Received: 1 August 2019 | Revised: 7 November 2019 | Accepted: 30 January 2020

DOI: 10.1002/qj.3765

Quarterly Journal of the Royal Meteorological Society

SPECIAL SECTION QBO MODELLING INTERCOMPARISON

Evaluation of the Quasi-Biennial Oscillation in global climate models for the SPARC QBO-initiative

A. C. Bushell[1] | J. A. Anstey[2] | N. Butchart[3] | Y. Kawatani[4] | S. M. Osprey[5] | J. H. Richter[6] | F. Serva[7] | P. Braesicke[8] | C. Cagnazzo[7] | C.-C. Chen[6] | H.-Y. Chun[9] | R. R. Garcia[6] | L. J. Gray[5] | K. Hamilton[10] | T. Kerzenmacher[8] | Y.-H. Kim[11,12] | F. Lott[13] | C. McLandress[14,2] | H. Naoe[15] | J. Scinocca[2] | A. K. Smith[6] | T. N. Stockdale[16] | S. Versick[8] | S. Watanabe[4] | K. Yoshida[15] | S. Yukimoto[15]

[1]Met Office, Exeter, UK

[2]Canadian Centre for Climate Modelling and Analysis (CCCma), Victoria, Canada

[3]Met Office Hadley Centre, Exeter, UK

[4]Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokohama, Japan

[5]Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, UK

[6]National Center for Atmospheric Research (NCAR), Boulder, Colorado, USA

[7]Institute of Marine Sciences, National Research Council (ISMAR-CNR), Rome, Italy

[8]Karlsruher Institut für Technologie (KIT), Karlsruhe, Germany

[9]Yonsei University, Seoul, South Korea

[10]International Pacific Research Center and Department of Atmospheric Sciences, University of Hawaii, Honolulu, Hawaii

[11]Institut für Atmosphäre und Umwelt, Goethe-Universität, Frankfurt am Main, Germany

[12]Ewha Womans University, Seoul, South Korea

[13]Laboratoire de Météorologie Dynamique (LMD), Paris, France

[14]University of Toronto, Toronto, Canada

[15]Meteorological Research Institute (MRI), Tsukuba, Japan

[16]European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK

Abstract

Quasi-biennial oscillations (QBOs) in thirteen atmospheric general circulation models forced with both observed and annually repeating sea surface temperatures (SSTs) are evaluated. In most models the QBO period is close to, but shorter than, the observed period of 28 months. Amplitudes are within ±20% of the observed QBO amplitude at 10 hPa, but typically about half of that observed at lower altitudes (50 and 70 hPa). For almost all models, the oscillation's amplitude profile shows an overall upward shift compared to reanalysis and its meridional extent is too narrow. Asymmetry in the duration of eastward and westward phases is reasonably well captured, though not all models replicate the observed slowing of the descending westward shear. Westward phases are generally too weak, and most models have an eastward time mean wind bias throughout the depth of the QBO. The intercycle period variability is realistic and in some models is enhanced in the experiment with observed SSTs compared to the experiment with repeated annual cycle SSTs. Mean periods are also sensitive to this difference between SSTs, but only when parametrized non-orographic gravity wave (NOGW) sources are coupled to tropospheric parameters and not prescribed with a fixed value. Overall, however, modelled QBOs are very similar whether or not the prescribed SSTs vary interannually. A portrait of the overall ensemble performance is provided by a normalized grading of QBO metrics. To simulate a QBO, all but one model used parametrized NOGWs, which provided the majority of the total wave forcing at altitudes above 70 hPa in most models.

Correspondence
A.C. Bushell, Met Office, FitzRoy Road,
Exeter EX1 3PB, UK.
Email: andrew.bushell@metoffice.gov.uk

Funding information
Agence Nationale Recherche, Grant/Award Number: ANR-15-JCLI-0004-01; Baden-Wuerttemberg, Grant/Award Number: bwHPC; Biological and Environmental Research, Grant/Award Number: IA1947282; Department for Business, Energy and Industrial Strategy; Department for Environment, Food and Rural Affairs; European Commission, Grant/Award Number: Copernicus Climate Change Service; Japan Agency for Marine-Earth Science and Technology; Japan Science and Technology Agency; Japan Society for the Promotion of Science, Grant/Award Number: JP15KK0178, Grant/Award Number: JP17K18816, Grant/Award Number: JP18H01286; Ministry of Education, Culture, Sports, Science and Technology, Grant/Award Number: Integrated Research Program for Advancing Climate Models; National Center for Atmospheric Research, Grant/Award Number: Cooperative Agreement No. 1852977; National Centre for Atmospheric Science; Natural Environment Research Council, Grant/Award Number: NE/M005828/1, Grant/Award Number: NE/P006779/1

Hence the representation of NOGWs either explicitly or through parametrization is still a major uncertainty underlying QBO simulation in these present-day experiments.

KEYWORDS

general circulation models, gravity waves, Quasi-Biennial Oscillation, stratosphere, tropical variability

1 INTRODUCTION

A key objective of the Stratosphere-troposphere Processes And their Role in Climate (SPARC) Quasi-Biennial Oscillation initiative (QBOi) is to improve confidence in general circulation and earth system model (GCM and ESM) simulations of the QBO, a prominent feature of tropical variability in the middle atmosphere first identified nearly sixty years ago (Ebdon and Veryard, 1961; Reed *et al.*, 1961). Understanding and predicting this variability is important for accurate representation of tropical to extratropical teleconnections (e.g., Huntingford *et al.*, 2014), seasonal forecasts in the extratropics (e.g., Scaife *et al.*, 2014) and the assessment of earth system model responses to climate change (e.g., Kawatani and Hamilton, 2013).

Observations, theory and modelling of the QBO were detailed in a major review (Baldwin *et al.*, 2001), which noted that very few Atmospheric GCMs (AGCMs) had been able to simulate such internal oscillations since the QBO was first modelled in a GCM by Takahashi (1996). Scaife *et al.* (2000) showed that including parametrized forcing from unresolved waves was sufficient to simulate a realistic QBO. Today there is improved understanding of the shortfalls in AGCM momentum budgets when contributions from small-scale waves are missing (e.g., Pulido and Thuburn, 2008) and more AGCMs have gained the capacity to generate a QBO, both by ensuring adequate vertical resolution in the stratosphere and by parametrizing accelerations due to subgrid non-orographic gravity waves (NOGWs). Nonetheless, only five out of 47 models contributing to the Coupled Model Intercomparison Project Phase 5 (CMIP5) had spontaneous QBOs (Schenzinger *et al.*, 2017; Butchart *et al.*, 2018). By comparison, the new CMIP6 generation of models is expected, among other improvements, to have generally higher resolution and higher upper boundaries. Hence, simulated QBOs are expected to become more common. However, the complexity of CMIP6 models and their forcing scenarios will impede use of the CMIP6 multimodel ensemble to analyse modelling uncertainties that are specific to the QBO and its impacts. Limited availability of stratospheric diagnostics from some models may also further restrict the usefulness of the CMIP6 ensemble for in-depth QBO analysis.

The QBOi multimodel ensemble represents an alternative approach in which modelling uncertainties related to the QBO are assessed by performing coordinated experiments with AGCMs that have simplified external forcings and boundary conditions, designed to characterize QBO representation and its response to idealized future climate scenarios (Butchart *et al.*, 2018). While a companion paper addresses the climate change simulations (Richter *et al.*, 2020, in this issue), this paper focuses on the present-day experiments and:

1. assesses the representation of tropical stratospheric climatology and variability by the participating models,
2. evaluates the impact on simulated QBOs of imposed interannual variability in sea-surface temperatures, and
3. explores the sensitivity of simulated QBOs to model characteristics and in particular differences in how the models represent NOGWs and their associated wave sources (e.g., specified versus parametrized).

The paper builds on an assessment (Butchart *et al.*, 2011) of tropical variability in 16 simulations for the Chemistry Climate Model Validation (CCMVal) project and a QBO-focused multimodel comparison of ten CMIP3/5 and CCMVal-2 models (Schenzinger *et al.*, 2017). The QBOi ensemble represents an important advance on such previous studies because it collates QBO simulations from a larger number of AGCMs than have previously been analysed together, and evaluates them using coordinated experiments that permit a cleaner comparison by eliminating effects of intermodel differences in external forcing or boundary conditions. Section 2 provides a brief overview of the models and experiments analysed (the reader is referred to Butchart *et al.*, 2018, for full details) along with a description of methods. Section 3 summarizes the mean state of the tropical stratosphere in simulations and evaluates basic characteristics of modelled QBOs. Section 4 applies a set of metrics to compare models with each other and with reanalysis. Normalized gradings of these metrics in Section 5 provide an integrated assessment of the ensemble performance over selected metrics. QBO forcing by resolved and parametrized waves is examined in Section 6 before concluding remarks are given in Section 7.

2 | EXPERIMENT DATASETS AND METHODS

[AUTHOR: The following comment has been made on the Richter et al. article to be included in the QBO Special Section.

'It is QJ convention that acronyms should be fully defined, either in the text or in an Appendix. We would normally expect that the model names used here should be explained (or a reference given to a list elsewhere) to allow the non-specialist to identify them. Butchart et al. (2018) have a good Table 5, but the acronyms are really not fully explained there. Perhaps you may know of a suitable unpublished table which could be readily adapted as an Appendix to this paper?'

It is desirable that one of the papers in the group should list the full names of the models referred to.]

2.1 | Model and reanalysis datasets

In order to enable validation against reanalyses, Experiment 1 of QBOi (hereafter Exp 1) specifies a 1- to 3-member ensemble of AGCM simulations over the 30-year period 1 January 1979 to 28 February 2009. Exp 1 is based on CMIP5 experiment 3.3, which uses observed sea-surface temperatures (SSTs) and sea-ice amounts prescribed under the Atmospheric Model Intercomparison Project (AMIP), as well as contemporaneous external forcings. (Butchart *et al.*, 2018 give design details of all the QBOi experiments.) Experiment 2 (Exp 2) specifies identical model configurations with those in Exp 1, except that for SSTs and sea-ice amounts a repeated annual cycle is constructed from Exp 1 data and used, along with fixed prescriptions for other external forcings. Exp 2 acts as control for two idealized climate change experiments (Experiments 3–4), which are analysed in a companion paper (Richter *et al.*, 2020). Table 1 summarizes details specific to Exp 1 and Exp 2, whereas a comprehensive list of GCMs participating in QBOi and information relevant to them is found in table 5 of (Butchart *et al.*, 2018). In total, output was analysed from thirteen Exp 1 models and eleven Exp 2 models uploaded to the QBOi archive at the time.

TABLE 1 Models participating in QBOi Experiments 1 and 2, showing the primary institute responsible for model data, chosen parametrizations for non-orographic gravity waves (NOGW) and gravity wave sources (NOGW Source), the number of latitude points to which zonal mean data are gridded (N_{lat}), the number of model levels (N_{lev}), the highest data pressure level[a] used (p_{lim} in hPa), and the ensemble size × number of years presented for each experiment

Model	Institute	NOGW	NOGW source	N_{lat}	N_{lev}	[a]p_{lim}	Exp 1	Exp 2
60LCAM5	NCAR	Li	Richter *et al.* (2010)	192	60	3.0	3×30	3×30
AGCM3-CMAM	CCCma	WM	Fixed	48	113	0.4	3×30	3×30
CESM1(WACCM5-110L)	NCAR	Li	Richter *et al.* (2010)	192	110	0.4	3×30	3×30
ECHAM5sh	CNR	Hi	Fixed	96	95	0.4	1×30	1×30
EMAC	KIT	Hi	Fixed	64	90	0.4	1×30	1×100
HadGEM2-A	Yonsei Univ.	WM	Fixed	145	60	0.4	1×28	—
HadGEM2-AC	Yonsei Univ.	WM	Choi and Chun (2011)	145	60	0.4	1×28	—
LMDz6	IPSL-LMD	Lo	Lott and Guez (2013)	143	79	0.4	1×30	1×70
MIROC-AGCM-LL	MIROC	None	N/A	160	72	5.0	3×30	3×30
MIROC-ESM	MIROC	Hi	Fixed	64	80	0.4	3×30	3×100
MRI-ESM2	MRI-JMA	Hi	Fixed	160	80	0.4	1×30	1×30
UMGA7	Met Office	WM	Fixed	145	85	0.4	3×30	1×100
UMGA7gws	Met Office	WM	Bushell *et al.* (2015)	145	85	0.4	3×30	1×100

Note: NOGW schemes are abbreviated as: Hines (1997) [Hi]; Warner and McIntyre (1999) [WM]; Lindzen (1981) [Li]; Lott [Lo]. Butchart *et al.* (2018) give more information on model characteristics.

[a] Protocol levels used: 300, 250, 200, 175, 150, 120, 100, 85, 70, 60, 50, 40, 30, 20, 15, 10, 7, 5, 3, 2, 1.5, 1, 0.4 hPa.

Although some differences exist in how the QBO is represented among reanalyses, Schenzinger *et al.* (2017) found that results for diagnostics like those considered in this study were largely independent of the reanalysis chosen. Hence, ERA-Interim (Dee *et al.*, 2011) for the 30-year period 1979–2009 is used here for validating the simulated QBOs.

2.2 | Methods

This study characterizes the QBO using a set of metrics that are similar, though not completely converged with, those used in Schenzinger *et al.* (2017). For instance, adaptation was required when the identification of a single dominant QBO peak from Fourier transform (FT) spectra proved problematic in the warming climate experiments (Richter *et al.*, 2020). Basing metrics instead upon QBO periods defined by transitions between wind phases allows the same methods to be used for both present and future experiments. Nontheless, to aid comparison with Schenzinger *et al.* (2017), QBO periods derived from peaks in the FT power spectra are also evaluated (Table 2) and correlate strongly (coefficient =0.93) with the transition periods. An added benefit from the transition cycle approach is the ability to derive intercycle properties such as range and standard deviation as, for instance, is required to calculate grades in Section 5.

2.2.1 | Transitions between eastward and westward QBO wind phases

A fixed *reference level* of 10 hPa was chosen as it is closest to the level where the QBO amplitude is a maximum (metric h_{max} in Schenzinger *et al.*, 2017) in a majority of the models for Exp 1 and Exp 2 (Table 3; Section 4.2). Transitions between QBO eastward and westward wind phases at the 10 hPa reference level are identified by applying the following method across all models, including the reanalysis. Near-equatorial zonal and monthly mean zonal wind, $\bar{u}_{\mathrm{eq}}$, is defined as the mean over latitudes within the range 5°S–5°N weighted by the cosine of latitude. $\bar{u}_{\mathrm{eq}}$ is first smoothed with a five-month running mean to reduce variability at the shortest time-scales, which can result in spurious phase reversals of 1 month or more that degrade the QBO period statistics. Sometimes in Exp 1 even smoothed winds remain close to zero between adjacent transitions and on two occasions (October 2006 in the third 60LCAM5 simulation and March 1980 in the MRI-ESM2 simulation) change sign for just a single month. These single-month excursions are detected and ignored when identifying

TABLE 2 QBO period metrics (see Section 2.2.2) evaluated at the 10 hPa reference level for Exp 1 models and ERA-Interim

		QBO transition period (months)				
Model	**FT period**	**Min**	**Max**	***N***	**Mean ± SD**	**Pt[E : W]**
ERA-Interim	28	22.0	35.0	12	27.8 ± 3.6	37:63
60LCAM5	26	19.0	35.0	38	26.2 ± 3.6	48:52
AGCM3-CMAM	28	24.0	32.0	36	27.9 ± 2.0	28:72
CESM1(WACCM5-110L)	28	23.0	39.0	33	29.6 ± 4.2	37:63
ECHAM5sh	27	23.0	32.0	13	25.9 ± 2.4	60:40
EMAC	26	23.0	32.0	13	25.4 ± 2.4	56:44
HadGEM2-A	26	23.0	28.0	12	(**25.0 ± 1.5**)	38:62
HadGEM2-AC	27	18.0	33.0	12	25.6 ± 4.1	41:59
LMDz6	28	25.0	32.0	11	29.3 ± 1.8	53:47
MIROC-AGCM-LL	20	17.0	25.0	50	(**20.0 ± 1.7**)	48:52
MIROC-ESM	24	18.0	31.0	42	(**24.5 ± 2.8**)	49:51
MRI-ESM2	24	12.0	27.0	15	(**22.4 ± 3.7**)	43:57
UMGA7	26	22.0	33.0	39	(**25.8 ± 2.3**)	35:65
UMGA7gws	26	23.0	32.0	40	(**25.8 ± 2.2**)	37:63

Note: FT period is the period (months) derived from the Fourier spectrum peak. Min, Max indicate the range of periods, and *N* is the number of periods identified. Mean±SD is the mean and standard deviation of periods (values in bold indicate that mean biases against ERA-Interim are significant at the 5% level with (..) indicating that all such biases are negative). Pt[E : W] is the ratio of durations of eastward and westward phases expressed as percentages of the mean period.

QBO cycles. A transition between QBO phases is defined as the time (TT) when $\bar{u}_{\mathrm{eq}}$ first passes from westward to eastward (i.e., through zero) or vice versa (results in Section 3.2).

2.2.2 | QBO periods, mean cycles and multimodel means

The duration of each full QBO cycle is calculated as the difference in months between subsequent eastward phase onsets, that is, the westward-to-eastward transition times, at the reference level. In a given model, the mean QBO period is simply the average of all available full-cycle durations. Means, standard deviations and period ranges for Exp 1 and ERA-Interim are catalogued in Table 2 (the table caption gives more details) and for each model are quite similar to equivalent metrics for Exp 2 (not shown). The values for ERA-Interim do not differ significantly from their equivalents in table 3 of Schenzinger *et al.* (2017). Further analysis of the periods is presented in Sections 3 and 4.

A mean QBO cycle can be computed from model ensemble data by transforming the time axis of each QBO cycle in the timeseries from t to

$$\tilde{t} = t \times \left(\frac{\text{normalizing period}}{\text{cycle period}} \right),$$

such that the duration of every cycle in terms of $\tilde{t}$ becomes identical (Section 3.2 and following). Normalizing periods herein are set either to the model mean period or, if combining models, the multimodel mean period. In order for each model to receive equal weight, multimodel means for any given metric are calculated after first obtaining the mean for each model.

2.2.3 | QBO amplitude

Three methods are used in this study to calculate QBO amplitudes (or half the QBO peak-to-peak signal) and each utilises $\bar{u}_{\mathrm{eq}}$. The QBO is assumed to dominate variability in $\bar{u}_{\mathrm{eq}}$, thereby guaranteeing that consecutive QBO period transition times (TTs; Section 2.2.1) straddle a clearly identified single cycle, though this assumption will break down at latitudes away from the Equator where the QBO is not sustained.

- **TTeq**. For the eastward and westward phases of each QBO cycle (delineated by TT) the amplitude of a given

TABLE 3 Metrics for QBO amplitudes (10 hPa; peak; 50 hPa) and extents (see Section 2) evaluated for Exp 1 models and ERA-Interim.

