Atmos. Chem. Phys., 8, 1–15, 2008
www.atmos-chem-phys.net/8/1/2008/

Atmospheric Chemistry and Physics

Quantitative performance metrics for stratospheric-resolving chemistry-climate models

D. W. Waugh[1] **and V. Eyring**[2]

[1]Department of Earth and Planetary Science, Johns Hopkins University, Baltimore, MD, USA
[2]Deutsches Zentrum für Luft- und Raumfahrt, Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany

Received: 11 April 2008 – Published in Atmos. Chem. Phys. Discuss.: 6 June 2008
Revised: 27 August 2008 – Accepted: 27 August 2008 – Published: 16 September 2008

Abstract. A set of performance metrics is applied to stratospheric-resolving chemistry-climate models (CCMs) to quantify their ability to reproduce key processes relevant for stratospheric ozone. The same metrics are used to assign a quantitative measure of performance ("grade") to each model-observations comparison shown in Eyring et al. (2006). A wide range of grades is obtained, both for different diagnostics applied to a single model and for the same diagnostic applied to different models, highlighting the wide range in ability of the CCMs to simulate key processes in the stratosphere. No model scores high or low on all tests, but differences in the performance of models can be seen, especially for processes that are mainly determined by transport where several models get low grades on multiple tests. The grades are used to assign relative weights to the CCM projections of 21st century total ozone. For the diagnostics used here there are generally only small differences between weighted and unweighted multi-model mean and variances of total ozone projections. This study raises several issues with the grading and weighting of CCMs that need further examination. However, it does provide a framework and benchmarks that will enable quantification of model improvements and assignment of relative weights to the model projections.

1 Introduction

There is considerable interest in how stratospheric ozone will evolve through the 21st century, and in particular how ozone will recover as the atmospheric abundance of halogens continues to decrease. This ozone recovery is likely to be influenced by changes in climate, and to correctly simulate the evolution of stratospheric ozone it is necessary to use models that include coupling between chemistry and climate processes. Many such Chemistry-Climate Models (CCMs) have been developed, and simulations using these models played an important role in the latest international assessment of stratospheric ozone (WMO, 2007).

Given the importance of CCM simulations there is a need for process-oriented evaluation of the CCMs. A set of core processes relevant for stratospheric ozone, with each process associated with one or more model diagnostics and with relevant datasets that can be used for validation, has been defined by the Chemistry-Climate Model Validation Activity (CCMVal) for WCRP's (World Climate Research Programme) SPARC (Stratospheric Processes and their Role in Climate) project (Eyring et al., 2005). Previous studies have performed observationally-based evaluation of CCMs (e.g., Austin et al., 2003; Eyring et al., 2006) using a subset of these key processes. However, although these studies compared simulated and observed fields they did not assign quantitative metrics of performance ("grades") to these observationally-based diagnostic tests.

Assigning grades to a range of diagnostics has several potential benefits that will also improve the input of the CCM community to international assessments. For example, it will

- Allow easy visualization of the model's performance for multiple aspects of the simulations.
- Allow, in the case of a systematic bias for all models, identification of missing or falsely modeled processes.
- Enable a quantitative assessment of model improvements, both for different versions of individual CCMs and for different generations of community-wide collections of models used in international assessments.
- Make it possible to explore the value of weighting the predictions by models based on their abilities to reproduce key processes, and to form a best estimate plus uncertainties that takes into account these differing abilities.

Correspondence to: D. W. Waugh
(waugh@jhu.edu)

In this paper we perform a quantitative evaluation of the ability of CCMs to reproduce key processes for stratospheric ozone. Our starting point is the recent study by Eyring et al. (2006) (hereinafter "E06") who evaluated processes important for stratospheric ozone in thirteen CCMs. We consider the same CCMs, diagnostics, and observational datasets shown in E06. This has the advantage that the model simulations, diagnostics, and graphical comparisons between models and observations have already been presented and don't need to be repeated here. We focus on the diagnostics of processes shown in E06 rather than diagnostics of past and present ozone, as such diagnostics might be a better predictor of a models ability to make reliable projections. We further simplify our approach by using the same metric to quantify model-observations differences for all diagnostics. This quantification of each CCM's ability to reproduce key observations and processes is then used to weight ozone projections for the 21st century from the same CCMs, which were analyzed in Eyring et al. (2007) (hereinafter "E07").

Several previous studies have performed similar quantitative evaluation of atmospheric models, although not stratospheric CCMs. For example, Douglass et al. (1999) and Strahan and Douglass (2004) performed a quantitative evaluation of stratospheric simulations from an off-line three-dimensional chemical transport model (CTM). In these two studies they assigned grades to multiple diagnostics to assess simulations driven by different meteorological fields. Brunner et al. (2003) compared model simulations of tropospheric trace gases with observations. In contrast to the comparison of climatological fields considered here, they focused on model-observations comparisons at the same time and location as the measurements. This is possible for CTMs driven by assimilated meteorological fields, but not for CCMs. More recently, several studies have performed quantitative evaluations of coupled ocean-atmosphere climate models, and formed a single performance index that combines the errors in simulating the climatological mean values of many different variables (e.g., Schmittner et al. (2005), Connolley and Bracegirdle (2007), (Reichler and Kim, 2008). Our approach draws on several features of the above studies. Several of the diagnostics considered here were considered in Douglass et al. (1999) and Strahan and Douglass (2004), and, in a similar manner to Schmittner et al. (2005), Connolley and Bracegirdle (2007), Reichler and Kim (2008) and Gleckler et al. (2008), we form a single performance index for each model.

The methods used to evaluate the models and weight their projections are described in the next section. The models and diagnostics considered are then described in Sect. 3. Results are presented in Sect. 4, and conclusions and future work discussed in the final section.

2 Method

2.1 General framework

The general framework used in this paper to evaluate the models and weight their predictions involves the following steps.

1. A suite of observationally-based diagnostic tests are applied to each model.

2. A quantitative metric of performance (grade) is assigned to the application of each observations-model comparison (diagnostic) to each model, i.e., g_{jk} is the grade of the j-th diagnostic applied to the k-th model.

3. Next, the grades for each diagnostic are combined together to form a single performance index for each model, i.e., the single index of model k is

$$\bar{g}_k = \frac{1}{W} \sum_{j=1}^{N} w_j g_{jk} \tag{1}$$

where $W = \sum_{j=1}^{N} w_j$, N is the number of diagnostics, and w_j is the weight (importance) assigned to each diagnostic. If all diagnostics have equal importance then w_j is the same for all j.

4. Finally, the model scores are used to weight the predictions of a given quantity X from M models, i.e.

$$\hat{\mu}_X = \frac{1}{\sum \bar{g}_k} \sum_{k=1}^{M} \bar{g}_k X_k. \tag{2}$$

If $\bar{g}_k$ are the same for all models then this reduces to the normal multi-model mean μ_X. The model scores can also be used to form a weighted variance:

$$\hat{\sigma}_X^2 = \frac{\sum \bar{g}_k}{(\sum \bar{g}_k)^2 - \sum \bar{g}_k^2} \sum_{k=1}^{M} \bar{g}_k (X_k - \hat{\mu}_X)^2. \tag{3}$$

Again, if $\bar{g}_k$ is the same for all models then this reduces to the normal multi-model variance σ_X^2.

If the focus is solely on assessing model performance then only steps 1 and 2 are required. However, if an overall model grade and weighted mean projections are required steps 3 and 4 are also needed.

The above framework is not fully objective as several subjective choices need to be made to apply it. For example, decisions need to be made on the diagnostics to apply, the observations to be used, the grading metric to be used, and the relative importance of the different diagnostics for predictions of quantity X. These issues are discussed below (see also discussion in (Connolley and Bracegirdle, 2007).

2.2 Grading metric

To implement the above framework a grading metric needs to be chosen. Several different metrics have been used in previous model-observation comparisons. For example, Reichler and Kim (2008) used the squared difference between model and observed climatological mean values divided by the observed variance, whereas Gleckler et al. (2008) focus on the root mean squared difference between the model and observed climatological mean values. In this study we wish to use a grading metric that can be applied to all diagnostics, and can easily be interpreted and compared between tests. We choose the simple diagnostic used by Douglass et al. (1999)

$$g=1-\frac{1}{n_g}\frac{|\mu_{\rm model}-\mu_{\rm obs}|}{\sigma_{\rm obs}} \quad (4)$$

where $\mu_{\rm model}$ is the mean of a given field from the model, $\mu_{\rm obs}$ is the corresponding quantity from observations, $\sigma_{\rm obs}$ is a measure of the uncertainty in the observations (see Sect. 3 for further discussion), and n_g is a scaling factor. If g=1 the simulated climatological mean matches the observations, and smaller g corresponds to a larger difference between model and observations. If g<0 then the model-observations difference is greater than n_g times σ. In our analysis we use, as in Douglass et al. (1999), n_g=3, for which g=0 if the model mean is 3σ from the observed climatological mean value. We reset negative values of g to zero, so g is always non-negative. As with the metrics used by Reichler and Kim (2008) and Gleckler et al. (2008), the metric g provides a measure of the difference in model and observed climatological means.

