On the lack of stratospheric dynamical variability in low-top versions of the CMIP5 models

Andrew J. Charlton-Perez,[1] Mark P. Baldwin,[2] Thomas Birner,[3] Robert X. Black,[4] Amy H. Butler,[5] Natalia Calvo,[6] Nicholas A. Davis,[3] Edwin P. Gerber,[7] Nathan Gillett,[8] Steven Hardiman,[9] Junsu Kim,[10] Kirstin Krüger,[16] Yun-Young Lee,[4] Elisa Manzini,[11] Brent A. McDaniel,[12] Lorenzo Polvani,[13] Thomas Reichler,[10] Tiffany A. Shaw,[13] Michael Sigmond,[14] Seok-Woo Son,[15] Matthew Toohey,[16] Laura Wilcox,[17] Shigeo Yoden,[18] Bo Christiansen,[19] François Lott,[20] Drew Shindell,[21] Seiji Yukimoto,[22] and Shingo Watanabe[23]

[1] We describe the main differences in simulations of stratospheric climate and variability by models within the fifth Coupled Model Intercomparison Project (CMIP5) that have a model top above the stratopause and relatively fine stratospheric vertical resolution (high-top), and those that have a model top below the stratopause (low-top). Although the simulation of mean stratospheric climate by the two model ensembles is similar, the low-top model ensemble has very weak stratospheric variability on daily and interannual time scales. The frequency of major sudden stratospheric warming events is strongly underestimated by the low-top models with less than half the frequency of events observed in the reanalysis data and high-top models. The lack of stratospheric variability in the low-top models affects their stratosphere-troposphere coupling, resulting in short-lived anomalies in the Northern Annular Mode, which do not produce long-lasting tropospheric impacts, as seen in observations. The lack of stratospheric variability, however, does not appear to have any impact on the ability of the low-top models to reproduce past stratospheric temperature trends. We find little improvement in the simulation of decadal variability for the high-top models compared to the low-top, which is likely related to the fact that neither ensemble produces a realistic dynamical response to volcanic eruptions.

All supporting information may be found in the online version of this article.

[1]Department of Meteorology, University of Reading, Reading, UK.
[2]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK.
[3]Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado, USA.
[4]School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, Georgia, USA.
[5]Climate Prediction Center, College Park, Maryland, USA.
[6]Departamento de Fisica de la Tierra II, Universidad Compultense de Madrid, Madrid, Spain.
[7]New York University, New York, New York, USA.
[8]Canadian Centre for Climate Modelling and Analysis, Victoria, British Columbia, Canada.
[9]Met Office, Exeter, UK.
[10]Department of Atmospheric Sciences, University of Utah, Utah, USA.
[11]Max Planck Institut für Meteorologie, Hamburg, Germany.
[12]Department of Biology and Physics, Kennesaw State University, Kennesaw, Georgia, USA.
[13]Department of Applied Mathematics and Applied Physics, Columbia University, New York, New York, USA.
[14]Department of Physics, University of Toronto, Toronto, Canada.
[15]School of Earth and Environmental Sciences, Seoul National University, South Korea.
[16]GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany.
[17]National Centre for Atmospheric Science, University of Reading, Reading, UK.
[18]Kyoto University, Kyoto, Japan.
[19]Danish Meteorological Institute, Copenhagen, Denmark.
[20]Laboratoir de Meteorologie Dynamique, Ecole Normale Superieure, Paris, France.
[21]NASA Goddard Institute for Space Studies, New York, New York, USA.
[22]Meteorological Research Institute, Tsukuba, Japan.
[23]Research Institute for Global Change, JAMSTEC, Japan.

Corresponding author: A. J. Charlton-Perez, Department of Meteorology, University of Reading, Reading, UK. (a.j.charlton@reading.ac.uk)

1. Introduction

[2] One major change in coupled-climate modeling between the third (CMIP3) and fifth (CMIP5) Coupled Model Intercomparison Projects is an increase in the number of models with model tops above the stratopause and general progress toward a more realistic representation of the stratosphere in coupled climate models. As an example of this trend, only 5 of the 23 CMIP3 models considered in *Cordero and Forster* [2006] had tops above 1 hPa. In the CMIP5 archive, this ratio has increased to 15 models of 45. Furthermore, very few models now place their model lid in the middle stratosphere near 10 hPa, thereby reducing the number of models that are likely to severely distort stratospheric dynamics. In the present study, we seek to understand the benefits of a model lid above 1 hPa to the simulation of stratospheric climate and variability by comparing the simulation of stratospheric climate by a subset of models submitted to the CMIP5 archive.

[3] The move to increased stratospheric representation in coupled climate models has been motivated in part by the large body of recent work providing evidence that both internal stratospheric climate variability and external stratospheric climate forcing can be important drivers of tropospheric climate (as discussed by *Gerber et al.* [2012]). The climate models that make up the CMIP5 ensemble might be crudely divided into two subensembles; one representing models that attempt to fully represent stratospheric processes (for which we use the shorthand, "high-top") and the other representing models that do not (for which we use the shorthand, "low-top").

[4] In this study, we attempt the first broad-scale assessment of the performance of the high-top and low-top ensembles in CMIP5 in simulating stratospheric climate and variability. Many previous assessments of the simulation of stratospheric climate by stratosphere-resolving models [*Pawson et al.*, 2000; *Cordero and Forster*, 2006] and chemistry-climate models [*Austin et al.*, 2003; *Eyring et al.*, 2006; *Butchart et al.*, 2011] have shown an improving simulation of stratospheric mean climate over time. However, several persistent biases remain in most models. One example is the cold bias in spring temperatures in the polar lower stratosphere, associated with a delay in the stratospheric final warming. This bias is present in both hemispheres and in high-top and low-top models. The key questions in this study are if biases in the lower and middle stratosphere are reduced in high-top models compared to low-top models and if low-top models exhibit additional stratospheric biases.

[5] Ultimately, for most climate modeling centers the value of enhancing the stratospheric representation in their models will be measured in terms of any improvement to tropospheric biases and variability and the simulation of tropospheric climate change. Several studies have shown that the pattern and magnitude of regional tropospheric climate change can be significantly affected by changes in the stratospheric climate [e.g., *Sigmond and Scinocca*, 2010; *Scaife et al.*, 2011; *Karpechko and Manzini*, 2012]. In a companion paper, *Manzini et al.* [2012] examine how the multimodel simulation of tropospheric climate change in CMIP5 is sensitive to stratospheric climate change. A key prerequisite to this analysis is the diagnosis of the simulation of stratospheric climate and variability presented in this paper.

2. High-Top and Low-Top Models in the CMIP5 Ensemble

[6] The subset of models from the CMIP5 experiment considered in this study are listed in Table 1. In considering this large ensemble, it is clear that the CMIP5 models have a wide variety of lid heights, vertical resolutions, and parameterized physical processes in the stratosphere. The models were classified into a high-top and low-top ensemble based primarily upon their lid height, with a threshold between high-top and low-top at 1 hPa. This choice is motivated by previous studies [*Cagnazzo and Manzini*, 2009; *Maycock et al.*, 2011; *Shaw and Perlwitz*, 2010], which have suggested that models with a top below the stratopause fail to properly simulate episodic stratospheric variability such as stratospheric sudden warmings. We might therefore expect a similar difference between models in the CMIP5 ensemble segregated by this threshold.

[7] Because this study is primarily concerned with assessing how the models reproduce past observed climatology, our focus is solely on the historical runs of the CMIP5 models, which have observed climate forcings (including forcing from greenhouse gasses, ozone depletion, land-use change, tropospheric and stratospheric aerosol, and solar variability). More details of the CMIP5 experimental design can be found in *Taylor et al.* [2012]. Typically, model simulations are compared with the MERRA [*Rienecker et al.*, 2011] and ERA-Interim [*Dee et al.*, 2011] reanalysis data sets over the modern satellite era (1979 to present) when confidence in stratospheric reanalysis is highest, but for some diagnostics that require longer climate records, the ERA-40 reanalysis data set is used as an additional or alternative point of reference. The methodology used for each diagnostic and the reanalysis data set used is described in each subsection.

