Multimodel assessment of the upper troposphere and lower stratosphere: Extratropics


Received 17 January 2010; revised 7 June 2010; accepted 21 June 2010; published 23 October 2010.

[1] A multimodel assessment of the performance of chemistry-climate models (CCMs) in the extratropical upper troposphere/lower stratosphere (UTLS) is conducted for the first time. Process-oriented diagnostics are used to validate dynamical and transport characteristics of 18 CCMs using meteorological analyses and aircraft and satellite observations. The main dynamical and chemical climatological characteristics of the extratropical UTLS are generally well represented by the models, despite the limited horizontal and vertical resolution. The seasonal cycle of lowermost stratospheric mass is realistic, however with a wide spread in its mean value. A tropopause inversion layer is present in most models, although the maximum in static stability is located too high above the tropopause and is somewhat too weak, as expected from limited model resolution. Similar comments apply to the extratropical tropopause transition layer. The seasonality in lower stratospheric chemical tracers is consistent with the seasonality in the Brewer-Dobson circulation. Both vertical and meridional tracer gradients are of similar strength to those found in observations. Models that perform less well tend to use a semi-Lagrangian transport scheme and/or have a very low resolution. Two models, and the multimodel mean, score consistently well on all diagnostics, while seven other models score well on all diagnostics except the seasonal cycle of water vapor. Only four of the models are consistently below average. The lack of tropospheric chemistry in most models limits their evaluation in the upper troposphere. Finally, the UTLS is relatively sparsely sampled by observations, limiting our ability to quantitatively evaluate many aspects of model performance.

Copyright 2010 by the American Geophysical Union.
0148-0227/10/2010JD013884

1Department of Physics, University of Toronto, Toronto, Ontario, Canada.
2Atmospheric Chemistry Division, National Center for Atmospheric Research, Boulder, Colorado, USA.
3Max Planck Institut für Chemie, Mainz, Germany.
4Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA.
5New Mexico Institute of Mining and Technology, Socorro, New Mexico, USA.
6Department of Atmospheric and Oceanic Sciences, McGill University, Montreal, Quebec, Canada.
7Karlsruhe Institute of Technology, Institute for Meteorology and Climate Research, Karlsruhe, Germany.
8Department of Chemistry, University of Waterloo, Waterloo, Canada.
9Deutsches Zentrum für Luft- und Raumfahrt, Institut für Physik der Atmosphäre, Oerberpfaffenhausen, Germany.
10Johns Hopkins University, Baltimore, Maryland, USA.
11National Institute for Environmental Studies, Tsukuba, Japan.
12Environmental Physics Laboratory, Universidade de Vigo, Ourense, Spain.
13Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, New Jersey, USA.
14LATMOS, Institut Pierre-Simone Laplace, UVSQ, UMPC, CNRS/INSU, Paris, France.
15Department of Chemistry, Cambridge University, UK.
16Met Office, Exeter, UK.
17School of Earth and Environment, University of Leeds, Leeds, UK.
18Global Modelling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, Maryland, USA.
19Dipartimento di Fisica, Universita degli Studi di L'Aquila, L'Aquila, Italy.
20GAME/CNRM, Météo-France, CNRS, Toulouse, France.
21National Institute of Water and Atmospheric Research, Lauder, New Zealand.
22Environment Canada, Toronto, Ontario, Canada.
23Physikalisch-Meteorologisches Observatorium Davos, Davos, Switzerland.
24Institute for Atmospheric and Climate Science, ETH, Zurich, Switzerland.
25Canadian Centre for Climate Modelling and Analysis, Victoria, British Columbia, Canada.
26Meteorological Research Institute, Tsukuba, Japan.
1. Introduction

The upper troposphere/lower stratosphere (UTLS) plays an important role in the radiative forcing of the climate system and in chemistry-climate coupling [Shepherd, 2007]. Changes in the extratropical UTLS determine the stratospheric influence on the troposphere through for example the transport of stratospheric ozone into the troposphere, UV fluxes [Hegglin and Shepherd, 2009], or radiative forcing of the surface climate [Solomon et al., 2010]. It is therefore important that chemistry-climate models (CCMs) used for the prediction of the ozone layer and climate change represent accurately the chemical and dynamical structures of the UTLS. For the first time, a multimodel assessment with focus on the extratropical UTLS has been performed within phase 2 of the Chemistry–Climate Model Validation activity (CCMVal-2) of the World Climate Research Programme’s Stratospheric Processes and their Role in Climate (SPARC) project. The tropical UTLS and global UTLS trends have been assessed in a companion paper by Gettelman et al. [2010].

The focus of the multimodel assessment presented here is the extratropical UTLS, which is here defined as the region between the mid-troposphere (approx. 5 km) and the upper boundary of the tropically controlled transition region (around 22 km [Rosenlof et al., 1997]). It includes the lowermost stratosphere (LMS), the region between the extratropical tropopause and the 380 K potential temperature surface [Holton et al., 1995], which is equivalent to the stratospheric part of the ’middleworld’ [Hoskins, 1991]. Chemical trace gas distributions in the extratropical UTLS are shaped by the tropopause and determined by transport processes on various time and length scales. The effect of transport is greatest on the shorter-lived greenhouse gases water vapor and ozone (with lifetimes of up to one year), and much less on the longer-lived, well-mixed greenhouse gases such as CH$_4$ or CFCs (with lifetimes of several decades). Thus, it is most critical to characterize the distributions of ozone and water vapor. A defining characteristic of the LMS is that isentropes intersect the tropopause, thereby potentially connecting the troposphere and the stratosphere via rapid quasi-adiabatic motion. The slower diabatic circulation in the stratosphere (the Brewer-Dobson circulation) is predominantly downward in the extratropics, which on its own would transport aged stratospheric air into the LMS. However, meridional mixing from the tropical UTLS transports younger air masses to mid and high latitudes and ‘rejuvenates’ air as it slowly descends into the LMS [Rosenlof et al., 1997; Bregman et al., 2000; Hegglin and Shepherd, 2007], an effect quantified by Hoor et al. [2005] and Bönisch et al. [2009]. The lower boundary of the LMS is defined by the tropopause, which is here defined by a dynamical quantity. Distributions of chemical tracers that are affected by transport exhibit strong spatial gradients across the tropopause in a layer of finite depth referred to as the extratropical tropopause transition layer (ExTL) [Fischer et al., 2000; Zahn et al., 2000; Hoor et al., 2002]. The ExTL is a global feature with increasing depth towards high latitudes, and has been found to be different for different tracers [Hegglin et al., 2009]. The transition has been interpreted as the result of recurrent wave-breaking events, forced by synoptic-scale baroclinic disturbances, which stir tropospheric and stratospheric air masses with very different chemical and radiative characteristics [Shepherd, 2007]. Indeed, Berthet et al. [2007] have shown that stratosphere-troposphere exchange events traced by Lagrangian backward trajectories calculated from large-scale wind fields of the ECMWF reanalysis reveal the layer of tropospheric influence just above the tropopause. Small-scale processes such as three-dimensional turbulence and ultimately molecular diffusion then act to reduce these inhomogeneities.

The radiative properties of the UTLS are determined by the distributions of greenhouse gases, aerosols, and clouds. The low temperatures in this region lead to a net radiative cooling that is close to zero around the tropopause, which implies a strong temperature sensitivity to radiative changes [Clough and Iacono, 1995]. Moreover the large contrast between the low temperatures around the tropopause and the higher temperatures at the surface means that changes in the radiative properties of the tropopause region make a particularly strong contribution to the greenhouse effect. This sensitivity has been quantified by Forster and Shine [1997], who found that the Earth’s surface temperature response to an ozone perturbation is maximized when the perturbation is introduced around the tropopause.

The dynamical properties of the UTLS are directly dependent on the radiative properties of this region since the prevailing latitudinal temperature gradients affect the strength and location of the subtropical jet, which organizes eddy fluxes and surface pressure distributions [Randel and Held, 1991]. These eddy fluxes appear to play a key role in stratosphere–troposphere coupling [Baldwin and Dunkerton, 2001; Thompson et al., 2006], and suggest a way in which stratospheric processes may affect tropospheric weather and regional climate [e.g., Thompson and Wallace, 1998].

Chemical, radiative, and dynamical processes are all important for maintaining the structure in chemical tracer distributions and physical quantities in the UTLS. These structures can therefore be used to validate indirectly the CCMs’ representation of these processes. In this multimodel evaluation, we focus in a first step on how well the CCMs represent the main dynamical and chemical climatological structures of the UTLS. The evaluation is only capable of revealing the weaknesses or strengths of the models; it does not give us the reasons why a model performs well or badly, since these reasons are invariably model-specific. In Section 2 we introduce the participating models and the observations used for the comparisons. Section 3 describes the different diagnostics and metrics used to qualitatively and quantitatively gauge the models’ performance in reproducing key features observed in the extratropical UTLS. The results are shown in Section 4, before we come...
to the discussion of the performance of each model in Section 5, and a summary and recommendations in Section 6.

