Multimodel climate and variability of the stratosphere

N. Butchart,[1] A. J. Charlton-Perez,[2] I. Cionni,[3] S. C. Hardiman,[1] P. H. Haynes,[4] K. Krüger,[5] P. J. Kushner,[6] P. A. Newman,[7] S. M. Osprey,[8] J. Perlwitz,[9] M. Sigmond,[6] L. Wang,[6] H. Akiyoshi,[10] J. Austin,[11] S. Bekki,[12] A. Baumgaertner,[13] P. Braesicke,[4] C. Brühl,[13] M. Chipperfield,[14] M. Dameris,[3] S. Dhomse,[14] V. Eyring,[3] R. Garcia,[15] H. Garny,[3] P. Jöckel,[13] J.-F. Lamarque,[15] M. Marchand,[12] M. Michou,[16] O. Morgenstern,[17] T. Nakamura,[10] S. Pawson,[7] D. Plummer,[18] J. Pyle,[4] E. Rozanov,[19] J. Scinocca,[20] T. G. Shepherd,[6] K. Shibata,[21] D. Smale,[17] H. Teyssèdre,[16] W. Tian,[14] D. Waugh,[22] and Y. Yamashita[10]

Received 1 September 2010; revised 26 November 2010; accepted 28 December 2010; published 3 March 2011.

[1] The stratospheric climate and variability from simulations of sixteen chemistry-climate models is evaluated. On average the polar night jet is well reproduced though its variability is less well reproduced with a large spread between models. Polar temperature biases are less than 5 K except in the Southern Hemisphere (SH) lower stratosphere in spring. The accumulated area of low temperatures responsible for polar stratospheric cloud formation is accurately reproduced for the Antarctic but underestimated for the Arctic. The shape and position of the polar vortex is well simulated, as is the tropical upwelling in the lower stratosphere. There is a wide model spread in the frequency of major sudden stratospheric warnings (SSWs), late biases in the breakup of the SH vortex, and a weak annual cycle in the zonal wind in the tropical upper stratosphere. Quantitatively, "metrics" indicate a wide spread in model performance for most diagnostics with systematic biases in many, and poorer performance in the SH than in the Northern Hemisphere (NH). Correlations were found in the SH between errors in the final warming, polar temperatures, the leading mode of variability, and jet strength, and in the NH between errors in polar temperatures, frequency of major SSWs, and jet strength. Models with a stronger QBO have stronger tropical upwelling and a colder NH vortex. Both the qualitative and quantitative analysis indicate a number of common and long-standing model problems, particularly related to the simulation of the SH and stratospheric variability.

Citation: Butchart, N., et al. (2011), Multimodel climate and variability of the stratosphere, *J. Geophys. Res.*, *116*, D05102, doi:10.1029/2010JD014995.

1. Introduction

[2] The accurate representation of the climate of the middle atmosphere is important for modeling both the effects of climate change on ozone recovery and the stratosphere-troposphere teleconnections which can have a

[1]Met Office Hadley Centre, Exeter, UK.
[2]Department of Meteorology, University of Reading, Reading, UK.
[3]Deutsches Zentrum für Luft und Raumfahrt, Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany.
[4]National Centre for Atmospheric Science, University of Cambridge, Cambridge, UK.
[5]IFM-GEOMAR, Leibniz Institute of Marine Sciences, Kiel University, Kiel, Germany.
[6]Department of Physics, University of Toronto, Toronto, Ontario, Canada.
[7]NASA Goddard Space Flight Center, Greenbelt, Maryland, USA.
[8]National Centre for Atmospheric Science, Department of Physics, University of Oxford, Oxford, UK.
[9]Physical Sciences Division, CIRES, University of Colorado and NOAA Earth System Research Laboratory, Boulder, Colorado, USA.
[10]National Institute for Environmental Studies, Tsukuba, Japan.
[11]Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, New Jersey, USA.
[12]LATMOS-IPSL, UPMC, France.
[13]Max Planck Institut für Chemie, Mainz, Germany.
[14]School of Earth and Environment, University of Leeds, Leeds, UK.
[15]National Center for Atmospheric Research, Boulder, Colorado, USA.
[16]GAME/CNRM, (Météo-France, CNRS), Toulouse, France.
[17]National Institute of Water and Atmospheric Research, Lauder, New Zealand.
[18]Environment Canada, Toronto, Ontario, Canada.
[19]Physical-Meteorological Observatory/World Radiation Centre, Davos, Switzerland.
[20]Meteorological Service of Canada, University of Victoria, Victoria, British Columbia, Canada.
[21]Meteorological Research Institute, Tsukuba, Japan.
[22]Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, Maryland, USA.

0148-0227/11/2010JD014995

significant impact on the surface climate and its variability [e.g., *Gillett and Thompson*, 2003; *Baldwin et al.*, 2007]. This study assesses and compares the abilities of a multimodel ensemble of chemistry-climate models (CCMs) to reproduce the climate, circulation, and associated variability of the stratosphere, over the period 1980–1999. The aim of the assessment is to describe in detail the current state-of-the-art in the modeling of stratospheric climate. For this study the focus is on the overall performance of the multimodel ensemble, rather than on the performance of any individual model. In particular, the parts of the stratospheric climate system where the models suffer from common biases are highlighted while the spread in model behavior, relative to the sampling uncertainty of a given parameter, is used to diagnose differences in model performance across the ensemble.

[3] Assessment of the current state-of-the-art in stratospheric climate modeling is important for a number of reasons. First, it is important to understand how deficiencies in the representation of stratospheric climate might influence projections of stratospheric ozone. Second, several authors [e.g., *Baldwin et al.*, 2003; *Baldwin et al.*, 2007; *Shaw and Shepherd*, 2008] have advocated the inclusion of a well-resolved stratosphere in models used for a variety of purposes including seasonal and decadal prediction and the simulation of longer-term changes in surface climate. Decisions about the inclusion of a well-resolved stratosphere in predominately tropospheric climate and earth system models are better informed by a clear assessment of the strengths and weaknesses of current stratosphere-resolving models.

[4] Multimodel assessments of the ability of stratosphere-resolving general circulation models (GCMs) have occurred at frequent intervals during the last decade. The GCM-Reality Intercomparison Project (GRIPS) of the Stratospheric Processes and their role in Climate (SPARC) core project of the World Climate Research Programme (WCRP), assessed short runs of 13 GCMs [*Pawson et al.*, 2000]. This intercomparison found that the main climatological features of the stratosphere were well simulated by most models but that significant cold biases existed throughout the extratropical lower stratosphere and were particularly acute in the Southern Hemisphere (SH). Additionally, there was a large divergence in the simulation of the annual cycle in the zonal mean temperature of the lower stratosphere.

[5] The performance of longer simulations from eight CCMs, which included coupled stratospheric chemistry, was reported by *Austin et al.* [2003]. In this intercomparison, models which incorporated a nonorographic gravity wave drag (NOGWD) parameterization were found to have much reduced temperature biases in both the northern and southern high latitudes compared both to models without a NOGWD parameterization and to the models in the *Pawson et al.* [2000] intercomparison. *Austin et al.* [2003] also demonstrated that the relationship identified by *Newman et al.* [2001] between polar temperatures and the meridional heat flux at 100 hPa could be used to evaluate the model responses to tropospheric wave forcing.

[6] More recent intercomparisons of CCMs have been conducted as part of the SPARC Chemistry Climate Model Validation (CCMVal) activity [*Eyring et al.*, 2005]. In CCMVal phase 1 (CCMVal-1), 13 CCMs, run with near identical climate and chemical forcings, were compared by *Eyring et al.* [2006]. In particular they found significant improvements in the simulation of both global mean and high-latitude temperatures relative to the earlier studies, though large differences among models still existed in the temperature and meridional heat flux diagnostics. Eyring et al. also pointed to a significant bias in the mean breakup date of the Southern Hemisphere polar vortex in most of the models. Further aspects of the dynamics of the CCMVal-1 models such as the driving of the Brewer-Dobson circulation and the threshold temperatures for polar stratospheric cloud (PSC) formation were assessed by *Butchart et al.* [2010a].

[7] The present study builds on these earlier assessments and compares 16 CCMs run with near identical climate and chemical forcings for CCMVal phase 2 (CCMVal-2). The simulation of the stratospheric climate is assessed in more detail than in the previous studies using a larger ensemble of CCMs and a more extensive range of diagnostics. In addition to examining the mean stratospheric climate and seasonal cycle, a detailed comparison of the model's abilities to model intraseasonal variability, stationary waves, tropical variability, and annular mode dynamics is made. The performance of the individual models is explored by *Butchart et al.* [2010b]; here the focus is on the multimodel mean climatology together with an assessment of the generic model biases and uncertainties (i.e., model spread). In the present study reference to individual model results is generally excluded. Nonetheless the individual model results shown in the figures are identified by the model names for cross referencing with the companion study of *Butchart et al.* [2010b].

2. Models and Simulations

[8] The 16 models used in this study are listed in Table 1, along with their horizontal and vertical resolution, top level, and references. For a more extensive description of these models, see *Morgenstern et al.* [2010]. The models vary greatly in their representation of key processes and sophistication though all include coupled stratospheric chemistry. Many of the models have been involved in one or more of the previous assessments but may have undergone significant modification and development even in the relatively short period between the CCMVal-1 and CCMVal-2 projects (see *Morgenstern et al.* [2010] and appropriate references in Table 1). The models considered here are those that uploaded dynamical diagnostics from the CCMVal-2 reference simulations [*Eyring et al.*, 2008] to the central data base at the British Atmospheric Data Centre (BADC) though, because the focus is on dynamical processes, only results from models formulated using the primitive equations or using a representation of the fluid equations of motion at least as accurate as the primitive equations, are used. Also note that two of the models have an upper boundary below 1 hPa (see Table 1) and hence for those diagnostics presented in section 3 as vertical profiles the curves for these two models stop below 1 hPa.

[9] *Eyring et al.* [2008] defined two reference simulations: REF-B1 and REF-B2. The "historical" REF-B1 simulation covers the period 1960–2005. This simulation generally includes all anthropogenic and natural forcings based on observed changes in the abundance of trace gases (i.e.,

Table 1. Resolution, Number of Levels, and Upper Boundary of the Models Used in This Study[a]

Model	Horizontal Resolution, Number of Levels/Top Level	Reference
AMTRAC3	~200 km (cube sphere grid), 48 L, 0.017 hPa	*Austin and Wilson* [2010]
CAM3.5	1.9° × 2.5°, 26 L, 3.5 hPa	*Lamarque et al.* [2008]
CCSRNIES	T42 (2.8° × 2.8°), 34 L, 0.012 hPa	*Akiyoshi et al.* [2009]
CMAM	T31 (3.75° × 3.75°), 71 L, 0.00081 hPa	*Scinocca et al.* [2008]; *de Grandpré et al.* [2000]
CNRM-ACM	T42 (2.8° × 2.8°), 60 L, 0.07 hPa	*Déqué* [2007]; *Teyssèdre et al.* [2007]
E39CA	T30 (3.75° × 3.75°), 39 L, 10 hPa	*Dameris et al.* [2005]; *Garny et al.* [2009]; *Stenke et al.* [2009]
EMAC	T42 (2.8° × 2.8°), 90 L, 0.01 hPa	*Jöckel et al.* [2006]
GEOSCCM	2° × 2.5°, 72 L, 0.015 hPa	*Pawson et al.* [2008]
LMDZrepro	2.5° × 3.75°, 50 L, 0.07 hPa	*Jourdain et al.* [2008]
MRI	T42 (2.8° × 2.8°), 68 L, 0.01 hPa	*Shibata and Deushi* [2008a, 2008b]
Niwa_SOCOL	T30 (3.75° × 3.75°), 39 L, 0.01 hPa	*Schraner et al.* [2008]; *Egorova et al.* [2005]
SOCOL	T30 (3.75° × 3.75°), 39 L, 0.01 hPa	*Schraner et al.* [2008]; *Egorova et al.* [2005]
UMSLIMCAT	2.5° × 3.75°, 64 L, 0.01 hPa	*Tian and Chipperfield* [2005, 2006]
UMUKCA_METO	2.5° × 3.75°, 60 L, 84 km	*Morgenstern et al.* [2008, 2009]; *Hardiman et al.* [2010b]; *Osprey et al.* [2010]
UMUKCA_UCAM	2.5° × 3.75°, 60 L, 84 km	*Morgenstern et al.* [2008, 2009]; *Hardiman et al.* [2010b]; *Osprey et al.* [2010]
WACCM	1.9° × 2.5°, 66 L, 5.96 × 10^{-6} hPa	*Garcia et al.* [2007]

[a]The models are listed alphabetically by name. For grid point models the horizontal resolution is given as the latitude×longitude grid spacing in degrees. For spectral models the horizontal resolution is given as the triangular truncation of the spectral domain, with the equivalent grid point resolution in brackets.

greenhouse gases (GHGs) and ozone depleting substances (ODSs)) solar variability, volcanic eruptions, and sea surface temperature and sea ice distributions (SSTs) (see Table 1 of *Eyring et al.* [2008]). In addition, several models included an extra artificial zonal momentum forcing in the equatorial stratosphere to constrain the model to reproduce the observed quasi-biennial oscillation (QBO) over this period. These models are therefore not strictly "free-running" GCMs but are still considered in this assessment.

[10] REF-B2 is a self consistent simulation from the past into the future (1960–2100). Observed changes in the concentrations of GHGs and ODSs are prescribed for the past period. For the future, GHG amounts follow the A1B scenario given by *Nakicenovic and Swart* [2000], and the surface halogens follow the adjusted A1 scenario given by *World Meteorological Organization* [2007]. External forcings such as solar variability and volcanic eruptions are not included to maintain consistency in the time series from the past to the future. Similarly, to avoid a possible discontinuity in the SST forcing between the past and future, the SSTs are taken from ocean-atmosphere model simulations (without the coupled chemistry) following the same A1B GHG scenario, apart from the Canadian Middle Atmosphere Model (CMAM) which includes a fully coupled ocean. Further details of the design of the REF-B1 and REF-B2 simulations and the rationale behind these simulations is given by *Eyring et al.* [2008] and *Morgenstern et al.* [2010].

[11] The main focus of this study is on the period 1980–1999 when the global stratosphere was extensively observed by instruments on artificial satellites and space ships. In addition, high-quality stratospheric (re)analyses of dynamical quantities [*Swinbank and O'Neill*, 1994; *Kalnay et al.*, 1996; *Uppala et al.*, 2005] are available for all or part of this period. The emphasis will be on the REF-B1 simulations which were specifically designed to provide the best possible representation of the stratospheric climate and variability over the period 1960–2006. However, because of the lack of emissions data for the simulations after 2000 [*Eyring et al.*, 2008], only the first 20 years of the extensively observed period from 1980 to the present day is analyzed from the simulations. In addition comparison with the corresponding period from the REF-B2 simulations will be used for some of the diagnostics to help elucidate the role of SST variability.

3. Qualitative Assessment

3.1. Polar Night Jet

[12] The starting point for this assessment is the mean structure and interannual variability of the stratospheric polar night jet (PNJ). Two aspects are considered: the strength of the stratospheric PNJ and its latitudinal position (Figure 1). The model ensemble performs extremely well in these diagnostics in the Northern Hemisphere (NH), though not quite so well in the SH. The NH jet is generally both well positioned and of the correct strength in almost all models, and the multimodel mean is very close to the reanalysis data. Apart from three obvious outliers, the spread in the jet strength is slightly larger than the observational range with no systematic bias toward strong or weak jets. The one outlying model with too weak of a jet also positions the jet about 20° too close to the equator (note the other incorrectly positioned jet at 10 hPa is almost certainly a consequence of that particular model having an upper boundary below 1 hPa).