[8] To put our results in context with other generations and types of GCMs, we compare some of the diagnostics with those from the CMIP3 [*Meehl et al.*, 2007] and CCMVal-2 [*SPARC CCMVal*, 2010] ensembles. The CMIP3 models, which were developed in the early 2000s, generally have a lower vertical resolution than the CMIP5 models. The CCMVal-2 models are coupled chemistry-climate models, which are stratosphere resolving and have a vertical resolution comparable to the CMIP5 high-top ensemble, but are mostly run without coupling to the ocean. There is also a degree of commonality between the CCMVal-2 and CMIP5 models, because many share a similar or identical dynamical core.

[9] In section 3, we assess stratospheric climate by first calculating broad-scale metrics of the overall skill of the two model ensembles in simulating stratospheric climate and variability, following the work of *Reichler and Kim* [2008]. This approach allows us to directly and compactly assess the overall skill of the two ensembles and to compare

Table 1. Models Used in the Study and Their Stratospheric Properties. An N in the Stratospheric Physics Column Indicates Nonorographic Gravity Wave Drag is Included, a C Indicates Stratospheric Heterogeneous Chemistry is Included, a Line Indicates Neither is Included

Model	Lid Height	Levels	Above 200 hPa	Physics	Ens.
BCC-CSM1.1	2.917 hPa	26	13	–	Low
CCSM4	2.194067 hPa	27	13	–	Low
CNRM-CM5	10 hPa	31	9	–	Low
CSIRO-mk3.6.0	4.5 hPa	18	5	–	Low
GFDL CM3	0.01 hPa	48	28	NC	High
GFDL-ESM2G	3 hPa	24	5	–	Low
GFDL-ESM2M	3 hPa	24	5	–	Low
GISS-E2-R	0.1 hPa	40	19	NC	High
GISS-E2-H	0.1 hPa	40	19	NC	High
HadCM3	10 hPa	19	3	–	Low
HadGEM2-ES	40 km	38	15		Low
HadGEM2-CC	85 km	60	37	N	High
INMCM4	10 hPa (0.01 sigma)	21	8	N	Low
IPSL-CM5A-LR	0.04 hPa	39	22	N	High
IPSL-CM5A-MR	0.04 hPa	39	22	N	High
MIROC-ESM	0.0036 hPa	80	63	N	High
MIROC-ESM-CHEM	0.0036 hPa	80	63	NC	High
MIROC5	3 hPa	40	17	–	Low
MPI-ESM-LR	0.01 hPa	47	25	N	High
MRI-CGCM3	0.01 hPa	48	25	NC	High
NorESM1-M	3.54 hPa	26	13	–	Low

them with the CMIP3 and CCMVal-2 ensembles. However, the broad-scale metric does not help us to understand the reasons for differences in the performance of the two CMIP5 ensembles. Hence, in section 4 we use a small selection of process-based diagnostics to illustrate why the model ensembles produce a similarly skillful simulation of mean climate but a very different simulation of stratospheric variability. In section 5, we then assess the impact of the lack of stratospheric variability in the low-top ensemble on the simulation of stratosphere-troposphere coupling and the reproduction of past stratospheric trends by the models.

[10] It is not possible to construct a consistent high-top and low-top ensemble for all of the diagnostics of stratospheric climate presented in this study and the supporting information. The coverage of different diagnostics in the CMIP5 archive is quite variable and hence we have chosen to construct the largest possible ensemble for each of the diagnostics available. The models and ensemble members used for each diagnostic are shown in Table 2. Although this approach is far from ideal, it does allow us to make a broad-scale assessment of the simulation of stratospheric climate by the high-top and low-top ensembles within the CMIP5 archive. In almost all cases, each ensemble is made up of a large number of different GCMs and the removal of individual models from the ensemble does not change the qualitative structure of the high-top or low-top ensemble mean or their difference where this is significant.

3. A Broad Metric of Stratospheric Climate

[11] A simple way to assess the performance of the CMIP5 models in the stratosphere and compare them to previous generations of models is to compute broad-scale climate metrics. In this section, model performance is evaluated in terms of zonal averages of temperature (T), zonal wind (U), and specific humidity (q). The analysis domain for the model metrics extends from 90°S to 90°N and from 100 to 10 hPa. Some models in the low-top ensemble may have artificial numerical momentum damping near the

model top extending into this domain, but this information is not provided in the CMIP5 meta-data. This information is important to future evaluation of model performance in the stratosphere and we recommend that future multimodel intercomparison experiments include this type of meta-data.

[12] Four different aspects of climate are examined: long-term mean (MEAN), and variability on synoptic (DAILY), interannual (INTA), and decadal (DCDL) time scales. All four aspects are calculated individually for the four seasons. The calculation of mean climate as well as interannual and decadal variability is based on monthly mean input data. Synoptic variability is based on daily data after removing a low-frequency component with a temporal smoother using Gaussian weighting with a full-width at half maximum of 15 days [Baldwin et al., 2003].

[13] Variability is defined as the standard deviation over the given period of years. For example, interannual variability is the standard deviation of seasonal means, and decadal variability is the standard deviation of band-pass filtered monthly anomalies, using the fast Fourier transform technique and only retaining periods between 5 and 15 years. Decadal variability is calculated for the period 1961–2000, using the ERA-40 reanalysis as validation data. Mean climate, interannual variability, and daily variability are all based on 1979–2000 data and validated against ERA-Interim reanalysis. The model data are taken from the 20C3M (CMIP3), REF-B1 (CCMVal-2), and HISTORICAL (CMIP5) experiments; only one ensemble member is included from each model to avoid biases in the calculation of the metric toward models with larger ensembles.

[14] We examine two different measures of error: the pattern correlations (r) and the normalized root mean square error (E) [Reichler and Kim, 2008]. The procedure to compute E follows the method employed in Chapter 10 of SPARC CCMVal [2010]. In short, we first square the grid-point error between simulated and observed climate, normalize on a grid-point basis with the observed interannual variance, average spatially over a certain domain, and then take the square root. The grid-point error in simulating

Table 2. Membership of High and Low-Top Ensembles for Each Section[a]

Model	Metric	Zon. Temp.	Trends	SSWs	AM	Volc.
High-top models						
GFDL CM3	1	5	-	1	1	5
GISS-E2-R	1	5	5	5	-	5
GISS-E2-H	1	-	-	15	-	5
HadGEM2-CC	1	3	1	1	1	3
IPSL-CM5A-LR	1	4	-	4	4	-
IPSL-CM5A-MR	1	1	-	1	1	-
MIROC-ESM	1	3	-	1	3	3
MIROC-ESM-CHEM	1	1	1	1	1	1
MPI-ESM-LR	1	3	3	3	3	3
MRI-CGCM3	1	3	3	3	1	3
Total models (EM)	10(10)	9(28)	8(30)	7(14)	8(15)	8(28)
Low-top models						
BCC-CSM1.1	-	3	3	-	1	-
CCSM4	1	1	5	1	1	-
CNRM-CM5	1	1	-	1	1	-
CSIRO-mk3.6.0	1	10	10	5	-	10
GFDL-ESM2M	1	1	1	1	1	1
GFDL-ESM2G	-	-	-	1	-	-
HadCM3	1	-	10	10	-	10
HadGEM2-ES	1	4	4	4	4	3
INMCM4	1	1	-	1	-	-
MIROC5	1	4	4	4	4	3
NorESM1-M	1	3	3	3	3	3
Total models (EM)	9(9)	9(28)	6(22)	9(27)	6(11)	7(40)

[a]Because all models do not provide all diagnostics to the archive, the high-top and low-top ensembles have varying composition in each section as shown here. Numbers indicate the number of ensemble members used in each case. Abbreviations used for each section are metric, a broad metric of stratospheric climate (Figure 1); Zon. Temp., Zonal Mean Temperature Bias (Figure 2); Trends, Trends in Lower Stratospheric Temperature (Figure 6); SSWs, Sudden Stratospheric Warmings (Figure 3); AM, Annular Modes (Figure 5); Volc., Response to Volcanic Eruptions (Figure 4).

variability is based on the log2 variability ratio between model and observation. Because the computed values of E are nondimensionalized, errors from different climate quantities can be combined into a single measure of overall model performance. The values of r and E are calculated separately for each quantity, season, and model. We then take appropriate averages, e.g., for the four seasons and the three quantities (T, U, and q).