There are several other possible metrics. One is the statistic t used in the standard t-test (Wilks, 1995):

$$t=\frac{\mu_{\rm model}-\mu_{\rm obs}}{\sigma_{\rm T}\sqrt{\frac{1}{n_{\rm model}}+\frac{1}{n_{\rm obs}}}}, \quad (5)$$

where

$$\sigma_{\rm T}^2=\frac{(n_{\rm model}-1)\sigma_{\rm model}^2+(n_{\rm obs}-1)\sigma_{\rm obs}^2}{(n_{\rm model}+n_{\rm obs}-2)}.$$

Unlike the metric g, or the above metrics, the t-statistic involves the variance in both the observations and models. The t-statistic has the advantage over the metric (4) in that there is a standard procedure to determine the statistical significance of the differences between models and observations from it. However, the value of t depends on the number of elements in the data sets, and it is not as easy to compare t from different tests that use datasets with a different number of elements as for the metric g. Also, most importantly it cannot be applied to all our diagnostics as some lack long enough data records for calculation of variance in the observations.

There is in fact a close relationship between g and t, and the statistical significance of the model-observations difference can be estimated from the value of g. To see this consider the idealized case where the models and observations have the same number of data elements and also the same standard deviations. Then from (4) and (5) we have

$$t=\sqrt{\frac{n}{2}}n_g(1-g), \quad (6)$$

where n=$n_{\rm model}$=$n_{\rm obs}$. This relationship holds only in the above special case, but as shown in Sect. 4.2 it is a good approximation for (at least some of) the more general cases considered here.

Given Eq. (6) we can estimate the value of t, and hence the statistical significance of the model-observations difference, from the grade g. For example, a model is statistically different from the observations at the p% confidence level if $g<g*$, where $g* = 1-\sqrt{\frac{2}{n}}\frac{t_p}{n_g}$ and t_p is the critical value for the two-sided t test with $2n-2$ degrees of freedom. For n_g=3 and p = 5%, this yields $g*$=0.70 and $g*$=0.78 for n=11 and n=20 (the cases shown in Fig. 6 below), respectively. So assuming decadal or longer datasets a value of g<0.7 indicates the difference between model and data are statistically significant (at 5% level).

The above relationship can also be used to estimate where the grades from two models are statistically different. As we have assumed n and σ are the same for all models and the observations, it can be shown that the grades from two models are statistically different if the difference in their grades exceeds 1−$g*$ (i.e., significant at the 5% level if the grades differ by 0.3 and 0.22 for n=11 and 20, respectively).

The above values are exact only in the idealized case of equal standard deviations, but provide useful estimates of the significance of differences in g.

In the metric (4) the errors for different diagnostics are normalized by the uncertainty in the observations. This means that the mean grade over all models for each diagnostic will vary if the models overall are better/poorer at simulating a particular process or field. Also, some quantities may be more tightly constrained by observations than others, and this can be captured (by variations in $\sigma_{\rm obs}$) by the metric (4). A different approach was used by Reichler and Kim (2008) and Gleckler et al. (2008), who normalized the error by the "typical" model error for each quantity. This approach means that the average grade over all models will be roughly the same for all diagnostics (around zero).

In summary, we use the metric (4) in our analysis because it is simple, can be applied to all the diagnostic tests, is easy to interpret, and can easily be compared between tests. Also, as shown above the statistical significance can be estimated from it.

One limitation with any metric, and in fact any comparison between models and observations, is uncertainties in the observations (and in particular, unknown biases in the observations). If the observations used are biased, then an unbiased model that reproduces the real atmosphere may get a low grade, while a model that is biased may get a high grade. The potential of a bias in the observational dataset used can

Table 1. CCMs used in this study. The models discussed in this paper are numbered alphabetically.

Name	Reference
AMTRAC	Austin et al. (2006)
CCSRNIES	Akiyoshi et al. (2004)
CMAM	Fomichev et al. (2007)
E39C	Dameris et al. (2005)
GEOSCCM	Pawson et al. (2008)
LMDZrepro	Lott et al. (2005)
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)

be assessed for diagnostics where there are several sources of data, and we consider several such cases in Sect. 4.2 below. However, for most diagnostics considered here multiple data sets are not available.

3 Models and diagnostics

As discussed in the Introduction we consider the CCM simulations, diagnostics, and observations that were evaluated in E06. The thirteen models considered are listed in Table 1, and further details are given in E06 and the listed reference for each model.

The simulations considered in the E06 model evaluation, and used here to form model grades, are transient simulations of the last decades of the 20th century. The specifications of the simulations follow, or are similar to, the "reference simulation 1" ("REF1") of CCMVal, and include observed natural and anthropogenic forcings based on changes in sea surface temperatures (SSTs), sea ice concentrations (SICs), surface concentrations of well-mixed greenhouse gases (GHGs) and halogens, solar variability, and aerosols from major volcanic eruptions. The simulations considered in E07 are projections of the 21st century ("REF2" simulation), in which the Intergovernmental Panel on Climate Change (IPCC) Special Report on Emission Scenarios (SRES) A1B GHG scenario and the WMO (2003) Ab surface halogens scenario are prescribed. SSTs and SICs in REF2 are taken from coupled atmosphere-ocean model projections using the same GHG scenario.

The diagnostic tests applied to the past CCM simulations are listed in Table 2. Each diagnostic is based on a model-observations comparison shown in E06, and the figures in E06 showing the comparison are listed in Table 2. The exception is the middle latitude Cl_y test where the comparison is shown in E07, and also Figure 1 below. Note that E06 also

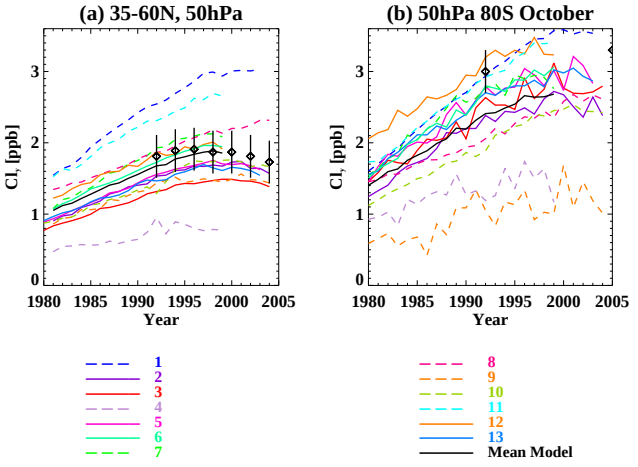

Fig. 1. Times series of **(a)** annual-mean, 35 dg–60 dg N, and **(b)** October-mean, 80 dg S Cl_y at 50 hPa from CCM simulations (curves) and observations (symbols plus vertical bars). See Table 1 for model names.

compared the simulated ozone with observations, but we do not include these comparisons here. We focus on diagnostics of processes rather than diagnostics of past and present ozone, as such diagnostics might be a better predictor of a model's ability to make reliable projections.

Many other diagnostic tests could be used in this analysis, such as those defined in Table 2 of Eyring et al. (2005). However, for this study we focus on a relatively small number of diagnostics that have already been applied. These diagnostics were chosen by E06 as they test processes that are key for simulating stratospheric ozone. In particular, the diagnostics were selected to assess how well models reproduce a) polar dynamics, b) stratospheric transport, and c) water vapor distribution. Correctly simulating polar ozone depletion (and recovery) requires the dynamics of polar regions, and in particular polar temperatures, to be correctly simulated. Assessing the reality of these aspects of the CCMs is thus an important component of a model assessment. Another important aspect for simulating ozone is realistic stratospheric transport. Of particular importance is simulating the integrated transport time scales (e.g. mean age), which plays a key role in determining the distributions of Cl_y and inorganic bromine (Br_y). Changes in water vapor can have an impact on ozone through radiative changes, changes in HO_x, or changes in formation of polar stratospheric clouds (PSCs), and it is therefore also important to assess how well models simulate the water vapor distribution.

Although the model-observations comparisons have already been presented further decisions still need to be made to quantify these comparisons. For example, choices need to be made on the region and season to be used in the grading metric (4). This choice will depend on the process to be examined as well as the availability of observations.

