Chapter 5: Implications of results

5.1 Implications for model-measurement intercomparison

Satellite trace gas datasets are crucial for the evaluation of transport and chemistry in numerical models. Datasets available from different satellite instruments vary in terms of measurement method, geographical coverage, spatial and temporal sampling and resolution, time period, and retrieval algorithm and thus have different strengths and shortcomings. Comparing numerical model output to different chemical datasets can lead to conflicting results depending on the particular application. Issues arising from the use of different observational datasets for model evaluations have been identified in the CCMVal report [*SPARC*, 2010]. It became clear that the characteristics of the satellite datasets, including quality, resolution, and representativeness, need to be known prior to their use and prior to the interpretation of model evaluation results. The CCMVal report's recommendations that "*A systematic comparison of existing observations is required in order to underpin future model evaluation efforts, by providing more accurate assessments of measurement uncertainty*" directly motivated the work for the SPARC Data Initiative presented in this report. While *Chapter 4* provides basic information on quality and consistency of the various data products, the following *Chapter 5* focuses on summarizing some implications of the results for model-measurement inter-comparisons. Examples of how knowledge of uncertainty and inter-instrument differences can be used to improve comparisons are given and particular diagnostics appropriate for model evaluations are recommended.

For the CCMVal report, the observational mean values and uncertainty range served as input for the performance metrics. Such metrics are used to quantify the ability of models to reproduce key stratospheric processes. One widely applied metric:

$$g = 1 - \frac{1}{n_g}\frac{|\mu_{mod} - \mu_{obs}|}{\sigma_{obs}} \qquad (5.1)$$

uses a scaling factor n_g as well as the observational uncertainty σ_{obs} and climatological mean μ_{obs} for the evaluation of the model climatological mean μ_{mod} [*Douglass et al.*, 1999; *Waugh and Eyring*, 2008; *SPARC*, 2010]. In the past, the observational uncertainty has most often been derived using the interannual variability of a single instrument only.

Our approach is to provide an alternative, more comprehensive uncertainty range derived from all available datasets, instead of recommending one particular satellite dataset for the model-measurement comparison. The selection of the data points suitable for the construction of the new climatological mean values and uncertainty range is based on their agreement with the mean state of the atmosphere as given by all instruments and on the specific satellite characteristics such as sampling patterns. The following general guidelines are applied for the selection process.

- The agreement of each individual dataset with the mean state of the atmosphere is determined based on the 1σ standard deviation over all instruments. For trace gases observed by more than five instruments, individual data points will be removed if they are outside of the ±1σ standard deviation range. For trace gases observed by five or less instruments, the data points will be removed if they are outside of the ±2σ standard deviation.
- Further specific criteria used to calculate the mean state and uncertainty range are chosen based on the instrument/retrieval performance identified in the different chapters of this report, and will change depending on the diagnostic and the trace gas. Detailed information on the evaluations is provided in the following paragraphs, structured according to evaluation diagnostic.
- For each diagnostic and even within one diagnostic, the datasets selected for the construction of the uncertainty range can be different depending on latitude, altitude, or time period considered. One example of this approach can be given for the evaluation of the ozone seasonal cycle: if one instrument presents a clear outlier for *e.g.*, March, then only the March value of this instrument is removed while the values for all other months stay included in the uncertainty range.
- The climatological mean μ_{obs} is defined as the multi-annual, multi-instrument mean value of all data points selected as suitable for the construction of the uncertainty range. Note that the climatological mean is different from the MIM used in the previous chapters which was based on all available datasets.
- The uncertainty range σ_{obs} is defined as the spread over all selected datasets and years. In general, the interannual spread needs to be accounted for when producing the uncertainty range, so that the free-running models can be compared against the observational mean state. Note however, that for model simulations nudged to meteorological reanalysis, the comparisons focus on the same years and the uncertainty range can be solely based on the spread over all selected datasets and not include interannual variations.

In summary, we derive an observational mean state and uncertainty range from multiple datasets from the SPARC Data Initiative for selected examples of evaluation diagnostics. For all evaluations listed in the following *Section 5.1*, the uncertainty range and climatological mean will be made publicly available through the SPARC Data Center.

5.1.1 Seasonal cycles

The seasonal cycle of long-lived atmospheric trace gases such as water vapor and ozone is often used as a diagnostic of transport processes in the stratosphere and in particular in the UTLS.

Ozone – O_3

The ozone seasonal cycle in the UTLS in mid-latitudes is determined by the seasonality of two processes: air mass transport with the Brewer-Dobson circulation and mixing with tropical air masses. In order to evaluate the model's representation of these large-scale transport and mixing processes, a comparison of the ozone seasonal cycle for the latitude bands 40°S/N-60°S/N at 100 and 200 hPa has been used [*SPARC*, 2010; *Hegglin et al.*, 2010]. While the calculation of the quantitative performance metric in the CCMVal report was based on MIPAS data alone, we will provide a new climatological mean state and uncertainty range derived from multiple datasets. The method (illustrated in **Figure 5.1.1**) is explained below for 40°N-60°N, 200 hPa.

Step 1: The ozone seasonal cycles for satellite datasets are derived from 2005-2010 multi-annual mean values. The time period has been chosen based on a maximum number of active satellite limb instruments. The uncertainty range (grey shading in **Figure 5.1.1**) is calculated as the ±1σ standard deviation over all instruments' multi-annual mean values.

Step 2: All data points outside of the ±1σ standard deviation from step 1 are removed. Additionally, data points from instruments with a very large interannual spread need to be excluded. Therefore, all multi-annual mean values with an interannual variability (vertical bars in the uppermost left panel of **Figure 5.1.1**) larger than the ±2σ standard deviations from step 1 are removed. The new mean values and uncertainty range are calculated.

