Prediction of the quasi-biennial oscillation with a multi-model ensemble of QBO-resolving models

Timothy N. Stockdale[1], Young-Ha Kim[2], James A. Anstey[3], Froila M. Palmeiro[4], Neal Butchart[5], Adam A. Scaife[5,6], Martin Andrews[5], Andrew C. Bushell[7], Mikhail Dobrynin[8,9], Javier Garcia-Serrano[4,10], Kevin Hamilton[11], Yoshio Kawatani[12], Francois Lott[13], Charles McLandress[14,3], Hiroaki Naoe[15], Scott Osprey[16], Holger Pohlmann[17,9], John Scinocca[3], Shingo Watanabe[12], Kohei Yoshida[15] and Seiji Yukimoto[15]

1 European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK

2 Goethe University Frankfurt am Main, Frankfurt am Main, Germany

3 Canadian Centre for Climate Modelling and Analysis (CCCma), Victoria, Canada

4 Barcelona Supercomputing Center (BSC), Barcelona, Spain

5 Met Office Hadley Centre (MOHC), Exeter, UK

6 College of Engineering Maths and Physical Sciences, University of Exeter, UK

7 Met Office, Exeter, UK

8 Universität Hamburg, Hamburg, Germany

9 Deutscher Wetterdienst (DWD), Hamburg, Germany

10 Group of Meteorology, Universitat de Barcelona (UB), Barcelona, Spain

11 International Pacific Research Center (IPRC), Honolulu, USA

12 Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokohama, Japan

13 Laboratoire de Météorologie Dynamique, Ecole Normale Supérieure, Paris, France

14 University of Toronto, Toronto, Canada

15 Meteorological Research Institute (MRI), Tsukuba, Japan

16 National Centre for Atmospheric Science (NCAS), University of Oxford, Oxford, UK

17 Max-Planck-Institut für Meteorologie (MPI), Hamburg, Germany

Correspondence: Tim Stockdale, ECMWF, Shinfield Park, Reading RG2 9AX, UK

Email: tim.stockdale@ecmwf.int

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/qj.3919

Abstract

A multi-model study is carried out to investigate the ability of models to predict the evolution of the quasi-biennial oscillation (QBO) up to 12 months in advance. All models are initialised from common reanalysis data, and forecasts run for a common set of 30 start dates over 15 years. All models have high skill in predicting the phase evolution of the QBO at 20-30 hPa, with slightly more variable results at higher and lower levels. Other aspects of the predicted QBO are of variable quality, and in some cases are consistently poor. QBO easterlies are too weak in all models at 20-50hPa, while westerlies can be either too strong or too weak. This results in both a reduced amplitude of the QBO and a westerly bias in zonal-mean winds, notably at 30 hPa. At 70 hPa models tend to have reduced QBO amplitude and an easterly bias. Despite these failings, a multi-model ensemble of bias- and variance-corrected forecasts can be used to give accurate and reliable QBO forecasts up to at least a year ahead. Analysis of the zonal momentum budget during the first month of the forecast shows that large scale forcing from Eliassen-Palm flux divergence and vertical advection are handled fairly well by the models, although vertical advection terms tend to be weaker than re-analysis estimates. Total tendencies show common errors, suggesting common failings in gravity wave drag treatments. Teleconnections from the QBO to Northern Hemisphere winter circulation are also examined, and do not appear to be realistic beyond the first month. Analysis of initialised forecasts is a powerful tool for diagnosing the accuracy of model processes driving the QBO.

Key Words: Quasi-Biennial Oscillation, predictability, stratosphere, initialised forecasts, momentum budget, QBO teleconnections

Funding information: See Acknowledgements

1. Introduction

The stratospheric quasi-biennial oscillation (QBO) is the dominant mode of interannual variability in the stratosphere. Questions about the mechanisms that drive it and what is needed to successfully reproduce it have traditionally been studied using idealised and free-running general circulation models, and have given us a good understanding of the essential dynamics of the QBO and how to reproduce it in models – the basic requirements being an adequate representation of the wave forcing, and an adequate vertical resolution (Takahashi, 1996; Scaife et al., 2000; Hamilton et al., 2001; Giorgetta et al., 2002, 2006; Kawatani et al., 2010).

However, there are further questions about the QBO which need more detailed understanding. Observations suggest that the QBO has significant influences on other aspects of stratospheric and tropospheric circulation, notably the stratospheric winter polar vortex (Holton and Tan, 1980; Anstey and Shepherd, 2014). The mechanisms of these influences are not fully understood (Garfinkel et al., 2012; Watson and Gray, 2014), nor are the modelling requirements to reproduce them. This is important because the QBO has predictability out to lead times of several years (Pohlmann et al., 2013, Scaife et al., 2014), making it a potential contributor to seasonal forecast skill, and the QBO phase has also been shown to be important for sub-seasonal prediction (Garfinkel et al., 2018). The QBO has the potential to influence the troposphere directly via interactions with tropical deep convection and the MJO as well as via the stratospheric polar vortex (Butler et al., 2019). Variability in the vortex leads to surface climate effects, particularly in winter, due to interactions with the tropospheric storm tracks. The strongest effects are seen over the North Atlantic and project strongly onto the North Atlantic Oscillation and Arctic Oscillation (Kidston et al., 2015). These effects span a broad range of timescales from monthly (Baldwin and Dunkerton 1999, 2001; Kolstad et al., 2010) to multidecadal periods (Scaife et al., 2005, Omrani et al., 2016). Predicting and accurately representing the QBO and its teleconnections is thus expected to be important for forecasting on a range of timescales. In this study we report for the first time on how well models are able to predict not just the phase and amplitude but also the structure of the QBO, which is likely to be important for the accurate representation of teleconnection mechanisms (Gray et al., 2018).

An additional area of concern is how the QBO might change in the future. Changes in atmospheric composition are leading to significant changes in the stratospheric climate,

which might plausibly lead to changes in the behaviour of the QBO and its teleconnections. A companion paper (Richter et al., 2020) looks at what existing QBO models tell us about possible future changes. However, confidence in projection of future changes is limited by uncertainty about how accurately existing models represent the detailed balance of processes in the present day QBO and its influences on the rest of the climate system.

One method to assess the accuracy of QBO representation in models is to use them to make initialised predictions of the QBO over relatively short periods (e.g. up to a year). Initialised forecasts allow us to make a detailed comparison between models and reality, in a context where we expect to be able to find very close agreement. It is also possible to undertake detailed studies of the momentum budget to try to understand the sources of any discrepancies between modelled and observed behaviour. This is the concept behind the fifth experiment in the Stratosphere-troposphere Processes And their Role in Climate (SPARC) Quasi-Biennial Oscillation initiative (QBOi) protocol (Butchart et al., 2018), the initial results of which are reported in this paper.

This study builds on the few existing studies of actual predictability of the QBO and its extratropical climate impacts (Pohlmann et al., 2013; Scaife et al., 2014; Butler et al., 2016). It uses coordinated experiments with initialised ensemble predictions from eight different general circulation models that simulate the QBO spontaneously; Section 2 briefly summarises the experiments. In Section 3 we analyse forecasts as a function of lead time and show detailed analysis of mean square errors, biases, amplitude and phase errors and the vertical structure of the QBO. Section 4 considers forecast performance in terms of phase space trajectories, and Section 5 considers the achieved skill in terms of estimated predictability limits, looking also at the potential for multi-model forecasts of the QBO.

In Section 6 we present a diagnostic analysis, considering the zonal momentum budget during the first month of initialised QBO forecasts, and what can be learned about inaccuracies in model representation of the processes driving the QBO. In Section 7 we assess teleconnections from the QBO to the northern hemisphere, focussing on the stratosphere, and looking both at the seasonal mean strength of the winter vortex and the statistics of Sudden Stratospheric Warmings. We then finish with a summary of our findings and a consideration of the conclusions which can be drawn.

2. Experiment summary

The initial phase of the SPARC QBOi project proposed five coordinated numerical experiments designed to better understand and improve the representation of the QBO in global models (Butchart et al., 2018). The first four experiments are multi-year un-initialised simulations and results from these are reported elsewhere in this special issue (Bushell et al., 2020; Richter et al., 2020; Holt et al., 2020; Anstey et al., 2020). In contrast, this paper focuses on Experiment 5/5a of the QBOi protocol which requires an ensemble of initialised hindcasts.

QBOi Experiment 5 uses initialised hindcasts of 9-12 months length using observed sea surface temperatures (SSTs) and sea ice conditions. External forcing is specified to correspond to observations (see Butchart et al. 2018 for full details). Thirty start dates from 1 November and 1 May for the years 1993 to 2007 cover fairly well the full range of QBO phases as well as two different points in the stratospheric seasonal cycle. For each hindcast start date, an ensemble of at least three members is specified. As an alternative to Experiment 5, Experiment 5a allows for an appropriately initialised coupled ocean model to be included instead of prescribing observed SSTs. We analyse six models that performed Experiment 5 and two models that performed Experiment 5a. Four of the six models that performed Experiment 5 also performed the un-initialised QBOi experiments (see Butchart et al., 2018 for details) that are analysed elsewhere in this special issue, using the exact same model versions as used here.

Details of the forecast systems that have so far taken part are summarised in Table 1. A brief recap of the resolution of the underpinning models is provided in column 1. It is worth emphasising that all these models are stratosphere resolving. Length of hindcast simulations, number of ensemble members and ensemble generation method are also given, as are details of initialisation method, and whether the hindcasts are atmosphere only (with prescribed SST; Experiment 5) or from fully coupled ocean-atmosphere systems (Experiment 5a).

