Atmos. Chem. Phys., 10, 1385–1400, 2010
www.atmos-chem-phys.net/10/1385/2010/

Atmospheric Chemistry and Physics

Quantitative assessment of Southern Hemisphere ozone in chemistry-climate model simulations

A. Yu. Karpechko[1,*], **N. P. Gillett**[2], **B. Hassler**[3,4,5], **K. H. Rosenlof**[4], **and E. Rozanov**[6,7]

[1]Climatic Research Unit, School of Environmental Sciences, University of East Anglia, UK
[2]Canadian Centre for Climate Modelling and Analysis, Environment Canada, Canada
[3]National Institute of Water and Atmospheric Research, Lauder, New Zealand
[4]NOAA, Earth System Research Laboratory, Boulder, USA
[5]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, USA
[6]Institute for Atmospheric and Climate Science, ETH Zürich, Switzerland
[7]Physical-Meteorological Observatory/World Radiation Center, Davos, Switzerland
[*]now at: Finnish Meteorological Institute, Arctic Research, Helsinki, Finland

Received: 6 July 2009 – Published in Atmos. Chem. Phys. Discuss.: 17 September 2009
Revised: 15 January 2010 – Accepted: 25 January 2010 – Published: 8 February 2010

Abstract. Stratospheric ozone recovery in the Southern Hemisphere is expected to drive pronounced trends in atmospheric temperature and circulation from the stratosphere to the troposphere in the 21st century; therefore ozone changes need to be accounted for in future climate simulations. Many climate models do not have interactive ozone chemistry and rely on prescribed ozone fields, which may be obtained from coupled chemistry-climate model (CCM) simulations. However CCMs vary widely in their predictions of ozone evolution, complicating the selection of ozone boundary conditions for future climate simulations. In order to assess which models might be expected to better simulate future ozone evolution, and thus provide more realistic ozone boundary conditions, we assess the ability of twelve CCMs to simulate observed ozone climatology and trends and rank the models according to their errors averaged across the individual diagnostics chosen. According to our analysis no one model performs better than the others in all the diagnostics; however, combining errors in individual diagnostics into one metric of model performance allows us to objectively rank the models. The multi-model average shows better overall agreement with the observations than any individual model. Based on this analysis we conclude that the multi-model average ozone projection presents the best estimate of future ozone evolution and recommend it for use as a boundary condition in future climate simulations. Our results also demonstrate a sensitivity of the analysis to the choice of reference data set for vertical ozone distribution over the Antarctic, highlighting the constraints that large observational uncertainty imposes on such model verification.

Correspondence to: A. Yu. Karpechko
(alexey.karpechko@fmi.fi)

1 Introduction

In the last two decades of the 20th century stratospheric ozone, which accounts for about 90% of the total ozone, has declined significantly as a result of chemical destruction by anthropogenic halogen-containing compounds (WMO, 2007). In the Southern Hemisphere (SH) where ozone depletion is particularly severe in high-latitudes in spring, ozone changes have led to cooling of the lower stratosphere and an increase in the lifetime of the Antarctic polar vortex (e.g., Randel and Wu, 1999; Zhou et al., 2000). These changes have further led to intensification of the tropospheric circumpolar circulation (Thompson and Solomon, 2002; Gillett and Thompson, 2003). Among other impacts, the intensification of the tropospheric circulation has contributed to significant decrease of rainfalls in southwest Australia (Cai et al., 2005) and to dramatic warming of the Antarctic Peninsula (Marshall et al., 2006).

Stratospheric ozone is expected to recover during the 21st century as a result of declining halogen abundances (WMO, 2007) and ozone recovery is expected to influence the position and strength of SH tropospheric westerlies, likely reversing the strengthening of westerlies caused by ozone depletion (Thompson and Solomon, 2002). However due to continuing increase of the greenhouse gas concentration, which act to further strengthen westerlies (e.g. Fyfe et al., 1999), the net change to the tropospheric circulation is unclear (Miller et al., 2006; Perlwitz et al., 2008; Son et al., 2008, 2009). This implies that details of the ozone recovery need to be predicted well in order to reliably simulate future SH climate.

Presently, full representation of stratospheric chemistry in climate models is quite expensive and the majority of coupled atmosphere-ocean climate models use prescribed ozone fields. Models that were used for the Intergovernmental Panel of Climate Change Forth Assessment Report (IPCC AR4) used either a simplified ozone recovery scenario or even assumed constant ozone (i.e. annual cycle not varying from year to year) throughout the 21st century (Miller et al., 2006). More physically sound future ozone scenarios are provided by coupled Chemistry-Climate Models (CCMs). These models account for interactions between stratospheric ozone chemistry and atmospheric physics and dynamics which may change due to projected greenhouse gases (GHGs) increases. However ozone projections by these models differ from model to model (Eyring et al., 2007) raising the question of which ozone scenario is more reliable.

Information on model performance in simulating present climate may be used to decide which model's projection is more reliable (Reichler and Kim, 2008; Gleckler et al., 2008). However models are tuned to represent the present climate and the best tuned model may not simulate future climate more correctly. Yet, without a better alternative, model ranking based on their ability to simulate present climate and observed trends looks like a reasonable approach and is widely employed (e.g. Connolley and Bracegirdle, 2007; Bracegirdle et al., 2008).

Eyring et al. (2006) assessed different aspects of performance of several CCMs including their ozone simulation skill; however they did not derive any quantitative metric of agreement between simulations and observations. Waugh and Eyring (2008) (hereinafter "WE08") carried out a quantitative assessment of CCMs' ability to simulate several key processes relevant to stratospheric ozone since they argued that a process-oriented evaluation might be a better predictor of a models' ability to make reliable ozone projections; because of that they did not assess models' skill at simulating ozone itself. Nevertheless, it is important to assess models' ability to simulate ozone climatology and trends; our study may be considered a complimentary to that of WE08. It is also of interest to look at how models skill in simulating ozone-related processes correlates with their skill in simulating ozone itself. The goal of this study is to provide climate modellers with a guideline for choice of future ozone scenario for simulations with prescribed ozone fields. To achieve this we perform a quantitative assessment of CCM skills in simulating observed ozone climatology and ozone trends with a focus on the SH, where the largest impacts of ozone recovery on the climate are expected. As a reference, we employ several available up-to-date observational data sets, which allow us to evaluate uncertainties associated with the observations.

Table 1. CCMs used in this study.

Model name	Reference
CCSRNIES	Akiyoshi et al. (2004)
CMAM	Fomichev et al. (2007)
E39C	Dameris et al. (2005)
GEOSCCM	Pawson et al. (2008)
LMDZrepro	Jourdain et al. (2008)
MAECHAM4CHEM	Steil et al. (2003)
MRI	Shibata and Deushi (2005)
SOCOL	Egorova et al. (2005)
ULAQ	Pitari et al. (2002)
UMETRAC	Austin (2002)
UMSLIMCAT	Tian and Chipperfield (2005)
WACCM	Garcia et al. (2007)

2 Data

We use output of twelve CCMs assembled in the Chemistry-Climate Model Validation (CCMVal) Archive at the British Atmospheric Data Center (BADC). These twelve CCM groups contributed data to the first round of CCMVal (CCMVal-1). The goal of CCMVal is "to improve understanding of CCMs and their underlying GCMs (General Circulation Models) through process-oriented evaluation" (CCMVal: http://www.pa.op.dlr.de/CCMVal/) and the first round was accomplished in support of WMO ozone assessment 2006 (WMO, 2007). The models are listed in Table 1 together with a reference for each model. We consider simulations of the last two decades of the 20th century based on forcings described in Eyring et al. (2006). These include observed sea surface temperature, sea ice concentrations, surface concentrations of well-mixed GHGs and halogens, solar variability, and aerosol from major volcanic eruptions. For all the models except MRI and SOCOL, outputs from the simulations performed in support of WMO ozone assessment 2006 are used. For MRI we use data from an updated run with an improved transport scheme (see http://www.pa.op.dlr.de/CCMVal/CCMVal_ErrataBADC.html). For SOCOL we use simulations from the model V2.0 described in Schraner et al. (2008). Compared with the model V1.0 used by Eyring et al. (2006), this version has improved parameterization of stratospheric water vapour condensation, a more sophisticated heterogeneous chemistry scheme, and improved transport scheme (Schraner et al., 2008). Here we mainly use ozone outputs from the first simulation of each model and restrict our attention to the period of 1980–1999 for which outputs from all the models and sufficient observations are available. Additional simulations started from different initial conditions are available for SOCOL, MRI, and WACCM and are used to study sensitivity of the results to sampling errors. In addition to the individual models we also consider ensemble averaged ozone time series (MULTI).

Observational data sets used for model performance validation include total ozone and ozone profiles data sets from several sources. The merged satellite total ozone data set (TOMS/SBUV) is based on individual Total Ozone Mapping Spectrometer (TOMS) and Solar Backscatter Ultraviolet 2 (SBUV/2) data sets (Stolarski and Frith, 2006). Another total ozone data set used in this study is that compiled by Karen Rosenlof from satellite (SME, SAGE-II, MLS, HALOE and TOMS/SBUV) and standard ozone climatology data (Dall'Amico et al., 2010). The Rosenlof data set also provides ozone profiles. Two other ozone profile data sets used here are those described in papers by Randel and Wu (2007) and Hassler et al. (2009). The former (Randel data set) is based on a regression model fitted to SAGE 1 and 2 and ozonesonde profiles combined with a seasonally varying ozone climatology. Over the Antarctic region, which is of interest here, the model utilises only data from Syowa station located at 69° S and may not adequately represent the ozone field further south. Implications of this will be discussed below. The Randel data set is provided on the height levels. These was converted to pressure levels applying the equation 1013.25*exp($-z$/7), where z is height expressed in kilometres. The latter data set (Hassler data set) is based on satellite (SAGE 1 and 2, POAM 2 and 3, HALOE) and ozonesonde profiles. Due to the lack of the observations, the Rosenlof and the Hassler data sets also apply different techniques to fill in the gaps, which will be discussed in more detail below.