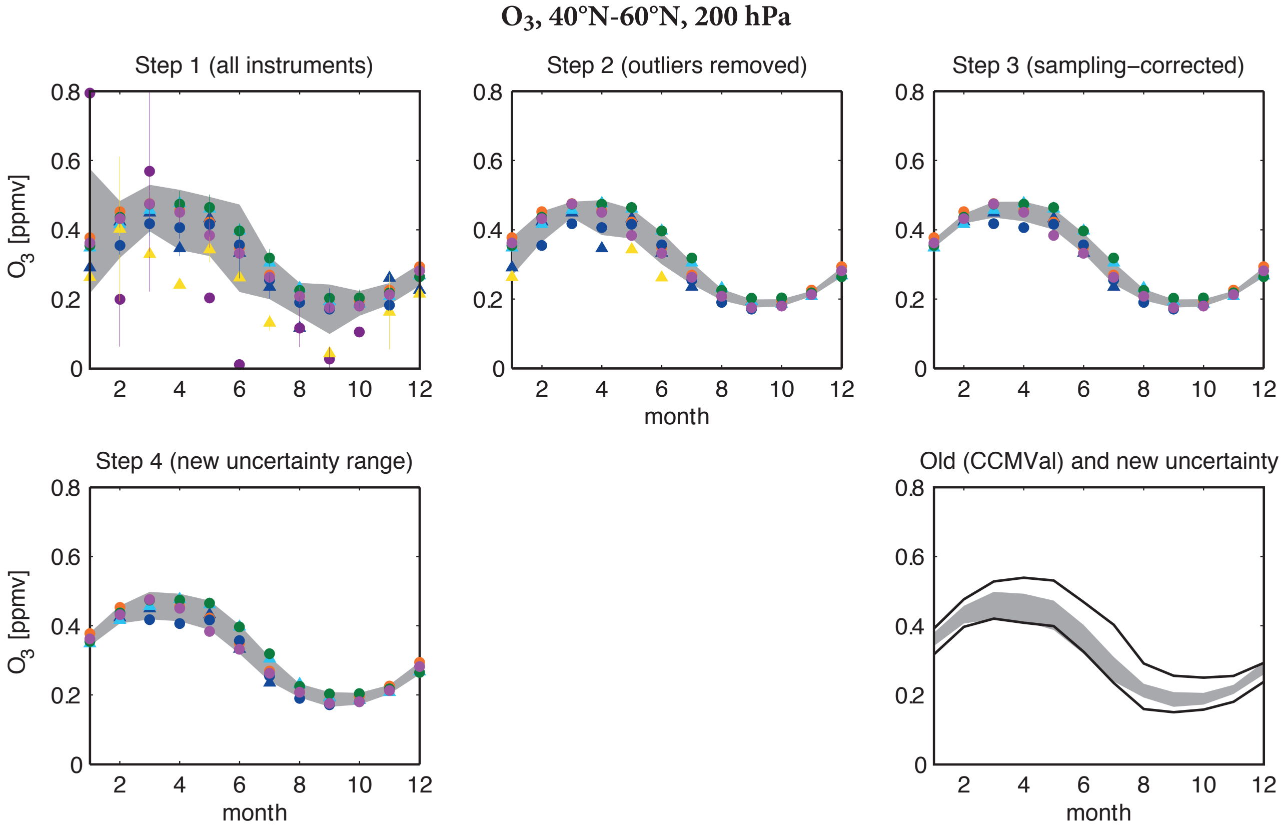

Figure 5.1.1:* *Ozone seasonal cycle diagnostic for 40°N-60°N at 200 hPa. *The individual steps of deriving the ozone seasonal cycle diagnostic are shown. The uncertainty range (grey shading) is given for each month by the standard deviation over all multi-annual means of the selected datasets. In the uppermost left panel the vertical bars indicate the interannual spread of each instrument calculated as the standard deviations over all years. For the selection of the datasets, outliers and data points strongly impacted by sampling are removed as illustrated in steps 1 to 3 and explained in detail in the text. In step 4 the uncertainty due to interannual variations is added to the uncertainty range. In the lower rightmost panel the old uncertainty range given in the CCMVal report and the new uncertainty range are compared.*

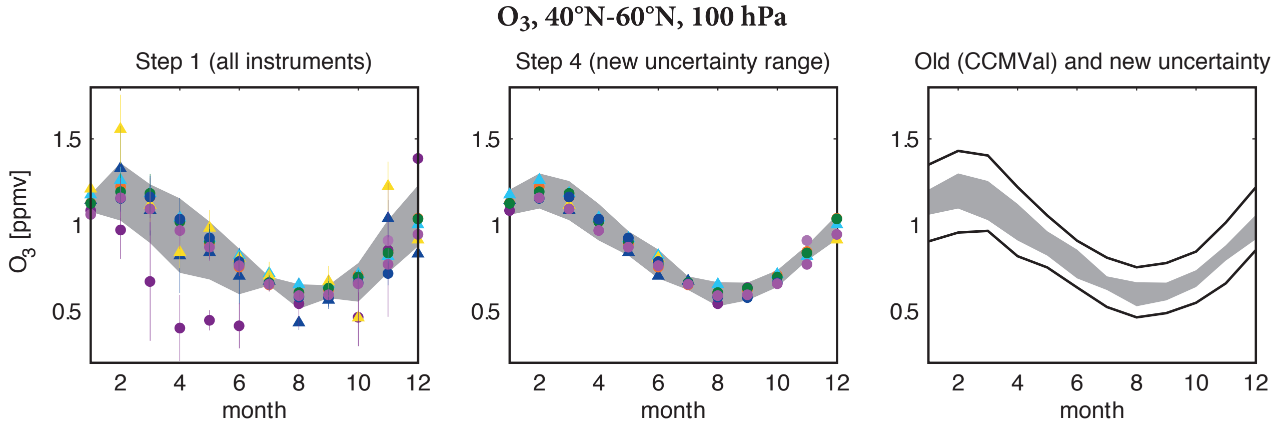

Figure 5.1.2: Ozone seasonal cycle diagnostic for 40°N-60°N at 100 hPa. *Steps 1 and 4 of deriving the ozone seasonal cycle diagnostic are shown. The uncertainty range (grey shading) is given for each month by the standard deviation over all selected datasets. In the rightmost panels the old uncertainty range given in the CCMVal report and the new uncertainty range are compared.*

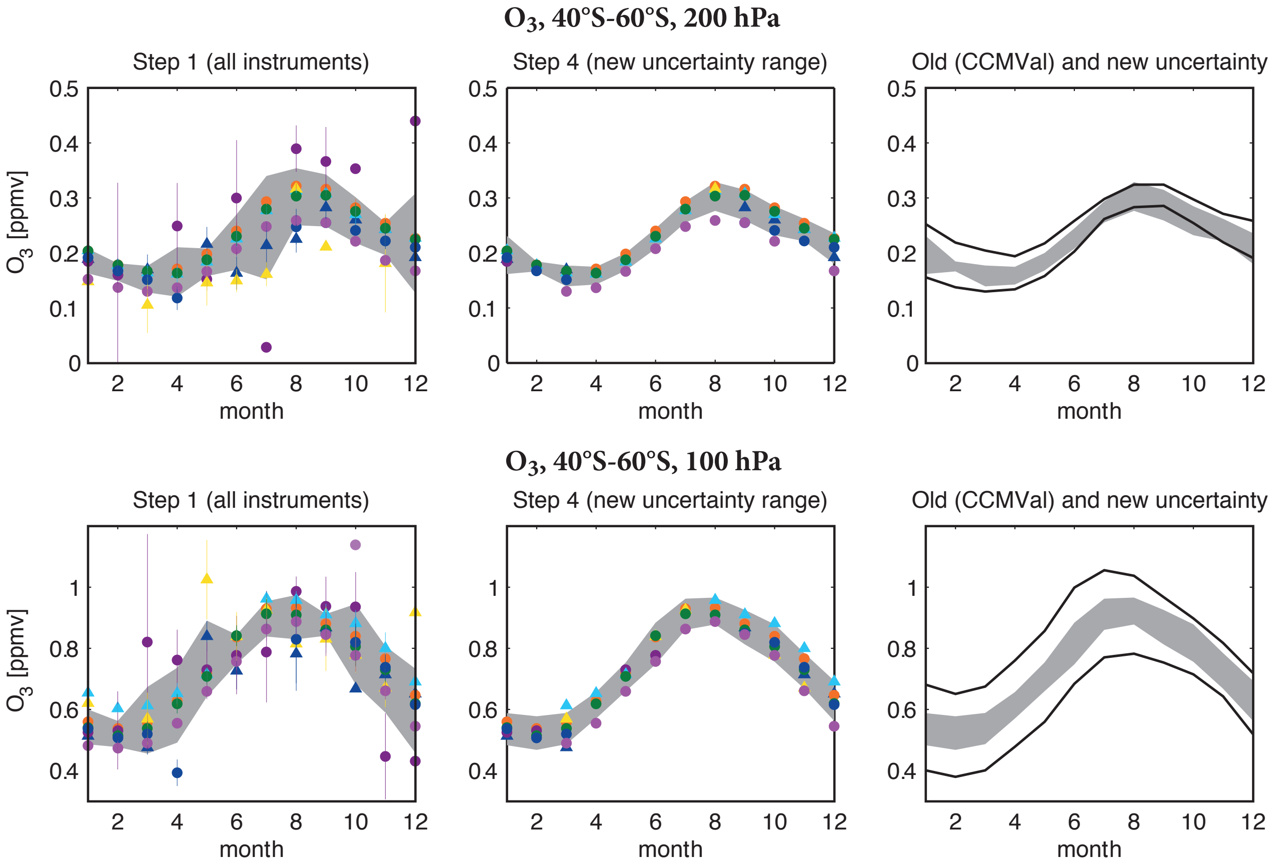

Figure 5.1.3: Ozone seasonal cycle diagnostic for 40°S-60°S at 200 and 100 hPa. *Same as Figure 5.1.2 but for 200 and 100 hPa at 40°S-60°S.*