All models were initialised using ERA-Interim reanalyses (hereafter ERAI), but the exact method by which this was done varied between models. Although this might account for some slight differences in results, we note that those models using relaxation generally used a fast timescale (6 hours or less) for winds, so differences in the initial state are expected to be small. MRI-ESM2 used a lagged method extending 5 days either side of the nominal start

date, so differences in initial conditions are bigger in this case. Overall, though, differences in behaviour are expected to be dominated by differences in the models themselves.

3 Forecast skill of the QBO in initialised models

The realism of the QBO in model forecasts can be measured and described in a variety of ways. We start by looking at the behaviour of zonal-mean zonal winds as a function of height and forecast lead time, providing a local view of the extent to which a realistic QBO is maintained and progressed as model integrations run forward. The simplest measure of the QBO is its phase at a given height. Previous studies of the QBO have generally picked a particular pressure level, and then measured the skill of forecasts in predicting the phase of the QBO at that height, typically using a metric of anomaly correlation of the zonal-mean zonal winds. Although this tells us something about the model's prediction of QBO phase, it does not tell us about other elements affecting the accuracy of the predicted winds, such as bias or the amplitude of the model QBO signal. We thus look first at a measure of the absolute accuracy of the forecasts, the Mean Square Skill Score against climatology (MSSS), and then obtain some physical insight by examining the factors that contribute to forecast accuracy.

3.1 Mean Square Skill Score of zonal-mean zonal wind forecasts

We work with monthly mean data at a set of standard pressure levels (10, 20, 30, 50 and 70 hPa), and calculate zonal-mean zonal winds averaged from 5N-5S. We use ERAI winds for verification, although any modern reanalysis would give very similar results for the pressure levels considered due to the observational constraint provided by assimilation of tropical radiosonde wind data, which extends up to ~10 hPa. For each of the thirty forecast dates, we calculate the difference of the ensemble mean forecast from ERAI (i.e. the error) as a function of lead time. We then calculate the mean square of this set of thirty forecast errors to give a mean square error (MSE) as a function of lead time. We also calculate the corresponding mean square error of a climatological forecast (which is equal to the mean square of the anomalies, in other words the interannual variance of ERAI). Finally, we obtain the Mean Square Skill Score against climatology: $MSSS = 1 - MSE_{\text{model}} / MSE_{\text{clim}}$. MSSS is a normalised function of the root mean square error (RMSE) that represents the fraction of the observed variance correctly predicted by the model forecast. If the amplitude of the

anomalies is correct, MSSS is equal to the square of the anomaly correlation. The MSSS of the individual models for different pressure levels is shown in Figure 1.

In the core regions of the QBO, 20-50 hPa, model forecasts have consistently good skill during the first months of the forecast, with most models having a MSSS above 0.7 after 5 months. At higher and lower levels (10hPa and 70 hPa) there is substantial variation in model skill: the best performing models retain a good level of skill, but some models become mediocre (around 0.5) or in some cases poor (near 0) after just a few months. It is also notable that no single model has the best performance at all levels. For example, the EC-EARTH3.1 model (light blue) is one of the better performers over the first 6 months at 10 hPa but is the worst performer at 30 hPa and 50 hPa; the MIROC-ESM model (red) is the best at 30 hPa, but poor at 70 hPa; the IFS43r1 model (blue) is best at 50 and 70 hPa but has average performance at 20 hPa. Not all models were integrated for 12 months, but those that were suggest that a substantial level of skill (MSSS>0.6) is achievable up to 12 months, not just in the core QBO region but over the full 10-70 hPa range.

3.2 Decomposition of performance – bias, anomaly correlation, amplitude

Although MSSS tells us about the overall accuracy of the forecasts, it does not give much insight into which aspects of the forecast are doing well, and which are sources of error. We therefore continue our analysis by decomposing QBO zonal wind forecasts into bias, and then amplitude and phase of the QBO signal.

Figure 2 shows the ensemble-mean zonal-mean zonal winds at two levels where bias is a significant feature of the model error. Models tend to have a positive bias at 30 hPa, especially in the second half of the year. This is a large error and implies a substantial error in the momentum budget in all models. At 70 hPa, the zonal winds tend to be too strong (i.e. too easterly) at all times of year for most but not all models. If the aim of a forecasting system is simply to predict the QBO then bias is not a first-order problem, since the forecasts can be bias-corrected. However, bias is unwelcome in a general seasonal forecasting system since the dynamics of wave propagation often depend on absolute wind velocity. Moreover, if we are using initialised forecasts to diagnose errors in our models, then the existence of a well-defined bias is important information.

If we remove the mean bias from the forecasts, we can look at the behaviour of the forecasts in terms of the amplitude of the anomalies and the agreement of the temporal variation or

phase, as measured by the anomaly correlation. Figure 3 shows the amplitude of the QBO, calculated after the removal of bias as the standard deviation of the anomalies of the individual ensemble members, relative to the standard deviation of the ERAI anomalies.

Most of the models are underactive – the amplitude of the QBO is too small. At 20 hPa only UMGC2 and MPI-ESM-MR have a realistic level of variability, the others being damped by between 10 and 40%. The two most active models happen to be those with coupling to the ocean, but experience with other models such as IFS43r1 suggest that coupling *per se* has only a very small impact on the QBO, so this may be coincidental. At 50 hPa the models have split into two groups: three models with about the right amplitude, and all the others damped by 40% over the first 3-4 months. At 10hPa (not shown), some models have slightly more variance than observed, but at this height the winds start to have an appreciable amount of non-QBO related variance, so a simple variance ratio cannot be interpreted purely in terms of the QBO. It is possible that the increase in grid spacing with height is causing spurious breaking of waves, leading to excessive momentum deposition and larger amplitude variability. The higher variance at 10 hPa is consistent with the QBO behaviour in the free running models (Bushell et al, 2020), where the amplitude tends to peak at around 10 hPa, higher than the observed peak at 15 hPa.

Finally, we can measure the forecast skill in terms of anomaly correlation. Figure 4 shows the anomaly correlation skill of the bias-corrected forecasts at different levels.
Anomaly correlation is high in all models in the core QBO region, being above 0.9 out to seven months at 30 hPa. Anomaly correlation drops off faster during the first 3 months at 10 hPa and 70 hPa, perhaps associated with a higher level of noise at these levels, but in the better models then stabilises and can remain above 0.8 for the full 12 months. However, at these levels there is again a range of performance between models, although the spread in performance is less dramatic than was seen in MSSS. In almost all cases the correlation skill remains statistically significant: the 5% level according to a 1-tailed test based on a Fisher z-transform for 30 cases is 0.307. The plot axis starts at 0.4, so all plotted values are significant. The ability to maintain good phase skill at 10 hPa over a period of 12 months implies that those models are able to predict the onset of phase reversals at higher levels and are not simply propagating downwards existing wind structures. Indeed, models with good skill at 10 hPa out to 12 months predict well the QBO evolution at 5 hPa, and in some cases (IFS43r1, AGCM3-CMAM) capture the QBO signal at 2 hPa (not shown). Although most models

maintain a good QBO phase prediction in the core region throughout the 9-12 month length of the forecast, two models (MIROC-AGCM-LL and EC-EARTH3.1) suffer substantial drops in performance at 20-30 hPa. Curiously these models (and all models) maintain good phase performance at 50 hPa, despite the phase problems higher up, and despite the substantial loss of amplitude at 50 hPa. It may be that these models have the right fundamental periodicity but are not reproducing the correct timing of events at intermediate levels at longer lead times.

3.3 A composite view of QBO forecast performance

The above analysis looked at the zonal-mean winds in terms of bias and errors in amplitude. Inspection of the full time-series (not shown) makes it clear that there is an inherent ambiguity between mean bias and amplitude error, and a QBO-state dependent bias. Thus we also examine model performance as a function of the QBO state.

We form composites of strong QBO easterly and westerly phases and compare forecast values with corresponding analyses. The composites are based not on the QBO phase at the start of the forecast, but on the observed phase at the target verification time. For each target lead time and level, we choose the 10 most extreme easterly and westerly reanalysis winds from our sample of 30 cases. We then take the forecasts that verify on the selected dates, and form composites of both datasets. The choice of dates is different at different levels: we are looking at the amplitude of the extreme phases per level, not at wind profiles for a given QBO state.

Figure 5 shows the forecast and reanalysis composites for different lead times. The observed composite is identical for lead times separated by 6 months (e.g. T+1 and T+7 months, or T+3 and T+9 months) because (almost) the same set of dates are considered. For the first month, the winds are well-predicted by all models over the full 10 to 70 hPa range. As lead time increases, the models increasingly fail to capture the strength of the easterlies in the 20 to 50 hPa range, while westerlies can be either under- or over-estimated depending on the model. It is the weakness of model easterlies which results in both the generally reduced amplitude of the QBO and a pronounced mid-level westerly bias in all models that was seen in the previous section. The models which appear to have the correct amplitude at 50hPa (UMGC2, IFS43r1, MPI-ESM-MR) still have easterlies which are too weak but have westerlies which are too strong. The biases that develop over the first five months are

remarkably similar to the biases seen in free-running simulations of the QBO, shown in Fig. 5a of Bushell et al., 2020 (this volume).