3 Method

To assess model performance we calculate a metric similar to that used by Reichler and Kim (2008) and Gleckler et al. (2008). First we calculate normalized root mean square (RMS) differences e_{jkl} between the j-th model and k-th reference observations for the l-th diagnostic

$$e^2_{jkl} = \frac{1}{W}\sum_i\sum_m (w_{im}(x_{imjl} - y_{imkl})^2/\sigma^2_{imkl}), \tag{1}$$

where x_{imjl} is the simulated variable and y_{imkl} is the observed variable at month m and grid point i, w_{im} is the weight assigned to each data point, $W = \sum w$, is a sum of individual weights and σ_{imkl} is a measure of the uncertainty in the observed variable y_{imkl}. In the following the value e_{jkl} will be referred to as model error in l-th diagnostic with respect to k-th reference observations. Calculations of the weights and the observation uncertainty are described below in this section.

Following Reichler and Kim (2008) we scale the errors in all diagnostics by the average error across the individual models to ensure that different diagnostics receive similar weights when calculating the combined metric of model performance

$$e'^2_{jkl} = \frac{e^2_{jkl}}{\frac{1}{J}\sum\limits_j e^2_{jkl}}, \tag{2}$$

where J is the number of models. Model errors are calculated with respect to several available observation-based data sets in order to reduce possible influence of biases in the observation-based data sets. However, the observation data sets are not completely independent since they share some of the same input data and therefore may suffer from similar biases. We next average the model errors with respect to all available reference data sets for each diagnostic

$$e'^2_{jl} = \frac{1}{K}\sum_k e'^2_{jkl}, \tag{3}$$

where K is the number of reference data sets. Finally, a model performance index (I) is calculated as an average across errors in all individual diagnostics

$$I^2_j = \frac{1}{L}\sum_l e'^2_{jl}, \tag{4}$$

where L is the number of diagnostics. A lower value of I indicates better overall agreement with the observations and is interpreted as a better model performance.

The choice of diagnostics and grading metric is inevitably subjective, which is a shortcoming of this approach (Connolley and Bracegridle, 2007; WE08). The sensitivity of our result to the choice of grading metric and diagnostics will be discussed in Sect. 4.3. The diagnostics we employ are listed in Table 2. Two of them are (1) monthly mean zonal mean total ozone climatology in the period 1980–1984, when the influence of ozone depletion was minimal; and (2) monthly mean zonal mean total ozone linear trend during the period 1980–1999. The period 1980–1984 is meant to represent pre-ozone hole climatology, although some ozone loss took place already then (see Fig. 1). Also, a five year period is somewhat short for calculating climatology; however using a longer period is restricted by absence of observations (and also some model data) before 1980 and by the increasing influence of ozone depletion after 1984. The influence of sampling errors on our results will be assessed by using additional simulations available for some models. Trends are calculated using linear, least squares regression of the ozone time series on time. We consider monthly zonal mean values and calculate model errors according to Eq. (1) by summation over months and latitudes. The weights w are cosine of the latitude. While the total ozone climatology errors are calculated globally the total ozone trend errors are only calculated over the SH. Ozone abundance in the NH extratropics during winter-early spring, when observed total ozone trends are largest, is strongly controlled by dynamics, in particular the

Table 2. Diagnostics used in this study.

Diagnostic	Diagnostic description
Global total ozone climatology	Total ozone climatology in 1980–1984, zonal mean monthly mean values, domain: 90° S–90° N, resolution: 5°
SH total ozone trend	Total ozone linear trend between 1980–1999, zonal mean monthly mean values, domain: 90° S–0° N, resolution: 5°
Polar SH vertical ozone distribution climatology	Ozone partial pressure profile climatology in 1980-1984, monthly mean values averaged over 90° S–60° S, levels: 500, 400, 300, 250, 200, 150, 130, 115, 100, 90, 80, 70, 50, 30, 20, 15, 10 hPa
Polar SH vertical ozone distribution trend	Ozone partial pressure profile linear trend between 1980–1999, monthly mean values averaged over 90° S–60° S, levels: 500, 400, 300, 250, 200, 150, 130, 115, 100, 90, 80, 70, 50, 30, 20, 15, 10 hPa

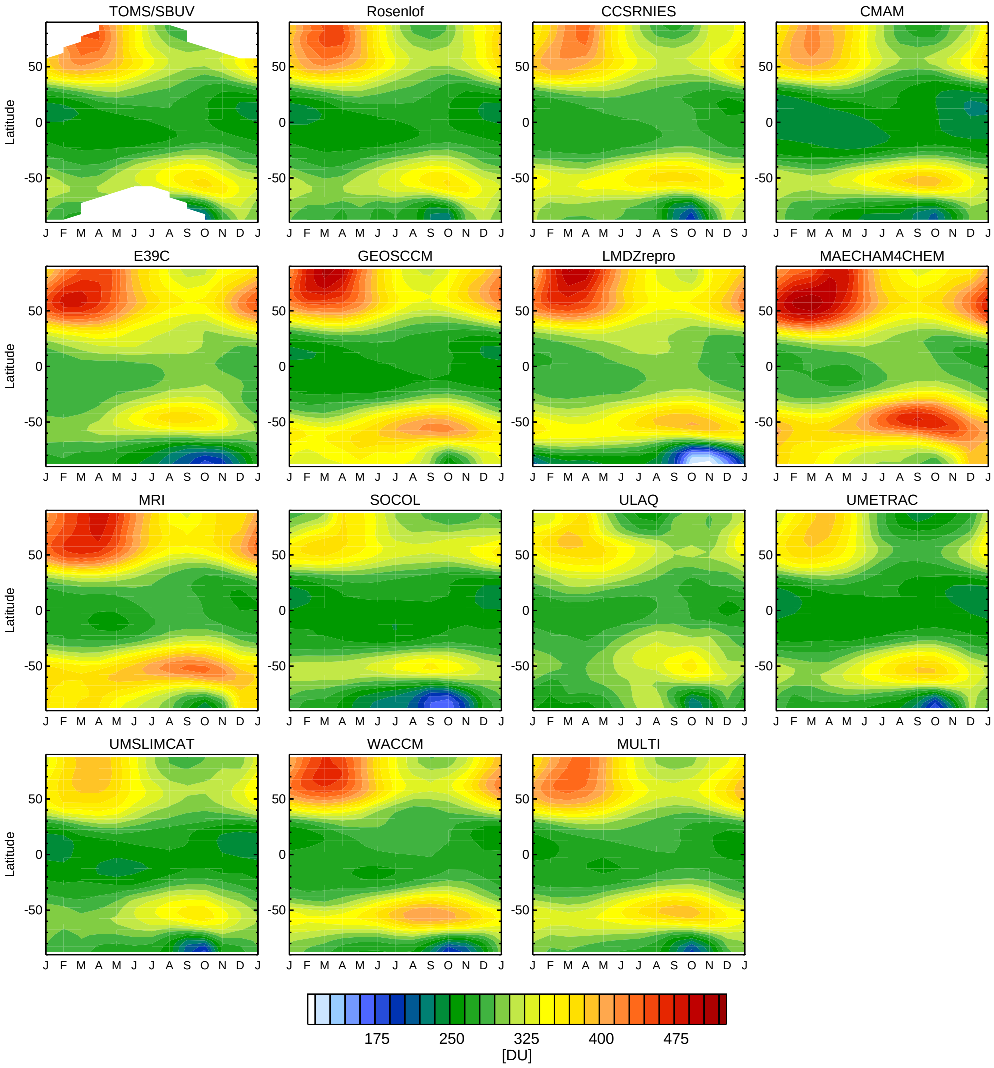

Fig. 1. Total ozone climatology (1980–1984) in observational data sets (TOMS/SBUV, Rosenlof), individual CCMVal models, and multi-model average (MULTI).

Brewer-Dobson circulation. The latter has experienced a significant change during the last two decades of the 20th century (e.g. Hu and Tung, 2002, Karpetchko and Nikulin, 2004) due to reasons which are not completely understood (Hu et al., 2005). This leaves a possibility that natural decadal variability, not related to external forcing, has considerably contributed into the NH trends. It is therefore not reasonable to expect that the models simulate the NH trends over 20 years correctly.

The simulation of realistic climate and climate trends depends not only on a correct simulation of total column ozone, but also on the vertical distribution of ozone. Therefore two additional diagnostics are considered here: (3) the monthly mean zonal mean vertical ozone distribution climatology over the period 1980–1984 and (4) the monthly mean zonal mean vertical ozone distribution trend over the period 1980–1999 at several pressure levels (see Table 2 for the list of pressure levels). As discussed in the Introduction the largest influence of stratospheric ozone changes on climate in the 21st century is expected in the SH associated with Antarctic ozone hole recovery. Therefore, to put more weight on model skill in simulating ozone over the Antarctic we average vertical ozone distributions only over the SH polar cap (60°–90° S) and weight the errors by the annually-average ozone profile, i.e. according to their contribution to the total ozone.

