Manuscript # WCD-2022-14 Response to Reviewers

García-Franco, JL., Gray LJ., Osprey S., R. Chadwick and Z. Martin

Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford. Parks Road Oxford, United Kingdom OX1 3PU email: jorge.garcia-franco@physics.ox.ac.uk

---

1. Reviewer 1

The authors provided detailed responses to the points I raised in the previous review stage and explained many aspects of their model configurations.

Since the paper is quite long to read, I would suggest to shorten the text wherever possible (some examples below), otherwise I just have some minor remarks (by line number in the track-changed paper):

We thank this reviewer for their time reviewing our manuscript in this second round of review. We have shortened the manuscript removing the text suggested by this reviewer. Additionally, following the suggestion by Reviewer 2, we have shortened section 3.1 and also removed a couple of sentences in the introduction, all of which amounts to about 15 lines less in the manuscript. Below we present a detailed point-by-point response to your concerns and questions. Your comments are shown in black,our responses in blue and the changes made to the manuscript, where appropriate in violet.

L58-9, since this is also written below, it can be deleted here

Deleted.

L95, daily or monthly?

Monthly, thank you for pointing out this ambiguos statement, we clarified this by changing these lines to the following:

Monthly mean data was used for all observational datasets.

L107, sentence is a bit repetitive

This paragraph was reworded for the (shorter) following text:

The three simulations (GC3 N216-pi, UKESM N96-pi and GC3 N96-pi) used in this study are 500-yr long and have the same experimental design with only 1 ensemble member. The simulations were run with constant year-1850 external forcing, further detail about the MOHC piControl experiments can be found in Menary et al. (2018) and about the UKESM1 model in Sellar et al. (2019).

L122, since this is not the main topic, the sentence is not necessary

Removed.

Fig 3e, how do you combine ensemble members to calculate this time series?

The simulation shown in Fig3e is only GC3 N216-pi which consists of only 1 ensemble member as indicated in section 2.2 and in the paragraph above.

L231, I guess this analysis should be done accounting for underlying trends, so maybe the sentence can be omitted and the topic left for other studies

Rephrased.

L234, 'a multidecadal modulation of...'?

We have removed this text, in order to highlight the advantages of these simulations compared to observations, as suggested by Chaim Garfinkel.

L268, you may want to note that model indices amplitudes are 2-3 times smaller than observations

We have noted so in the text.

L343, NN ENSO -¿ NN

Changed.

L358, extra ')'?

Removed.

L378, for me the 'constant' is this one `https://en.wikipedia.org/wiki/Gravitational_constant`, not 'small g'

We have referred to the 'g' in this context as gravitational acceleration in the revised text.

L459, please rephrase to avoid repetition

Removed.

2. Reviewer 2

I thank the authors for their detailed response to my initial comments, including a careful discussion of the differences between their study and that of Rao et al 2020. I appreciate their focus on the role of the QBO level (70hpa vs. 30hPa) and signal to noise ratio (i.e. how many years are included in the composite). I'm going to sign my review, as it is about to become fairly obvious who wrote this review (if it wasn't already obvious from the first round). I also want to apologize if my initial review was a bit too harsh and dismissive, as I really like this paper!

Dear Chaim,

Thank you for your time reviewing our manuscript and your comments which have greatly improved the revised version. No apology was necessary as we understand the reasons for your earlier concerns given the disagreement between our results and those of Rao et al. (2020) and we believe that by investigating these differences and discussing them directly in our manuscript we have improved the paper substantially.

I have one more general comment on section 3.1, and also a few remaining comments on the difference between Rao et al 2020 and this paper.

Section 3.1 still seems to be overly precise when trying to compare the model to observations. The GPCP response shown in figure 1a and the left column of figure 2 reflects the fact that the data is available only from 1979 to the near present. Over this period there were more EN during WQBO. If the observational precip data product was available for more years, then the observed signal would be different (as the authors show shortly for SST). In other words, there is substantial uncertainty on the observed response.

