on regressions of the PNJ against QBO winds, instead of a compositing perspective. A regression approach will "correct" for any deficiency in the QBO and instead focus directly on whether teleconnection processes are too weak. It would be interesting to compare regression coefficients of the QBO with the PNJ in Exp1 vs. in Exp1-obsQBO, and if they differ this would imply that the nudging is somehow interfering with teleconnection processes in models (e.g., the nudging could be messing with the waves, despite the fact it is zonal mean only, by perhaps inducing artificial wave reflection, though I would think the 5-day timescale is weak enough to mostly avoid this).

Response: The protocol paper (https://egusphere.copernicus.org/preprints/2026/egusphere-2026-1165/egusphere-2026-1165.pdf) demonstrates that the nudging has no direct effect in the models on the climatology and interannual standard deviation in the extra-tropics. In response to both your suggestion and that of Reviewer 4, we calculated regressions between the PNJ and QBO winds when sub-sampling the multi-model ensembles with a QBOE-W sample size equivalent to observations. A comparison of Exp1 and Exp1-ObsQBO indicates that the strength of the model teleconnections are largely unchanged, and that the enhanced PNJ response in Exp1-ObsQBO is primarily the result of the increased amplitude of the nudged equatorial QBO. This is now reflected in the main text.

Lines: 47 – 49, 378 – 384, 772 - 774

On a related noted note, some QBOi-phase2 models use full-field nudging and not zonal-mean nudging, and it might be helpful to see if those models have a systematically different QBO->PNJ effect than the ones using zonal-mean nudging only. Some of the downward impacts from the polar vortex to the near-surface in the SNAPSI runs are stronger in the full-field nudging runs (e.g., Australia T2m in Feng et al 2025) for reasons that aren't yet fully clear. To be specific, it would be helpful to add additional panels to figure 2 where you separately show the multimodel mean for full-nudged vs. zonal-mean nudged.

Response: Differences between the full field and zonal mean nudging were considered but obtaining useful conclusions was limited by the small sample size of the full field results .

Likewise, it is difficult to generalise from the specific case studies of SNAPSI to the multi-year simulations used in QBOi phase 2.

2 (but to some degree an extension of 1): I'm a bit confused as to how to interpret figure 2. Doesn't this figure show that the correlation of the QBO winds with U1060 is nearly perfect in the multi-model mean? There are still issues with January specifically, but if I mentally perform a wintertime average the correlations seem very reasonable. See comment #1 about using a regression approach rather than a compositing approach - based on Figure 2 I would guess that a regression approach would likely show that the models are doing a very good job in the winter mean, and that other than a specific issue in January, there is no evidence for a systematic inability to capture QBO->PNJ teleconnection. Rather, the bias is in whether the models are simulating the correct amount of variance, either in the PNJ or in the QBO. (A regression version of Figure 2 in addition to the correlation version currently used would also help clarify this point.) If the authors agree with this interpretation, then the overarching conclusion reached in this paper (i.e., models struggle with the PNJ teleconnection) may need to be toned down.

Response: We are not sure what is meant by the correlation been "nearly perfect" in the multi-model mean? Sure, the correlation is of similar size to that for ERA5 but that is not unexpected given that the multi-model averaging removes most of the internal variability. Nonetheless the signal is much weaker as can be deduce from the results shown in Figure 6. The protocol paper (now available at https://egusphere.copernicus.org/preprints/2026/egusphere-2026-1165/egusphere-2026-1165.pdf) confirms that the models do simulate the correct amount of variance in the PNJ. The nudge ensemble members have near identical QBOs (again see the protocol paper) which of course have the same QBO variance as ERA5.

3. I'm also confused as to how to interpret the results of Table 2 in the context of figure 4 and 5. Figure 4 and 5 indicate that the QBO in Exp1 leads to a significant vortex response (albeit weaker than that of Exp1-ObsQBO). On the other hand, Table 2 indicates this experiment nonetheless has no skill. Is it surprising that skill is now essentially zero even as the HT effect is still there, just weaker? To be specific, I would be interested in seeing an analogous panel on Figure 2 but for the Exp1 runs: is the correlation weaker than for Exp1-ObsQBO? If it isn't weaker, then this would indicate that the primary problem is that the QBO itself is too weak in Exp1.

Response: Figure 4 considers the difference in the composite response of QBOE and QBOW. The correlations in Table 2 assess the ability of the experiments to capture interannual variability of the ensemble-mean timeseries (which includes times when the QBO is neutral). Figure 5 simply assesses whether a subset of QBOE and QBOW randomly chosen from the across the experiment period are capable of matching the observed PNJ response. Hence there is no conflict between the results in Figures 4 and 5, and Table 2. The text has been revised to highlight that the approaches differ and that a significant composite PNJ response does not necessarily lead to significant predictability.

Lines: 386 – 388, 431 – 433

4. Section 5: it would be nice to see a map of slp or near surface U (or similar). It could be the signal doesn't project strongly onto the NAO pattern you predefine, but there still could be a surface signal. Also, I couldn't find the description of the NAO definition used in this paper. Are you allowing for each model to have a slightly different NAO loading pattern? Does the QBO downward impact project onto a slightly modified version of the NAO loading pattern?

Response: The connection of the zonal mean wind towards the middle troposphere is shown in Figure 3 (ERA5) and Figure 4 (QBOi experiments). The NAO definition was contained within the caption of Figure 10 but has now been moved to the start of section 5. The lack of MME NAO signal is most likely not because of the chosen definition of the NAO but more the fact that the models show different seasonality of the QBO-PNJ coupling and its propagation to the surface. We intend to perform a detailed analysis of the NAO response, or lack thereof, in the future. This would stretch to a whole paper and is beyond the scope of the current manuscript.