2. Models and Observations

2.1. Models

[7] Eighteen chemistry-climate models (CCMs; see Table 1) are evaluated in this multimodel assessment focusing on the extratropical UTLS. All the CCMs used here participated in the CCMVal-2 intercomparison [Eyring et al., 2008]. The CCMs are fully interactive models with a comprehensive stratosphere which aim to represent the coupling between chemistry and climate in order to simulate and predict the evolution of the stratospheric ozone layer over the past 50 years and the 21st century. For this purpose different past and future long-term simulations have been run using specified greenhouse gas and halogen scenarios. Details on the model specifications and the simulations are given by Morgenstern et al. [2010]. Note that CMAM is the only model coupled to a dynamical ocean. For the diagnostics presented in this study, we used model data obtained from past simulations extending from 1960 to 2007 and using observed SSTs (REF-B1; see Morgenstern et al. [2010] for a detailed explanation of the simulation setup).

2.2. Observations

[8] Observations of chemical species in the UTLS are still relatively sparse considering the large temporal and spatial variations and gradients in tracer concentrations in this region. In-situ observations are difficult to obtain due to the low pressures and temperatures. Satellite measurements in the upper troposphere are often obscured by clouds, and are moreover subject to significant spatial smearing. For this reason, different observations had to be compiled and validated prior to their use in this multimodel validation effort.

2.2.1. Aircraft Data

[9] Aircraft observations are generally characterized by high accuracy, high precision, and high resolution data in the UTLS, but are restricted in their representativeness due to limited sampling in time and space. The data set covers a broad altitude range up to 22 km. The spatial coverage ranges over all latitudes in the NH for most of the four seasons, but coverage is predominantly over North America and Europe. The precision and accuracy of the ozone data are ±5%. CO observations taken by different instruments have a precision of <1% and an accuracy of <3%. The precision of H$_2$O data is estimated to be <5% and the accuracy is between 0.5 ppmv and values of 10% depending on the instrument. A key purpose of the aircraft climatology is to serve as a tool to evaluate the representation of chemistry and transport by CCMs in the UTLS.

[10] A subset of those high-resolution and high-precision observations used in this paper separately has been provided by the German SPURT (Trace Gas Transport in the Tropopause Region) aircraft campaign [Engel et al., 2006]. The campaign consisted of 8 deployments distributed seasonally...
over the course of three years (2001–2003), with a total of 36 flights, each yielding around 2–5 hours of observations. The flights were carried out between around 35°N and 75°N over Europe and reached potential temperature levels between 370 K and 375 K. CO typically showed a total uncertainty of 1.5% [Hoor et al., 2004].

[12] The other subset of high-resolution data is from the NASA POLARIS (Photochemistry of Ozone Loss in the Arctic Region in Summer) campaign [Newman et al., 1999]. During the campaign, 35 flights using the NASA ER-2 research aircraft were deployed from three locations at mid- to high-latitudes with flights covering latitudes between \(\pm 20^\circ\) and \(70^\circ\), a vertical range of 5 to 18 km, and March to September 1997. A large suite of in-situ measurements was made on board the ER-2. For the \(O_3\) and \(H_2O\) data used in this study, the estimated accuracies are \(\pm 3\%\) and \(5\%\), respectively [Hintsa et al., 1999]. The use of these data to characterize the ExTL has been described by Pan et al. [2004, 2007].

2.2.2. Satellite Data
[13] Satellite instruments have recently achieved the technological maturity to measure the UTLS from space, offering unprecedented temporal and spatial coverage of this region. Investigating the accuracy and precision of these measurements is the focus of intensive validation efforts. We describe here the data from four different satellite instruments used in this paper.

2.2.2.1. ACE-FTS on SCISAT-1
[14] The Atmospheric Chemistry Experiment Fourier Transform Spectrometer (ACE-FTS) on Canada’s SCISAT-1 satellite features high resolution (0.02 cm\(^{-1}\)) and broad spectral coverage in the infrared (750 to 4400 cm\(^{-1}\)) [Boone et al., 2005; Bernath et al., 2005]. The instrument has operated since February 2004 in solar occultation mode providing seasonally varying coverage of the globe, with an emphasis on mid-latitudes and the polar regions. Up to 30 occultation events occur per calendar day. The very high signal-to-noise ratio characterizing the ACE-FTS infrared spectra makes it possible to measure more than 30 chemical trace gas species with high accuracy and precision throughout the stratosphere and lower mesosphere [Clerbaux et al., 2008; Dupuy et al., 2009], and also in the UTLS [Hegglin et al., 2008]. This, together with vertical sampling ranging from about 3 km to less than 1 km in the UTLS, provides the first global view of tracer distributions in the extratropical tropopause region [Hegglin et al., 2009].

2.2.2.2. Aura MLS
[15] The Microwave Limb Sounder (MLS) on the EOS Aura satellite measures millimeter- and submillimeter-wavelength thermal emission from the limb of Earth’s atmosphere [Waters et al., 2006]. Aura MLS has data coverage from 82°S to 82°N latitude on every orbit, providing comprehensive information on UTLS tracer distributions. Vertical profiles are measured every 165 km along the suborbital track and have a horizontal resolution of \(\approx 200–300\) km along-track and \(\approx 3–9\) km across-track. Vertical resolution of the Aura MLS data is typically \(\pm 3–4\) km in the lower and middle stratosphere [Livesey et al., 2007]. \(O_3\) has been used successfully in studies to examine transport in the UTLS, although some biases still exist in the version 2.2 which is used in the evaluations presented here. Validation of UTLS \(O_3\) is discussed by Livesey et al. [2008].

2.2.2.3. MIPAS
[16] MIPAS is a limb-viewing Fourier transform emission spectrometer covering the mid-infrared spectral region between 685 and 2410 cm\(^{-1}\) [Fischer et al., 2008] on board Envisat in a sun-synchronous polar orbit. MIPAS has provided data since 2002 at about 1000 geo-locations per day from pole to pole during day and night. It covers the atmosphere from the upper troposphere to the mesosphere (6 to 70 km), thus MIPAS provides global distributions of a large number of species. In its original observation set-up from July 2002 to March 2004 it measured one limb radiance profile every 500 km along track with a vertical sampling of 3 km and a spectral resolution of 0.035 cm\(^{-1}\). Validation of these data products is provided by Milz et al. [2005, 2009], Wang et al. [2007], and Steck et al. [2007]. Since January 2005, the observation set-up has been changed to slightly reduced spectral resolution (0.0625 cm\(^{-1}\)), but improved vertical (1.5 km) and horizontal along-track (400 km) sampling. Description of these data products is given by von Clarmann et al. [2009]. In this study we use MIPAS \(O_3\), \(H_2O\), and \(HNO_3\) observations which have been processed at the Institute for Meteorology and Climate Research (IMK) [von Clarmann et al., 2003].

2.2.2.4. Global Positioning System Radio Occultation Data
[17] The Global Positioning System Radio Occultation (GPS RO) data used in this study were obtained from the COSMIC/FORMOSAT-3 (Formosa Satellite Mission 3) mission, which is a collaborative project between Taiwan and the United States [Anthes et al., 2008]. The mission placed six micro-satellites in different orbits at 700–800 km above the ground. These satellites form a low-orbit constellation that receives signals from US GPS satellites, providing approximately 2500–3000 soundings per day almost evenly distributed over the globe. The mission has a relatively short data record since its mission launch was only in 2006. In this study, we use GPS RO retrieved temperature data between 2006 and 2009.

2.2.3. Meteorological Reanalyses
[18] For the comparisons of dynamical fields we use meteorological reanalyses such as the ERA-40 data set from the European Centre for Medium Range Weather Forecasts (ECMWF) and NCEP from the National Centers for Environmental Prediction and National Center for Atmospheric Research. ERA-40 and NCEP both have a horizontal resolution of \(2.5^\circ \times 2.5^\circ\), and exhibit 8 and 6 vertical layers in the UTLS between 300 and 100 hPa, respectively. For a more detailed description of these data sets we refer to Uppala et al. [2005], Kalnay et al. [1996], Randel et al. [2003], and SPARC [2002].

3. Diagnostics and Performance Metrics
[19] Diagnostics and performance metrics are used to evaluate model performance in the UTLS in a qualitative and quantitative way, respectively.

3.1. Diagnostics
[20] The diagnostics are chosen to evaluate the main characteristics of dynamics and transport in the UTLS in the models. The main characteristics include the seasonality in the background LMS tracer distributions and LMS mass...
determined by the seasonally varying strength of the Brewer-Dobson circulation, as well as the fine-scale structure of the transition between the troposphere and the stratosphere determined by synoptic and smaller scale eddies. Both these classes of characteristics are important for determining surface climate and stratosphere-troposphere coupling, and can be tested with different diagnostics. The full list of diagnostics used for the validation of the CCMVal-2 models in the extratropical UTLS is given in Table 2. The following sections provide in addition short descriptions of each diagnostic:

3.1.1. Diagnostic 1
[21] The seasonal zonal mean zonal wind is used to test the models’ realism in representing the meridional gradients of the thermal structure.

