Review of Andrews et al. (2026) submitted to WCD

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.

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.

Specific comments

I 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, …”

94: “This results in …”

96: “This modulates the concentration of …”

99: “Regardless of the …”

105: “If the PNJ …”

118: “For example …”

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

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.

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.

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

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”.

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

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

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?

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

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?

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?

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

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.

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.

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?

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

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.

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.

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

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?

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.

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

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?

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.

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

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

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.

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.

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

643: Again, you mention very unspecific model biases.

Technical comments

84: “… lies **at** approximately …”

125: “… **on** either side …”

234: “through to” ?

302: “observe**d**”

378: “~~winter~~ frequency”

400: during > between

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

539: “…, on either side …”