Total ozone from all the data sets is linearly interpolated onto a 5° latitude grid (87.5° S…87.5° N). Ozone profiles are interpolated linearly in the logarithm of pressure onto the pressure levels specified in Table 2. As a measure of the observational uncertainty σ, the standard error of the mean is used in the case of the ozone climatology, and the standard error of the slope parameter from a linear regression is used in the case of ozone trends. Measurement errors are only available for the Hassler data set. We calculate measurement errors for our diagnostics using the law of combination of errors:

$$\sigma'^2_{iml} = \sum_n \left(\frac{\partial f_l}{\partial y_{imn}}\right)^2 \sigma^2_{imn}, \qquad (5)$$

where y_{imn} and σ_{imn} are individual observations and errors at month m and grid point i and f_l is either the function for the mean in the case of the climatology, or the function for the slope parameter in the case of the trend. These measurement errors are combined with the sampling errors (i.e. with the standard error of the mean or with the standard error of the slope parameter) at each month and grid point by root mean squares and the resulting σ are used for all the three profile data sets.

4 Results

4.1 Total ozone

Figure 1 shows 5-year total ozone climatology for the period 1980-1984 from the individual models, MULTI, TOMS/SBUV, and the Rosenlof data set. This picture is similar to Fig. 14 from Eyring et al. (2006) except that they show 20-year total ozone climatology (1980–1999) and NIWA data set instead of the Rosenlof data set shown here. Also we show an updated MRI simulation. All the models simulate familiar features of the ozone distribution including the wintertime build-up in both hemispheres, and also the early stage of the Antarctic springtime ozone depletion. Figure 2 shows models and TOMS/SBUV relative errors $(x_{im} - y_{im})/\sigma_{im}$ with respect to the Rosenlof data set. Agreement between the two observational data sets is excellent, as might be expected since prior to 1985 the Rosenlof data set employs only data from TOMS and SBUV.

Figures 1–2 show that some models (MAECHAM4-CHEM, MRI) strongly overestimate total ozone globally while others (SOCOL, UMETRAC) strongly underestimate it in the extratropics. Some models (E39C, LMDZrepro) underestimate total ozone in SH mid- and high-latitudes while overestimating it elsewhere. In many models the errors typically exceed 3σ, and are therefore very unlikely to be explained by sampling variability associated with the particular 5-yr period chosen for comparison. Eyring et al. (2006) identified the causes of some model errors, like the positive biases in the extratropics in some models which is likely due to the simulated Brewer-Dobson circulation being too strong. However in most cases the causes are not straightforward to identify. MULTI tends to overestimate total ozone, especially in the SH mid-latitudes where several models show strong positive biases. In general there are no consistent biases in total ozone across the models.

20-year linear trends in total column ozone are shown in Fig. 3. The largest negative trends according to the observations are in the SH high-latitudes in November. All the models simulate a maximum negative trend in the SH high-latitudes but the time varies between September and December. Also the magnitude of the trend differs considerably between the models. The smallest simulated trend is only a half of the observed trend (E39C) while the largest trend exceeds the observed trend almost by factor 2 (MAECHAM4CHEM). Eyring et al. (2006) showed that the simulated Antarctic ozone trends are consistent with the trends in Antarctic stratospheric halogen loading. The largest trends in Cl_y are simulated by UMETRAC while the smallest trends are simulated by E39C and SOCOL. Accordingly, these models simulate too large and too small ozone trends (Eyring et al., 2006). According to WE08 assessment UMETRAC has high grade in simulating Antarctic Cl_y while E39C and SOCOL have low grade. Note that in the newer version of SOCOL used here the simulated Antarctic

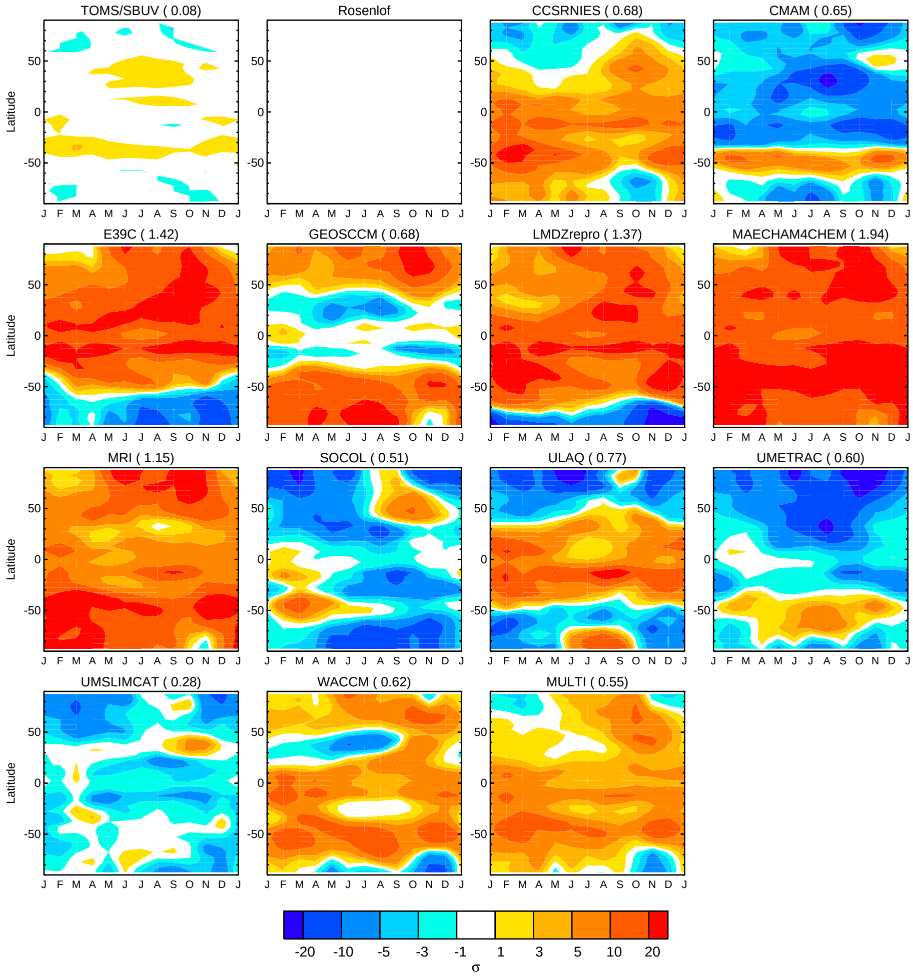

Fig. 2. Normalised errors with respect to Rosenlof data set in total ozone climatologies shown in Fig. 1. Numbers next to data set names indicate area-weighted globally averaged errors normalised by the average error across the individual models according to Eq. (2).

Cl_y values are considerably closer to the observed ones than those used by WE08 but still remain smaller than those in observations and in the majority of CCMVal-1 models (not shown).

Figure 4 shows relative errors with respect to the Rosenlof data set. The TOMS/SBUV biases are small, indicating consistency between TOMS/SBUV and the other satellites (SAGE-II, MLS, HALOE) employed in the Rosenlof data set after 1985. Models that overestimate the magnitude of the trends typically show the largest errors. As a result MULTI trends are biased negative. However the MULTI total error with respect to the Rosenlof data set (and also to TOMS/SBUV) is smaller than in any individual model. MULTI errors are everywhere within 3σ of the observed trends. The CCMVal models almost all show too much ozone depletion in the tropics, but elsewhere biases are not consistent in sign amongst the models.

To test the sensitivity of our results to the trend period we calculated the trends for the period 1980–2001 for observations and for those models for which the data are available. The observed trends for this period are typically smaller than those shown in Fig. 3; however the errors patterns do not change much and our conclusions are unaffected by these changes.

4.2 Vertical ozone distribution

The climatology of the vertical distribution of ozone is shown in Fig. 5. All three observational data sets exhibit an ozone minimum in October and a lifting of the ozone maximum layer in November–December when the polar vortex breaks up and mid-latitude air is mixed into high-latitudes. The majority of the models reproduce these features, however some models (ULAQ, UMETRAC) simulate a comparable minimum at the end of summer, a feature typically observed in the NH seasonal cycle and attributed to ozone depletion by summertime NOx chemistry (Brühl et al., 1998). Excessive ozone simulated by MAECHAM4CHEM and MRI is