Because of this, I don't think it makes sense in the text to compare the model to observations in a quantitative sense nor to focus on the details of the response, as the observational signal is fundamentally unknown (e.g. Deser et al 2017; Journal of Climate on ENSO teleconnections) and the model SSTs are not the same as obs SSTs.

Stated another way, I would expect the model response to be weaker than obs because the SST response shown in figure 2 is weaker in the ENSO region.

Stated a third way, if we had a gridded, observed precip product for the period 1953 to the near present, I speculate that the agreement with the model would be better.

If the authors agree with my interpretation, the text itself in section 3.1 needs to be modified, though the main conclusions will be generally unchanged (and in fact, the model would actually become more suitable for the analysis the authors subsequently perform).

Thank you for this suggestion. We agree with you that if there was an observational dataset for precipitation available for an earlier period, 1953-2021 or 1921-2021, Figures 1 and 2 would look different. As a result, we have decided to shorten section 3.1 by removing text that compared model and observed responses. However, the Figures themselves set the scene for the rest of the paper so we have not changed

the Figures and we have kept some sentences that indicate that the observational results agree with previous studies. Half-way through section 3.2 we have added the following text:

The observed multidecadal changes to the ENSO-QBO relationship (Fig. 3) means that the precipitation response (Figs. 1 and 2) would likely be different if a longer record of precipitation was available. While our analysis of the observed record is affected by statistical uncertainty (e.g. Deser et al., 2017), this is likely not the case in the pre-industrial control simulations given their length and constant external forcing. This result further highlights the advantage of using these model experiments to understand QBO tropical teleconnections, including ENSO relationships, in the remainder of this paper.

I also have a few comments on the discrepancy between Rao et al and this paper. The first is that Rao et al considered many models where the QBO would be ill-defined at 70hPa. Hence it would be impossible to consider the role of the QBO at 70hPa on impacts outside of the QBO region in such models. While the Met Office models do indeed have a too-weak QBO at 70hPa, this model was actually one of the better ones in this regard (though its periodicity was too long as the authors acknowledge). In order to have a common definition for all models, Rao et al adopted the 30hPa level for all models.

Second, Rao et al identified a tropical convective signal associated with the QBO at 30hPa which differs from the one in this paper in its pattern. Rao et al also identified a robust signal in 100hPa buoyancy frequency for this phase of the QBO in observations and in most models, including the Met Office models which were among the best performing (Figure 9 of Rao et al). My interpretation is not that the winds at 30hPa have a direct effect on buoyancy frequency and convection, rather that this is a convenient way to pick a particular phase of the QBO whose downward extension has a direct impact on the TTL. For this specific phase of the QBO, the Met Office models struggle to represent the convective impact even as they did a reasonable job with the buoyancy frequency anomalies at 100hPa. This could be because of biases in the QBO itself (e.g. downward propagation to the lower stratosphere, or the overly long stalling of lower stratospheric anomalies), or a small signal to noise ratio that a single ensemble member may miss (as the authors point out).

My own speculation/intuition based on the results from Rao et al and the current paper is that there may be multiple QBO regimes with an impact on tropical convection, but future work is clearly needed to sort out whether this indeed the case and why. While I agree that the 70hPa level is best to diagnose a direct impact on the TTL, the unfortunate reality is that nearly all models still struggle with the downward extension of the QBO to the lower stratosphere with very little progress having been made recently and with few ideas on how to improve the situation (other than substantially more resolution, as suggested in Garfinkel et al 2022; JAMES). Hence a focus on 70hPa necessarily excludes many models which may still have teleconnections from the QBO higher up. I would suggest that as a community, we should consider teleconnections associated with different QBO levels (e.g. both 70hPa and 30hPa), so as to be able to include models with relatively larger biases in the QBO in the lowermost stratosphere.

Performing such an analysis is well outside the scope of the authors' paper, and specifically the authors

could decide to not include any of it. However, the authors may want to include more about this sensitivity to QBO level and the nature of biases in most models in their discussion section.