Table 2. Diagnostic tests used in this study.

Short Name	Diagnostic	Quantity	Observations	Fig. E06
Temp-SP	South Polar Temperatures	SON, 60 dg −90 dg S, 30–50 hPa	ERA-40	1
Temp-NP	North Polar Temperatures	DJF, 60 dg −90 dg N, 30–50 hPa	ERA-40	1
U-SP	Transition to Easterlies	U, 20 hPa, 60 dg S	ERA-40	2
HFlux-SH	SH Eddy Heat Flux	JA, 40 dg −80 dg S, 100 hPa	ERA-40	3
HFlux-NH	NH Eddy Heat Flux	JF, 40 dg −80 dg N, 100 hPa	ERA-40	3
Temp-Trop	Tropical Tropopause Temp.	T, 100 hPa, EQ	ERA-40	7a
H_2O-Trop	Entry Water Vapor	H_2O, 100 hPa, EQ	HALOE	7b
CH_4-Subt	Subtropical Tracer Gradients	CH_4, 50 hPa, 0–30 dg N/S, Mar/Oct	HALOE	5
CH_4-SP	Polar Transport	CH_4, 30/50 hPa, 80 dg S, Oct	HALOE	5
CH_4-EQ	Tropical Transport	CH_4, 30/50 hPa, 10 dg S-10 dg N, Mar	HALOE	5
Tape-R	H_2O Tape Recorder Amplitude	Amplitude Attenuation R	HALOE	9
Tape-c	H_2O Tape Recorder Phase Speed	Phase Speed c	HALOE	9
Age-50 hPa	Middle Stratospheric Age	10 hPa , 10 dg S-10 dg N and 35 dg −55 dg N	CO_2 and SF_6	10
Age-10 hPa	Lower Stratospheric Age	50 hPa , 10 dg S-10 dg N and 35 dg −55 dg N	ER2 CO_2	10
Cl_y-SP	Polar Cl_y	80 dg S, 50 hPa , Oct	UARS HCl	12
Cl_y-Mid	Mid-latitude Cl_y	30 dg −60 dg N, 50 hPa , Annual mean	multiple	–

Another important issue in the calculation of the grade is the assignment of σ_{obs}. For some quantities there are multi-year observations and an interannual standard deviation can be calculated, and in these cases we use this in metric (4). For other quantities these observations do not exist and an estimate of the uncertainty in the quantity is used as σ_{obs}. This is not very satisfying, and it would be much better if in all cases σ_{obs} included both measurement uncertainty and variability. Even if estimates of different measurements uncertainties and the variability are available, combining these to form a single uncertainty estimate is not straightforward, and should be examined in future studies. Note, however, that σ_{obs} does not impact the ranking of models for a particular test, it only impacts comparisons of the grades for different tests.

The regions, seasons, and observations used for each diagnostic are listed in Table 2 and are described in more detail below.

– For the polar temperature diagnostic we focus on the lower stratosphere during winter and spring, as these temperatures are particularly important for modeling polar ozone depletion. Specifically we consider polar average (60–90 dg N or S) temperatures averaged over 50 to 30 hPa and January to March (60–90 dg N) or September to November (for 60–90 dg S). These tests will be referred to as "Temp-NP" and "Temp-SP", respectively. Climatological mean and interannual standard deviation of ERA-40 reanalyses (Uppala et al., 2005) for 1980–1999 are used for the observations in metric (4), and the same period is used to calculate the model climatology. The biases of the models relative to ERA-40 reanalyses are shown in Fig. 1 of E06.

– The transition to easterlies diagnostic ("U-SP") measures the timing of the break down of the Antarctic polar vortices in the CCMs. It is based on Fig. 2 of E06, which shows the timing of the transition from westerlies to easterlies for zonal-mean zonal winds at 60 dg S. The grade is determined using the date for the transition at 20 hPa, and climatological ERA-40 reanalysis for the observations.

– The vertical propagation of planetary waves into the stratosphere plays a significant role in determining polar temperatures during winter and spring (Newman et al., 2001). This wave forcing can be diagnosed with the mid-latitude 100 hPa eddy heat flux for the regions and periods shown in Fig. 3 of E06: 40 dg–80 dg N for January–February ("HFlux-NH"), or 40 dg–80 dg S for July–August ("HFlux-SH"). This is only one aspect of the information shown in this figure, and grades could also be based, for example, on the slope of the heat flux–temperature relationship.

– The tropical tropopause temperature diagnostic ("Temp-Trop") is based on the tropical temperature at 100 hPa, shown in Fig. 7a of E06. For this diagnostic Eq. (4) is applied for each month separately, using ERA-40 climatological mean and interannual standard deviation for the observations, and then the average of these 12 values is used as the single grade for this diagnostic.

– The entry water vapor diagnostic ("H_2O-Trop") is based on the tropical water vapor at 100 hPa, see Fig. 7b of E06. As for the tropical tropopause temperature diagnostic Eq. (4) is applied for each month separately, this

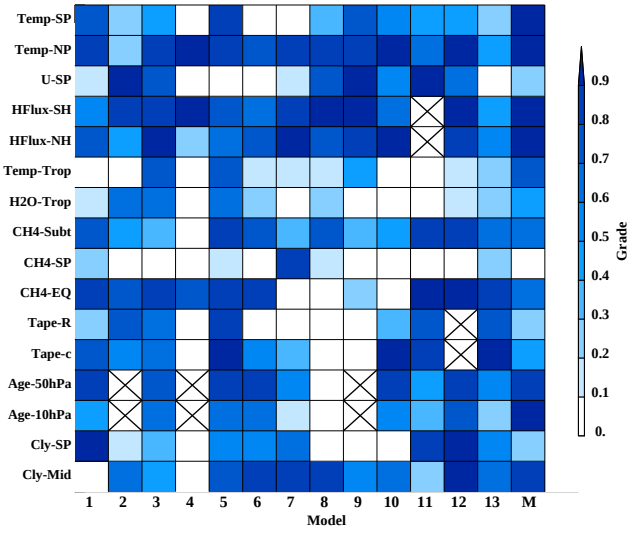

Fig. 2. Matrix displaying the grades (see color bar) for application of each diagnostic test to each CCM. Each row shows a different test, and each column a CCM. The right most column is the "mean model". A cross indicates that this test could not be applied, because the required output was not available from that model. See Table 1 for model names.

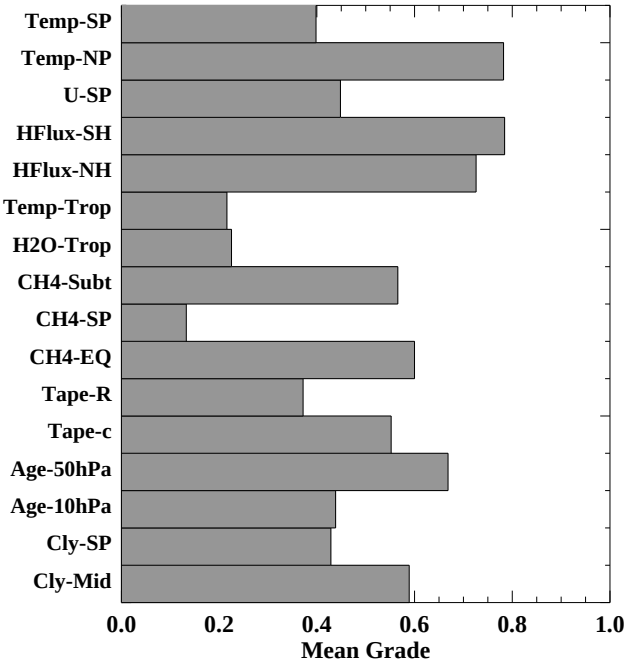

Fig. 3. Average grade over all models for each diagnostic test.

time using HALOE climatological mean and interannual standard deviation for the observations (Grooßand Russell, 2005), and the average of these 12 values is used as the single grade. The model climatology is for 1990–1999, whereas the HALOE observations are for 1991–2002.

– Three diagnostics that are mainly determined by transport are based on comparisons of the simulated and observed methane (CH_4) distributions, see Fig. 5 of E06. We focus on the lower stratosphere, and use these diagnostics to assess the lower stratospheric transport in the tropics and polar regions. As CH_4 at the tropical tropopause (100 hPa) is very similar in all models and observations we use the tropical (10 dg S–10 dg N) averaged values between 30 and 50 hPa to quantify differences in transport in the tropical lower stratosphere ("CH_4-EQ"). Similarly, we use October CH_4 at 80 dg S averaged values between 30 and 50 hPa to quantify differences in transport in the Antarctic lower stratosphere ("CH_4-SP"). Note there is limited coverage by HALOE in southern polar regions leading to increased uncertainty in the observed climatological mean values. The coverage is even worse at 80 dg N in winter-spring, even if equivalent latitude is used, which is why we do not include a diagnostic for 80 dg N. To test subtropical meridional gradients we use the difference in 50 hPa CH_4 between 0 dg N and 30 dg N, for March and between 0 dg N and 30 dg S for October ("CH_4-Subt"). A grade is determined for each month separately, and then averaged together to form a single grade for subtropical gradients.