3.1.2. Diagnostic 2
[22] The seasonal cycle of LMS mass ([Appenzeller et al., 1996]) is a basic test of the response of the atmosphere to a combination of direct radiative forcing through changes in UTLS temperatures as well as of dynamical forcing through the wave-driven Brewer-Dobson circulation. It can be seen as a basic measure of the vertical structure in a model.

3.1.3. Diagnostic 3
[23] Interannual anomalies in extratropical tropopause pressure are used as a measure of the response of the models to different forcings such as volcanoes, ENSO, etc. They are strongly related to the LMS mass.

3.1.4. Diagnostic 4
[24] The tropopause inversion layer (TIL) ([Birner et al., 2002; Birner, 2006]) is a distinctive feature of the thermal structure of the tropopause, which reflects the balance between radiative and dynamical processes. The formation of the TIL has been interpreted as a result of large-scale dynamics or convection, but also seems to be forced or maintained through the distributions of the radiatively active species O$_3$ and H$_2$O. This diagnostic extends diagnostic 6 from [Gettelman et al., 2010] to the extratropics.

3.1.5. Diagnostic 5
[25] The seasonal cycles of O$_3$, HNO$_3$, and H$_2$O at 100 and 200 hPa are used to test the models’ representation of the seasonally varying strength in large-scale transport through the diabatic Brewer-Dobson circulation, and in quasi-horizontal transport (and subsequent mixing) between the tropics and the extratropics within the tropically controlled transition region (380–420 K, or ≈100 hPa) or across the subtropical jet (340–380 K, or ≈200 hPa).

3.1.6. Diagnostic 6
[26] The meridional gradient of O$_3$ is a measure of the chemical distinctiveness (i.e., different air masses can be readily distinguished by their chemical composition) of the UTLS in latitude, and therefore of the degree of isolation of different regions such as the tropics and the extratropics.

3.1.7. Diagnostic 7
[27] Vertical profiles of normalized CO in tropopause coordinates (here potential temperature units relative to the potential temperature at the 2 PVU tropopause) allow us to separate transport into the LMS across the extratropical tropopause on short time scales from transport from the tropical and subtropical UT on longer time scales. It thereby helps to determine the tropospheric influence on the lowermost stratospheric background. The CO is normalized with respect to the mean tropospheric value, in order to remove discrepancies between model and measurements arising from the different tropospheric boundary specification of CO used in the models.

3.1.8. Diagnostic 8
[28] Mean annual profiles of H$_2$O and CO in tropopause coordinates (here kilometers relative to the thermal tropopause height) at mid-latitudes and northern hemisphere polar regions provide a critical diagnostic for testing the degree of vertical distinctiveness of the UT and LS, as well as to obtain information on chemical processes that determine the tracer profiles.

---

**Table 2. Diagnostics Used for the Multimodel Assessment**

<table>
<thead>
<tr>
<th>Process</th>
<th>Diagnostic</th>
<th>Variables</th>
<th>Observations</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamics</td>
<td>Zonal mean zonal wind @200 hPa$^a$</td>
<td>U (Zonal Wind)</td>
<td>ERA-40, NCEP</td>
<td>Appenzeller et al. [1996]</td>
</tr>
<tr>
<td></td>
<td>in LMS mass$^b$</td>
<td>M (Mass)</td>
<td>ERA-40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TP pressure anomalies</td>
<td>P (Pressure)</td>
<td>ERA-40, NCEP</td>
<td>Birner et al. [2002], Birner [2006], Randel et al. [2007]</td>
</tr>
<tr>
<td></td>
<td>TP inversion layer</td>
<td>T (Temperature)</td>
<td>GPS</td>
<td></td>
</tr>
<tr>
<td>Transport and mixing</td>
<td>Seasonal cycle of tracers @100 and 200 hPa$^a$</td>
<td>O$_3$, HNO$_3$, H$_2$O</td>
<td>MIPAS, MLS</td>
<td>Shepherd [2002]</td>
</tr>
<tr>
<td></td>
<td>Meridional tracer gradients @200 hPa</td>
<td>O$_3$</td>
<td>ACE-FTS</td>
<td></td>
</tr>
<tr>
<td></td>
<td>relative to TP$^c$</td>
<td>CO</td>
<td>MLS</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vertical profiles in TP coordinates$^c$</td>
<td>H$_2$O and CO</td>
<td>Aircraft, ACE-FTS</td>
<td>Pan et al. [2004, 2007]; Hegglín et al. [2009]</td>
</tr>
<tr>
<td></td>
<td>ExTL depth from tracer-tracer correlations</td>
<td>H$_2$O/O$_3$</td>
<td>Aircraft</td>
<td>Pan et al. [2007]; Hegglín et al. [2009]</td>
</tr>
</tbody>
</table>

$^a$TP denotes tropopause.

$^b$Additional literature with information on the data sets or the diagnostics.

$^c$Diagnositics which are used for final grading of the CCMs.
3.1.9. Diagnostic 9

The depth of the extratropical tropopause transition layer (ExTL) based on the tracer–tracer correlation method using H₂O–O₃ by Pan et al. [2007] and its location relative to the thermal tropopause are used to diagnose the mixing and transport characteristics of the models in the tropopause region.

3.2. Performance Metrics and Grading

Performance metrics provide statistical measures of agreement between the model field and observational data sets. The following metrics are used in this study to quantitatively assess the ability of the CCMs to reproduce patterns and mean state of the extratropical UTLS. While errors in the mean are simply defined (see Section 3.2.2), errors in the patterns are more complicated since they can arise as a result of errors in either phase or amplitude. How can we visualize this information in a simple way?

3.2.1. Taylor Diagrams

An answer to the above question is provided by Taylor diagrams (Taylor [2001] and Figure 1). Taylor diagrams provide a statistical summary of how well two patterns from a test (i.e. the model) field \( f \) and a reference (i.e. the observational) field \( r \) of the same quantity match each other in terms of their correlation \( R \), their pattern root-mean-square (RMS) difference \( \bar{E} \), and the ratio of their standard deviations \( \sigma_f/\sigma_r \). Taylor diagrams have been widely used to test various aspects of model performance, as for example in the IPCC TAR [Intergovernmental Panel on Climate Change, 2001] or in chemistry transport model inter-comparisons [Brunner et al., 2003, 2005]. The Pearson’s correlation \( R \) used here as the first metric is given by

\[
R = \frac{1}{N} \sum_{n=1}^{N} (f_n - \bar{f})(r_n - \bar{r}) \quad (1)
\]

For seasonal cycles \( n \) denotes the month of the year and \( N \) is 12, while for latitudinal structure \( n \) indexes the latitude and \( N \) is the total number of latitude bins. For the calculation of the spatial correlation, the model data need to be interpolated onto the latitude of the observations. In statistics the quantification of the difference between the two fields \( r \) and \( f \) is most often given by the root-mean-square (RMS) difference

\[
E = \sqrt{\frac{1}{N} \sum_{n=1}^{N} [f_n - r_n]^2} \quad (2)
\]

and the full-mean square difference between \( f \) and \( r \) by

\[
E^2 = E^2 + E_f^2 \quad (3)
\]

where

\[
E = f - r \quad (4)
\]

is the mean bias and

\[
E_f = \sqrt{\frac{1}{N} \sum_{n=1}^{N} [f_n - f - (r_n - \bar{r})]^2} \quad (5)
\]

quantifies the magnitude of the mismatch between the two patterns (the pattern RMS difference). Normalizing \( E^2 \) by \( \sigma_r \), i.e., \( \bar{E} = E^2/\sigma_r \), it is easily shown that

\[
\bar{E}_f^2 = \left( \frac{\sigma_f^2 + \sigma_r^2 - 2\sigma_f \sigma_r R}{\sigma_r^2} \right)^2 = \frac{\sigma_f^2}{\sigma_r^2} + 1 - 2\frac{\sigma_f}{\sigma_r} R \quad (6)
\]

so that \( \bar{E}_f \) is determined once \( \sigma_f/\sigma_r \) and \( R \) are known. Smaller \( \bar{E}_f \) (i.e. the closer \( f \) and \( r \) are to each other) represents a better fit between the test field and the reference field. The Taylor diagram uses \( \sigma_f/\sigma_r \) and \( R \) as radial and azimuthal coordinates, respectively. By scaling \( R \) by its inverse cosine, it turns out that \( \bar{E}_f \) exactly corresponds to the distance in the diagram from the reference point [Taylor, 2001].

An illustrative example of a Taylor diagram is provided in Figure 2. Different test fields (\( T1, T2, \ldots, T5 \), in color) are depicted together with a reference field (in black) in Figure 2 (left). The pattern statistics between the test and the reference field are then transferred into the Taylor diagram on the right. \( T1 \) has the exact phase and amplitude as the reference field, and therefore lies on top of the reference field in the Taylor diagram, despite the fact that the test field is offset by 2 units in the \( y \)-direction. Taylor diagrams do not provide information on the mean error, and we need therefore to have an additional grading of the mean to get full information on the model performance (see Section 3.3.2). \( T2 \) and \( T3 \) have the exact phase (and therefore maximal correlation \( R \)), but different amplitudes which correspond to normalized standard deviations greater and less than unity.
moving them further away and closer to the origin, respectively. It follows that $E'$ increases. $T4$ and $T5$ have both different phases and different amplitudes, which decreases the correlation $R$ and increases $E'$ even further.