	TTeq amplitude 10 hPa		**DD amplitude**			**TT-based extent metrics**			**TTeq amplitude 50 hPa**	
	Mean ± SD	**Pa[E : W]**	**A_{QBO}**	**p_{Max}**	**A_{Max}**	**$dZ \pm$ SD**	**$W_{10} \pm$ SD**	**$W_{50} \pm$ SD**	**Mean ± SD**	**Pa[E : W]**
Model	**($m \cdot s^{-1}$)**	**(%)**	**($m \cdot s^{-1}$)**	**(hPa)**	**($m \cdot s^{-1}$)**	**(km)**	**(°)**	**(°)**	**($m \cdot s^{-1}$)**	**(%)**
ERA-Interim	23.9 ± 3.0	28:72	25.5	15.0	28.4	8.7 ± 2.1	27.5 ± 2.5	23.4 ± 2.0	14.2 ± 1.66	37:63
60LCAM5	23.2 ± 1.8	40:60	22.0	20.0	26.5	8.7 ± 1.5	26.8 ± 2.1	**(21.0 ± 1.8)**	**16.8 ± 3.04**	64:36
AGCM3-CMAM	23.0 ± 1.0	28:72	23.4	7.0	24.3	9.0 ± 0.7	**(23.7 ± 2.7)**	**(17.8 ± 2.0)**	**(8.3 ± 0.74)**	58:42
CESM1(WACCM5-110L)	**26.0 ± 1.3**	29:71	29.0	10.0	29.0	10.9 ± 1.0	26.2 ± 2.6	**(18.6 ± 1.6)**	**16.3 ± 1.50**	40:60
ECHAM5sh	**27.8 ± 2.0**	46:54	29.4	5.0	30.9	7.0 ± 1.3	**(23.4 ± 1.6)**	**(16.4 ± 2.1)**	**(8.9 ± 1.60)**	89:11
EMAC	**26.1 ± 1.2**	40:60	27.7	7.0	28.3	8.2 ± 1.1	**(24.8 ± 1.3)**	**(18.2 ± 2.5)**	**(8.9 ± 1.68)**	60:40
HadGEM2-A	24.0 ± 1.0	33:67	27.0	10.0	27.0	10.5 ± 0.8	25.9 ± 2.7	**(18.5 ± 2.4)**	**(11.6 ± 1.21)**	29:71
HadGEM2-AC	24.0 ± 2.0	38:62	27.0	10.0	27.0	9.3 ± 0.7	**(25.9 ± 0.9)**	**(17.4 ± 2.3)**	**(9.5 ± 1.55)**	34:66
LMDz6	22.2 ± 2.3	41:59	23.0	1.0	26.6	8.4 ± 0.9	**(22.1 ± 1.8)**	**(16.6 ± 2.6)**	**(7.1 ± 1.49)**	33:67
MIROC-AGCM-LL	**(19.6 ± 0.8)**	39:61	20.8	10.0	20.8	7.8 ± 1.2	**(21.7 ± 1.7)**	**(20.1 ± 3.4)**	**(6.4 ± 1.16)**	49:51
MIROC-ESM	22.8 ± 1.6	33:67	24.6	10.0	24.6	8.5 ± 1.3	**(22.8 ± 1.6)**	**(19.1 ± 2.6)**	**(7.9 ± 1.71)**	38:62
MRI-ESM2	**(19.1 ± 2.4)**	39:61	20.7	5.0	22.8	8.1 ± 1.5	**(22.5 ± 2.2)**	**(18.5 ± 2.7)**	**(6.4 ± 1.66)**	26:74
UMGA7	**28.5 ± 1.6**	37:63	32.5	10.0	32.5	9.5 ± 0.8	**(25.9 ± 1.3)**	**(18.6 ± 2.6)**	**(12.1 ± 1.13)**	40:60
UMGA7gws	23.8 ± 1.4	36:64	26.1	10.0	26.1	9.3 ± 0.8	**(24.9 ± 1.6)**	**(19.3 ± 2.5)**	**(9.5 ± 1.22)**	38:62

Note: Mean±SD is the mean and standard deviation of amplitudes; values in bold indicate that mean biases against ERA-Interim are significant at the 5% level and (..) indicate such biases are negative. Pa[E : W] is the percentage of peak to peak (2×total amplitude) contributed by mean phase amplitudes eastward : westward. A_{QBO}, A_{Max} are amplitudes (based on Section 2.2.3, method DD) evaluated at the 10 hPa reference level and the pressure level, p_{Max}, at which the maximum QBO amplitude is detected. dZ±SD is the mean and standard deviation of the attenuation depth metric (Section 2.2.4). W_{10}±SD, W_{50}±SD are mean and standard deviation of the QBO width metric at 10 hPa, 50 hPa (Section 2.2.4) phase at a specified level is defined to be the maximum value of $|\bar{u}_{\text{eq}}|$ at that level, after a 5-month centred binomial smoothing is applied to the timeseries, which preferentially reduces any systematic impact of short-period fluctuations.

- **TTmceq**. For most variables considered here (e.g., $\bar{u}_{\text{eq}}$, temperature, wave forcing), a mean QBO cycle is calculated by averaging individual cycles (defined by TT at the reference level) using the method of Section 2.2.2. Mean cycle amplitude profiles are defined, at each altitude, as half the (maximum minus minimum) value over the cycle time duration.
- **DD**. Dunkerton and Delisi (1985) argued that the root mean square of the deseasonalised $\bar{u}_{\text{eq}}$ timeseries (approximately, the standard deviation) multiplied by $\sqrt{2}$ provides a good estimate of QBO amplitude. An advantage of this method is that it does not require explicit calculation of transition times between QBO phases, which is useful when the QBO signal becomes less discernible, such as occurs in some future climate simulations (Richter *et al.*, 2020).

Further analysis and comparison of the amplitudes is presented in Sections 3.2 and 4, while amplitudes calculated by methods TTeq and DD for Exp 1 are catalogued in Table 3 (the caption gives more details) and are strongly correlated with a coefficient of 0.94. For each model, results for Exp 2 (not shown) are again quite similar to those for Exp 1. Amplitude metrics that employ the same 10 hPa transition times but are evaluated at a lower level (50 hPa) also appear in Table 3.

2.2.4 | QBO vertical and latitudinal extent

The metric for vertical extent is based upon the rate of decrease with altitude on descent through the stratosphere from an upper level defined as the level of maximum amplitude or 10 hPa, whichever is lower. As a key difference from Schenzinger *et al.* (2017), where a Gaussian fit to the FT amplitude peak is used, the method in this study is focused on QBO cycle amplitudes (TTeq; TTmceq). A gradient of amplitude with altitude is defined between the upper level and the first data level on which the amplitude is less than half the amplitude at the upper level, and this gradient is used to calculate a metric of altitude difference that would yield an exact halving of amplitude. This is similar to taking half the depth derived from the Schenzinger *et al.* (2017) Gaussian fit and, indeed, the multicycle mean value of vertical attenuation for ERA-Interim (Table 3; dZ) is close to half the Schenzinger *et al.* (2017) estimate of 15.1 km.

As for Schenzinger *et al.* (2017), a metric for QBO cycle width is defined as the *full width at half maximum* (FWHM) from a Gaussian fit to QBO amplitude between 20°N and 20°S (Table 3; W_{10} and W_{50}). Here the latitudinal profile of QBO amplitude (denoted by TText) is defined simply as half the difference in latitude profiles of zonal wind evaluated at times in each cycle when winds at the Equator are maximum (eastward) or minimum (westward) – Section 3.3 gives more detail. Employing a fit function greatly reduces the impact on the width metric of different latitude resolutions in the datasets as noted in Table 1. Width metrics evaluated at both 10 and 50 hPa are also used to calculate grades for Exp 1 models in Section 5.

3 | QBO CHARACTERIZATION

The typical structure and evolution of the QBO in each model are first examined by focusing on Exp 1. Comparison of Exp 1 and Exp 2, as well as consideration of variations between QBO cycles in a given model, is deferred until Section 4.

3.1 | Equatorial climatology

Before addressing QBO behaviour, the climatological mean state of the equatorial zone (5°S–5°N) in the models is considered as this can influence propagation of waves driving the QBO. Time means of zonal and monthly mean zonal wind and temperature are calculated for periods from 1979 to at latest 2009 for Exp 1 and for up to 100 years for Exp 2.

Climatological equatorial zonal mean winds in ERA-Interim are westward throughout the upper troposphere and stratosphere (Figure 1a, thick black line) and, on average, the models underpredict the strength of these winds at all levels in both Exp 1 (thick blue solid line) and Exp 2 (thick dark blue dashed line). This eastward bias with respect to ERA-Interim is common to most models, except in the region near the tropical tropopause where models generally have a westward bias. In contrast, the multimodel mean difference between the two experiments is considerably smaller than the bias, and the equivalent differences for each individual model are much smaller than the spread among models (compare solid and dashed coloured lines in Figure 1). This spread is particularly large above 20 hPa, where climatological winds in four of the models (ECHAM5sh, EMAC, LMDz6 and 60LCAM5) reach zero or even eastward. Below 30 hPa climatological wind biases are mostly smaller, although in relative terms still comparable to those above 20 hPa as the ERA-Interim wind is smaller at these altitudes. However, notable

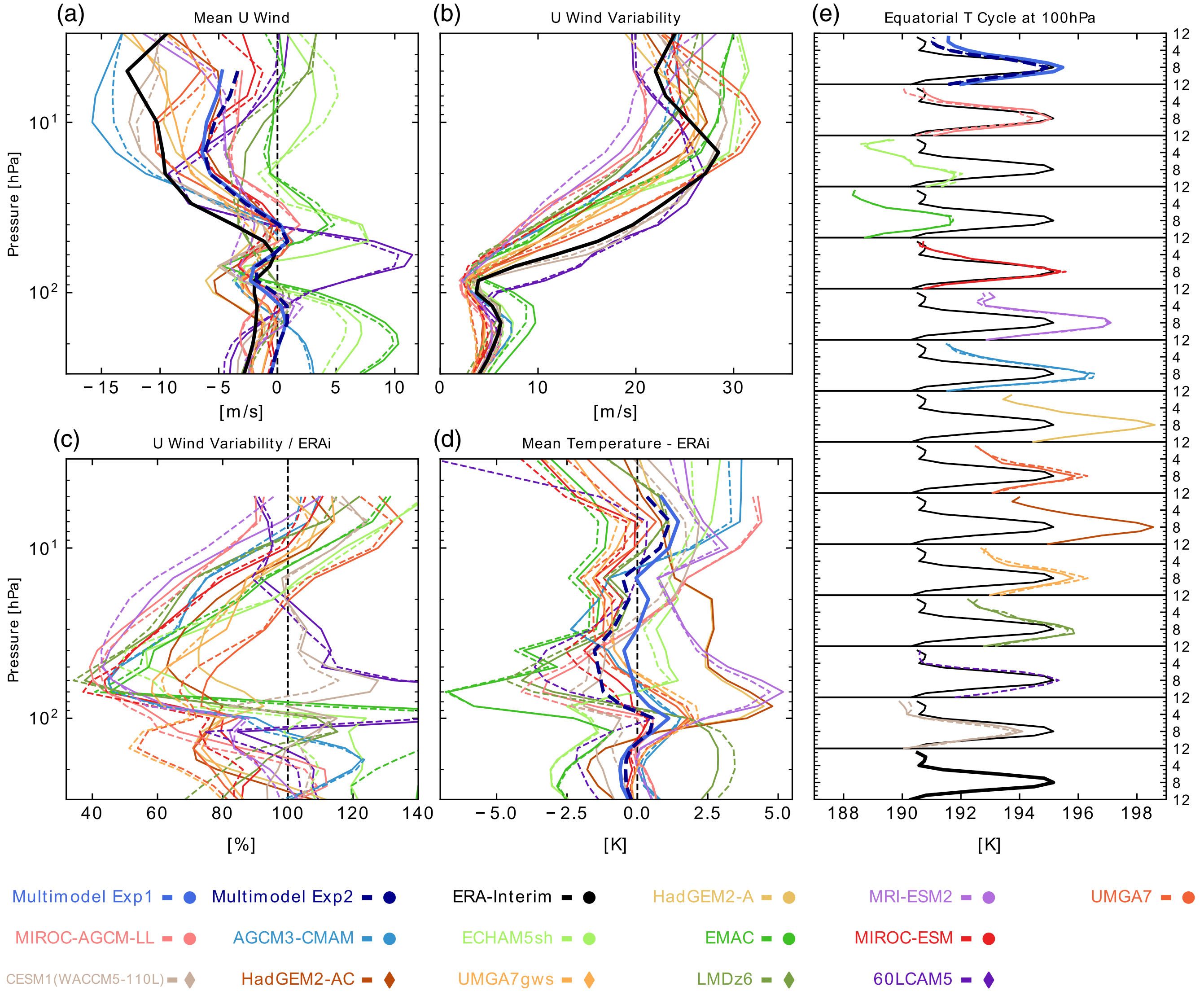


FIGURE 1 Vertical profiles of equatorial (5°S–5°N) zonal mean between 300 and 3 hPa for models in Exp 1 (solid coloured lines) and Exp 2 (dashed coloured lines), ERA-Interim (black) and multimodel means (dark blue solid, dashed) of: (a) climatological mean zonal wind (zero wind grey dashed), (b) standard deviation of zonal wind monthly timeseries $\times\sqrt{2}$, (c) is as (b) but as percentage relative to ERA-Interim, (d) climatological mean temperature difference from ERA-Interim (zero difference grey dashed), (e) mean seasonal cycle in the 100 hPa equatorial zone mean temperature plotted with ERA-Interim (thin black) for reference. Diamonds indicate models that parametrize the source of NOGWs

exceptions ECHAM5sh, EMAC, MIROC-AGCM-LL and 60LCAM5 all have clear eastward biases.

Vertical profiles of variability about the time mean zonal wind (Figure 1b) are defined as the standard deviation of $\bar{u}_{\mathrm{eq}}$ multiplied by $\sqrt{2}$ to aid comparison with the DD amplitude (2.2.3) profiles which are the deseasonalised counterpart of Figure 1b (see discussion in Section 4.2). The first generation of stratosphere-resolving GCMs generally underpredicted variability in the tropical stratosphere (e.g., figures 4 and 10 in Butchart and Austin, 1998; figure 9 in Manzini and Bengtsson, 1996). In contrast, the QBOi models have peak variability similar to or larger than that for ERA-Interim. However, as they generally peak at a higher altitude (~10 hPa compared with 15 hPa for ERA-Interim), models' variability is mostly larger than ERA-Interim at 10 hPa and above, and underpredicted by most models in the lower stratosphere (Figure 1b). The ratio of model variability to reanalysis variability (Figure 1c) shows that the largest disagreements, in relative terms, occur in the lower stratosphere between 80 and 30 hPa. In most models the variability at these altitudes is approximately half that seen in ERA-Interim, but there are two exceptions, 60LCAM5 and CESM1(WACCM5-110L), with larger variability.

As tropical tropopause temperatures are sensitive to many model processes, related not just to radiation but also to advection, cloud and ice microphysics, and diffusion (Hardiman *et al.*, 2015), there is a range of biases with respect to ERA-Interim in the QBOi models (Figure 1d). Nonetheless the magnitudes of these biases are not exceptional (e.g., figure 3 in Kim *et al.*, 2013), even for the notable outliers: EMAC with a significant cold bias, and HadGEM2-A, HadGEM2-AC and MRI-ESM2 with significant warm biases.

Both the mean temperature and amplitude of the mean seasonal cycle for the multimodel means (solid and dashed dark blue lines in Figure 1e) are in good agreement with ERA-Interim (black lines, repeated for reference). Good agreement in individual models, such as 60LCAM5, CESM1(WACCM5-110L), MIROC-AGCM-LL and MIROC-ESM suggests those models represent the extratropical large-scale waves reasonably well, as these waves drive the seasonal cycle of lower stratospheric equatorial temperatures through their impact on equatorial upwelling (Yulaeva *et al.*, 1994). Interestingly the three outlying models (HadGEM2-A, HadGEM2-AC and MRI-ESM2) with significant warm biases all have seasonal cycles of reasonable amplitude. On the other hand, UMGA7 and UMGA7gws which are closely related to HadGEM2-A and HadGEM2-AC, albeit with finer resolution, have reduced mean biases but seasonal cycles that are too weak. This illustrates that good simulations of the mean do not automatically guarantee a good simulation of the seasonal cycle and *viceversa*.

3.2 | QBO vertical structure

Timeseries of equatorial mean zonal wind as a function of pressure in the upper troposphere and stratosphere confirm the presence in all Exp 1 models of QBOs that are broadly similar to the QBO in ERA-Interim (Figure 2). Characteristic bands of alternating eastward and westward winds descend from around 2 hPa through much of the depth of the stratosphere and terminate as they approach the tropopause, especially in the westward phase. This occurs in all models, although there are individual differences in detail. For instance, the eastward mean biases seen in Figure 1a for 60LCAM5, ECHAM5sh, EMAC and LMDz6 are clearly evident in Figure 2.

Individual QBO cycles in Figure 2 are bounded by vertical black lines that indicate the transition from westward to eastward winds at 10 hPa (horizontal line). Mean cycles for Exp 1 model ensembles and ERA-Interim (Section 2.2.2) are shown in Figure 3, where numbers at the top right in each panel and the time range of the plotted region denote that model's mean QBO period. CESM1(WACCM5-110L) and LMDz6 have the longest mean periods and MIROC-AGCM-LL the shortest. In general the mean periods cluster around a multimodel mean of 25.3 months, which agrees reasonably well with the 27.8 month mean calculated for ERA-Interim (Table 2). This is perhaps not too surprising as the Exp 1 specification (Butchart *et al.*, 2018) was for models to be configured to produce their best simulation of the QBO.

Similarities and differences in the simulated QBOs are easier to identify in Figure 3, where averaging the QBO cycles removes the additional complication of intercycle variability that is present in Figure 2. All models in the figure clearly replicate the descent of the shear zones from the upper stratosphere to between 120 hPa and 70 hPa (100 hPa for ERA-Interim). In addition, most models reproduce the observed asymmetry between QBO phases, with the 10 hPa transition to the westward phase taking place less than half a period after the initial transition to the eastward phase (Table 2), and the eastward phase descends faster. However, there is quite a large spread in the phase asymmetry among models, especially with respect to the slowing descent of the westward shear zone when it reaches the lower stratosphere. This deceleration of the descent results in part from the mean meridional circulation induced by the QBO causing anomalously strong (weak) upwelling where there is westward (eastward) shear (Plumb and Bell, 1982). In addition, episodic stalling events, such as those seen in the observations when the descent of the westward shear halts briefly around 30 hPa before continuing downward (e.g., Yang and Yu, 2016), can also influence the slower descent of westward shear zones seen for mean QBO cycles in Figure 3. At 10 hPa, westward phases in the models are all stronger than the eastward phases, as seen for ERA-Interim (Table 3). Westward and eastward phases have more similar amplitudes at lower QBO altitudes, also as seen in ERA-Interim, although the eastward strength exceeds the westward in some models (e.g., ECHAM5sh, 60LCAM5, EMAC and AGCM3-CMAM).