– Diagnostics of the water vapor tape recorder (Mote et al., 1996) test the ability of models to reproduce the amplitude and phase propagation of the annual cycle in tropical water vapor, which in turn tests the model's tropical transport. As in Hall et al. (1999), we use the phase speed c ("Tape-c") and attenuation of the amplitude R ("Tape-R") to quantify the ability of the models to reproduce the observed propagation of the H_2O annual cycle. The attenuation $R=H/\lambda$, where $\lambda=c\times 1$ yr is the vertical wavelength and H is the attenuation scale height of the amplitude, $A=\exp(-z/H)$. The values of c (R) are determined from linear (exponential) fits to the simulated phase lag (relative amplitude) from the level of maximum amplitude to 10 km above this level, and compared with similar calculations using HALOE observations (see Fig. 9 of E06).

– The mean age diagnostics are based on comparisons of the simulated with observed mean age at 50 hPa and 10 hPa shown in Fig. 10 of E06. The mean age is an integrated measure of the transport in the stratosphere, and together with tape recorder diagnostics place a stringent test on models transport (e.g. Waugh and Hall, 2002). At each pressure level Eq. (4) is applied separately for the tropics (10 dg S–10 dg N) and northern mid-latitudes (35 dg N–55 dg N), and then the average

of these 2 values is used as the single grade for each pressure level ("Age-10 hPa" and "Age-50 hPa "). Balloon observations are used for mean values and uncertainty at 10 hPa (see symbols in Fig. 10b of E06), whereas ER2 observations are used for 50 hPa (Fig. 10c of E06).

– The Cl_y diagnostics are based on comparisons with the observed lower stratospheric (50 hPa) Cl_y shown in Fig. 12 of E06 and Fig. 1 of E07, and repeated in Fig. 1. We calculate grades separately for spring in the southern polar region (80 dg S, October; "Cly-SP") and for annual-mean values in northern mid-latitudes (30 dg - 60 dg N; "Cly-Mid"). The observed mean values and uncertainties used are the same as shown in Fig. 1. As the REF1 simulation in some models stops at the end of 1999, only observations in the 1990s are used in calculating the grade.

4 Results

4.1 Model grades

The diagnostic tests listed in Table 2 have been applied to the thirteen CCMs listed in Table 1 and grades g determined using metric (4). We also calculate the grade for the "mean model", i.e., the mean over all models is calculated for the various quantities listed in Table 2, and then a grade is calculated using this mean value in the metric (4).

We consider first, as an example, the grades of the Cl_y tests. Figure 1 shows the time series of mid-latitude and polar Cl_y at 50 hPa from the 13 models, together with observations and the mean of the models. As discussed in E06 there is a large spread in the modeled Cl_y, and some large model-observations differences. For polar Cl_y, models 1, 11, and 12 produce values close to the observations, and these models have grades around 0.9 for n_g=3 (see second to bottom row in Fig. 2). However, most of the models produce polar Cl_y much lower than observed, and several models are more than 3σ from the mean observations and have a grade of 0 (as noted above, values of g less than zero are reset to zero). Models 4 and 9 are in fact more than 5σ from the mean observations. For mid-latitude Cl_y the model-observations differences are not as large as for polar Cl_y. Several models are within σ of the mean value (and have g>0.66) and only one model (model 4) is 3 σ away from the mean value. Note the model 8 has a high grade for mid-latitude Cl_y because it agrees with observations before 2000, but Figure 1 shows that the Cl_y in this model continues to increase and deviates from observations after 2000. If these later measurements were used the grade for this model would be much lower.

We now consider the grades for all diagnostics. The results of the application of each test to each model are shown in Fig. 2. In this matrix ("portrait" diagram) the shading of each element indicates the grade for application of a particular diagnostic to a particular model (a cross indicates that this test could not be applied, because the required output was not available from that model). Each row corresponds to a different diagnostic test (e.g. bottom two rows show g for the two Cl_y tests), and each column corresponds to a different model. The grades for the "mean model" are shown in the right-most column.

Figure 2 shows that there is a wide range of grades, with many cases with $g\approx$0 and also many cases with g>0.8. As discussed in Sect. 2.2, model-data differences are significant at the 5% level for g less than around 0.7 for n=10 (or less than 0.8 for n=20). These large variations in g can occur for different diagnostics applied to the same model (e.g., most columns in Fig. 2) or for the same diagnostic applied to different models (e.g., most rows in Fig. 2). The wide range in the ability of models to reproduce observations, with variations between models and between different diagnostics, can be seen in the figures in E06. This analysis quantifies these differences and enables presentation in a single figure.

The wide range of grades for all diagnostic tests shows that there are no tests where all models perform well or all models perform poorly. However, the majority of models perform well in simulating north polar temperatures and NH and SH heat fluxes (mean grades over all models are larger than 0.7), and, to a lesser degree, mid-latitude age (mean grade greater than 0.6), see Figure 3. At the other extreme the majority of models perform poorly for the Tropical Tropopause Temperature, Entry Water Vapor, and Polar CH_4 tests (mean grades less than or around 0.2). Note that caution should be applied when comparing grades from different diagnostics as g is sensitive to the choice of σ, e.g., use of a smaller σ in a test results in lower grades, and some of the variations between diagnostics could be due to differences in the assigned σ.

Figure 2 also shows that there are no models that score high on all tests or score low on all tests. However, differences in the performance of models can be seen and quantified. For example, several models get low grades on multiple tests, i.e., models 4, 7, 8, and 9 have g near zero for 4 or more transport tests. The poorer performance of these models for several of the transport diagnostics was highlighted in E06.

To further examine the difference in model performance we compare a single performance index calculated from the average grades ($\bar{g}_k$) for each model. This is shown in Fig. 4a, where the average grade is calculated assuming that all diagnostic tests are equally important (i.e., w_j=1 in Eq. 1). If a grade is missing for a particular test and model (crosses in Fig. 2) then this test is not included in the average for that model. There is a large range in the average performance of the models, with $\bar{g}$ varying from around 0.2 to around 0.7. The value of $\bar{g}$ changes with n_g but there is a very similar variation between models and the ranking of models for different n_g. For example, using n_g=5 results in a grade around 0.1 larger, for all diagnostics.

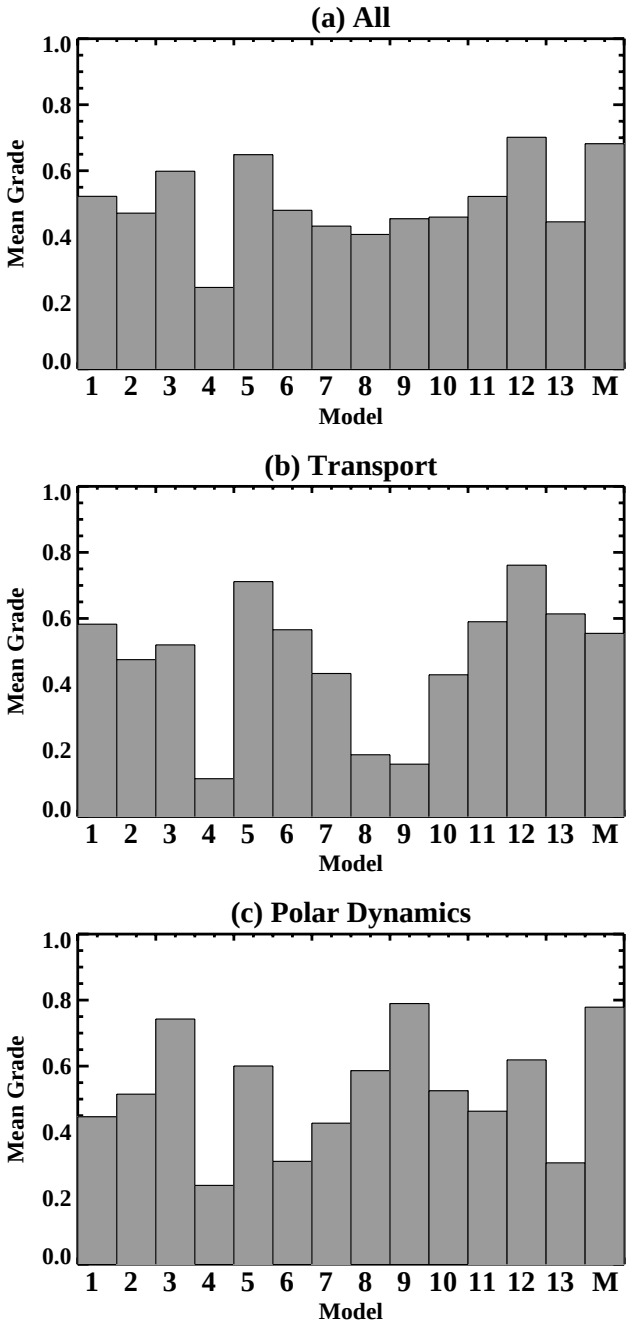

Fig. 4. Average grade for each model for **(a)** average over all diagnostic tests used in this study (see Table 2), **(b)** average only over transport diagnostics, and **(c)** average only over polar dynamics diagnostics. Note, we have evaluated only a subset of key processes important for stratospheric ozone.