[13] In the SH winter, clear biases exist for the majority of the models in the upper stratosphere. The model ensemble fails to capture the observed tilt of the jet toward the equator between 10 and 1 hPa, with most models producing a jet with an untilted profile. In the upper stratosphere there is a large spread in the strength of the SH midwinter jet produced by the models with a systematic bias toward jets which are too strong. Only one model produces a jet which is too weak. In contrast, below 10 hPa the spread in the jet strength is smaller in the SH than in the NH.

[14] The interannual variability of the wintertime extratropical stratospheric circulation is mainly characterized by variations in the strength and location of the PNJ. This

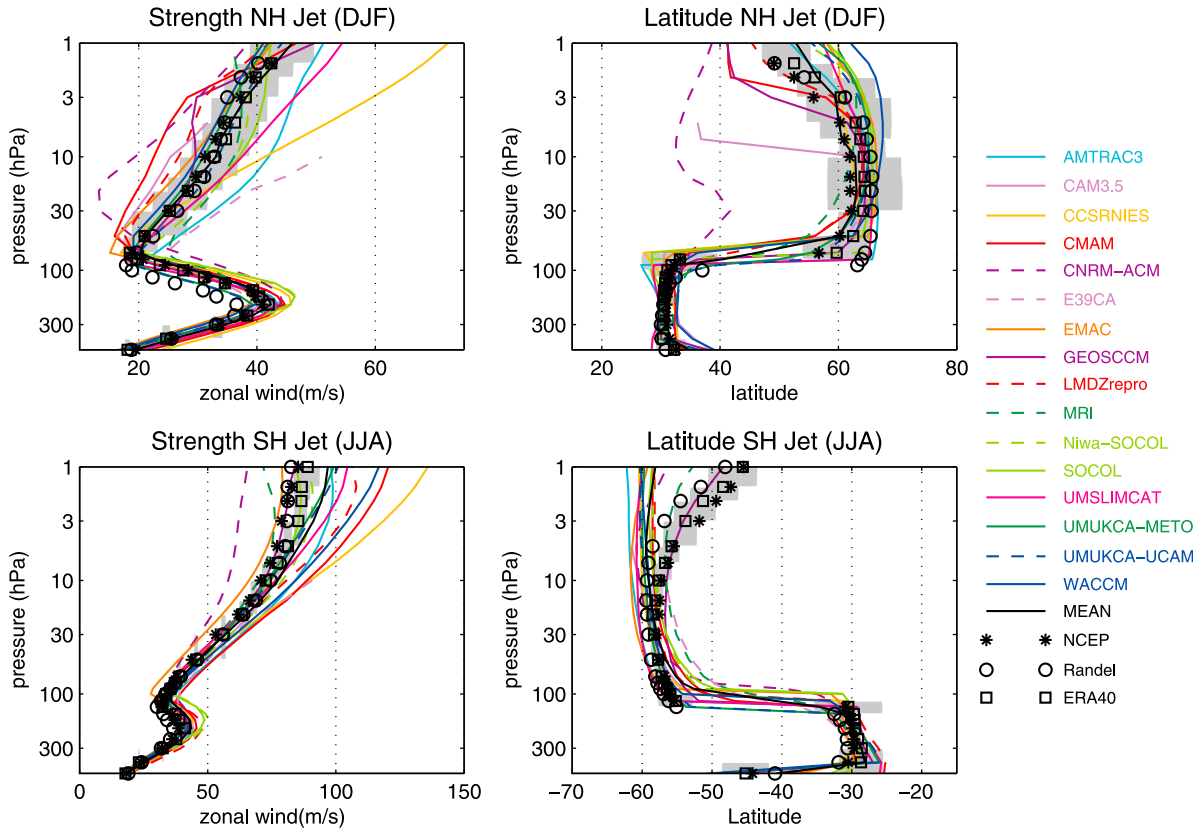

Figure 1. Zonal wind speed and latitude of the jet maxima (top) of the Northern Hemisphere (NH) December to February (DJF) climatology and (bottom) of the Southern Hemisphere (SH) June to August (JJA) climatology in the REF-B1 simulations. Data are based on climatological means for the models, ERA-40 and NCEP data from 1980 to 1999 and on the *Randel et al.* [2004] climatology that represents the time period 1992–1997. The grey shading indicates a 95% confidence interval for the 20-year mean ERA-40 climatology based on a t-distribution. Where an ensemble of simulations is available for a model, quantities are calculated for the ensemble mean zonal-mean zonal wind field.

variability is again assessed by considering the maximum in the interannual standard deviation of the zonal wind and its latitude. Because the maximum interannual variability occurs in high latitudes in the NH winter but is displaced toward midlatitudes in the SH winter (e.g., see Figures 5 and 11 of *Butchart and Austin* [1998]), results are shown in Figure 2 for the regions 45–90°N, and 30–80°S. On average the variability is not as well simulated by the models as the mean climate. For the NH winter, the observations show maximum variability close to the climatological mean jet maximum. All the models fail to capture the equatorward tilt with height for the maximum variability, and in two models the maximum is displaced to the lower middle latitudes in the upper stratosphere. There is also a wide spread among the models in the amplitude of the jet variability with several obvious outliers, most of which have too much variability especially in the upper stratosphere. Only one model exhibits a distinct lack of variability compared to the observations.

[15] For the SH winter, the observations show maximum variability on the equatorward side of the jet, fairly close to the region of the QBO. Most of the models show variability that is too weak and located too far poleward compared to observations.

[16] The nature of the variability of the PNJ can be further isolated by applying an Empirical Orthogonal Function (EOF) analysis to the extratropical zonal-mean zonal wind [e.g., *Feser et al.*, 2000; *Black and McDaniel*, 2009]. Here an EOF analysis is applied at 50 hPa. By considering all months, this analysis captures seasons when the variability maximizes: January to March in the NH and mid-October to mid-December in the SH [*Thompson and Wallace*, 2000]. In general, the models capture this seasonality reasonably well though the period when there is large variability is extended in several of the models compared to the reanalysis (not shown).

[17] In both the reanalysis and the models, the extratropical variability of the zonal-mean zonal wind in the stratosphere can be mainly described by two modes with the first mode dominating. In the reanalysis data the leading mode explains 87% of the variance in the NH. In the SH, both modes contribute, explaining 59% and 35% of the variance, respectively. The leading mode describes the variations in the strength of the eastward PNJ while the second mode represents the meridional shift of the jet. Moreover, because theses two leading modes describe the same two processes (i.e., variations in the jet strength and a meridional shift of

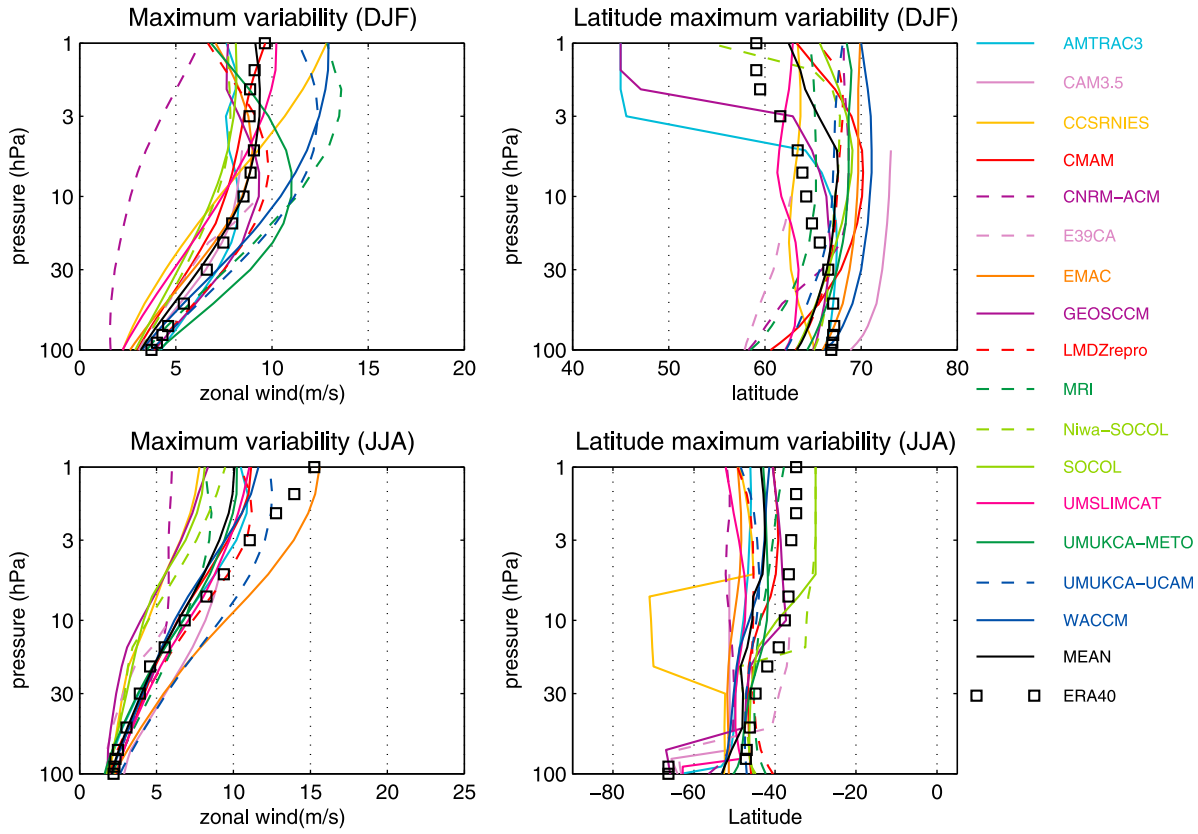

Figure 2. Location and amplitude of the maximum interannual standard deviation of the zonal-mean zonal wind (top) in the NH in DJF poleward of 45°N and (bottom) in the SH in JJA between 30 and 80°S. Data are based on the period 1980–1999 for the models and ERA-40. Where an ensemble of simulations is available for a model, quantities are calculated for the ensemble mean interannual zonal wind standard deviation field.

the jet, respectively) in both the observations and in all the models, meaningful comparisons can be made.

[18] The eigenvalues of the first mode of variability (Figure 3) indicate that for the NH this mode explains a similar amount of the variance in the models as in the reanalysis data, although there is a large intermodel spread. The model ensemble broadly reproduces the structure of the leading EOF and is particularly successful in reproducing the structure of the second EOF (Figure 4).

[19] In the SH, a more significant bias can be identified with the eigenvalue of the leading mode generally much larger for the models than for the reanalysis data (Figure 3), indicating that on average there is too much variance in the strength of the model PNJs. This large variance is accompanied by an overall equatorward bias of the leading EOF pattern in the SH (Figure 4). These results contrast with those for the midwinter interannual variability shown in Figure 2 where the model variability is generally too weak and too far poleward compared to the reanalysis. The differences are a consequence of EOF analysis being dominated by the variability in the late winter and spring.

3.2. Polar Temperatures Biases and PSC Threshold Temperatures

[20] Figure 5 shows the climatological temperature biases over the polar cap in the winter and spring seasons in the NH and SH. *Eyring et al.* [2006] highlighted the contrast between the upper and lower stratosphere in the CCMVal-1 ensemble which remains in the CCMVal-2 ensemble analyzed here. In the upper stratosphere most models lie within the range of temperatures shown in the different analyses in both hemispheres, though there is a large intermodel spread. In the lower stratosphere, where the range of the analyses is much smaller, strong contrasts exist between the two hemispheres, with a clear cold bias for most of the models in the SH spring and a more vertically confined cold bias between 300 and 100 hPa in the NH spring. In the midwinter seasons, the model ensemble generally performs better than in the spring seasons, although there is a cold bias below 200 hPa in the SH.

[21] High-latitude temperature biases can have a large impact on the formation and occurrences of PSCs in the models which are critical for the accurate simulation of polar ozone loss [e.g., *Austin et al.*, 2010a]. Following *Pawson et al.* [1999] and *Austin et al.* [2003], the potential for PSC formation in the models and ERA-40 reanalysis is estimated by calculating for each day the percentage of the horizontal area of the hemisphere where the 50 hPa daily mean temperatures poleward of 60° are below the nitric acid trihydrate (NAT) and ice PSC formation thresholds (195 K and 188 K, respectively). These daily percentage areas are then accumulated over the course of the winter and spring

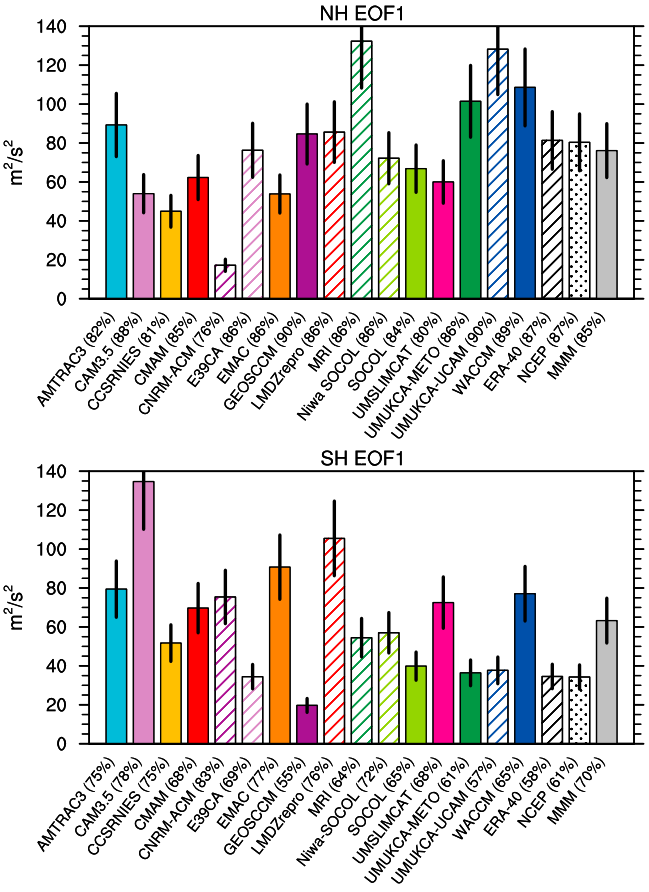

Figure 3. Eigenvalue of the leading mode of variability of the 50 hPa zonal-mean zonal wind ($m^2\ s^{-2}$) for the (top) NH and (bottom) SH. Numbers in brackets (tick labels of the x-axes) indicate the fraction of the total variance explained by the leading mode. Error bars $2\Delta\lambda$ indicate the sampling error determined after *North et al.* [1982]: $\Delta\lambda = \sqrt{2/N}$ where N is the sample size. With $N = 60$, a conservative estimate of the effective sample size is used considering long persistence (2 months) in the stratosphere and weak zonal wind variations during 50% of the year. The EOF analysis was carried out for the NH (SH) 50 hPa zonal-mean zonal wind anomalies poleward of 45°N (S). Monthly mean fields for all months from 1980 to 1999 are included with seasonal cycle and linear trends removed. Data are also weighted with the square root of the cosine of latitude.