Although the annual cycle of $\bar{u}_{\mathrm{eq}}$ is generally retained because the QBO is regarded not as a textbook harmonic oscillation but as a product of uneven eastward and westward wave forcings (Dunkerton, 2016), when the mean seasonal cycle is removed from each year of the original timeseries, QBO anomalies appear rather more regular with reductions in phase asymmetry and spread among the model amplitudes (compare Figures 3 and 4). Subtle differences that are more apparent in the deseasonalised QBO cycles include not just the amplitudes but the relative descent rates. Models with particularly slow descent rates (e.g., ECHAM5sh and MIROC-ESM, where the time for the QBO phase transitions to descend through the stratosphere exceeds their respective QBO phase durations) commence month 0 with three-cell vertical structures that

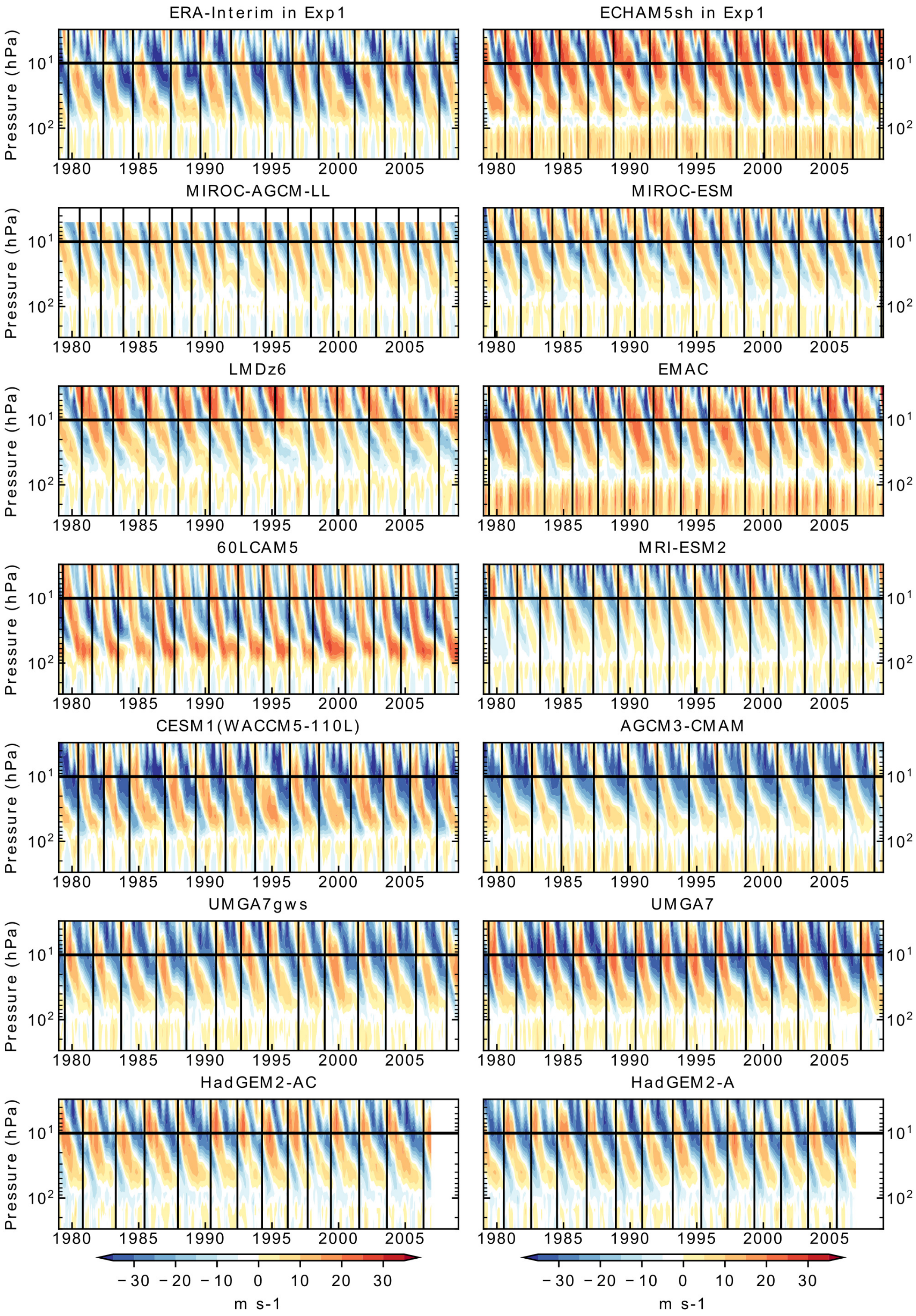


FIGURE 2 Pressure (300 to 3 hPa) versus time of equatorial (5°S–5°N) monthly and zonal mean zonal winds (colour shading) from 1979 to 2009 for Exp 1 models (first ensemble members only) and ERA-Interim. Black vertical lines indicate the times of transitions from westward to eastward winds at 10 hPa and define individual QBO cycles. Models with fixed source NOGWs are presented in the right column and, where possible, models which are to some degree related (MIROC-AGCM-LL, MIROC-ESM; 60LCAM5, CESM1(WACCM5-110L); UMGA7, UMGA7gws; HadGEM2-A, HadGEM2-AC) are placed adjacent to each other for easier comparison

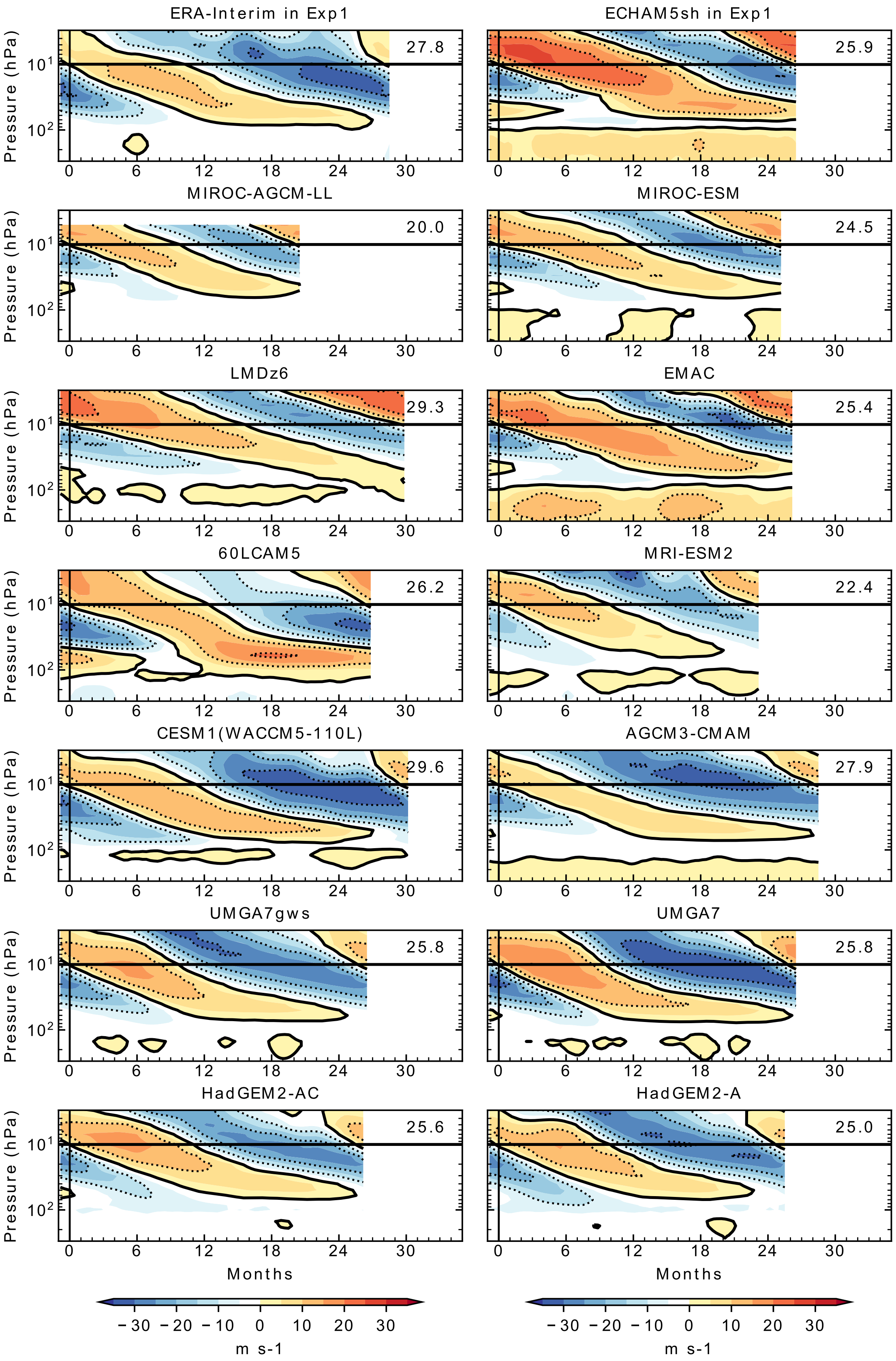
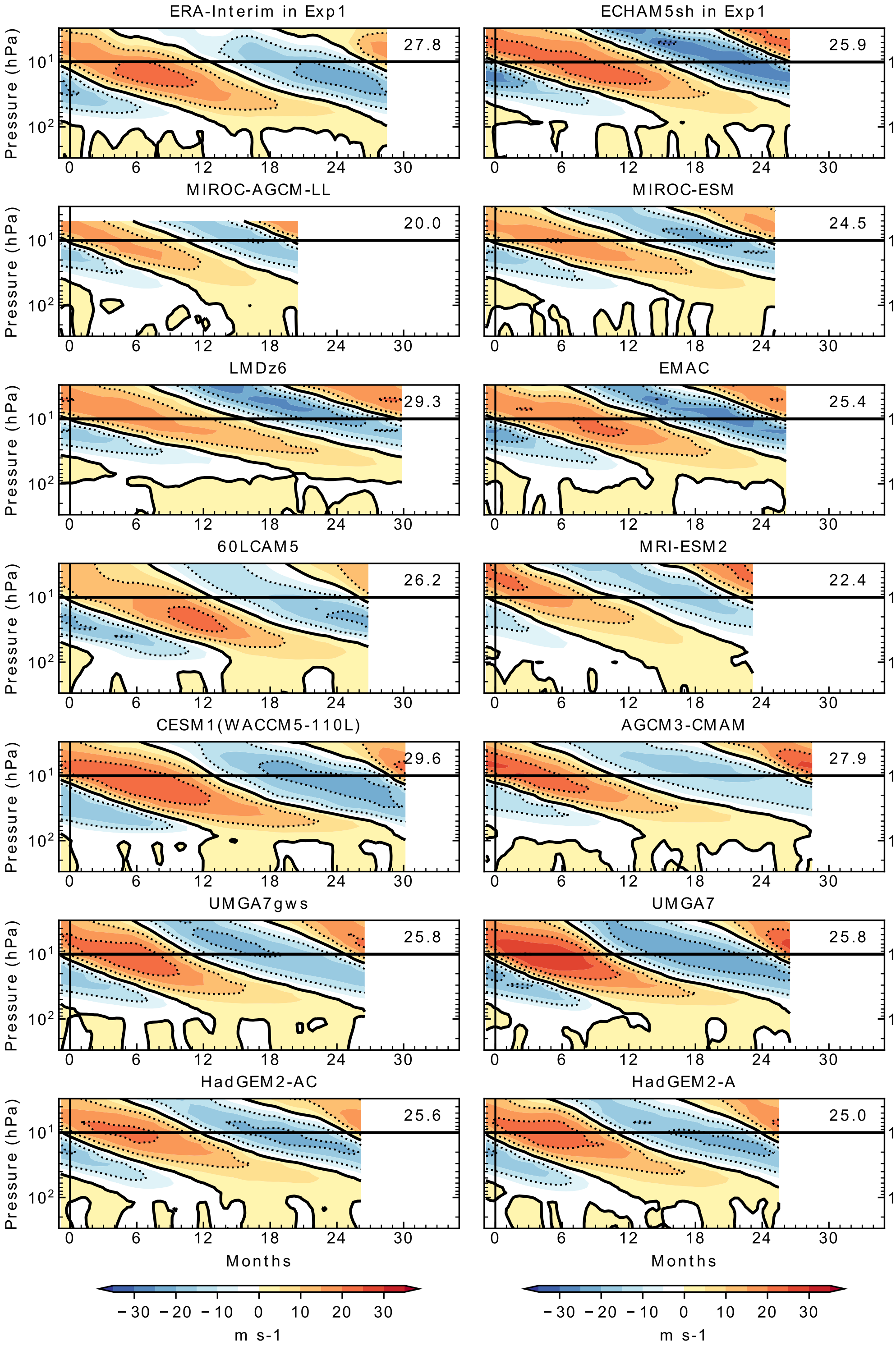


FIGURE 3 Pressure (300 to 3 hPa) versus time of mean cycles of equatorial (5°S–5°N) monthly and zonal mean zonal wind (colour shading) for Exp 1 ensembles and ERA-Interim, expressed as mean for individual model of QBO period in months defined by westward to eastward QBO wind transitions at 10 hPa, as seen in respective timeseries in Figure 2. Numbers at the top right in each panel and the time range of the plotted region denote that model's mean QBO period

FIGURE 4 As Figure 3, but with mean seasonal cycle removed



restrict the depth of the westward phase. Nonetheless, in all models the descent rates for deseasonalised westward and eastward shear zones are more uniform throughout the depth of the stratosphere than was seen in Figure 3.

Profiles of the mean QBO cycle (TTmceq method; Section 2.2.3) maxima (eastward phase) and minima (westward) at each pressure level in Figure 3, together with the total amplitude (0.5 × [|eastward| + |westward|]) and mean deseasonalised QBO cycles (Figure 5a–c, respectively), permit a quantitative comparison of QBO amplitudes across models. As with the standard deviation of monthly mean zonal wind shown in Figure 1b, mean QBO cycle amplitudes (Figure 5b) peak at higher altitudes than ERA-Interim (15 hPa) for most models, with the profile for the Exp 1 multimodel mean cycle (Section 2.2.2) peaking at 10 hPa (thick dark blue line). Only 60LCAM5 has

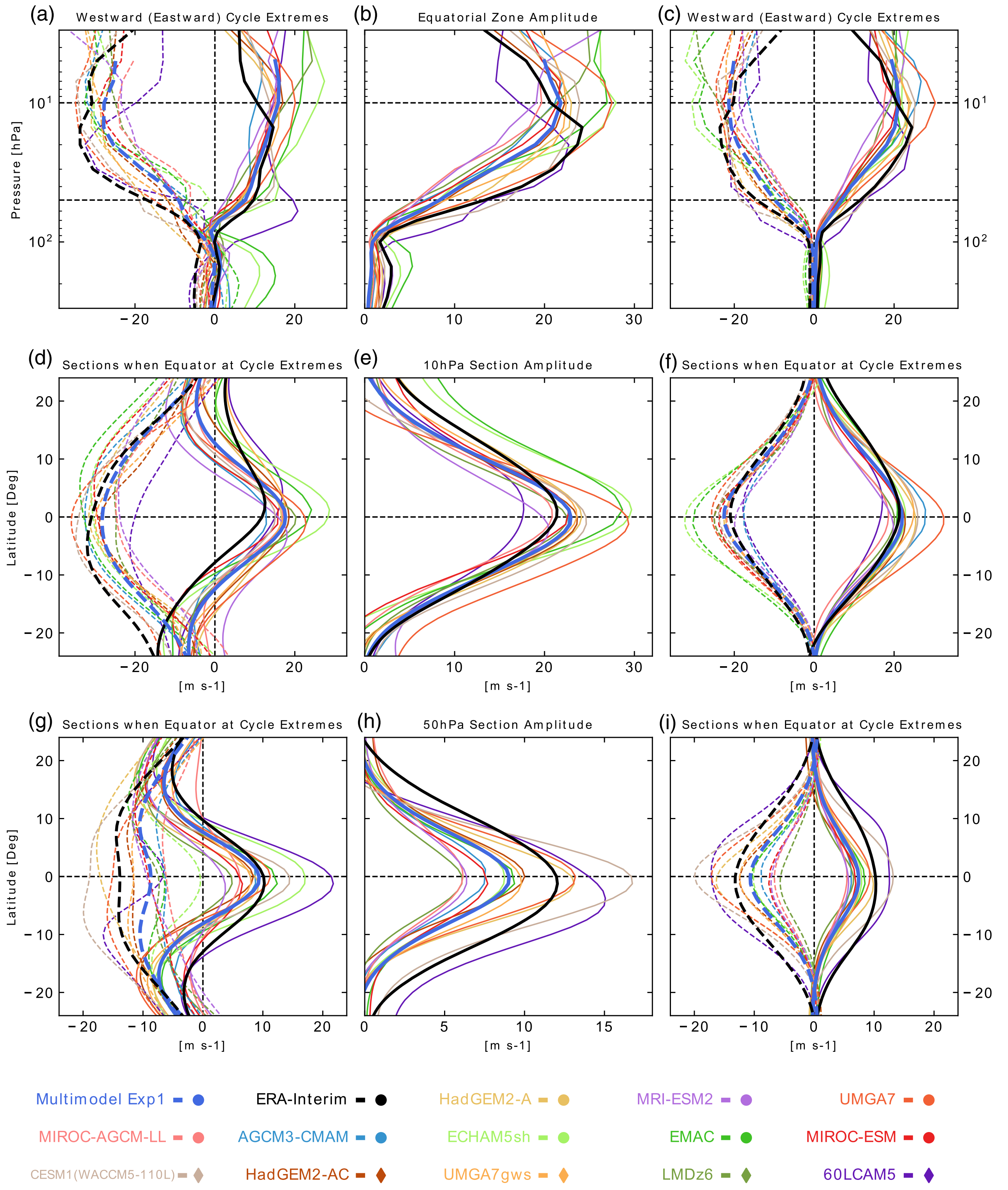


FIGURE 5 QBO zonal mean zonal wind for Exp 1 ensembles and ERA-Interim. (a, b, c) vertical profiles of equatorial wind amplitudes determined using method TTmceq of Section 2.2.3, with (a) QBO cycle maxima (eastward phases; solid) and minima (westward phases; dashed), and (b) QBO cycle total amplitudes. (c) is as (a), but for the deseasonalised wind. (d, e, f) meridional cross-sections for 10 hPa wind. (g, h, i) meridional cross-sections for 50 hPa wind. (d, g) show sections at instants when the cycle on the Equator reaches a maximum (eastward phases; solid) or minimum (westward phases; dashed). (e, h) show QBO cycle total amplitudes TText estimated from differences between sections in (d, g). (f, i) show sections as (d, g) but for the deseasonalised wind. In the legend, diamonds indicate models that parametrize the source of NOGWs

a maximum (at 20 hPa) below ERA-Interim. In general, model westward winds peak at higher speeds than the eastward winds (Figure 5a) consistent with ERA-Interim 10 hPa amplitude metrics (–33.8 m·s^{-1} versus 14.6 m·s^{-1}). However, the simulated westward phase throughout the stratosphere is nearly always too weak compared to ERA-Interim, with only CESM1(WACCM5-110L) and UMGA7 having westward amplitude metrics greater than ERA-Interim at 36.9 and 34.5 m·s^{-1}, respectively. Conversely, the simulated eastward phase is too strong above ~15 hPa in the model ensemble mean, and also too strong at lower altitudes in some models (Figure 5a).