Step 3: All data points impacted by a sampling bias estimated to be larger than 10% are removed. Such sampling bias can arise when averaging binned atmospheric measurements due to non-uniform sampling in time or space. These sampling biases have been identified by applying the sampling patterns of the satellite instruments to O$_3$ fields from coupled chemistry climate models (see *Chapter 3*; *Toohey et al.* [2013]). In the tropics, comparisons to ozonesondes are used to remove data points that show large deviations. The new mean values and uncertainty range are calculated.

Step 4: The uncertainty range is recalculated as the ±1σ standard deviation over all remaining instruments and years, now taking not only the inter-instrument but also the inter-annual spread into account. Including the latter in this final step increases the uncertainty range for most cases, but is nevertheless important in order to produce an uncertainty that free-running models can be compared against.

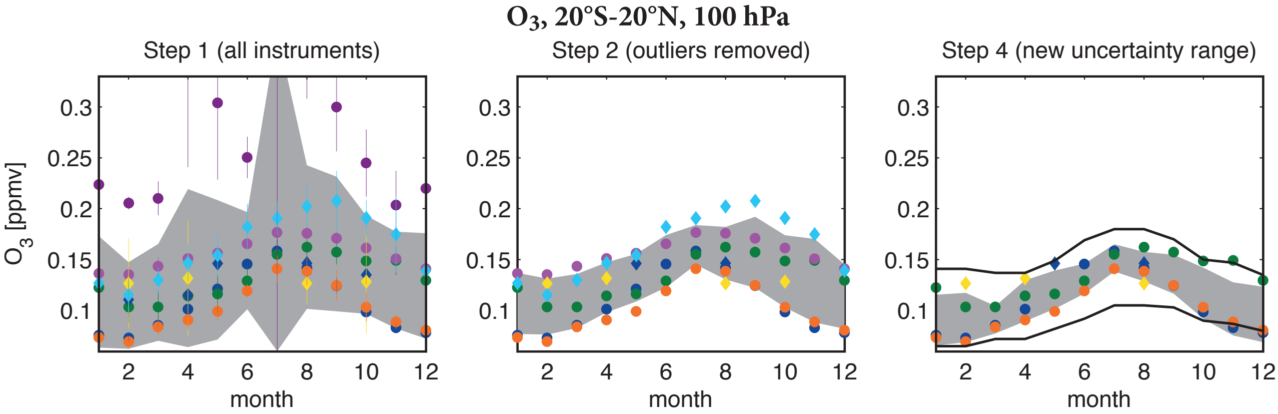

Figure 5.1.4: ***Ozone seasonal cycle diagnostic for 20°S-20°N at 100 hPa.*** *Steps 1, 2 and 4 of deriving the ozone seasonal cycle diagnostic are shown. The uncertainty range (grey shading) is given for each month by the standard deviation over all selected data points.*

To summarise, the final ozone seasonal cycle for the 2005-2010 period is calculated from the instruments' multi-annual mean values remaining after the removal of outliers and data points impacted by sampling bias (steps 1-3). The uncertainty range is calculated accordingly as the ±1σ standard deviation over all those instruments and over all years. The new uncertainty range is generally smaller than the old uncertainty range used in the CCMVal report (lower right panel in **Figure 5.1.1**). For some months, the uncertainty has been reduced by more than 50%. This reduced uncertainty range applied to the quantitative performance metric (**Eq. 5.1**) will provide a powerful constraint on the model results and thus the model representations of the ozone seasonal cycle can be differentiated more clearly than before. Additionally, the climatological mean is now shifted to lower values. The new lower mean values agree better with the CCMVal models whose multi-model mean values were found to be too low compared to the old climatological mean values (see Figure 7.22 in the *CCMVal report*; also Figure 11 in *Hegglin et al.* [2010]). The improved agreement suggests that most of the CCMVal models perform better than previously thought with regard to the ozone seasonal cycle in the UTLS.

For the presentation of the improved seasonal cycle diagnostic for other regions and trace gases, only step 1 and 4 and for some regions also step 2 as well as the comparison with the old CCMVal uncertainty range will be displayed. Figures containing each step of the derivation of the new uncertainty range are provided in *Appendix A5*. At 100 hPa in the Northern Hemisphere (NH) mid-latitudes (40°N-60°N), the ozone seasonal cycle is derived from the 2005-2010 multi-annual mean of 9 satellite instruments (**Figure 5.1.2**). Reducing the satellite datasets according to their agreement with the multi-instrument mean value and their sampling biases results in a much reduced uncertainty range in particular during NH winter and spring. For these months, most of the new uncertainty is caused by inter-annual variations and not by inter-instrument variations as becomes clear from the multi-annual mean values clustering in the center of the new uncertainty range. Similar to our results for 200 hPa, the new uncertainty range is much reduced when compared to the one used in the CCMVal report, and hence will be much better suited to identify badly performing models. For the NH summer and autumn, the reduction is about 2/3 of the old range. The climatological mean value, however, did not change systematically.

The evaluation of the Southern Hemisphere (SH) mid-latitude ozone seasonal cycle (**Figure 5.1.3**) follows the same steps as described above for the NH based on 2005-2010 multi-annual mean datasets from 9 satellite instruments. At 200 hPa, the new uncertainty range is very similar to the old one based on MIPAS observations only. Except for January, this uncertainty results mostly from the inter-instrument spread (and not from the inter-annual spread) with one instrument having particularly lower values than all other datasets. Despite this instrument seeming to be an outlier for some months, it agrees very well with SAGE II and HALOE ozone for the overlap time period 2003 (not included here) confirming this as the lower end of our uncertainty range. At 100 hPa, the new uncertainty range is reduced over the whole year compared to the old one based on MIPAS observations, with strongest improvement for SH summer, autumn and winter. The climatological mean values increase slightly for August-October, but do not change systematically for the rest of the year.