At 70 hPa, model easterly phases have varying rather than consistent biases, while the amplitude of the westerly phase is underestimated by most models and overestimated by none. Inspection of individual forecast trajectories shows that the substantial model underestimation of the observed westerly extremes at leads of 2-5 months is not because the models are incapable of generating sufficiently strong westerlies; rather it is because the models tend not to sustain strong westerlies for long enough on average. The transition to easterlies occurs either at the right time or too soon, but never too late, thus causing the composite to be biased toward too weak westerlies.

4. A comparison of QBO phase space trajectories

We can also examine the QBO evolution as a function of the initial wind profiles, and we choose to do this in a phase space of zonal winds at two complementary levels. We present the trajectory of pairs of 20- and 50-hPa monthly zonal winds for selected cases, chosen to best characterize the model biases at different stages of the QBO cycle. Figure 6 shows the trajectories of monthly-mean zonal wind starting from the first to the ninth months of forecasts in each model (except MPI-ESM-MR, which is to the seventh month). The trajectory of the ERAI analysis is also presented for nine months (black) and then, to indicate the evolution of the remaining QBO cycle in the phase space, extended further to 24 months (grey). The zonal wind pair on the initial date is indicated by the open circle using ERAI, and that for the third forecast month is indicated by the filled circle on each trajectory. The trajectory of pairs of the 30- and 70-hPa wind is also presented in the same way. The trajectory rotates clockwise with time, following the QBO phase.

Figure 6a shows the evolution of zonal wind for the case initialised on 1 Nov 2001. The initial QBO state for this case is westerly (easterly) at 20 hPa (50 hPa). During the first 5–6 forecast months, the easterly-to-westerly phase transition occurs at 50 hPa in all models, although somewhat too quickly, while the evolution of the 20-hPa westerly shows a spread between models. The 20-hPa westerly in MPI-ESM-MR and UMGC2 is intensified by about 5 m/s following ERAI, while that in AGCM3-CMAM and the MIROC-ESM models is weakened as the forecast time increases. The magnitude of 50-hPa westerly after the phase transition also shows spread among models. In general, the 50-hPa westerly tends to have

large magnitudes in the models in which the 20-hPa westerly magnitude has been large. There are seven other cases with a similar phase evolution of the QBO (i.e., easterly-to-westerly transition at 50 hPa), and the features in the evolution of the case shown in Fig. 6a are also found in these other cases, with similar inter-model differences (not shown). Differences between ensemble members from the same model are generally smaller than the differences between the ensemble means of different models: plotting the full set of individual ensemble members leads to the same conclusions. The spread in strength of the westerly phase at both 20 and 50 hPa seen in the phase trajectory plots corresponds to the spread in strength of the westerlies seen in Figure 5.

Figure 6b presents a case with initial westerly at both 20 and 50 hPa on 1 May 2002. In this case, the westerly-to-easterly phase transition occurs at 20 hPa at about the right speed, while the westerly persists at 50 hPa. However, similarly to the evolution in Figure 6a, the strength of the 50-hPa westerly develops a spread among models over the first five forecast months or so, after which the 50 hPa wind tends to maintain its magnitude in each model. At the same time, the zonal wind at 70 hPa in most models tends to rapidly and incorrectly change direction to easterly within the first five forecast months, except in IFS43r1 and UMGC2 (lower panel). This is consistent with the weak amplitude of the westerly at 70 hPa commonly found in the models (Figure 5). Similar behaviour is also found in later parts of the QBO cycle with 70 hPa westerlies, as illustrated in the lower panel of Figure 6c.

The case initialised on 1 November 1999 (Figure 6c) involves a later stage of the 20-hPa westerly-to-easterly phase transition. All models end the transition at smaller magnitudes of the easterly compared to ERAI. This is related to the common bias in easterlies in the models shown in Figure 5. The bias in easterlies at lower levels can be seen in the case initialised on 1 May 1994 (Figure 6d). In this case, the initial wind is easterly at 20 and 30 hPa and westerly at 70 hPa. In the first 2–3 forecast months, the 50-hPa wind reaches its maximum easterly strength in the models but with smaller magnitudes than ERAI by 5–15 m/s. In the fifth forecast month, the simulated wind at 70 hPa, which has become easterly, stops the transition, whereas ERAI continues to be intensified. Consequently, most models exhibit weaker easterly magnitudes at 70 hPa as well (except MPI-ESM-MR and UMGC2). In addition, large spread of the easterly at 30 hPa among models is also seen in the early period of the forecast, while the 30-hPa easterly becomes weaker in all models during the 70-hPa transition, compared to ERAI.

5. Forecast skill and predictability

For a fuller picture, we examine the forecast skill in the context of model estimates of QBO predictability. In an ideal ensemble forecast system, ensemble members are equally likely and are statistically indistinguishable from reality. The ensemble standard deviation is then equal to the expected distance between the ensemble mean and the observations, i.e. the RMSE of the ensemble mean. The ensemble spread is thus an estimate of the predictability limit for the RMSE, and in a nearly optimal forecast system the two would be close. Most of the forecast uncertainty in the stratosphere is linked to forcing from the troposphere. Errors in the troposphere will grow rapidly regardless of how they are initiated and so we might expect (or at least hope for) a realistic amount of forecast uncertainty regardless of the exact initialisation method used.

Figure 7 shows the forecast skill of individual models in terms of RMSE compared to the unbiased estimator of ensemble spread, at various levels. At 70 hPa (lower right), the ensemble spread grows quickly for the first two months in all models, but then generally stabilises within a range of 1-1.5 m/s, growing only slowly. At this level the QBO amplitude is weak (both in reanalysis and even more so in most of the models), and it seems that unpredictable noise originating in the troposphere and unrelated to the QBO injects uncertainty into the lower stratospheric winds quite quickly. However, uncertainty in the QBO phase, which develops slowly over time, adds only a limited amount of additional uncertainty over a 12-month period, even in those models which have a realistic QBO amplitude at this level. In terms of RMSE, the models exhibit a very large range, consistent with what has been seen in other scores. The best performing model at this level (IFS43r1) has errors which, while substantially worse than the predictability limit estimated from its corresponding ensemble spread, are not dramatically so.

At 20 hPa and 30 hPa, the gap between the predictability limit and the actual model performance is clearer, with errors very much larger than the spread. At these levels there is less diversity in model RMSE scores, certainly over the first 6 or 7 months, and compared to the estimated predictability limit we can say that all models are poor. These are levels where anomaly correlation is high, so it might seem strange to characterise the forecast performance as poor. However, the amplitude of the QBO is very large, and small timing errors in a phase transition will produce large wind errors. The very small spread in all models indicates that the timing of the QBO evolution at these core levels during the first six months of a forecast

should be extremely predictable, with little influence from sources of noise such as tropospheric weather. It is possible that the models underrepresent sources of noise such as variations in tropical wave forcing or (as in the disruption event of 2016) occasional mid-latitude influences. As also noted by Scaife et al. (2014), we thus cannot be certain that the implied exceptionally high predictability is correct. Nonetheless, this is the estimate given by all the models in this study.

At 10 hPa RMSE varies substantially between models – from 7.5 to 25 m/s at 7 months – and the gap between the RMSE of the best performing models and the spread is less, in relative terms, than the gap seen at 20 and 30 hPa. The better agreement between spread and error at 10 hPa is more due to an increase in spread than a reduction in error. Why the model winds show increased ensemble spread at this higher level is unclear.

We note that at all levels, the MRI-ESM2 model has larger spread in the first month than other models. This is due to the lagged average method used to construct the MRI-ESM2 ensemble, which introduces a large spread in the initial conditions. At 10 hPa the extra spread persists into the second month, but otherwise it is rapidly lost as the forecasts progress. This is consistent with our assumption that forecast spread is driven more by noise in the tropospheric forcing than by growth of initial uncertainties in the stratospheric state itself.

For a final reflection on forecast skill versus predictability, we look at what can be achieved from a multi-model approach. Multi-model forecasts, where forecasts from different models are averaged together to reduce the impact of model-specific errors, have been shown to increase reliability and skill of seasonal forecasts (Palmer et al., 2004, Weisheimer et al., 2009). A multi-model QBO forecast has been constructed from the four models with the longest lead-times (AGCM3-CMAM, IFS43r1, EC-EARTH3.1 and UMGC2). Each model is bias-corrected and then variance-corrected (i.e. the forecast anomalies are scaled such that their variance matches that observed), using cross-validation so that the correction for each year uses data only from other years. Variance correction has been shown to be very helpful in ENSO forecasting (Stockdale, 2013), and given the amplitude problems that many models have with the QBO is also appropriate here. The corrected forecasts from the models are then averaged with an equal weighting to form the multi-model ensemble mean forecast. We do not formally estimate the forecast uncertainty, but we take the multi-model ensemble spread (calculated from the corrected individual forecasts, with the ensemble size of UMGC2

limited to 11 to weight it comparably to other models) as indicative of the forecast uncertainty.

Figure 8 compares the multi-model skill (in red) with that of the best performing single model (in blue) in terms of RMSE and anomaly correlation. We have chosen two levels that are representative of the range of behaviour seen. At 30 hPa (left), the best single model is AGCM3-CMAM, and we see as before that despite a fairly high level of correlation skill (at or above 0.8 for 12 months), the RMSEs for a single model are much larger than the ensemble spread. However, the multi-model forecast gives a large reduction in RMSE and improves the anomaly correlation, particularly in the 6-8 months range. This demonstrates that the large errors seen in Figure 7 at 20 and 30 hPa are in large part due to model error and can be reduced by a multi-model approach. The relationship between the multi-model error and the predictability limit estimated from single model spread is now roughly comparable to that seen at 10 hPa and 70 hPa, for the first seven months. This supports the idea that the QBO is highly predictable in its core region. The drop in skill after 7 months seen in both RMSE and anomaly correlation may be due to an insufficient number and quality of models (noting especially the performance of EC-EARTH3.1 anomaly correlation at these lead times) rather than indicating a fundamental predictability problem.