[15] Figure 1 summarizes the outcome of the validation exercise. Shown are the mean values of r and E for T, U, and q and for the four seasons. The oval shapes show the two standard deviation uncertainty intervals for the mean performance of the different model groups and aspects of climate, obtained by boot-strapping results from individual models within each ensemble. The clustering and location of the same colored ovals indicate that the simulation of synoptic variability (green) is generally associated with the highest skill scores. On the other hand, model performance is lowest for the simulation of decadal variability (light blue; note that this includes both forced and unforced variability), which may be in part related to the relatively short 40 year long data record and the uncertainty in observing this aspect of climate (particularly in the Southern Hemisphere (SH)).

[16] It is most notable that the high-top CMIP5 ensemble (thick solid line) simulates all three time scales of climate variability considerably better than the low-top CMIP5 counterpart (thin solid line). The uncertainty ovals for low-top and high-top CMIP5 models are well separated from each other. In other words, models with an increased vertical resolution and a higher model top presumably resolve stratospheric processes better, which leads to improved simulations of stratospheric climate variability. The mean

climate (orange) of the high-top ensemble has slightly better correlation with the reanalysis than the low-top ensemble (i.e., it reproduces better the horizontal and vertical structure of the climate), but the root mean square error is comparable (i.e., the size of climate biases is broadly similar).

[17] Of the four model ensembles considered, the CMIP3 ensemble (thin dashed line) has the worst performance in simulating mean climate and interannual variability. When considering daily and decadal variability the CMIP3 and CMIP5 low-top ensembles have comparable performance, which is significantly worse than either the CMIP5 high-top ensemble or the CCMVal-2 ensemble.

[18] Comparing the two high-top ensembles (CMIP5-high and CCMVal-2 (thick dashed line)) one finds that CCMVal-2 simulates decadal and interannual climate variability significantly better, but that the simulation quality for daily variability and mean climate are essentially the same. However, when interpreting these results it is important to know that most CCMVal-2 models are run with observed SST forcing and that many CCMVal-2 models include the quasi-biennial oscillation (QBO), an important phenomenon of interannual variability in the tropical stratosphere. In most cases, however, the QBO simulation is due to "nudging" to observations and thus does not represent a true simulation. On the other hand, CMIP5 models do not use such nudging. Nudging the QBO mostly improves the simulation of the CCMVal-2 models over the tropics, whereas over the extratropics the CCMVal-2 and the CMIP5 models perform very similarly (not shown). It is important to note, of course, that most of the models in the CMIP5 ensemble are run with a prescribed stratospheric ozone field, potentially reducing intermodel spread in
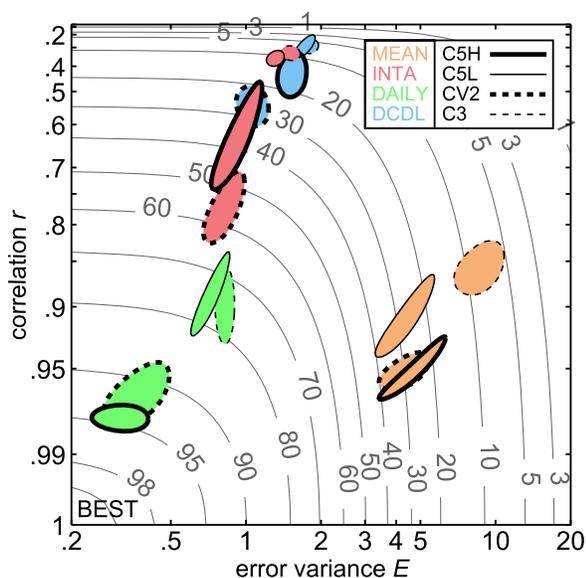
Figure 1. Simulation performance (90°S–90°N, 100–10 hPa) for different model ensembles and aspects of climate. Best performing ensembles are located at the lower left. Gray contours show the skill score S (in %), which combines E and r into a single index [*SPARC CCMVal*, 2010]. Oval shapes indicate ± 2 standard deviation uncertainty intervals, derived by bootstrapping results from individual models within a specific ensemble (estimates of the ensemble mean uncertainty were derived by resampling the existing estimates with replacement). C5H is the CMIP5 high-top model ensemble, C5L is the CMIP5 low-top model ensemble, CV2 is the CCMVal-2 model ensemble and C3 is the CMIP3 model ensemble. MEAN is the skill of the mean climate simulation, INTA is the skill of the internannual variability, DAILY is the skill of the daily variability and DCDL is the skill of the decadal variability.

comparison to the CCMVal-2 models, which have their own internally generated ozone fields.

[19] This analysis clearly shows that there are significant differences in the simulation of climate variability in the lower stratosphere between the high-top and low-top ensembles. The aim of the remainder of the study is to analyze the stratosphere in the two model ensembles in more detail to discover the origin of these differences.

4. Possible Causes of the Differing Performance of the Two Ensembles

4.1. Mean Climate

[20] The similarity of the mean temperature biases in the two model ensembles can be shown by simply calculating the difference between the multimodel mean, zonal-mean, annual-mean temperature as a function of latitude and pressure, and the same quantity in the ERA-Interim reanalysis data set (Figure 2). In a separate procedure to that used for calculation of the model metrics above, each models' climatology is determined for individual realizations, then averaged for all available ensemble members. The resulting temperature field is interpolated onto T42 latitudes and

standard pressure levels, and averaged over all available models within the high-top and low-top ensembles. For both model ensembles there is model bias in the region of the tropopause, with warm biases near the tropical tropopause (around 100 hPa) and cold biases near the extratropical tropopause (near 250 hPa). These differences are consistent with a low bias in tropical tropopause heights and a high-bias in extra-tropical tropopause heights, with somewhat stronger biases in the low-top models (see Figures S1 and S5 in the supporting information for details). One difference between the two model ensembles is in the high-latitude middle stratosphere (10 to 30 hPa). In the low-top model ensemble there are large cold biases. This is consistent with previous generations of low-top models [e.g., *Cordero and Forster*, 2006]. In the Northern Hemisphere (NH) this is generally thought to be associated with a lack of episodic stratospheric dynamical variability driven by wave-mean flow interactions. It is explicitly shown in the next section that there is clear distinction between the high-top and low-top ensembles in CMIP5 in this regard. In the SH, where such episodic dynamical variability is weaker, differences in the indirect heating by the residual circulation driven by the gravity wave drag parameterization may play a role in the large cold bias in the low-top ensemble. We cannot explicitly address this issue because tendencies from the gravity wave parameterizations of the CMIP5 models are not available.

[21] More detailed analysis of the mean temperature and zonal wind biases shown in the supporting information (Figures S1 and S2) reveals that the model ensembles diverge most in the extratropical middle stratosphere in the seasons in which the stratosphere is dynamically active (DJF in the NH and JJA and SON in the SH). This suggests that differences in mean climate are largely associated with differences in dynamical variability between the two ensembles.