The average grade for a model can also be calculated separately for diagnostics that are mainly determined by transport or polar dynamics diagnostics, see Figure 4b and c. The tropical tropopause temperature and H_2O diagnostics are not included in either the polar dynamics or transport averages. The range of model performance for the transport diagnostics is larger than the performance for dynamics diagnostics, with $\bar{g}$ varying from around 0.1 to around 0.8. Figure 4 also shows that some models simulate the polar dynamics much better than the transport (e.g., models 4, 8, and 9) while the reverse is true for others (e.g., models 6 and 13). Part of this could be because the dynamics tests focus on polar dynamics, whereas the majority of the transport diagnostics measure, or are dependent, on tropical lower stratospheric transport.

It is of interest to compare the model average grades shown in Fig. 4 with the segregation of models made by E07. In the plots in E06 and E07 solid curves were used for fields from around half the CCMs and dashed curves for the other CCMs, and E07 stated that CCMs shown with solid curves are those that are in general in good agreement with the observations in the diagnostics considered by Eyring et al. (2006). In making this separation E06 put more emphasis on the transport diagnostics than the temperature diagnostics, with most weight on Cl_y comparisons. As a result the separation used in the E06 and E07 papers is not visible in the mean grades over all diagnostics but can be seen in the average of the transport grades. The models shown as solid curves in E06 and E07 (models 2, 3, 5, 6, 12, and 13) all have high average transport grades, while nearly all those shown with dashed curves have low average transport grades. The exceptions are models 1 and 11 which were shown as dashed curves in E06 and E07 but whose average transport grades in Fig. 4 are high. The reasons for this difference is, as mentioned above, that E06 put high weight on the comparisons with Cl_y. Models 1 and 11 significantly overestimate the mid-latitude Cl_y and have a very low score for the mid-latitude Cl_y test (Fig. 2).

Another interesting comparison is between the grades of individual models and that of the "mean model". Analysis of coupled atmosphere-ocean climate models has shown that the "mean model" generally scores better than all other models (e.g. Gleckler et al., 2008). This is not however the case for the CCMs examined here (see right most column of Fig. 2). For some of the diagnostics the grade of the mean model is larger than or around the grade of the best individual model, e.g., the NH polar temperatures, heat flux, 10 hPa mean age, and mid-latitude Cl_y diagnostics (see right most column in Fig. 2). However, for most diagnostics the grade for the mean model is smaller than that of some of the individual models, with a large difference for the transition to easterlies, south polar CH_4, tape recorder attenuation, and polar Cl_y diagnostics. In these latter diagnostics there is significant bias in most, but not all, of the models, and this bias dominates the calculation of the mean of the models and the grade of the mean model is less than 0.3. But for each of these diagnostics there is at least one model that performs well (e.g., g around or greater than 0.8), and has a higher grade than the mean model. The contrast between a diagnostic where the grade for the mean model is higher than or around the best individual models and a diagnostic where

the grade for the mean model is lower than many individual models can be seen in Fig. 1. In panel a the Cl_y for individual models is both above and below the observations, and the mean of the models is very close to the observations (and has a high grade). In contrast, in panel b most models underestimate the observed Cl_y and the mean Cl_y is much lower than the observations (and several models).

4.2 Sensitivity analysis

As discussed above several choices need to be made in this analysis. A detailed examination of the sensitivity to these choices is beyond the scope of this study. However, a limited sensitivity analysis has been performed for some tests.

We first consider the sensitivity of the grades to the source of the observations. We focus on diagnostics based on the temperature field, as data are available from different meteorological centers. Fig. 5 shows the grades for the a) Temp-NP, b) Temp-SP, and c) Temp-Trop tests when different meteorological analyses are used for the observations. The grades for Temp-NP (panel a) are not sensitive to whether the ERA40, NCEP stratospheric analyses (Gelman et al., 1996) or UK Met Office (UKMO) assimilated stratospheric analyses (Swinbank and O'Neill, 1994) are used as the observations. This is because the climatological values from the three analyses are very similar (within 0.2 K), and the differences between analyses is much smaller than model-observations differences (see Fig. 1 of E06). There is larger sensitivity for Temp-SP (panel b) as there are larger differences between the analyses. However, the general ranking of the models is similar which ever meteorological analyses are used in the grading, i.e., models 1, 5 and 9 have high grades and models 4, 6 and 7 have low grades for all 3 analyses.

The above insensitivity to data source does not, however, hold for the Temp-Trop diagnostic. Here there are significant differences between the meteorological analyses, and the model grades vary depending on which data source is used. The climatological mean UKMO values are around 1 to 2 K warmer than those of ERA40, depending on the month (see Fig. 7 of E06), and as a result very different grades are calculated for some models, see Figure 5c. For models that are colder than ERA40 lower grades are calculated if UKMO temperatures are used in the metric (e.g., models 3, 5, 9, 12, and 13), whereas the reverse is true for models that are warmer than ERA40 (models 4, 6, 7, 8 and 10).

The above sensitivity to meteorological analyses highlights the dependence of the grading, and any model-data comparison, on the accuracy of the observations used. It is therefore important to use the most accurate observations in the model-data comparisons. With regard to the temperature datasets used above, intercomparisons and comparisons with other datasets have shown that some biases exist in these meteorological analyses. In particular, the UKMO analyses have a 1–2 K warm bias at the tropical tropopause, and the NCEP analyses have a 2–3 K warm bias at the tropical tropopause and a 1–3 K warm bias in Antarctic lower stratosphere during winter-spring (Randel et al., 2004). Given these biases, it is more appropriate to use, as we have, the ERA40 analyses for the model grading.

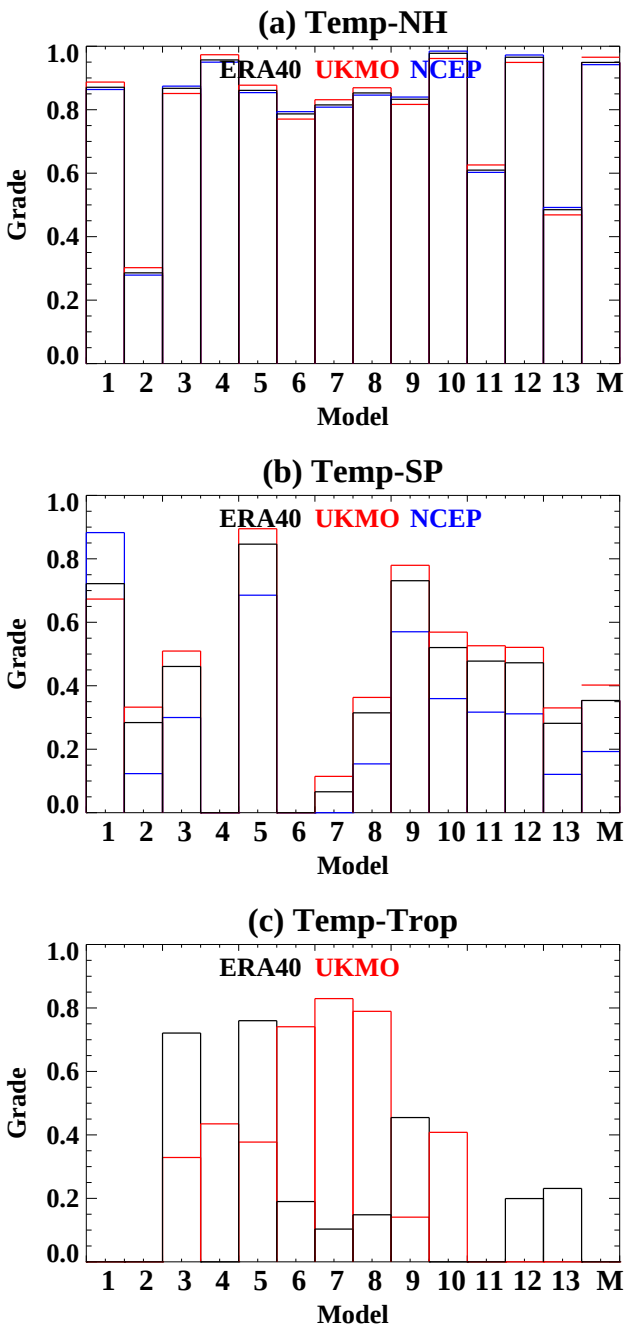

Fig. 5. Comparison of model grades for the **(a)** Temp-NP, **(b)** Temp-SP, and **(c)** Temp-Trop tests when ERA40 (black), UKMO (red), or NCEP (blue) meteorological analyses are used for the observations in the metric (4).