Lines: 583 – 584

5. The discussion near line 575-580 of the causes of the subtropical jet response seems very speculative to me - please explain in more detail why these tropical precip anomalies should matter. Garfinkel et al 2011 (already cited) instead argue that the subtropical Pacific response is related to downward arching of zonal wind anomalies which then interacts with baroclinic eddies, and actually the Introduction (line 124-130) and Discussion section (line 677-680) of this paper seems to focus on this downward arching mechanism and not the tropical precip mechanism. Note that Garfinkel et al 2011 (the dry model one) explicitly rejects the possibility that the subtropical jet modulation arises from a tropical convection route, but of course in other models or in observations this pathway cannot be ruled out just based on the dry model results. If you want to focus on this tropical convection mechanism, I suggest including it in the Introduction and Discussion.

To be specific, can you make figure 12 (or 13) for a range of pressure levels from 70hPa to 300hPa? If the Garfinkel et al mechanism is relevant, there should be a continuous connection from the QBO region itself to the subtropical Pacific in the upper troposphere.

Response: We thank the reviewer for highlighting the omission of any mention in the Introduction of the tropical convection route. Text has been added to the introduction highlighting this as a possibility and the reason why tropical precipitation anomalies might matter. Like most of the teleconnections in this paper we consider the co-variability with the QBO but, for the sake of brevity, do not deep-dive into the detailed analysis. Follow-up studies that explore the mechanisms in depth are planned.

Lines: 134 – 141, 801 - 802

6. I have a stylistic suggestion, motivated by the fact that the discussion of what may cause the remaining difference in teleconnection strength/position between the nudged runs and observations seems somewhat disjointed. Namely, some parts of the text focus on issues in teleconnection processes (PNJ response) while others focus on the possibility of internal variability or other processes aliasing the observed response (subtropical jet). My guess is that this reflects the intuition of whichever co-author wrote that particular section, though Figure 5

makes the case that internal variability isn't enough for the SPV response and there is a genuine bias in dynamical processes. Either way, it would seem that both possibilities are relevant to all of the teleconnections (though maybe the authors think otherwise! I would welcome being corrected.) If that is indeed the case, the paper could probably be better organized by laying down both possibilities already in the introduction section, then shortening/cutting as much as possible the discussion of these two possibilities from the Results section, and instead revisiting it in detail in the Discussion section in a more systematic way. [I think the paper would be clearer if this suggestion was followed, but it's entirely up to the authors.]

Response: Unfortunately, this paragraph wasn't very clear to the authors. We surmise that the reviewer is concerned with separating the impact of biases in the QBO from other model biases. As a result of this remark, we have revised the introduction to make it clearer that our study focuses on the role of the biases in the QBO. In turn, the discussion section already acknowledged the possible role of other model biases, though a detailed analysis of these is quite beyond the scope of the present study.

Lines: 177, 199, 755 – 756, 765

minor comments

line 49: "robust and insignificant" is an oxymoron. I wouldn't use this expression in the abstract, as there isn't space to explain the subtleties behind why this combination makes sense. (I think you intend for "robust" to refer to across model agreement, and "insignificant" whether anomalies meet 95% significance using standard statistical tests, right? You might want to state this explicitly in the methods section.)

Response: The sentence has been reworded to remove reference to 'robust' as you are right, the two words are challenging to reconcile. Have also done similar in section 4 the Conclusions section.

Lines: 50, 780

line 209: also shown in the ozone-only LESFMIP experiments (Garfinkel et al 2025)

Response: Reference to Garfinkel et al., 2025 added.

Lines: 211, 1055 - 1057

line 302: observe -> observed

Response: Corrected.

Line: 348

line 327: "is an overestimate" could be worded more carefully. Do you intend "might reflect aliasing from unrelated variability"? See general comment 6 above.

Response: We have considered your suggestion as well as a similar comment from reviewer 2 and have changed the text as recommended.

Lines 373 – 374.

line 350: on S2S timescales, there is no evidence for a S2N paradox in the stratospheric polar vortex (Garfinkel et al 2024)

Response: It is true that Garfinkel et al 2024 demonstrates no evidence of a S2N paradox in the stratospheric polar vortex out to around 7 weeks, however, the QBOi phase 2 experiments consist of 40-year model runs implying that the period of influence of the initial conditions is not a significant factor. It is clear from Figure 6 that the PNJ signal for the ensemble mean is much smaller than the model spread for these long experiments. The text has been modified to highlight that these experiments are of much longer timescale than sub-seasonal forecasts, that Eade et al's paper covers seasonal-to-decadal, and a reference to Garfinkel et al. 2024 has been added.

Lines: 403 – 407, 1051 - 1053

line 414: "QBOE vs QBOE" one of these should be QBOW

Response: The paragraph containing this text has been deleted in response to a change in Figure 8.

Line: 504 – 509

Figure 10 seems to have lower resolution than the other figures.

Response: Figure 10 has been improved with a higher resolution version with colours that match those in Figure 11.

Lines 610 – 615.

Figure 13: it would be interesting to include the eastern Indian Ocean in these figures. Overall the pattern matches very well the recent simulations of Schwartz et al 2026, including a tripole pattern in the tropics

Response: It would be interesting, but we prefer to maintain the focus of the discussions on the North Pacific.

Feng, Kexiang, Jian Rao, Chaim I. Garfinkel, Amy H. Butler, Weihua Jie, Tongwen Wu, Peter Hitchcock, Eun-Pa Lim, Andrew J. Dowdy, and William Seviour. "Can stratospheric nudging improve surface predictability? Insights from the 2019 Southern Hemisphere sudden stratospheric warming." npj Climate and Atmospheric Science 8, no. 1 (2025): 353.

Garfinkel, Chaim I., Jeff Knight, Masakazu Taguchi, Chen Schwartz, Judah Cohen, Wen Chen, Amy H. Butler, and Daniela IV Domeisen. "Development of the signal-to-noise paradox in subseasonal forecasting models: When? Where? Why?." Quarterly Journal of the Royal Meteorological Society 150, no. 764 (2024): 4417-4436.

Response: Reference added.

Garfinkel, C. I., Avisar, D., Osprey, S., & Smith, D. (2025). The response of the QBO to external forcings: implications for disruption events. Journal of Geophysical Research: Atmospheres, 130(22), e2025JD044438.

Response: Reference added.