(92 days from July to September in the SH and 90 days from December to February in the NH) to provide, for that year, an estimate of the amount of NAT ($\tilde{A}_{NAT}$) and ice ($\tilde{A}_{ice}$) PSCs in units of %-days.

[22] In the Antarctic, the multimodel mean $\tilde{A}_{ice}$ (Figure 6, grey bars) agrees well with the ERA-40 estimate, but the multimodel mean $\tilde{A}_{NAT}$ is significantly smaller than the ERA-40 estimate over the same period. For both $\tilde{A}_{ice}$ and $\tilde{A}_{NAT}$ the spread between the models is small in the SH. In contrast, in the Arctic there are large differences in the simulation of these quantities (Figure 6, right). In general, the models simulate lower values of $\tilde{A}_{NAT}$ and $\tilde{A}_{ice}$ than those derived from the ERA-40 reanalysis with the exception of one model which had a large cold bias in the NH winter (cf. Figures 5 and 6). An important caveat to these conclusions is, however, the known difficulties [e.g., *Manney et al.*, 2003, 2005a, 2005b] in deriving PSC quantities from global assimilation data and the dependence on the analyses or reanalysis dataset used [*Austin and Wilson*, 2010].

3.3. Stationary Waves

[23] At each altitude in the extratropical troposphere and stratosphere the climatological stationary wave field (i.e., the zonally asymmetric part of the climatological mean circulation) is observed to have a well-defined peak in latitude. For the geopotential the latitude of this peak is generally well simulated by the model ensemble during December to February (DJF) in the NH, and September to November (SON) in the SH (see Figure 7, top).

[24] The models have more difficulty in simulating the stratospheric stationary wave amplitude (Figure 7, bottom) with a tendency for the waves to be too weak in the NH winter and a large model spread in amplitudes in the SH spring. The bias in the NH winter extends throughout the year resulting in a relatively weak seasonal cycle of stationary planetary wave amplitude (see Figure 8, which shows the seasonal cycle at 10 hPa). In the SH the amplitude of the seasonal cycle is too large and peaks too early for many of the models. The differences in the seasonal timing in the SH are the main reason for the large spread in the simulations seen in Figure 7. For many of the models, the peak stationary wave is weaker in the NH than in the SH, in contrast to the observations.

[25] The structure of the polar vortex is reflected in the stratospheric stationary wavefield when decomposed into its dominant wave-1 component, which governs the location of the center of the vortex relative to the pole, and its weaker wave-2 component, which further governs the orientation and distortion of the vortex. Figure 9 (top) shows in polar coordinates the amplitude and phase of these components for the 50–70° latitude climatological stationary wave at 10 hPa, for the NH and SH peak periods (the wave-2 amplitude is multiplied by a factor of four for graphical display). The amplitude biases in the figure are consistent with Figures 7 and 8. In the observations, the NH wave-1 component leads to a polar vortex centered off the pole between 0 and 30°E. Most of the models simulate this. The SH wave-1 component is more poorly simulated, corresponding to the fact that the orientation of the Antarctic polar vortex varies significantly among the models. The wave-2 component in both hemispheres is more variable among the models.

[26] A measure of the distortion of the vortex from a simple shifting off the pole is given by the ratio of the wave-2 to wave-1 amplitudes which in the observations is about 25% in the NH and 10% in the SH (see Figure 9, bottom). This ratio is generally well simulated in the NH, with a moderate bias toward small values, but is generally overestimated in the SH, suggesting that the SH vortex in the models is unrealistically distorted from circularity.

3.4. Stratospheric Response to Wave Driving

[27] Probably the most prominent feature of the stratospheric response to wave driving is the Brewer-Dobson

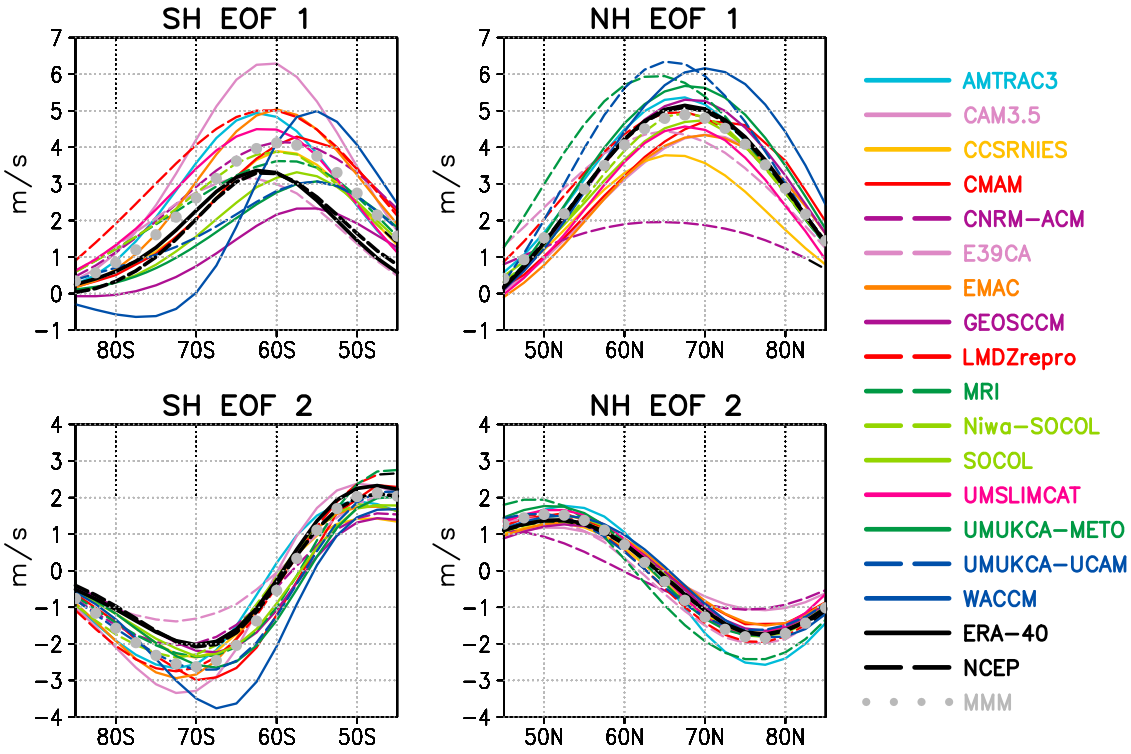

Figure 4. Regression patterns (m s^{-1}) of the (top) first and (bottom) second mode of variability of the 50 hPa zonal-mean zonal wind determined for regions poleward of 45°; (left) SH and (right) NH.

circulation and associated transformed Eulerian mean residual circulation ($\overline{v}^*$, $\overline{w}^*$) in the models [*Andrews et al.*, 1987, chap. 3; *Hardiman et al.*, 2010a, equations (22) and (23)]. A useful measure of the overall strength of this overturning meridional mass circulation is the mass flux entering the stratosphere deduced from the residual vertical velocity, $\overline{w}^*$, just above the tropical tropopause [*Butchart and Scaife*, 2001]. In the REF-B1 simulations the latitudinal distributions of $\overline{w}^*$ at 70 hPa and between 40°S and 40°N are remarkably similar to that derived from the ERA-Interim reanalysis (not shown) and also the UKMO analyses (see thick dashed line in Figure 2 of *Butchart et al.* [2006]), though in the models $\overline{w}^*$ is more symmetric across the equator. All but one of the models accurately reproduce the locations of the "turn-around latitudes" where $\overline{w}^*$ is zero (i.e., the latitudes where the tropical upwelling changes to extratropical downwelling) and the annual cycle in the integrated upward mass flux between these turn-around latitudes was also generally well reproduced.

[28] On average the annual mean tropical upwelling mass fluxes at 70 and 10 hPa in the REF-B1 simulations agree with the mass fluxes derived from the ERA-Interim reanalysis (Figure 10, black bars, see caption for details), with the standard error in the multimodel mean less than the interannual variability in the analyzed mass fluxes (not shown). Following *Butchart et al.* [2010a], corresponding "Downward Control" [*Haynes et al.*, 1991] estimates of the upwelling mass fluxes are shown by the grey bars in Figure 10 (again see caption for details) and agree reasonably well with actual mass fluxes derived from the residual vertical velocities $\overline{w}^*$ shown by the black bars. Apart from in one outlying model, parameterized orographic gravity wave drag (OGWD) contributes significantly to the downward control estimate (for the five models that supplied OGWD data) and, on average, accounts for 21.1% of the driving of the upwelling at 70 hPa decreasing to 4.7% at 10 hPa (Figure 10). At 70 hPa the resolved waves accounted for 70.7% (71.6% at 10 hPa) and NOGWD 7.1% (10.9% at 10 hPa) of the driving, again with the NOGWD contribution averaged only over the four models which provided these diagnostics. In general, however, there was a wide spread between the models in the contributions from the different types of wave drag (i.e., drag from the resolved waves, OGWD and NOGWD). At 70 hPa the contributions from the resolved waves ranged from 51.0% to 102.7% (74.7% if the one outlying model is excluded) while the range for OGWD and NOGWD was 2.0 to 40.9% and −3.4 to 16.8%, respectively.

[29] For each model the ratio of the upwelling (as calculated from $\overline{w}^*$) at 10 hPa to that at 70 hPa (weighted by the multimodel mean at each altitude) provides a measure of the net entrainment out of the tropical pipe in the lower stratosphere with respect to the multimodel mean [*Neu and Plumb*, 1999]. When there is no mixing from midlatitudes into the tropics the ratio reduces to a measure of the horizontal transport across the subtropical barrier. In the models the ratio ranges from 90% to 115% of the multimodel average (Figure 10b, see caption for details) indicating much less spread between models than is obtained from tracer-based measures of subtropical transport [*Neu et al.*, 2010].

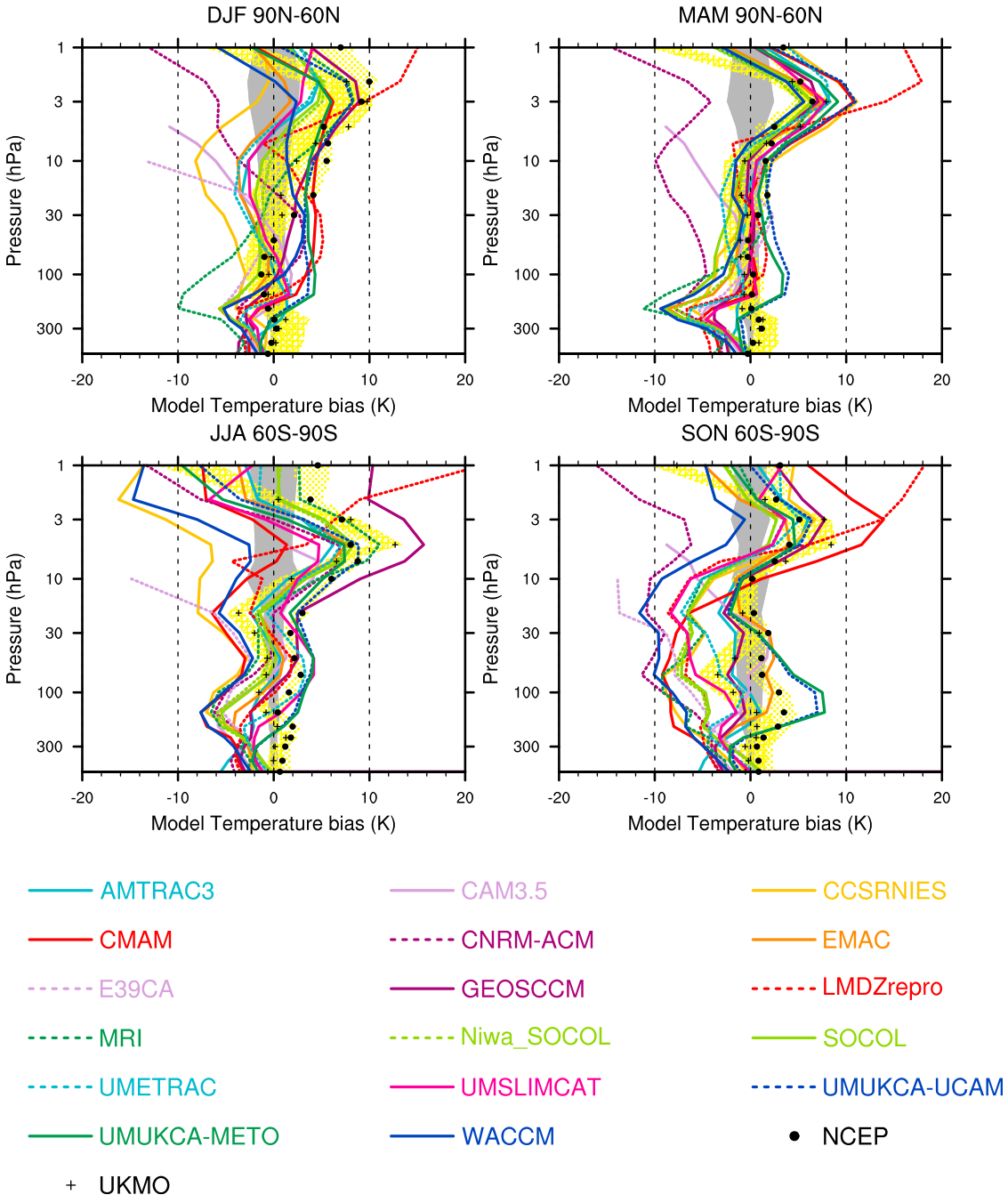

Figure 5. Climatological mean temperature biases for (top) 60–90°N and (bottom) 60–90°S for the (left) winter and (right) spring seasons. The climatological means for the models, ERA-40 and NCEP data from 1980 to 1999 and for UKMO from 1992 to 2001 are included. Biases are calculated relative to ERA-40 reanalyses for 1980–1999. The grey (yellow) area shows a 95% confidence interval for the 20-year mean (10-year mean for UKMO) from the ERA-40 (NCEP and UKMO) reanalyses based on a t-distribution.

Most likely this is a consequence of differences in the relative leakiness of the tropical pipes in the models.

[30] For all the tropical upwelling diagnostics presented above broadly similar conclusions were obtained if the REF-B2 rather than the REF-B1 simulations were used, suggesting that the conclusions for the multiyear mean upwelling are not sensitive to the choice of SST forcing data.