Tropical ozone exhibits a large annual cycle near and above the tropopause which is related to seasonal changes in vertical transport acting on the strong vertical ozone gradient in this region [*Randel et al.*, 2007] and in quasi-horizontal mixing [*Ploeger et al.*, 2012]. Although the annual cycle extends over only a narrow vertical range from approximately 100 to 50 hPa, it is an important characteristic of tropical ozone in the LS and has been used to analyse transport and mixing processes. The SPARC CCMVal evaluation of the seasonal cycle in tropical ozone is based on a comparison to the observational NIWA dataset [*Hassler et al.*, 2008] at 100 hPa for 20°S-20°N. A new uncertainty range based on the SPARC Data Initiative datasets is presented in **Figure 5.1.4.** After the removal of the outliers the uncertainty range (middle panel of **Figure 5.1.4**) is still relatively large and comparable to the

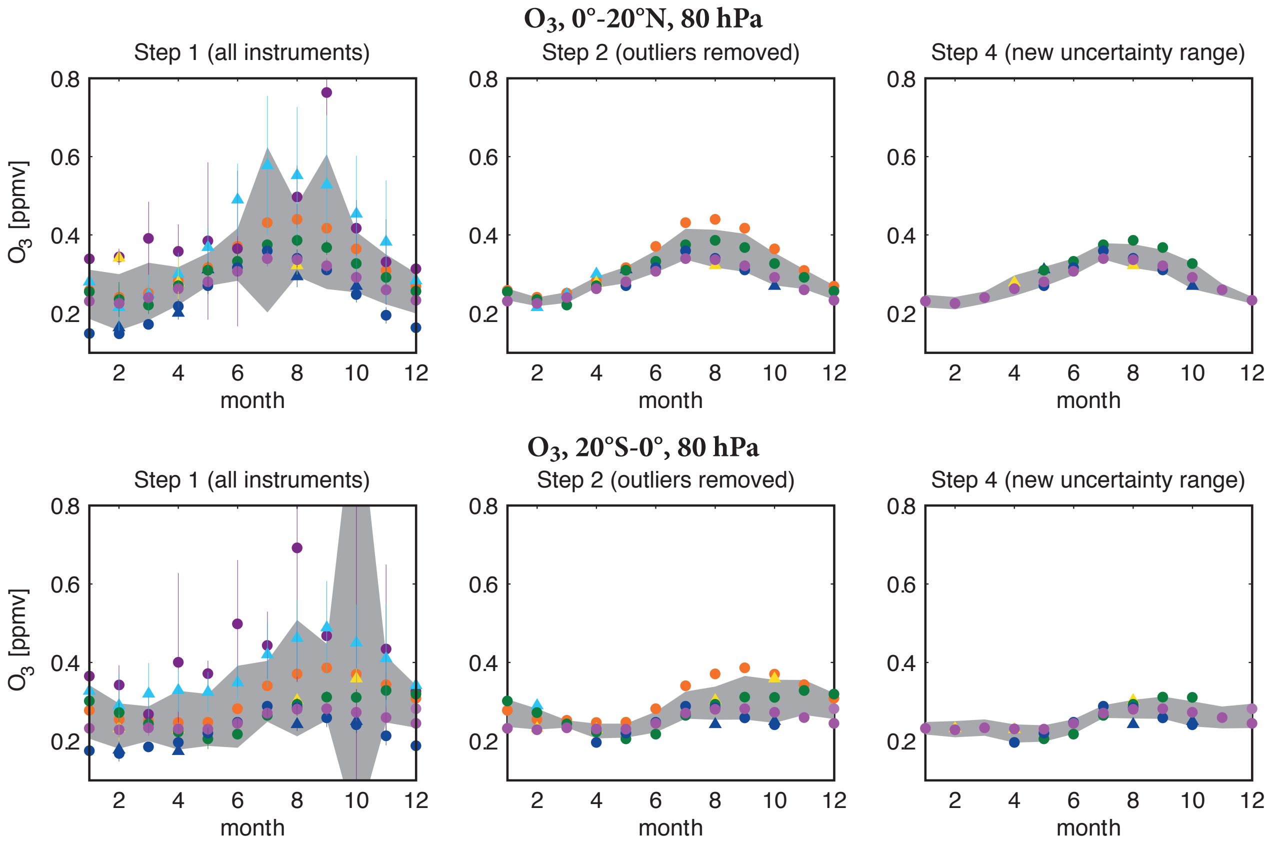

***Figure 5.1.5: Ozone seasonal cycle diagnostic for 0°-20°N and 20°S-0° at 80 hPa.** Same as Figure 5.1.4 but for 80 hPa at 0°-20°N and 20°S-0°.*

NIWA-based uncertainty range (black lines in right panel of **Figure 5.1.4**). Evaluations of UTLS ozone after the application of the TES observational operator [*Section 4.27*; *Neu et al.*, 2014a] including a comparison to a "zonal mean" ozonesonde climatology indicate that in the tropics below 100 hPa most instruments have a positive bias. Removing all datasets that are outside of the ±1σ standard deviation of the climatological ozonesonde measurements (see **Figure 4.27.6** for details) results in a lower mean and also a reduced uncertainty range (right panel of **Figure 5.1.4**). Note that for November, the criteria has not been applied in order to avoid inconsistencies with the October and December uncertainty ranges. In particular for the time period from March to October the uncertainty range has been substantially reduced and is now smaller than the NIWA-based one.

Most studies analyzing the seasonal cycle of long-lived trace gases treat the tropics as a horizontally homogeneous region without differentiating between NH and SH. Very recently differences between the ozone seasonal cycles in the NH and SH tropics, related to hemispheric differences in the seasonal strength of vertical transport and horizontal mixing, have been pointed out by *Stolarski et al.* [2014]. Here, we follow their approach and derive uncertainty ranges for the ozone seasonal cycle at 80 hPa for the NH tropics (0°-20°N) and the SH tropics (20°S-0°) as illustrated in **Figure 5.1.5**. For both regions, the number of overall applicable datasets decreases substantially when identifying outliers and comparing to ozonesondes resulting in a new, narrow uncertainty range. Confirming the results from *Stolarski et al.* [2014] the seasonal cycle of tropical ozone is substantially different in the two hemispheres with a less pronounced and later occurring maximum in the SH. Model-evaluations of the ozone seasonal cycle should thus be based on two diagnostics differentiating between the NH and SH tropics.

Nitric acid – HNO_3

The HNO_3 seasonal cycle in the UTLS is used to evaluate transport and mixing processes in the models on typical time scales of weeks to months. Like ozone, HNO_3 is mostly produced in the stratosphere and thus has a similar seasonal cycle with some differences caused by chemistry and microphysics. **Figure 5.1.6** shows the evaluation diagnostics of the HNO_3 seasonal cycle for 40°N-60°N at 100 and 200 hPa. The seasonal cycles of five satellite instruments are derived from multi-annual mean (2005-2010) values. At both levels, but in particular at 100 hPa, the instruments are clustered together with only little inter-instrument spread (left panels in **Figure 5.1.6**). Compared to the ozone seasonal cycle, we have fewer instruments available (the ozone evaluations are based on 9 instruments) and thus choose a different criterion to identify outliers. Only data points outside of the ±2σ standard deviation calculated in step 1 will be removed. Note that the grey shading in **Figure 5.1.6** corresponds to the ±1σ standard deviation. At both levels, the agreement between the multi-annual mean states of the five instruments is very