Schwartz, C., Garfinkel, C. I., & Chen, W. (2026). Simulated tropical troposphere response to the QBO: Effect of vertical resolution, gravity waves parameterization, and boundary forcing. Geophysical Research Letters, 53, e2025GL120711. https://doi.org/10.1029/2025GL120711

Rao, J., C. I. Garfinkel, and I. P. White, 2020: Impact of the Quasi-Biennial Oscillation on the Northern Winter Stratospheric Polar Vortex in CMIP5/6 Models. J. Climate, 33, 4787–4813, https://doi.org/10.1175/JCLI-D-19-0663.1.

Response to reviewer 2

Response to Reviewer #2 (anonymous)

This paper evaluates how well twelve climate models from the QBOi phase 2 project reproduce tropospheric teleconnections associated with the QBO. Three experiments are analyzed: one with a freely evolving (internally generated) QBO, one in which the QBO is nudged toward observations, and one in which the QBO is suppressed. The results show that prescribing a realistic QBO improves some teleconnections, particularly the Holton-Tan effect. However, most teleconnections remain too weak compared to observations. The models generally underestimate key relationships, including the QBO's influence on the frequency and timing of SSWs, and they fail to reproduce the observed connection between the QBO and the NAO. The impact of the QBO on the North Pacific subtropical jet is also poorly represented. Overall, the findings suggest that the primary limitation is not simply the simulation of the QBO itself, but the models' ability to accurately represent the dynamical pathways through which QBO signals are transmitted from the stratosphere to the troposphere.

I find the paper to be well written and of clear interest to the readership of WCD. The analysis is relevant and thorough, and most of the results are insightful. I did not identify any major flaws in the study. I therefore recommend that the paper be accepted after the authors address the comments outlined below.

The authors would like to thank Reviewer 2 for kindly taking time to carefully review our manuscript. Responses to specific points are given below.

Main comments

I find the presented analysis somewhat unsatisfying in its current form. The central question—namely, why models struggle to reproduce QBO teleconnections—is not addressed. In particular, the paper does not fully exploit the potential of a multi-model framework by examining differences between individual models and assessing whether these differences are related to how each model represents the QBO itself or to other model characteristics (e.g., resolution).

Instead, the analysis focuses predominantly on multi-model means, which provides no insight into the underlying reasons for the relatively weak teleconnection signals. The manuscript frequently attributes discrepancies to "model biases," but this term is vague. It would be very helpful if the authors could identify which specific biases (e.g., in QBO amplitude, vertical structure, wave forcing, or background circulation) are most relevant, and whether these can be linked to differences in teleconnection strength across models. I recognize that such an analysis is challenging and potentially beyond the scope of the present study, and I appreciate that this limitation is acknowledged as an important direction for future work.

The reviewer is correct we did just focus on the multi-model mean as now noted in the final paragraph of the Introduction. Inter-model differences were considered in our analysis though in general the findings were inconclusive. Therefore, we thought it best not to include that analysis in the current paper. Nonetheless, the paper in our opinion makes an important advance as for the first time we are able to identify the role QBO biases have in the poor representation of QBO teleconnections in models.

Lines: 198 - 200

Specific commentsI find some of the explanations given in the introduction somewhat unclear or unphysical. For example:

81: The entire sentence starting with “This zero-wind line, …”

Response: We have removed this sentence.

Lines: 85 - 87

94: “This results in …”

Response: The sentence has been rewritten for clarity.

Lines: 98 - 101

96: “This modulates the concentration of …”

Response: The sentence has been rewritten for clarity.

Lines: 101 - 103

99: “Regardless of the …”

Response: We have removed this sentence.

Lines: 105 – 106.

105: “If the PNJ …”

Response: We have removed this sentence.

Lines 111 – 112.

118: “For example …”

Response: This sentence has been rewritten for clarity.

Lines 125 – 127.

83: “Under QBOE …”: Please explain at what level the zero-wind line is measured and provide a reference for this statement.

Response: Explanation and reference added.

Lines: 87 – 90, 982 - 984

147: “Of relevance to this study …”. I would make a new paragraph here and keep it together with the next paragraph. Also, check the language/grammar of this paragraph.

Response: Suggested change made.

Line: 162

175: A single-sentence paragraph is insufficient to motivate the study. This section should be expanded to better frame the scientific objectives and relevance of the work. For example, this could be done by better connecting it to the previous paragraphs.

Response: Single sentence paragraph removed, and the relevant content integrated into the next paragraph.

Lines: 192 – 194

187: Explain “AMIP-like” and/or provide a reference.

Response: The experiment description has been completely revised and now includes suitable reference for ‘AMIP’.

Line 207, 205 – 243, 1194 – 1195.

196: The current naming convention (e.g., “Exp1”, “Exp1-ObsQBO”, “Exp1-NoQBO”) is somewhat confusing as the repeated use of “Exp1” does not add meaningful distinction in the context of this paper. While there may be internal or project-specific reasons for this nomenclature, it reduces readability. It would improve clarity if the authors adopted simpler and more intuitive labels throughout the manuscript, such as “FreeQBO”, “ObsQBO”, and “NoQBO”.

Response: The experiment description has been completely revised.

We prefer to maintain the experiment short names for consistency with the companion papers. Exp1 is a generic term used through all QBOi phases to describe the base (control) experiment similar to the way CMIP uses the term “AMIP.” For continuity we would prefer to keep it. Also, it is not a requirement for a model to spontaneously generate a QBO in Exp1 hence replacing it with “FreeQBO” would be a misnomer.

231: Define boreal mid-winter. Also, in the caption of Fig. 2a you say early-mid boreal winter.

Response: “boreal mid winter” changed to “December and January” . Caption also revised as appropriate.

Lines: 264, 283 – 286.

235: What is meant with “climatological zonal-mean zonal wind”. A climatology over all months and years, or over each month individually?

Response: This has been clarified. It is the climatological monthly zonal-mean zonal wind.

Line: 270.

Figure 2: The caption should explain what is shown in the figure. In particular, the contour lines in the right panel. Also, how is early-mid boreal winter defined?