At 50 hPa (right) the situation is different. The best single model at this level (IFS43r1) already has a very high level of skill and RMSEs that are only moderately worse than the estimated predictability limit. Here, the multi-model forecast matches the best-performing model but does not exceed it in terms of skill of the ensemble mean. Nonetheless, the multi-model approach does allow a much better estimate of forecast uncertainty than a single model, which is important for forecast reliability. Adjusting RMSE scores to compensate for the different ensemble sizes (not shown) confirms for both levels that the benefit of the MM comes from the diversity of models, not the increased ensemble size.

If we wanted to produce operational forecasts of the QBO, these results suggest that a multi-model system, using appropriate calibrations, would be able to produce highly accurate and reliable QBO forecasts out to at least 12 months.

6. Momentum budget

The large-scale atmospheric state in all models is close to that of ERAI during the first month of each forecast. We can therefore meaningfully compare the momentum budget between

models during the first month, and also compare models against reanalysis for those terms which can be derived from the reanalysis. The zonal-wind tendency in the transformed Eulerian mean (TEM) equation is expressed as

$$\frac{\partial \overline{u}}{\partial t} = \overline{v}^* \hat{f} - \overline{w}^* \frac{\partial \overline{u}}{\partial z} + \frac{\nabla \cdot \boldsymbol{F}}{\rho_0 \, a \cos\phi} + \overline{X}_{\mathrm{GW}} + R$$

following the notation by Andrews et al. (1987), where $\overline{v}^*$, $\overline{w}^*$, and $\boldsymbol{F}$ are residual-mean meridional and vertical velocities, and Eliassen–Palm (EP) flux, respectively, defined in the TEM equation (see Andrews et al., 1987, Eqs. 3.5.1 and 3.5.3, for their definitions), and $\hat{f} \equiv f - (a\cos\phi)^{-1}\partial(\overline{u}\cos\phi)/\partial\phi$. The second to fifth terms in the righthand side are vertical advection, EP flux divergence, parameterized gravity wave drag (GWD), and the residual of the equation, respectively. Figure 9 presents these terms, along with the zonal-wind tendency, averaged over 5N-5S in each model for a number of cases (the first term in the righthand side, the sum of Coriolis force and meridional advection, is negligibly small and thus not shown). The corresponding terms from ERAI are also presented, except for the GWD which is not available.The residual in ERAI therefore includes the effect of unresolved waves (GWD), and might be expected to be dominated by them most of the time. Although one of our cases is an exception, as discussed below, we thus choose to plot the ERAI residual in the same panel as the GWD from the models. Note that estimates of the residual vertical velocity differ between different re-analyses (*e.g.* Abalos et al., 2015), so the ERAI estimate of the vertical advection term has some uncertainty, which in turn may affect the estimate of the residual. Note also that MIROC-AGCM-LL does not include a GWD parametrization.

The first case, initialised on 1 November 1999 (the same case as in Figure 6c), involves upper-level easterlies and an easterly shear layer at 10–40 hPa (Figure 9, upper, wind profile shown with a dashed line in all panels). The models all capture the total tendency of increasing easterlies in the 10-30 hPa layer but are all missing the slight westerly tendencies between 30 and 70 hPa (top left panel). We first consider the momentum budget in the region of upper-level easterlies and will return to the low-level westerly jet later. The models generally produce EP flux divergence forcing very similar to that in ERAI up to 20 hPa, but then diverge, with a majority of models giving stronger easterly forcing than reanalysis. One model is different to the others – MIROC-AGCM-LL, which has no GWD parametrization, has a large easterly forcing from the EP forcing. This appears to largely compensate for the

lack of GWD forcing, such that the total tendency is similar to other models. It is not immediately obvious why the difference in terms is so well balanced. Forcing by vertical advection, which is in the opposite direction to that by the EP flux divergence, is again rather well produced by the models at lower levels, but with some spread in values at 20hPa and above, with models tending to have weaker (less westerly) values than ERAI. The sum of advective and EP flux forcing results in westerly forcing in ERAI and MRI-ESM2 and easterly or near-zero forcing in the others at around 15 hPa. It is seen that the residual in ERAI is comparable to the models' GWD up to 20 hPa (8–11 m/s/month). However, model GWD is reduced above this level and becomes smaller than the residual in ERAI at 15 hPa by 4–9 m/s/month. This may well be responsible for the difference in the easterly tendency below the easterly maximum (10-15 hPa) between the models and reanalysis. Indeed, the vertical structures of GWD in the models and residual in ERAI in the easterly shear layer resemble those of wind tendency in the models and ERAI, respectively (Figure 9).

A qualitatively similar result is found also in the case initialised on 1 May 1994 (Figure 9, middle row, see also Figure 6d) when the easterly shear is shifted downward compared to the first case. In this case, the differences in the wind tendency between ERAI and the models is large at 50 hPa, as well as the difference between the residual in ERAI and models' GWD. The smaller easterly tendency in the models is associated with a lack of easterly forcing, and this is presumably what leads to the weak easterly magnitudes during the phase progression shown in Figures 6d. The magnitude of the residual obtained in the models is relatively small in the easterly shear layer in these cases.

The third case, initialised on 1 November 1996, involves westerly and easterly jets above and below 20 hPa, respectively (Figure 9, lower). The EP flux divergence produced in the models shows a large spread above 30 hPa. The large spread of this forcing is also seen in other cases of westerly shear phase (not shown), implying that the representation of Kelvin waves and resolved-scale gravity waves differs significantly between models. The westerly GWD in the westerly shear layer also shows a difference in its magnitude between the models, which may come from the differences in the parametrization scheme used or in the simulated state of the troposphere where the parameterised gravity waves are partly filtered out. The westerly shear of the QBO is large in magnitude and, furthermore, in the deep tropics mean vertical motion often changes to be downward in the westerly shear. Vertical advection in this shear layer is sensitive to any errors in mean vertical motion, which might possibly explain the opposite

sign of advection around 20 hPa shown by MRI-ESM2. The other models capture the vertical structure of the advection term as represented by ERAI, although tending to underestimate the magnitude. In contrast to our first two examples of easterly shear phases, the residual is very large in the westerly shear layer in some models, a feature seen in other cases with westerly shears (not shown). The residual tends to be positive just above the easterly jet core and negative below the westerly core as seen in Figure 9, which resembles the vertical structure that might be expected from vertical diffusion of momentum. The residual in ERAI shows a similar vertical structure in the lower stratosphere, indicating that forcing with such a structure has been required in ERAI during its assimilation in order to approach the observed state. The model results suggest that this forcing was not from GWD but from other processes. The reanalysis is consistent with the model results, but cannot separate out the unresolved processes contributing to the momentum budget to confirm their sources.

Finally, we return to the low-level westerly jet in our first case, initialised on 1 November 1999 (Figure 9, upper). The low-level jet is located at 40–50 hPa, and as already seen, the EP flux divergence and vertical advection at this level are in good agreement between ERAI and the models. ERAI exhibits a residual of about 2.5 m/s/month, which is responsible for the positive tendency in the westerly jet core. However, the models do not provide enough westerly forcing to reproduce the acceleration of the low-level westerly. As a result, the zonal wind in the lowermost stratosphere in the models evolves with smaller amplitudes as shown in Figure 6c. The lack of the low-level westerly forcing in the models might benefit from further study, since many models currently have a bias in the low-level westerly jet of the QBO, both in initialised forecasts (Figure 5) and in free-running simulations of the QBO (Bushell et al., 2020, Fig. 5a).

7. Teleconnections from the QBO to the northern hemisphere

7.1 Holton-Tan relationship

We look at the connection between the QBO and the Northern Hemisphere winter stratospheric vortex, examining the relationship as a function of both QBO phase and calendar month. Figure 10 shows the vortex response to the QBO for the ERAI reanalysis (1979-2016) and the model hindcasts initialised in November (1993-2007). The full ERAI period is used, rather than just the period corresponding to the QBOi hindcasts, to provide a more robust estimate of the observed response. The top panel in each sub-figure indicates with coloured boxes the vortex anomaly (defined by 10 hPa 60N deseasonalised monthly-

mean zonal-mean zonal wind) as a function of QBO phase and time in the season. The QBO phase is defined by a phase angle determined from the two leading empirical orthogonal functions (EOFs) of 10-70 hPa equatorial wind (Wallace et al., 1993), with the range [0,1] representing a full QBO cycle, and a single value of QBO phase is assigned to each individual hindcast based on the value of its phase angle in November. Each column in the figure thus represents the ensemble mean over a coherent group of winters, with the number of winters shown in green text at the bottom of the column. The group of winters belonging to each column are those with a QBO phase angle falling within a bin of width 0.25 centred on the phase angle indicated on the x-axis. (Since 0.25 exceeds the 0.1 resolution of the x-axis, there is some overlap in group membership between neighbouring columns.) Since the QBO phase angle is a slightly abstract representation of QBO phase, the bottom two panels in each sub-figure illustrate the vertical structure of QBO winds associated with each value of the phase angle. The middle panel shows equatorial zonal-mean zonal wind profiles composited (irrespective of lead time or calendar month) for times when the phase angle takes the values on the x-axis, and provides both a reference for the QBO structure as a function of phase and a (qualitative only) reminder that the model QBO structures diverge from the observed structure during the forecasts. The bottom panel shows the wind profile in November for each of the corresponding individual reanalysis or hindcast dates.