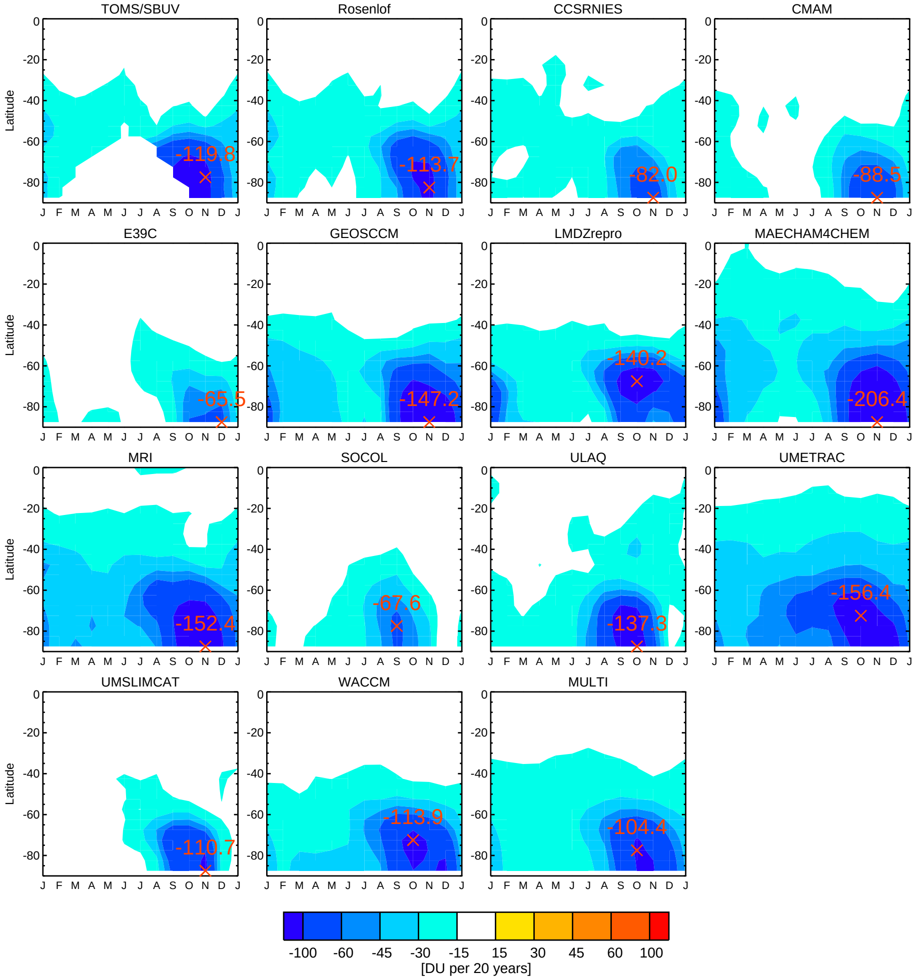

Fig. 3. 20-yr total ozone trends (1980–1999) in observational data sets (TOMS/SBUV, Rosenlof), individual CCMVal models and multi-model average (MULTI). Red numbers indicate minimum trends.

apparent throughout the year. Several other models simulate too much ozone during the winter build-up period, which may be an indication of either a too strong Brewer-Dobson circulation, or weak isolation of the lower stratospheric polar vortex from mid-latitude ozone-rich air, or both.

Figure 6 shows model errors in vertical ozone distribution climatology with respect to the Rosenlof data set. The differences between the observational data sets are striking. The Randel data set has typically larger values than the two other data sets especially in spring and summer. This maybe because the Randel data set comprises only data from Syowa station located relatively close to the polar vortex edge (Randel and Wu, 2007). The polar vortex edge region is more influenced by mixing with mid-latitude ozone-rich air while air from the vortex interior further south, impacted by chemical ozone depletion, remains more isolated. Reassuringly, both the Randel and the Hassler data set total biases are lower than those of individual models and MULTI, although the differences between the data sets often exceed 3σ. The differences between the Rosenlof and the Hassler data sets are largest in the troposphere where no satellite data is available and both data sets rely on a reconstruction to fill in the gaps. In the Rosenlof data set tropospheric ozone is obtained as a difference between total ozone and stratospheric ozone (Dall'Amico et al., 2010) while in the Hassler data set it is calculated it as a regression fit to ozonesonde data, mainly available after 1986 (Hassler et al., 2008), using equivalent effective stratospheric chlorine, QBO, solar cycle, El Nino Southern Oscillation, and stratospheric aerosol loading resulting from volcanic eruptions (Hassler et al., 2009). The large differences in the stratosphere during winter when satellite coverage of high-latitudes is limited, are also, most probably, related to the differences in the reconstruction techniques.

Almost all the models show lower values in the upper troposphere and in the lower stratosphere below 150 hPa throughout the year than the Rosenlof data set does. The differences in the troposphere with respect to the Hassler data

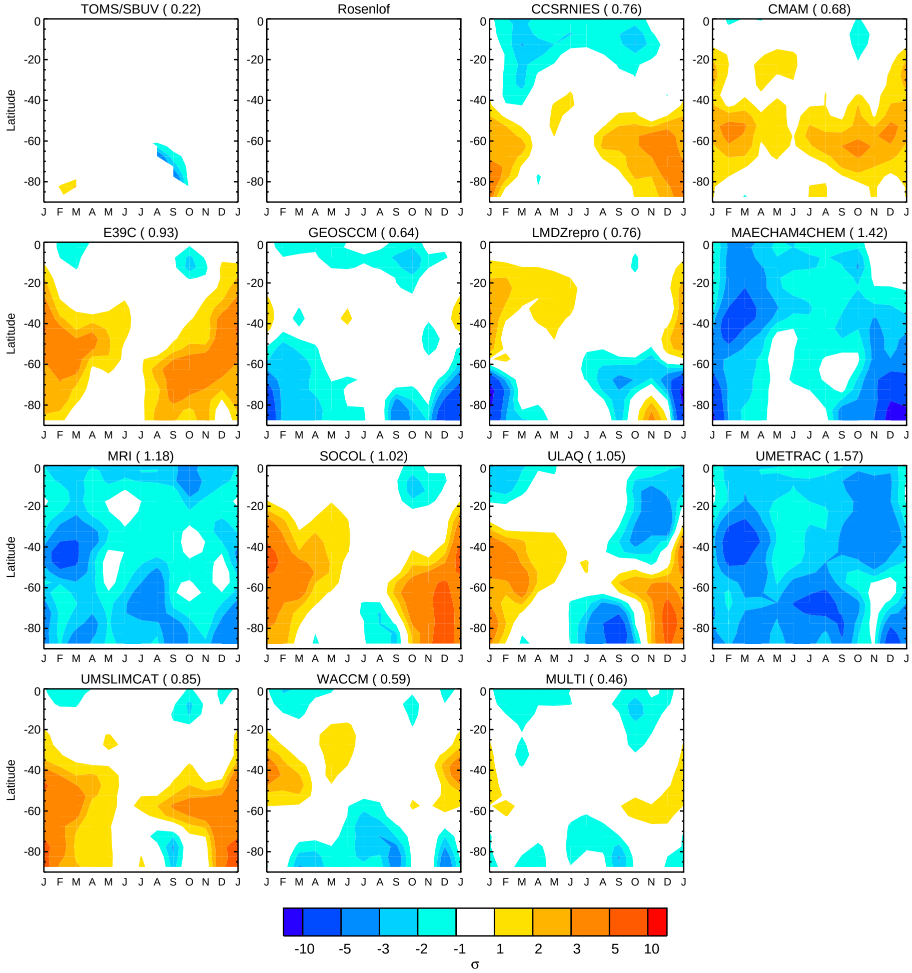

Fig. 4. Normalised errors with respect to the Rosenlof data set in total ozone trends shown in Fig. 3. Numbers next to data set names indicate area-weighted hemisphere-averaged errors normalised by the average error across the individual models according to Eq. (2).

set are smaller (Fig. 7) and mainly restricted to 200–300 hPa, with model values below this typically being higher than in the Hassler data set. In some models (E39C, ULAQ, UMETRAC) the lower values near the tropopause arise because of a too high ozonopause. Above 100 hPa models typically simulate higher ozone values than observed, particularly during the winter build-up period and above 50 hPa during summer, presumably due to a more vigorous exchange with mid-latitudes. Model errors with respect to both observational data sets typically exceed 3σ. MULTI shows the lowest total error among the models with respect to the Rosenlof data set but not with respect to the Hassler data set.

Vertical ozone distribution trends are shown in Fig. 8. The trends in the observational data sets differ considerably from each other and the differences between them are comparable to the differences between the observations and the models. The maximum negative trend in the Rosenlof data set is only 60% of that in the Hassler data set and lags it by two months. The differences between the time series arise largely after 1990 and are therefore attributable to the different data sources rather than to the methods used to construct the data sets.

Figures 9–10 show trend errors with respect to the Rosenlof and Hassler data sets correspondingly. The majority of the models underestimate the springtime depletion compared with the Hassler data set but not with the Rosenlof data set. Several models simulate too strong ozone depletion below 100 hPa in summer comparing with the three observation data sets. In some models (LMDZrepro, WACCM) this may be a result of a delayed polar vortex break up (Eyring et al., 2006). MULTI and several individual models show better agreement with the Rosenlof data set (and also with the Hassler data set) than the two other observation data sets.

Comparing Figs. 2, 6, 7 with Figs. 4, 9, 10 one can see that the trend errors are typically smaller than the climatology errors. However, since errors are normalised by the uncertainty σ, we caution against the interpretation that the models simulate trends better than the climatology.