Response: The caption has been rewritten.

Lines: 283 - 287

Figure 3: The individual panels are much too small and hard to see. Also, what is the sample size of QBOE and QBOW?

Response: The sample sizes have been added to each panel in figures 3 and 4 and the figures has been expanded to fill the page width.

Lines: 301 – 307, 323

269: What are the criteria for the composite analysis? Presumably, each model has a somewhat different QBO climatology. So, do you use thresholds specific to each model, or do you calculate the composite QBOE-W response across all models regardless of the intermodel differences?

Response: The text has been revised to clarify the details for the WCD reader. Note in the nudged simulations the inter-model differences in the QBO are very small.

Lines: 311 – 312.

327: “These results might …”: I do not agree with this interpretation. None of the model sub-samples reaches the observed PNJ response. So, why then should the relatively large observed response be due to sampling uncertainty?

Response: The reviewer makes a valid point. In light of this, and also reviewer 1’s suggestion we have modified the text.

Lines: 373 – 374.

329: “Anstey et al. …”: Why do you expect that each experiment has a different variability? Could you explain this better?

Response: It is possible that the nudging directly impacts on the amount of interannual variability of the PNJ and in turn the modulation of that variability by the QBO (i.e. the HTE) is likely to be impacted. However, Anstey et al. shows there is very little change in the

amount of interannual variability. The sentences have been revised in an attempt to improve clarity.

Lines: 376 – 378.

338: "This suggests that …". I do not entirely agree. The models do not faithfully reproduce the HT-effect. Also, some of the correlation could be due to having identical boundary conditions.

Response: All three experiments have the identical boundary conditions and forcings, so the comparison between ObsQBO and NoQBO is reasonable. The reviewer is correct that models do not faithfully reproduce the HT-effect, but the correlation is a measure of co-variability and is not affected by magnitude. We have softened the sentence to suggest that the observed interannual variance "may be accounted for" for rather than "is accounted for" and made a few associated changes.

Lines: 392 – 396.

345: I would also explain the PNJ correlations between the different experiments to shed light on the role of the boundary forcing in driving PNJ variability.

Response: The paragraph has been revised (please see above). An additional comment is included when the table is discussed highlighting that time-varying boundary conditions may play a role in driving PNJ variability.

Lines 428 – 429.

349: What do you mean with "This"? The "fact" from the previous sentence? Also, can you explain in more detail why this could be linked to the signal-to-noise paradox?

Response: The line beginning with 'The fact' has been removed to aid clarity. Added the pointers which are indicative of the signal-to-noise paradox, and changed 'likely linked' to a more accurate 'indicative of'.

Lines: 402 – 407.

359: This is not a good explanation. Why should "other influences" like ENSO only affect the observations and not the simulations?

Response: We agree, it more likely due to models not getting the mechanism correct and/or having the incorrect seasonality (e.g., timing and frequencies of SSWs). The sentence has been revised to reflect that the models aren't able to capture the weakening and reversal of the PNJ response seen in observations.

Lines: 415 – 418.

Table 2: Can you explain/comment on the rather large and significant correlations between the observed PNJ and the PNJ in Exp1 and Exp1-NoQBO? This must be due to the BCs.

Response: A comment has been inserted in the main text regarding the possibility of the time-varying boundary forcings playing a role in driving PNJ variability in November and December.

Lines: 428 – 429.

Figure 7: How are the SSWs defined? Also, the figure indicates that the SSW frequency in ERA5 is ~1. I find a SSW frequency which is much lower, more like 0.6-0.7, depending on the time period. Lastly, the DJF U1060 strength in the ERA5 is ~30 m/s, and not ~23 m/s.

Response: Thank you for highlighting the need to specify the SSW definition in this manuscript. Text has been added to explain the definition and alludes to the increase in SSW events compared to the 60°N 10 hPa definition (e.g. Baldwin et al. 2021).

Thank you, also, for your careful analysis of Figure 7. The ERA5 strength in the figure has been corrected. The plot now shows results for Exp1-ObsQBO (which yields similar correlations as for Exp1 and Exp1-NoQBO). As for the SSW frequency for ERA5: as noted in the text we are using a different, somewhat more flexible SSW definition than the traditional one based on 10hPa 60°N zonal wind reversal, which naturally increases our SSW count, but also accounts for different model biases in the vortex.

Lines: 440 – 444, 487

409: "If we …". Please provide a short explanation why doing this might be useful.

Response: Since Figure 8 has been significantly modified so that it does not show the ENSO-filtered results (in response to a comment by Reviewer 4), this part of the text has also been changed, and focuses on the Holton-Tan ensemble results instead, including a short explanation.

Lines: 493 – 495, 504

Figure 8: The figure is not very clear. Perhaps, to reduce the amount of information, could you just show differences in SSW-frequency between QBOE and QBOW? Also, perhaps use different signatures for the different models. This would allow identifying whether some models reproduce the observations better than others. Also, why is El Nino not shown, and La Nina shows only SVMM?

Response: In response to comments by yourself and Reviewer 4, Figure 8 has been significantly reworked to *only* show the QBOE-QBOW differences in SSW frequency. The individual experiments are now indicated with markers, the same as used in Figure 7. The ENSO-filtered results have also been removed leaving only the full multi-model and Holton-Tan ensembles.

We mention in the text that filtering by ENSO phase was performed but that it did not produce significant results due to the very large uncertainty introduced by sampling such a small subset of the 1979 – 2020 boreal winters.

Lines: 521 – 523, 499 – 503.

546: It is interesting that the observed and simulated latitudinal position of the STJ agree. Is this true when taking the zonal mean over the entire globe, or only over the North-Pacific? If this is for the entire globe, then perhaps calculate the observed and simulated STJ positions over the North-Pacific, and if there is disagreement, this may change your interpretation of Figure 12.

Response: The latitude of the climatological STJ in the manuscript is for the North Pacific. The text has been updated to clarify this point.

Lines: 656

Figure 12: Could you use the same experiment names as in the paper?

Response: We have replaced Figure 12 with a QBOE-W version with the correct experiment names.