The ERAI response (Figure 10a) shows the expected strengthening of the vortex during 50 hPa QBO westerlies, and vice versa during 50 hPa QBO easterlies, well known from previous studies (e.g. Holton and Tan 1980). High statistical significance is not apparent due to the short record considered here, but the response is indicative of that which appears in a longer record (Anstey et al., 2020), although the apparent late-winter response of opposite sign to that of early-to-midwinter may not be robust. Although the strongest anomalies tend to appear in midwinter (December and January), the early-winter response in November is appreciable compared to the smaller climatological level of variability occurring at that time in the seasonal cycle.

It is common for models to miss the November response (Anstey et al., 2020), and if the early-winter response plays a role in conditioning the vortex for its midwinter response, this error could plausibly affect the models' midwinter responses. The models shown in Figure 10b-i all show a November response due to the initialisation. In most models this response extends into December, consistent with a typical vortex decorrelation timescale (Gerber et al.,

2010). Beyond December there is no clear evidence in the models for a midwinter response that strongly resembles the observed one, although many models show a midwinter response of some sort. Both EC-EARTH3.1 and IFS43r1 show a fairly strong January-February response, although it is shifted in QBO phase with respect to the observations. In other models the midwinter response is weak or negligible (MIROC-AGCM-LL, MPI-ESM-MR) while in yet others it is apparent only for a small range of QBO phases (AGCM3-CMAM, UMGC2).

Note that the effective sample size differs between models, and in some cases is only 45 years, barely larger than the ERAI sample size. Given the large interannual variability of the polar vortex, samples of this size may be inadequate to robustly characterise the response in a given model. However other models approach or exceed 100 years of data, which is likely to provide a good characterization of the response (Anstey et al., 2020). It should also be noted that sampling of QBO phase by the November hindcast initialisation dates is not uniform, which is apparent as a gap in the vertical profiles at a QBO phase of roughly 0.5 shown in the bottom panel of each sub-figure. The location of the gap is unfortunate since it coincides with a strong response diagnosed from the ERAI 1979-2016 period, and some model results are also suggestive of a strong response to this QBO phase (e.g. UMGC2). The phase angle of about 0.5 corresponds to a deep westerly layer of tropical winds, which has been proposed to favour a quiescent vortex (e.g. Dunkerton et al., 1988; Gray 2003).

For the models analysed here, Figure 10 indicates that the QBO teleconnection to the polar vortex is mostly weaker than observed, certainly beyond the first month of integration. Even in the cases of the strongest response (IFS43r1, EC-EARTH3.1) it is not clear that the sign of the response to a given QBO phase is consistent with that observed – rather, the plots suggest that the model vortex response may correlate with winds at around the 20 hPa level, rather than the 50 hPa level in the observed correlation. Results from the free-running QBOi models also showed an apparent phase shift in the response of many models (Anstey et al., 2020), although the phase error here is perhaps worse. It is a little disappointing that the initialised models are not more successful in capturing the structure of the teleconnection, given that they are starting from a realistic state. One might hypothesize that the phase error is related to either the rapid development of incorrect structures in the QBO, to a lack of balance in the initial state of the models, or to errors in the model affecting the teleconnection process. The

present experiments are not sufficient to investigate this further in any detail. The large sampling variability means that the observed response is also not known with great certainty.

7.2 Sudden Stratospheric Warmings

The Northern Hemisphere winter stratospheric vortex exhibits variability at multiple timescales. The most impressive variations are Sudden Stratospheric Warmings (SSWs), where the vortex either splits in two or is displaced from the pole, leading to a dramatic fall in vortex strength and a corresponding dramatic rise in temperature. SSWs have been shown to act as precursors to changes in tropospheric circulation (Baldwin and Dunkerton, 2001) and long-lasting negative NAO regimes (Ayarzagüena et al., 2018). Although the onset of an individual SSW is not predictable more than two weeks in advance (e.g. Tripathi et al., 2016), the statistics of SSW occurrence and timing have been found to depend on the phase of the QBO (e.g. Gray et al., 2004). To complement the above analysis of the monthly-mean polar vortex strength, we consider the statistics of SSWs for the subset of models for which daily data are available, and compare their relation to the QBO to that seen in reanalysis data. Can the models reproduce the observed influence of the QBO on the statistics of SSWs?

Figure 11 shows the frequency of SSWs for easterly (QBOE) and westerly (QBOW) phases, classified according to the phase at 50 hPa in November. SSWs are defined here as zonal-mean zonal wind reversals at 10 hPa and at any latitude from 55 to 70°N from November to March as in Palmeiro et al. (2015) – note that this definition gives a higher climatological number of SSWs than some other definitions (e.g.Charlton and Polvani, 2007; see also Butler et al., 2015). Those events in which the winter circulation does not recover for at least 10 days before the 10th of April are classified as final warmings and discarded from the counting. Contrary to previous findings (e.g., van Loon and Labitzke, 1987; Gray et al., 2004), this analysis does not show a clear effect of the QBO phase on SSW frequency when using reanalysis data from 1979-2016, with a weak but non-significant enhancement of frequency in QBOE phases compared to QBOW. Most models agree that there is not much difference between QBOE and QBOW phases, the exception being MIROC-AGCM-LL which shows an apparent enhancement of SSWs during QBOE, although this is relative to a very low frequency during QBOW for this model.

Figure 12 shows the QBO impact on the timing of SSW onsets. Since the hypothesis is that QBOE favours SSWs to occur sooner (Gray et al., 2004), only the first SSW is considered

when more than one SSW occurs in the same winter. ERAI shows a significant advance of the SSW mean date of occurrence by more than a month. Most models fail to capture this, the exceptions being MIROC-AGCM-LL and EC-EARTH3.1. However, the effect disappears in the latter for the May initialisation, i.e. at longer lead times. This might be due to deterioration of the QBO structure or growth of errors in the model mean state, or it might be that the November initialisations benefit from predictability effects. Interestingly, it is the model with the lowest SSW frequency, MIROC-AGCM-LL, which shows the largest impact of the QBO phase in relation to SSW occurrence in both timing and total frequency. Overall, none of the models reproduce both accurate SSW statistics and the observed relationships to the QBO.

8. Summary and conclusions

We have examined the capability of a number of QBOi participating models to predict the evolution of the QBO and its teleconnections to the northern winter polar vortex. The models represent the state-of-the-art of QBO and stratospheric modelling, yet produce a mixture of successes and failures when assessed rigorously in this predictive challenge.

All models were successful in predicting the phase evolution of the QBO at 30 hPa over the first 6 months of the forecasts, with correlations at or above 0.9 at this range. Most but not all models managed a good level of skill (>0.8) out to 9-12 months for levels from 20 hPa to 50 hPa. At higher and lower levels (10 hPa and 70 hPa) there was more variety, with some models performing poorly, although some models managed good skill at these levels too.

Models were generally less successful at maintaining an adequate amplitude of the QBO, and all models proved incapable of developing sufficiently strong easterlies. This is a major common failure, and results in a westerly bias in the climatological winds at 30 hPa for all models. The amplitude of westerlies varied more between models, although at the lowest level (70 hPa) all models had a tendency to be unable to retain westerlies sufficiently long. These biases are very similar to the biases seen in free-running integrations of the QBO (Bushell et al, 2020, this issue).

The use of ensembles enabled estimates of predictability, providing a benchmark against which to assess forecast performance. Although models perform well at 30 hPa in terms of correlation, all models have an RMSE which is much larger than the estimated predictability limit. Multi-model averaging enables a large reduction in error at this height, bringing

forecast errors closer to (although still noticeably above) the estimated limit. At higher and lower levels, although model skill is typically less, individual ensembles are noisier, indicating that the potential skill is also reduced

Although they seem reasonable and are supported in general terms by the actual level of skill achieved by the best-performing model or multi-model combination, it is not certain how reliable the model estimates of predictability are. We note that the 15-year period in our study did not include 2016, when a rare disruption broke the quasi-regular cycling of the oscillation (Osprey et al., 2016, Newman et al., 2016). This disruption appears to have been unpredictable more than a few weeks ahead (Osprey et al., 2016; Watanabe et al., 2018), perhaps because of its mechanism, which originated from outside the tropics. Predictability is generally thought to be much lower in the extratropics, and the packets of equatorward propagating Rossby waves which fluxed easterly momentum close to the equator during the 2016 disruption (Osprey et al., 2016; Coy et al., 2017; Barton and McCormack 2017; Watanabe et al., 2018) may therefore be unpredictable at longer ranges. Nevertheless, the 2016 disruption of the QBO is a rare event (unprecedented in the previous 50-year record) and does not invalidate the high level of predictability usually seen for the QBO.

Momentum budgets show that although models correctly capture much of the large-scale forcing and vertical advection, there are also systematic errors in the parametrised processes when compared to residuals from the re-analysis. Some of the failings in the forecasts can be related to apparent errors in the (largely parametrised) gravity wave drag forcing, although in some situations vertical diffusion might also play a role in the momentum budget, complicating the comparisons between model terms and re-analysis residuals. Not all models provided the data needed for the momentum budget analysis, but the results here suggest that careful examination of the momentum budget over the first month of a forecast, when the model and observed wind profiles match well, will enable diagnosis of errors in the model processes. Initialised forecasts are particularly well suited for such studies.