Signed, Chaim Garfinkel

Thank you for these insightful remarks. We agree with your comment that a 70 hPa index would not work for those models in which the weak QBO amplitude bias in the lower stratosphere is so large that the QBO would be ill-defined at this level. The discussion in the manuscript aims to clarify differences between our findings and those of Rao et al. (2020) and not to indicate that only this level should be used for all models. Multi-model analyses should indeed consider multiple levels or an index that captures the state of the vertical profile (e.g. Schenzinger et al., 2017). In the discussion section, in the second-to-last paragraph, we have added the following sentences to highlight the role that stratospheric biases have for diagnosing teleconnections and the uncertainty introduced by the weak amplitude bias in our results.

Tropospheric biases, e.g., in the strength or position of the ITCZ (Fig. 9), may limit the robustness of these results and may mean that the impacts diagnosed in this study may be different in another model. Similarly, stratospheric biases such as the weak amplitude of the QBO in the lower stratosphere found in most models (Bushell et al., 2020, Rao et al., 2020), means that the simulated tropical pathway of QBO teleconnections may be weaker, nonexistent or difficult to diagnose in some models, highlighting the need to improve vertical resolution in GCMs (Garfinkel et al., 2022).

2.1. Minor comments

Line 227 please rewrite "for the most part of the simulation"

Reworded to: mostly positive.

Line 242: "However, the equatorial Atlantic and Pacific MAM responses are stronger when ENSO events are included." This isn't obvious to me from figure 4.

We have removed this sentence.

Table 1: I found the caption included for this table confusing. Are the stated units ("#months EN/# months W") correct? Shouldn't it be ("#months ENSO/# months QBO")? Also, "standard deviation of the PDF" is confusing as well – I think you mean you did a bootstrapping in order to quantify the uncertainty of #months ENSO/# months QBO, but maybe I misread.

We have changed the units in the caption as suggesed and yes, a boostrapping of the data was done to account for model and observational uncertainty. This is now clarified in the table caption but also in the text as follows:

Probability density functions (PDFs) were constructed, first, for the observations by bootstrapping with replacement to account for observational uncertainty, and for the model data using 39-yr samples to match the length of the ERA5 period.

Section 3.4 The word "explain" on line 373, 399, and 440 seems overstated. There is no casual explanation here, as the authors note later. Rather the authors are establishing a self-consistent framework or schematic that allows for connecting tropical anomalies in disparate regions.

These sentences have been rephrased.

Deser, Clara, Isla R. Simpson, Karen A. McKinnon, and Adam S. Phillips. "The Northern Hemisphere extratropical atmospheric circulation response to ENSO: How well do we know it and how do we evaluate models accordingly?." Journal of Climate 30, no. 13 (2017): 5059-5082.

Garfinkel, Chaim I., Edwin P. Gerber, Ofer Shamir, Jian Rao, Martin Jucker, Ian White, and Nathan Paldor. "A QBO Cookbook: Sensitivity of the Quasi-Biennial Oscillation to Resolution, Resolved Waves, and Parameterized Gravity Waves." Journal of Advances in Modeling Earth Systems 14, no. 3 (2022): e2021MS002568.

3. Reviewer 3

The authors have addressed most of my comments satisfactorily, and I am now ready to accept the paper for publication.

We thank this reviewer for their time reviewing our manuscript a second time and their constructive comments. Below we present a detailed point-by-point response to your concerns and questions. Your comments are shown in black,our responses in blue and the changes made to the manuscript, where appropiate in violet.

However, I am still confused about the different significance tests. For the composite analysis you use bootstrapping for the observations but a Welch t-test for the models? This is what you now write in the paper, but from the Reply you seem to also use the t-test for observations? Why not use the same test for both observations and models?