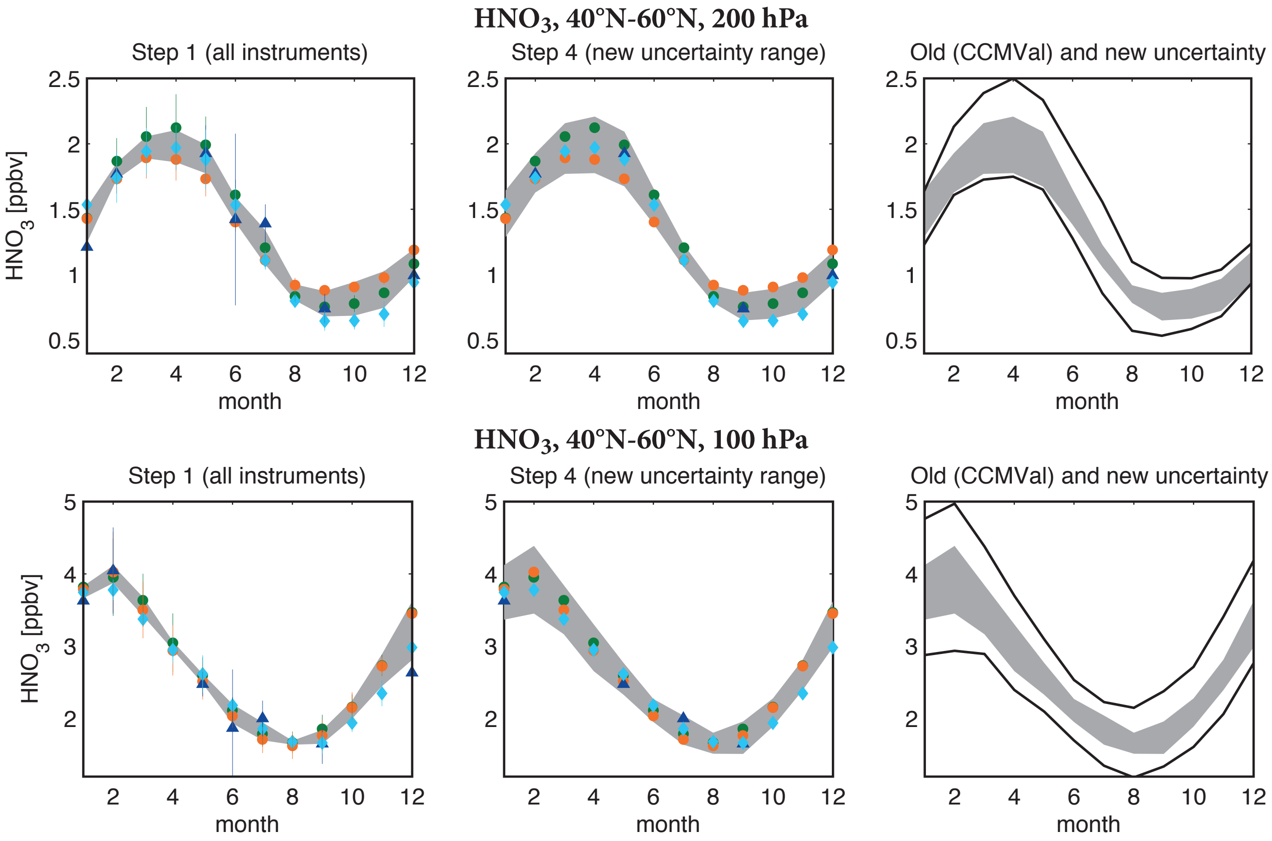

__Figure 5.1.6:__ __HNO__$_3$ __seasonal cycle diagnostic for 40°N-60°N at 200 and 100 hPa.__ Steps 1 and 4 of deriving the HNO$_3$ seasonal cycle diagnostic are shown. The uncertainty range (grey shading) is given for each month by the standard deviation over all selected data points. In the rightmost panels the old uncertainty range given in SPARC (2010) and the new uncertainty range are compared.

good and thus no data points are identified as outliers and excluded from the calculation of the uncertainty range. However, some instruments are removed due to their large interannual variability (illustrated by the vertical bars in the left panels). At 200 hPa, the uncertainty is more driven by the instrument spread than by the interannual variability and is smallest during NH summer. Compared to the CCMVal report the new climatological mean values during NH winter and spring are lower. The new lower mean values agree also better with most of the CCMVal models which were found to be too low when compared to the old climatological mean (see Figure 7.22 in *SPARC*, 2010). At 100 hPa, the new uncertainty range is largest during the NH winter as a result of the inter-annual variability of the remaining data. Comparisons to the uncertainty used in the CCMVal report and in *Hegglin et al.* [2010] show that the new uncertainty range is much reduced.

Similarly to the HNO$_3$ seasonal cycle in the NH UTLS, we derive a new uncertainty range and climatological mean (see **Figure A5.1.4** in *Appendix A5*) for the HNO$_3$ seasonal cycle in the SH UTLS (30°S-60°S, 100 and 200 hPa) as applied in the UTLS chapter of the CCMVal report. The strongest reduction of the uncertainty range with respect to the one used in the SPARC report is found at 100 hPa in the form of an 80% decrease.

Water vapour – H$_2$O

The H$_2$O seasonal cycle in the tropical tropopause region (at 80 hPa) is a key diagnostic to evaluate the amount of water vapour entering the stratosphere [*Gettelman et al.*, 2010]. Water vapour affects stratospheric ozone through HO$_x$-chemistry as well as the formation of polar stratospheric clouds, and also the radiative budget of the UTLS [*SPARC*, 2000]. The seasonal cycle in water vapour is closely related to the seasonal cycle in tropical coldpoint tropopause temperature, which in turn is dominated by the seasonally varying strength of the stratospheric Brewer-Dobson circulation.

Figure 5.1.7 shows the evaluation diagnostics of the H$_2$O seasonal cycle for 20°S-20°N at 80 hPa. The seasonal cycles of seven satellite instruments are derived from multi-annual means averaged over the time period 1996-2010. Choosing a shorter time period for which the instruments would show exact overlap does not improve the comparison between the instruments (see *Chapter 4*; *Hegglin et al.* [2013]), but would limit the number of instruments and information on interannual variability needed in step 4 to calculate an improved uncertainty range.

The instruments do not agree well on the mean values and hence the uncertainty range is relatively large.

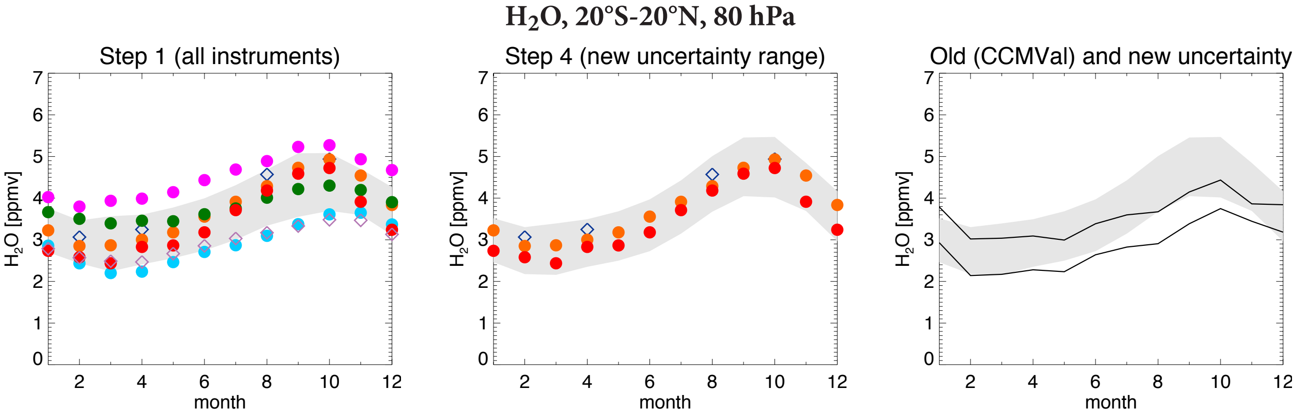

***Figure 5.1.7:** H_2O seasonal cycle diagnostic for 20°S-20°N at 80 hPa. Steps 1 and 4 of deriving the H_2O seasonal cycle diagnostic are shown. The uncertainty range (grey shading) is given for each month by the 1σ standard deviation over all selected datasets (left panel). The middle panel shows the sub-selected datasets, but with the new uncertainty range accounting for interannual variability. In the right panel the new uncertainty range is compared to the uncertainty range given in the CCMVal report.*