[33] From Figures 1 and 2 it is clear that a certain value of $E'$ can be obtained by different combinations of $R$ and $\sigma_f/\sigma_r$. How do we gauge the relative skills of two models that obtain the same $E'$ but where one has a better phase and the other a better amplitude? For this purpose, we can define a skill score ($S$) following Taylor [2001] which ensures that at a fixed phase or amplitude the skill increases monotonically with improvements in amplitude or phase, respectively, and with values between 0 and 1:

$$S = \frac{4(1 + R)}{(\hat{\sigma} + 1/\hat{\sigma})^2(1 + R_0)}$$  

(7)

Here, $\hat{\sigma}$ ($\sigma_f/\sigma_r$) denotes the standard deviation of the test field normalized by the standard deviation of the reference field, and $R_0$ is the maximum correlation that models can achieve. Choosing $R_0 < 1$ allows us to account for uncertainty in the observations or model limitations such as spatial and temporal resolution. The skill ($S$) approaches unity as the model variance approaches the observed variance (i.e., as $\hat{\sigma} \rightarrow 1$) and as $R \rightarrow R_0$. The skill score is indicated with gray lines in Figures 1 and 2.

### 3.2.2. Climatological Mean State Metrics

[34] As mentioned in the previous section, Taylor diagrams do not yield information on how close the mean of a given test field is to that of the reference field. We therefore use in addition to the skill factor as defined in equation (7), a grading for the mean values as introduced by Douglass et al. [1999], which also has been applied within CCMVal-1 [Waugh and Eyring, 2008] and by Gettelman et al. [2010]:

$$g_m = \max \left( 0, 1 - \frac{1}{N} \sum_{j=1}^{N} \left| \mu_{obs} - \mu_{mod} \right| n_g \sigma_{obs} \right)$$  

(8)

Here, $n_g$ is a scaling factor (usually taken to be 3 if not specified differently), $\mu$ the monthly mean values from the observations and the models at all latitudes for all months ($N$), and $\sigma_{obs}$ the standard deviation of the measurements. Note that $n_g$ and $\sigma_{obs}$ are often not straightforwardly determined, and therefore involve subjective choices. This needs to be taken into account when interpreting grades from this metric, as emphasized by Waugh and Eyring [2008] and Grewe and Sausen [2009].

#### 3.2.3. Combining Pattern and Mean State Metrics

[35] The final grade ($G_{tot}$) of model performance for each diagnostic is then defined by the linear combination of the skill score ($S$; see equation (7)) and the grading of the mean ($g_m$; see equation (8)):

$$G_{tot} = (S + g_m)/2.$$  

(9)

[36] The $G_{tot}$ of the different diagnostics are then listed in a summary matrix (Figure 19). We will derive total grades only for the diagnostics 1, 2, and 5 to 8 of Section 3.1, since these diagnostics are easy to define or more established in the literature than diagnostics 3, 4, and 9. The latter diagnostics therefore have to be seen as ‘experimental’ diagnostics.

### 4. Results

#### 4.1. Dynamical Structure of the Extratropical UTLS

#### 4.1.1. Zonal Mean Zonal Wind at 200 hPa

[37] The zonal mean zonal wind field is used to validate the representation of the thermal structure of the models, and therefore the basic dynamical state of the models’ atmospheres. For this diagnostic, monthly zonal mean wind fields averaged over the period 1979–1999 and for JJA (June–July–August) and DJF (December–January–February) are compared between the model simulations and ERA-40 re-analyses. NCEP re-analyses are also included for further comparison with ERA-40 and the models.
Figures 3, 4, and 5 illustrate that the models represent the strength and latitudinal behavior of the observed zonal mean zonal wind in a realistic way. This is to be expected since the zonal mean wind fields have been used extensively as a diagnostic to help improve gravity wave parameterizations in model development. ULAQ is the only model that shows clear deficiencies in resolving the latitudinal structure, especially during JJA. This lack of realism is also expressed in the Taylor diagrams by very low latitude-by-latitude correlation and skill values. ULAQ also scores very low in the climatological mean state metric, with 0.5 during DJF and 0.3 during JJA. The low performance of ULAQ might be attributable to the very low resolution of the model and the quasi-geostrophic dynamical core characterizing this model. In the mean state metric, the SOCOL-based models score slightly lower than the multimodel mean during both DJF and JJA.

The tight correspondence between NCEP and ERA-40 (the skill of NCEP is 0.98), with the models lying much farther away from the reference point, indicates that the reanalyses are consistent with each other and that the models may still have room for improvement. For example, several models displace the tropospheric ‘eddy-driven’ jet in the SH summer (DJF) when compared to the observations.

The total grading values ($G_{tot}$) obtained in Figures 4 and 5 for the two seasons are averaged and listed in the summary matrix (Figure 19).

### 4.1.2. Seasonal Cycle of LMS Mass

The seasonal cycle of the LMS mass is a basic test of the response of the atmosphere to a combination of direct radiative forcing through changes in UTLS temperatures as well as of dynamical forcing through the wave-driven Brewer-Dobson circulation, and represents an integrated measure for the tropopause behavior. Stratospheric mass variations due to seasonal tropopause height variations can contribute to stratosphere-troposphere exchange [Appenzeller et al., 1996]. This exchange transports ozone, reactive nitrogen, and other species into the troposphere, where it affects the tropospheric ozone budget and with it air quality. We test here the realism of the seasonal variations of the total LMS mass by comparing them to the NCEP reanalyses using a method similar to that of Appenzeller et al. [1996]. The LMS mass is defined as the mass of the stratospheric region that lies between the thermal tropopause, calculated using the WMO definition [World Meteorological Organization, 1957], and the 100 hPa pressure surface. The thermal tropopause is derived from monthly zonal mean temperature fields averaged over a time period between 1990 and 1999. Note that the 100 hPa pressure level, unlike the 380 K potential temperature level, does not reflect the seasonal shifts in the upper boundary of the LMS. The results are shown in Figure 6.

In the NH, most models show a very high skill (with values larger than 0.9) in reproducing amplitude and phase of the seasonal cycle of the LMS mass from the NCEP reanalyses. One exception is LMDZrepro which only scores 0.62. LMDZrepro’s variability is too low (with a standard deviation of about 0.5), however the model captures the seasonal evolution (correlation of 0.95). There are also quite a few models that have difficulty in simulating accurate mean values of the LMS mass as compiled in Figure 7. UMUKCA-METO and UMUKCA-UCAM exhibit too large LMS masses, indicating an average tropopause pressure that is too high. CCSRINES, CNRM-ACM, EMAC, Niwa-SOCOL, SOCOL, and ULAQ have smaller LMS mass values than
NCEP, indicating generally too low tropopause pressures. These findings are consistent with the evaluation of the annual mean tropopause heights by Gettelman et al. [2010], which show that the SOCOL-based (UMUKCA-based) models exhibit too low (high) tropopause pressures in both the tropics and the extratropics.

In the SH, the models’ overall performance is worse than in the NH. The skill based on the correlative metrics lies around 20–40% lower than in the NH for all models, with particular deficiencies for CAM3.5, CCSRNIES, GEOSCCM, ULAQ, UMSLIMCAT, and WACCM. The Taylor diagram reveals that almost all models exhibit standard deviations that are too large, which moves them further away from the reference point (note the different radial axis scale in the Taylor diagrams in Figure 6). The following models have major deficiencies in representing the mean values (see Figure 8): CNRM-ACM, CCSRNIES, UMKCA-METO, and UMKCA-UCAM. The differences between SH and NH are likely due to the much less pronounced seasonality seen in the SH, but may also stem from the fact that the reanalyses the models are being compared to are less consistent and therefore less reliable in the SH than in the NH [Schoeberl, 2004].

The total grading values ($G_{tot}$) obtained in Figures 7 (for the NH) and 8 (for the SH) are averaged and listed in the summary matrix (Figure 19). The models that perform the best (showing total scores [$G_{tot}$] larger than 0.8) are AMTRAC3, CMAM, E39CA, GEOSCCM, and MRI, which are the models that also reach the highest scores in the NH.

### 4.1.3. Tropopause Pressure Anomalies

The extratropical tropopause pressure is a basic measure of the vertical structure in a model. We focus here on interannual anomalies in the monthly and zonal mean tropopause pressure that yield insight into the models’ abilities to respond to forcing of the climate system. While the use of this diagnostic has proven to be useful in the tropics [Gettelman et al., 2010], we may not necessarily expect the same strong connection between sea surface temperatures and the tropopause pressure in the extratropics. This diagnostic has therefore to be seen as more explorative, for which reason we do not use it as a metric in the final evaluation.

The tropopause is calculated using the WMO-definition and averaged over 40°N–60°N and 40°S–60°S, respectively. The analysis is based on monthly mean temperature fields. The models are compared to five different analyses: ERA Interim, NCEP, NCEP2, JRA25, and ERA-40.