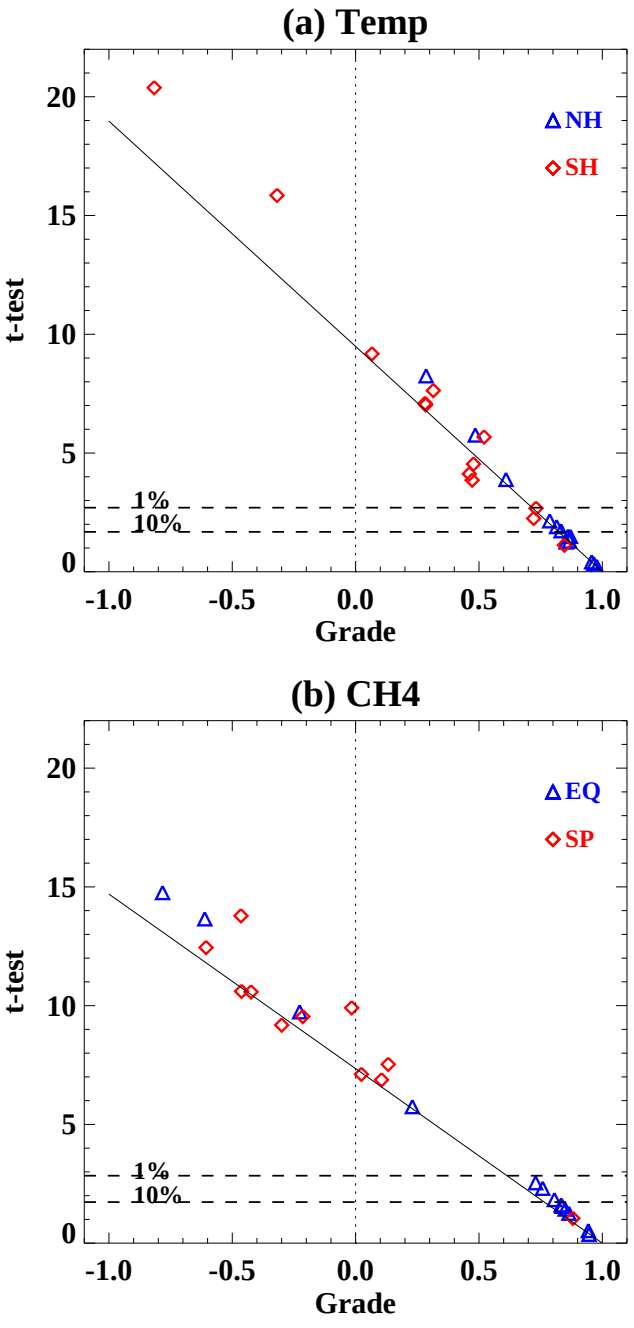

Fig. 6. Comparison of t-statistic with grading metric g for **(a)** Temp-NP and Temp-SP tests or **(b)** CH_4-EQ and CH_4-SP diagnostic tests. The solid lines show the theoretical relationship given by Eq. (6) for (a) n=20 and (b) n=11. The horizontal dashed lines show the critical values of t for statistically significant differences at the 1% and 10% level.

Another choice made in the analysis is the metric used to form the grades. As discussed in Sect. 2.2, an alternate metric to the one used here is the t-statistic. It was shown in Sect. 2.2 that there is a simple linear relationship between t and g in cases with the same number of data points and same standard deviations for the models and observations (see Eq. 6). To test this relationship in a more general case we compare t and g for tests using temperature and CH_4 fields. For these tests there are multi-year observational data sets available, and the mean and variance from these observations and from the models can be used to calculate t and g.

Figure 6 shows this comparison for the a) Temp-NP and Temp-SP, and b) CH_4-EQ and CH_4-SP diagnostic tests. For these comparisons we do not set negative values of g to zero, so that we can test how well the relationship (6) holds. If we set negative g to zero then the points left of the vertical dashed lines move to the left to lie on this line. As expected there is a very close relationship between the calculated values of t and g. For all four diagnostics, g and t are highly anti-correlated and models that have a high (low) value of g have a low (high) value of t. As a result, very similar ranking of the models is obtained for both metrics. Furthermore, the values of t are close to those predicted by (6), which are shown as the solid lines in Fig. 6. The deviations from this linear relationship are due to differences in σ between models and observations.

The horizontal dashed lines in Fig. 6 show the critical values of t for statistically significant differences at the 1% and 10% level, i.e. if t is above the lines then the model-data differences are significant at this significance level. Similarly, if the value of g is less than the value where these horizontal lines cross the solid line the model-data differences are significant at these levels. Both metrics show that statistically significant differences exist between some models and the data.

The above shows that a very similar ranking of the models will be obtained whether t or g is used, and that Eq. (6) can be used to estimate the statistical significance of values of g. Hence, the results presented here are not likely to be sensitive to our choice of g as the metric.

4.3 Relationships among diagnostics

We now examine what, if any, correlations there are between grades for different diagnostics. If a strong correlation is found this could indicate that there is some redundancy in the suite of diagnostics considered, i.e., two or more diagnostics could be testing the same aspects of the models. Identifying and removing these duplications from the suite of diagnostics would make the model evaluation more concise. However, an alternative explanation for a connection between grades for different diagnostics could be that the models that perform poorly for one process also perform poorly for another process. In this case the two diagnostics are not duplications, and the consideration of both might provide insights into the cause of poor performance.

Figure 7 shows the correlation between the grades for each of the different diagnostics. There are generally low correlations between the grades, which indicates that in most cases

the tests are measuring different aspects of the model performance. There are however some exceptions, and several notable features in this correlation matrix.

As might be expected, several of these high correlations occur between grades for diagnostics based on the same field, e.g. between the two mean age diagnostics and the two tape recorder diagnostics. There might, therefore, be some redundancy in including two grades for each of these quantities, e.g., it might be possible to consider a single grade for mean age which considers just a single region or averages over all regions. However, this is not the case for all fields, and there are low correlations between the two Cl_y diagnostics, between the two polar temperature diagnostics, and between the CH_4 diagnostics. Thus diagnostics using the same fields can measure different aspects of the simulations, and averaging into a single grade might result in loss of information.

High correlations between grades might also be expected for some pairs of diagnostics that are based on different fields but are dependent on the same processes. One example are the Temp-Trop and H_2O-Trop diagnostics. As the water vapor entering the stratosphere is dependent on the tropical tropopause temperature a high correlation is expected between these two diagnostics. There is indeed a positive correlation, but the value of 0.48 may not be as high as expected. The fact that the correlation is not higher is mainly because of two models, whose performance differs greatly for these two diagnostics. The 100 hPa temperature in model 2 is much colder than observed (with a zero grade for this diagnostic) but the 100 hPa water vapor is just outside 1σ of the observed value (g=0.61). The reverse is true for model 9, which has a low water vapor grade but reasonable temperature grade. For the remaining models, models with low grades for the tropical temperature tend to have low grades for the water vapor diagnostic, and there is a much higher correlation (0.88). A higher correlation is found between the tropical cold point and 100 hPa water vapor correlation (Gettelman et al., 2008), but the above two models are still anomalous. The sensitivity of the correlations to results of two models illustrates that care should be taken interpreting these correlations. It also suggests problems with, and need for further analysis of, the two anomalous models that do not display the physically expected relationship between tropical tropopause temperature and entry water vapor.

There are high positive correlations between many of the transport diagnostics, i.e., there are generally high correlations between the tape recorder, mean age, and equatorial CH_4 diagnostics. This is likely because these diagnostics measure, or are dependent, on tropical lower stratospheric transport, and a model with good (poor) transport in the tropical lower stratosphere will have good (poor) grades for all these diagnostics.