Differences between Figure 5a and c provide a quantitative measure of how biases in the mean wind (Figure 1a) impact the phase asymmetry. This asymmetry is most evident for the largest mean eastward wind biases, those of 60LCAM5, ECHAM5sh, EMAC and LMDz6, that strengthen QBO eastward maximum winds (up to 25.8 m·s^{-1} for ECHAM5sh). In addition, the QBO again appears to vary rather more regularly with altitude in Figure 5c, with the mean wind structure removed, than in Figure 5a. In Figure 5c, both eastward and westward winds in nearly all models are too weak below the ERA-Interim maximum and too strong above, as expected from the total amplitude (Figure 5b). This is consistent with insufficient deposition of wave momentum in the lower and middle stratosphere, and too much momentum deposition higher up, which may (Section 6.1) be common to many models. If this is the case, errors in QBO amplitude could result even when total momentum fluxes entering the stratosphere are correct. MIROC-AGCM-LL has no parametrized NOGWs and has the smallest maximum amplitude (20.8 m·s^{-1}). Errors in ERA-Interim above 10 hPa may account for some of the discrepancies in the upper stratosphere.

3.3 | QBO latitudinal structure

Intermodel differences in QBO amplitude and width are clearly evident in latitude cross-sections of the mean QBO cycle evaluated at the 10 hPa reference level (Figure 6). Strong westward background winds at 10 hPa in most models, and also ERA-Interim (Figure 1a), are responsible for much greater prominence of the westward phase in Figure 6, while the eastward winds are more tightly confined to the equatorial zone, as seen in observations (Dunkerton and Delisi, 1985). The majority of models have wind maxima of both phases located either on the Equator or within one model grid spacing of it, as does ERA-Interim for its maximum eastward winds, whereas its maximum westward winds are at 4.5°S. Only two models exhibit maximum westward winds clearly south of the Equator, 60LCAM5 (3.3°S) and CESM1(WACCM5-110L) (7.0°S), and both have sources of parametrized NOGWs that depend explicitly on convective heating.

Figure 5d–f compare the latitudinal structure of the 10 hPa mean QBO cycle for ERA-Interim with the corresponding cross-sections, taken at times when the cycle on the Equator is maximum (eastward) or minimum (westward), for Exp 1 ensembles. The phase asymmetry noted previously is again evident in the multimodel mean, for which the cycle maxima only remain eastward in a latitude band roughly 12° either side of the Equator, whereas the westward phase has a range of roughly 24° either side of the Equator. When deseasonalised zonal wind anomalies are used, both the multimodel mean and ERA-Interim are highly symmetric about the Equator, with cross-sections that separate into eastward and westward anomalies of roughly equal magnitude and both phases have similar ranges of roughly 24° either side of the Equator (Figure 5f). Hence, the multimodel mean biases in Figure 5d can largely be attributed to biases in the mean seasonal cycle. Taking a difference between the profiles of maximum eastward and westward winds also removes the asymmetry (Figure 5e), which suggests that the impact largely arises from the climatological mean component. The QBO amplitude (TText; Section 2.2.4) evaluated from the profile differences (where positive) is for many models too large at the Equator and for most falls off too rapidly away from the Equator, such that widths of the simulated QBOs tend to be too narrow by up to 6 ° (negative biases in Table 3). Similar results are obtained at 50 hPa (Figure 5g–i) where, however, the simulated peak winds are relatively weaker than ERA-Interim and the widths narrower. Near the Equator, eastward biases in climatological winds relative to ERA-Interim at 10 hPa in most models (Figure 1) are clearly evident in Figure 5d,g and, for the westward phase, are visible across the full width of the QBO. In contrast, during the eastward phase, eastward biases generally become weaker toward the edge of the QBO and in some cases even turn westward, especially in the Northern Hemisphere, and hence are possibly related to the horizontal influence on the QBO from extratropical planetary waves (e.g., Osprey *et al.*, 2016).

3.4 | Multimodel mean QBO cycle

A multimodel mean cycle is calculated for Exp 1 (Figure 7a, shaded with dotted contours) with time axis normalized to the Exp 1 multimodel mean period (Section 2.2.2), as is that for ERA-Interim to allow direct comparison (Figure 7a, thick red overplotted contours). The multimodel mean and ERA-Interim show similar amplitude cycles but with the simulated winds displaced to higher levels over most of the descending cycle, such that

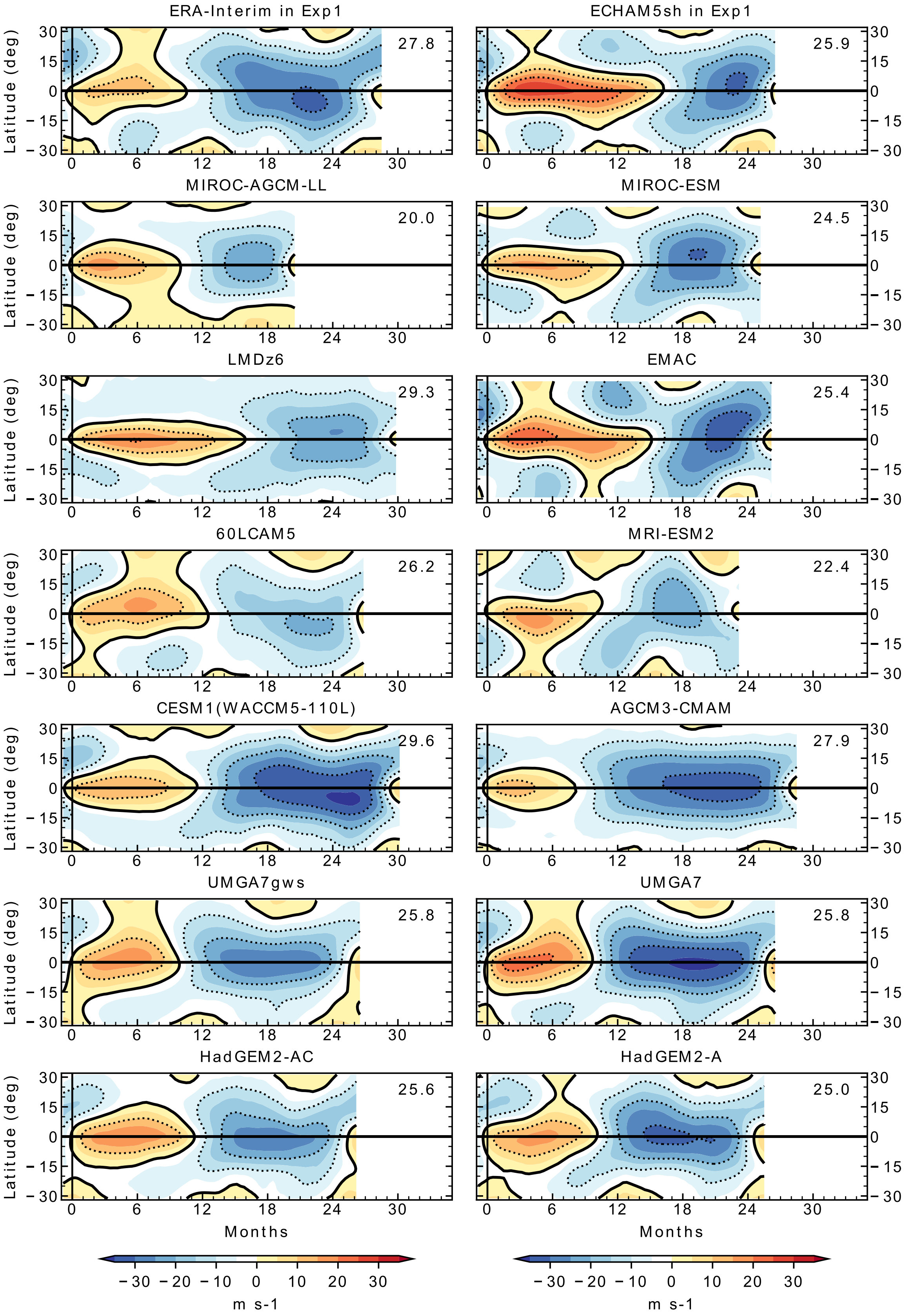


FIGURE 6 Latitude versus time mean cycles of 10 hPa monthly and zonal mean zonal wind for Exp 1 ensembles and ERA-Interim, expressed as mean for individual model of QBO period (months, colour shading) defined by westward to eastward QBO wind transitions at 10 hPa as in Figure 3

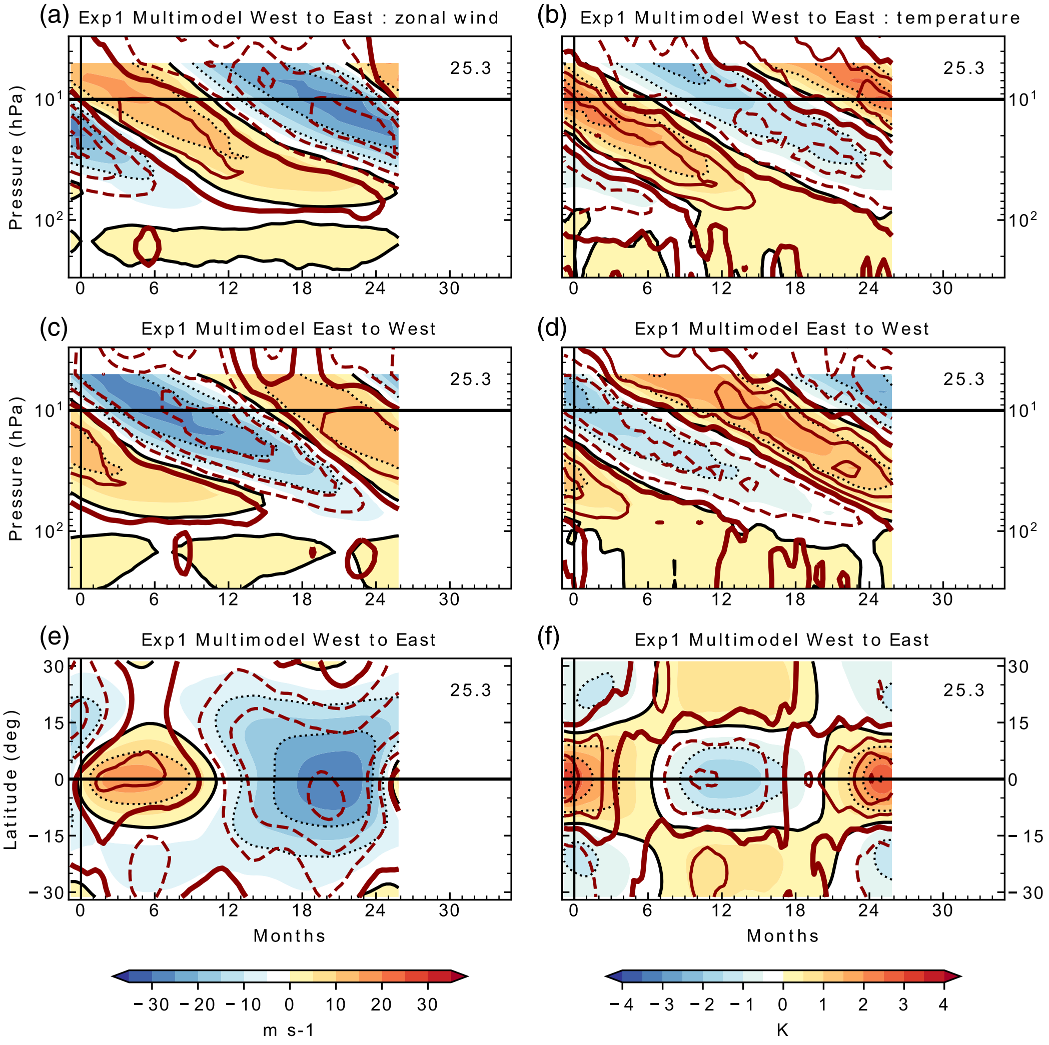


FIGURE 7 Multimodel mean of model QBO cycles in Exp 1 (colour shading, thin black contours) compared to the ERA-Interim mean cycle (bold red contours, with interval 10 m·s^{-1} for wind and 1 K for temperature) for the zonal and monthly mean of equatorial (5°S–5°N) (a, c) zonal wind, and (b, d) temperature anomalies as a function of pressure (300 to 3 hPa) versus time. (a, b) use westward to eastward transitions at 10 hPa, and (c, d) use eastward to westward transitions. (e, f) are as (a, b) but show latitudinal structure of the mean QBO cycles at 10 hPa versus time. The duration of all QBO cycles are scaled to the Exp 1 multimodel mean period to facilitate both the averaging and comparison.

the largest differences (not shown) are in the shear zones where gradients are strongest. As the individual cycles are aligned with the westward to eastward zonal wind transitions, and are generally not pure sinusoids, they are least coherent midway between transitions, which leads to broadened, reduced amplitude for the mean cycle toward its central time. However, for zonal winds this effect is small, as can be seen by a comparison with the mean cycles based on aligning eastward to westward transitions (Figure 7c). For Exp 2 the multimodel mean QBO cycle (not shown) is similar to that for Exp 1 with slightly elevated amplitudes around 10 hPa and reduced amplitudes toward the base of the descending eastward wind phase.

Multimodel means are also calculated for Exp 1 zonal mean deseasonalised temperature (Figure 7b,d, shaded with dotted contours) using the same transition times as for the zonal winds and compared with ERA-Interim (Figure 7b,d, thick red overplotted contours). Thermal wind balance in the Tropics ensures that the maximum temperature anomalies collocate with the strongest vertical wind shear. Hence the warmest anomaly at 10 hPa coincides with the westward to eastward wind transition, whereas the more pronounced westward wind maximum leads to the coldest anomaly lying slightly above 10 hPa at the eastward to westward wind transition (Figure 7d), which delays its arrival at 10 hPa. Though its amplitude diminishes in the lower stratosphere, the temperature QBO signal in both the multimodel mean and reanalysis propagates right down to the tropopause. As was the case for zonal wind, the multimodel mean temperature QBO is shifted upward relative to ERA-Interim. Results for Exp 2 are very similar and therefore are not shown.

The latitudinal distribution of the multimodel mean QBO cycle for zonal wind (Figure 7e) emphasizes the previously noted characteristics of maximum amplitude at the Equator and of models in general appearing rather more symmetric about it than ERA-Interim (red contours). However, as the ERA-Interim mean cycle is sensitive to the choice of transition (i.e., eastward to westward or westward to eastward) which is used to determine the mean cycle (not shown), this may partly be down to greater variability between cycles in the reanalysis. The equivalent multimodel mean cycle in temperature (Figure 7f) also reaches a maximum on the Equator, with the highest temperatures at the westward to eastward crossing and the lowest occurring slightly after the eastward to westward wind transition, with some indication of a delay of the positive temperature anomaly relative to that of ERA-Interim. The out-of-phase pattern of warm and cold anomalies in the subtropics is a consequence of the vertical component of the QBO-induced meridional circulation (Plumb and Bell, 1982) changing from upward to downward motion and vice versa roughly 20° poleward of the Equator.

4 | COMPARISON OF EXP 1 AND EXP 2

4.1 | Periods

Distributions of QBO periods from both Exp 1 and Exp 2 for each model are shown in Figure 8. All available Exp 1 data between 1979 and 2009 are used in order to provide the most accurate estimate of the period distribution for each individual model. In the distribution of all available periods from all ensemble members and all models (Figure 8a), models that provide more cycles, either by having more ensemble members or shorter mean periods, carry more weight. Exp 2 uses all ensemble members from models that repeat the 30-year format of Exp 1 but only uses the first 1×100 years ensemble member from models that choose an extended simulation length option allowed by the QBOi protocol. This limits only MIROC-ESM (for which 3×100 years of data are available whereas other models provide up to 100 years – Table 1) to prevent results from this model skewing the multimodel distribution. The period distributions are presented as percentages of the total number of QBO cycles in each histogram, and the multimodel distributions are repeated as outlines on each individual model plot for comparison. Also indicated on each panel are the mean period for each histogram (T) and the period derived from identifying QBO peaks in the Fourier spectrum (Tf) (as in Table 2). The difference in means of the multimodel distributions for Exp 1 and Exp 2 is not significant and much smaller than the bias of the Exp 1 mean against that of ERA-Interim, which is significant at the 5% level. However, intercycle variability in the periods for all models and for ERA-Interim is considerable to the extent that each individual model distribution has an overlap with that of ERA-Interim, and the ERA-Interim distribution lies within the multimodel distribution for Exp 1 (Figure 8a).

Models in the left column of Figure 8 either have no parametrized NOGWs (MIROC-AGCM-LL) or have variable sources of parametrized NOGWs, and these models all indicate a shorter mean QBO period in Exp 2 (grey) than in Exp 1 (orange): Table 4 shows that these differences are significant at the 5% level for all five variable source models. In contrast, no consistent difference is seen between Exp 1 and Exp 2 for the models with fixed sources of parametrized NOGWs (right column in Figure 8): MIROC-ESM, MRI-ESM2 and UMGA7 show little difference and, although ECHAM5sh and EMAC have longer mean periods in Exp 2 and AGCM3-CMAM has shorter mean period, these changes are not significant at the 5% level (Table 4). Of the eleven models that performed both Exp 1 and Exp 2, two models (UMGA7 and UMGA7gws) differ only in their specification of the source of parametrized NOGWs. With a fixed-source UMGA7 shows essentially no sensitivity of mean QBO period to the different boundary conditions and forcings in Exp 1 and Exp 2, while the variable-source UMGA7gws has a significantly shorter mean period in Exp 2.