[22] There are also other potentially important differences between the mean climate of the high-top and low-top ensembles, which could provide interesting examples for further study. One point of difference is in the strength of the tropical upwelling. Although both ensembles reasonably capture the Brewer-Dobson circulation, the low-top models produce anomalously strong upwelling at the equator but with a narrower tropical pipe (Figure S3a). A second difference concerns the well-known and persistent late bias in the final warming of the polar vortex in both hemispheres. Although both the high-top and low-top ensembles exhibit this bias, the high-top ensemble does show some improvement compared to the low-top ensemble (Figure S4). Finally, all models underestimate the pole-to-equator contrast in tropopause height (Figure S5), but this bias does appear to be somewhat alleviated in the high-top models.

[23] In summary, aside from the region near the model top, the high-top and low-top ensembles have very similar mean climate biases.

4.2. Daily and Interannual Variability

[24] As shown in Figure 1 there are large differences in the skill of the simulation of stratospheric daily variability between the high-top and low-top ensembles. This section shows that these differences are the result of a significant lack of episodic stratospheric variability in the low-top models.
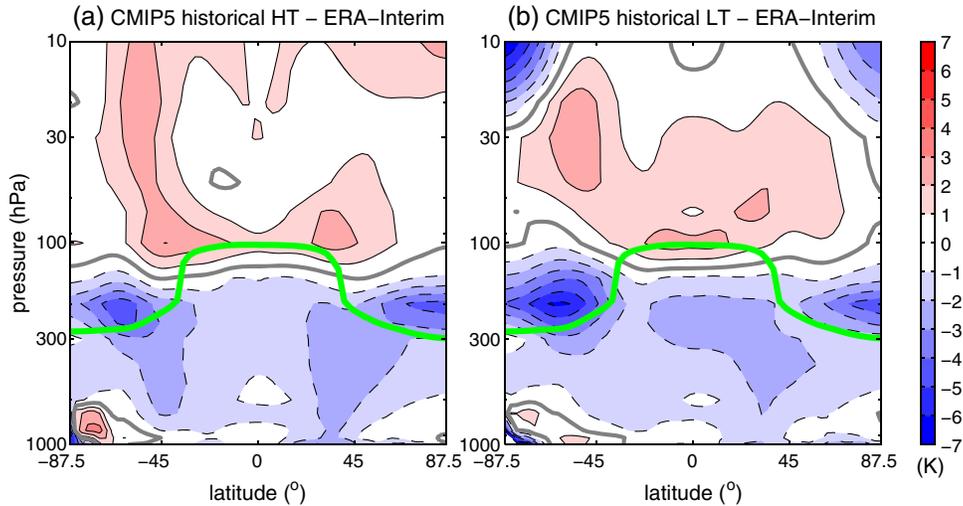
Figure 2. Zonal mean annual mean temperature difference between the multimodel average temperature and the ERA-Interim reanalysis over the period 1979–2000 for (a) the high-top models and (b) the low-top models. Contour interval is 1 K and zero lines are shown in bold grey.

[25] Stratospheric sudden warming events (SSWs) are the most dramatic examples of wintertime, extratropical stratospheric variability and are often followed by large perturbations to the tropospheric flow [e.g., *Baldwin and Dunkerton*, 2001]. To properly represent stratospheric climate variability, a climate model should be expected to simulate stratospheric warmings at approximately the same frequency as long-term reanalysis data sets and with a similar climatological distribution. *Charlton and Polvani* [2007] showed that, on an event by event basis, SSWs contribute to the daily variability in the middle and lower stratosphere and the interannual variability in the lower stratosphere. We confirm at the end of this section that, generally, models with a larger frequency of major SSW events also have large daily and interannual zonal mean zonal wind variance in the middle stratosphere.

[26] There are many different ways in which the occurrence of SSW events in the stratosphere can be detected. In the present analysis, we use the algorithm of *Charlton and Polvani* [2007], based on measuring the number of times that the zonal mean zonal wind at 60°N and 10 hPa crosses zero during midwinter, which has been used to evaluate SSW occurrence in both reanalysis data sets and a large number of previous GCM studies. While other studies have suggested potential modifications to the algorithm or alternative algorithms, we retain the algorithm in its original form to allow ease of comparison with other studies. In addition, there is growing evidence that substantial decadal variability in SSW occurrence may exist [*Schimanke et al.*, 2010]. Any analysis that seeks to characterize the SSW climatology of GCMs needs to take this into account. Therefore, in this analysis we use the period 1960–2005 for the models (the end of the historical simulation) so that a moderately large sample of SSW events in both models and reanalysis data sets can be considered. The calculated model SSW frequency is compared to the frequency from the ERA-40 reanalysis for the period 1958–2001. Previous studies of high-top models, both with and without coupled stratospheric chemistry [*Charlton et al.*, 2007; *Butchart*

et al., 2011] have highlighted the wide spread in the simulation of SSW frequency.

[27] Figure 3a shows the difference between the frequency of SSW events in the high-top and low-top ensembles by computing the mean decadal frequency of SSW events in each model, and provides an estimate of the 95% confidence interval for each frequency estimate. All the high-top models produce SSWs at a frequency consistent with the estimate from the ERA-40 climatology. On the other hand, almost all low-top models produce too few SSWs, with one model failing to produce any SSW events during the analyzed period. This difference is highlighted in the high-top and low-top ensemble averages, with high-top models on average producing more than double the frequency of SSW events as the low-top models.

[28] It is also interesting to compare the simulations of SSWs in the two versions of the IPSL model, which differ only in their horizontal resolution. The medium-resolution (MR) version of the IPSL model produces SSWs with almost double the frequency of the low-resolution (LR) version although note the large standard error on the MR estimate and the overlap between the confidence interval for the LR and MR estimates). Studies with idealized models have pointed to the potential sensitivity of stratospheric dynamical variability to horizontal resolution [*Scott et al.*, 2004], but to our knowledge this is the first evidence of that sensitivity in more comprehensive models. As a point of information, the horizontal resolution of the LR version is 1.875° × 3.75° and the MR version is 1.25° × 2.5°. *Scott et al.* [2004] found a significant underestimation of Rossby wave propagation and breaking for dynamical cores at resolutions coarser than T42 (approximately equivalent to a resolution of 2.8° × 2.8°), which is consistent with the fewer SSWs in the low-resolution compared to the medium-resolution IPSL model.