Another area where strong correlations might be expected is between the heat flux, polar temperatures, and transition to easterlies diagnostics. However, in general, the correlations between these fields is not high. The exception is the high correlation (0.75) between the eddy heat flux and transition to easterlies in the southern hemisphere. A high correlation between these diagnostics might be expected as weak heat fluxes might lead to a late transition. However, this is not likely the cause of the high correlation. First, the heat flux diagnostic is for mid-winter (July-August) and the late-winter/spring heat fluxes are likely more important for the transition to easterlies. More importantly, several of the models with late transitions, and very low grades for this diagnostic, actually have larger than observed heat fluxes (models 1, 5 and 13), which is opposite than expected from the above arguments. Again, this anomalous behavior suggests possible problems with these models and the need for further analysis.

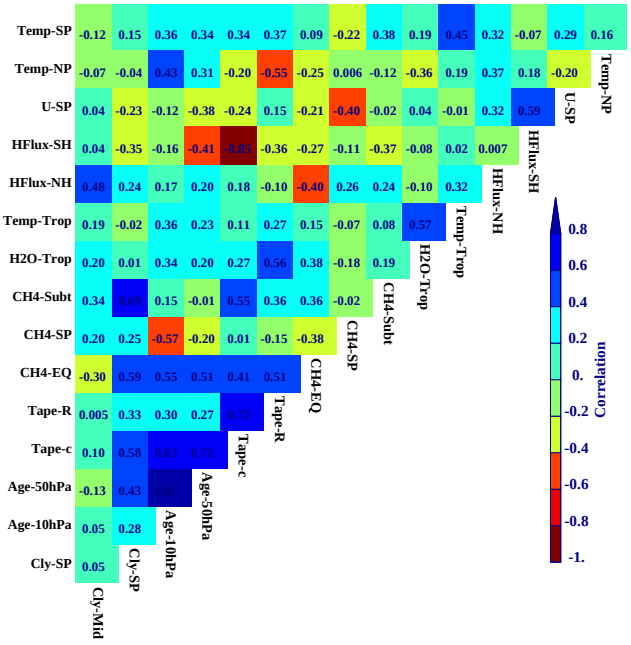

Fig. 7. Matrix displaying the correlation between the grades for different diagnostics.

Figure 7 also shows some correlation between the transport and dynamics diagnostics. In fact there are several cases where there is moderate to high negative correlations between a dynamics diagnostic and a transport diagnostic, suggesting that there are models that perform poorly for (tropical) transport diagnostics but perform well for (polar) dynamical diagnostics. As discussed above this is indeed the case for several models (most notably models 8 and 9), see Fig. 4).

4.4 Weighted ozone projections

The assignment of grades to the diagnostic tests enables relative weights to be assigned to the ozone projections from different models, and for a weighted mean to be formed that takes into account differing abilities of models to reproduce

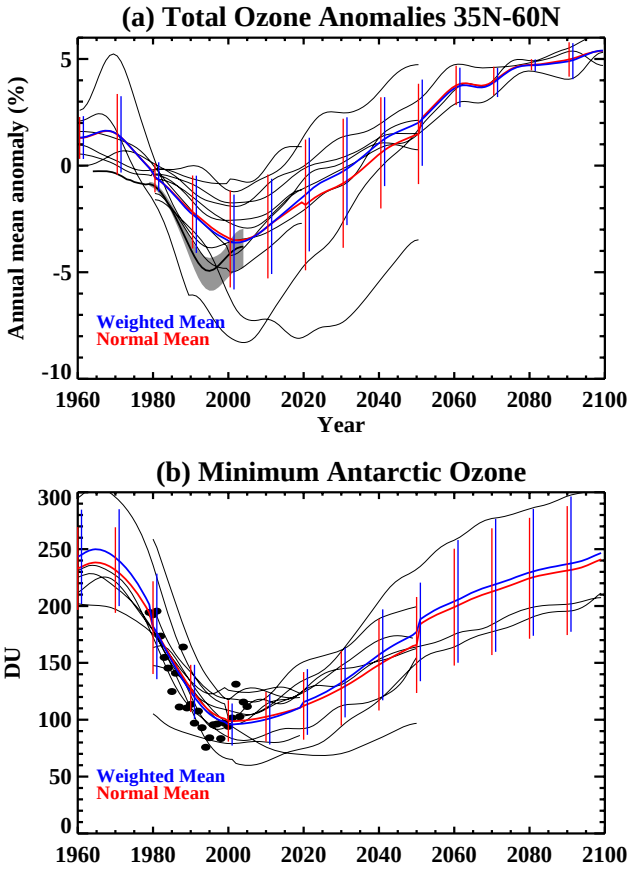

Fig. 8. Temporal variation of **(a)** annual mean anomalies for total ozone averaged over northern mid-latitudes (35 dg N to 60 dg N) and **(b)** minimum Antarctic ozone for individual models (black curves), unweighted mean (red) and weighted mean using performance indices based on the average transport grade (blue) of all models. The thick black curve and shaded region in (a) shows the mean and range of observed ozone anomalies, while the black dots in (b) show the observed minimum total ozone, see E06 for details.

key processes. We explore this issue using ozone projections for the 21st century made by the same CCMs, and shown in E07. Note, models 6 and 11 did not perform simulations into the future, and the analysis below is only for the other 11 models listed in Table 1. E07 examined the model projections of total column ozone for several different regions and for different diagnostics of polar ozone.

To form a weighted mean it is necessary to assign weights to each diagnostic so that a single model performance index can be formed (e.g. $\bar{g}_k$ in equation 1). We have calculated the weighted mean ozone, for each of the regions and diagnostics considered in E07, for a variety of weights and single model performance indices. This includes using equal weights for all diagnostics so that the performance index is the grade shown in Fig. 4a, as well as performance indices based only on transport diagnostics (Fig. 4b), based only on polar dynamics diagnostics (Fig. 4c), or based on a single diagnostic (e.g., mid-latitude or polar Cl_y). In all cases there are generally only small differences between the weighted and unweighted mean values, even when there is a large variation between the model performance indices.

This weak sensitivity to weighting is illustrated in Fig. 8 where the unweighted mean values (red curves) are compared with weighted mean values (blue curves) using performance indices based on the average transport grade, for a) annual mean northern mid-latitude total ozone and b) minimum Antarctic total column ozone for September to October. Following E07, the mid-latitudes monthly anomalies in Fig. 8 were calculated by subtracting a detrended mean annual cycle calculated over the period 1980–1989, from each time series. Ideally, a longer period of pre-1980 years should be used to define the zero line, but several model simulations only started in 1980. Shepherd (2008) showed that when the baseline is defined by the 1960–1975 mean, the annual mean northern mid-latitude total ozone time series in one of the participating models agree better with the observations than implied by Fig. 8. The jumps in the mean curves, e.g. at 2050, in Fig. 8 occur because not all model simulations cover the whole time period, and there is a change in the number of model simulations at the location of the jumps, e.g. eight models performed simulations to 2050, but only three models simulate past 2050.

For both the northern mid-latitude ozone and Antarctic minimum ozone there are generally only small differences between the weighted and unweighted mean values (and between weighted and unweighted variances), even though there is a large variation in the model performance indices used (see Figure 4b). This is because there is a wide range in the ozone projections, and for most time periods models that simulate ozone at opposite extremes have similar grades. For example, there are large differences ($\approx$100 DU) in minimum Antarctic ozone after 2050 between one of the three models and the other two, but the three models have similar average grades ($\bar{g}\approx$0.6–0.7) resulting in similar unweighted and weighted means and variances. The largest difference between unweighted and weighted means occurs between 2030 and 2050. This difference is primarily because of the ozone and index for model 8. During this period model 8 predicts much lower ozone than the other models (the black curve with lowest ozone between 2030 and 2050 is model 8), but this model has a low transport index ($\approx$0.2) which means that less weight is put on the ozone from this model in the weighted mean. As a result the weighted mean is larger than the unweighted mean, and the uncertainty smaller, for this time period.

Similar results are found for other regions and ozone diagnostics, and for different weights. The average transport grade is used only for illustrative purposes, and not to imply this is the best index to use.

The conclusion from the above analysis is that, at least for the grading and diagnostics applied to these CCM simulations, weighting the model results does not significantly influence the multimodel mean projection of total ozone. A similar conclusion was reached by Stevenson et al. (2006) in their analysis of model simulations of tropospheric ozone, and by Schmittner et al. (2005) in their analysis of the thermohaline circulation in coupled atmosphere-ocean models.