Lines: 676

Figure 12 caption: "The QBOW and QBOE groups are each deseasonalized …". It is not exactly clear what is done and what is meant here. I assume the zonal wind, but, model-by-model, or for the multi-model mean? Or did you deseasonalize the u-wind used to define the QBOW and QBOE groups?

Response: We have clarified what has been done in the manuscript by adding the following explanation to the caption of figure 12, 'The zonal wind for each ensemble member (within a given model and experiment) is deseasonalized to produce anomalies. These anomalies are then composited by QBOE and QBOW phases before subtraction. '

Lines: 679 – 681.

Figure 12 caption: Does the vertical dashed line show the latitude of the climatological (October-March?) STJ over the Pacific (?) seen in ERA5? Maybe clarify this.

Response: Thank you for highlighting the need for clarity. The latitude of the climatological STJ is calculated from October-to-March across all years over the Pacific in ERA5. The caption for Figure 12 has been updated accordingly.

Lines: 688.

579: Exp1 essentially shows no wind response that would be consistent with the obs. So, "weaker tropical and subtropical wind response" is an overstatement.

Response: We understand the potential for confusion that arises from imprecise wording. The text has been tightened up to state that there are weaker tropical and subtropical responses for individual models leading to no wind response in the ensemble mean.

Lines: 693 – 697.

612: What is "tropical (Nino3.4) precipitation"?

Response: This forms part of the text associated with Figure 14 which has now been removed following the suggestion of yourself and reviewer 4.

Line: 732.

Figure 14 and its discussion: This figure is confusing and difficult to understand. In my opinion, the figure shows that there is no relationship as the correlations depend on single outlier models. How significant are these correlations? In Fig. 14c, it seems the positive correlation would become negative if model EMS4 would be excluded. I strongly recommend excluding Figure 14 and its discussion from the manuscript. The paper is already relatively long.

Response: We agree with the suggestion and have removed Figure 14 and associated text.

Lines: 724 – 746.

635: It is stated that the sensitivity of QBO teleconnections to model biases in the equatorial zonal-mean winds has been investigated. This appears to be somewhat overstated, as the paper does not clearly identify or quantify the relevant model biases. It would be helpful if the authors clarified which specific biases are being referred to and how they are diagnosed.

Response: Thank you for highlighting this omission. The authors' involvement with this paper and also Anstey et al., (2026) have resulted in this oversight. The bias-correction alluded to in this paper is reducing biases in the amplitudes of the vertical QBO zonal wind structure.

One of the main outcomes of QBOi phase 1 was that models tend to show similar biases in the QBO zonal-mean zonal winds which includes an underestimated amplitude at the lower altitudes of the QBO around 50 hPa. In QBOi phase 2 the Exp1-ObsQBO experiment relaxes the model zonal winds in the tropical stratosphere toward reanalysis winds, thus imposing a realistic (unbiased) QBO (Anstey et al., 2026). I refer the reviewer to the contents of Anstey et al. (2026) including Figure 4.

Clarifications have been added to the Introduction, Section 2 . The line in the conclusion already mentioned that the model biases in the equatorial stratospheric zonal mean winds has been investigated. This line has been rewritten to more concisely reflect the investigations of this manuscript.

Lines: 175 – 177, 198 – 199, 256, 755 – 757.

640: “… exhibited a negative PNJ anomaly …”. You need to be more specific. This is for QBOE-QBOW conditions, I assume.

Response: Thank you. More specific wording has been inserted.

Line: 761.

643: Again, you mention very unspecific model biases.

Response: Thank you. More specific wording has been inserted.

Lines: 764 – 765.

Technical comments

84: “… lies at approximately …”

Removed in response to another comment.

Line: 89.

125: “… on either side …”

Fixed

Line: 132.

234: “through to” ?

Fixed

Line: 267.

302: “observed”

Fixed

Line: 348.

378: “winter frequency”

Fixed

Line: 447.

400: during > between

Changed to 'for' as this relates to the QBOE-W frequency difference not the difference between Exp1-ObsQBO and Exp1. Figure 8 has been modified to show QBOE-W difference in SSW frequencies.

Line: 478.

477: Please check this sentence, in particular the "consistent with the model's". Which model do you refer to?

Response: we have corrected the sentence by removing "model's" and replacing with "weak MME".

Line: 577 – 578.

539: "..., on either side ..."

Fixed

Line 648.

Response to reviewer 3

Response to Reviewer #3 (Prof. Lesley Gray)

This is a very useful paper documenting the impact of correcting biases in model representation of the lower stratospheric QBO on teleconnections to the extra-tropics and troposphere. The main conclusion of the paper is that several aspects of the teleconnections are improved by correcting the QBO bias but there remain many significant aspects that are still not reproduced by the models. The paper is therefore useful in highlighting areas that require further investigation and potential model improvements.

The authors thank Professor Gray for these positive remarks.

I believe the paper is suitable for publication but it requires fairly extensive (but hopefully not onerous) revision.

Major comments.

1. The title suggests only extra-tropical teleconnections are examined but there are some interesting results of teleconnections with the sub-tropical jet and tropical rainfall, so I suggest a slight revision of the title to reflect this.

Response: The subtropical jet is generally considered to be at ~30°N-35°N which is in the extra-tropics. There is also a companion paper analysing tropical teleconnections. Therefore, to avoid confusion we prefer to keep the title quite simple and general, leaving the abstract to provide additional detail.

2. With regard to figure 2, I think it would be useful for the reader to also see the equivalent plot for the multi-model Exp1 results. The text states that more models achieve the H-T effect in Exp1-ObsQBO but it would be good to show this in fig 2.

Response: Fig 2 shows the ensemble mean and is only viable because all ensemble members for Exp1-ObsQBO have near identical nudged QBOs. In Exp1 the QBO is part of the internal variability and would be removed or significantly reduced in the ensemble mean.