[29] Detailed analysis of the monthly climatology of SSW events (not shown) shows the wide range of behavior of GCMs in this diagnostic. Although the climatological distribution of SSW events is of course subject to significant sampling uncertainty, the results are robust enough to suggest that all models, but particularly low-top models, shift
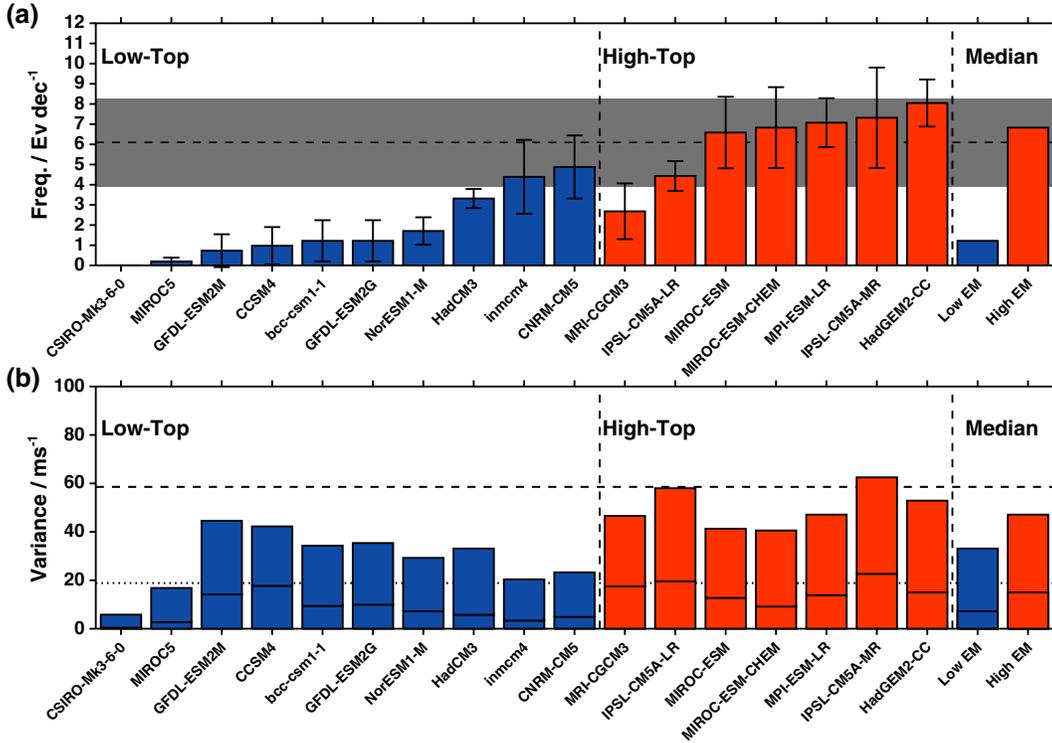
Figure 3. Daily and interannual variability in the models. (a) Climatological mean decadal frequency of stratospheric sudden warming events, 1960–2000 in 19 historical simulations of CMIP5 models. Colored bars show the number of SSW events per month calculated by the Charlton and Polvani algorithm, along with 95% confidence intervals for each estimate. Models shown in red are classified as high-top models, those shown in blue as low-top models. The climatological mean decadal frequency in the ERA-40 reanalysis data set is shown in the horizontal dashed black line and the 95% confidence interval for this estimate in gray. On the right of the plot, median estimates for the low-top and high-top ensembles are shown. (b) Ensemble mean total (full bar) and interannual (horizontal line crossing each bar) for the de-seasonalized zonal mean zonal wind at 60°N and 50 hPa during the period 1960–2000. The thick dashed line and thick dotted line show the total and interannual variance (respectively) for the ERA-40 reanalysis data set for the same period.

the majority of their SSW events toward the end of winter, and over-estimate the frequency of SSW events during March.

[30] To confirm that the frequency of major SSW events is a good indicator of the wintertime daily and interannual variability in the models, in Figure 3b we plot estimates of the total and interannual variance of the de-seasonalized zonal mean zonal wind at 50 hPa and 60°N for the same models. Although there is not a one-to-one relationship between SSW frequency and variance, it is clear that models with a higher frequency of SSW events also tend to have a larger total and interannual variance. In both measures, models in the high-top ensemble are closer to the value derived from the ERA-40 reanalysis.

[31] In summary, low-top models underestimate the frequency of major SSW events during midwinter and this is the main reason for the poorer skill of these models in simulating daily and interannual variability. As mentioned earlier, this is likely linked to the cold biases in the NH winter from 10–30 hPa in the low-top ensemble, in the sense that the cold bias and the limited number of SSW events are both likely related to the weak response of

the stratosphere to tropospheric wave driving (see Table S3). Given the well-known link between SSWs and seasonal variations in tropospheric climate [e.g., *Baldwin and Dunkerton*, 2001], it is also possible that tropospheric seasonal climate variability will be different between the low-top and high-top ensemble. We begin to explore this idea in section 5.1.

4.3. Decadal Variability

[32] A significant source of stratospheric decadal variability over the historical period studied is that due to the large volcanic eruptions of El Chichón (in 1982) and Mt. Pinatubo (in 1991) and the relative paucity of eruptions since.

[33] To assess the ability of the current generation models to reproduce postvolcanic dynamical anomalies in the stratosphere, we compare here anomalies of lower stratospheric (50 hPa) zonal mean geopotential height from the CMIP5 models and ERA-interim. Posteruption geopotential height anomaly composites are then produced by averaging seasonal geopotential height anomalies after the 1982 El Chichón and 1991 Mt. Pinatubo eruptions. In the NH, the two winters following each eruption are averaged, while

for the SH, since aerosol transport to the SH after El Chichón was quite weak, only the first spring after the El Chichón eruption is averaged together with the two springs following the Pinatubo eruption. Figure 4 shows the posteruption geopotential height anomalies for a number of CMIP5 historical simulation ensemble members (as detailed in Table 2) and from ERA-interim in SH spring (OND) and NH winter (DJF), seasons which have been identified as showing maximum posteruption circulation anomalies.

[34] Both high-top and low-top ensemble means show similar postvolcanic response, with a modest increase in 50 hPa geopotential height over most latitudes. No CMIP5 model simulation from either the low-top or high-top ensemble reproduces the strong volcanic response in the NH winter stratosphere, a result which is consistent with the recent analysis of *Driscoll et al.* [2012]. In SH spring, the high-top ensemble mean 50 hPa geopotential height anomaly has a slightly positive value in the high latitudes, consistent with the reanalysis, but the large ensemble spread indicates that the high-top model simulations, as an ensemble, are not significantly closer to reality than the low-top model results. The poor performance of both model ensembles in simulating decadal variability in the stratosphere (see Figure 1) is strongly tied to this deficiency but may also be related to the ability of the models to reproduce the smaller signal due to solar variability which has not been tested. The impact of the weak decadal variability in both ensembles on their simulation of past temperature trends is assessed in section 5.2.

5. The Impact of Reduced Stratospheric Variability

[35] It is clear from section 4.2 that the low-top CMIP5 model ensemble is deficient in its simulation of stratospheric variability on daily and interannual time scales. In this final section, we explore the impact of this deficiency on the

simulation of coupling between the stratosphere and troposphere by the models and on the simulation of stratospheric trends. We choose to focus on these two areas, because they are relevant to many different applications of CMIP5 data and should be of interest to a wide range of fellow climate scientists.

5.1. Stratosphere-Troposphere Coupling

[36] We attempt to characterize stratosphere-troposphere coupling in the models by examining the annular modes. The annular modes characterize variability of the tropospheric midlatitude jets and the stratospheric polar vortex [e.g., *Thompson and Wallace*, 2000]. Following *Baldwin and Dunkerton* [2001] and *Baldwin and Thompson* [2009], we compute the annular mode index separately at each pressure level to explore the vertical coupling of the atmosphere. We use the procedure of *Gerber et al.* [2010], defining the Northern and Southern Annular Modes (NAM and SAM) as the first Empirical Orthogonal Functions of daily zonal mean geopotential height anomalies from each hemisphere. The height fields are first filtered (by removing a 30 year low-pass filtered version of the time series) to separate climate trends, so that the remaining anomalies reflect the natural variations of the atmosphere.

[37] Figure 5 shows composites of the NAM index based on extreme events in the stratosphere. Negative NAM events in the stratosphere are associated with a weak polar vortex, and closely (though not entirely) correspond to stratospheric sudden warmings [*Charlton and Polvani*, 2007]. Figure 5a, based on ERA-40 and ERA-Interim, is an update of *Baldwin and Dunkerton* [2001]; following a weakening of the stratospheric polar vortex, the NAM in the troposphere tends to shift toward a negative index, associated with an equatorward shift of the midlatitude jet stream.