Following the approach introduced in this chapter, we exclude instruments that lie outside the ±1σ uncertainty range given by the seven available satellite datasets. The instruments excluded were already identified in *Chapter 4.2* to have weaknesses, with SMR and HALOE showing a distinct low bias and SCIAMACHY showing a high bias due to too low resolution in the altitude range considered. In addition, we also remove MIPAS, since its averaging kernels are state-dependent (measuring with better altitude resolution in a more humid atmosphere), which leads to a seasonal cycle that exhibits too small an amplitude. The remaining instruments agree very well with each other, even though ACE-FTS has very limited sampling in the tropics. Thus no further data points are removed from the calculation and the new uncertainty range is calculated including interannual variability. The new mean seasonal cycle and its uncertainty imply that the models have been evaluated in the CCMVal report against a too low water vapour reference in terms of both mean values and seasonal cycle amplitude, while the old uncertainty range may have underestimated the impact of interannual variability.

5.1.2 Vertical and meridional profiles

Ozone – O_3

Another important aspect of CCM validation is the evaluation of polar spring time ozone profiles. Climatological mean vertical profiles in March at 75°N-85°N and in October at 75°S-85°S are compared between models and observations in order to test the models' representation of transport and chemistry in the polar regions. In contrast to the strong ozone decline driven by anthropogenically emitted ozone depleting substances until the mid-1990s, the Antarctic ozone hole has been controlled primarily by variations in stratospheric temperature and dynamical processes since 1997 [*WMO*, 2011]. In order to avoid the impact of the strong trend before the mid-1990s, we choose the time period 1997-2010 for the ozone profile evaluation in the polar regions. Over this long time period eight ozone datasets provide profile information for the Antarctic spring (**Figure 5.1.8**, left panel). Although some of the datasets cover only part of the time period, most of the profiles cluster together closely. We find one clear outlier with large deviations on the positive side, which is removed in step 2 (**Figure 5.1.8**, middle panel). In the last step, interannual variations are included in the construction of the uncertainty range resulting in slightly larger uncertainties (**Figure 5.1.8**, right panel). Overall in the MS and US, a well-defined mean ozone profile with a relatively narrow uncertainty range is derived for the Antarctic spring. In the LS, however, the spread is quite large which given the overall very small ozone abundances during this time of the Antarctic ozone hole, results in very large relative differences (see also *Chapter 4.1.6*; *Tegtmeier et al.* [2013]). The ozone hole with near-zero ozone values extends from 300 to nearly 50 hPa. Particularly between 100 and 50 hPa, the uncertainty is much higher than in other altitude ranges with similarly low abundance (above 0.3 hPa) or during other times of the year (not shown here). Such differences might be related to the different sampling patterns of the individual instruments and for detailed evaluations of high-latitude ozone in the LS we recommend the use of coincident measurement comparisons, polar vortex coordinates and the use of *in-situ* measurements.

Ozone evaluations can depend on the time period chosen. If we limit the ozone profile comparisons to shorter time periods such as 2000-2010 or 2005-2010 we get very similar mean profiles but a somewhat smaller uncertainty range. In particular, for the latter time period, the uncertainty range in the lower and middle stratosphere can be substantially reduced (see **Figure A5.1.5** in *Appendix A5*). While this suggests a better agreement of the instruments covering the latter time period, one needs to keep in mind that fewer instruments go into this evaluation (five instead of eight) which have at the same time a denser sampling pattern. The evaluation of the earlier time period 1991-2000

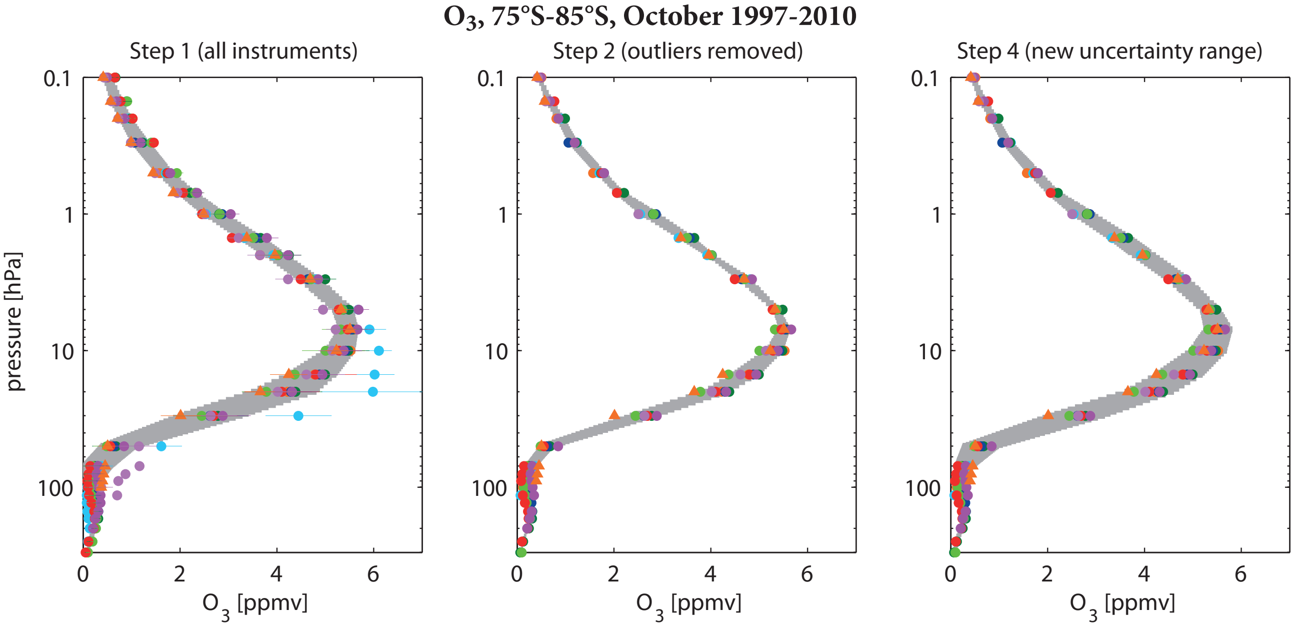

***Figure 5.1.8:** O_3 **vertical profile for 75°S - 85°S in October 1997-2010.** Steps 1, 2 and 4 of deriving the O_3 vertical profile diagnostic are shown. The uncertainty range (grey shading) is given for each level by the standard deviation over all selected datasets.*