3. Figures 2 and 3 clearly show a reversal of the H-T effect in late winter (Feb/March) in the ERA5 data which is not reproduced by the multi-model mean in figure 4. This needs mentioning / discussing both here and later (line 354) when you discuss table 2 (and probably also when you discuss the failure of the models to reproduce the timing of SSWs), including references to earlier studies that have examined this reversal. Based on figure 3, it also seems surprising to use the DJF comparison (lines 325-327) to compare ERA5 with the multi-model mean because the ERA5 data have strong negative PNJ values in Dec/Jan counterbalanced by very weak (possibly positive) values in February, whereas the MME clearly shows a weak negative value in all three months. This should perhaps be noted in the text. A better comparison would be to

compare DJ and FM averages, which I suspect would more clearly highlight the ERA5/MME differences.

Response: The reversal in sign response in ERA5 for March is now noted. We have included a reference to the paper by Lu et al. (2020) in Section 3, a paper which was already referenced elsewhere in the manuscript. We agree with the reviewer that the differences between ERA5 and MME are likely due to seasonality (e.g. timing of SSWs) therefore it is important to check that after removing the differences in seasonality with a seasonal mean (e.g. DJF) if ERA5 and the MME agree or disagree. The two-month average doesn't eliminate differences in seasonality hence it is not clear what it offers in addition to the monthly means.

Lines: 290, 297 – 298, 315, 332 – 333, 415 – 418, 769.

4. The analysis finds that the QBO influence on the NAO is weak and mostly insignificant in the models. Did you explore whether the results were sensitive to the definition of the NAO? It would be interesting to see maps of surface pressure correlation (as e.g. in Gray et al. 2018) to see if there is any correlation at all in the Atlantic. It may be that a signal is evident in the subtropics but absent at higher latitudes because of the higher background variability there.

Response: Given the relatively weak extra-tropical stratospheric response to the QBO in the MME we didn't consider it appropriate to devote too much space to the NAO response. As noted in our response to Reviewer 1, the lack of MME NAO signal is most likely not because of the chosen definition of the NAO but more the fact that the models show different seasonality of the QBO-PNJ coupling and its propagation to the surface. We looked at the two-station NAO index (Azores, Iceland) and found that the results were consistent. The two-station equivalent of Figure 11 has been included as Figure A4, and the text has been modified accordingly. A detailed analysis of the NAO response could stretch to a whole paper, including correlation maps, which is very much encouraged, but is beyond the scope of this manuscript.

Lines: 600 – 601, 620 – 621, 867 – 869.

5. All figure captions need to be re-visited and revised where appropriate. Figure captions should provide information on what is plotted. Currently, the majority include interpretive statements that are more appropriate in the main text.

Response: The interpretive statements were included to allow the reader to gain a quick understanding from the figures-alone without having to refer to the main text. These have been removed as requested.

Figure captions: 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13.

Minor comments

Line 51: 'influence the latitudinal profile …' – in what way? In good agreement with obs?

Response: Revised to "improve its influence on……"

Line: 53

Line 64: do you mean 1 hPa here? Surely this level is dominated by the SAO?

Response: Changed to "around the stratopause," consistent with the Anstey et al (2022) QBO review.

Line: 68.

Line 72: worth adding also that the QBO influences tropical precip, also less well understood, given that you mention it in the abstract.

Response: The influence of the QBO on tropical precipitation is now mentioned in the appropriate place further on in the introduction in response to a comment by reviewer 1.

Lines: 134 – 141.

Line 84: typo - 'it lies AT approximately 6 degrees N'. Also, how can the QBO be westerly if the zero wind line is in the N hemisphere? Do you mean 6 degrees S?

Response: The phrase has been removed from the revised text.

Line: 85 – 89.

Line 107: 'the frequency of occurrence for SSWs …. For QBOE winters …..' Which months do you use to define 'winter'? DJF? Which months you use matters, given that SSWs tend to occur later in QBOW years.

Response: This is just reporting results from Baldwin et al. (2021) and Gray et al. (2024) and the WCD reader is referred to those papers for the full details.

Line 115: I suggest replacing 'the slowly varying QBO might improve NAO predictability' by 'improved model representation of the slowly varying QBO and its teleconnections might improve NAO predictability'.

Response: Changed as suggested.

Lines: 121 – 122.

Line 147: I misunderstood this text at first, and didn't realise that the 'two sets of experiments ….' referred to the Anstey et al. (2022b) study and not to this current one. It would maybe help if you started a new paragraph at 'Of relevance….' and merged it with the following paragraph which is still discussing the Anstey paper.

Response: Revised as suggested.

Lines: 161 – 162.

Line 155: when you say 'ensemble-mean' here, do you mean model-mean too?

Response: No, we mean ensemble-mean.

Line 175: this paragraph was clearly written before the addition of analysis of the QBO impact on tropical rainfall and the subtropical jet. You also need an extra paragraph somewhere in the introduction to introduce the tropical rainfall / STJ issues and mention relevant previous studies.

Response: In response to reviewer 1 we have included additional text relating to QBO-related tropical precipitation anomalies earlier in the introduction. The paragraph you mention has now been removed, the essence of which has been incorporated in the following paragraph starting 'Section 2'.

Lines: 134 – 141, 196 – 202.

Line 211: 'suggested adopting ….' – this is too vague. Did all models follow this suggestion?

Response: This whole paragraph has been removed in light of Anstey et al. (2026) becoming available in the public domain. Additional explanation has been moved to the start of section 2.

Lines: 205+, 237 – 243.

Line 230: I think you need to say (in the text) how the PNJ is defined.

Response: The PNJ definition has been included.

Lines: 440 – 444.

Line 249: the figure caption is meant to explain what is plotted in the figure. It should not be a statement of interpretation.

Response: Caption revised as suggested.

Line: 283.

Line 252: I think you need another couple of sentences at the end of section 2 that describes how you derived the multi-model mean and also the statistical significance estimates. Also, perhaps how you defined the NAO and why you chose the method you used (other studies have used EOFs and/or pressure differences between Iceland / Azores); presumably you used your NAO method because it was easiest to extract from the model data?