The box–whisker plot (Figure 9a) shows that QBO periods in the models are mostly biased low compared to ERA-Interim: only AGCM3-CMAM, CESM1(WACCM5-110L) and LMDz6 have longer mean periods and then only for Exp 1. These negative biases are significant at the 5% level in both Exp 1 and Exp 2 for MIROC-AGCM-LL, MIROC-ESM, MRI-ESM2, and UMGA7gws, in Exp 1 for HadGEM2-A and UMGA7, and in Exp 2 for 60LCAM5 (Table 2). In contrast, no clear conclusions could be drawn for the variability of the QBO periods between models and experiments at least in terms of overall and 25–75% ranges, although many models have smaller 25–75%ile ranges than ERA-Interim. For the two pairs of models (HadGEM2-A, HadGEM2-AC and UMGA7, UMGA7gws) that differed only in their sources of parametrized NOGWs, only HadGEM2-AC showed increased QBO variability when variable sources were used (Table 2). For the other variable-source models, more QBO variability in Exp 1 than in Exp 2 is only seen in 60LCAM5 and CESM1(WACCM5-110L) (Table 4). All three of these models (60LCAM5, CESM1(WACCM5-110L) and HadGEM2-AC) use source parametrizations that are coupled directly to the model's

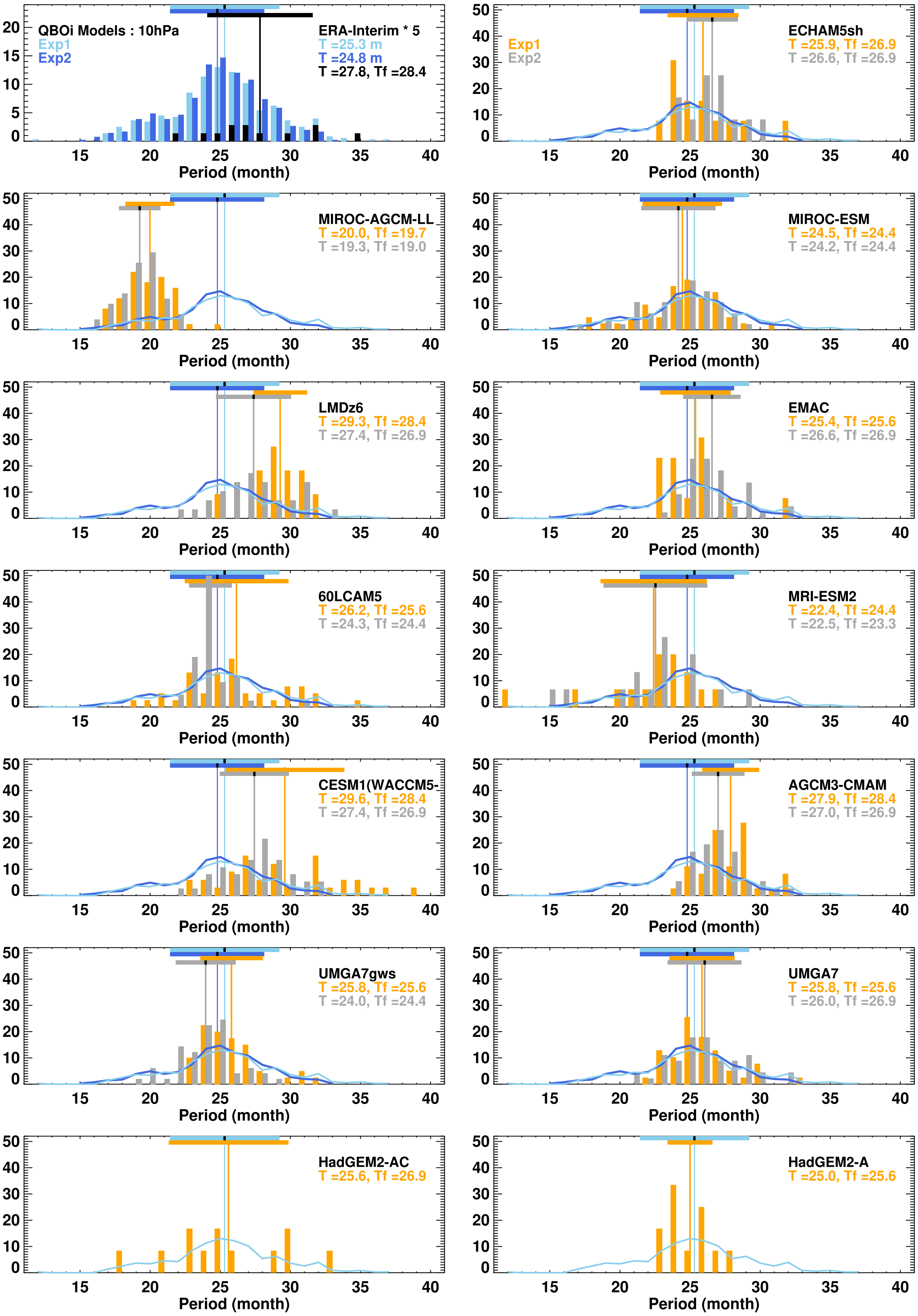


FIGURE 8 Histograms showing the distribution of QBO periods as a percentage of the total number of cycles in the distribution. The top left panel shows all periods simulated by all the model ensembles for Exp 1 (light blue), Exp 2 (dark blue) and ERA-Interim (black, scaled as Exp 1 ×5). Remaining panels show histograms for each individual model ensemble in Exp 1 (orange) and Exp 2 (grey) with the combined distributions from Exp 1 and Exp 2 models repeated as light and dark blue curves, respectively. Coloured vertical lines and horizontal bars indicate the mean, T, and standard deviation associated with the histogram of matching colour. Mean periods, T, and those inferred from the peak in the Fourier transform power spectrum, Tf, are indicated at top right on each panel

TABLE 4 Percentage differences between Exp 1 and Exp 2 for QBO period (Section 2.2.2) and amplitude (Section 2.2.3, TTeq) metrics evaluated at the 10 hPa reference level using the QBO transitions method (Section 2.2.1)

	Period mean			Period variability			Amplitude mean			Amplitude variability		
Model	**Full**	**Ewd**	**Wwd**	**Full**	**Ewd**	**Wwd**	**Full**	**Ewd**	**Wwd**	**Full**	**Ewd**	**Wwd**
60LCAM5	**8**	**17**	0	142	79	96	1	−3	3	−15	44	−23
AGCM3-CMAM	3	5	3	8	16	28	**3**	**8**	1	17	−14	29
CESM1(WACCM5-110L)	**7**	8	**6**	68	37	64	-0	−3	1	1	−1	6
ECHAM5sh	−2	−0	-6	39	−4	16	-0	−2	1	−21	79	−36
EMAC	-4	−4	−5	19	7	−4	1	2	1	−18	24	10
LMDz6	**7**	**16**	−2	−31	−14	−24	-0	3	−2	71	51	39
MIROC-AGCM-LL	**4**	6	1	18	27	−1	−1	1	−2	−1	−1	−2
MIROC-ESM	1	1	2	7	23	4	0	3	−1	20	26	13
MRI-ESM2	−1	9	−7	3	−3	2	**15**	19	12	−37	−42	−21
UMGA7	−1	**(−8)**	4	−12	1	−3	−1	−3	1	65	59	−15
UMGA7gws	**8**	1	**12**	4	7	5	0	2	−1	16	37	20
Multimodel	**2**	−1	**5**	16	3	18	−0	−1	1	−5	−5	−1

Note: Percentage difference is defined as $100 \times (x_1 - x_2)/x_2$ where x_1, x_2 are the values of metric x (means, standard deviations of periods and amplitudes) for Exp 1 and Exp 2 respectively. Values in bold indicate that mean $(x_1 - x_2)$ values are significant at the 5% level, with (..) identifying negative differences.

convective heating. The other two variable-source models (LMDz6 and UMGA7gws) use precipitation as a proxy for convective heating and show no significant change in QBO variability between Exp 1 and Exp 2. Likewise UMGA7 and UMGA7gws exhibit little difference in QBO variability when variable source (UMGA7gws) is used in place of fixed source (UMGA7). Thus it appears from these limited results that the impact of NOGW source parametrizations on QBO period is likely to depend on details of the source parametrization employed. Note that in Table 4 the larger standard deviation of QBO periods in Exp 1 than in Exp 2 in all but two models is consistent with the additional variability in SSTs leading to added variability in the QBO periods.

A complication in interpreting differences between Exp 1 and Exp 2 is that the sea-surface temperatures (SSTs) recommended for each experiment have different climatologies. As described in Butchart *et al.*, 2018 (2018, appendix A), climatological SSTs provided for Exp 1 cover the period 1979–2009, but those used when calculating the climatological annual cycle of SST for Exp 2 cover 1988–2007. Hence, Exp 2 has a slightly warmer SST climatology than the climatology for Exp 1 due to exclusion of the cooler 1979–1987 period (not shown). A warmer SST climatology could lead to increased tropical wave activity impacting the tropical stratosphere and increased parametrized wave activity in the variable-source models, which might be expected to reduce mean QBO periods in Exp 2 relative to Exp 1. However, the earlier result that all models with variable sources have shorter mean QBO periods in Exp 2 than in Exp 1 was unaffected by recalculation of the Exp 1 period histograms using only the 1989–2009 period (not shown). This suggests that shorter Exp 2 mean QBO periods in variable-source models result instead from differences in SST variability between Exp 1 and Exp 2, although the mechanism for this is unclear.

4.2 | Amplitudes

QBO amplitudes obtained at 10 hPa from individual QBO cycles (as described in Section 2.2.3) are summarized in Figure 9b. In contrast to the overall negative bias seen in QBO period, there is no clear systematic bias relative to ERA-Interim in the amplitudes, though amplitude variability is smaller in many models than that in ERA-Interim (Table 3). Across the models there is no systematic difference between Exp 1 and Exp 2 in either the mean or standard deviation of 10 hPa cycle amplitudes (Table 4), and for all but two models the mean differences are not statistically significant.

To check for QBO amplitude differences between Exp 1 and Exp 2 at other altitudes besides 10 hPa, Figure 10c displays (Exp 1 – Exp 2) differences in the vertical structure of QBO amplitude and compares these differences to the Exp 1 bias from ERA-Interim (Figure 10b). To test how robust the results are to choice of amplitude metric, Figure 10a shows amplitudes calculated from

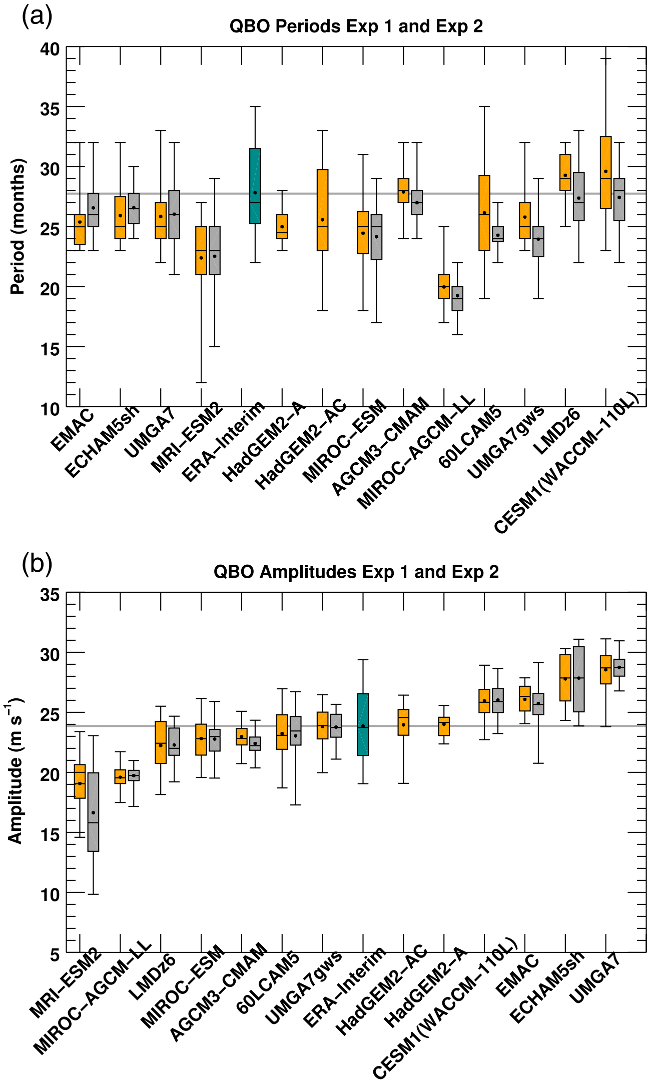


FIGURE 9 Summary of 10 hPa equatorial (5°S–5°N) monthly and zonal mean zonal wind QBO (a) period and (b) total amplitude distribution statistics, with maximum–minimum range (whisker) and 25th to 75th percentile range (box) for ERA-Interim (blue-green), Exp 1 models (orange) and Exp 2 models (grey). Periods are ranked in order of increasing Exp 1-Exp 2 mean (dot) values, and amplitudes are ranked in order of increasing Exp 1 mean values. Also indicated are the distribution medians (horizontal lines)

deseasonalised timeseries (method DD: Section 2.2.3). These are very similar to the standard deviation of the full timeseries (Figure 1b), confirming that QBO variability in zonal mean zonal wind dominates over the seasonal cycle in the equatorial middle stratosphere, with relatively small additional contributions from the annual oscillation (AO) and semiannual oscillation (SAO), though the latter starts to dominate toward the stratopause, consistent with current knowledge (figure 30 in Baldwin *et al.*, 2001).

As already noted in Section 3 for the TTmceq amplitude method (Figure 5b), the DD metric peak amplitudes (Figure 10a) occur near 10 hPa for the models and at 15 hPa for ERA-Interim, with similar spread between models. The DD amplitude peak for ERA-Interim (at 15 hPa) is enhanced by 11% relative to its amplitude at 10 hPa. This is around double the difference between the equivalent ERA-Interim peak (20 hPa) amplitude from Schenzinger *et al.* (2017) and the ERA-Interim 10 hPa TTeq amplitude here, which suggests that level choices may be the biggest factor separating the two metrics. The choice of amplitude method does not affect the vertical structure of Exp 1 biases: most models have low amplitude relative to ERA-Interim below 10 hPa and relatively high amplitude above, although 60LCAM5 and CESM1(WACCM5-110L) are exceptions displaying the opposite pattern (Figure 10b). Dashed lines in the figure suggest that models with variable sources of parametrized NOGWs have smaller biases overall, though given the small ensemble size this result is not conclusive. Exp 1 - Exp 2 differences for individual models (Figure 10c) are much smaller than the Exp 1 biases throughout the lower stratosphere (also for the TTmceq method metric, not shown). At altitudes below 10 hPa, amplitudes in Exp 1 are slightly larger than in Exp 2, though typically by less than 2 m·s^{-1} (Figure 10c).

5 | GRADING OF QBO METRICS

In order to provide a holistic assessment of model performance across a range of metrics, Waugh and Eyring (2008) proposed a measure of skill or "grading" of a metric based on the model bias against observations scaled by an estimate of the observed variability (typically taken as the standard deviation). For a metric from a model with mean m and standard deviation σ, this non-dimensional grade is

$$g = 1 - \frac{|m - m_{\mathrm{obs}}|}{3\sigma_{\mathrm{obs}}}.$$

Grades vary from 1 when there is perfect agreement between model and observations, down to an imposed cut-off at zero when the magnitude of the bias exceeds $3\sigma_{\mathrm{obs}}$ (i.e., the model mean m lies outside 99.7% of a normal distribution about the observed mean).

Grades for a range of metrics representing the period, amplitude and structure of the QBO in Exp 1 are presented in Figure 11. Each box depicts the grade (number and shading) for a metric (column) evaluated for a given model (row). Numbers in white indicate agreement between model and ERA-Interim within the 95% confidence level, based on Student's t-test. A metric can have the same grade from two models but different statistical

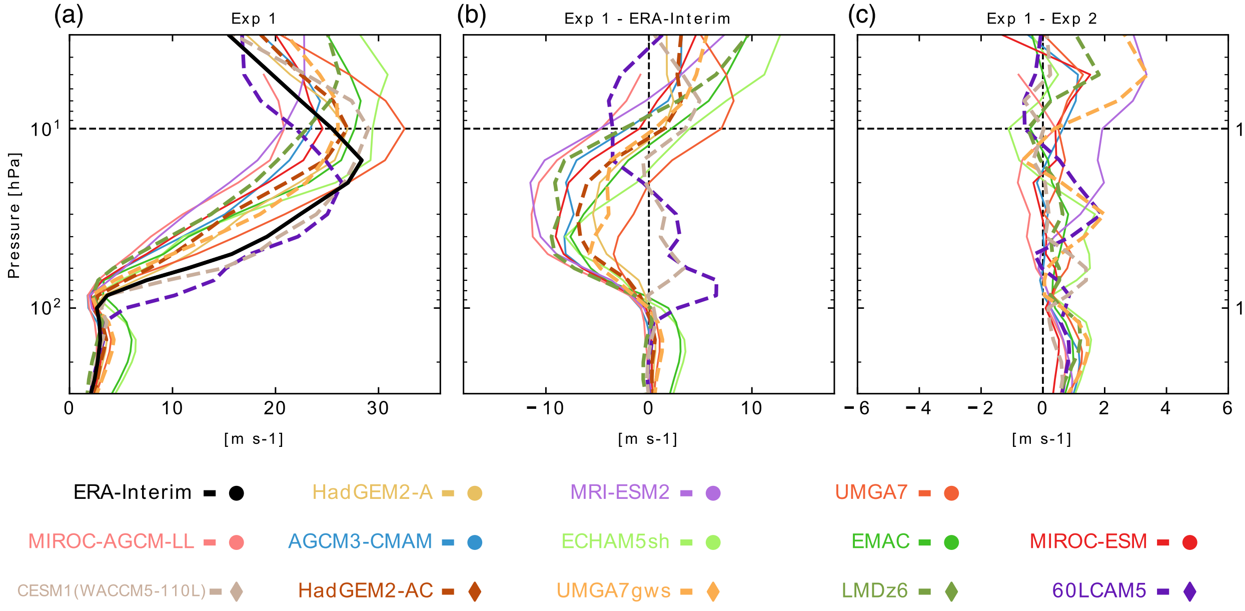
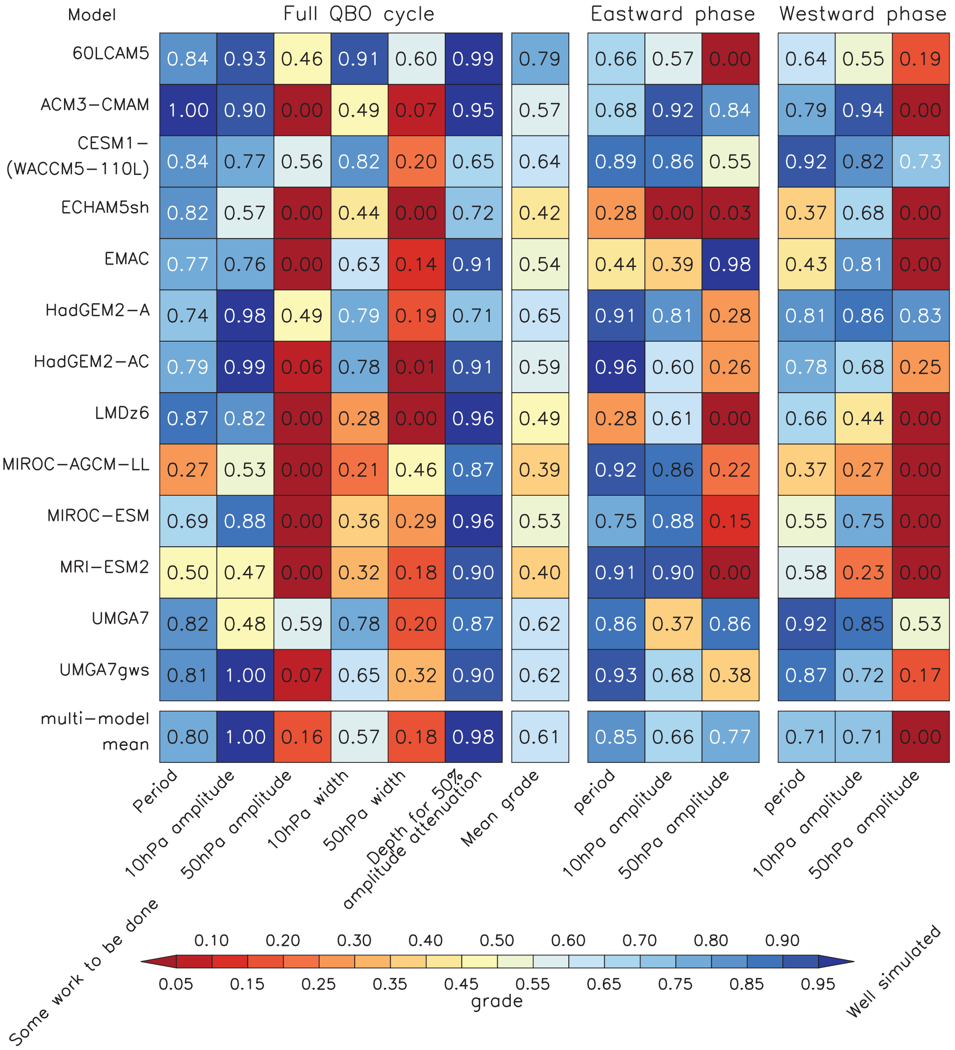