Although the models seem to reproduce the seasonal cycle of tropopause pressure well in the NH (based on the diagnostic for the LMS mass in Section 4.1.2), they show more problems in representing interannual variability. This can be seen from Figure 9. CNRM-ACM has an unrealistically large interannual variability and a low tropopause pressure, as also has been noted in the corresponding analysis in the tropics [Gettelman et al., 2010]. The excessive variability is due to a large signal from volcanic aerosol heating (note the correspondence of anomalies with recent major volcanic events). CCSRNIES, EMAC, ULAQ and WACCM
perform worst among the models due to both too high or low mean values and very small correlation with the observed variability structure. The low correlation is most obvious during years that show volcanic events, where these models do not show the expected drops in pressure.

In the SH, the models simulate the interannual variability somewhat better, except for CNRM-ACM which exhibits a too large interannual variability and a too low tropopause pressure as found also in the NH. CCSRNIES, MRI, ULAQ, and WACCM have a negative bias in the mean tropopause pressure, and ULAQ shows the worst correlation with the reanalyses.

4.1.4. Tropopause Inversion Layer

A strong temperature inversion with a depth of a few kilometers and located just above the tropopause (also called the tropopause inversion layer, TIL), has been extensively investigated using high-resolution radiosondes [Birner et al., 2002; Birner, 2006; Bell and Geller, 2008] and Global Positioning System (GPS) Radio Occultation (RO) data [Randel et al., 2007; Grise et al., 2010]. The TIL is characterized by a sharp and strong maximum in static stability ($N^2 = \frac{g}{C_1^8 \frac{d\theta}{dz}}$). The presence of the TIL is believed to be important for the impact of cross-tropopause exchange on chemical tracer distributions, wave-breaking and wave-generation [Miyazaki et al., 2010a], and for the dynamical coupling between the stratosphere and the troposphere.

$N^2$ has been calculated for 9 different models for which the necessary instantaneous data fields (i.e. on model levels) were available and compared to the COSMIC (Constellation Observing System for Meteorology, Iono-sphere, and Climate) GPS-RO data. Three years of data were used between 2006 and 2009, with up to 3000 profiles per day. The analysis is performed in tropopause coordinates using the tropopause pressure as reference ($p_{\text{TP}}$) and with height $z = -H \ln(p/p_{\text{TP}})$, where $H$ is a scale height of 8 km. The observed TIL is computed using both data at full vertical resolution and data only at CCMVal-2 standard levels, so as to ensure a fair comparison between the high resolution observations and the lower resolution model fields. The use of degraded observations reduces differences due to the models’ low vertical resolution, and the remaining differences may then be attributed more likely to missing processes in the models.

Zonal mean cross-sections and tropical profiles of $N^2$ are discussed in detail by Gettelman et al. [2010]. Here, we discuss in addition how models perform in simulating $N^2$ profiles at two latitude bands, representing the Northern Hemisphere TIL in winter and summer (Figure 10). In general, the models seem to be capable of reproducing the qualitative structure and seasonality of the TIL. The TIL has a weaker maximum during winter and a stronger maximum.
during summer, which increases also with latitude [Birner, 2006; Randel et al., 2007].

It can also be seen that maximum values of simulated $N^2$ are comparable to or larger than those derived from degraded GPS data. However, they are always weaker than those computed from full-level GPS data unless vertical resolution is sufficiently high as seen for the high-resolution version of WACCM with a vertical resolution of 300 m in the UTLS region. It is also evident that the location of maximum $N^2$ in the CCMVal-2 models is found always at somewhat greater distances from the tropopause and to be broader than in the full GPS data, but closer to that seen in the degraded GPS data. Again, the high-resolution WACCM version yields better results with a maximum $N^2$ location closer to the observed one, although the maximum is still too broad. Some of the shift in the $N^2$ maximum can therefore be explained by the limited resolution of the CCMs, and does not necessarily mean that they misrepresent the processes determining the TIL structure. Another more detailed comparison between a high-resolution (T213L256) and a low-resolution (T41L32) version of a global circulation model (GCM), which supports this conclusion, is given by Miyazaki et al. [2010b]. They show that the high-resolution version of their model (given all other model settings to be equal) was capable of reproducing the fine-scale structure and seasonality of the TIL much better than the low-resolution version, with the maximum in $N^2$ being more realistic and located closer to the tropopause.

The above comparisons indicate that the CCMs are qualitatively reproducing the TIL, but underestimate the strength of the TIL quantitatively, most likely due to their limited vertical resolution. A too weak TIL is likely to result in too weak potential vorticity gradients across the tropopause, which in turn may lead to too strong transport across the tropopause and may also alter wave-propagation into the stratosphere or wave-generation in the tropopause region.

4.2. Transport and Mixing Within the Extratropical UTLS

4.2.1. Seasonal Cycle of Tracers at 100 and 200 hPa

The large-scale Brewer-Dobson circulation consists of two main branches. The deeper branch is driven mainly by planetary wave breaking in the stratosphere and transports aged stratospheric air into the LMS. The shallow branch is driven by the breaking of both synoptic scale and planetary waves above the subtropical jet and transports (and ultimately mixes) younger tropical air masses to higher latitudes. Both transport processes exhibit a seasonally varying strength, and determine the chemical background composition of the LMS. It is crucial for CCMs to capture the relative strength and seasonality of these processes, since they determine the distribution of $O_3$ and $H_2O$ in the LMS (which through radiative heating have a strong impact on the temperature distribution and therefore on winds), and also the monthly input of stratospheric ozone into the troposphere.

The models’ representation of these large-scale transport and mixing processes, which take place on time-scales of weeks to a couple of months, is evaluated here using the seasonal cycles in $O_3$, $HNO_3$, and $H_2O$ at 100 and 200 hPa averaged between 40° and 60°N and S, respectively. While $O_3$ and $HNO_3$ are expected to exhibit similar seasonal cycles since their sources in the UTLS are mostly...
stratospheric, H$_2$O is a tropospheric tracer in the UTLS (since the contribution of CH$_4$-oxidation to the H$_2$O budget is small) and gives insight into a possible tropospheric influence. HNO$_3$ is further affected by chemistry and microphysics, which may cause some differences to the structure seen in the O$_3$ seasonal cycle. The monthly mean zonal mean tracer fields from 2000–2006 are compared to observations obtained by MIPAS between 2004 and 2008, MLS during 2006, and the ACE-FTS between 2004 and 2008.

The top two rows in Figures 11 and 12 show the results for the 100 hPa level with the corresponding Taylor diagrams in the NH and SH, respectively. In the NH, O$_3$ is relatively well represented in all the models despite a tendency to overestimate the mean (not shown) and the amplitude (i.e. standard deviation) of the seasonal cycle relative to MIPAS observations. The models’ variability is also too strong when compared to MLS observations; however the correlation improves slightly. The Taylor diagram also reveals slightly lower correlation values than average for CAM3.5, ULAQ, UMUKCA-UCAM, and WACCM. The seasonal cycle of HNO$_3$ mostly confirms this behavior, with the exception of UMUKCA-METO, which exhibits a very low correlation with MIPAS satellite observations. The behavior in the seasonal cycle of H$_2$O is similar to the tropics [see Gettelman et al., 2010], pointing toward a strong connection between the tropics and the extratropics. The performance of the models therefore strongly depends on their ability to represent tropical processes such as dehydration and the seasonal strength in the meridional mixing between the tropics and the extratropics. Models that score low in the tropical H$_2$O diagnostic [Gettelman et al., 2010] indeed score low also in the diagnostic presented here. The too large amplitude in O$_3$ may also have its origin in the tropics, which is supported by the finding of Gettelman et al. [2009] that most of the CCMVal-1 models’ O$_3$ in the tropics increases too quickly at and above the tropopause. This bias is somewhat less pronounced in the CCMVal-2 models, but several outliers still exist. The seasonal cycles in the SH show generally smaller amplitudes, reflecting the weaker influence of the Brewer-Dobson circulation. The models generally show the same behavior, however they tend to overestimate the mean O$_3$ values.