[31] The forcing from upward propagating waves also affects the polar stratosphere. One manifestation of this forcing is the approximate correlation between the eddy meridional heat flux (a proxy for the upward flux of wave activity) at 100 hPa averaged over a band between 40° and 80°N (40°–80°S) during January and February (July and August) and the subsequent temperature of the polar cap at 50 hPa in February and March (August and September), first

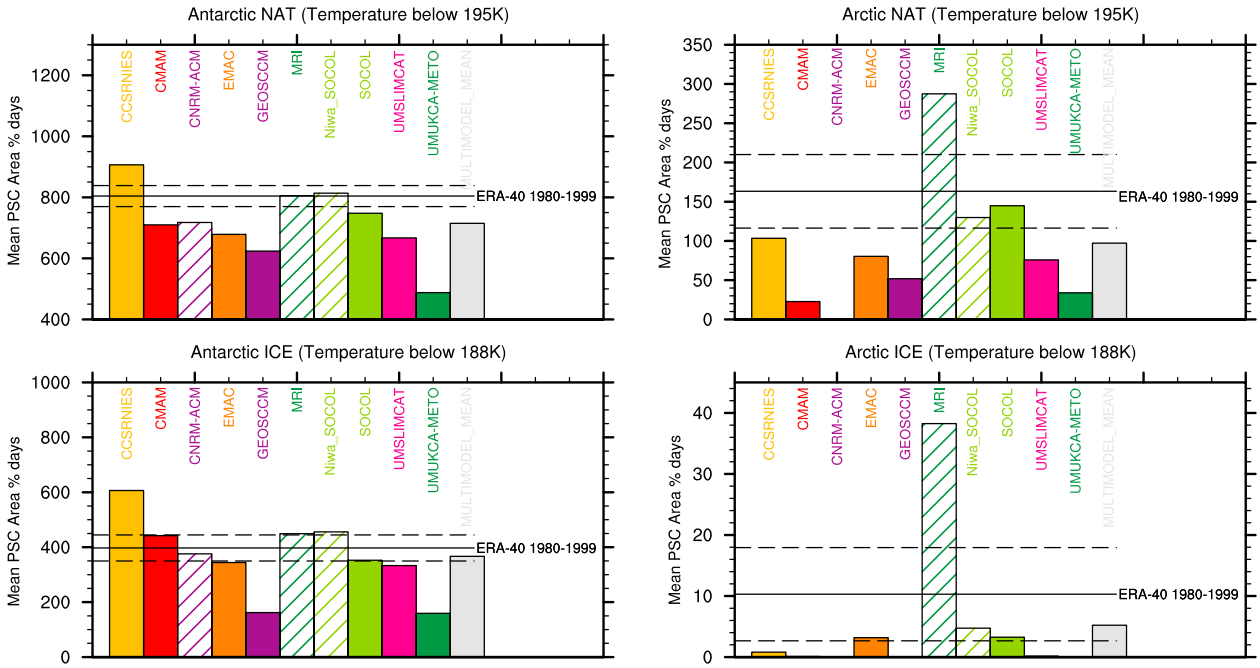

Figure 6. Mean (1980–1999) for (left) the Antarctic and (right) the Arctic of the seasonally accumulated area at 50 hPa where daily temperatures are (top) below 195 K (approximate threshold temperature for NAT formation) and (bottom) below 188 K (approximate threshold temperature for ice formation). Dashed black line is for ERA-40 reanalysis (1980–1999). The units are the percentage of the hemisphere where the daily temperature is below the threshold multiplied by the duration in days.

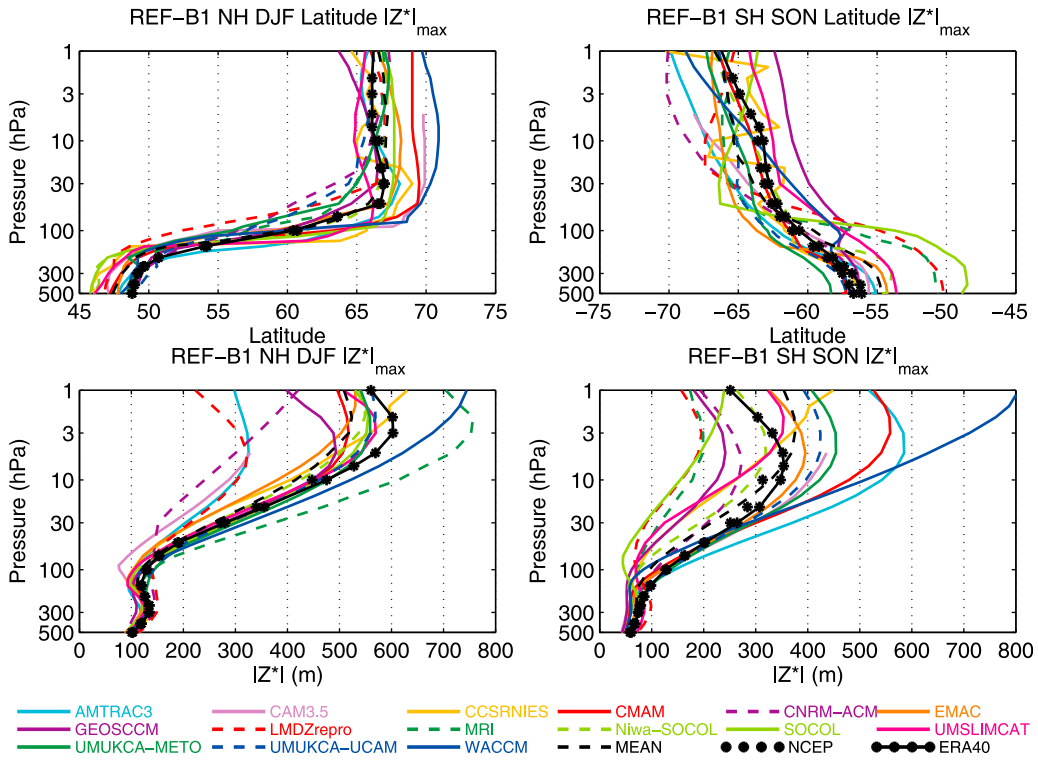

Figure 7. Latitudinal location and value of the maximum amplitude of the stationary wavefield (left) for the NH DJF climatology and (right) for the SH SON climatology. Data are based on climatological means for the models, ERA-40 and NCEP data from 1980 to 1999. The stationary wave amplitude is defined as the zonal root-mean square of the zonally asymmetric climatological geopotential height. Cubic spline interpolation is used to determine the latitude of the maximum and its value from the gridded data. The black dashed curve is the mean of all the model curves.

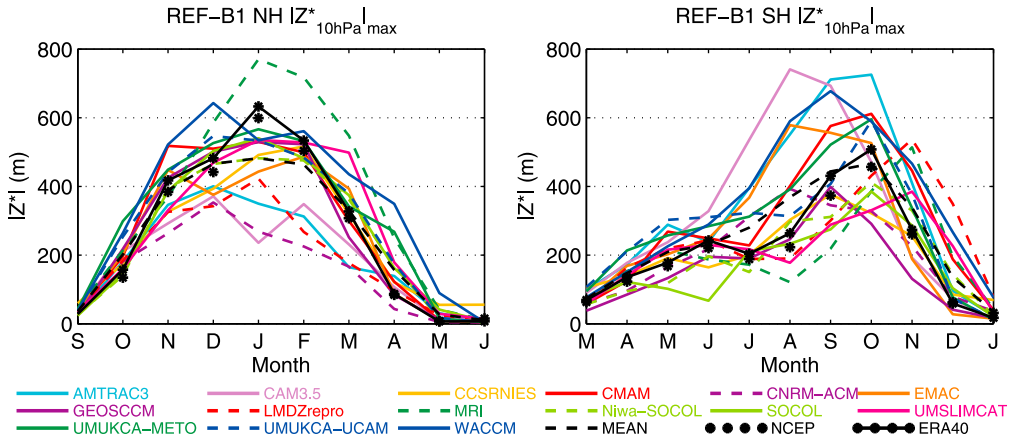

Figure 8. Seasonal variation of the maximum amplitude of the (left) NH and (right) SH 10 hPa climatological stationary wave. Data are based on climatological means for the models, ERA-40 and NCEP data from 1980 to 1999. Cubic spline interpolation is used to determine the maximum value, as in Figure 7. The black dashed curve is the mean of all the model curves.

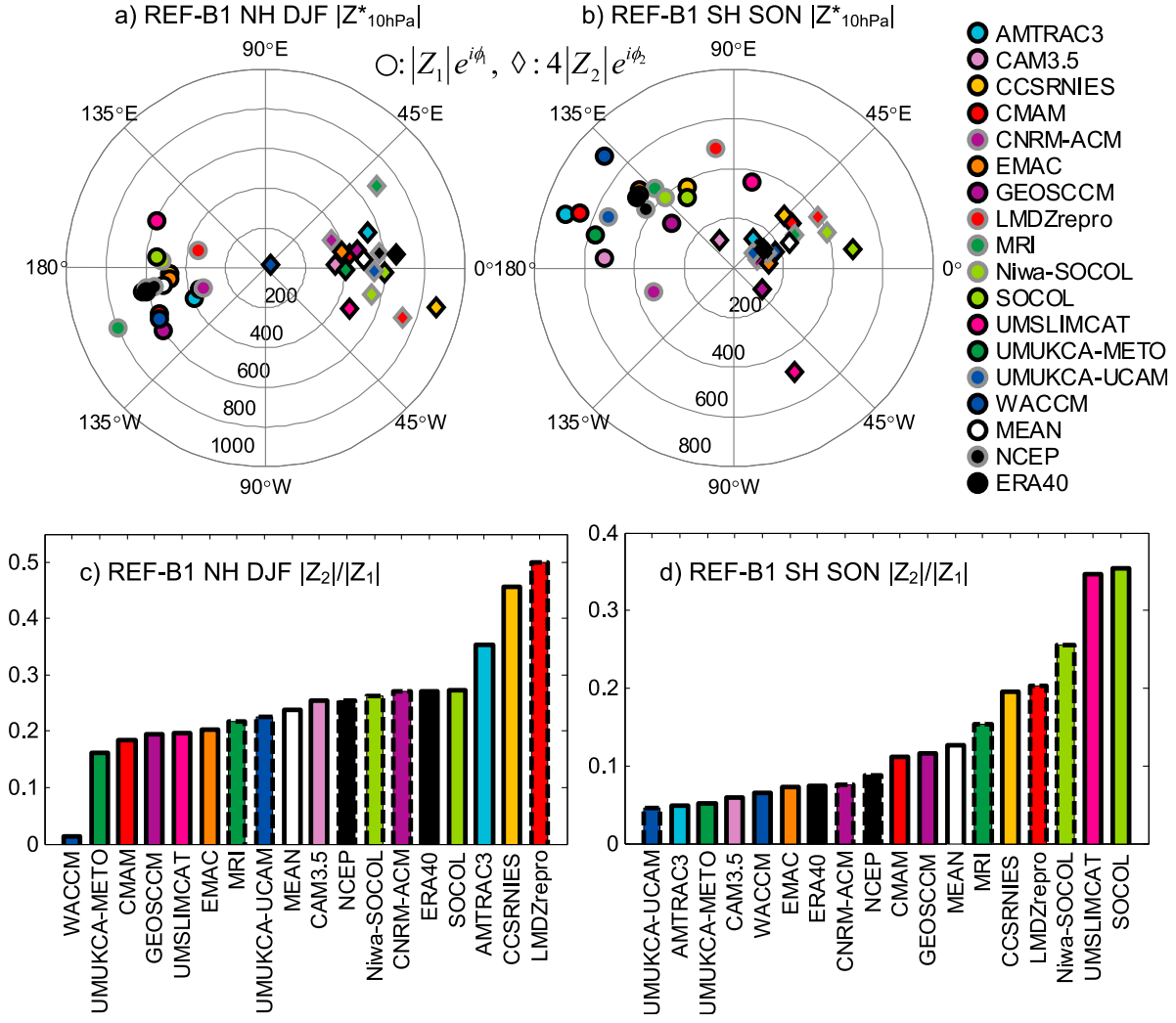

Figure 9. (a) Phase in degrees and amplitude (contour interval 200 m), in polar coordinates, of wave-1 (circles) and wave-2 (diamonds) 10 hPa DJF stationary waves for the NH. The wave-2 amplitude has been multiplied by a factor of four. (b) As in Figure 9a, for the SH SON. (c) Ratio of wave-2 to wave-1 amplitude at 10 hPa for the NH DJF. (d) As in Figure 9c, for the SH SON. Data are based on climatological means for the model REF-B1 simulations, ERA-40 and NCEP data from 1980 to 1999.

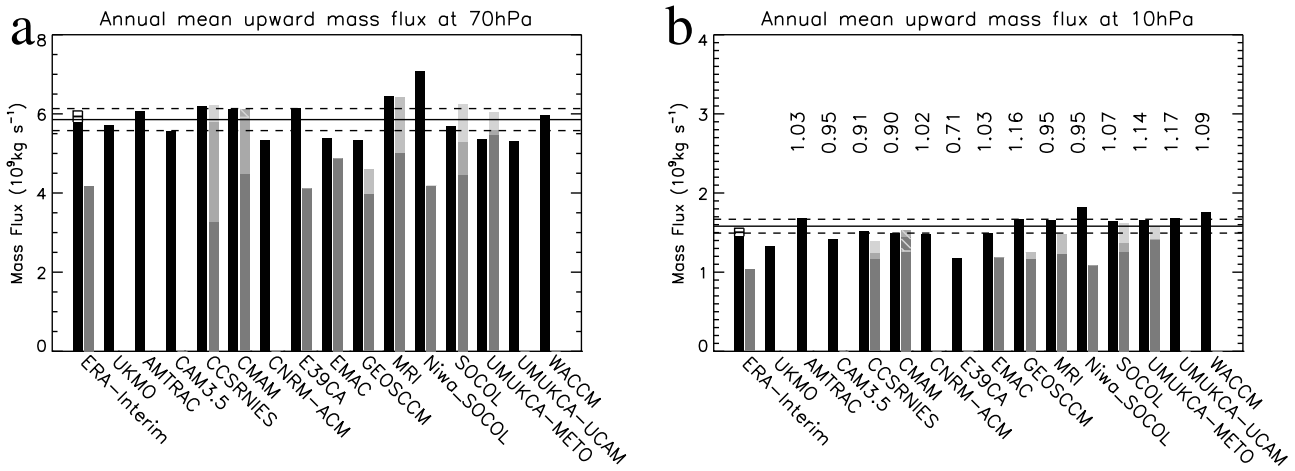

Figure 10. Annual mean upward mass flux averaged from 1980–1999 for the REF-B1 simulations, ERA-Interim reanalysis averaged from 1989–2009 and UKMO analyses from 1992–2001. Upwelling calculated from $\overline{w}^*$ is shown by black bars. Upwelling calculated by downward control is split into contributions from resolved waves (dark grey), orographic gravity wave drag (OGWD) (grey), and nonorographic gravity wave drag (NOGWD) (light grey). OGWD and NOGWD are shown combined for the GEOSCCM and MRI model. For some models and for the ERA-Interim reanalysis only the resolved wave contributions are shown. In the CMAM, NOGWD produces a negative upwelling and so cancels some of the upwelling produced by the OGWD and the resolved waves. This cancellation is shown by diagonal lines. The black horizontal lines show the multimodel mean and the intermodel standard error. The interannual standard error for the ERA-Interim reanalysis is shown by the unshaded part of the bar with the horizontal line at the midpoint being the multiyear mean. Values shown at (a) 70 hPa and (b) 10 hPa. The numbers above the bars in Figure 10b are the ratio for that model of the upwelling mass flux (normalized by the multimodel mean) at 10 hPa to upwelling mass flux (normalized by the multimodel mean) at 70 hPa.

noted by *Newman et al.* [2001] using reanalysis data for the NH.

[32] A succinct way of comparing and evaluating the different models is to plot the parameters of linear fits to scatterplots of 100 hPa meridional heat flux versus 50 hPa temperatures (Figure 11, see caption for details). The intercept of the regression line (x axis) gives an indication of the temperature that the polar cap would have if no resolved wave-driving were present. The slope of the regression line (y axis) gives an indication of the sensitivity of the stratospheric temperature response to changes in the wave forcing or, more particularly, the flux of wave activity from the troposphere.