FIGURE 10 QBO amplitude versus pressure (300 hPa to 3 hPa) vertical profiles for (a) Exp 1 model first ensemble members and ERA-Interim, (b) Exp 1 models minus ERA-Interim, (c) Exp 1 minus respective Exp 2 models, where amplitudes are calculated using method DD (Section 2.2.3). Solid and dashed lines indicate models using fixed and variable parametrized NOGW sources, respectively. (MIROC-AGCM-LL, which has no parametrized NOGW, is shown in solid.)

significance because the *t*-test takes account of intercycle spread in the simulated metric, which is absent in the grades. Hence some grades in a fairly narrow range may show agreement despite better performing grades (i.e., closer to one) showing significant bias, or *viceversa*. For the multimodel mean, statistical significance is based on a single sample *t*-test for the difference of the ERA-Interim mean metric (i.e., the value obtained by averaging over QBO cycles) from the mean of the multimodel distribution of mean metrics. Only the 50 hPa metrics showed more intermodel variability (i.e., larger standard deviation) in the mean metrics than the intercycle variability in the corresponding ERA-Interim metric (not shown).

With few exceptions, Figure 11 demonstrates a clear tendency for model performance to polarize when simulating the selected metrics, with most columns dominated either by blue or by red. Where metrics are generally well simulated by the majority of models (blue), on average the multimodel mean performs better than the individual models (i.e., the grade for the multimodel mean is higher than the average grade for the models, suggesting random biases). Likewise, where some work needs to be done to improve a metric (red), this applies to a majority of the models. The greatest spread in skill across the models is in the simulation of the 10 hPa widths and to a lesser extent in the individual phase durations and 10 hPa amplitudes for the westward phase.

Good grades (≥ 0.5) would be expected for the QBO period if modellers have adjusted the total launch momentum flux in their NOGW parametrizations to obtain realistic QBO periods. The fact that the only model with no NOGW parametrization (MIROC-AGCM-LL) has the poorest period grade is consistent with this assumption. Based on metrics (Period; Amplitude (h_{max}) easterly, westerly and total) presented in table 5 of Schenzinger *et al.* (2017) for a subset of CMIP5 models that excluded the worst performers, we estimate a four-grade mean of 0.84. Grades for corresponding 10 hPa metrics from the multimodel mean data in Figure 11 give a mean grade of 0.79, indicating that on average QBOi AGCMs have comparable performance with the better CMIP5 models for this somewhat selective subset of grades. Why models tend to perform similarly across the rest of the metrics is less clear, though it strongly suggests that poor performance (e.g., for 50 hPa amplitude and width) is due to models sharing similar deficiencies. Given generally good grades for 10 hPa amplitudes, but grades for the 50 hPa amplitudes that are zero or very close to zero in all but four models, a challenging question for future activity is whether more than simple parameter adjustments in the participating models would be needed to improve performance at 50 hPa without simultaneously degrading the good performance at 10 hPa. One issue might be an over-reliance in nearly all models on parametrized NOGWs which, in turn, may deposit momentum too high up (Section 6 gives a breakdown of wave driving between resolved and parametrized waves.). Approximately half the models have a mean grade greater than 0.5 and half



Model	Full QBO cycle							Eastward phase			Westward phase		
	Period	10hPa amplitude	50hPa amplitude	10hPa width	50hPa width	Depth for 50% amplitude attenuation	Mean grade	period	10hPa amplitude	50hPa amplitude	period	10hPa amplitude	50hPa amplitude
60LCAM5	0.84	0.93	0.46	0.91	0.60	0.99	0.79	0.66	0.57	0.00	0.64	0.55	0.19
ACM3−CMAM	1.00	0.90	0.00	0.49	0.07	0.95	0.57	0.68	0.92	0.84	0.79	0.94	0.00
CESM1−(WACCM5−110L)	0.84	0.77	0.56	0.82	0.20	0.65	0.64	0.89	0.86	0.55	0.92	0.82	0.73
ECHAM5sh	0.82	0.57	0.00	0.44	0.00	0.72	0.42	0.28	0.00	0.03	0.37	0.68	0.00
EMAC	0.77	0.76	0.00	0.63	0.14	0.91	0.54	0.44	0.39	0.98	0.43	0.81	0.00
HadGEM2−A	0.74	0.98	0.49	0.79	0.19	0.71	0.65	0.91	0.81	0.28	0.81	0.86	0.83
HadGEM2−AC	0.79	0.99	0.06	0.78	0.01	0.91	0.59	0.96	0.60	0.26	0.78	0.68	0.25
LMDz6	0.87	0.82	0.00	0.28	0.00	0.96	0.49	0.28	0.61	0.00	0.66	0.44	0.00
MIROC−AGCM−LL	0.27	0.53	0.00	0.21	0.46	0.87	0.39	0.92	0.86	0.22	0.37	0.27	0.00
MIROC−ESM	0.69	0.88	0.00	0.36	0.29	0.96	0.53	0.75	0.88	0.15	0.55	0.75	0.00
MRI−ESM2	0.50	0.47	0.00	0.32	0.18	0.90	0.40	0.91	0.90	0.00	0.58	0.23	0.00
UMGA7	0.82	0.48	0.59	0.78	0.20	0.87	0.62	0.86	0.37	0.86	0.92	0.85	0.53
UMGA7gws	0.81	1.00	0.07	0.65	0.32	0.90	0.62	0.93	0.68	0.38	0.87	0.72	0.17
multi−model mean	0.80	1.00	0.16	0.57	0.18	0.98	0.61	0.85	0.66	0.77	0.71	0.71	0.00

FIGURE 11 Quantitative measure of performance or *grade* for the comparison of QBO metrics between models and ERA-Interim for Exp 1. Values of the normalized grade, defined in Section 5, are given by numbers in the boxes and corresponding shading. A value of 1 and dark dark blue shading for a specific metric (column) indicates exact agreement of a given model (row) with ERA-Interim, while numbers in white indicate agreement between the model and ERA-Interim within the 95% confidence level. Numbers close to zero and red shading indicate poor agreement between a model and ERA-Interim: zero indicates that the magnitude of the model bias with respect to ERA-Interim is more than three times the standard deviation obtained from ERA-Interim for that metric. For the bottom row, the multimodel mean shows the grades for the mean of mean metrics from individual models: white indicates that the multimodel mean of the mean metrics agrees with the ERA-Interim mean metric with 95% confidence according to a single-sample two-sided *t*-test.

have a mean grade less than 0.5, while the corresponding mean of grades for the multimodel mean is 0.61. This variation in mean grades mostly reflects the models' gradings for the most variable metric, QBO width at 10 hPa.

High grades for attenuation depth suggest that models are better at representing the QBO's vertical structure than its latitudinal structure. However, this is only partly true: although the vertical attenuation of QBO amplitude is realistic, QBOs in the models are mostly shifted upward compared to reanalysis. Hence, in Figure 5b the agreement between models and reanalysis for 10 hPa amplitude appears to be improved because, although the models' peak amplitudes are typically smaller than the reanalysis peak amplitude, they mostly occur at higher altitudes where the reanalysis off-peak amplitude happens to have a similar value to the models' peak amplitudes. Thus, even with realistic vertical attenuation of amplitude, a model can have a good grade for 10 hPa amplitude but a poor one for 50 hPa amplitude.

In general, models are better at simulating the eastward phase than the westward phase (Figure 11, right-hand columns). The multimodel mean grades (bottom row) show that partial period durations of both eastward and westward phases are represented well, albeit slightly better for the eastward phase. Overall, 10 hPa amplitudes for both phases are well represented, though for the westward phase several models have grades less than 0.5, all of which had low grades for 10 hPa widths and the strongest correlation of 10 hPa width with other metrics (0.64) is with the westward phase amplitude. This suggests that deficiencies in simulating 10 hPa width may be due to deficiencies in simulating the westward phase, which possibly relate to poor representation of mixed Rossby-gravity waves by the models (Holt *et al.*, 2020). Interestingly the multimodel mean does rather well at simulating the eastward phase amplitude at 50 hPa, despite over half the models having grades below 0.4 for this metric, which is due to a range of both positive and negative model biases against ERA-Interim.

The corresponding 50 hPa westward phase amplitudes are poorly simulated across the ensemble, apart from three models, two of which – CESM1(WACCM5-110L) and HadGEM2-A – reproduced the amplitudes rather well, while the remainder have amplitudes biased systematically low relative to ERA-Interim.

In summary, the models have a lot in common in terms of performance in representing the QBO with shared strengths and weaknesses across a majority of models. While the range of metrics used here is clearly not exhaustive, the common deficiencies already identified for this current choice of metrics provide compelling motivation for further investigation and model development in these areas.

6 | QBO WAVE FORCING

As the results of Sections 4 and 5 suggest the need for a more detailed investigation into the differences in wave driving between models, the relative contributions of resolved and parametrized wave forcing to the QBO are briefly explored. Vertical structure and relative contributions of resolved and parametrized wave forcing are examined in Section 6.1, followed in Section 6.2 by a comparison between models using parametrized NOGWs against the single model in the ensemble (MIROC-AGCM-LL) that does not use parametrized NOGWs.

6.1 | Relative contributions by resolved and parametrized wave forcing

Figure 12 shows for Exp 1 composite zonal wind and wave forcing profiles for the first month in each QBO cycle when the phase is eastward (a,b,c) and westward (d,e,f). As described in Section 2.2.1, the phases are defined by transitions of the mean zonal wind $\bar{u}_{\text{eq}}$ at the 10 hPa reference level. Accelerations due to resolved waves, that is, divergence of the Eliassen–Palm (EP) flux, and parametrized NOGWs are shown in Figure 12(b,e) and (c,f), respectively. Westward acceleration near the tropopause in Figure 12b,e almost certainly results from the horizontal component of the EP flux (e.g., Yoshida *et al.*, 2018) and not the vertically propagating waves that contribute to driving the QBO. However, this persistent westward forcing throughout the QBO cycles affects the climatological background zonal wind, which in turn can impact on the vertically propagating waves and also the parametrized NOGWs.

Accelerations from both the resolved and parametrized forcing peak, on average, at roughly the level of zero zonal wind for the eastward onsets (Figure 12b,c), and just above the level of zero wind for westward onsets (Figure 12e,f). The most likely cause of this effect is different filtering of eastward and westward waves below 10 hPa due to the phase asymmetry in QBO winds noted in Section 3. In particular, in the mid-stratosphere the westward winds in Figure 12a peak at 30 m·s^{-1}, roughly three times the peak eastward wind speed in Figure 12d, while near the tropopause winds are, on average, westward in both panels (indicative of divergence of the horizontal EP-flux component noted above). A consequence of these differences is that more westward waves with low phase speeds are filtered or damped below 10 hPa, during the onset of the westward phase, than is the case for eastward waves during the onset of the eastward phase. Hence, during the westward onset, westward waves with higher zonal phase speeds are more dominant. As critical levels for these faster-moving waves are where their westward phase speeds match the zonal wind, this will tend to shift wave forcing to above 10 hPa, further into the region of westward winds (Figure 12d). A similar contrast between westward and eastward peak forcing was found when the models' NOGW parametrizations were used offline to calculate forcings for opposing QBO phases (Butchart *et al.*, 2018, figures 7b,e). Again the peak forcing for the westward onset occurred, on average, above the zero zonal wind level, whereas for the eastward QBO phase the peak forcing was closer to the level of zero wind. The offline calculations all specified the same momentum flux at 100 hPa and, as their background wind profiles did not evolve, the contrasting behaviour can be directly attributed to differences in the filtering by the two wind profiles.

In general, the peak accelerations due to NOGWs are larger than those due to resolved waves, for both QBO phases. The one model without parametrized NOGWs, MIROC-AGCM-LL, has peak values of resolved wave forcing much larger than those in the other models, and similar in magnitude to the peak NOGW forcing seen in most models (≈ 0.4 m·s^{-1}·day^{-1}; Figure 12). There are large variations in NOGW forcing across models, with peak values ranging from 0.2 to 0.6 m·s^{-1}·day^{-1} for eastward onsets and from –0.4 to –0.9 m·s^{-1}·day^{-1} for westward onsets. This intermodel spread is similar to that obtained from the offline calculations with specified momentum fluxes at 100 hPa that were the same for each parametrization (figures 7b,e of Butchart *et al.*, 2018), though overall peak values are roughly a factor of two smaller in Figure 12b,c, which show zonal and monthly averages. When the launch fluxes and launch altitudes from their respective GCMs were used in the offline calculations, larger intermodel spread was seen (figure 7c,f of Butchart *et al.*, 2018). Differences between Exp 1 models in the wind profiles through which the NOGWs propagate (Figure 12a,d) are also likely to contribute to intermodel variations in NOGW forcing. Conversely, adjusting NOGW schemes to

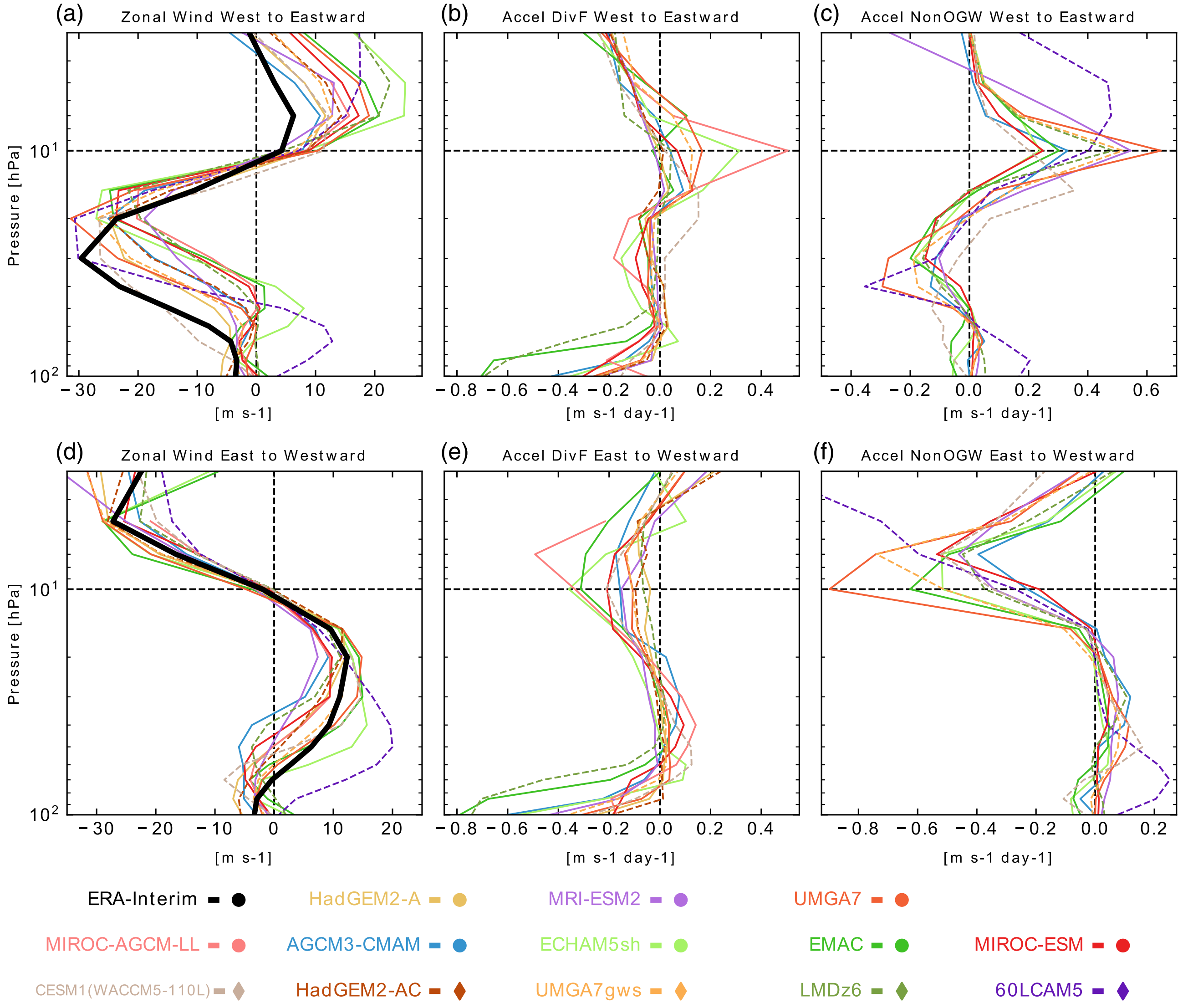


FIGURE 12 Composited equatorial (5°S–5°N) monthly and zonal mean zonal wind and wave forcing profiles (100 hPa to 3 hPa) from Exp 1 model first ensemble members and ERA-Interim for the first month after (a, b, c) 10 hPa westward to eastward QBO wind transitions and (d, e, f) eastward to westward transitions. (a, d) zonal wind, (b, e) mean-flow accelerations due to resolved waves, and (c, f) mean-flow accelerations due to parametrized NOGWs. Due to data availability issues, (b, e) are missing the 60LCAM5 model and (c, f) are missing HadGEM2-A and HadGEM2-AC. Solid and dashed lines indicate fixed and variable NOGW sources as in Figure 10. (Note the different horizontal scales in the panels.)

drive more realistic QBOs in models can lead to an artificial reduction of intermodel spread (i.e., models may be overtuned) which most likely accounts for the absence of stronger differences between models in Figure 12c,f.