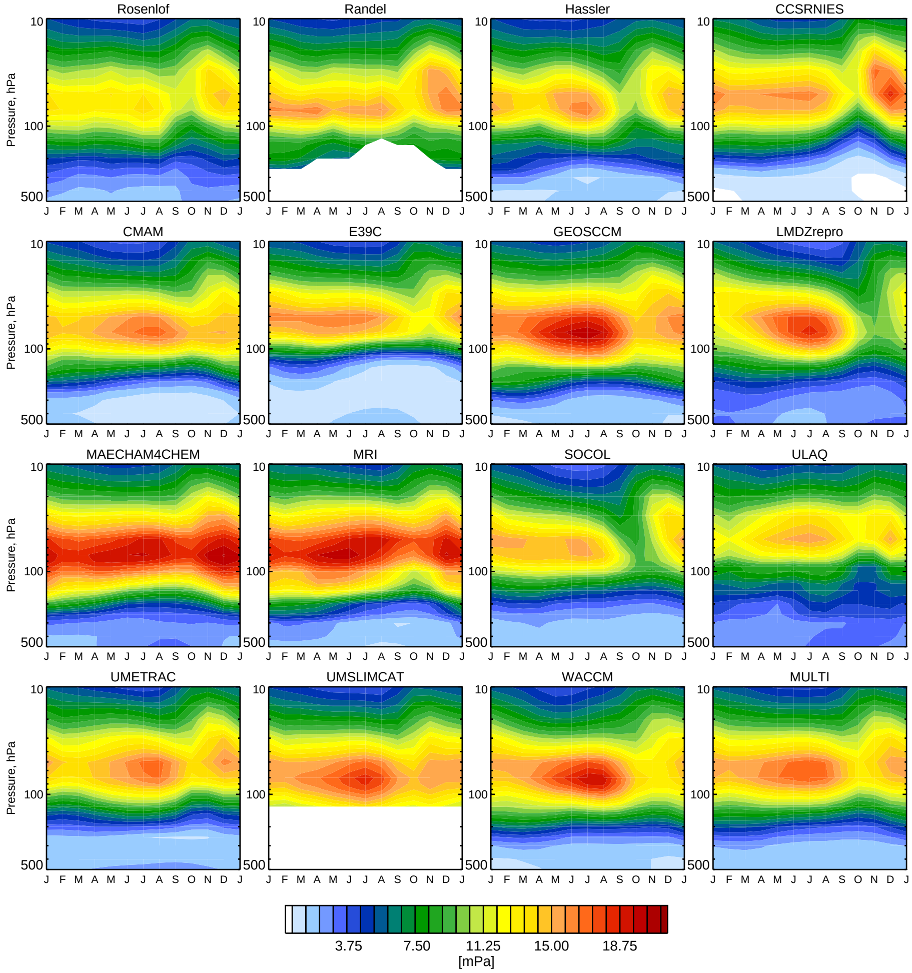

Fig. 5. Vertical ozone distribution climatology (1980–1984) in observational data sets (Rosenlof, Randel, Hassler), individual CCMVal models, and multi-model average (MULTI).

To investigate whether the choice of reference data set has a large impact on the model ranking we calculate Spearman's rank correlation coefficient between model ranks in the same diagnostic but calculated with respect to different observation data sets. There is a high correlation between the model ranks in ozone profile trends with respect to the Rosenlof and Randel data sets ($r = 0.93$) however the model ranks with respect to the Hassler data set are poorly correlated with those with respect to either the Rosenlof ($r = 0.41$) or the Randel ($r = 0.56$) data sets. Similar tests performed for the model ranks in the ozone profile climatology showed a high correlation between the ranks with respect to the Rosenlof and Hassler data sets ($r = 0.93$) but lower correlations between the model ranks with respect to the Randel data set and either the Rosenlof ($r = 0.60$) or the Hassler ($r = 0.54$) data sets. Model ranks in the total ozone climatology and trends were very similar with respect to both datasets ($r > 0.9$).

4.3 Performance index and its uncertainty

Performance indices calculated using Eq. (4) are shown in Fig. 11 together with errors for the individual diagnostics. Comparing individual models shows that no one model performs better than the others in all diagnostics however some models have generally low errors while other have generally high errors. MULTI does not perform better than the individual models in all diagnostics however its combined error is the lowest.

To make sure that our diagnostics do not duplicate each other we have correlated model ranks obtained in different diagnostics. The correlation between model ranks in total ozone climatology and ozone profile climatology is found to be the only one significant at the 5% level ($r = 0.7$), with the other correlation coefficients being less than 0.52. Therefore our diagnostics provide complementary information on different aspects of model performance.

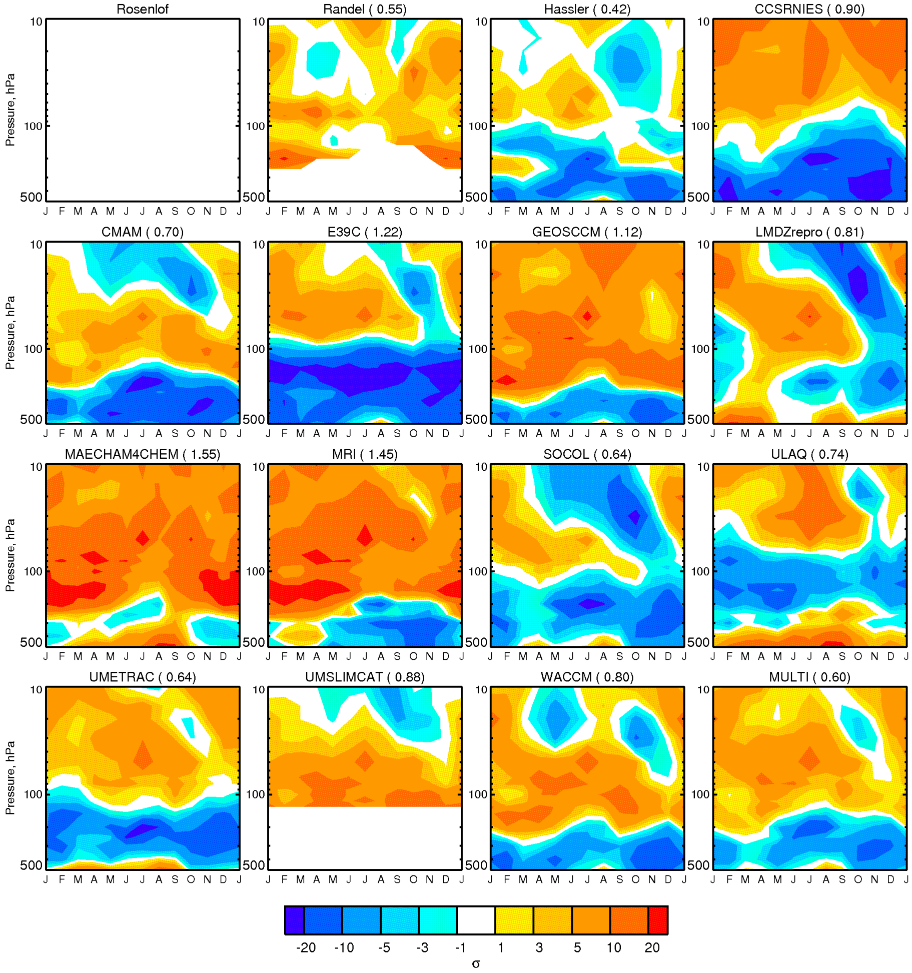

Fig. 6. Normalised errors with respect to the Rosenlof data set in vertical ozone distribution climatology shown in Fig. 5. Numbers next to data set names indicate domain-averaged errors normalised by the average error across the individual models according to Eq. (2).

To assess the robustness of our model ranking we perform several sensitivity tests. The sensitivity of the ranking to the choice of the reference data sets is assessed by calculating the performance indices using the individual reference data sets separately. Since for the Randel and Hassler data sets we do not have corresponding total ozone timeseries, these data sets are used in combination with the TOMS/SBUV data set. To study sensitivity of the results to sampling errors we use additional runs available for SOCOL, MRI, and WACCM. The calculations were repeated for two additional simulations for each of these models. Also we apply small modifications to the original diagnostics, which include restricting the domain to above 200 hPa, weighting the ozone profile errors according to the mass or geometric thickness of the corresponding layer, or calculating the total ozone climatology diagnostic over the SH only. The performance indices calculated in these sensitivity tests are shown in Fig. 11 and provide an estimate of the ranking uncertainty. The changes in the performance indices in these experiments are up to 15% (about 0.1 in absolute units), suggesting that smaller differences in performance index between models may be insignificant. In terms of ranking, these changes resulted in models ranking changes by 0–2 positions. In all the tests MULTI gets the highest rank.

We also test the sensitivity of the ranking to the choice of model performance metric. Here, instead of applying Eqs. (1)–(4) we use an index similar to the one used in WE08:

$$g_{jkl} = 1 - \frac{1}{nW} \sum_i \sum_m \left(w_{im} \left| x_{imjl} - y_{imkl} \right| / \sigma_{imkl} \right), \tag{6}$$

where n is a scaling factor, and g_{jkl} is the grade of j-th model in l-th diagnostic with respect to k-th reference data set. Negative values of g, wherever they are obtained, are set to zero. The maximum possible grade is 1. A zero grade means that the differences between the model and the observation in average exceed $n\sigma$. The model grade for an individual diagnostic g_{jl} is calculated as an average over model grades with respect to all reference data sets; and the overall model grade