[38] Figures 5b and 5c suggest a difference in the response of the high- and low-top ensembles to extreme stratospheric events. The high-top ensemble composite (Figure 5b) is
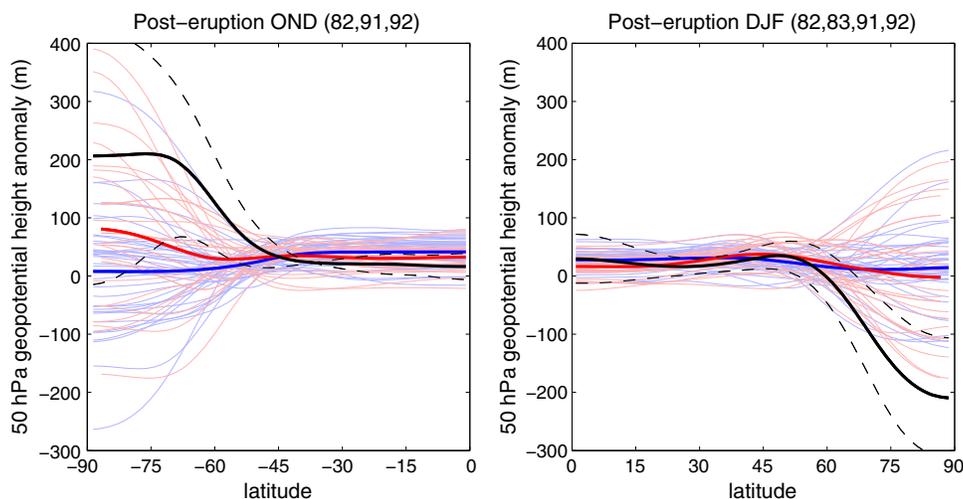

Figure 4. Averaged geopotential height anomaly at 50 hPa for (right) two boreal winters following the El Chichón and Pinatubo eruptions in the historical simulations in the NH and for (left) one OND period following the El Chichón and two OND periods following the Pinatubo eruptions in the SH. Thin lines show individual models, thick lines show the ensemble average (except black line which shows anomaly in the ERA-Interim reanalysis). Blue lines show low-top models and red lines show high-top models. Black dashed lines show the two standard error estimate for the ERA-Interim data.
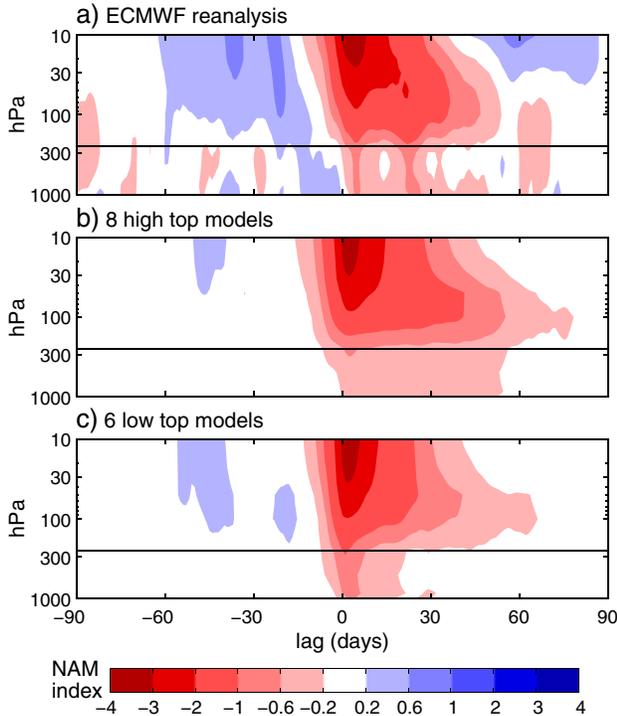
Figure 5. Downward propagation of negative NAM events in ECMWF reanalyses and CMIP5 models. Events are defined by instances when the NAM index at 10 hPa drops below –3 standard deviations, and the composites are constructed from NAM anomalies in the 90 days preceding (negative lags) and proceeding (positive lags) the dip at 10 hPa. The multimodel composites are based on all events in the CMIP5 historical simulations in either the high or low-top ensembles, but weighted so that each model makes an equal contribution to the composite.

quite similar to the reanalyses; the NAM throughout the atmosphere shifts toward a negative index following an event at 10 hPa, with a slight lag in the troposphere, and persists for approximately 60 days in the lower stratosphere and troposphere. In the low-top ensemble, the initial response to stratospheric perturbations is similar, but with a slightly stronger response in the troposphere. The events, however, fail to persist as long, particularly in the troposphere. Composites based on extreme positive NAM events, when the polar vortex becomes very strong, reveal a similar bias in the low-top ensemble; the magnitude of the tropospheric response is correct, but short lived (not shown).

[39] Further analysis suggests that the differences in the models' tropospheric response is due to a difference in the temporal variability of the lower stratosphere. Following *Baldwin et al.* [2003], the e-folding time scale of the NAM was computed as a function of height. The time scales of variability in the troposphere are fairly similar in the reanalysis and both model ensembles (all about 10–15 days). In the lower stratosphere, however, time scales were reduced in the low-top models, peaking at approximately 20 days, compared to approximate 30 days in the reanalyses and 35 days in the high-top ensemble. Both model ensembles show a bias in the timing of the seasonal peak, which occurs around January in the reanalyses, compared to March in the

models. This bias in the timing is consistent with CCMVal2 and CMIP3 models [*Gerber et al.*, 2010]. The persistence of anomalies in the lower stratosphere appears critical to sustaining the tropospheric response.

[40] It is important to point out that extreme NAM variability is defined relative to the background variability in each model separately. As the variance of the middle stratosphere extratropics is smaller in the low-top models than in the high-top models (not shown), their NAM composites are based on smaller absolute changes to the geopotential height field. Nonetheless, the diagnostic highlights potential links between biases in natural stratospheric variability and the dynamical coupling between the stratosphere and troposphere in the low-top models.

5.2. Reproduction of Past Stratospheric Trends

[41] Because the CMIP5 models will be used extensively to understand and predict future climate, a key test of their fidelity in the stratosphere is their ability to reproduce past stratospheric trends. In this section, we test if the missing stratospheric variability in the low-top models has any impact on their ability to reproduce past trends in the lower stratosphere. We focus here on the satellite period for which coverage is almost globally-complete, and compare observed remote sensing system microwave sounding unit (MSU) temperature of the lower stratosphere (TLS) [*Mears and Wentz*, 2009] (a layer temperature with a weighting function maximum located at close to 80 hPa), with simulated temperatures from CMIP5 models. We compare with an MSU observational data set, because stratospheric trends are not robust in reanalyses [*Xu and Powell*, 2011], and radiosonde coverage is spatially incomplete. We focus on the lower stratosphere, because several models considered do not resolve the region covered by the stratospheric sounding units. Each historical simulation to 2005 was concatenated with either the corresponding extended historical simulation where available, or with the corresponding RCP 4.5 simulation to compare with observations up to 2011 and to improve the signal-to-noise ratio. In both cases, radiative forcings are continuous through the simulations and close to those observed over the past 6 years [*Gillett et al.*, 2012]. A fixed weighting function with a maximum near 80 hPa [*Mears and Wentz*, 2009] was applied to zonal mean monthly mean data. Each individual simulation was given equal weight in the analysis, meaning that models with larger ensembles were given more weight (see Table 2). IPSL-CM5A-LR, IPSL-CM5A-MR, and INMCM4 were excluded from the analysis because they did not include volcanic aerosols, and CNRM-CM5 was excluded because its simulated stratospheric ozone changes were underestimated compared to those observed.