The main reason is because the same test would not be well suited for both observations and models. Note the very different sample sizes of the observations and the long simulations. The composite sizes of a difference such as QBO W-E (NN) (Fig 11c) would be too small (N 10 or less) for anything meaningful to be drawn from a t-test. However, the bootstrap with replacement test as done in the manuscript allows to evaluate the likelihood of a difference being affected by observational uncertainty. We have also used a bootstrap test in the model simulations for most of our results, but this bootstrapping is different by design. In this case we compute the QBO W-E difference for model samples of randomly sampled 39-yr periods, i.e., subsampling bootstrapping, repeated 10,000 times, and we evaluated how likely it is that our result is of the same sign as the mean QBO W-E difference. In other words, this process is very different to the observed bootstrapping, repeating the same process for the model would mean boostrapping with replacement in a composite with such a large size that this test would not render meaningful results. Our investigation shows that the modelling results discussed in the manuscript are not sensitive to the choice of test (Welch-test versus boostrapping into 39-yr chunks). The revised text in this section now reads:

The significance level is then interpreted as QBO W-E differences that are outside of the 95% of the distribution of randomly generated differences. The significance of the differences in the simulations is estimated using a Welch two-sided t-test, but other bootstrap methods were tested without significantly changing the results.

When considering correlations (and regressions?) a bootstrap method is also used. But I would think serial correlations should be taken into account. You can do this by using block-bootstrap.

Thank you for pointing this out, the text was misleading, the reported significance in Figure 5 uses the standard p-values from the regression model, which are based on the t distribution and the standard error. Thank you for the suggestion of using a btlock-bootstrap for our regression analyses. While the QBO index is definitely autocorrelated, the deseasonalized precipitation time-series at each grid point are most frequently not autocorrelated. We have computed the regression coefficients and the p-values using block bootstrapping as follows. Circular block bootstraping was used following (Politis and Romano,

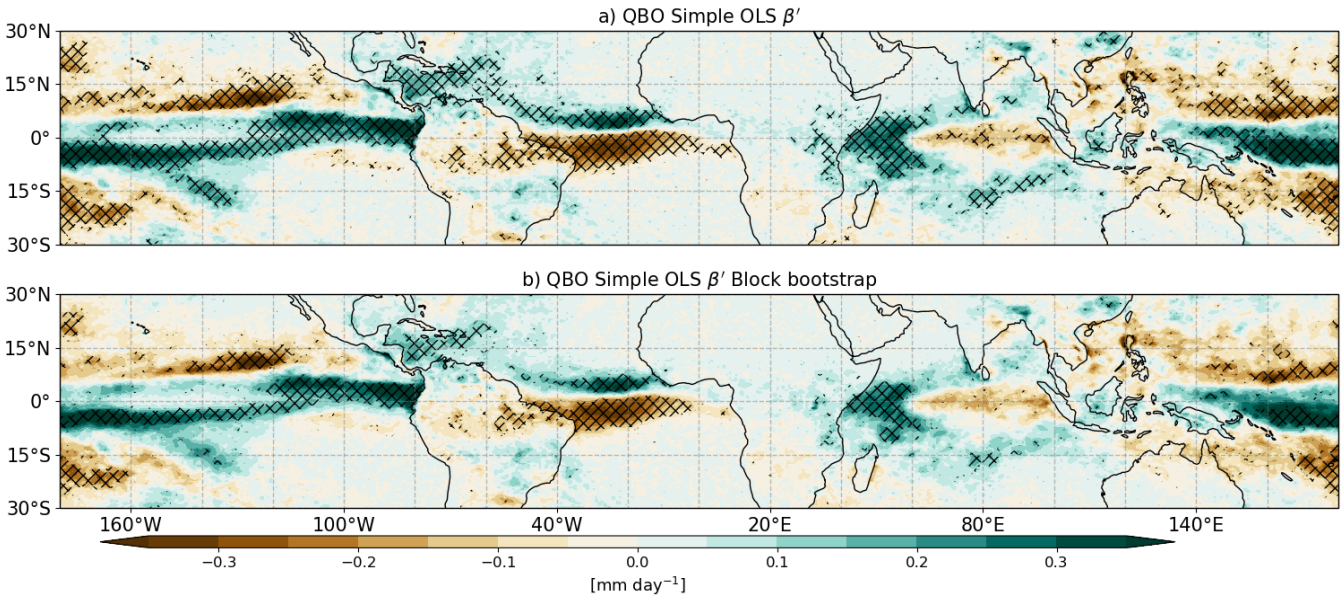

Figure 1: Regression coefficients between the QBO index and deseasonalized convective precipitation (a) as in Figure 5 of the main manuscript, (b) as (a) but showing the mean regression coefficients from a circular block-bootstrap method (see text) using a block length of 61 months.