Relationships between the QBO and the Northern Hemisphere winter polar vortex are present in the first month of the initialised forecasts of all models, but then rapidly diverge. Some models produce moderately strong teleconnections but not matching the observed phase relationships. Other models show disparate responses or else very little at all beyond the

second month. There is also diversity in the model SSW statistics, with no model capturing the reanalysis-estimated relationships of QBO to both number and timing of SSWs.

The lack of agreement concerning interactions with the polar vortex between observations and the initialised models is disappointing. There is a range of possible mechanisms by which the QBO can influence the high latitudes (e.g. Lu et al. 2014; Gray et al. 2018), and it may be that even more accurate modelling of both the structure of the QBO and the rest of the stratosphere is required. Alternatively or additionally, interactions with the troposphere or other processes affecting the stratosphere may be inadequately represented. It will require more work, including appropriately designed experimentation, to better understand the range of behaviour seen in the models.

Acknowledgements

HP was supported by the German Federal Ministry for Education and Research (BMBF) project MiKlip (FKZ 01LP1519A)

YK was supported by Japan Society for Promotion of Science (JPSP) KAKENHI Grant Numbers JP15KK0178, JP17K18816 and JP18H01286. YK and KH were supported by the Japan Agency for Marine-Earth Science and Technology (JAMSTEC) through its sponsorship of research at the International Pacific Research Center.

SW and YK were partly supported by a Japan Science and Technology Agency (JST) as part of the Belmont Forum, and by the "Integrated Research Program for Advancing Climate Models (TOUGOU program)" from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. The Earth Simulator was used for MIROC-ESM and MIROC-AGCM-LL simulations.

FMP and JG-S have been funded by the Spanish DANAE project (CGL2015-68342-R) and 'Ramón y Cajal' programme (RYC-2016-21181), respectively, and supported by Red Española de Supercomputación (RES projects AECT-2017-3-0015 and AECT-2018-2-0023).

A.S. and M.A. were supported by the Hadley Climate Centre Program.

References

Abalos, M., Legras, B., Ploeger, F. and Randel, W. J., 2015: Evaluating the advective Brewer-Dobson circulation in three reanalyses for the period 1979–2012. *J. Geophys. Res.*, **120**, 7534–7554. doi:10.1002/2015JD023182, 2015.

Andrews, D.G., Holton, J.R. and Leovy, C.B. (1987) *Middle Atmosphere Dynamics*. San Diego, CA: Academic Press.

Anstey, J. A. and Shepherd, T. G. (2014) High-latitude influence of the quasi-biennial oscillation. *Quart. J. Roy. Meteor. Soc.* **140** (678). pp. 1-21. ISSN 1477-870X doi: 10.1002/qj.2132

Anstey, J. A., Scinocca, J. F. and Keller, M. (2016) Simulating the QBO in at Atmospheric General Circulation Model: Sensitivity to Resolved and Parameterized Forcing. *J. Atmos. Sci.*, **73**, 1649–1665.

Anstey, J. A., Simpson, I. R., Richter, J. H., Naoe, H., Taguchi, M., Serva, F., Gray, L. J., Bellprat, O., Braesicke, P., Bushell, A. C., Butchart, N., Cagnazzo, C., Chen, C.-C., Chun, H.-Y., Garcia, R. R., Hamilton, K., Holt, L., Kawatani, Y., Kerzenmacher, T., Kim, Y.-H., Lott, F., McLandress, C., Osprey, S., Scinocca, J., Versick, S., Watanabe, S., Yoshida, K., and Yukimoto, S., (2020) Teleconnections of the quasi-biennial oscillation in a multi-model ensemble of QBO-resolving models. *Quart. J. Roy. Meteor. Soc.*, this issue.

Ayarzagüena, B., Barriopedro, D., Garrido-Perez, J. M., Abalos, M., de la Cámara, A., García-Herrera, R., et al. (2018). Stratospheric connection to the abrupt end of the 2016/2017 Iberian drought. *Geophysical Research Letters*, **45**, 12,639– 12,646. doi: 10.1029/2018GL079802

Baldwin, M. P., & Dunkerton, T. J., 1999: Propagation of the Arctic Oscillation from the stratosphere to the troposphere. *J. Geophys. Res. Atmospheres*, **104**(D24), 30937–30946. https://doi.org/10.1029/1999JD900445

Baldwin, M. P., & Dunkerton, T. J., 2001: Stratospheric harbingers of anomalous weather regimes. *Science*, **294**, 581–584.

Barton, C. A., and J. P. McCormack, 2017: Origin of the 2016 QBO disruption and its relationship to extreme El Niño events. *Geophys. Res. Lett.*, **44**, 11,150–11,157. doi: 10.1002/2017GL075576.

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. *Quart. J. Roy. Meteor. Soc.*, this issue.

Butchart, N., Anstey, J. A., Hamilton, K., Osprey, S., McLandress, C., Bushell, A. C., Kawatani, Y., Kim, Y.-H., Lott, F., Scinocca, J., Stockdale, T. N., Andrews, M., Bellprat, O., Braesicke, P., Cagnazzo, C., Chen, C.-C., Chun, H.-Y., Dobrynin, M., Garcia, R. R., Garcia-Serrano, J., Gray, L. J., Holt, L., Kerzenmacher, T., Naoe, H., Pohlmann, H., Richter, J. H., Scaife, A. A., Schenzinger, V., Serva, F., Versick, S., Watanabe, S., Yoshida, K., and Yukimoto, S. (2018) Overview of experiment design and comparison of models participating in phase 1 of the SPARC Quasi-Biennial Oscillation initiative (QBOi), *Geosci. Model Dev.*, 11, 1009-1032. doi: 10.5194/gmd-11-1009-2018

Butler, A. H., Seidel, D. J., Hardiman, S. C., Butchart, N., Birner, T., and Match, A., 2015: Defining sudden stratospheric warmings. *Bulletin of the American Meteorological Society*, **96**, 1913-1928.

Butler, A. H. et al., 2016: The Climate-system Historical Forecast Project: do stratosphere-resolving models make better seasonal climate predictions in boreal winter? *Q. J. R. Meteorol. Soc.* **142**: 1413–1427, DOI:10.1002/qj.2743

Butler, A., Andrew Charlton-Perez, Daniela I.V. Domeisen, Chaim Garfinkel, Edwin P. Gerber, Peter Hitchcock, Alexey Yu. Karpechko, Amanda C. Maycock, Michael Sigmond, Isla Simpson, Seok-Woo Son, 2019: Sub-seasonal Predictability and the Stratosphere, in *Sub-Seasonal to Seasonal Prediction*, Eds. Andrew W. Robertson, Frédéric Vitart. Elsevier, Pages 223-241, ISBN 9780128117149. doi:10.1016/B978-0-12-811714-9.00011-5.

Charlton, A. J., and Polvani, L. M. (2007). A new look at stratospheric sudden warmings. Part I: Climatology and modeling benchmarks. *J. Climate*, **20**, 449-469.

Coy, L., P.A. Newman, S. Pawson, and L.R. Lait, 2017: Dynamics of the Disrupted 2015/16 Quasi-Biennial Oscillation. *J. Climate*, **30**, 5661–5674, doi: 10.1175/JCLI-D-16-0663.1

Dobrynin, M., Domeisen, D. I. V., Müller, W. A., Bell, L., Brune, S., Bunzel, F., et al. (2018). Improved teleconnection-based dynamical seasonal predictions of boreal winter. *Geophys. Res. Lett.*, 45, 3605– 3614. doi: 10.1002/2018GL077209

Dunkerton T.J., Delisi D.P. and Baldwin M.P. (1988) Distribution of major stratospheric warmings in relation to the quasi-biennial oscillation, *Geophys. Res. Lett.*, **15**(2), 136-139.

Dunstone, N., D. Smith, A. Scaife, L. Hermanson, R. Eade, N. Robinson, M. Andrews, and J. Knight, 2016: Skilful predictions of the winter North Atlantic Oscillation one year ahead. *Nat. Geosci.*, **9**, 809–814, doi: 10.1038/ngeo2824.

ECMWF, 2016: IFS documentation CY43R1, published by ECMWF, available at: https://www.ecmwf.int/en/publications/ifs-documentation .

Garfinkel, C. I., Shaw, T. A., Hartmann, D. L., & Waugh, D. W., 2012: Does the Holton-Tan mechanism explain how the quasi-biennial oscillation modulates the arctic polar vortex? *J. Atmos. Sci.*, **69**, 1713–1733. doi:10.1175/JAS-D-11-0209.1

Garfinkel, C. I., Schwartz, C., Domeisen, D. I. V., Son, S.-W., Butler, A. H., & White, I. P. (2018). Extratropical atmospheric predictability from the quasi-biennial oscillation in subseasonal forecast models. *J. of Geophysical Research: Atmospheres*, 123, 7855– 7866. https://doi.org/10.1029/2018JD028724

Giorgetta, M. A., Manzini, E., and Roeckner, E., 2002: Forcing of the quasi-biennial oscillation from a broad spectrum of atmospheric waves, *Geophys. Res. Lett.*, **29**(8), doi:10.1029/2001GL014756

Giorgetta, M. A., E. Manzini, E. Roechner, M. Esch, and L. Bengtsson, 2006: Climatology and forcing of the quasi-biennial oscillation in the MAECHAM5 model. *J. Climate*, **19**, 3882–3901.