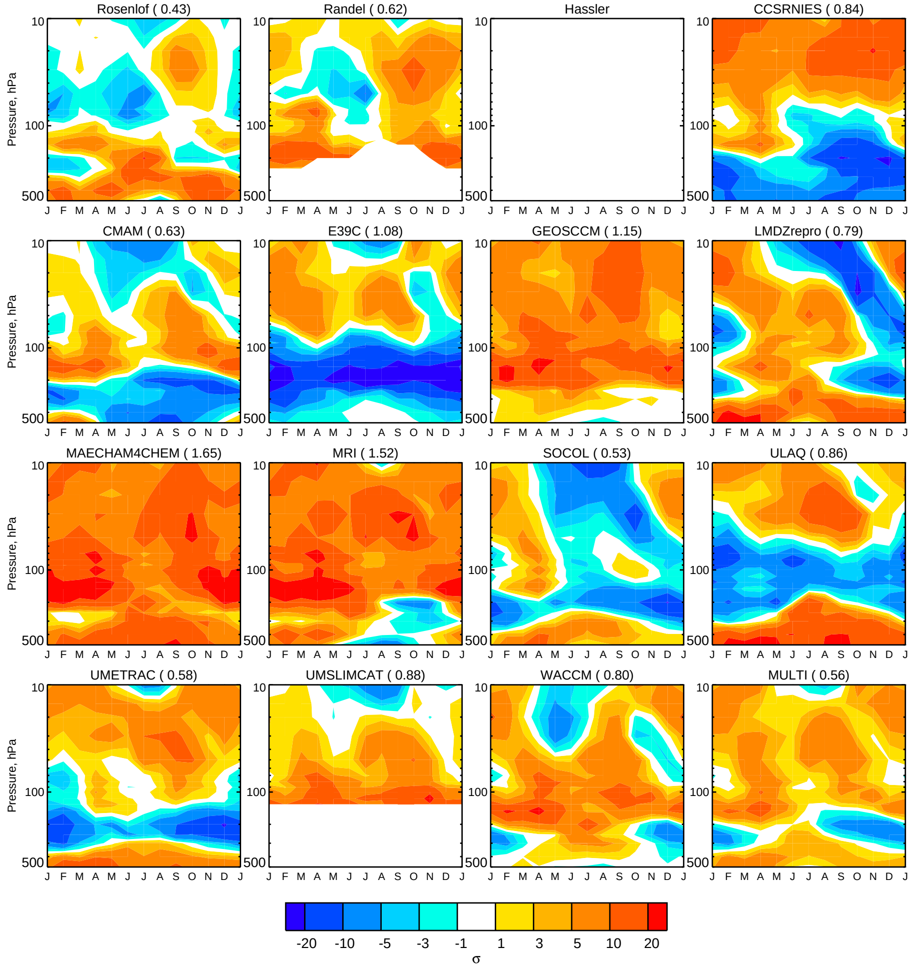

Fig. 7. The same as in Fig. 6 but with respect to the Hassler data set.

g_j is calculated by averaging over the grades for the individual diagnostics.

The model grading calculated using Eq. (6) depends on the choice of the scaling factor n. If we choose n =3, as in WE08, then almost all the grades in the both climatology diagnostics are zero, the only exception being the total ozone climatology diagnostic in UMSLIMCAT. As a result, the combined model grades are determined by the trend diagnostics only. Nevertheless, Spearman's correlation coefficient between the model ranks calculated using Eq. (6) and the original ranks is high ($r=0.86$), although some individual models change their ranks by 3–4 positions. Increasing n to 5 results in a small improvement of the correlation coefficient ($r=0.89$) but further increases lead to converging of the grades in the trend diagnostics towards 1, thus decreasing the differences between the model grades in the trend diagnostics. As a result, the overall model grades are largely determined by the climatology diagnostics. MULTI typically gets the highest rank, except when n varies between 4 and 6. In these cases, MULTI gets the second rank, and UMSLIMCAT gets the highest rank. This is because UMSLIMCAT gets high grade in the total ozone climatology while the majority of the models get zero grades in this diagnostic. Despite these discrepancies the ranking obtained using Eq. (6) is in a reasonable agreement with the ranking obtained using Eqs. (1)–(4), with correlation coefficients exceeding 0.8.

4.4 Comparison with previous studies

We will now see how our ranking agrees with results by Eyring et al. (2006) and WE08. For this comparison we use the same simulations as those used by them. Eyring et al. (2006) highlighted 6 out of 13 models which agree better with the observations based on analysis of several transport, temperature and chemistry diagnostics. Five individual models with the highest performance index according to our original analysis are among the six models highlighted by Eyring et al. (2006) while the four models with the lowest performance index are among the seven non-highlighted models. Taking into account that margins between the models with

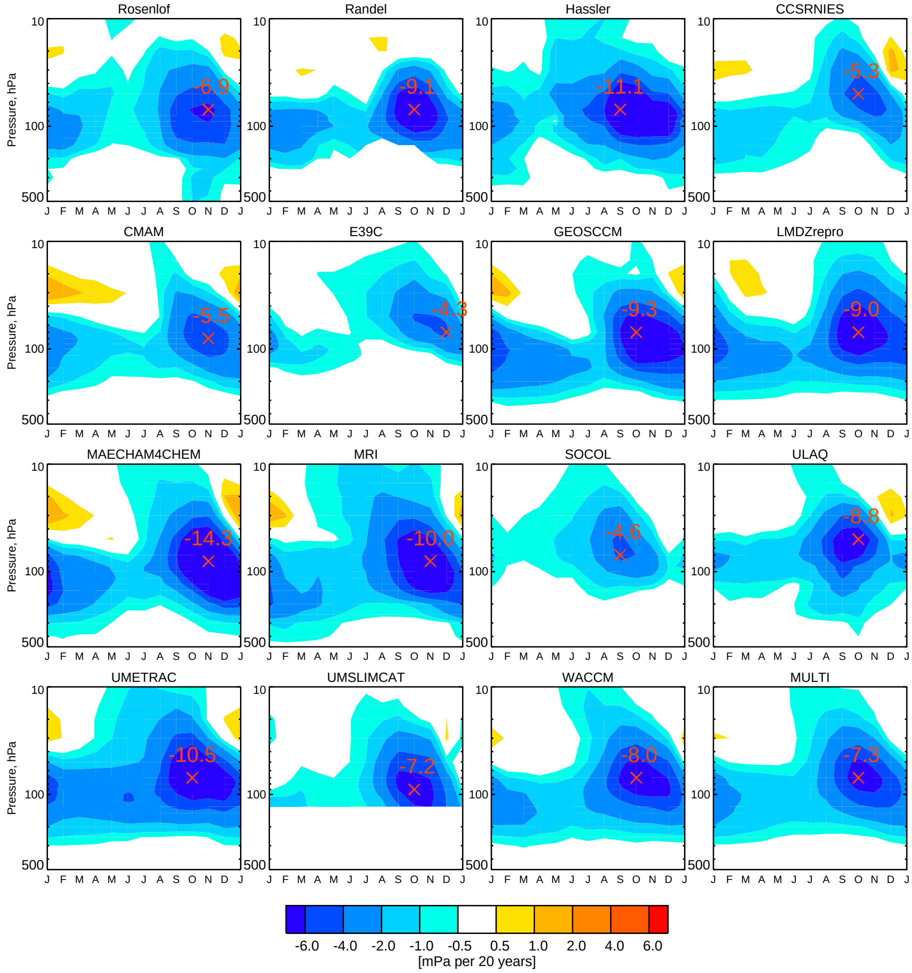

Fig. 8. 20-yr vertical ozone distribution trends (19808–1999) in observational data sets (Rosenlof, Randel, Hassler), individual CCMVal models, and multi-model average (MULTI). Red numbers indicate minimum trends.

intermediate performance index are small compared to the overall spread of the errors we conclude that our ranking is in a reasonable agreement with results by Eyring et al. (2006).

WE08 evaluated a subset of key processes important for stratospheric ozone, with the focus mainly on diagnostics to evaluate transport and dynamics in the CCMs. Spearman's rank correlation coefficient between our model ranking and theirs is 0.59. WE08 also provide model grades based separately on transport diagnostics or polar dynamic diagnostics. The agreement with our ranking gets worse if we consider either grades based on transport diagnostics ($r = 0.44$) or grades based on polar dynamic diagnostics ($r = 0.54$). In the case of polar dynamics diagnostics the most considerable difference is that WACCM which has the highest performance index among the individual models in our analysis gets low grade in the polar dynamics diagnostics because of low grade in the zonal wind diagnostic (Eyring et al., 2006; WE08). The differences between our grades may be explained by, first, taking into account that some important diagnostics (e.g. those related to polar chemistry) may be not considered in their study, and second, that those diagnostics that were considered may need to be given different weights depending on their importance for polar ozone.

WE08 found that, in many diagnostics, MULTI does not get better grades than the best individual models. This is because significant biases in these diagnostics are shared by many but not all the models. We also found that MULTI does not have the smallest error in all the diagnostics. However the overall performance of MULTI according to the original test (Fig. 11) is considerably better than that of any individual model. This result holds also if available alternative model realisations are used, and is also robust to small modifications in the original diagnostics (see Sect. 4.3). Therefore, assuming that the ability of the models to simulate observed ozone climatology and trends is a reliable indicator of their ability to simulate future ozone we conclude that the multi-model average of ozone projections appears to be the best choice as a future ozone scenario for usage in climate model