The results for the 200 hPa level are shown in the two bottom rows in Figures 11 and 12. At 200 hPa in the NH, the O$_3$ seasonal cycle is well represented in almost all the models. However, the mean values and amplitudes tend to be rather low compared to those in the observations. CNRM-ACM, Niwa-SOCOL, and SOCOL score worst; these models show, apart from too low amplitudes, also relatively low month-by-month correlations in comparison with the MIPAS observations. HNO$_3$ again shows a consistent behavior in almost all the models. The low mean values in both O$_3$ and HNO$_3$ can be explained by too much
Figure 11. Seasonal cycles in monthly mean (left) O₃, (middle) HNO₃, and (right) H₂O between 40°N and 60°N and corresponding Taylor diagrams at (top two rows) 100 hPa and (bottom two rows) 200 hPa for different models (color code; see Figure 6 legend) and compared to MIPAS satellite data (brown solid lines) ±1σ (gray shading) over the years 2004–2008. In addition to the MIPAS data, MLS O₃ data (brown dashed line and brown dots) and ACE-FTS H₂O data (brown diamonds) are also shown. The multimodel mean is denoted in black.
transport across the subtropical and extratropical tropopause. This is reflected also in too large amplitudes (standard deviations) in the $\text{H}_2\text{O}$ seasonal cycle, the tropospheric tracer which is most sensitive to mixing due to its exponential decrease across the tropopause. Tropospheric influence seems to be particularly high during late summer and autumn. The lack of a sophisticated tropospheric chemistry in most models may also contribute to some of the observed differences. As the analysis is done on fixed pressure levels, the model biases could in principle originate from biases in the tropopause altitude. However, this does not seem to be the case. MRI for example exhibits a too low tropopause,
but also too strong mixing, while UMSLIMCAT exhibits a too high tropopause, but not enough mixing (inferred from too high O$_3$, since H$_2$O is not available). SOCOL and Niwa-SOCOL both are too diffusive, possibly due to their semi-Lagrangian transport scheme. In the SH, the observed seasonal cycles of all the three tracers at 200 hPa show smaller amplitudes consistent with the finding on the 100 hPa level. Again, the models’ means and amplitudes (standard deviations) in O$_3$ and HNO$_3$ are shifted to values smaller than expected from the observations with H$_2$O indicating too strong cross-tropopause transport or too high tropopause temperatures. Note that there is some evidence that the seasonal cycle in MIPAS H$_2$O at 200 hPa exhibits a smaller amplitude than other satellite observations like HALOE and MLS. Indeed, comparison with the ACE-FTS measurements, which yielded a good agreement with high-resolution aircraft measurements [Hegglin et al., 2008], indicates that MIPAS H$_2$O might be somewhat low especially during summer on the 200 hPa level and in both hemispheres, a retrieval or instrument issue which is currently under investigation. However, the large noise and standard deviations in the ACE-FTS data imply that the sampling from the ACE-FTS is not sufficient to determine the seasonal cycle of H$_2$O accurately. The discrepancies between the two satellite observations reveal how problematic H$_2$O measurements in the UTLS from space are, which is mainly due to the high variability and large spatial and temporal gradients in H$_2$O found in the UTLS. At the moment, the metrics presented here are therefore not defined accurately. Additional measurements with higher (spatial and temporal) resolution and higher accuracy will be needed to resolve this issue and to gain more confidence in this metric in the future. Comparison

Figure 13. Meridional gradient in O$_3$ at 200 hPa and corresponding Taylor diagrams for (top and bottom left) JJA and (middle and bottom right) DJF. Brown thick lines and gray shading show MLS observations averaged over 2006–2008 with 1σ uncertainty.
[58] For this diagnostic we derive two different grades for each model. One grade is based on the O$_3$ seasonal cycle, and calculated as the average over all mean grades ($g_m$) and skill scores obtained for both pressure levels and hemispheres. The other grade is based on the H$_2$O seasonal cycle, calculated as the average of all skill scores obtained for both pressure levels and hemispheres. We do not include a grade for the mean value of H$_2$O, since it is already used as a metric in the vertical profiles (see Section 4.2.4). The model grades are then listed in the summary matrix (Figure 19).

### 4.2.2. Meridional Tracer Gradients at 200 hPa

[59] Useful information on mixing barriers and therefore the degree of isolation and chemical distinctness of different regions such as the tropics and the extratropics is provided by the sharpness of meridional gradients of long-lived species. Here we use the meridional gradient in O$_3$ at 200 hPa (which is long-lived compared to the transport time scales in this region). We use seasonal means for JJA and DJF derived from monthly mean zonal mean O$_3$ fields from all models and compare them to a multiyear seasonal climatology derived from MLS data (averaged over 2004–2008). Ideally one would examine the gradient of the mode of the probability density function rather than the mean [Shepherd, 2002], as this provides a more robust representation of the tracer gradient. However, this would require 3D instantaneous data which are not available for all the models.

[60] The models generally reproduce the meridional gradients in both JJA and DJF well (Figure 13), which implies that they can reproduce the separation between the tropical UT and the extratropical LMS. In JJA (Figure 13, bottom left), the correlations are mostly higher than 0.9 except for CNRM-ACM, ULAQ and UMETRAC, which show correlations between 0.5 and 0.8. However, there is a substantial spread in the models in terms of standard deviations, resulting in decreased skill scores (also see Figure 14). Too low amplitudes in the meridional gradient are found in CNRM-ACM, EMAC, the SOCOL-models, ULAQ and UMETRAC, and too high amplitudes in UMSLIMCAT and UMUKCA-METO, diminishing their skill score to values below 0.85. In DJF (Figure 13, bottom right, and Figure 15), the models perform slightly worse, especially CNRM-ACM, ULAQ, UMETRAC and UMUKCA-METO.

[61] A relation between this diagnostic and the zonal mean zonal wind might be expected, however this seems to be true only for ULAQ, which shows low grades for both the zonal mean zonal wind and for the meridional gradient. It is noteworthy that the multimodel mean produces in both seasons the most realistic picture of the meridional ozone gradient with skills reaching values of 0.94.

### 4.2.3. Normalized CO in Tropopause Coordinates

[62] To evaluate the representation of tropospheric influence on the background LMS in the models, and to distinguish between transport across the extratropical tropopause on short time scales and transport from the tropics and subtropics on longer time scales, we use CO which has an approximate lifetime of 3 months in the LMS. In the middle stratosphere above $\Theta = 500$ K, CO is nearly constant with an observed background value of 10–15 ppbv [Flocke et al., 1999], due to the chemical equilibrium between methane

---

**Figure 14.** Same as Figure 4 but for meridional gradient in O$_3$ at 200 hPa for JJA.

**Figure 15.** Same as Figure 4 but for meridional gradient in O$_3$ at 200 hPa for DJF.
Profiles of CO* (normalized CO) for winter/spring and summer/autumn in layers of dΘ = 30 K from Hoor et al. − CO* is the stratospheric background value of each model year. We assume that the choice of the model year has a negligible impact on the results of this evaluation. Data were analyzed in layers of 30 K relative to the 2 PVU surface (represented by the centered layer means at −15, 15, 45, and 75 K in Figure 16). The tropospheric fraction of CO in the stratosphere (CO*) is determined by

\[
CO^* = \frac{(CO - CO_{strat})}{(CO_{trop} - CO_{strat})} \tag{10}
\]

where CO_{strat} is the stratospheric background value of each individual model defined as the mean value within the 500–600 K layer, and CO_{trop} is the mean value for the layer between −30 and 0 K below the tropopause. The normalization using the factor 1/(CO_{trop}−CO_{strat}) accounts for the varying boundary specifications of CO in the models, which we do not want to test here. Models that did not provide instantaneous fields were not included in the comparison.

Two properties were tested and graded as follows.

1. The abundance of CO* between 30 and 60 K above the tropopause as a measure for tropospheric influence. A model was given a grade W1 of 3, 2, 1, 0 if the difference between the SPURT observations and the model was <1σ, <2σ, <3σ, and >3σ of the observational standard deviation, respectively.

2. The decreasing coupling to the local 2 PVU tropopause found in the layers 0–30 K and 30–90 K above the tropopause as represented by the different gradients in CO* in the respective layers. A model was given a grade W2 following the same scheme as in (1) but comparing the ratio of the gradients in CO* at the 60 K and 0 K levels obtained for the model and the observations.

High values for both weights (Table 3) therefore indicate a good separation from the extratropical troposphere and mainly weak influence from the subtropics (as seen for UMUKCA-METO and EMAC). Low values for W1 (abundance), but high values for W2 (separation) reflect too much tropospheric CO (CO*) in the LMS, but the transition from the local troposphere to the LMS occurs correctly within dΘ = 0–30 K (AMTRAC3 and CCSRNIES). Low values for both W1 and W2 indicate that the coupling to the extratropical tropopause extends too deep into the stratosphere (LMDZrepro and Niwa-SOCOL during winter/spring), leading to an unrealistic tropospheric contribution.

### Table 3. Grading for the Normalized CO Gradient in Tropopause Coordinates

<table>
<thead>
<tr>
<th>Model</th>
<th>Winter/Spring</th>
<th>Summer/Autumn</th>
<th>TOTAL (ΣW/12)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W1</td>
<td>W2</td>
<td>W1</td>
</tr>
<tr>
<td>AMTRAC3</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>CAM3.5</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>CCSRNIES</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>CMAM</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>CNRM-ACM</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>EMAC</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>LMDZrepro</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Niwa-SOCOL</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SOCOL</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ULAQ</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>UMUKCA-METO</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>WACCM</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
due to overestimation of transport across the extratropical tropopause. In fact, the SOCOL models score low in $W_1$ in both seasons.

In general models tend to transport too much of the tropospheric tracer into the LMS in winter as indicated by the low values of $W_1$ in Table 3. However, most models capture the separation (i.e. the change of gradient) around $d\Theta = 30$ K indicated by $W_2$. Thus, most models are able to separate between transport across the local tropopause in the extratropics and processes involving other time scales and source regions. During summer, the models are capable of capturing the increase in tropospheric influence from the subtropics. The high summer values of $W_2$ are a result of weaker differences in the vertical gradients in the two layers due to the enhanced transport from the subtropics accompanied with larger variability found in the measurements.