The same method that was applied to create Figure 5b amplitude profiles of $\bar{u}_{\text{eq}}$ from Figure 3 mean QBO cycles defined by westward to eastward phase transitions (TTm-ceq from Section 2.2.3) was used to process the forcings by resolved waves and NOGWs. Forcing by NOGWs is larger than forcing by resolved waves at all QBO altitudes (Figure 13a,b) and both are weaker in the multimodel mean than individual models, consistent with spread among models in timing of forcing events relative to the QBO cycle. For most models the relative contribution to the total forcing from NOGWs (Figure 13c) exceeds 50% above ≈70 hPa, and does not vary strongly with altitude. Below 70 hPa the small relative size of NOGW forcing is due to large resolved wave forcing near the tropopause (Figure 13a) which, as noted above, is unlikely to result from vertically propagating waves. The amplitude of both resolved and parametrized wave forcing increases rapidly with altitude (Figure 13a,b) with the NOGW forcing growing slightly more rapidly in most models. On average 80% of the total forcing at 10 hPa is from NOGWs compared to 70% at 70 hPa, though in some models, such as MIROC-ESM, the change in relative

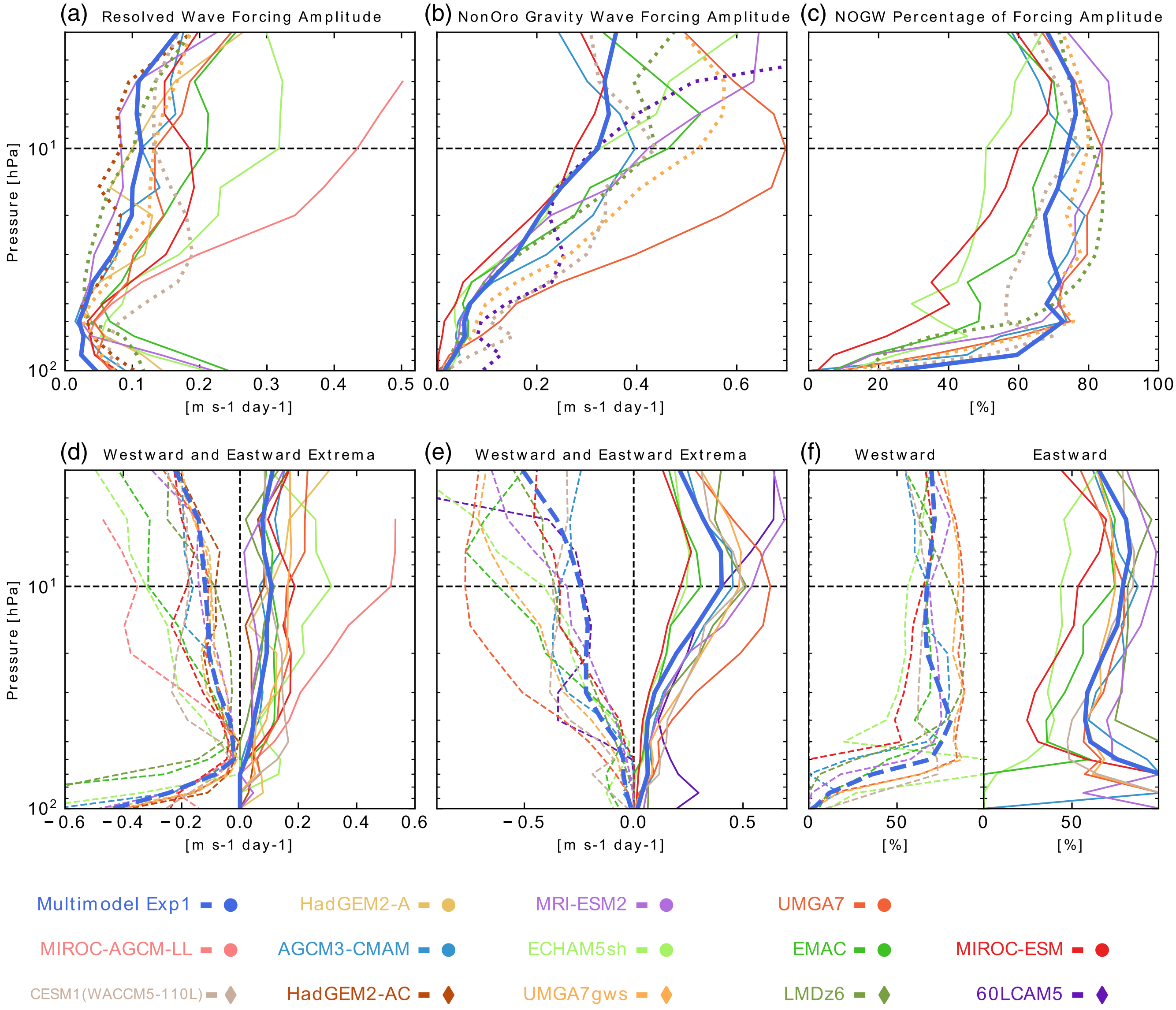


FIGURE 13 Vertical profiles (100 hPa to 3 hPa) of QBO mean cycles of equatorial (5°S–5°N) monthly and zonal mean accelerations for Exp 1 ensembles. (a) TTmceq amplitude of resolved wave acceleration (dotted lines are for models with variable source for parametrized NOGW; solid lines are for the remaining models). (b) is as (a) but for NOGW acceleration. (c) shows percentage of NOGW to the sum of resolved and NOGW acceleration amplitudes. (d–f) are as (a–c), but for separate eastward (solid) and westward (dashed) QBO phases. Models included are as in Figure 12

forcing with altitude is greater (Figure 13c). An increase with altitude in the proportion of forcing that comes from NOGWs suggests that these are depositing their momentum at too high an altitude, which would lead to a poorly simulated vertical distribution of momentum deposition and might explain why most simulated QBOs are too strong above 10 hPa and too weak lower down (Section 3.2).

Insufficient wave forcing at lower altitudes could also be due to problems with resolved waves which can result, for example, from coarse vertical resolution (Anstey *et al.*, 2016). Using a wave-213 spectral truncation, 256-level (T213L256) AGCM, Kawatani *et al.* (2010) found that small-scale gravity waves with zonal wavenumber greater than 107 ($\lesssim$180 km) were crucial for driving the westward phase in the lower-middle stratosphere. With the exception of MRI-ESM2, 60LCAM5 and CESM1(WACCM5-110L), models used here have horizontal resolution that is similar to or coarser than this limit (figure 5 of Butchart *et al.*, 2018) and cannot resolve these smaller-scale waves. Nonetheless, Figure 13f shows that the relative contributions of NOGWs to the forcing of the westward and eastward phases are quite similar, though the relative contribution to the westward phase is more constant with altitude. The growing importance with altitude of the eastward forcing from NOGWs would

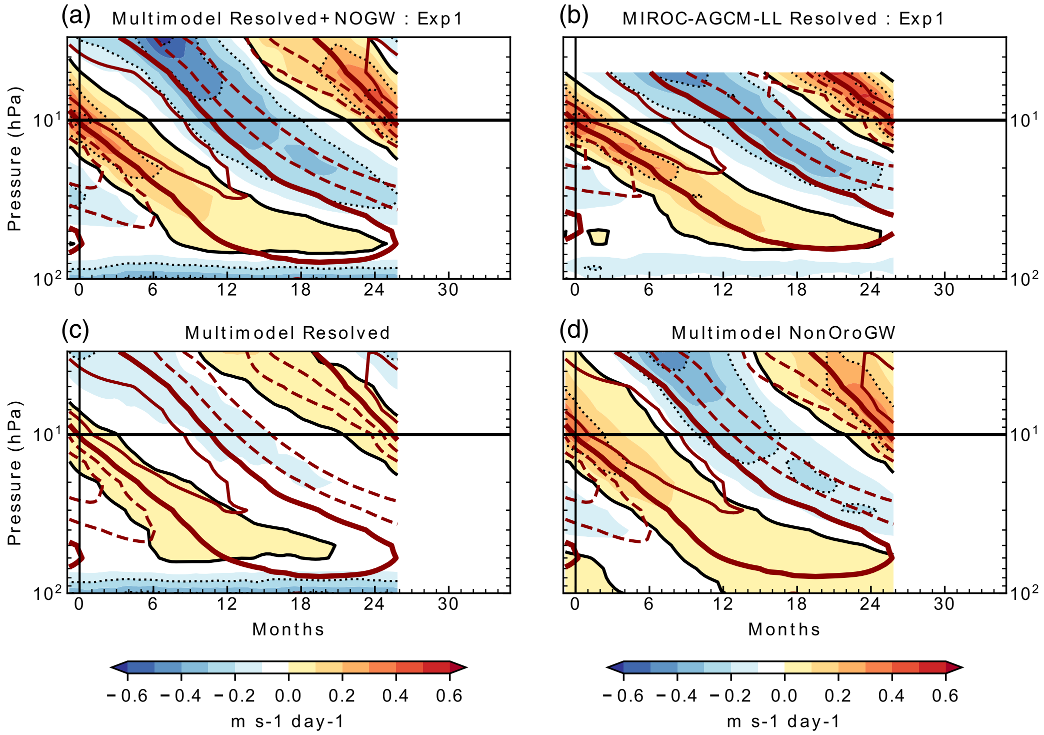


FIGURE 14 Pressure (100 hPa to 3 hPa) versus time mean QBO cycle, defined by westward to eastward wind transitions at 10 hPa and normalized to the multimodel mean period, for Exp 1 equatorial (5°S–5°N) monthly and zonal mean accelerations of the mean flow (colour shading, thin black contours) with corresponding zonal wind for reference (thick red contours, every 0.2 $\mathrm{m{\cdot}s^{-1}{\cdot}day^{-1}}$) due to (a) sum of resolved and subgrid NOGWs for all model ensembles except MIROC-AGCM-LL, (b) resolved waves for MIROC-AGCM-LL model only, (c) resolved waves for all models except MIROC-AGCM-LL, (d) subgrid NOGWs for all models except MIROC-AGCM-LL. Models included are as in Figure 12

be expected in models that poorly represent the eastward forcing from large-scale Kelvin waves (Holt *et al.*, 2020) though it is also consistent with a simple overestimate of parametrized fluxes in models with substantial eastward biases.

6.2 | Comparison with resolved gravity waves

Unlike for diagnostics shown in Sections 3 and 4, evaluating the forcings from NOGWs is complicated by a lack of observations. An alternative is to compare the parametrized results with those from models, ideally with much finer resolution, which explicitly resolve all the QBO wave forcings. However in QBOi to date there is only one model, MIROC-AGCM-LL, that is configured without parametrized NOGWs and it lacks the resolution to resolve the small-scale gravity waves identified in Section 6.1. As there is an interest in improving understanding of these contrasting modelling strategies, wave forcings over the depth and width (at 10 hPa) of the mean QBO cycle in MIROC-AGCM-LL are briefly compared below to the multimodel means of available forcings from models with parametrized NOGWs. Nonetheless, more robust conclusions will clearly depend on more results becoming available from models with better resolution that can explicitly represent the QBO wave forcings; this is an area for future work.

The change in wave forcing over the QBO cycle at each altitude is shown in Figure 14 averaged over all cycles and model ensembles for Exp 1. As in Figure 7, the averaging was achieved by scaling the time dependence of each QBO cycle such that its duration maps onto the Exp 1 multimodel mean period (i.e., all cycles are of the same duration and can be averaged). Shading in Figure 14a,c,d shows the total wave forcing as well as forcing from resolved and parametrized NOGWs averaged over models that include parametrized NOGWs (i.e., excluding MIROC-AGCM-LL). Contours denote the corresponding zonal mean zonal wind. As MIROC-AGCM-LL is able to simulate a realistic QBO, its wave forcing (Figure 14b) at most levels and throughout the QBO cycle is, as expected, very similar to the total wave forcing averaged for the models that include parametrized NOGWs (Figure 14a). This shows that the NOGW parametrizations used in these models are able to compensate for the shortfall in resolved wave forcing, which is weaker than the forcing by parametrized NOGW (compare Figure 14c,d). However, MIROC-AGCM-LL is still expected to underestimate resolved wave forcing as its resolution is too coarse to capture the aforementioned NOGWs with zonal wavenumber greater than 107. It is also notable that the occurrence of peak westward forcing at a higher altitude relative to the zero-wind line than eastward forcing, which was evident earlier, is clearly seen in both the MIROC-AGCM-LL and multimodel NOGW results and extends down to about 50 hPa where the descent of the westward shear zones terminate (Figure 14d). This again indicates realistic behaviour of the parametrized NOGWs, at least as judged by comparison with waves that are resolved in MIROC-AGCM-LL. Furthermore, it suggests that westward winds just above the tropopause (100–70 hPa) play an important role in filtering westward waves throughout the QBO cycle (not only during 10 hPa onsets as were shown in Figure 12).

Figure 15 is similar to Figure 14 except that it displays the latitudinal structure of the mean QBO cycle at 10 hPa rather than vertical structure at the Equator. As in Figure 14, the MIROC-AGCM-LL resolved wave forcing (Figure 15b) is seen to resemble total forcing in the mean of other models in most respects (Figure 15a). However a difference between them is seen at the QBO edges (10°–15° poleward of the Equator) where the eastward forcing persists somewhat farther past the peak in the eastward winds in MIROC-AGCM-LL than it does for total forcing in the mean of the remaining models. A similar latitudinal structure appears, weakly, in the multimodel resolved wave forcing (Figure 15c). The multimodel parametrized wave forcing, in contrast, shows a more uniform latitudinal distribution at about six months (Figure 15d) than is seen in MIROC-AGCM-LL, despite substantial intermodel variability (not shown) in the amplitude of the latitudinal structure earlier in the mean cycle. This suggests that parametrized eastward NOGWs may be too strong at the Equator, as would be consistent with the eastward biases seen in many models (Figure 5a) and narrow latitudinal extents (Figure 5f). However, as noted above, more robust conclusions will require further comparison against other models that include a spectrum of resolved small-scale NOGWs.

7 | CONCLUSIONS

In this paper, simulations of the QBO by thirteen AGCMs that performed two present-day experiments have been compared and evaluated against a widely used reanalysis product, ERA-Interim. The first experiment used observed SSTs for 1979–2009 while the second was a time-slice simulation with repeated annual cycle for the SSTs. Model configurations were kept the same as in the first experiment (i.e., only the external forcings differ). Comparison of the two experiments indicates only a small impact on the QBO due to these differences in interannual variability of the SSTs. Given this similarity, the focus of this study has been primarily on the simulations forced with observed SSTs in order to obtain the best possible validation of model QBOs against ERA-Interim. When observed SSTs are used, QBO mean periods range from 20 to 30 months, or −29% to +7% of the 28-month ERA-Interim mean period. One change in behaviour between the two experiments is that models using variable-source parametrized NOGWs have shorter QBO periods in the time-slice simulations. Despite differences in the climatology of observed SSTs compared to that used for the repeated annual cycle, this change appears to be an outcome of the reduced interannual variability. Although this sensitivity in models with variable-source parametrized NOGWs merits more detailed investigation, the differences are sufficiently small that an absence of interannual variability in the SSTs is unlikely to compromise time-slice projections of future QBO behaviour using these models (e.g., Richter *et al.*, 2020).

For CMIP5, a couple of metrics (e.g., mean period, amplitude) were sufficient to distinguish the few models capable of producing internally generated QBOs from the overwhelming majority that could not. The prospect of data becoming available (e.g., under CMIP6) from many more models with QBOs brings both the promise of improved statistical significance and a need to expand the range of metrics routinely deployed for characterizing QBO performance, in order to identify common traits that might be associated with predicting QBO impacts beyond the equatorial stratosphere. Such considerations motivated previous work to define a suite of metrics (Schenzinger *et al.*, 2017), which this paper has extended by also evaluating grades for a selection of metrics, based on the QBO zonal mean zonal wind, which summarize performance of the QBOi multimodel ensemble relative to recent QBO observations. A subset of these grades indicates that QBOi models on average simulate QBO periods and 10 hPa amplitudes at least as well as those CMIP5 models with the most realistic QBOs.