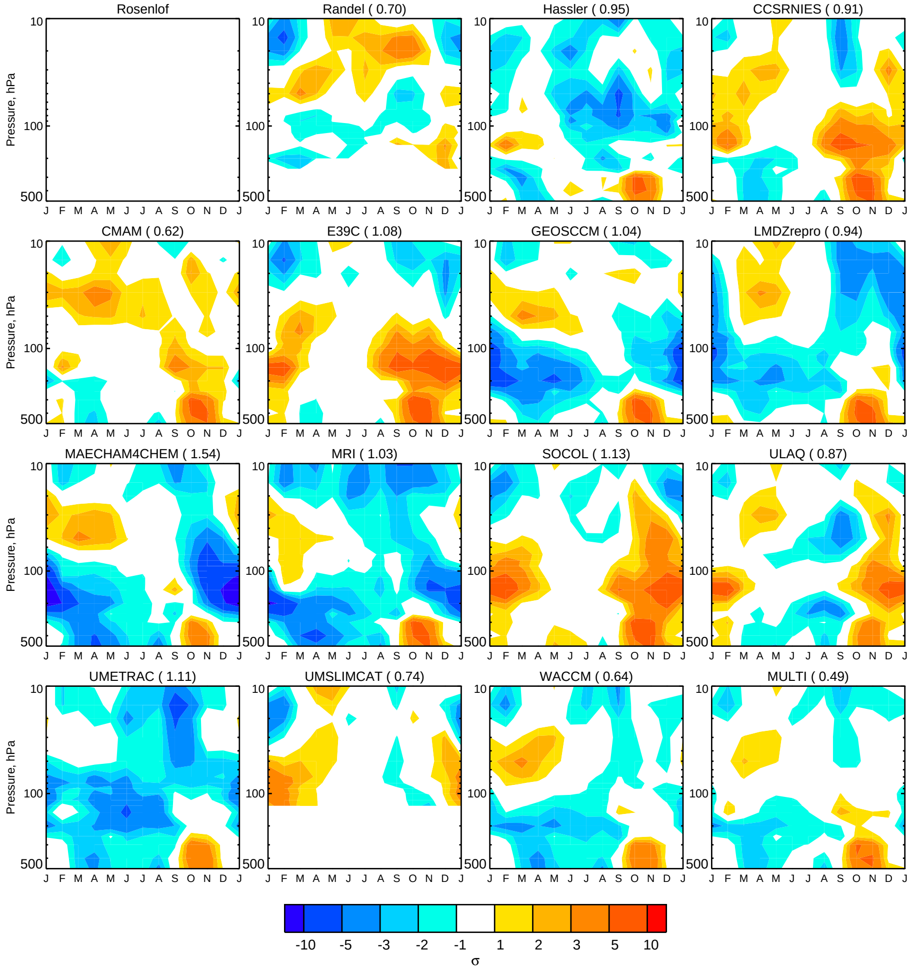

Fig. 9. Normalised errors with respect to the Rosenlof data set in vertical ozone distribution trends shown in Fig. 8. Numbers next to data set names indicate domain-averaged errors normalised by the average error across the individual models according to Eq (2).

simulations that require prescribed ozone fields. WE08 also noticed that weighting the model results according to their performance does not change significantly the multi-model average projection.

5 Conclusions

The goal of this study is to provide the climate modelling community with some recommendations regarding the choice of future ozone scenario for implementation in climate simulations. We have validated the abilities of twelve CCMs to simulate the observed total ozone climatology and trends and also the Antarctic ozone profile climatology and trends and ranked the models according to their errors averaged across four chosen diagnostics. No one model performs better than the others in all four diagnostics; however combining errors in individual diagnostics into one metric of model performance allowed us to objectively rank the models. The highest rank is obtained by the multi-model ensemble average. Sensitivity tests performed to assess the robustness of the ranking showed that the individual models may change their rank by several positions but that the multi-model ensemble average gets the highest rank in the majority of the experiments. Therefore we argue that the multi-model averaged projection, which is less sensitive to individual model biases, provides the best estimate of future ozone.

The model ranks obtained are further compared with those made earlier by Eyring et al. (2006) and WE08. A rather good agreement is found with the results of Eyring et al. (2006) who only separated the models into two groups, with models in one group being generally in a better agreement with the observations than models in the other group. However, comparison with the other model ranking based on evaluation of dynamics and transport in the models showed only modest correlation, probably because some processes important for polar ozone, which is given large weight in our

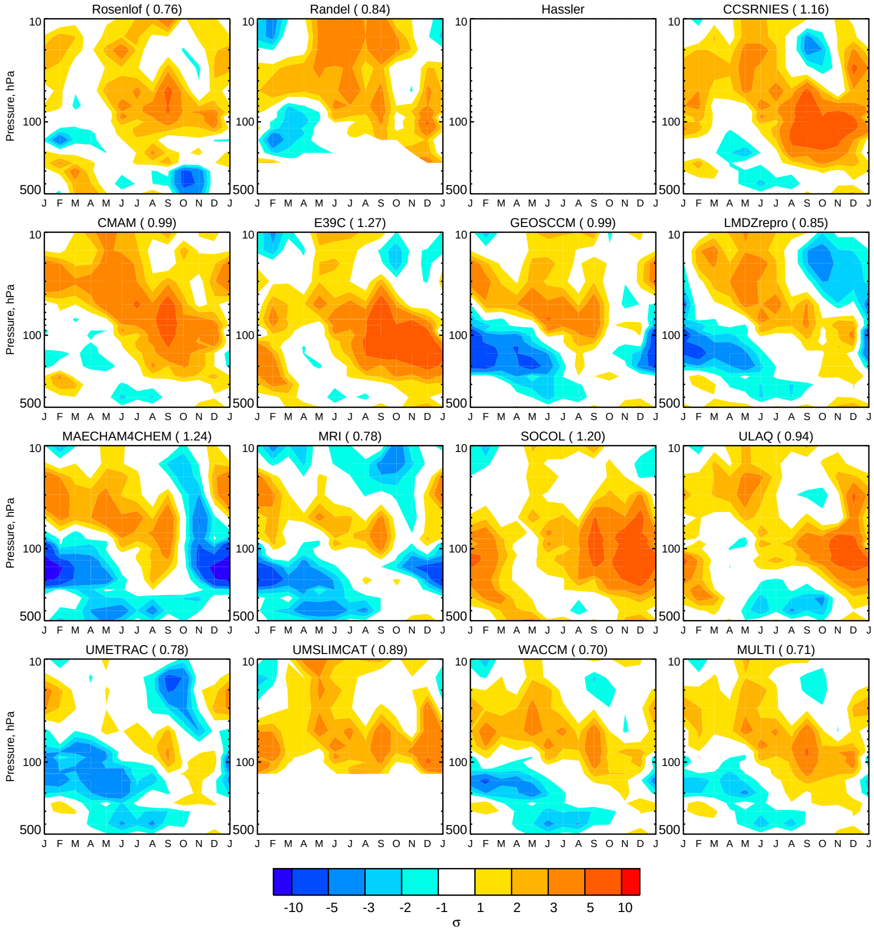

Fig. 10. The same as in Fig. 9 but with respect to the Hassler data set.

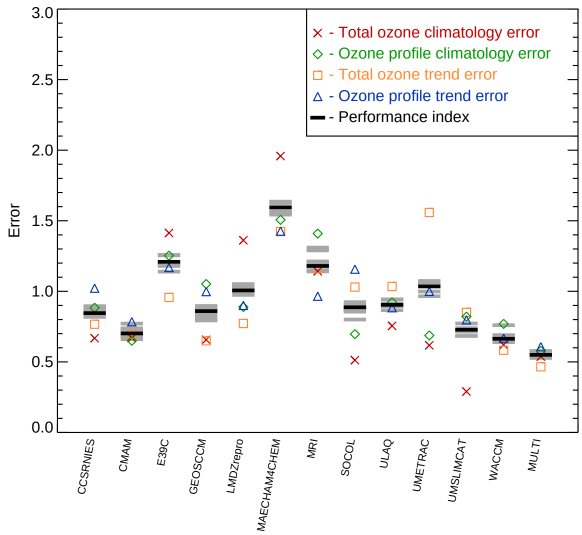

Fig. 11. Model performance index and errors in individual diagnostics. Performance indices from sensitivity tests are marked by grey bars.

study, were not considered by WE08 or because those diagnostics that were considered may need to be given different weights depending on their importance for polar ozone. This comparison shows that the selection of diagnostics for process-oriented CCM validation remains a challenging task and depends on the region of interest.

In order to reduce the influence of possible biases in ozone observations more than one up-to-date observational data set has been used in this evaluation and the true model error is estimated as an average across errors with respect to individual observational data sets. While diagnostics based on total ozone are found to be insensitive to choice of observational data set, partly because they share the same data sources, the evaluation of simulated Antarctic ozone profiles provides significantly different results depending on which data set is used as a reference. In the ozone profile trend diagnostic differences between the observations are found to be comparable to or even to exceed model errors. In particular, the maximum negative trend in the Rosenlof data set in the Antarctic lower stratosphere during spring is only 60% of that in the Hassler data set; presumably due to different data sources employed in the data sets. This result stresses the need for the compilation of a unified reliable vertically resolved ozone reference data set.

Our assessment has at least two practical applications. First, modellers can make a choice of future ozone scenario based on the quantitative evaluation. Second, ozone simulations by future model generations can be validated in the same way as done here and model improvements can be quantitatively assessed.

Acknowledgements. This work is supported by NERC Project NE/E006787/1. We acknowledge the modeling groups for making their simulations available for this analysis, the Chemistry-Climate Model Validation Activity (CCMVal) for WCRP's (World Climate Research Programme) SPARC (Stratospheric Processes and their Role in Climate) project for organizing and coordinating the model data analysis activity, and the British Atmospheric Data Center (BADC) for collecting and archiving the CCMVal model output. We thank TOMS/SBUV team and W. Randel for making their data sets publicly available and also V. Eyring, M. Dameris, J. Scinocca, T. Reichler, and an anonymous reviewer for useful comments.

Edited by: L. Carpenter

References

Akiyoshi, H., Sugita, T., Kanzawa, H., and Kawamoto, N.: Ozone perturbations in the Arctic summer lower stratosphere as a reflection of NO_x chemistry and planetary scale wave activity, J. Geophys. Res., 109, D03304, doi:10.1029/2003JD003632, 2004.

Austin, J., Wilson, R. J., Li, F., and Vomel, H.: Evolution of water vapor concentrations and stratospheric age of air in coupled chemistry-climate model simulations, J. Atmos. Sci., 64, 905–921, 2006.