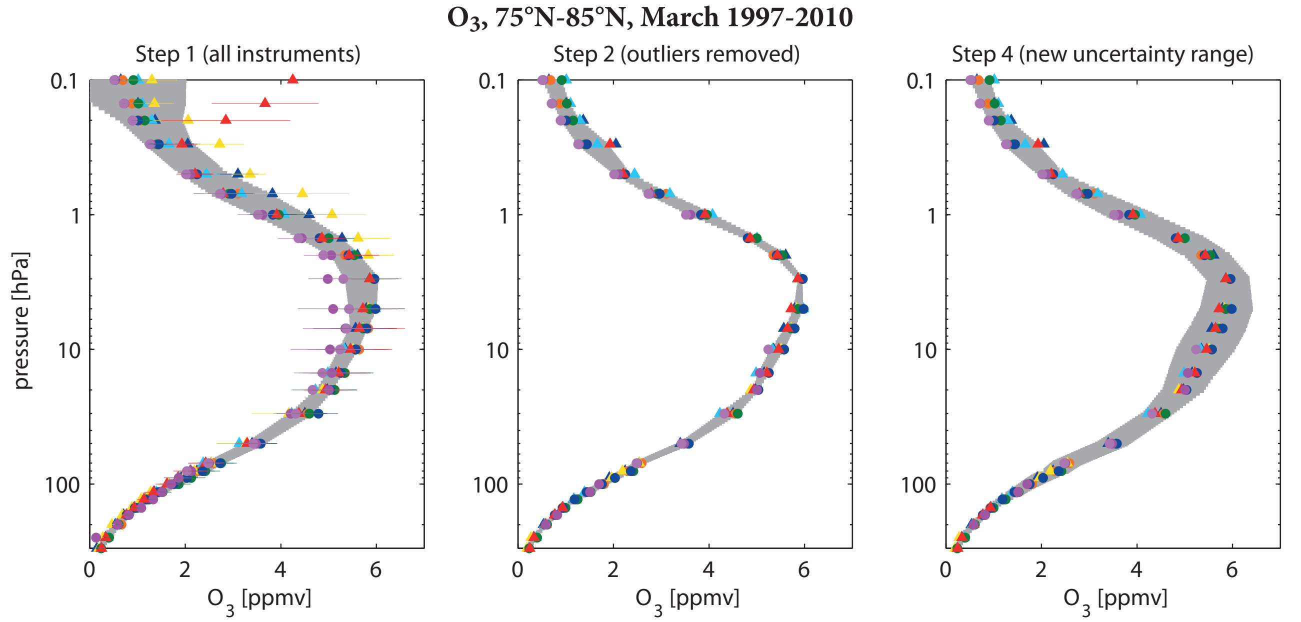

***Figure 5.1.9:** O_3 **vertical profile for 75°N-85°N in March 1997-2010.** Same as Figure 5.1.8 but for 75°N-85°N in March.*

(**Figure A5.1.6** in *Appendix A5*), on the other hand, gives a different mean profile and a slightly larger uncertainty range due to larger interannual variability and, in comparison to 2005-2010, larger instrument-spread. In previous model evaluations focusing also on the 1990s [*SPARC*, 2010; *Eyring et al.*, 2006] the uncertainty range, based on the HALOE climatology and interannual standard deviations, was much larger than the new, multi-instrument uncertainty range introduced above.

Evaluation of the Arctic spring time ozone (here 75°N-85°N in March) shows a large inter-instrument spread, in particular in the MS/US (**Figure 5.1.9**, left panel). The spread is in most cases based on 1-2 outliers which are removed in step 2 resulting in a very narrow uncertainty range (**Figure 5.1.9**, middle panel). Due to the larger dynamical variability at the NH high latitudes, including the interannual standard deviation in the construction of the uncertainty range leads to much larger uncertainties, in particular in the MS. In contrast to the Antarctic, the inter-instrument spread in the LS is quite small leading to a well-defined profile with low uncertainties in this region.

Methane – CH_4

Methane (CH_4) meridional profiles are similarly used in model evaluation to study stratospheric transport characteristics (see *Eyring et al.* [2006]). As mentioned above, transport in the stratosphere involves both the residual mean circulation and isentropic mixing, with the

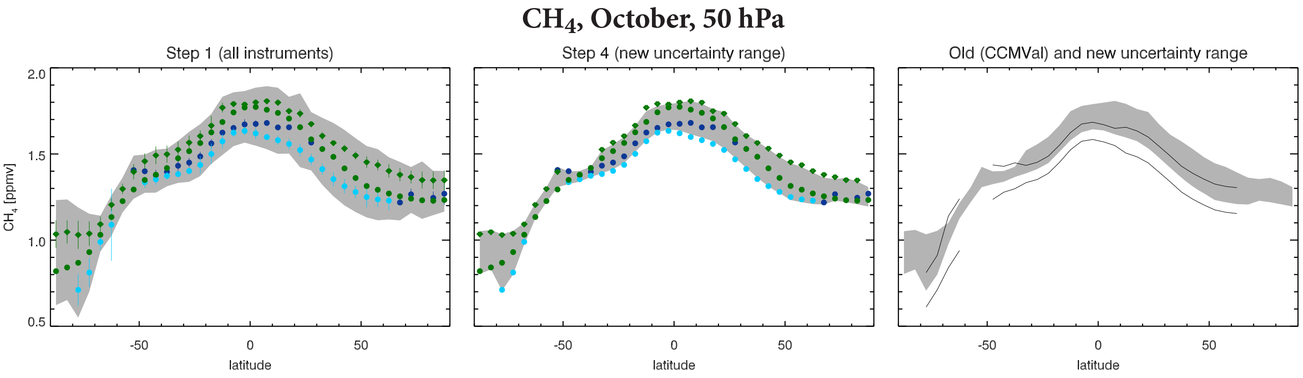

Figure 5.1.10: ***Climatological CH_4 meridional profile at 50 hPa in October over the time period 1998-2010.*** *Steps 1 and 4 of deriving the meridional profile of CH_4 are shown in the upper two panels. The uncertainty range (grey shading) is given for each month by the ±2σ standard deviation over all selected datasets (left panel). In the middle panel the newly derived uncertainty range (accounting for interannual variability) is shown, and in the right panel it is compared to the old uncertainty range given in the CCMVal report.*

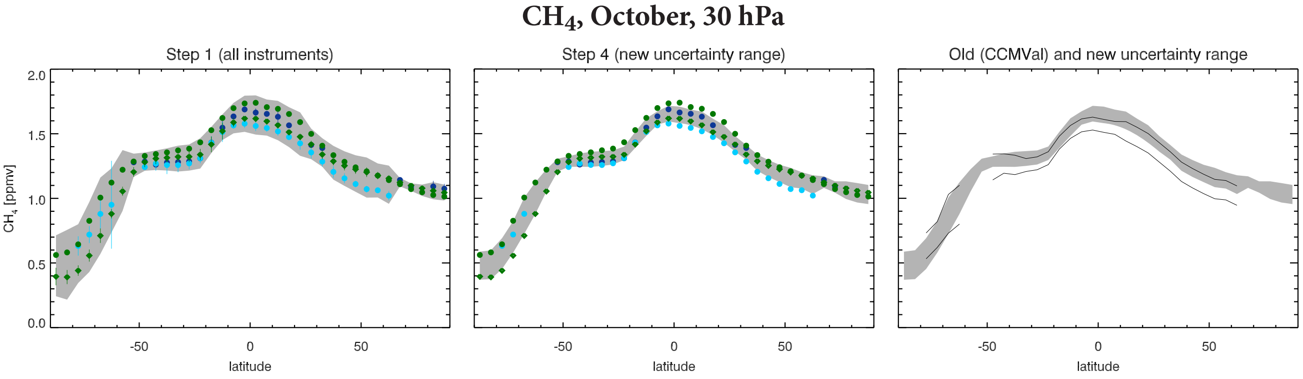

Figure 5.1.11: ***Climatological CH_4 meridional profile at 30 hPa in October over the time period 1998-2010.*** *Same as previous Figure, but for the 30 hPa level.*

latter being highly inhomogeneous in space and time. The winter hemisphere surf zone thereby constitutes a region of strong stirring and mixing, whereas the subtropical edges and the polar vortex are barriers to transport and mixing processes. Failing to reproduce the strength of these mixing barriers can lead to wrong distributions of long-lived and reactive trace gas species with potentially significant impacts on the ozone chemistry. The meridional profile of methane (or any other long-lived trace gas) reveals the existence of transport and mixing barriers in regions where tracer gradients are large. On the other hand, small tracer gradients indicate regions of strong mixing.