However, model grades for other metrics show that most perform poorly at 50 hPa and in representing widths (where meridional structure is generally too narrow) at both 10 and 50 hPa, indicating a need for further model development. Better grades for a metric based on amplitude attenuation suggest that the vertical profile of amplitude relative to an individual model's peak amplitude tends to be well represented, though the the peak amplitude (and hence profile scale) tends to be somewhat underestimated relative to ERA-Interim and its location translated upward. For both experiments, QBO amplitudes

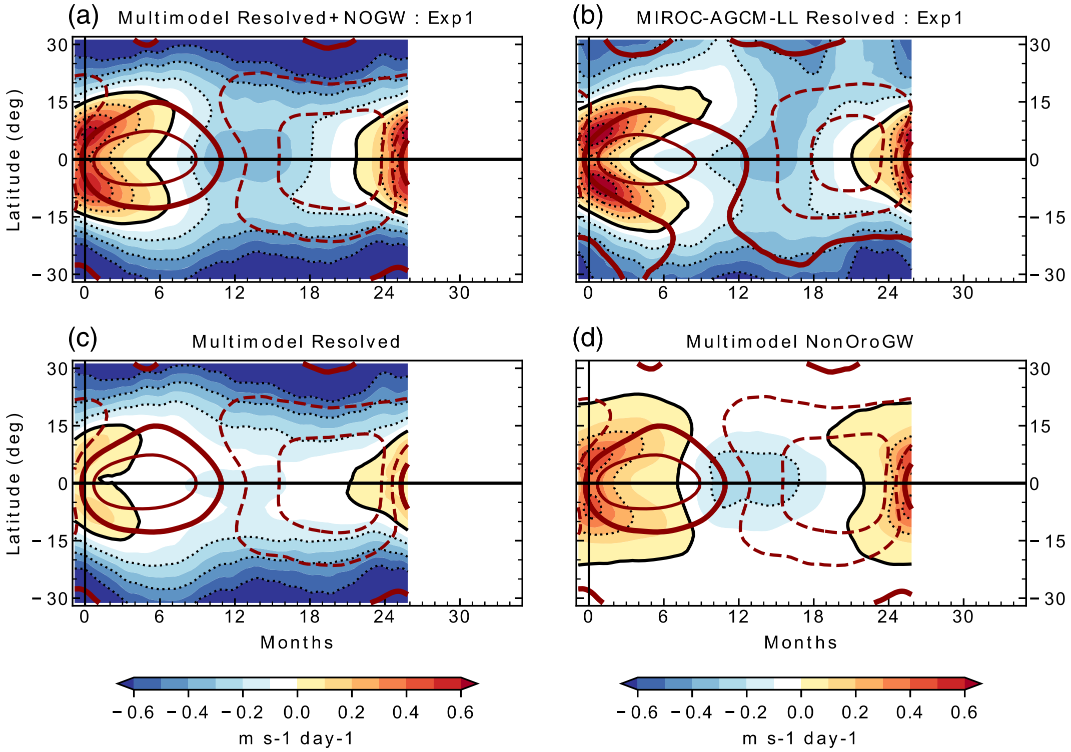


FIGURE 15 Subtropical latitude versus time mean QBO cycle, defined by westward to eastward wind transitions at 10 hPa and normalized to the multimodel mean period, for 10 hPa monthly and zonal mean Exp 1 accelerations of the mean flow (colour shading, thin black contours) with corresponding zonal wind (thick red contours, every 0.2 $\mathrm{m \cdot s^{-1} \cdot day^{-1}}$) due to (a) sum of resolved and subgrid NOGWs for all model ensembles except MIROC-AGCM-LL, (b) resolved waves for MIROC-AGCM-LL model only, (c) resolved waves for all models except MIROC-AGCM-LL, (d) subgrid NOGWs for all models except MIROC-AGCM-LL. Models included are as in Figure 12

reach their maximum in the vertical at or about 10 hPa in most models, where their amplitudes agree best with reanalysis, though the peak for ERA-Interim lies near 15 hPa. As is consistent with the weaker zonal wind shear, the QBO in temperature is weaker than in ERA-Interim, but penetrates down to the tropopause in the multimodel mean. A challenge for future model development is to improve the QBO simulation as assessed not just against individual (e.g., period) metrics, but across a range of metrics such as those used in Figure 11.

Forcing by parametrized NOGWs exceeds that by resolved waves at all altitudes, indicating that the simulated QBOs are strongly dependent on parametrized forcing for both eastward and westward phases of the QBO. The relative strength of resolved and parametrized forcing does not vary strongly with altitude, but the slight increase with altitude in the relative contribution of NOGWs could play a role in the overestimate of QBO amplitude at high altitudes by most models. However, given a lack of observational constraints on reanalysis winds at altitudes above 10 hPa in the Tropics, it is possible that ERA-Interim underestimates the amplitude above 10 hPa. MERRA and MERRA-2 reanalyses (not shown), which unlike ERA-Interim include parametrized NOGW (Fujiwara *et al.*, 2017), have by comparison larger QBO amplitudes above 10 hPa. If true, this would imply a smaller overestimate by QBOi models than ERA-Interim suggests.

Comparison between MIROC-AGCM-LL, the sole model configured without a NOGW parametrization, and the other models suggests that parametrized NOGWs successfully compensate for low resolved wave forcing in many respects, allowing these other models to simulate the QBO in a physically plausible way. This comparison also suggests that the parametrized NOGWs may produce eastward forcing that is too strong at the Equator. However, any conclusions regarding the realism of parametrized gravity waves based on a comparison with the single model without parametrized NOGWs (MIROC-AGCM-LL) are strictly tentative, as that model lacks the resolution believed necessary to represent explicitly a full spectrum of gravity waves. Moreover, MIROC-AGCM-LL shows similar QBO biases to the other models and has resolved gravity waves that are not necessarily realistic (e.g., they are likely to depend on the model's parametrized deep convection). Nevertheless, as climate model resolutions move into territory previously occupied by numerical weather prediction, where more of the GW spectrum is explicitly represented, QBOs are likely to be less dependent on NOGW parametrizations and there is continued interest in exploring the development potential of both modelling strategies. In these present-day experiments, parametrized forcing by NOGWs remains a major uncertainty underlying QBO simulation, as it does for future projections of the QBO (Richter *et al.*, 2020). One aspect that might be explored in more detail is vertical resolution, which can influence wave propagation through the QBO region due to its impact on dissipation processes in the stratosphere. Other topics relevant for future work include the representation of the temperature QBO and its associated meridional mean circulation, especially near the tropical tropopause; the QBO in ozone and ozone feedbacks on the dynamical QBO (e.g., Butchart *et al.*, 2003; Naoe *et al.*, 2017); and synchronization of the QBO with ENSO or the annual cycle (e.g., Christiansen *et al.*, 2016).

In summary, all thirteen QBOi models simulate reasonable QBOs, though the graded metric portrait of model performance indicated that this success centres on selected metrics that are most commonly associated with the tuning of NOGW parametrization schemes. Further analysis is thus needed to better understand the origin of common model deficiencies that have been identified, in particular around the characterization of width and amplitude at 50 hPa which may be important for teleconnections with the Extratropics. With improved understanding of the strength and weakness of simulated QBOs, the choice of graded metrics used to quantify model performance will likewise evolve. However, this evolution will need to be balanced with a measure of continuity to allow for a comparison between different generations of models.

ACKNOWLEDGEMENTS

We acknowledge the scientific guidance of the World Climate Research Programme (WCRP) for helping motivate this work, coordinated under the framework of the Stratosphere-troposphere Processes and their Role in Climate (SPARC) QBO initiative (QBOi) led by JA, NB, KH and SO. FL, SW, YK, LG and SO acknowledge funding for the Belmont Forum JPI-Climate GOTHAM project from Agence Nationale Recherche (ANR-15-JCLI-0004-01), the Japan Science and Technology (JST) Agency and the Natural Environment Research Council (NE/P006779/1). YK was supported by Japan Society for Promotion of Science (JPSP) KAKENHI grant nos. 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 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. LG and SO acknowledge support from the UK Natural Environment Research Council (NERC) grant NE/M005828/1 and the National Centre for Atmospheric Science (NCAS). PB, TK, SV acknowledge support by the state of Baden-Württemberg through bwHPC. NB was supported by the Met Office Hadley Centre Programme funded by BEIS and Defra. This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement no. 1852977. The CESM project is supported primarily by the National Science Foundation. Portions of this study were supported by the Regional and Global Model Analysis (RGMA) component of the Earth and Environmental System Modeling Program of the U.S. Department of Energy's Office of Biological & Environmental Research (BER) via National Science Foundation IA 1947282. 60LCAM5 and CESM1(WACCM5-110L) simulations were carried out on the Yellowstone high-performance computing platform (CISL, 2012). CC and FS have been supported by the Copernicus Climate Change Service, funded by the EU and implemented by the European Centre for Medium-Range Weather Forecasts (ECMWF). We acknowledge the Centre for Environmental Data Analysis for use of the Jasmin analysis platform and group workspace environment.

ORCID

A. C. Bushell https://orcid.org/0000-0001-5683-4387
S. M. Osprey https://orcid.org/0000-0002-8751-1211
J. H. Richter https://orcid.org/0000-0001-7048-0781
F. Serva https://orcid.org/0000-0002-7118-0817
T. Kerzenmacher https://orcid.org/0000-0001-8413-0539
Y.-H. Kim https://orcid.org/0000-0003-4014-073X
F. Lott https://orcid.org/0000-0003-2126-5510
H. Naoe https://orcid.org/0000-0002-6261-0854
A. K. Smith https://orcid.org/0000-0003-2384-5033
T. N. Stockdale https://orcid.org/0000-0002-7901-0337

REFERENCES

Anstey, J.A., Scinocca, J.F. and Keller, M. (2016) Simulating the QBO in an atmospheric general circulation model: sensitivity to resolved and parameterized forcing. *Journal of the Atmospheric Sciences*, 73, 1649–1665.

Baldwin, M.P., Gray, L.J., Dunkerton, T.J., Hamilton, K., Haynes, P., Randel, W.J., Holton, J.R., Alexander, M.J., Hirota, I., Horinouchi, T., Jones, D.B.A., Kinnersley, J.S., Marquardt, C., Sato, K. and Takahashi, M. (2001) The Quasi-Biennial Oscillation. *Reviews of Geophysics*, 39, 179–229.

Bushell, A.C., Butchart, N., Derbyshire, S.H., Jackson, D.R., Shutts, G.J., Vosper, S.B. and Webster, S. (2015) Parameterized gravity wave momentum fluxes from sources related to convection and large-scale precipitation processes in a global atmosphere model. *Journal of the Atmospheric Sciences*, 72, 4349–4371.

Butchart, N. and Austin, J. (1998) Middle atmosphere climatologies from the troposphere–stratosphere configuration of the UKMO's Unified Model. *Journal of the Atmospheric Sciences*, 55, 2782–2809.

Butchart, N., Scaife, A.A., Austin, J., Hare, S.H.E. and Knight, J.R. (2003) Quasi-Biennial Oscillation in ozone in a coupled chemistry–climate model. *Journal of Geophysical Research; Atmospheres*, 108(15). https://doi.org/10.1029/2002JD003004.

Butchart, N., Charlton-Perez, A.J., Cionni, I., Hardiman, S.C., Haynes, P.H., Krüger, K., Kushner, P.J., Newman, P.A., Osprey, S.M., Perlwitz, J., Sigmond, M., Wang, L., Akiyoshi, H., Austin, J., Bekki, S., Baumgaertner, A., Braesicke, P., Brühl, C., Chipperfield, M., Dameris, M., Dhomse, S., Eyring, V., Garcia, R., Garny, H., Jöckel, P., Lamarque, J.-F., Marchand, M., Michou, M., Morgenstern, O., Nakamura, T., Pawson, S., Plummer, D., Pyle, J., Rozanov, E., Scinocca, J., Shepherd, T.G., Shibata, K., Smale, D., Teyssèdre, H., Tian, W., Waugh, D. and Yamashita, Y. (2011) Multimodel climate and variability of the stratosphere. *Journal of Geophysical Research; Atmospheres*, 116(D5). https://doi.org/10.1029/2010JD014995.

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). *Geoscientific Model Development*, 11, 1009–1032.

Choi, H. and Chun, H. (2011) Momentum flux spectrum of convective gravity waves. Part I: an update of a parameterization using mesoscale simulations. *Journal of the Atmospheric Sciences*, 68, 739–759.

Christiansen, B., Yang, S. and Madsen, M.S. (2016) Do strong warm ENSO events control the phase of the stratospheric QBO?. *Geophysical Research Letters*, 43, 10489–10495.

CISL (2012) *Yellowstone: IBM iDataPlex /FDR-IB*. Computational and Information Systems Laboratory, NCAR/UCAR, Boulder, CO. Available at: http://n2t.net/ark:/85065/d7wd3xhc; accessed 29 January 2020.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L., Kållberg, P., Koehler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N. and Vitart, F. (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137, 553–597.

Dunkerton, T.J. (2016) The Quasi-Biennial Oscillation of 2015–2016: hiccup or death spiral?. *Geophysical Research Letters*, 43, 10547–10552.

Dunkerton, T.J. and Delisi, D.P. (1985) Climatology of the equatorial lower stratosphere. *Journal of the Atmospheric Sciences*, 42, 376–396.

Ebdon, R.A. and Veryard, R.G. (1961) Fluctuations in equatorial stratospheric winds. *Nature*, 189, 791–793.

Fujiwara, M., Wright, J.S., Manney, G.L., Gray, L.J., Anstey, J.A., Birner, T., Davis, S., Gerber, E.P., Harvey, V.L., Hegglin, M.I., Homeyer, C.R., Knox, J.A., Krüger, K., Lambert, A., Long, C.S., Martineau, P., Molod, A., Monge-Sanz, B.M., Santee, M.L., Tegtmeier, S., Chabrillat, S., Tan, D.G.H., Jackson, D.R., Polavarapu, S., Compo, G.P., Dragani, R., Ebisuzaki, W., Harada, Y., Kobayashi, C., McCarty, W., Onogi, K., Pawson, S., Simmons, A., Wargan, K., Whitaker, J.S. and Zou, C. (2017) Introduction to the SPARC Reanalysis Intercomparison Project (S-RIP) and overview of the reanalysis systems. *Atmospheric Chemistry and Physics*, 17, 1417–1452.

Hardiman, S.C., Boutle, I.A., Bushell, A.C., Butchart, N., Cullen, M.J.P., Field, P.R., Furtado, K., Manners, J.C., Milton, S.F., Morcrette, C., O'Connor, F.M., Shipway, B.J., Smith, C., Walters, D.N., Willett, M.R., Williams, K.D., Wood, N., Abraham, N.L., Keeble, J., Maycock, A.C., Thuburn, J. and Woodhouse, M.T. (2015) Processes controlling tropical tropopause temperature and stratospheric water vapor in climate models. *Journal of Climate*, 28, 6516–6535.

Hines, C.O. (1997) Doppler spreading parametrization of gravity-wave momentum deposition in the middle atmosphere. 2. Broad and quasi-monochromatic spectra, and implementation. *Journal of Atmospheric and Solar-Terrestrial Physics*, 59, 387–400.

Holt, L., Lott, F., Garcia, R.R., Kiladis, G.N., Anstey, J.A., 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. *Quarterly Journal of the Royal Meteorological Society*.

Huntingford, C., Marsh, T., Scaife, A.A., Kendon, E.J., Hannaford, J., Kay, A.L., Lockwood, M., Prudhomme, C., Reynard, N.S., Parry, S., Lowe, J.A., Screen, J.A., Ward, H.C., Roberts, M., Stott, P.A., Bell, V.A., Bailey, M., Jenkins, A., Legg, T., Otto, F.E.L., Massey, N., Schaller, N., Slingo, J. and Allen, M.R. (2014) Potential influences on the United Kingdom's floods of winter 2013/14. *Nature Climate Change*, 4, 769–777.

Kawatani, Y. and Hamilton, K. (2013) Weakened stratospheric Quasi-Biennial Oscillation driven by increased tropical mean upwelling. *Nature*, 497, 478–481.

Kawatani, Y., Watanabe, S., Sato, K., Dunkerton, T., Miyahara, S. and Takahashi, M. (2010) The roles of equatorial trapped waves and internal inertia-gravity waves in driving the Quasi-Biennial Oscillation. Part I: zonal mean wave forcing. *Journal of the Atmospheric Sciences*, 67, 963–980.

Kim, J., Grise, K. and Son, S. (2013) Thermal characteristics of the cold-point tropopause region in CMIP5 models. *Journal of Geophysical Research: Atmospheres*, 118, 8827–8841.

Lindzen, R.S. (1981) Turbulence and stress owing to gravity wave and tidal breakdown. *Journal of Geophysical Research; Oceans*, 86, 9707–9714.

Lott, F. and Guez, L. (2013) A stochastic parameterization of the gravity waves due to convection and its impact on the equatorial stratosphere. *Journal of Geophysical Research: Atmospheres*, 118, 8897–8909.

Manzini, E. and Bengtsson, L. (1996) Stratospheric climate and variability from a general circulation model and observations. *Climate Dynamics*, 12, 615–639.

Naoe, H., Deushi, M., Yoshida, K. and Shibata, K. (2017) Future changes in the ozone Quasi-Biennial Oscillation with increasing GHGs and ozone recovery in CCMI simulations. *Journal of Climate*, 30, 6977–6997.

Osprey, S.M., 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, 1424–1427.

Plumb, R.A. and Bell, R.C. (1982) A model of the Quasi-Biennial Oscillation on an equatorial beta-plane. *Quarterly Journal of the Royal Meteorological Society*, 108, 335–352.

Pulido, M. and Thuburn, J. (2008) The seasonal cycle of gravity wave drag in the middle atmosphere. *Journal of Climate*, 21, 4664–4679.

Reed, R.J., Campbell, W.J., Rasmussen, L.A. and Rogers, D.G. (1961) Evidence of a downward-propagating, annual wind reversal in the equatorial stratosphere. *Journal of Geophysical Research*, 66, 813–818.

Richter, J.H., Sassi, F. and Garcia, R.R. (2010) Toward a physically based gravity wave source parameterization in a general circulation model. *Journal of the Atmospheric Sciences*, 67, 136–156.

Richter, J.H., Butchart, N., Kawatani, Y., Bushell, A.C., Holt, L., Anstey, J.A., 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. *Quarterly Journal of the Royal Meteorological Society*.

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. *Geophysical Research Letters*, 27, 3481–3484.

Scaife, A.A., Athanassiadou, M., Andrews, M., Arribas, A., Baldwin, M.P., Dunstone, N., Knight, J., MacLachlan, C., Manzini, E., Müller, W.A., Pohlmann, H., Smith, D., Stockdale, T.N. and William, A. (2014) Predictability of the Quasi-Biennial Oscillation and its northern winter teleconnection on seasonal to decadal timescales. *Geophysical Research Letters*, 41, 1752–1758.

Schenzinger, V., Osprey, S., Gray, L.J. and Butchart, N. (2017) Defining metrics of the Quasi-Biennial Oscillation in global climate models. *Geoscientific Model Development*, 10, 2157–2168.

Takahashi, M. (1996) Simulation of the stratospheric Quasi-Biennial Oscillation using a general circulation model. *Geophysical Research Letters*, 23, 661–664.

Warner, C.D. and McIntyre, M.E. (1999) Toward an ultra-simple spectral gravity wave parametrization for general circulation models. *Earth Planets Space*, 51, 475–484.

Waugh, D.W. and Eyring, V. (2008) Performance metrics for stratosphere resolving chemistry–climate models. *Atmospheric Chemistry and Physics*, 8, 5699–5713.

Yang, M. and Yu, Y. (2016) Attribution of variations in the Quasi-Biennial Oscillation period from the duration of easterly and westerly phases. *Climate Dynamics*, 47, 1943–1959.

Yoshida, K., Mizuta, R. and Arakawa, O. (2018) Intermodel differences in upwelling in the tropical tropopause layer among CMIP5 models. *Journal of Geophysical Research: Atmospheres*, 123, 13658–13675.

Yulaeva, E., Holton, J.R. and Wallace, J.M. (1994) On the cause of the annual cycle in tropical lower-stratospheric temperatures. *Journal of the Atmospheric Sciences*, 51, 169–174.

How to cite this article: Bushell AC, Anstey JA, Butchart N, Kawatani Y, Osprey SM, Richter JH, Serva F, Braesicke P, Cagnazzo C, Chen C-C, Chun H-Y, Garcia RR, Gray LJ, Hamilton K, Kerzenmacher T, Kim Y-H, Lott F, McLandress C, Naoe H, Scinocca J, Smith AK, Stockdale TN, Versick S, Watanabe S, Yoshida K, Yukimoto S. Evaluation of the Quasi-Biennial Oscillation in global climate models for the SPARC QBO-initiative. *QJR Meteorol Soc*. 2020;1–31. https://doi.org/10.1002/qj.3765