Response: How the multi-model mean and statistical significance are derived is given in Section 3. The NAO definition has been moved from the caption in Figure 10 to Section 5 in

response to the previous request. The broader area definition of NAO can help to capture model NAO better in general, given that the centres of action in different models will differ. We also looked at the QBOE-W results using a two-station NAO index and now include those results in Figure A4 – they are consistent with the original results. Using an EOF based definition would be better but that is difficult to do as a multi-model mean.

Lines: 292 – 294, 583 – 584, 600 – 601, 620 – 621, 867 – 869.

Line 262: see previous comment about caption 2. The first sentence of this caption is an interpretation rather than description of what is plotted. Also, the text describing how the significance is derived should be in the main text of section 2.

Response: The caption of Figure 3 has been revised. The derivation of statistical significance has been moved to where it first occurs in the text.

Lines: 292 – 294, 302 – 307.

Line 271: the text 'of -2ms-1 December-March' does not make sense.

Response: Changed to "of -2ms-1 over the four months December to March"

Lines: 313 – 314.

Line 272: I think you should also comment that the MME does not capture the H-T reversal at end of the winter (Feb/March).

Response: Comments have been added which note that the models (both free running and with a nudged QBO) failure to capture the reversal in the sign of response in March.

Lines: 315 – 316.

Line 290: did any of the separate models capture the late-winter H-T reversal?

Response: None of the individual models for Exp1 managed to capture the reversal in the sign of PNJ response in March. For Exp1-ObsQBO, only one model (BCC-CSM-MR) managed to capture the transition from negative to positive PNJ response from February to March.

Line 302: typo – observe

Response: corrected

Line: 348.

Line 314: see previous comments on figure captions. The first sentence of the caption is an interpretive statement, not a description of the figure.

Response: All figure captions have been revisited and revised in response to the reviewer's earlier suggestion.

Line 338: this is a nice result.

Response: Thank you for the compliment.

Line 348: do any of the models display inter-annual variability as large as the observed?

Response: Broadly, all the models have the correct interannual variability as shown in Anstey et al. (2026) which is now in the public domain. A lot of this variability is removed with the ensemble averaging even when the simulations all have the same nudged QBO. In reality the cycle-to-cycle variability of the QBO is part of the natural internal variability and would be lost when averaging over realizations.

Line 354: please see major comments (point 3) on the late-winter reversal. I don't understand the comment that the late-winter response in observations is being masked by other influences – are you suggesting that the late-winter reversal is not real?

Response: No, the observed late winter reversal is indeed real. The text has been revised to remove the suggestion.

Line: 417 – 418.

Line 361: see previous comments on figure captions.

Response: All figure captions have been revisited and revised according to the reviewer's earlier suggestion.

Line 366: these early winter correlations in Exp1 and Exp1-noQBO are interesting / puzzling. Do they suggest that SST/SI influences may be important in early winter?

Response: Yes. The SST/SI and other forcings are the only other external source of predictability in these experiments. An extra sentence has been added.

Lines: 428 – 429.

Line 376: I suggest you reverse the caption heading, since you discuss frequency of SSWs first.

Response: Section heading modified.

Line: 439.

Line 436: is your conclusion then that the HTE is not simply due to the occurrence (or otherwise) of SSWs? Do you think it more likely the accumulation of relatively smaller wave impacts?

Response: It is generally accepted that the HTE is not solely due to the occurrence of SSWs. More likely this result indicates that there are issues with the morphology of SSWs in the model and/or deficiencies in mechanism producing a HTE in the models.

The paragraph containing the text from lines 436-440 (original manuscript) has been removed, and Figure 8 and its discussion have been significantly reworked. As discussed in the new text, filtering the models according to those that produced a Holton-Tan effect does result in a marginally significant, positive QBOE-W SSW frequency difference for both Exp1 and Exp1-ObsQBO.

Lines: 468+, 535 – 539.

Lines 468-475: this is consistent with figures 3/4 that show the models do not capture the late-winter H-T reversal.

Response: See our response to major comment #3 above.

Line 482: I think this section contains a useful set of analyses, even though the results are weak and mostly insignificant. I think it's important to report null results. You could also point out the differences between figures 3 and 4 in the extent to which the responses penetrate into the troposphere at high latitudes. The model responses do not penetrate as deeply as the ERA5 response. In fact, the responses in January and February appear to penetrate further into the troposphere in Exp1 than in Exp1ObsQBO, which is interesting.

Response: This paragraph has been revised to make the important link to figures 3 and 4.

Lines: 586 – 587, 590 – 591.

Line 485: describe how you define the NAO (it does not belong in the caption to figure 10).

Response: The definition of the NAO has been moved from the figure caption to the main text.

Lines: 583 – 584.

Line 528: please see major comment 4 regarding an additional surface level pressure analysis, which would also be helpful to this section.

Response: Please see response to major comment 4.

Line 536: please tell the reader here that you have switched to showing W-minus-E rather than E-minus-W, and explain why.

Response: We have changed to QBOE-W, modifying figures 12 and 13, and the text as appropriate.

Lines: 646 – 655, 669 – 671, Figure 12, 690 – 694, 703 – 704, Figure 13, 718 – 720, 800.

Line 541: presumably when you say neither are significant, you mean significantly different from zero?

Response: Yes. The text has been updated to confirm that 'neither are significantly different from zero.'

Line: 650 – 651.

Lines 556-563: nice results here.

Response: Thank you. We appreciate the compliment.

Line 584: maybe say that it's an E-W dipole?

Response: Thank you for highlighting the need for a more precise description. The text has been modified to reflect that is an east-west dipole.

Lines: 703 – 704.

Line 601: is this a weakening of the Walker circulation under QBOW?

Response: We thank the reviewer for identifying a misleading nuance in the text. The responses are associated with QBOE-W and not a specific QBO phase. The text has been modified to reflect this.

Lines: 717 – 720.

Line 635: you state 'extra-tropical' here, but you have also looked at the STJ and tropical rainfall.

Response: Please see our response to major comment 1.

Line 646: I think a sentence to say that the models do not reproduce the late-winter H-T reversal would be appropriate here.

Response: Sentence added.

Line: 769.