Gerber, E. P., et al. (2010), Stratosphere-troposphere coupling and annular mode variability in chemistry-climate models, *J. Geophys. Res.*, 115, D00M06, doi:10.1029/2009JD013770.

Gray, L. J. (2003), The influence of the equatorial upper stratosphere on stratospheric sudden warmings, *Geophys. Res. Lett.*, **30**, 1166, doi:10.1029/2002GL016430

Gray, L. J., Crooks, S., Pascoe, C., Sparrow, S., and Palmer, M. (2004). Solar and QBO influences on the timing of stratospheric sudden warmings. *J. Atmos. Sci.*, **61**(23), 2777-2796.

Gray, L. J., Anstey, J. A., Kawatani, Y., Lu, H., Osprey, S., and Schenzinger, V., 2018: Surface impacts of the Quasi Biennial Oscillation, *Atmos. Chem. Phys.*, **18**, 8227-8247. doi: 10.5194/acp-18-8227-2018

Hamilton, K., R.J. Wilson, and R.S. Hemler, 2001: Spontaneous Stratospheric QBO-like Oscillations Simulated by the GFDL SKYHI General Circulation Model. *J. Atmos. Sci.*, **58**, 3271–3292. doi: 10.1175/1520-0469(2001)058<3271:SSQLOS>2.0.CO;2

Holt, L., Lott, F., Garcia, R. R., Kiladis, G. N., Anstey, J., Braesicke, P., Bushell, A. C., Butchart, N., Cagnazzo, C., Chen, C., Chun, H., Kawatani, Y., Kerzenmacher, T., Kim, Y., McLandress, C., Naoe, H., Osprey, S. M., Richter, J. H., Scaife, A. A., Scinocca, J., Serva, F., Versick, S., Watanabe, S., Yoshida, K. and Yukimoto, S. (2020) An evaluation of tropical waves and wave forcing of the QBO in the QBOi models. *Q. J. R. Meteorol. Soc.*, this issue.

Holton, J. R., and H. C. Tan, 1980: The influence of the equatorial quasi-biennial oscillation on the global circulation at 50 mb, *J. Atmos. Sci.*, **37**, 2200–2208.

Kawatani, Y., K. Sato, T. J. Dunkerton, S. Watanabe, S. Miyahara and M. Takahashi, 2010: The roles of equatorial trapped waves and internal inertia-gravity waves in driving the quasi biennial oscillation. Part I: Zonal mean wave forcing. *J. Atmos. Sci.*, **67**, 963-980. doi:10.1175/2009JAS3222.1

Kawatani, Y., Hamilton, K., and Watanabe, S., 2011: The quasi-biennial oscillation in a double CO2 climate, *J. Atmos. Sci.*, **68**, 265–283, doi:10.1175/2010JAS3623.1

Kidston J, Scaife AA, Hardiman SC, Mitchell DM, Butchart N, Baldwin MP, Gray LJ. 2015. Stratospheric influence on tropospheric jet streams, storm tracks and surface weather. *Nat. Geosci*. **8**: 433– 440, doi: 10.1038/ngeo2424

Kolstad, E. W., Breiteig, T. and Scaife, A. A. (2010), The association between stratospheric weak polar vortex events and cold air outbreaks in the Northern Hemisphere. *Q.J.R. Meteorol. Soc.*, **136**: 886-893. doi:10.1002/qj.620

Lu, H., T. J. Bracegirdle, T. Phillips, A. Bushell, and L. Gray (2014), Mechanisms for the Holton-Tan relationship and its decadal variation, *J. Geophys. Res. Atmos.*, **119**, 2811–2830, doi:10.1002/2013JD021352.

Newman, P. A., Coy, L., Pawson, S., and Lait, L. R. (2016), The anomalous change in the QBO in 2015–2016, *Geophys. Res. Lett.*, **43**, 8791– 8797, doi:10.1002/2016GL070373

Omrani, N., Bader, J., Keenlyside, N.S. *et al.* (2016) Troposphere–stratosphere response to large-scale North Atlantic Ocean variability in an atmosphere/ocean coupled model. *Clim Dyn* **46,** 1397–1415. https://doi.org/10.1007/s00382-015-2654-6

Osprey, S., Butchart, N., Knight, J. R., Scaife, A. A., Hamilton, K., Anstey, J. A., Schenzinger, V. and Zhang, C. (2016). An unexpected disruption of the atmospheric quasi-biennial oscillation. *Science*, **353**(6306), 1424–1427. DOI: 10.1126/science.aah4156

Palmer, T. N., et al., 2004: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bull. Am. Meteorol. Soc.*, **85**, 853–872, doi:10.1175/BAMS-85-6-853.

Palmeiro, F. M., Barriopedro, D., García-Herrera, R., and Calvo, N. (2015). Comparing sudden stratospheric warming definitions in reanalysis data. *J. Clim.*, **28**(17), 6823-6840.

Palmeiro F. M., J. García-Serrano, O. Bellprat, P. A. Bretonnière and F.J. Doblas-Reyes 2020: Boreal winter stratospheric variability in EC-EARTH: High-Top versus Low-Top. *Climate Dynamics*, **54**, 3135-3150.

Pohlmann, H., W. A. Müller, K. Kulkarni, M. Kameswarrao, D. Matei, F. Vamborg, C. Kadow, S. Illing, and J. Marotzke, 2013: Improved forecast skill in the tropics in the new MiKlip decadal climate predictions, *Geophys. Res. Lett.*, **40**, 5798–5802, doi:10.1002/2013GL058051

Richter, J. H., Butchart, N., Kawatani, Y., Bushell, A. C., Holt, L., Anstey, J., Serva, F., Simpson, I. R., Osprey, S. M., Hamilton, K., Braesicke, P., Cagnazzo, C., Chen, C., Garcia,

R. R., Gray, L. J., Kerzenmacher, T., Lott, F., McLandress, C., Naoe, H., Scinocca, J., Stockdale, T. N., Watanabe, S., Yoshida, K. and Yukimoto, S. (2020) Response of the quasi-biennial oscillation to a warming climate in global climate models. *Q. J. R. Meteorol. Soc.*, this issue.

Scaife, A.A., Butchart, N., Warner, C.D., Stainforth, D., Norton, W. and Austin, J., 2000: Realistic quasi-biennial oscillations in a simulation of the global climate. *Geophys. Res. Lett.* **27**, 3481-3484. doi: 10.1029/2000GL011625

Scaife AA, Knight JR, Vallis GK, Folland CK, 2005: A stratospheric influence on the winter NAO and North Atlantic surface climate. *Geophys. Res. Lett.* **32**: L18715, doi: 10.1029/2005GL023226

Scaife, A.A., Athanassiadou, M., Andrews, M., Arribas, A., Baldwin, M., Dunstone, N., Knight, J., Maclachlan, C., Manzini, E., Müller, W.A., Pohlmann, H., Smith, D., Stockdale, T. and Williams, A., 2014: Predictability of the quasi-biennial oscillation and its northern winter teleconnection on seasonal to decadal timescales. *Geophys. Res. Lett.*, **41**, 1752-1758, doi:10.1002/2013GL059160

Scinocca, J. F., McFarlane, N. A., Lazare, M., Li, J. and Plummer, D. (2008) Technical note: The CCCma third generation AGCM and its extension into the middle atmopshere. *Atmos. Chem. Phys.*, **8**, 7055–7074. doi:10.5194/acp-8-7055-2008

Stockdale, T., 2013: The EUROSIP system-a multi-model approach. In: *Seminar on seasonal prediction: science and applications, 3–7 September 2012. ECMWF, Reading, UK*, pp 257–268. https://www.ecmwf.int/node/12429

Takahashi, M., 1996: Simulation of the stratospheric quasi-biennial oscillation using a general circulation model. *Geophys. Res. Lett.*, **23**, 661–664.

Tripathi, O. P., and Coauthors, 2016: Examining the Predictability of the Stratospheric Sudden Warming of January 2013 Using Multiple NWP Systems. *Mon. Wea. Rev.*, **144**, 1935–1960. doi: 10.1175/MWR-D-15-0010.1.

Van Loon, H., and Labitzke, K., 1987: The Southern Oscillation. Part V: The anomalies in the lower stratosphere of the Northern Hemisphere in winter and a comparison with the quasi-biennial oscillation. *Mon. Weath. Rev.*, **115**, 357-369.

Wallace, J.M., Panetta, R.L. and Estberg, J., 1993: Representation of the equatorial stratospheric quasi-biennial oscillation in EOF phase space. *J. Atmos. Sci.*, **50**, 1751–1762.

Watanabe, S., Hajima, T., Sudo, K., Nagashima, T., Takemura, T., Okajima, H., Nozawa, T., Kawase, H., Abe, M., Yokohata, T., Ise, T., Sato, H., Kato, E., Takata, K., Emori, S., and Kawamiya, M., 2011: MIROC-ESM 2010: model description and basic results of CMIP5-20c3m experiments, *Geosci. Model Dev.*, **4**, 845–872. doi: 10.5194/gmd-4-845-2011

Watanabe, S., K. Hamilton, S. Osprey, Y. Kawatani and E. Nishimoto, 2018: First Successful Hindcasts of the 2016 Disruption of the Stratospheric Quasibiennial Oscillation. *Geophys. Res. Lett.*, **45**, 1602-1610. doi: 10.1002/2017GL076406

Watson, P. A. G., and L. J. Gray, 2014: How does the quasi-biennial oscillation affect the stratospheric polar vortex?, *J. Atmos. Sci.*, **71**, 391–409. doi:10.1175/JAS-D-13-096.1

Weisheimer, A., F. J. Doblas-Reyes, T. N. Palmer, A. Alessandri, A. Arribas, M. Déqué, N. Keenlyside, M. MacVean, A. Navarra, and P. Rogel, 2009: ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions—Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophys. Res. Lett.*, **36**, L21711, doi:10.1029/2009GL040896.