[33] In the NH, the multimodel mean linear fit parameters are within sampling uncertainty of the linear fit parameters in the ERA-40 reanalysis, with only one outlier. In general in the NH, the cluster of model points is shifted toward the upper left quadrant of the plot, indicating a tendency toward lower polar temperatures and an enhanced response of the lower stratosphere to tropospheric wave-driving. The tendency toward a cold bias in the lower stratosphere during spring is consistent with previous model assessments and with Figure 5.

[34] In the SH, although there is a much larger spread than in the NH, the multimodel mean linear fit parameters are again within sampling uncertainty of the linear fit parameters in the ERA-40 reanalysis. Several of the models show properties statistically distinct from those in the ERA-40 reanalysis and the large spread is probably due to the large differences in the simulated midwinter ozone during 1980–1999 [*Austin et al.*, 2010b] affecting the dynamics of the models.

3.5. Intraseasonal Variability

[35] In the extratropical regions major stratospheric sudden warmings (SSWs) are an important component of the intraseasonal variability which contribute significantly to determining the mean climate. In the simulations major SSWs are identified using the methodology of *Charlton and Polvani* [2007], based on reversals of the zonal-mean zonal wind at 60°N and 10 hPa, for the months November to March. Unlike previous model intercomparisons [*Charlton et al.*, 2007] most models produce approximately the correct number of major SSWs over the period 1960 to 2000 (note the use of the longer period to account for the large interannual standard deviation), with the model ensemble mean frequency very close to the ERA-40 climatological frequency (see Figure 12). At the 95% confidence level two models had a lower frequency of major SSWs compared to the reanalysis and one a higher frequency with a resultant midwinter stratospheric jet of significantly reduced strength. Apart from one model there was little systematic difference between the frequency of major SSWs in the REF-B1 and REF-B2 simulations, suggesting little sensitivity to the choice of SST forcing. On the other hand there were large differences between the models, though in all cases, 95% confidence intervals for the major SSW frequency analyzed for ERA-40 overlap for the REF-B1 and REF-B2 simulations (again see Figure 12). The SH winter period was also analyzed between 1960 and 2000 but no examples of a

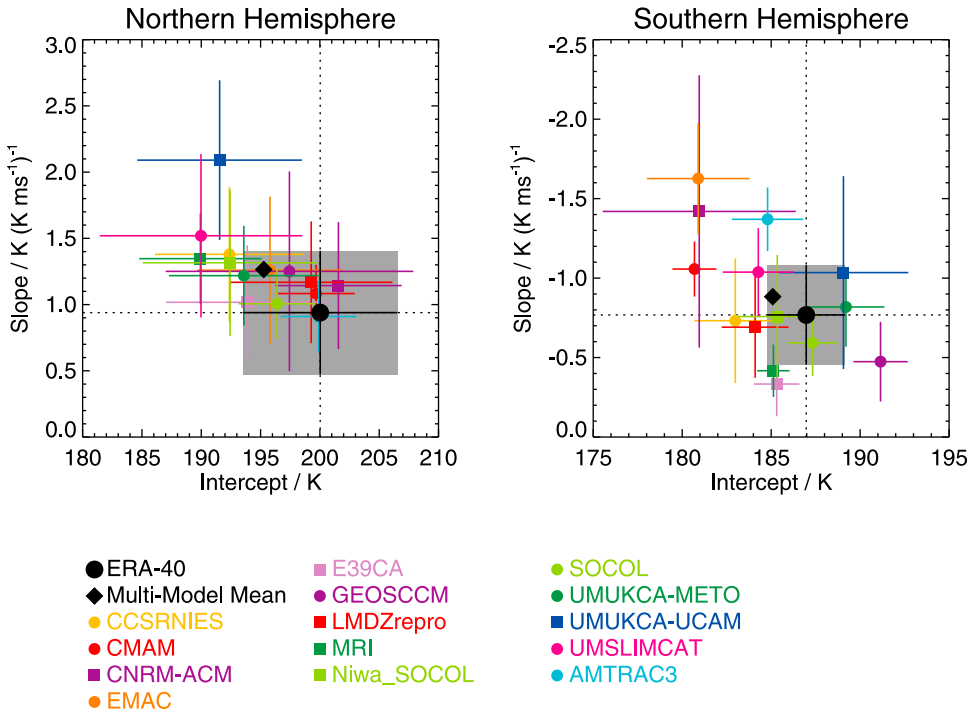

Figure 11. Parameters of the linear fit to the scatterplot of the 100 hPa meridional heat flux versus the 50 hPa temperature (for more details of the procedure see *Newman et al.* [2001] and *Eyring et al.* [2006]). Shown is the intercept of the linear fit (x-axis) plotted against the slope of the regression line (y-axis) for the NH and the SH. Black symbols show the same diagnostic for the ERA-40 reanalysis data. Estimates of 95% confidence limits for the two parameters are included for each estimate in the thin colored lines. Grey shading indicates the 95% confidence estimates for the ERA-40 reanalysis data.

major SSW, similar to that observed during September 2002, were simulated in any model (using the same criteria for major SSW occurrence as for the NH).

[36] Studies by *Black et al.* [2006] and *Black and McDaniel* [2007a, 2007b] (hereafter BM) have shown that there is an important dynamical link between the stratosphere and troposphere as the final warming takes place and that the timing of the final warming is highly variable from year to year. Final warming dates in both hemispheres were calculated using the BM method which defines the final warming as occurring when zonal-mean zonal winds at a specified latitude fall below a low-wind threshold (0 ms^{-1} in the NH and 10 ms^{-1} in the SH) and do not return to values above the threshold before the next winter (see BM for further details). For some models, the zonal-mean zonal winds never become westward in some years; these years are ignored in the analysis. In both hemispheres the models generally have final warming dates either at or later than the date obtained from the ERA-40 re-analysis data for the period 1980–1999 (see Figure 13). In SH over half the models shown in Figure 13 had mean final warmings later than observed and in both hemispheres the multimodel mean estimate of the final warming date is significantly later than observed.

[37] A useful comparison in the SH can be made with diagnostics of the climatological descent of the zero wind line (Figure 14), which was calculated for the previous intercomparison by *Eyring et al.* [2006]. Results from the CCMVal-2 model ensemble and the CCMVal-1 ensemble shown by Eyring et al. are very similar, both showing a delayed or missing transition to westward winds in the zonal wind climatology in the SH spring in many of the models which is consistent with the spring time temperature biases noted in section 3.2. Models in which a late final warming is observed in the SH generally also have a late climatological transition of the zonal winds at 60°S.

3.6. Tropical Variability

[38] Vertical profiles of the interannual standard deviation in the detrended zonal-mean zonal wind averaged between 10°S and 10°N in the REF-B1 simulations are shown in Figures 15a and 15b. Below ~48 km (~1 hPa) nearly all the models underestimate tropical variability in comparison to ERA-40. Five models exhibit particularly low stratospheric variability, largely due to the absence of either an internally generated or artificially prescribed QBO.

[39] Figures 15c and 15d show the vertical profiles of the amplitude of the variability in zonal wind at periods between 2 and 5 years (see caption for details). This range of periods captures possible QBO-like variability and it is evident from the figure which models neither prescribe nor internally generate a QBO (see *Morgenstern et al.* [2010] for details of the models). Interestingly enough, there are still differences seen between those models which prescribe a QBO, possibly related to the fact that these models do not include any feedback mechanisms between the simulated ozone and the imposed artificial forcings. Furthermore, nearly all models

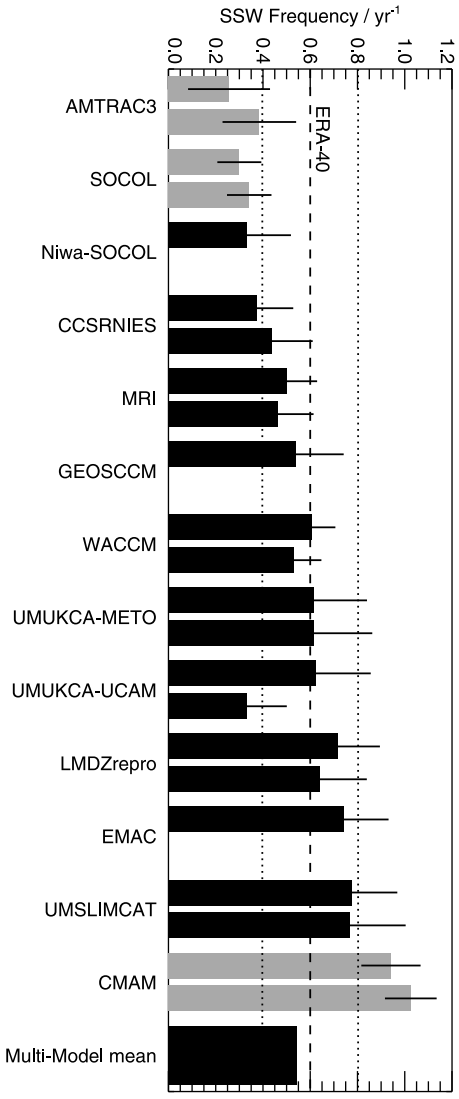

Figure 12. Mean frequency of major SSWs per year for REF-B1 and REF-B2 simulations between 1960 and 2000. Dotted black line shows the mean frequency in the ERA-40 data set (1960–2000) and 95% confidence interval (dotted lines). For each model, the upper bars indicate REF-B1 simulations and the lower bars indicate REF-B2 simulations. Where ensemble simulations are available, the mean frequency is calculated by combining all ensemble members. Bars are sorted according to the major SSW frequency in the REF-B1 simulations. Where the frequency of SSWs in the model and ERA-40 data set is significantly different at 95% confidence the bars are shown in grey. Whiskers on each bar indicate a 95% confidence interval for the major SSW frequency.

show a weaker peak amplitude for the QBO compared with ERA-40.

[40] Unlike for the QBO, peak amplitudes of the SAO in the models are spread about the amplitude seen for ERA-40 (Figures 15e and 15f). For the two models which overestimate the SAO amplitude by the largest amount the bias is most likely a consequence of their lack of a QBO: the QBO in the lower stratosphere winds would act periodically to filter out small-scale gravity waves, which would otherwise drive the eastward phase of the SAO. However, the significance of any net model bias above ~32 km (10 hPa) has to be treated with caution due to the paucity of observations assimilated there by ERA-40.

[41] The amplitude of the annual cycle in tropical zonal-mean zonal wind in the REF-B1 simulations is shown in Figures 15g and 15h. The amplitude of the ERA-40 annual cycle shows two peaks: in the upper troposphere and at the stratopause. All the models exhibit a peak in the amplitude in the upper troposphere but with one model having unrealistically small amplitudes. All the models significantly underestimate the amplitude of the annual cycle near the stratopause. On the basis of results from a high-top version of the Met Office's global climate model *Osprey et al.* [2010] argue this may be linked to an overly strong SAO and SH summer jet and stronger than observed westward circulation during June to August (JJA), though this may be only relevant to those models having an overly strong SAO.

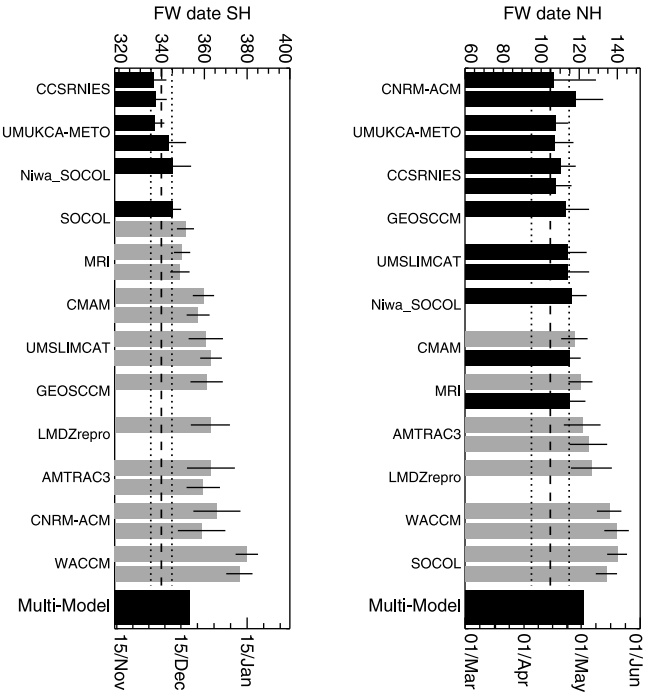

Figure 13. Mean date of the final warming (day number) for the REF-B1 (upper bars for each model) and REF-B2 (lower bars for each model) simulations (1980–1999). Black dashed line shows the mean final warming date for the ERA-40 data with 95% confidence estimates shown in dotted lines. Models are ordered by the mean date of their final warming in the REF-B1 simulation. Where a significant difference between models and the ERA-40 reanalysis estimate is observed the bar is plotted in grey. Where an ensemble of simulations is available, the statistic reflects the mean of all three ensemble members. Black whiskers on each bar indicate twice the standard error for each estimate. Approximate comparable calendar dates for a nonleap year are included on the bottom axis.

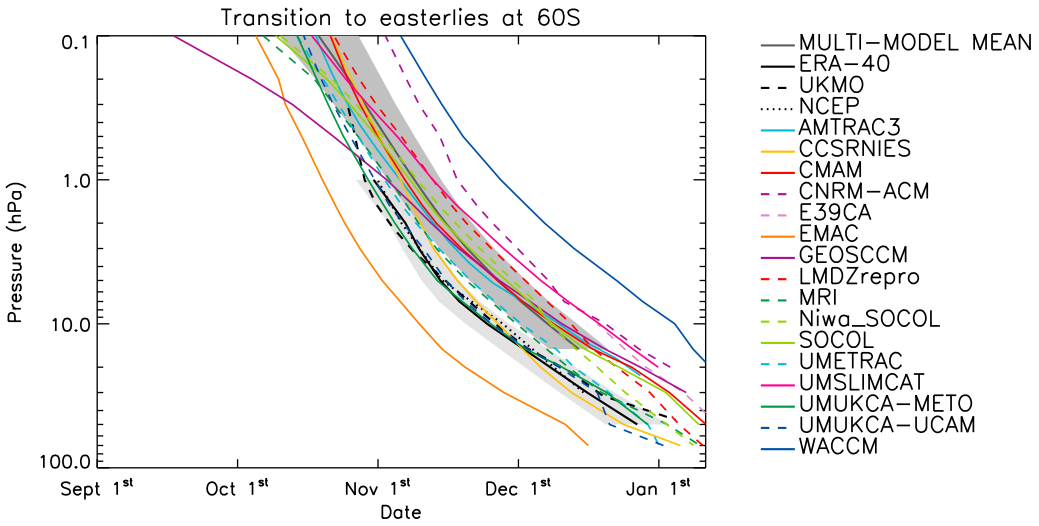

Figure 14. Descent of the zero zonal-mean zonal wind at 60°S based on the climatological mean annual cycle calculated from the monthly and zonal-mean zonal winds. The dark grey area shows a 95% confidence interval for the intermodel standard error, and the light grey area shows a 95% confidence interval for the 20-year mean ERA-40 transition, based on a t-distribution. Climatological means are calculated for the same period as in Figure 1.

[42] A brief comparison of the variability in the zonal wind in the tropics in the REF-B1 and REF-B2 simulations from 1980–2000 shows differences throughout the stratosphere, which are associated with a lack of a QBO in most of the REF-B2 simulations and a strengthened SAO (not shown). Like the REF-B1 ensemble, all REF-B2 simulations exhibit a poor annual cycle in the upper stratosphere.