Line 693: ditto comments about the model failing to capture the late-winter H-T reversal, which is likely linked to their failure to capture the SSW frequency and timing response.

Response: Given that this is speculation we prefer not to put too much emphasis on this in the final paragraph. It has been mentioned a number of times previously.

Line 701: I suggest removing the word ‘quite’ here.

Response: Removed as suggested.

Line: 829.

Response to reviewer 4

Response to Reviewer #4 (Prof. David Battisti)

Dear Dr. Andrews,

We have received three thoughtful reviews of "Extratropical teleconnections in a multimodel ensemble nudged towards the observed QBO" (WCD 2026-737). The ball is now in your court to post your response to how you and your co-authors intend to respond to these three reviews. In the way of some guidance, I note that all three reviewers have significant concerns that must be addressed in a revised manuscript for it to be publishable in WCD.

Regards, David Battisti

PS. I include below some of my own comments on the manuscript.

Response: The authors thank Professor Battisti for these positive remarks.

Editor's additional comments:

Figure 2: what months are included in "mid-winter"?

Response: "mid-winter" removed and instead the specific months are now given.

Lines: 283 – 284.

Line 270 and elsewhere: Please provide information on how the BCs differ; how the experimental periods differ; how the models changed. Have they all changed? Are some of the models used in phase 2 identical to that used in phase1? If so, which ones?

Response: The experiment description has been rewritten (Section 2) plus we added reference to the QBOi phase 2 protocol paper (Anstey et al., 2026) which wasn't available in the public domain at the time of submission. We believe that together these changes address all the issues raised here.

Lines: 205 – 246.

Line 285: what is meant by significant? Statistically significant? If so, at what level?

Response: Details of the "significance" is now given.

Lines: 292 – 294.

Line 330: please also state the regression coefficient for the models and for observations.

Response: Thank you for highlighting this important omission. The regression coefficients for the experiments and observations have now been included. They highlight that the strengths of the ensemble-mean QBO-PNJ teleconnections are largely unchanged for Exp1

and Exp1-ObsQBO, and that the improvement is linked to a more realistic amplitude of the nudged QBO. Additional text has been included in section 3, the abstract and conclusions.

Lines: 378 – 384.

Lines 339-41: Rather than look at the relationship in the MMM, perhaps a better indication of how well the models are doing is to string together the time series from all models and then correlate the PNJ with the QBO.

Response: The reason for using the MMM is to eliminate or dampen internal variability. It likely that the correlation for an extended timeseries will already be close to its asymptotic values after 40 years.

Line 390-91: How different are the SSW frequencies in the warm vs. cold phases of ENSO. Please quantify this (in observations and models) by calculating the p value for the difference between two means, taking into account the number of events in QBOE and QBOW. Put 2 sigma uncertainty bars on the mean values for the ERA results, for the ALL ENSO and La Nina cases.

Response: As noted in response to your next comment, Figure 8 has been greatly revised and now shows uncertainties (error bars) on all QBOE-W SSW frequencies. In addition, we have removed parts of the text referenced above which talk about difference in SSW frequencies during different ENSO phases as these were found to be broadly small and insignificant, and, instead, mentioned the neutral-ENSO phase results which were not significant.

Lines: 523 – 533, 467 – 520.

Lines 388-395 and figure 8. In addition to what is shown in this figure, you might consider looking at the difference between the SSW frequencies in QBOE and QBOW for each model, and then plot the model spread in the difference. If it is interesting, you could plot this as panel b in figure 8 (or simply add more columns to the existing figure, but that might be confusing).

Response: Figure 8 has been significantly reworked based on your comment to only show QBOE-QBOW SSW frequency difference for each experiment, for the multi-model ensemble, Holton Tan ensemble (models with a significant HTE), and ERA5. The spread in these values is also shown, using Poisson uncertainty for ERA5 and standard deviations for the multi-model ensembles.

Lines: 523 – 533, 467 – 520.

Figure 12 is difficult to read. Please split the figure into panel a (Exp1) and panel b (Exp1-ObsQBO) so we can see the difference.

Response: The co-authors prefer the combined version of the figure as it allows a direct comparison of the ensemble mean zonal-wind anomalies between Ex1-ObsQBO and Exp1. We also prefer not to increase the number of figures. We have, in response to a different reviewer comment, flipped the diagram to show QBOE-W which is consistent with the other sections.

Lines: 676 – 682.

Figure 14 and the discussion. I agree with Anonymous Reviewer #2. This figure is very difficult to read, and correlations are very likely not significant. I strongly recommend you remove this figure and the related discussion.

Response: We agree with the suggestion and have removed Figure 14 and associated text.

Lines: 723 – 746.

Additional modifications not directly related to reviewer comments:

Line 51: Minor editorial revision.

Lines 56 – 58: Minor editorial revision.

Line 91: Minor editorial revision.

Lines 184 – 190: Minor editorial revision.

Lines 258 - 259: Minor editorial revision.

Lines 262 – 265: Minor editorial revision.

Line 350: Minor editorial revision.

Line 537: Minor editorial revision.

Lines 458 – 453: Minor editorial revision relating to Figure 7 now showing Exp1-ObsQBO results (which gives similar results for Exp1 and Exp1-NoQBO, Lines 452 – 453).

Lines 453 – 466: Replaced existing text with a more concise analysis relating to ENSO-filtering of SSWs.

Lines 561 – 569: Minor editorial revision.

Lines 625 – 626: Minor editorial revision.

Lines 781 – 782: Minor editorial revision.

Line 841: Minor editorial revision.

Lines 844 - 845: Minor editorial revision.

The ordering of each set of bars for Figures A1 and A2 has been revised to maintain consistency with Figure 9. Lines: 850 – 860.

Lines 933 – 935: Removed reference.

Lines 937 – 939: Added reference.

Lines 950 – 957: Updated Anstey et al., (2026) reference with latest information.

Lines 959 – 960: Removed reference.

Lines 999 – 1000: Added reference.

Lines 1152 – 1153: Added reference.

Lines 1201 – 1202: Removed reference.