Bracegirdle, T. J., Connolley, W. M., and Turner J.: Antarctic climate change over the twenty first century, J. Geophys. Res., 113, D03103, doi:10.1029/2007JD008933, 2008.

Brühl, C., Crutzen, P. J., and Grooß, J. U.: High-latitude, summertime NOx activation and seasonal ozone decline in the lower stratosphere: Model calculations based on observations by HALOE on UARS, J. Geophys. Res., 103, 3587–3597, 1998.

Cai, W., Shi, G., and Li, Y.: Multidecadal fluctuations of winter rainfall over southwest Western Australia simulated in the CSIRO Mark 3 coupled model. Geophys. Res. Lett., 32, L12701, doi:10.1029/2005GL022712, 2005.

Connolley, W. M. and Bracegirdle, T. J.: An Antarctic assessment of IPCC AR4 climate models, Geophys. Res. Lett., 34, L22505, doi:10.1029/2007GL031648, 2007.

Dall'Amico, M., Gray, L. J., Rosenlof, K. H., Scaife, A. A., Shine, K. P., and Stott, P. A.: Stratospheric temperature trends: impact of ozone variability and the QBO, Clim. Dynam., 34(2–3), 381–398, doi:10.1007/s00382-009-0604-x, 2010.

Dameris, M., Grewe, V., Ponater, M., Deckert, R., Eyring, V., Mager, F., Matthes, S., Schnadt, C., Stenke, A., Steil, B., Brühl, C., and Giorgetta, M.: Long-term changes and variability in a transient simulation with a chemistry-climate model employing realistic forcings, Atmos. Chem. Phys., 5, 2121–2145, 2005, http://www.atmos-chem-phys.net/5/2121/2005/.

Egorova, T., Rozanov, E., Zubov, V., Manzini, E., Schmutz, W., and Peter, T.: Chemistry-climate model SOCOL: a validation of the present-day climatology, Atmos. Chem. Phys., 5, 1557–1576, 2005, http://www.atmos-chem-phys.net/5/1557/2005/.

Eyring, V., Butchart, N., Waugh, D. W., et al.: 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, 2006.

Eyring, V., Waugh, D. W., Bodeker, G. E., et al.: Multimodel projections of stratospheric ozone in the 21st century, J. Geophys. Res., 112, D16303, doi:10.1029/2006JD008332, 2007.

Fomichev, V. I., Jonsson, A. I., de Grandpr'e, J., et al.: Response of the middle atmosphere to CO2 doubling: Results from the Canadian Middle Atmosphere Model, J. Climate, 20, 1121–1144, 2007.

Fyfe, J. C., Boer, G., and Flato, G.: The Arctic and Antarctic Oscillations and their projected changes under global warming, Geophys. Res. Lett., 26, 1601–1604, 1999.

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

Gillett, N. P. and Thompson, D. W. J.: Simulation of recent Southern Hemisphere climate change, Science, 302, 273–275, 2003.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res., 113, D06104, doi:10.1029/2007JD008972, 2008.

Hassler, B., Bodeker, G. E., and Dameris, M.: Technical Note: A new global database of trace gases and aerosols from multiple sources of high vertical resolution measurements, Atmos. Chem. Phys., 8, 5403–5421, 2008

Hassler, B., Bodeker, G. E., Cionni, I., and Dameris, M.: A vertically resolved, monthly mean, ozone database from 1979 to 2100 for constraining global climate model simulations, International Journal of Remote Sensing, Int. J. Remote Sens., 30(15), 4009-4018, doi:10.1080/01431160902821874, 2009.

Hu, Y. and Tung, K. K.: Interannual and decadal variations of planetary wave activity, stratospheric cooling, and Northern Hemisphere annular mode, J. Climate, 15**,** 1659–1673, 2002.

Hu, Y., Tung, K. K., and Liu, J.: A Closer Comparison of Early and Late-Winter Atmospheric Trends in the Northern Hemisphere, J. Climate, 18, 3204–3216, 2005.

Jourdain, L., Bekki, S., Lott, F., and Lefèvre, F.: The coupled chemistry-climate model LMDz-REPROBUS: description and evaluation of a transient simulation of the period 1980–1999, Ann. Geophys., 26, 1391–1413, 2008, http://www.ann-geophys.net/26/1391/2008/.

Karpetchko, A. and Nikulin, G.: Influence of early winter upward wave activity flux on midwinter circulation in the stratosphere and troposphere, J. Climate, 17, 4443–4452, 2004.

Marshall, G. J., Orr, A., van Lipzig, N. P. M., and King, J. C.: The Impact of a Changing Southern Hemisphere Annular Mode on Antarctic Peninsula Summer Temperatures, J. Climate, 19**,** 5388–5404, 2006.

Miller, R. L., Schmidt, G. A., and Shindell, D. T.: Forced annular variations in the 20th century Intergovernmental Panel on Climate Change Fourth Assessment Report models, J. Geophys. Res., 111, D18101, doi:10.1029/2005JD006323, 2006.

Pawson, S., Stolarski, R. S., Douglass, A. R., Newman, P. A., Nielsen, J. E., Frith, S. M., and Gupta, M. L.: Goddard Earth Observing System Chemistry-ClimateModel Simulations of Stratospheric Ozone-Temperature Coupling Between 1950 and 2005, J. Geophys. Res., 113, D12103, doi:10.1029/2007JD009511, 2008.

Pitari, G., Mancini, E., Rizi, V., and Shindell, D.: Feedback of future climate and sulfur emission changes an stratospheric aerosols and ozone, J. Atmos. Sci., 59, 414–440, 2002.

Perlwitz, J., Pawson, S., Fogt, R. L., Nielsen, J. E., and Neff, W. D.: Impact of stratospheric ozone hole recovery on Antarctic climate, Geophys. Res. Lett., 35, L08714, doi:10.1029/2008GL033317, 2008.

Randel, W. J. and Wu F.: Cooling of the Arctic and Antarctic polar stratospheres due to ozone depletion, J. Climate, 12, 1467–1479, 1999.

Randel, W. J. and Wu F.: A stratospheric ozone profile data set for 1979–2005: Variability, trends, and comparisons with column ozone data, J. Geophys. Res., 112, D06313, doi:10.1029/2006JD007339, 2007.

Reichler, T. and Kim, J.: How well do coupled models simulate today's climate?, B. Am. Meteorol. Soc., 89, 303–311, 2008.

Schraner, M., Rozanov, E., Schnadt Poberaj, C., Kenzelmann, P., Fischer, A. M., Zubov, V., Luo, B. P., Hoyle, C. R., Egorova, T., Fueglistaler, S., Bronnimann, S., Schmutz, W., and Peter, T.: Technical Note: Chemistry-climate model SOCOL: version 2.0 with improved transport and chemistry/microphysics schemes, Atmos. Chem. Phys., 8, 5957–5974, 2008, http://www.atmos-chem-phys.net/8/5957/2008/.

Shibata, K. and Deushi, M.: Partitioning between resolved wave forcing and unresolved gravity wave forcing to the quasi-biennial oscillation as revealed with a coupled chemistry-climate model, Geophys. Res. Lett., L12820, doi:10.1029/2005GL022885, 2005.

Son, S.-W., Polvani, L. M., Waugh, D. W., Akiyoshi, H., Garcia, R., Kinnison, D., Pawson, S., Rozanov, E., Shepherd, T. G., and Shibata, K.: The Impact of Stratospheric Ozone Recovery on the Southern Hemisphere Westerly Jet, Science, 320, 1486–1489, 2008.

Son, S.-W., Tandon, N. F., Polvani, L. M., and Waugh, D. W.: Ozone hole and Southern Hemisphere climate change, Geophys. Res. Lett., 36, L15705, doi:10.1029/2009GL038671, 2009.

Steil, B., Brühl, C., Manzini, E., Crutzen, P. J., Lelieveld, J., Rasch, P. J., Roeckner, E., and Krüger, K.: A new interactive chemistry climate model, 1: Present day climatology and interannual variability of the middle atmosphere using the model and 9 years of HALOE/UARS data, J. Geophys. Res., 108, 4290, doi:10.1029/2002JD002971, 2003.

Stolarski, R. S. and Frith, S.: Search for evidence of trend slowdown in the long-term TOMS/SBUV total ozone data record: The importance of instrument drift uncertainty, Atmos. Chem. Phys., 6, 4057–4065, 2006, http://www.atmos-chem-phys.net/6/4057/2006/.

Thompson, D. W. J. and Solomon, S., Interpretation of recent Southern Hemisphere climate change, Science, 296, 895–899, 2002.

Tian, W. and Chipperfield, M. P.: A new coupled chemistry-climate model for the stratosphere: The importance of coupling for future O3-climate predictions, Q. J. Roy. Meteor. Soc., 131, 281–304, 2005.

Waugh, D. W. and Eyring, V.: Quantitative performance metrics for stratospheric-resolving chemistry-climate models, Atmos. Chem. Phys., 8, 5699–5713, 2008, http://www.atmos-chem-phys.net/8/5699/2008/.

World Meteorological Organization (WMO)/United Nations Environment Programme (UNEP): Scientific Assessment of Ozone Depletion: 2006, World Meteorological Organization, Global Ozone Research and Monitoring Project, Report No. 50, Geneva, Switzerland, 2007.

Zhou, S. T., Gelman, M. E., Miller, A. J., and McCormack J. P.: An inter-hemisphere comparison of the persistent stratospheric polar vortex, Geophys. Res. Lett., 27, 1123–1126, 2000.