5 Conclusions

The aim of this study was to perform a quantitative evaluation of stratospheric-resolving chemistry-climate models (CCMs). To this end, we assigned a quantitative metric of performance (grade) to each of the observationally-based diagnostics applied in Eyring et al. (2006), and quantified the ability of the thirteen CCMs to simulate a range of processes important for stratospheric ozone. The metric used is not a standard statistical quantity, but has previously been used by Douglass et al. (1999) and, more importantly, can be applied to each of the diagnostics in Eyring et al. (2006) even though some lack long enough data records for calculation of variance in the observations.

This analysis quantified several features noted in Eyring et al. (2006). A wide range of grades were obtained, showing that there is a large variation in the ability of the CCMs to simulate different key processes. This large variation in grades occurs both for different diagnostics applied to a single model and for the same diagnostic applied to different models. No model scores high or low on all tests, but differences in the performance of models can be seen. This is especially true for processes that are mainly determined by transport where several models get low grades on multiple diagnostics, as noted in Eyring et al. (2006).

The assignment of grades to diagnostic tests enables a single performance index to be determined for each CCM, and for relative weights to be assigned to model projections. Such a procedure was applied to the CCMs' projections of 21st century ozone (Eyring et al., 2007). However, except for some time periods, only small differences are found between weighted and unweighted multi-model mean and variances of ozone projections, and weighting these model projections based on the diagnostic tests applied here does not significantly influence the results.

Although the calculation of the grades and weighting of the ozone projections is relatively easy, there are many subjective decisions that need to be made in this process. For example, decisions need to be made on the grading metric to be used, the source and measure of uncertainty of the observations, the set of diagnostics to be used, and the relative importance of different processes/diagnostics. We have performed some limited analysis of the sensitivity to these choices, and Gleckler et al. (2008) have discussed these issues in the context of atmosphere-ocean climate models. However, further studies are required to address this in more detail. In particular, determining the relative importance of each diagnostic is not straightforward, and more research is needed to determine the key diagnostics and the relative importance of each of these diagnostics in the weighting.

This study provides only an initial step towards a quantitative grading and weighting of CCM projections. For example, quantitative metrics could be used to explore the value of weighting model projections based on their performance in simulating the present day climate. The presented grades are for a single version of each model, and we have evaluated only a subset of key processes important for stratospheric ozone, in particular we have focused on diagnostics to evaluate transport and dynamics in the CCMs and to a lesser extent the representation of chemistry and radiation. Furthermore, we have only evaluated the climatological mean state and not the ability of the models to represent seasonal and interannual variability and trends. However, this study does provide a framework, and benchmarks, for the evaluation of new CCMs simulations, such as those being performed for the upcoming SPARC CCMVal Report (http://www.pa.op.dlr.de/CCMVal/). It is expected that the grades will change with improvements in the models and observations, and once other diagnostics are evaluated.

Acknowledgements. We thank David Fahey, Steven Pawson, and the rest of the CCMVal community for helpful discussions of this concept at the CCMVal 2007 workshop. We acknowledge the CCM groups of AMTRAC (GFDL, USA), CCSRNIES (NIES, Tsukuba, Japan), CMAM (MSC, University of Toronto and York University, Canada), E39C (DLR, Oberpfaffenhofen, Germany), GEOSCCM (NASA/GSFC, USA), LMDZrepro (IPSL, France), MAECHAM4CHEM (MPI Mainz, Hamburg, Germany), MRI (MRI, Tsukuba, Japan), SOCOL (PMOB/WRC and ETHZ, Switzerland), ULAQ (University of LAquila, Italy), UMETRAC (UK Met Office, UK, NIWA, NZ), UMSLIMCAT (University of Leeds, UK), and WACCM (NCAR, USA) for providing their model data for the CCMVal Archive. Co-ordination of this study was supported by the Chemistry-Climate Model Validation Activity (CCMVal) for WCRP's (World Climate Research Programme) SPARC (Stratospheric Processes and their Role in Climate) project. We thank the British Atmospheric Data Center for assistance with the CCMVal Archive. This work was supported by grants from NASA and NSF.

Edited by: W. Lahoz

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.: A three-dimensional coupled chemistry-climate model simulation of past stratospheric trends, J. Atmos. Sci., 59, 218–232, 2002.

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

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.

Brunner, D., Staehlin, J., Rogers, H. L., et al.: An evaluation of the performace of chemistry transport models by comparison with research aircraft observations, Part 1: Concepts and overall model performance, Atmos. Chem. Phys., 3, 1609–1631, 2003

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

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.

Douglass, A. R., Prather, M. J., Hall, T. M., Strahan, S. E., Rasch,P. J., Sparling, L. C., Coy, L., and Rodriguez, J. M.: Choosing meteorological input for the global modeling initiative assessment of high-speed aircraft, J. Geophys. Res., 104, 27 545–27 564, 1999.

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.

Eyring, V., Harris, N. R. P., Rex, M., et al.: A strategy for process-oriented validation of coupled chemistry-climate models, Bull. Am. Meteorol. Soc., 86, 1117–1133, 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 Grandpré, J., Beagley, S. R., et al.: Response of the middle atmosphere to CO_2 doubling: Results from the Canadian Middle Atmosphere Model, J. Climate, 20, 1121–1144, 2007.

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.

Gelman, M. E., Miller, A. J., Johnson, K. W., and Nagatani, R.: Detection of long-term trends in global stratospheric temperature from NMC analyses derived from NOAA satellite data, Adv. Space Res., 6, 17–26, 1996.

Gettelman, A., Birner, T., Eyring, V., Akiyoshi, H., Plummer, D. A., Dameris, M., Bekki, S., Lefevre, F., Lott, F., Brühl, C., Shibata, K., Rozanov, E., Mancini, E., Pitari, G., Struthers, H., Tian, W., and Kinnison, D. E.: The Tropical Tropopause Layer 1960–2100, Atmos. Chem. Phys. Discuss., 8, 1367–1413, 2008.

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

Grooß, J.-U. and Russell III, J. M.: Technical note: A stratospheric climatology for O_3, H_2O, CH_4, NO_x, HCl and HF derived from HALOE measurements, Atmos. Chem. Phys., 5, 2797–2807, 2005.

Hall, T. M., Waugh, D. W., Boering, K. A., and Plumb, R. A.: Evaluation of transport in stratospheric models, J. Geophys. Res., 104, 18 815–18 840, 1999.

Lott, F., Fairhead, L., Hourdin, F., and Levan, P.: The stratospheric version of LMDz: Dynamical Climatologies, Arctic Oscillation, and Impact on the Surface Climate, Climate Dynamics, 25, doi:10.1007/s003820050064, 2005.

Mote, P.W., K.H. Rosenlof, McIntyre, M. E., Carr, E. S., Gille, J. C., Holton, J. R., Kinnersley, J. S., Pumphrey, H. C., Russell III, J., and Waters, J. W.: An atmospheric tape recorder: The imprint of tropical tropopause temperatures on stratospheric water vapor, J. Geophys. Res., 101, 3989–4006, doi:10.1029/95JD03422, 1996.

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

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-Climate Model 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.

Randel, W., Udelhofen, P., Fleming, E., et al.: The SPARC Intercomparison of Middle-Atmosphere Climatologies. J. Climate, 17, 986–1003, 2004.

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

Schmittner A., Latif, M., and Schneider, B.: Model projections of the North Atlantic thermohaline circulation for the 21st century, Geophys. Res. Lett., 32, doi:10.1029/2005GL024368, 2005.

Shepherd, T.: Dynamics, Stratospheric Ozone and Climate Change, Atmos.-Ocean, 46, 117–138, 2008.

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.

Stevenson, D. S., Dentener, F. J., Schultz, M. G., et al.: Multimodel ensemble simulations of present-day and near-future tropospheric ozone. J. Geophys. Res., 111, D08301, doi:10.1029/2005JD006338, 2006.

Strahan, S. E. and Douglass, A. R.: Evaluating the credibility of transport processes in simulations of ozone recovery using the Global Modeling Initiative three-dimensional model, J. Geophys. Res., 109, D05110, doi:10.1029/2003JD004238, 2004.

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

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res., 106, 7183–7192, 2001.

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

Uppala, S., et al. [redacted]**please name at least 3 authors**: The ERA-40 reanalysis, Quart. J. Roy. Meteor. Soc., 131, 2961–3012, 2005.

Waugh, D. W. and Hall, T. M.: Age of stratospheric air: theory, observations, and models, Rev. Geophys., 40, 1010, doi:10.1029/2000RG000101, 2002.

Wilks, D. S.: Statistical methods in the atmospheric sciences, Academic Press, London, UK, 467 pp., 1995.

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.