Yukimoto, S., Adachi, Y., Hosaka, M., Sakami, T., Yoshimura, H., Hirabara, M., Tanaka, T. Y., Shindo, E., Tsujino, H., Deushi, M., Mizuta, R., Yabu, S., Obata, A., Nakano, H., Koshiro, T., Ose, T., and Kitoh, A.: A new global climate model of the Meteorological Research Institute: MRI-CGCM3 -Model descrip- tion and basic performance-, *J. Meterol. Soc. Jpn.*, 90A, 23–64, https://doi.org/10.2151/jmsj.2012-A02, 2012.

Figure captions

Figure 1: MSSS relative to climatology of 5N-5S monthly mean zonal-mean zonal wind forecasts for individual pressure levels from 10 to 70hPa. Coloured lines are the scores for individual models, black dashed line is the score for an anomaly persistence forecast. Scores are calculated from a set of 30 forecasts starting in May and November of the years 1993-2007.

Figure 2: Mean zonal-mean zonal wind forecasts at 30hPa and 70 hPa, shown separately for May and November starts. Individual models are plotted in colours, the observed mean climatology is the black dashed line. All models have a substantial positive bias at 30 hPa in June-December; most models also have a negative bias at 70 hPa.

Figure 3: Ratio of modelled to observed amplitude of zonal-mean zonal wind anomalies as a function of lead time for individual models at 20hPa and 50 hPa.

Figure 4: Anomaly correlation of zonal-mean zonal wind forecasts for pressure levels from 10 to 70 hPa. Colours show individual models, black dashed line shows anomaly persistence obtained by persisting the analysed monthly-mean anomaly from the month immediately before the start of the forecast.

Figure 5: Easterly and westerly wind composited respectively for the cases of ten strongest analysed monthly-mean easterly and westerly winds at each level and forecast verification time.

Figure 6: Upper row: trajectory of pairs of the 20- and 50-hPa monthly zonal wind starting from the first to the ninth simulated months in each model (except MPI-ESM-MR whose trajectory is up to seventh month). The same trajectory but using ERAI is also presented (black), continued in grey to the 24th month. The zonal wind pair on the initial date is indicated by the open circle using ERAI, and that for the third simulated month is by the filled circle on each trajectory. Lower row: the same as the upper row, except the trajectories are of the 30- and 70-hPa wind, with filled circles representing the fifth simulated month.

Figure 7: Root mean square error (solid lines) and ensemble spread (dashed lines) for monthly zonal-mean zonal wind forecasts from individual models for pressure levels from 10 to 70 hPa. Note the change of vertical scale with height. Black dashed line represents the r.m.s. error of anomaly persistence.

Figure 8: Top row: Root mean square error (solid) and ensemble spread (dashed) for monthly zonal-mean zonal wind forecasts comparing a multi-model forecast (red) with the best-performing individual model at that level (blue). Left is at 30 hPa, where the comparison is with AGCM3-CMAM, right is at 50 hPa, where the comparison is with IFS43r1. Bottom row shows the corresponding anomaly correlation scores. Black dashed lines represent the scores of anomaly persistence forecasts.

Figure 9: Monthly-mean zonal-wind tendency, vertical advection, Eliassen–Palm (EP) flux divergence, parameterised gravity wave drag (GWD), and the residual of the zonal-wind tendency equation averaged over 5N–5S in each model (from left to right) for the first month of each of the hindcasts initialised on (upper) 1 Nov 1999, (middle) 1 May 1994, and (lower) 1 Nov 1996, along with those in ERAI except GWD. The monthly- and zonal-mean zonal wind in ERAI is also presented in each panel (dash). The residual in ERAI is plotted in the panel for GWD in models. Note that MIROC-AGCM-LL does not include GWD parametrization.

Figure 10: Stratospheric NH winter vortex response to QBO phase. (a) Top panel: Seasonal variation of observed NH vortex response to QBO phase, using ERAI reanalysis, 1979-2016. QBO phase defined by phase angle determined from the two leading EOFs of 10-70 hPa equatorial wind, following Wallace et al. 1993. Range [0,1] represents a full QBO cycle. Each year is assigned a QBO phase according to the monthly-mean Nov value of the phase angle. Vortex is defined by deseasonalised monthly-mean 10 hPa 60N zonal-mean zonal wind. Each box shows mean of the zonal wind values falling within a box of width 0.25 in phase angle. Circles indicate where absolute value of the mean exceeds $1.5\sigma/\sqrt{n}$, where σ is the standard deviation of the zonal wind about its climatological mean for a given month and n is the population of the box, indicated in green at the bottom of each column. Circle size is scaled by σ, and boxes for which the mean exceeds $2\sigma/\sqrt{n}$ have white circles. (For a normal distribution, 87% and 95% of values lie within 1.5σ and 2.0σ of the mean, respectively.) Middle panel: Equatorial zonal wind profiles composited for times when phase angle takes the values on the x-axis. Bottom panel: Equatorial wind profiles in Nov of each year. The value of phase angle corresponding to a given profile is indicated by the vertical line that has the same colour as the wind profile. (b-i): As (a), but for the full set of QBOi hindcasts initialised in November for each model.

Figure 11: Impact of QBO phases on frequency of Sudden Stratospheric Warmings (SSWs). For each model and for ERAI, the frequency of SSWs per decade are plotted for QBOE (blue) and QBOW years (red). Whiskers represent the $\pm\sigma$ interval. The overall decadal frequency of SSWs is also given in brackets. Results are shown for November initialisations (left) and additionally for May initialisations for two models (right).

Figure 12: Impact of QBO phases on the timing of Sudden Stratospheric Warmings (SSWs). For each model and for ERAI, the mean and standard deviation of the SSW onset dates are plotted for QBOE and QBOW years. Only the first SSW in considered for winters with multiple SSWs (see text for details). Results are shown for November initialisations (upper) and additionally for May initialisations for two models (lower).

Model (resolution, model top)	Forecast length (months)	Ensemble members & generation	Ocean	Initialisation method	References
AGCM3-CMAM (T47L113, 0.0007hPa)	12	8, physics parameter change during spinup	Prescribed SST	Nudging to ERAI, vorticity, divergence,T, ⍰=6hr	Scinocca et al. 2008, Anstey et al. 2016
EC-EARTH3.1 (T255L91, 0.01hPa)	12	5, Singular vector perturbations	Prescribed SST	ERAI interpolation div,vort, t, INSP GG: q, cc, crwc, cswc, clwc, ciwc	Palmeiro et al. 2020
MIROC-AGCM-LL (T106L72, 1.2hPa)	9	5	Prescribed SST	ERAI interpolation U,v,T,q,ps	Kawatani et al. 2011
MIROC-ESM (T42L80, 0.004 hPa)	9	3, differing nudging strengths	Prescribed SST	Nudging troposphere and strat. to ERAI (τ= 0.1, 0.11 and 0.12 days)	Watanabe et al. 2011
MRI-ESM2 (T159L80, 0.01 hPa)	9	3, lagged method with -5, 0, 5 days	Prescribed SST	ERAI interpolation, u,v,T,q,Z,ps	Yukimoto et al, 2012
MPI-ESM-MR (T63L95, 0.01 hPa)	7	10, perturbations to diffusion and ocean breeding	Coupled Ocean-Atmosphere	Nudging to ERAI, ⍰=48h div, 6h vort, 24h T, 24h ps; Ocean nudging to ORAS4: T and S (10d), Sea-ice nudging to NSIDC (1d)	Dobrynin et al. 2018
UMGC2 (N216L85, ~50km, 0.005 hPa)	12	40, stochastic physics	Coupled Ocean-Atmosphere	ERAI, weakly coupled nudging, τ=6hr in the atmosphere. Iterative global covariances and EN3 in ocean.	Dunstone et al. 2016
IFS43r1 (T255L137, 0.01 hPa)	12	5, stochastic physics	Prescribed SST	ERAI interpolated to L137	ECMWF, 2016

Abbreviations: AGCM3-CMAM, Third generation Atmospheric General Circulation Model (Canadian Middle-Atmosphere Model); EC-EARTH3.1, EC-Earth Consortium Earth system model, version 3.1; MIROC-AGCM-LL, Model for Interdisciplinary Research on Climate, Atmospheric General Circulation Model (Low Lid); MIROC-ESM, Model for Interdisciplinary Research on Climate, Earth System Model; MRI-ESM2, Meteorological Research Institute Earth System Model, version 2; MPI-ESM-MR, Max Planck Institute for Meteorology Earth System Model, Mixed Resolution; UMGC2, Met Office Unified Model, Global Coupled configuration, version 2; IFS43r1, ECMWF Integrated Forecasting System, Cycle 43r1.

Table 1: Summary of QBOi models used, and the initialisation method. More information on the models is given in Butchart et al, 2018.