[42] Figure 6a shows the generally good agreement between simulated and observed TLS in both high-top and low-top simulations. Following Pinatubo in 1992, the positive temperature anomaly is overestimated in both sets of simulations. Given that the models tend to underestimate the 50 hPa geopotential height anomalies associated with volcanic eruptions (Figure 4), it appears that models may have some difficulty capturing the structure of the warming or other feedbacks with the circulation. The overestimation of the temperature response is larger for the high-top simulations and is explained mainly by an overestimation of the response to Pinatubo in MRI-CGCM3, even though its
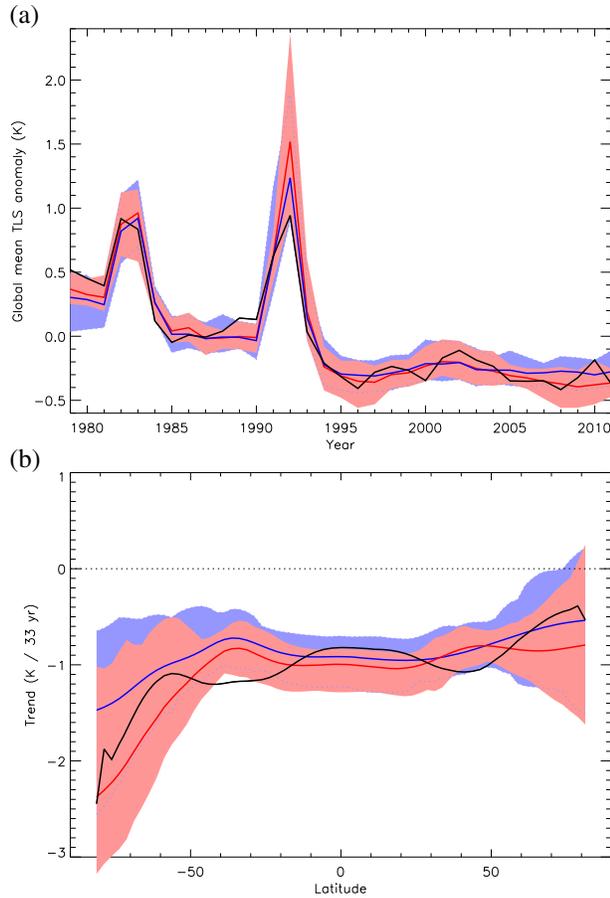
(a)

(b)

Figure 6. Comparison of simulated and observed lower stratospheric temperatures. (a) Global mean annual mean lower stratospheric temperature anomalies from the 1979–2011 mean, based on the remote sensing solutions data set (black), the mean of the high-top models (red), and the mean of the low-top models (blue). Global means are calculated over the observed latitudes 82.5°S–82.5°N. Pink and blue shaded bands show the approximate 5–95% ensemble range. (b) Corresponding zonal mean temperature trends in °C over the 1979–2011 period. Volcano years (1982, 1983, 1991, 1992, and 1993) were excluded from the analysis, and linear trends were fitted by least squares.

response to El Chichón is relatively consistent with observations. MRI-CGCM3 is the only model considered here that interactively simulated the stratospheric aerosol distribution and its radiative properties following volcanic eruptions, unlike the other models in which the aerosol distribution and properties are prescribed from observations [*Driscoll et al.*, 2012]. Our results suggest that this model's stratospheric aerosol scheme leads to an overestimation of the stratospheric temperature response to Pinatubo. A further apparent difference between the observations and simulations is in the first 3 years (1979–1981), in which observed TLS temperatures are higher than those in either the high-top or low-top ensembles. Overall, agreement between simulations and observations appears better for the CMIP5 simulations than for the CMIP3 simulations [*Cordero and Forster*, 2006], with the volcanic response less overestimated than in CMIP3, and perhaps also because of the more realistic ozone changes specified in the

CMIP5 simulations. Agreement with observations also appears better for the CMIP5 models than for the CCMVal models [*Forster et al.*, 2011; *Austin et al.*, 2009], perhaps because the former mostly use an observational data set of stratospheric ozone changes usually derived from *Cionni et al.* [2011], rather than the internally simulated ozone concentration of the CCMVal simulations. It is important to note, however, that the observed decreases in ozone concentrations following major volcanic eruptions are not included in the forcing data set, and hence the impact of stratospheric aerosol changes on the lower-stratospheric temperature trends may be slightly amplified.

[43] Figure 6b compares the pattern of zonal mean temperature trends in simulations and observations, with the 2 years following El Chichón and Pinatubo excluded. Cooling is seen at all latitudes in observations and high-top and low-top ensemble means. The level of agreement with observations is generally good, and there is no significant difference between the level of agreement with observations of the high-top and low-top simulations. Whereas observed trends calculated over the 1979–2005 period show little latitudinal variation [*Gillett et al.*, 2011], in part due to anomalous warmth over the Antarctic in 2002 [*Varotsos*, 2002; *Allen et al.*, 2003], the trend calculated over the 1979–2011 period shows enhanced Antarctic cooling, consistent with the simulations. This is associated with the years since 2005 all being colder than average over the Antarctic, with 2011 particularly cold [*Seidel et al.*, 2011].

[44] Overall, we find that the CMIP5 models considered are able to simulate observed temperature changes in the lower stratosphere rather well. The CMIP5 models appear to perform better in this respect than either the CMIP3 models or the CCMVal models, which may relate to the realistically prescribed stratospheric ozone variations. Overall, although there are few significant differences between the temperature trends simulated by the low-top and high-top models the simulation of stratospheric trends by the low-top models seems unaffected by their reduced stratospheric variability.

6. Discussion and Conclusions

[45] In this study, the simulation of stratospheric climate and variability by two subensembles of the full CMIP5 ensemble, one containing models with a high-top and one containing models with a low-top, is compared. Away from the model top, the simulation of mean stratospheric climate by the two ensembles is similar and better than in the CMIP3 ensemble. However, the low-top ensemble has a poorer simulation of stratospheric daily and interannual variability than the high-top ensemble. In the Northern Hemisphere, this difference is due to a lack of episodic stratospheric variability in midwinter in the low-top models.

[46] The lack of dynamical variability in the low-top models has an impact on stratosphere-troposphere coupling, resulting in relatively short-lived anomalies in the Northern Annular Mode, which do not persist in either the stratosphere or troposphere. In this respect, the high-top models more faithfully reproduce the stratosphere-troposphere coupling seen in the reanalysis data. Conversely however, the lack of daily and interannual stratospheric variability in the low-top models does not seem to affect their ability to reproduce

historical trends in lower stratospheric temperature. Both the high-top and low-top ensembles are able to reproduce global stratospheric trends with good fidelity.

[47] Of the three time scales of variability examined, decadal variability is the least well represented by CMIP5 models compared to observations, with little difference between the high-top and low-top ensembles. Examining the stratospheric response to volcanic eruptions, a significant source of forced decadal variability, we find realistic global mean temperature responses in the models, but a lack of dynamical response as seen in observations.

[48] Few of the models in CMIP5 place their model top at 10 hPa, thereby severely truncating stratospheric dynamical behavior, but it is clear from our analysis that making a modeling choice to place the model top below the stratopause still has the potential to severely limit stratospheric dynamical variability. Simulating realistic stratospheric dynamical variability remains a challenge for all climate models, even those with high-tops. For example, work with the suite of Canadian climate models has shown that the simulation of stratospheric dynamical variability is very sensitive to small changes in the stratospheric mean state and it is therefore very important to continue work to understand biases in the stratospheric climate or climate models and seek to improve them [see for example, *Sigmond and Scinocca*, 2010]. The single model studies by *Hardiman et al.* [2012] and *Shaw and Perlwitz* [2010] come to a similar conclusion that the current generation of low-top and high-top models have a very similar simulation of stratospheric mean climate and trends but differ in their simulation of stratospheric variability.

References

Allen, D. R., R. M. Bevilacqua, G. E. Nedoluha, C. E. Randall, and G. L. Manney (2003), Unusual stratospheric transport and mixing during the 2002 Antarctic winter, *Geophys. Res. Lett.*, 30, 1599, doi:10.1029/2003GL017117.