3.7. Synopsis and Comparison to Previous Multimodel Assessments

[43] It is clear from the above results that the models, on average, perform well in simulating most aspects of mean climate of the stratosphere. There are, however, some stratospheric processes and phenomena in which there are significant consistent biases in most of the models. In particular, these include the springtime cold bias in the lower stratosphere and general delay in the winter to summer transition in many of the models. In comparison with previous multimodel assessments, the overall simulation of stratospheric climate has on average improved over the 10 years or so since *Pawson et al.* [2000], most notably due to the introduction of parameterized NOGWD [*Austin et al.*, 2003]. On the other hand, there is no clear evidence that there has been a systematic improvement in the simulation of stratospheric climate between the current generation of CCMs and those assessed by *Eyring et al.* [2006], i.e., between CCMVal-1 and CCMVal-2.

[44] The present study, nonetheless, advances that of *Eyring et al.* [2006] with a comprehensive intercomparison of the intraseasonal to interannual variability and the zonally asymmetric component of the circulation. In general, the variability was not as well reproduced by the models as the time-mean climate. This was a particularly acute problem in the tropics where nearly all the models under represent the strength of the QBO despite many of them artificially imposing it. Indeed even when the QBO was imposed there was an unexpected spread in tropical zonal wind variability. A weak tropical annual cycle in the zonal-mean zonal wind was common across all models too. In the extratropics there are some clear links between diagnostics of stratospheric variability and persistent biases in the models, for example between the late final warming in many models and the cold bias in the spring time lower stratosphere. The multimodel assessment also indicated common deficiencies and uncertainties in simulating the zonally asymmetric component of the flow. In the NH the circulation is on average too zonal whereas in the SH there was a wide spread in the orientation of the polar vortex.

Figure 15. Profiles of (a and b) the interannual standard deviation, (c and d) the amplitude of the "QBO" (i.e., coherent variability with periods between 2 and 5 years), (e and f) the amplitude of SAO, and (g and h) the amplitude of the annual cycle in the detrended zonal-mean zonal wind averaged from 10°S–10°N for the full period of the REF-B1 simulations and ERA-40 reanalysis. Methodology is similar to that in the work of *Pascoe et al.* [2005]. The amplitude is the ratio of the definite integral of the zonal mean power spectrum to the standard deviation of the zonal-mean zonal wind for periods between 2 and 5 years (Figures 15c and 15d), the 6 month harmonic (Figures 15e and 15f), and the 12 month harmonic (Figures 15g and 15h). Linear trends were first fitted to and then removed from the data. An asterisk after a model name indicates that the model has an externally forced (i.e., artificial) QBO. For clarity the model results are split into right-hand and left-hand panels.

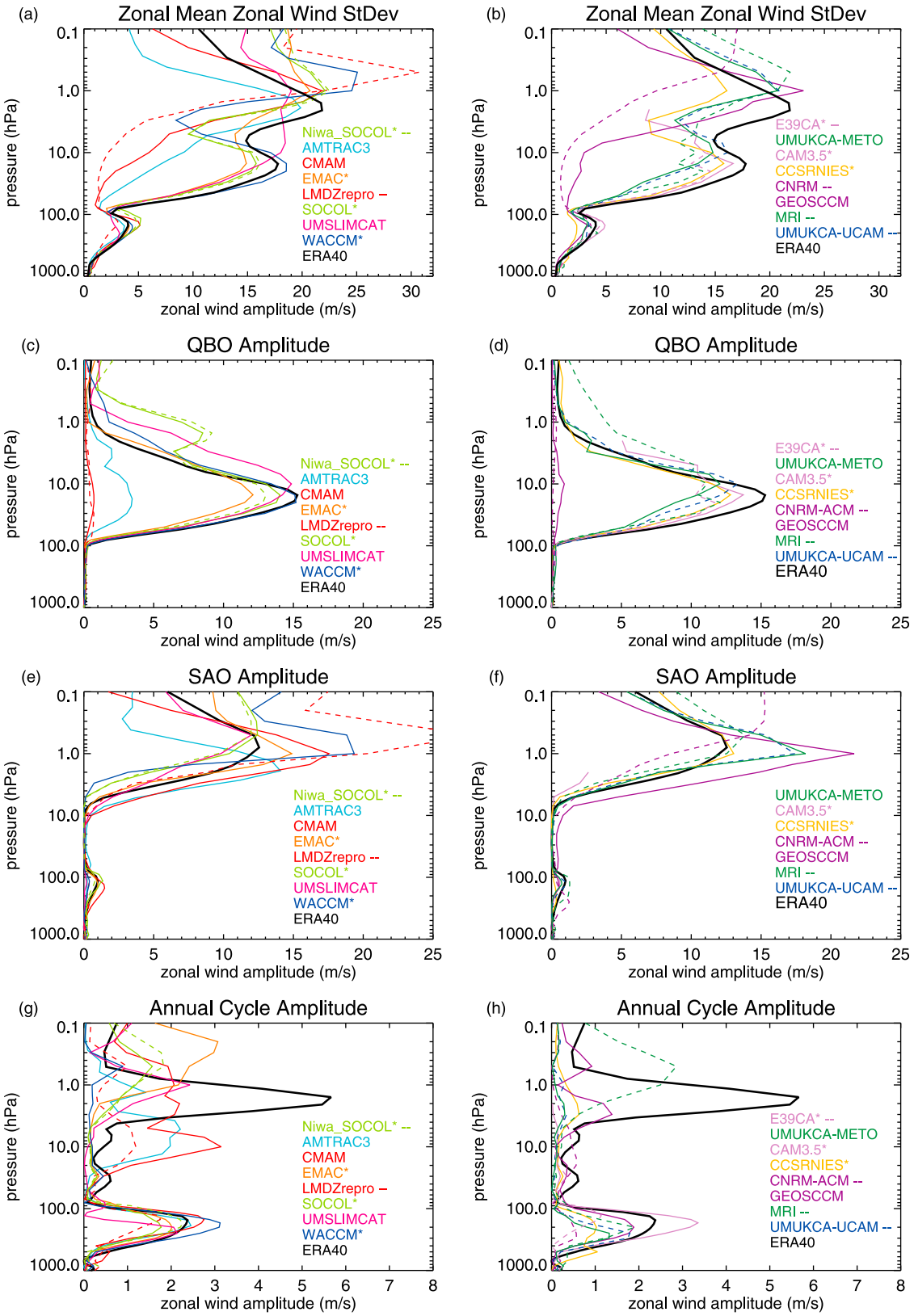
(a)
Zonal Mean Zonal Wind StDev
pressure (hPa)
zonal wind amplitude (m/s)
Niwa_SOCOL*
AMTRAC3
CMAM
EMAC*
LMDZrepro
SOCOL*
UMSLIMCAT
WACCM*
ERA40
(b)
Zonal Mean Zonal Wind StDev
E39CA*
UMUKCA-METO
CAM3.5*
CCSRNIES*
CNRM
GEOSCCM
MRI
UMUKCA-UCAM
ERA40
(c)
QBO Amplitude
(d)
QBO Amplitude
(e)
SAO Amplitude
(f)
SAO Amplitude
(g)
Annual Cycle Amplitude
(h)
Annual Cycle Amplitude

Figure 15

Table 2. Processes and/or Phenomena Validated Using Metrics[a]

Phenomena Process	Metric: Description	Metric: Name
	Mean Climate	
Zonal mean climatology	60–90°N DJF temperatures at 50 hPa	**tmp_nh**
	60–90°S SON temperatures at 50 hPa	**tmp_sh**
	Maximum NH eastward wind in DJF at 10 hPa	**umx_nh**
	Maximum SH eastward wind in JJA at 10 hPa	**umx_sh**
Brewer-Dobson circulation	Tropical upwelling mass flux at 70 & 10 hPa	**up_70**
		up_10
Extratropical wave driving	Slope of the regression of the February and March 50 hPa temperatures 60–90°N on the 100 hPa January and February heat flux 40–80°N	**PW_nh**
	Slope of the regression of the August and September 50 hPa temperatures 60–90°S on the 100 hPa July and August heat flux 40–80°S	**PW_sh**
	Climate Variability (Intraseasonal–Interannual)	
Extratropical variability	Amplitude of the leading mode of variability (EOF) of the 50 hPa zonal-mean zonal wind for NH and SH	**fev_nh** **fev_sh**
Tropical variability	Amplitude of the annual-cycle at 2 hPa in the zonal-mean zonal wind, 10°S–10°N	**tann**
	Amplitude of the SAO at 1 hPa in the zonal-mean zonal wind, 10°S–10°N	**sao**
	Amplitude of "QBO" at 20 hPa in the zonal-mean zonal wind, 10°S–10°N	**qbo**
Stratospheric sudden warmings	Frequency per year of major stratospheric sudden warmings, defined using reversal of the zonal-mean zonal wind at 10 hPa, 60°N	**SSW**
Final warming	Mean date of the NH final warmings at 50 hPa, 60°N defined using the criteria of *Black and McDaniel* [2007a, 2007b]	**fw_nh**
	Mean date of the SH final warmings at 50 hPa, 70°S defined using the criteria of *Black and McDaniel* [2007a, 2007b]	**fw_sh**

[a]The first column lists the processes and phenomena with the chosen metrics given in columns 2 and 3. Abbreviations: NH=Northern Hemisphere; SH=Southern Hemisphere; DJF=December–January–February; JJA=June–July–August; SON=September–October–November; EOF=empirical orthogonal function; SAO=semiannual oscillation; QBO=quasi-biennial oscillation.

[45] While the above qualitative analysis enables a detailed examination of individual processes within the models, assessment of the relative severity of model biases and possible links between biases is difficult. One approach to comparing model performance across a range of different processes is to define and calculate metrics of model performance [e.g., *Waugh and Eyring*, 2008].

4. Quantitative Assessment: Metrics

[46] To establish the fidelity and quantify the assessment of the simulations "metrics" representing many of the key stratospheric dynamical processes have been identified (see Table 2). The list has some metrics in common with *Waugh and Eyring* [2008] but also extends that list particularly in the area of stratospheric variability. A pragmatic approach has, however, been used and for many diagnostics the metrics opted for require the least input of dynamical fields or complex analysis and thus are available for a greater range of models.

[47] As in the previous section the aim is to assess the performance of the model ensemble and provide a guide to the overall performance of the models in several key areas. Again the analysis is not concerned with identifying the performance of any of the models in particular. Because of this, models which did not provide enough data to fully assess a significant proportion of the metrics in Table 2 (particularly the CAM3.5 and E39CA model) are excluded from this analysis. This minimizes any potential bias between metrics which might result from changing the composition of the multimodel ensemble for each diagnostic.

4.1. Metric Calculation

[48] Model validation metrics are calculated using equation (4) of *Waugh and Eyring* [2008]:

$$g = 1 - \frac{1}{n_g} \frac{|\mu_{\text{model}} - \mu_{\text{obs}}|}{\sigma_{\text{obs}}}$$

where μ_{model} and μ_{obs} are the model and observational estimates of each diagnostic, respectively, σ_{obs} is the interannual standard deviation of the observations, and n_g is a scaling parameter. For consistency with the *Waugh and Eyring* [2008] analysis, scores are standardized using the interannual standard deviation of the observed quantity in question and the parameter n_g is set to 3. Where $g \leq 0$ (i.e., the diagnostic is different from the observational estimate by more than three standard deviations), the value of the metric is set to zero.

[49] Perhaps a more natural normalization to use instead of the interannual standard deviation would be to use the standard error inherent in an estimate of the quantity and include some estimate of the observational uncertainty. Since in this study all of the metrics, except for the one for SSWs and the tropical variability metrics, are calculated for the same 20 year period (1980–1999), using the standard

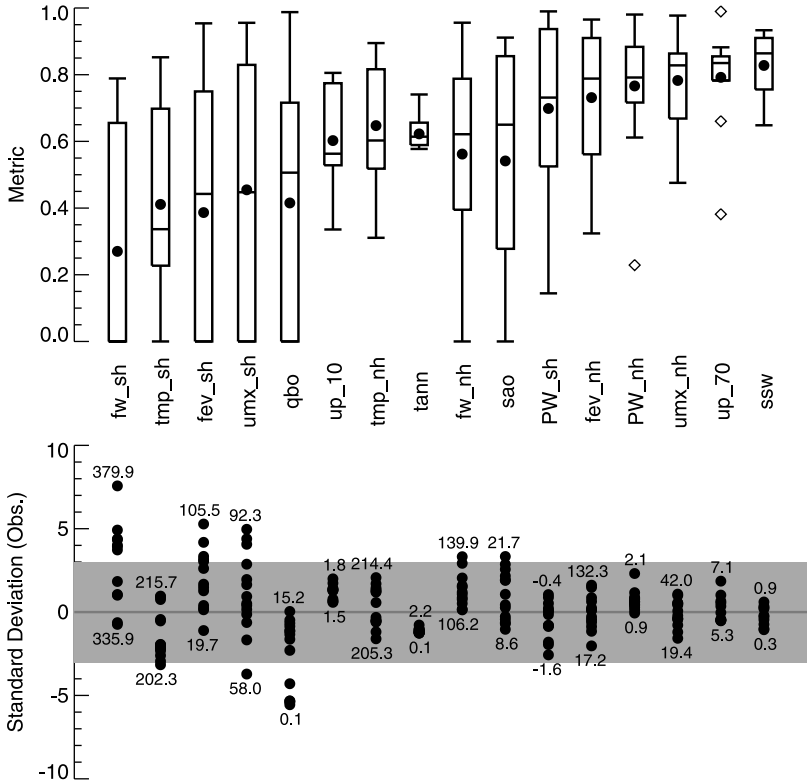

Figure 16. (top) Box and whisker plots of g-metrics. For details of metrics, see text and Table 2. Box shows the 25th to 75th percentile of the distribution of the validation metrics, the central horizontal line shows the median of the validation metrics and the black dot shows the arithmetic mean. Whiskers show the range of the data excluding outliers (plotted with open diamonds). Metrics are ordered by median. Also shown is (bottom) the distribution of j metrics (see text) and are plotted relative to the same diagnostic calculated from reanalysis data and scaled by its standard deviation. The ±3 region used to define models which would achieve a zero metric in the validation metric calculation is shaded in grey. The absolute value of the maximum and minimum for each diagnostic in the model ensemble is printed at the top and bottom of each group of points.

error in place of the standard deviation would have little effect on the comparison of performance revealed by the calculations presented here (it would simply tend to make all metric scores lower and these could then be renormalized be changing the value of n_g). In the case of the SSW metric, the relative skill of the models will be slightly overestimated in the current analysis (i.e., using the interannual standard deviation instead of the standard error). It has not been possible at this stage to incorporate estimates of observational uncertainty in the calculation of the metrics. Obtaining an estimate of the observational uncertainty is far from trivial since most observational estimates are derived from complex reanalysis products and hence any simple comparison between reanalysis datasets would incorporate both true observational uncertainty and that due to the details of the particular model/data assimilation system used. Consequently, for this study it was considered preferable not to incorporate this term in the analysis. For the tropical variability metrics, estimating the uncertainty in the ERA-40 reanalysis is more complex. To estimate the uncertainty, the data set was sampled for several 10-year periods, and the range of possible values of annual cycle, SAO, and QBO amplitudes was used in the metric calculation.