The best overall representation of transport and coupling is seen in CAM3.5, CMAM, CNRM-ACM, EMAC, UMUKCA-METO and WACCM, whereas LMDZrepro and Niwa-SOCOL seem to be too diffusive or permeable across the tropopause, confirming the results of the previous diagnostic using seasonal cycles. Most models tend to get the separation between the different regimes in the LMS (UT, transition layer, background LMS) right within the measurements’ variability.

### 4.2.4. Vertical Profiles in Tropopause Coordinates

The vertical structure of $H_2O$ and CO across the tropopause is evaluated using profiles in tropopause coordinates [Logan et al., 1999; Pan et al., 2004, 2007; Considine et al., 2008; Hegglin et al., 2006, 2009]. Tropopause coordinates have been shown to effectively decrease tracer variability due to geophysical variability (i.e. day-to-day variations in tropopause height) in the region of $\pm 3$ km and 5 km from the tropopause in summer/autumn and winter/spring, respectively [Hegglin et al., 2008]. The diagnostic requires instantaneous model output. For consistency with the coverage of the aircraft data [Tilmes et al., 2010], models are evaluated for the years between 1995 and 2005.

The region of analysis is chosen to be the part of the extratropics that is not strongly influenced by the subtropical tropopause break and double tropopauses. Selection criteria were: a tropopause height of $\leq 325$ K in winter and spring and $\leq 335$ K in summer and fall, the absence of double tropopauses, and a latitude equatorward of 80°N. The profiles selected are largely within 40–80°N. The chemical composition of the lower stratosphere, as already discussed in the introduction, is largely controlled by the downward transport of aged stratospheric air via the Brewer-Dobson circulation, with seasonally varying contribution from isentropic mixing between tropical and high latitudes. The region is therefore well suited for evaluating how well models represent the two competing processes. The vertical structure is examined using $H_2O$ and CO using annual mean distributions. The CO evaluation can be seen as an extension of the metric using normalized vertical profiles of CO (see Section 4.2.3). A mean value is derived from observations and models for both the UT (with data between 1 and 5 km below the tropopause) and LS (with data between 1 and 5 km above the tropopause) and used to calculate grades according to equation (8).
The CO and H$_2$O vertical structures and the grading values are shown in Figure 17. Most models are not yet designed to simulate tropospheric chemistry and anthropogenic fuel or biomass burning sources of CO, resulting in a significant underestimation of the simulated CO in the UT compared to the observations (Figure 17, left), but good agreement is generally found for the LS. This can be seen in Figure 17 (bottom left) where grades are shown to be close to zero for the UT, but generally higher than 0.5 for the LS. However, even the models including tropospheric chemistry (CAM3.5, EMAC, and ULAQ) underestimate the CO vertical profiles substantially. H$_2$O is well simulated by the models in both the UT and LS, except for MRI which shows too high values in the LS. The corresponding grades (Figure 17, bottom right) are clustered at values of 0.7 for the UT and 0.8 for the LS. The differences between observations and models, although smaller than those observed in CO, may have significant implications for the simulation of surface climate [Forster and Shine, 1997; Solomon et al., 2010]. The comparison with the ACE-FTS satellite data indicates a good agreement between the two data sets, within their uncertainties. Differences between satellite and aircraft observations may be due to the coarser vertical resolution in the satellite data and/or the smaller regional coverage of aircraft observations. A final grade is calculated by averaging the UT and LS H$_2$O grades for each model and is listed in the summary matrix.

4.2.5. ExTL Depth From Tracer-Tracer Correlations

The extratropical tropopause transition layer (ExTL) exhibits air masses with a mixed composition of partly tropospheric, partly stratospheric air [Fischer et al., 2000; Haynes and Shepherd, 2001; Hoor et al., 2002; Pan et al., 2004]. The representation of the ExTL characterizes how well the models reproduce the tropopause as a transport barrier and its sharpness. The transition layer depth and center location is examined using the tracer-tracer correlation method applied to O$_3$ and H$_2$O [Pan et al., 2007; Hegglin et al., 2009]. A stratospheric branch is identified using a fit to a polynomial function of second order to all data points in the LS (<20 km) with H$_2$O < 10 ppmv. Similarly, a tropospheric branch is represented by a linear function derived by fitting all data points with O$_3$ < 100 ppbv for both observations and models. Mixed air masses are identified as those points outside the 3σ range of both the stratospheric branch and the tropospheric branch. The fractional distribution of the identified air parcels is then plotted as a function of their distance from the thermal tropopause (in Figure 18). The observed transition layer is derived using POLARIS aircraft data in the NH, which include measurements in spring, summer and fall [Pan et al., 2007]. The model output was chosen for the same latitudes and from the same seasons. Note that the ExTL depth derived from the O$_3$-H$_2$O pair is different from the one obtained using the O$_3$-CO pair [Hegglin et al., 2009]. The different ExTL depths are explained by the fact that H$_2$O is strongly affected by the minimum temperature an air parcel experiences during its travel across the tropopause into the stratosphere, and its longer life-time compared to CO once it enters the stratosphere.

Two parameters are used to quantify the comparisons: a) the center of the ExTL, defined as the center point of the distribution at the half maximum, and b) the ExTL width, defined as the width at half maximum of the probability density function. These criteria are influenced by the bin size, which was chosen as 0.5 km for the observations and for the models adjusted to their vertical resolution (0.5 or 1 km). ExTL width and center are shown in Figure 18 (right). The ExTL is well manifested in most models. However, in all cases the ExTL is broader, between 2 and 4 km, compared to 1 km derived from the observations. Further, the models’ layer centers are shifted upward by about 1 km in most cases. Observations from the ACE-FTS satellite are also shown (observations taken from Figure 3 of Hegglin et al. [2009]). ACE-FTS has an effective resolution similar to the CCMs (around 1 km) and shows a behavior similar to that of AMTRAC3, CAM3.5, CMAM,
GEOSCCM, and WACCM, exhibiting a layer width of 2 km and a layer center at 1 km above the thermal tropopause. The limited resolution of the models is therefore likely the primary cause for this shift. Indeed, a higher resolution model by Miyazaki et al. [2010a] is capable of better resolving the fine-scale structure of the ExTL. Uncertainties in tropopause locations derived from the relatively coarse vertical resolution of the models or lack of representativeness of the aircraft observations may also have contributed to the discrepancies.

5. Discussion of Model Performance

5.1. Quantitative Discussion

The quantitative results of the metrics listed in Section 3.2 and discussed in more detail in the main body of this paper are compiled in the summary matrix (Figure 19). Note that the grading often involves subjective choices and is sensitive to observational errors. However, while the models often reach very high scores, indicating a good model performance in the extratropical UTLS, there are also some obvious deficiencies. The models score better in metrics testing dynamics (zonal mean zonal wind and LMS mass) rather than transport and mixing (seasonal cycle of O$_3$ and H$_2$O, meridional gradient in O$_3$, normalized and absolute vertical profiles), and better in metrics focusing on the LS (seasonal cycle of O$_3$ at 100 hPa, meridional gradient in O$_3$, and transition from the troposphere to the stratosphere). At least for the distribution of O$_3$, the lack of tropospheric chemistry or limitations in its representation may also contribute to the poorer performance of the models in the transition region. Note, however, that models including a more comprehensive tropospheric chemistry are not seen to perform better than the others. The multimodel mean generally scores higher than any individual model, except for the seasonal cycle of H$_2$O. In fact, most models seem to score lower in this latter metric, which is likely due to the uncertainty in the observations as discussed in Section 4.2.1. The fact that the multimodel mean scores so well on all diagnostics suggests that there are no significant missing processes in the models, although particular models may have significant deficiencies in their representation of the processes.

Two models score consistently higher than 0.7 throughout the different extratropical UTLS metrics: AMTRAC3 and CMAM. CMAM may deserve a special note, since it has a relatively low horizontal resolution (T31), which would usually be expected to limit model performance. However, the ratio of vertical to horizontal resolution may be more important [Fox-Rabinovitz and Lindzen, 1993]. Also the spectral transport scheme is known not to be diffusive. Other models that perform well are CAM3.5, E39CA, EMAC, GEOSCCM, the UMUKCA-models, and WACCM with the only grade lower than 0.7 being the metric of the seasonal cycle of H$_2$O. Moderately performing models are CCSRNIES, LMDZ-repro, and MRI. The lowest scoring models are CNRM-ACM, the SOCOL-models, and ULAQ showing lower values than average for most of the extratropical UTLS metrics. The SOCOL-models may score low due to a combination of their hybrid transport scheme and a relatively low horizontal (T30) and vertical resolution (5 levels in the UTLS), and ULAQ due to its very low horizontal (11.5° × 22.5°) and vertical resolution (3 levels in the UTLS) and its quasi-geostrophic dynamical core. CNRM-ACM also uses a cubic semi-Lagrangian transport

Figure 19. Summary of metrics used to quantitatively evaluate the different models. ‘MMM’ denotes the multimodel mean. The diagnostics used as metrics are: ‘U@200’ the zonal mean zonal wind at 200 hPa (Section 4.1.1), ‘SEAS CYC O3’ and ‘SEAS CYC H2O’ the seasonal cycle of the tracers (Section 4.2.1), ‘GRAD@200’ the meridional gradient in O$_3$ at 200 hPa (Section 4.2.2), ‘CO NORM’ the normalized vertical CO profiles in tropopause coordinates (Section 4.2.3), and ‘H2O PROF’ H$_2$O profiles in tropopause coordinates (Section 4.2.4). Numbers indicate the lower bound of the color-range. Models with grades larger than 0.6 are considered to perform satisfactorily.
scheme. UMETRAC and UMSLIMCAT are not rated here since they lack grades for more than two diagnostics.