Austin J., D. Shindell, S. R. Beagley, C. Brühl, M. Dameris, T. Nagashima, P. Newman, S. Pawson, G. Pitari, E. Rozanov, C. Schnadt, and T. G. Shepherd (2003), Uncertainties and assessments of chemistry-climate models of the stratosphere, *Atmos. Chem. Phys.*, 3, 1–27.

Austin, J., et al. (2009), Coupled chemistry climate model simulations of stratospheric temperatures and their trends for the recent past, *Geophys. Res. Lett.*, 36, L13809, doi: 10.1029/2009GL038462.

Baldwin, M. P., and T. J. Dunkerton (2001) Stratospheric Harbingers of Anomalous Weather Regimes, *Science*, 294, doi:10.1126/science.1063315.

Baldwin M. P., and D. W. J. Thompson (2009), A critical comparison of stratosphere-troposphere coupling indices, *Q. J. Roy. Met. Soc.*, 135, 1661–1672.

Baldwin, M. P., D. B. Stephenson, David 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.

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

Cagnazzo C., and E. Manzini (2009), Impact of the Stratosphere on the Winter Tropospheric Teleconnections between ENSO and the North-Atlantic European Region, *J. Climate*, 22, 1223–1238.

Charlton A. J., and L. M. Polvani (2007), A new look at stratospheric sudden warmings. Part I. Climatology and modeling benchmarks, *Journal of Climate*, 20, 449–471.

Charlton A. J., L. M. Polvani, J. Perlwitz, F. Sassi, E. Manzini, S. Pawson, J. E. Nielsen, K. Shibata, and D. Rind (2007), A new look at stratospheric sudden warmings. Part II. Evaluation of Numerical Model Simulations, *Journal of Climate*, 20, 471–488.

Cionni, I., V. Eyring, J. F. Lamarque, W. J. Randel, D. S. Stevenson, F. Wu, G. E. Bodeker, T. G. Shepherd, D. T. Shindell, and D. W. Waugh (2011), Ozone database in support of CMIP5 simulations: results and corresponding radiative forcing, *ect Atmos. Chem. Phys.*, 11, 11267–11292, doi:10.5194/acp-11-11267-2011.

Cordero, E. C., and P. M. de F. Forster (2006), Stratospheric variability and trends in models used for the IPCC AR4, *Atmos. Chem. Phys.*, 6, 5369–5380, doi:10.5194/acp-6-5369-2006.

Dee D. P., et al. (2011), The ERA-Interim re-analysis: configuration and performance of the data assimilation system, *Quarterly Journal of the Royal Meteorological Society*, 137, 553–597, doi:10.1002/qj.828.

Driscoll, S., A. Bozzo, L. J. Gray, A. Robock, and G. Stenchikov (2012), Coupled Model Intercomparison Project 5 (CMIP5) Simulations of Climate Following Volcanic Eruptions, *J. Geophys. Res.*, in press, doi:10.1029/2012JD017607.

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

Forster P. M., et al. (2011), Evaluation of radiation scheme performance within chemistry climate models, *J. Geophys. Res.*, 116, D10302.

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

Gerber E. P., et al. (2012), Assessing and Understanding the Impact of Stratospheric Dynamics and Variability on the Earth System, *Bull. Am. Met. Soc.*, http://dx.doi.org/10.1175/BAMS-D-11-00145.1.

Gillett, N. P., et al. (2011), Attribution of observed changes in stratospheric ozone and temperature, *Atmos. Chem. Phys.*, 11, 599–609.

Gillett, N. P., V. K. Arora, G. M. Flato, J. F. Scinocca, and K. von Salzen (2012), Improved constraints on 21st-century warming derived using 160 years of temperature observations, *Geophys. Res. Lett.*, 39, L01704, doi:10.1029/2011GL050226.

Hardiman, S. C., N. Butchart, T. J. Hinton, S. M. Osprey, and L. J. Gray (2012), The effect of a well resolved stratosphere on surface climate: Differences between CMIP5 simulations with high and low top versions of the Met Office climate model, *protect J. Climate*, 25, 7083–7099, doi: http://dx.doi.org/10.1175/JCLI-D-11-00579.1.

Karpechko, A. Y., and E. Manzini (2012), Stratospheric influence on tropospheric climate change in the Northern Hemisphere, *J. Geophys. Res.*, 117, D05133, doi:10.1029/2011JD017036.

Manzini E., et al. (2012), Role of the stratosphere in Northern winter climate change as simulated by the CMIP5 models, *J. Geophys. Res.*, submitted.

Maycock A. C., S. P. E. Keeley, A. J. Charlton-Perez, and F. J. Doblas-Reyes (2011), Stratospheric circulation in seasonal forecasting models: implications for seasonal prediction, *Clim. Dyn.*, 36, 309–321.

Mears C. A., and F. J. Wentz (2009), Construction of the RSS V3.2 Lower-Tropospheric Temperature Dataset from the MSU and AMSU Microwave Sounders, *J. Atmos. Oceanic Technol.*, 26, 1493–1509, doi: 10.1175/2009JTECHA1237.1.

Meehl G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor (2007), The WCRP CMIP3 Multimodel Dataset: A new era in Climate Change Research, *Bull. Am. Met. Soc.*, 136(652), 1313–1394, doi:10.1175/BAMS-88-9-1383.

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

Reichler T., and J. Kim (2008), How well do coupled models simulate today's climate?, *Bull. Am. Met. Soc.*, doi:10.1175/BAMS-89-3-303.

Rienecker, M. M., et al. (2011), MERRA - NASA's Modern-Era Retrospective Analysis for Research and Applications, *J. Climate*, 24, 3624–3648, doi: 10.1175/JCLI-D-11-00015.1.

Scaife A. A., et al. (2011), Climate-change projections and stratosphere-troposphere interaction, *Climate Dynamics*, 10.1007/s00382-011-1080-7.

Schimanke S., J. Koerper, T. Spangehl, and U. Cubasch (2010), Multi-decadal variability of sudden stratospheric warmings in an AOGCM, *Geophysical Research Letters*, 38, L01801.

Scott, R. K., D. G. Dritschel, L. M. Polvani, and D. W. Waugh (2004), Enhancement of Rossby Wave breaking by Steep Potential Vorticity Gradients in the Winter Stratosphere, *J. Atmos. Sci.*, 61, 904–918.

Seidel, D. J., N. P. Gillett, J. R. Lanzante, K. P. Shine, and P. W. Thorne (2011), Stratospheric temperature trends: our evolving understanding, *Wiley Interdiscip. Rev. Clim. Change*, 2, 592–616.

Shaw, T. A., and J. Perlwitz (2010), The impact of stratospheric model configuration on planetary scale waves in northern hemisphere winter, *J. Climate*, 22, 3369–3389.

Sigmond M., and J. F. Scinocca (2010), The Influence of Basic State on the Northern Hemisphere Circulation Response to Climate Change, *J. Clim.*, 23, 1434–1446.

SPARC CCMVal (2010), SPARC Report on the Evaluation of Chemistry-Climate Models, V. Eyring, T. G. Shepherd, D. W. Waugh (Eds.), SPARC Report No. 5, WCRP-132, WMO/TD-No. 1526, http://www.atmosp.physics.utoronto.ca/SPARC

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012), An Overview of CMIP5 and the Experimental Design, *Bull. Am. Met. Soc.*, 93, 485–498, doi: http://dx.doi.org/10.1175/BAMS-D-11-00094.1

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

Varotsos, C. (2002), The Southern Hemisphere ozone hole split in 2002, *Environ. Sci. Pollut. Res.*, 6, 375–376.

Xu, J., and A. M. Powell Jr. (2011), Uncertainty of the stratospheric/tropospheric temperature trends in 197920132008: multiple satellite MSU, radiosonde, and re-analysis datasets, *Atmos. Chem. Phys.*, 11, 10727–10732, doi:10.5194/acp-11-10727-2011.