[50] While the metric g is a useful way of validating the performance of the model ensemble against reanalysis data, it does not provide any information about the sign of biases in the models. This information is an important component of the assessment of model performance, since in some diagnostics the model ensemble shows a systematic negative or positive bias indicative of a common deficiency in the models. Therefore an additional metric which retains the sign information removed in the calculation of g is also considered:

$$j = \frac{\text{model} \quad \text{obs}}{\text{obs}}$$

j is then simply the difference between model and observational estimates of each diagnostic, normalized by the standard deviation of the observational estimate. Note that in this metric, large absolute values indicate a mismatch between model and observations.

4.2. Results

[51] The distribution of g and j for the metrics in Table 2 is shown in Figure 16. Several broad conclusions about the performance of the models can be drawn from this figure.

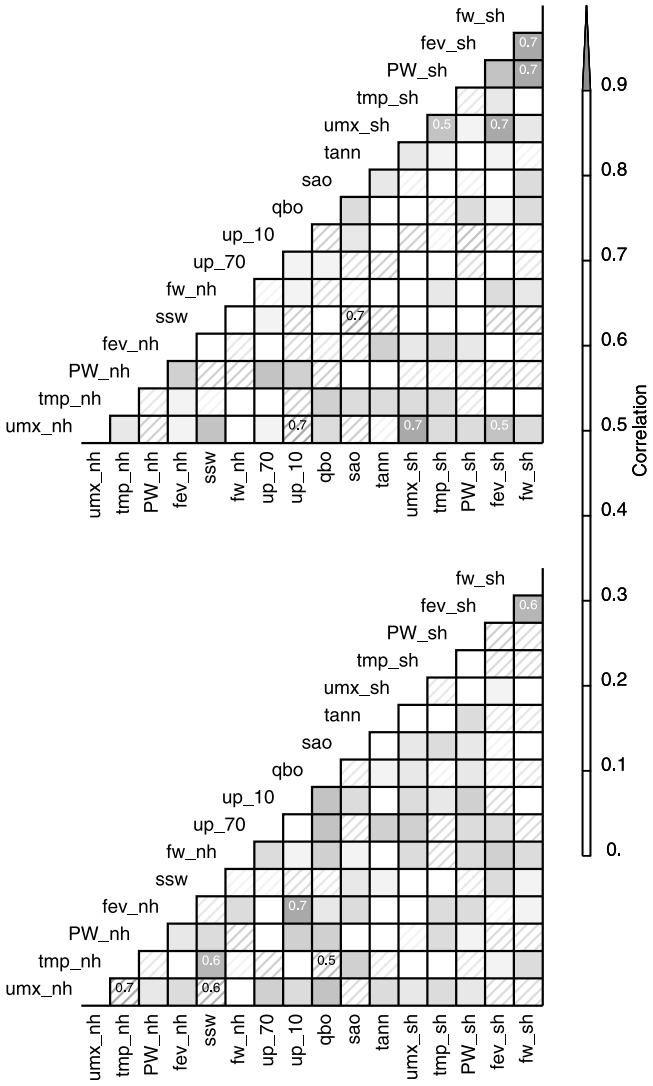

Figure 17. Pearson rank correlation matrix for quantitative metrics. Shown are (top) the correlation between *g* metrics and (bottom) the correlation between scaled model diagnostics (*j* metrics). Correlation between metrics is printed in each square where the correlation is significant at p=0.05. Solid shading and printing indicates positive correlation and hatched shading and printing indicates negative correlation.

[52] 1. For most diagnostics, there is a large spread in the performance of the models. This is particularly apparent for diagnostics in which the 25th percentile line overlaps zero in the box and whisker plots in Figure 16 (top). This indicates that a significant number of the models are graded with $g = 0$, or in other words have biases greater than three standard deviations when compared to observational estimates.

[53] 2. The SH diagnostics of both climate and variability generally show poorer performance than similar diagnostics for the NH. The four diagnostics with the smallest median value of *g* are those for SH final warming date, springtime polar cap temperature, variance of the first EOF, and the strength of the midwinter midstratosphere jet.

[54] 3. For some metrics in which model performance is generally poor, model biases tend to have the same sign indicating a systematic difference between the models and the observations. For the metrics considered here, systematic negative biases are found for the SH temperature, and the amplitudes of the tropical annual cycle and QBO (although some caution is necessary for the tropical diagnostics). Metrics with a systematic positive bias are those for upwelling at 10 hPa, the final warming dates in the NH and SH, the amplitude of the first EOF in the SH and the slope of the fit between lower stratospheric heat flux and lower stratospheric temperature in the NH. For other metrics, there are large numbers of models with significant biases, but these tend to be evenly distributed between positive and negative signs and hence while indicating poor performance for individual models, they do not indicate systematic biases amongst the multimodel ensemble.

[55] The relationship between diagnostics can be characterized further using the correlation between different metrics (Figure 17). Since the calculation of *g* uses a cutoff for differences greater than n_g the Spearman rank correlation is used in the analysis presented here rather than the standard Pearson correlation coefficient (sensitivity tests with the Pearson correlation showed broadly similar results). The correlation between diagnostics is calculated for both the *g* and *j* metrics, however, for the two metrics the correlation should be interpreted slightly differently.

[56] Large positive correlations between diagnostics in the *g* metric (Figure 17, top) indicate that models that perform well when compared to reanalysis in one diagnostic also tend to perform well in another diagnostic. Large negative correlations in the *g* metric indicate that models that perform well when compared to reanalysis in one diagnostic also tend to perform poorly in another diagnostic. In other words, cross correlation in the *g* metric indicates pairs of diagnostics where good performance is or is not related.

[57] Large positive correlations between diagnostics in the *j* metric (Figure 17, bottom) indicate that models tend to have a similar position in the model ensemble. The performance of the model relative to observations is not considered. Large negative correlations between diagnostics in the *j* metric indicate that models tend to have an opposing position in the model ensemble. In other words, cross correlation in the *j* metric indicates that the diagnostics are related, or linked to each other by a dynamical and/or physical process.

[58] Several interesting relationships between the diagnostics considered are revealed by this analysis. In the SH, where model performance is generally poor, there are positive correlations in the *g* metric between several diagnostics including the springtime temperature, the midwinter jet maximum, the final warming date and the amplitude of the first EOF. However, only weak correlations exist between the *j* metrics of the same variables. This suggests that an additional external factor may be responsible for the correlations between model validation in the SH. Also note that *j* metrics of the amplitude of the first EOF and the date of the final warming in the SH are positively correlated as noted previously by *Fogt et al.* [2009]; models with too much variability in the first EOF tend to have a delayed final warming.

[59] In the NH, coherent behavior for the *j* metrics of springtime temperature, SSW frequency and midwinter jet maximum is observed, a stronger, less variable midwinter

vortex tending to lead to a colder vortex in springtime. There is also a positive correlation in model performance (i.e., g metrics) for the strength of the midwinter jet and SSW frequency.

[60] As expected, there is little correlation between model performance between hemispheres, however there is a strong positive correlation between the strength of the midwinter jet in the NH and SH.

[61] In the tropics, understanding the relationship between metrics is complicated because of the different design of the models, with some imposing a QBO, some generating an internal QBO and others with no QBO at all. However, models with a stronger QBO tend to have stronger upwelling at both 70 hPa and 10 hPa and a stronger and colder NH vortex. Interestingly at higher tropical altitudes, there are negative correlations using the g metric between SSW frequency and the model simulation of the SAO and 1 hPa annual cycle. This suggests that improved model performance in the tropical upper stratosphere tends to degrade the simulation of major midwinter warmings.

[62] Analysis of dynamical performance using performance metrics provides a useful way of summarizing the performance of the ensemble of current models. It should be noted however that there are many caveats to the way in which metrics are calculated and the choice of diagnostics used for model assessment. The metrics and analysis presented here are simply one way of assessing the performance of models.

5. Concluding Remarks

[63] This study both updates and more importantly extends previous evaluations of multimodel simulations of the stratospheric climate. The main conclusion from the updating of the earlier assessments was that in terms of simulating the time-mean, zonally averaged stratospheric climate the models have, on average, not improved significantly since the last comprehensive assessment by *Eyring et al.* [2006]. Nonetheless, with the notable exception of some key phenomena, the extratropical temperatures and zonal mean climate were, in general, qualitatively well reproduced with little uncertainty or spread between the models.

[64] The extension of the assessment to include intraseasonal and interannual variability indicated that this aspect was, on average, less well simulated. On the other hand, the zonal asymmetries which determine the shape and position of the polar vortex were reproduced reasonably well.

[65] A major difference of the present assessment from the previous multimodel assessments of *Pawson et al.* [2000], *Austin et al.* [2003], and *Eyring et al.* [2006] is the use of quantitative metrics for evaluating the models. The choice of metrics used to "rate" models is subject to some implicit assumptions about errors in observed data and can lead to a lack of differentiation between good and bad models if not considered carefully [*Grewe and Sausen*, 2009]. Nonetheless, considering both the spread, sign, and correlation of metrics provides a useful tool for examining any link in model performance between the different dynamical processes considered and across the multimodel ensemble. Interestingly, the metrics suggest a wider spread in model performance than would be inferred from the qualitative analysis and quantitatively confirm that overall model performance is poorer in the SH than the NH. Moreover the use of metrics indicated little correlation in model performance between the two hemispheres apart from in the jet strength. This suggest that model development should perhaps be focused more on the hemispheric scale rather than the global parameters and setup, though it is also possible that the individual metrics in each hemisphere could be sensitive to the choice of global parameters.

[66] Although the metrics do provide a useful additional tool for identifying links in model performance between the different dynamical processes and identifying common model deficiencies, the metrics themselves provide little or no useful information on the underlying physical processes within the models. Therefore the metrics are of most scientific value when combined with a more conventional analysis of physical quantities, as was done in section 3. The combined use of the two approaches in this study indicates that there are long-standing and significant common biases in models which remain poorly understood. Particularly challenging are the biases associated with the springtime breakup of the polar vortex in both hemispheres and the generally poor performance of the models in the Southern Hemisphere. In the tropics, the majority of models are still unable to reproduce anything like a realistic quasi-biennial oscillation, though many partially circumvent the problem by artificially prescribing this variability even though it is unclear if this approach actually leads to an overall model improvement away from the tropics. Clearly, it restricts the applications for which these models can be used for.

[67] A key outstanding question of this study is how to address some of the persistent dynamical biases and problems which bedevil stratosphere-resolving climate and Earth system models. While massive coordinated multimodel assessments such as that of the *Eyring et al.* [2010] report have proved extremely valuable both for identifying common model strengths and weakness and also for singling out those which are most relevant they have not been quite so successful at addressing many of the long-standing and persistent model problems, at least from a dynamical perspective. A transfer of effort to a coordinated focus on specific process such as the SPARC DynVar [*Kushner et.*, 2007] initiative on stratospheric variability and stratosphere-troposphere coupling is a potentially useful way forward.

[68] **Acknowledgments.** We acknowledge the Chemistry-Climate Model Validation Activity (CCMVal) of the World Climate Research Programme's (WCRP) Stratospheric Processes and their Role in Climate (SPARC) project for organizing and coordinating the model data analysis, and the British Atmospheric Data Centre (BADC) for collecting and archiving the CCMVal model output. ECMWF ERA-Interim data used in this study have been provided by ECMWF. The contribution from Neal Butchart and Steven Hardiman was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101). The ERA-Interim results shown in Figure 10 were kindly provided by William Seviour of the University of Cambridge. CCSRNIES research was supported by the Global Environmental Research Fund of the Ministry of the Environment of Japan (A-071 and A-0903) and the simulations were completed with the super computer at CGER, NIES. The MRI simulation was made with the supercomputer at the National Institute for Environmental Studies, Japan. The contribution from the LATMOS-IPSL was supported by the European Commission through the funding of the RECONCILE and GEOMON projects.

References

Akiyoshi, H., L. B. Zhou, Y. Yamashita, K. Sakamoto, M. Yoshiki, T. Nagashima, M. Takahashi, J. Kurokawa, M. Takigawa, and T. Imamura (2009), A CCM simulation of the breakup of the Antarctic polar vortex in the years 1980–2004 under the CCMVal scenarios, *J. Geophys. Res.*, *114*, D03103, doi:10.1029/2007JD009261.

Andrews, D. G., J. R. Holton, and C. B. Leovy (1987), *Middle Atmosphere Dynamics*, 489 pp., Academic. San Diego, Calif.

Austin, J., and R. J. Wilson (2010), Sensitivity of polar ozone to sea surface temperatures and halogen amounts, *J. Geophys. Res.*, *115*, D18303, doi:10.1029/2009JD013292.

Austin, J., et al. (2003), Uncertainties and assessments of chemistry-climate models of the stratosphere, *Atmos. Chem. Phys.*, *3*, 1–27.

Austin, J., et al. (2010a), Chemistry climate model simulations of spring Antarctic ozone, *J. Geophys. Res.*, *115*, D00M11, doi:10.1029/2009JD013577.

Austin, J., et al. (2010b), The decline and recovery of total column ozone using a multi-model time series analysis, *J. Geophys. Res.*, *115*, D00M10, doi:10.1029/2010JD013857.

Baldwin, M. P., D. B. Stephenson, D. W. J. Thompson, T. J. Dunkerton, A. J. Charlton, and A. O'Neill (2003), Stratospheric memory and skill of extended-range weather forecasts, *Science*, *301*, 636–640.

Baldwin, M. P., M. Dameris, and T. G. Shepherd (2007), How will the stratosphere affect climate change?, *Science*, *316*, 1576–1577.

Black, R. X., and B. A. McDaniel (2007a), The dynamics of Northern Hemisphere stratospheric final warming events, *J. Atmos. Sci.*, *64*, 2932–2946.

Black, R. X., and B. A. McDaniel (2007b), Interannual variability in the Southern Hemisphere circulation organized by stratospheric final warming events, *J. Atmos. Sci.*, *64*, 2968–2974.

Black, R. X., and B. A. McDaniel (2009), Submonthly polar vortex variability and stratosphere-troposphere coupling in the Arctic, *J. Clim.*, *22*, 5886–5901.

Black, R. X., B. A. McDaniel, and W. A. Robinson (2006), Stratosphere-troposphere coupling during spring onset, *J. Clim.*, *19*, 4891–4901.

Butchart, N., and J. Austin (1998), Middle atmosphere climatologies from the troposphere-stratosphere configuration of the UKMO's unified model, *J. Atmos. Sci.*, *55*, 2782–2809.

Butchart, N., and A. A. Scaife (2001), Removal of chlorofluorocarbons by increased mass exchange between the stratosphere and troposphere in a changing climate, *Nature*, *410*, 799–802.

Butchart, N., et al. (2006), Simulations of anthropogenic change in the strength of the Brewer-Dobson circulation. *Climate Dyn.*, *27*, 727–741.

Butchart, N., et al. (2010a), Chemistry-climate model simulations of 21st century stratospheric climate and circulation changes, *J. Clim.*, *23*, 5349–5374.

Butchart, N., et al. (2010b), Stratospheric dynamics, in *Evaluation of Chemistry-Climate Models, Rep. 5*, edited by V. Eyring, T. G. Shepherd, and D. W. Waugh, *WCRP-132*, pp. 109–148, WCRP, Geneva, Switzerland.

Charlton, A. J., and L. M. Polvani (2007), A new look at stratospheric sudden warmings. Part I. Climatology and modelling benchmarks, *J. Clim.*, *20*, 449–469.