5.2. Qualitative Discussion

5.2.1. Seasonal Zonal Mean Zonal Wind

[76] Most models perform well in reproducing this quantity, reflecting that the thermal structure and therefore the basic dynamical state of the atmosphere is well represented in the models. One exception to this is ULAQ. This might be due to an insufficient horizontal and/or vertical resolution and the quasi-geostrophic dynamical core which introduces errors in the thermal wind balance.

5.2.2. Seasonal Cycle of LMS Mass

[79] Most models represent well the phase and amplitude of the seasonal cycle of LMS mass, but not so the annual mean value in LMS mass. Overall scores are generally higher for the NH than for the SH. Models scoring high are AMTRAC3, CAMM, E39CA, and GEOSCCM. Models that perform poorly are CCSRNIES, CNRM-ACM, and ULAQ. The diagnostic yields insight into the strength and seasonality of the Brewer–Dobson circulation which will affect stratosphere–troposphere exchange and therefore both UT and LS tracer distributions.

5.2.3. Tropopause Pressure Anomalies

[80] Although the models seem to reproduce the seasonal cycle of tropopause pressure well in the NH (which is supported by the metric of LMS mass), they show more problems in representing interannual variability. CNRM-ACM has unrealistically large interannual variability and low tropopause pressure. CCSRNIES, EMAC, ULAQ and WACCM perform worst among the models, with both too high/low mean values and very small correlation with the structure of the observed variability.

5.2.4. Tropopause Inversion Layer

[81] Most models simulate a clear TIL with a strong maximum in $N^2$ just above the extratropical tropopause, except CCSRNIES which shows a more monotonic increase in $N^2$ across the tropopause. The models also simulate correctly the seasonal cycle of the TIL with stronger maxima in summer than during winter. However, the maximum in the TIL is shifted to slightly higher altitudes above the tropopause than expected from full-resolution observations. Degrading the resolution of the observations shows a closer match with the models and yields a more meaningful model-measurement comparison.

5.2.5. Seasonal Cycles in $O_3$, HNO$_3$, and H$_2$O at 100 and 200 hPa

[82] Most models perform reasonably well for $O_3$ in the NH, however the amplitude is consistently too high at 100 hPa and too low at 200 hPa. The latter finding indicates that the models exhibit too much transport from the tropics at and above 100 hPa, and across the tropopause at 200 hPa. The spread in skill in representing H$_2$O is larger than for $O_3$, with models doing better at 100 hPa than at 200 hPa. At 200 hPa, strong tropospheric influence causes too large amplitudes in the seasonal cycle. Note that there exist large observational uncertainties in H$_2$O at the 200 hPa level. Better measurements are needed to gain more confidence in this quantitative metric. The spread in skill at representing HNO$_3$ in the SH at 200 hPa is even larger. UMSLIMCAT shows the highest score, but most other models have a low correlation with observations and underestimate the annual cycle amplitude. The CCSRNIES and the SOCOL-models seem to perform worst.

5.2.6. Sharpness of Meridional Gradients in $O_3$

[83] The results of this metric are largely consistent with those from the seasonal cycle of $O_3$ at 200 hPa. Models generally are capable of maintaining a clear distinction between the tropical UT and the extratropical LS, as indicated in strong maxima in the gradient at the location of the subtropical jet. However, some models overestimate (UMSLIMCAT and UMUKCA–METO) and some underestimate (CNRM–ACM and ULAQ) the maximum value in the gradient.

5.2.7. Normalized CO in Potential Temperature Relative to the Tropopause Height

[84] Most models perform reasonably well in this diagnostic, except LMDZrepro, SOCOL, and Niwa–SOCOL, which have too much transport across the extratropical tropopause. The SOCOL-models have a semi-Lagrangian transport scheme which together with a relatively low vertical and horizontal resolution may contribute to the models being too diffusive.

5.2.8. Vertical Profiles of H$_2$O and CO in Tropopause Coordinates

[85] The models show some difficulties simulating the seasonal mean vertical profiles of the different tracers. Models perform better for H$_2$O, possibly because it is primarily affected by the tropopause temperature and less affected by chemistry and global-scale transport than CO. CO is represented poorly, primarily because most models do not include tropospheric chemistry or treat it in a simplified way. Note however that CAM3.5 which includes tropospheric chemistry shows about the same CO profile as WACCM which does not include tropospheric chemistry. EMAC and ULAQ also have a comprehensive tropospheric chemistry, however they do not perform remarkably better than the other models. The lack of a more sophisticated tropospheric chemistry will likely result in poor UT O$_3$ distributions, which remains to be tested in a future assessment.

5.2.9. Depth of the Extratropical Tropopause Transition Layer

[86] The models simulate an extratropical tropopause transition layer (ExTL) that is deeper than observed in aircraft observations, and shifted above the thermal tropopause. This is likely due to the models’ limited vertical resolutions, as comparison with the ACE–FTS satellite observations indicates, which have a resolution more similar to that of the models. On the other hand, the aircraft measurements may lack representativeness. CMAM scores best in this metric, which is noteworthy since CMAM has a relatively low horizontal resolution compared to other models. The ratio between vertical and horizontal resolution might matter more. Models that show most difficulties in reproducing the ExTL are SOCOL, UMUKCA–METO, and CNRM–ACM, whose transport schemes may be too diffusive. Also rather poor performance is seen for CCSRNIES, LMDZrepro, and MRI.

6. Conclusions

[87] We have presented here the first multimodel assessment of chemistry–climate models (CCMs) in the extratropical
upper troposphere/lower stratosphere (UTLS) – a region of the atmosphere important to different aspects of chemistry-climate coupling. Different diagnostics have been used to characterize model performance and the results are discussed in more detail in Sections 5.1 and 5.2. A main finding of our evaluations is that the CCMs represent the main dynamical and chemical characteristics relatively well despite their limited horizontal and vertical resolution.

[88] Why do the CCMs perform relatively well in the UTLS despite resolution issues? The models get the basic wave-driven dynamics right of (a) the stratospheric Brewer-Dobson circulation, and (b) the tropospheric baroclinic general circulation in the extratropics. The basic (stratospheric) chemistry for ozone and condensation of water vapor is also captured. These abilities allow the models to capture the basic dynamical climatology, and with appropriate transport schemes, can properly approximate transport and thereby tracer distributions on a climatological basis. For the tropopause inversion layer (TIL), it is the large scale dynamics, a good representation of the distributions of radiatively-active tracers, and a proper representation of radiative transfer, that allow the models to represent a stability inversion above the tropopause. However, as discussed above, the low vertical resolution in CCMs biases the TIL with respect to high resolution observations. The ExTL is a product of the large scale dynamics and resulting transport with basic tracer chemistry, and in the UTLS the combination of large scale diabatic transport, quasi-isentropic transport and mixing with the tropics across seasonally variable subtropical transport barriers, and cross-isentropic transport and mixing at the extratropical tropopause. In the models these features are represented with sufficient fidelity to reproduce tracer gradients in the ExTL.

[80] For a better representation of the UTLS in the future, CCMs will need to improve in a few key areas. Increased vertical resolution will help represent the fine-scale structures of the TIL and ExTL. Tropospheric chemistry and increased horizontal resolution are also important for improving model representation of gradients across the tropopause. As part of this development, CCMs should add a range of very short-lived species (VSLS) that represent major reservoirs of source gases for stratospheric bromine. VSLS can be used as process-specific transport diagnostics. These tracers provide a range of life times and can discriminate transport from marine and continental source regions into the UTLS.

[90] Our results further demonstrate that the UTLS is still relatively sparsely sampled by observations which unfortunately limits confidence in the quantitative evaluation of model performance in the UTLS (see Figures 11 and 12 and related discussion). It will be necessary to compare the metrics applied here with future measurements to reduce uncertainty in the model comparison associated with potential measurement errors. The observational data-base for the UTLS needs to be expanded by measurements with both higher spatial and temporal resolution and sampling, but also reasonable accuracy. New observations are needed especially for O₃ and H₂O in the UTLS with a vertical resolution better than 1 km and a horizontal resolution better than 100 km, especially in the SH, the tropics, and the upper troposphere.

References


Cheverux, C., et al. (2008), CO measurements from the ACE-FTS satellite instrument: Data analysis and validation using ground-based, airborne and spaceborne observations, Atmos. Chem. Phys., 8, 2569–2594.


Engel, A., et al. (2006), Highly resolved observations of trace gases in the lowermost stratosphere and upper troposphere from the SPURT project: An overview, Atmos. Chem. Phys., 6, 283–301.