1991) using block lengths of 21, 61 and 121 months and using 10,000 repetitions. For each repetition, the regression coefficient and the p-value are computed and stored, then, the mean values are plotted for the block length of 61 months (Figure 1). The original results hold, although several sparse regions do show a change in the magnitude of the regression coefficient and overall less significant regions are diagnosed in block-bootstrapping results. The main possible reason for these results is the lack of autocorrelation in the deseasonalized precipitation which means the regression analysis renders effectively equal results using block bootstrap methods or straightforward regression analyses. The figure caption (Fig. 5) in the manuscript now specifies:

"..and the hatching indicates significance to the 95% confidence level based on a t-test."

In Equation A1 the subscript i should be deleted.

Done.

References

Bushell, A. C., Anstey, J. A., Butchart, N., Kawatani, Y., Osprey, S. M., 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., Smith, A. K., Stockdale, T. N., Versick, S., Watanabe, S., Yoshida, K. and 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* pp. 1–31.
URL: *https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3765*

Deser, C., Simpson, I. R., McKinnon, K. A. and Phillips, A. S. (2017), 'The northern hemisphere extratropical atmospheric circulation response to enso: How well do we know it and how do we evaluate models accordingly?', *Journal of Climate* **30**(13), 5059–5082.

Garfinkel, C. I., Gerber, E. P., Shamir, O., Rao, J., Jucker, M., White, I. and Paldor, N. (2022), 'A qbo cookbook: Sensitivity of the quasi-biennial oscillation to resolution, resolved waves, and parameterized gravity waves', *Journal of Advances in Modeling Earth Systems* **14**(3), e2021MS002568.

Menary, M. B., Kuhlbrodt, T., Ridley, J., Andrews, M. B., Dimdore-Miles, O. B., Deshayes, J., Eade, R., Gray, L., Ineson, S., Mignot, J., Roberts, C. D., Robson, J., Wood, R. A. and Xavier, P. (2018), 'Preindustrial control simulations with HadGEM3-GC3. 1 for CMIP6', *Journal of Advances in Modeling Earth Systems* **10**(12), 3049–3075.

Politis, D. N. and Romano, J. P. (1991), *A circular block-resampling procedure for stationary data*, Purdue University. Department of Statistics.

Rao, J., Garfinkel, C. I. and White, I. P. (2020), 'How does the quasi-biennial oscillation affect the boreal winter tropospheric circulation in cmip5/6 models?', *Journal of Climate* **33**(20), 8975–8996.

Schenzinger, V., Osprey, S., Gray, L. and Butchart, N. (2017), 'Defining metrics of the Quasi-Biennial oscillation in global climate models', *Geoscientific Model Development* **10**(6).

Sellar, A. A., Jones, C. G., Mulcahy, J., Tang, Y., Yool, A., Wiltshire, A., O'Connor, F. M., Stringer, M., Hill, R., Palmieri, J., Woodward, S., de Mora, L., Kuhlbrodt, T., Rumbold, S., Kelley, D. I., Ellis, R., Johnson, C. E., Walton, J., Abraham, N. L., Andrews, M. B., Andrews, T., Archibald, A. T., Berthou, S., Burke, E., Blockley, E., Carslaw, K., Dalvi, M., Edwards, J., Folberth, G. A., Gedney, N., Griffiths, P. T., Harper, A. B., Hendry, M. A., Hewitt, A. J., Johnson, B., Jones, A., Jones, C. D., Keeble, J., Liddicoat, S., Morgenstern, O., Parker, R. J., Predoi, V., Robertson, E., Siahaan, A., Smith, R. S., Swaminathan, R., Woodhouse, M. T., Zeng, G. and Zerroukat, M. (2019), 'UKESM1: Description and evaluation of the UK Earth System Model', *Journal of Advances in Modeling Earth Systems* **11**(12), 4513–4558.