Charlton, A. J., L. M. Polvani, J. Perlwitz, F. Sassi, E. Manzini, K. Shibata, S. Pawson, J. E. Nielsen, and D. Rind (2007), A new look at stratospheric sudden warmings. Part II. Evaluation of numerical model simulations, *J. Clim.*, *20*, 470–488.

Dameris, M., et al. (2005), Long-term changes and variability in a transient simulation with a chemistry-climate model employing realistic forcing, *Atmos. Chem. Phys.*, *5*, 2121–2145.

de Grandpré, J., S. Beagley, V. Fomichev, E. Griffioen, J. McConnell, A. Medvedev, and T. G. Shepherd (2000), Ozone climatology using interactive chemistry: Results from the Canadian middle atmosphere model, *J. Geophys. Res.*, *105*, 26,475–26,491.

Déqué, M. (2007), Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values, *Global Planet. Change*, *57*, 16–26.

Egorova, T., E. Rozanov, V. Zubov, E. Manzini, W. Schmutz, and T. Peter (2005), Chemistry-climate model SOCOL: A validation of the present-day climatology, *Atmos. Chem. Phys.*, *5*, 1557–1576.

Eyring, V., et al. (2005), A strategy for process-oriented validation of coupled chemistry-climate models, *Bull. Am. Meteorol. Soc.*, *86*, 1117–1133.

Eyring, V., et al. (2006), Assessment of temperature, trace spices, and ozone in chemistry-climate model simulations of the recent past, *J. Geophys. Res.*, *111*, D22308, doi:10.1029/2006JD007327.

Eyring, V., M. P. Chipperfield, M. A. Giorgetta, D. E. Kinnison, E. Manzini, K. Matthes, P. A. Newman, S. Pawson, T. G. Shepherd, and D. W. Waugh (2008), Overview of the new CCMVal reference and sensitivity simulations in support of upcoming ozone and climate assessments and the planned SPARC CCMVal report, *SPARC Newsl. 30*, 20–26.

Eyring, V., T. G. Shepherd, and D. W. Waugh (Eds.) (2010), *Evaluation of Chemistry-Climate Models, Rep. 5*, *WCRP-132*, WCRP, Geneva, Switzerland.

Feser, F., H.-F. Graf, and J. Perlwitz (2000), Secular variability of the coupled tropospheric and stratospheric circulation in the GCM ECHAM 3/LSG, *Theor. Appl. Climatol.*, *65*, 1–15.

Fogt, R. L., J. Perlwitz, S. Pawson, and M. A. Olsen (2009), Intra-annual relationships between polar ozone and the SAM, *Geophys. Res. Lett.*, *36*, L04707, doi:10.1029/2008GL036627.

Garcia, R. R., D. Marsh, D. Kinnison, B. A. Boville, and F. Sassi (2007), Simulations of secular trends in the middle atmosphere 1950–2003, *J. Geophys. Res.*, *112*, D09301, doi:10.1029/2006JD007485.

Garny, H., M. Dameris, and A. Stenke (2009), Impact of prescribed SSTs on climatologies and long-term trends in CCM simulations, *Atmos. Chem. Phys.*, *9*, 6017–6031.

Gillett, N. P., and D. W. J. Thompson (2003), Simulation of recent southern hemisphere climate change, *Science*, *302*, 273–275.

Grewe, V., and R. Sausen (2009), Comment on "Quantitative performance metrics for stratosphere-resolving chemistry-climate models" by Waugh and Eyring (2008), *Atmos. Chem. Phys.*, *9*, 9101–9110.

Hardiman, S. C., D. G. Andrews, A. A. White, N. Butchart, and I. Edmond (2010a), Using different formulations of the transformed Eulerian-mean equations and Eliassen-Palm diagnostics in general circulation models, *J. Atmos. Sci.*, *67*, 1983–1995.

Hardiman, S. C., N. Butchart, S. M. Osprey, L. J. Gray, A. C. Bushell, and T. J. Hinton (2010b), The climatology of the middle atmosphere in a vertically extended version of the Met Office's climate model. Part I: Mean state, *J. Atmos. Sci.*, *67*, 1509–1525.

Haynes, P. H., C. J. Marks, M. E. McIntyre, T. G. Shepherd, and K. P. Shine (1991), On the "downward control" of the extratropical diabatic circulation by eddy-induced mean zonal forces, *J. Atmos. Sci.*, *48*, 651–678.

Jöckel, P., et al. (2006), The atmospheric chemistry general circulation model ECHAM5/MESSy1: consistent simulation of ozone from the surface to the mesosphere, *Atmos. Chem. Phys.*, *6*, 5067–5104.

Jourdain, L., S. Bekki, F. Lott, and F. Lefévre (2008), The coupled chemistry-climate model LMDz-REPROBUS: description and evaluation of a transient simulation of the period 1980–1999, *Ann. Geophys.*, *26*, 6, 1391–1413.

Kalnay, E., et al. (1996), The NCAR/NCEP 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*, 437–471.

Kushner, P. J., et al. (2007), The SPARC DynVar project: A SPARC project on the dynamics and variability of the coupled stratosphere-troposphere system, *SPARC Newsl.*, *29*, 9–14.

Lamarque, J.-F., D. E. Kinnison, P. G. Hess, and F. M. Vitt (2008), Simulated lower stratospheric trends between 1970 and 2005: Identifying the role of climate and composition changes, *J. Geophys. Res.*, *113*, D12301, doi:10.1029/2007JD009277.

Manney, G. L., J. L. Sabutis, S. Pawson, M. L. Santee, B. Naujokat, R. Swinbank, M. E. Gelman, and W. Ebisuzaki (2003), Lower stratospheric temperature differences between meteorological analyses in two cold Arctic winters and their impact on polar processing studies, *J. Geophys. Res.*, *108*(D5), 8328, doi:10.1029/2001JD001149.

Manney, G. L., K. Krüger, J. L. Sabutis, S. A. Sena, and S. Pawson (2005a), The remarkable 2003–2004 winter and other recent warm winters in the Arctic stratosphere since the late 1990s, *J. Geophys. Res.*, *110*, D04107, doi:10.1029/2004JD005367.

Manney, G. L., D. R. Allen, K. Krüger, J. L. Sabutis, S. Pawson, R. Swinbank, C. E. Randall, A. J. Simmons, and C. Long (2005b), Diagnostic comparison of meteorological analyses during the 2002 Antarctic winter, *Mon. Weather Rev.*, *133*, 1261–1278.

Morgenstern, O., P. Braesicke, M. M. Hurwitz, F. M. O'Connor, A. C. Bushell, C. E. Johnson, and J. A. Pyle (2008), The world avoided by the Montreal Protocol, *Geophys. Res. Lett.*, *35*, L16811, doi:10.1029/2008GL034590.

Morgenstern, O., P. Braesicke, F. M. O'Connor, A. C. Bushell, C. E. Johnson, S. M. Osprey, and J. A. Pyle (2009), Evaluation of the new UKCA climate-composition model — Part 1: The stratosphere, *Geosci. Model Dev.*, *2*, 43–57.

Morgenstern, O., et al. (2010), Review of present-generation stratospheric chemistry and associated external forcings. *J. Geophys. Res.*, *115*, D00M02, doi:10.1029/2009JD013728.

Nakicenovic, N., and R. Swart (Eds.) (2000), *Special Report on Emissions Scenarios: A Special Report of Working Group III of the Intergovernmental Panel on Climate Change*, 570 pp., Cambridge Univ. Press, New York.

Neu, J. L., and R. A. Plumb (1999), Age of air in a "leaky pipe" model of stratospheric transport, *J. Geophys. Res.*, *104*, 19,243–19,255.

Neu, J., S. Strahan, P. Braesicke, A. Douglas, P. Huck, L. Oman, D. Pendlebury, and S. Tegtmeier (2010), Transport, in *Evaluation of Chemistry-Climate Models, Rep. 5*, edited by V. Eyring, T. G. Shepherd, and D. W. Waugh, *WCRP-132*, pp. 149–190, WCRP, Geneva, Switzerland.

Newman, P. A., E. R. Nash, and J. E. Rosenfield (2001), What controls the temperature of the Arctic stratosphere during spring?, *J. Geophys. Res.*, *106*, 19,999–20,010.

North, G. R., T. L. Bell, R. F. Cahalan, and F. J. Moeng (1982), Sampling errors in the estimation of empirical orthogonal functions, *Mon. Weather Rev.*, *110*, 699–706.

Osprey, S. M., L. J. Gray, A. C. Bushell, N. Butchart, S. C. Hardiman, and T. J. Hinton (2010), The climatology of the middle atmosphere in a vertically extended version of the Met Office's climate model. Part II: Variability, *J. Atmos. Sci.*, *67*, 3637–3651.

Pascoe, C., L. Gray, S. Crooks, M. Juckes, and M. Baldwin (2005) The quasi-biennial oscillation: Analysis using ERA40 data, *J. Geophys. Res.*, *110*, D08105, doi:10.1029/2004JD004941.

Pawson, S., K. Krüger, R. Swinbank, M. Bailey, and A. O'Neill (1999), Intercomparison of two stratospheric analyses: temperatures relevant polar stratospheric cloud formation, *J. Geophys. Res.*, *104*, 2041–2050.

Pawson, S, et al. (2000), The GCM-reality intercomparison project for SPARC (GRIPS): Scientific issues and initial results, *Bull. Am. Meteorol. Soc.*, *81*, 781–796.

Pawson, S., R. S. Stolarski, A. R. Douglass, P. A. Newman, J. E. Nielsen, S. M. Frith, and M. L. Gupta (2008), Goddard Earth Observing System chemistry-climate model simulations of stratospheric ozone-temperature coupling between 1950 and 2005, *J. Geophys. Res.*, *113*, D12103, doi:10.1029/2007JD009511.

Randel, W., et al. (2004), The SPARC intercomparison of middle-atmosphere climatologies, *J. Clim.*, *17*, 986–1003.

Schraner, M., et al. (2008), Technical Note: Chemistry-climate model SOCOL: Version 2.0 with improved transport and chemistry/microphysics schemes, *Atmos. Chem. Phys.*, *8*, 5957–5974.

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

Shaw, T. A., and T. G. Shepherd (2008), Raising the roof, *Nat. Geosci.*, *1*, 12–13.

Shibata, K., and M. Deushi (2008a), Long-term variations and trends in the simulation of the middle atmosphere 1980–2004 by the chemistry-climate model of the Meteorological Research Institute, *Ann. Geophys.*, *26*, 1299–1326.

Shibata, K., and M. Deushi (2008b), Simulation of the stratospheric circulation and ozone during the recent past (1980–2004) with the MRI chemistry-climate model, *CGER's Supercomputer Monogr. Rep. 13*, 154 pp., Natl. Inst. for Environ. Stud., Tsukuba, Japan.

Stenke, A., M. Dameris, V. Grewe, and H. Garny (2009), Implications of Lagrangian transport for simulations with a coupled chemistry-climate model, *Atmos. Chem. Phys.*,*9*, 5489–5504.

Swinbank, R., and A. O' Neill (1994), A stratosphere-troposphere data assimilation system, *Mon. Weather Rev.*, *122*, 686–702.

Teyssèdre, H., et al. (2007), A new tropospheric and stratospheric chemistry and transport model MOCAGE-Climat for multi-year studies: evaluation of the present-day climatology and sensitivity to surface processes, *Atmos. Chem. Phys.*, *7*, 5815–5860.

Thompson, D. W. J., and J. M. Wallace (2000), Annular modes in the extratropical circulation. Part I: Month-to-month variability, *J. Clim.*, *13*, 1000–1016.

Tian, W., and M. P. Chipperfield (2005), A new coupled chemistry-climate model for the stratosphere: The importance of coupling for future O_3-climate predictions, *Q. J. R. Meteorol. Soc.*, *131*, 281–304.

Tian, W., and M. P. Chipperfield (2006), Stratospheric water vapor trends in a coupled chemistry-climate model, *Geophys. Res. Lett.*, *33*, L06819, doi:10.1029/2005GL024675.

Uppala, S. M., et al. (2005), The ERA-40 re-analysis, *Q. J. R. Meteorol. Soc.*, *131*, 2961–3012, doi:10.1256/qj.04.176.

Waugh, D. W., and V. Eyring (2008), Performance metrics for stratosphere-resolving chemistry-climate models, *Atmos. Chem. Phys.*, *8*, 5699–5713.

World Meteorological Organization (2007), *Scientific Assessment of Ozone Depletion: 2006, Global Ozone Research and Monitoring Project, Rep. 50*, 572 pp., Geneva.

H. Akiyoshi, T. Nakamura, and Y. Yamashita, National Institute for Environmental Studies, Tsukuba, 305-8506 Japan.

J. Austin, Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, NJ 08540, USA.

A. Baumgaertner, C. Brühl, and P. Jöckel, Max Planck Institut fũr Chemie, D-55128 Mainz, Germany.

S. Bekki and M. Marchand, LATMOS, Institut Pierre Simon Laplace, Universite Pierre et Marie Curie, F-75252 Paris CEDEX 05, France.

P. Braesicke, P. H. Haynes, and J. Pyle, National Centre for Atmospheric Science, University of Cambridge, Cambridge CB2 1SZ, UK.

N. Butchart and S. C. Hardiman, Met Office Hadley Centre, FitzRoy Road, Exeter, Devon, EX1 3PB, UK. (neal.butchart@metoffice.gov.uk)

A. J. Charlton-Perez, Department of Meteorology, University of Reading, Earley Gate, PO Box 243, Reading, RG6 6BB, UK. (a.j.charlton@reading.ac.uk)

M. Chipperfield, S. Dhomse, and W. Tian, School of Earth and Environment, University of Leeds, Leeds, LS2 9JT, UK.

I. Cionni, M. Dameris, V. Eyring, and H. Garny, Deutsches Zentrum fũr Luft und Raumfahrt, Institut für Physik der Atmosphäre, D-82234 Oberpfaffenhofen, Germany.

R. Garcia and J. F. Lamarque, National Center for Atmospheric Research, Boulder, CO 80307, USA.

K. Krüger, Leibniz Institute of Marine Sciences, IFM-GEOMAR, D-24148 Kiel, Germany.

P. J. Kushner, M. Sigmond, T. G. Shepherd, and L. Wang, Department of Physics, University of Toronto, Toronto, ON M5S 1A1, Canada.

M. Michou and H. Teyssèdre, GAME/CNRM, (Météo-France, CNRS), F-31057 Toulouse CEDEX 01, France.

O. Morgenstern and D. Smale, National Institute of Water and Atmospheric Research, Private Bag 50061, Lauder, New Zealand.

P. A. Newman and S. Pawson, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA.

S. M. Osprey, National Centre for Atmospheric Science, Department of Physics, University of Oxford, Oxford OX1 2JD, UK.

J. Perlwitz, Physical Sciences Division, Cooperative Institute for Research in Environmental Sciences, University of Colorado and NOAA Earth System Research Laboratory, Boulder, CO 80309, USA.

D. Plummer, Environment Canada, Toronto, ON M3H 5T4, Canada.

E. Rozanov, Physical-Meteorological Observatory/World Radiation Centre, CH-7260 Davos, Switzerland.

J. Scinocca, Meteorological Service of Canada, University of Victoria, Victoria, BC V8W 2Y2, Canada.

K. Shibata, Meteorological Research Institute, Tsukuba, 305-0052, Japan.

D. Waugh, Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, MD 21218, USA.