Only three instruments participating in the SPARC Data Initiative measured CH_4. The conclusions of *Chapter 4*, supported by other validation studies from the literature, suggest to treat the two MIPAS retrievals (high-spectral and high-spatial resolution) as two different instruments, hence **Figures 5.1.10** and **5.1.11** include four datasets each.

Figure 5.1.10 shows the meridional profile of methane at 50 hPa. The uncertainty range in step 1 is relatively large, especially in the SH polar vortex region, where the diagnostic is used to test the relative strengths of mixing across the polar vortex edge versus descent within the polar vortex. Removing multi-annual mean values with an interannual variability larger than the ±2σ standard deviations from step 1 and accounting for interannual variability yields a much smaller uncertainty range. This uncertainty range compared to the one used in *Eyring et al.* [2006] is shown to have improved in two aspects. First, the strong gradient across the polar vortex edge is much better defined than by using HALOE measurements alone. Second, HALOE mean values are much lower than the new multi-instrument mean values, in particular within the polar vortex region. The models (from Figure 5 in *Eyring et al.* [2006]) would hence compare much more favorably to the new instrument mean than to the old measurement diagnostic derived from HALOE. Note that the HALOE reference does not improve using a more limited range of years (*e.g.*, 2003-2005), but loses latitudinal coverage due to increasing sampling limitations towards the end of the mission.

Figure 5.1.11 shows the meridional profile of methane at 30 hPa. This level is chosen in order to illustrate that the comparison between the HALOE reference (as calculated in an equivalent way to that used in the CCMVal report at 50 hPa) and the multi-instrument mean and standard deviation from the SPARC Data Initiative datasets is altitude dependent. The comparison has much improved in terms of latitudinal structure, although the HALOE mean values are still generally somewhat lower than those of the other instruments.

5.1.3 Recommendations for short-lived species

Short-lived species are characterised by chemically driven variations linked to the local solar time (LST). Limb-viewing instruments measure at LSTs that can differ from instrument to instrument, and between seasons and latitudes for the same instrument. Most of the instruments measure two distinct LSTs per latitude. These instruments are in polar sun-synchronous orbits, with one LST for the ascending portion of the orbit and one forthe descending portion. In the case of solar occultation sounders, measurements correspond to sunrise and sunset as seen from the satellite and the LSTs shift with the day of year.

The SPARC Data Initiative produced two types of climatologies for the diurnally varying species; climatologies from observations binned by LST (unscaled), and climatologies from observations scaled to a common LST. The climatologies from instruments in a sun-synchronous orbit are generally based on measurements separated into am and pm data. Climatologies from instruments that observe from non sun-synchronous orbits are generally separated into daytime and night-time measurements. Exceptions are the climatologies from solar occultation measurements which are based on data separated into local sunrise and sunset measurements. Additional climatologies are compiled using a photochemical box model to scale the measurements to a common LST, as explained in detail in *Section 3.1.2*.

When evaluating short-lived species from chemistry-climate models with the SPARC Data Initiative climatologies, the comparisons will be meaningless in most cases, if the monthly zonal mean model output is constructed in the traditional way by averaging over all longitudes at each output time step. Since most of the SPARC Data Initiative climatologies correspond to specific LSTs or times of day, the model output needs to be sampled in a similar manner. Even for instruments like SMILES, that observe species at varying LST because of their non sun-synchronous orbit, the constructed zonal mean climatologies are biased towards particular LSTs as a result of the non-homogeneous sampling patterns [*Kreyling et al.*, 2013]. Ideally, model data should be sampled with the satellite sampling patterns including the position and LST of each measurement. Trace gas climatologies derived from thus sampled model fields can be directly compared to the trace gas climatologies from the respective satellite instrument. While this approach is well suited for the comparison of short-lived species, it also means a lot of effort given that each satellite instrument has a different sampling pattern. Alternatively, the model output could be filtered according to LST in a manner similar to the SPARC Data Initiative climatologies in order to construct datasets corresponding to a particular LST, am/pm, day/night, or local sunrise/local sunset conditions. Another possibility is to restrict comparisons between model and satellite climatologies of short-lived species to latitude and altitude regions where the diurnal variations are small. Guidelines for appropriate comparisons of the individual short-lived species are given below.

- NO measurements show strong gradients at sunrise and sunset and model output should be filtered to construct sunrise and sunset (comparable to ACE-FTS or HALOE) or 10am LST (comparable to MIPAS or scaled ACE-FTS) climatologies.
- NO_2 diurnal variations are also most pronounced during sunset/sunrise. Model data should be filtered in order to construct sunrise/sunset (comparable to HALOE, SAGE II, POAM III, SAGE III and ACE-FTS) or 10am/10pm LST (comparable to MIPAS, SCIAMACHY, GOMOS or scaled OSIRIS, HIRDLS, and ACE-FTS) climatologies. If the model output is binned into daytime or night-time data instead (comparable to MIPAS, SCIAMACHY, GOMOS, OSIRIS, HIRDLS am/pm) differences of up to 20-30% can arise from the diurnal variations.
- NO_x is longer lived and has small diurnal variations in the MS. Data should be filtered to construct sunrise/sunset (comparable to HALOE and ACE-FTS) or 10am/10pm (comparable to MIPAS, SCIAMACHY, or scaled OSIRIS, and ACE-FTS) climatologies. Comparison of unfiltered monthly zonal mean climatologies can result in differences of around 20%. Binning the model output into daytime/night-time will not improve the comparison since there are no pronounced gradients at sunrise/sunset.
- HNO_3 is fairly long-lived in the UT to MS and shows a weak diurnal cycle in the US which increases further in the LM. Zonal mean climatologies can be compared directly at altitudes below 3 hPa.
- HNO_4, $ClONO_2$ and N_2O_5 climatologies show strong diurnal cycles above 10 hPa (100 hPa for N_2O_5) where model data needs to be binned according to sunrise/sunset (comparable to ACE-FTS) or 10am/10pm data (comparable to MIPAS). Below 10 hPa (100 hPa for N_2O_5), diurnal variations are weak allowing for a direct comparison of the datasets corresponding to different LSTs.
- ClO and BrO exhibit strong diurnal variations most pronounced during sunset/sunrise and with decreasing amplitude towards the USLM. Daytime variations are much smaller than night-time variations. For ClO, model data should be filtered in order to construct sunrise/sunset (comparable to SMR) or daytime climatologies (comparable to Aura-MLS pm, SMILES daytime, MIPAS am, or scaled daytime SMR climatologies). Comparisons should focus on the tropical/mid-latitude US. For BrO, model output should be filtered to construct daytime climatologies (comparable to scaled OSIRIS, scaled SCIAMACHY, or daytime SMILES climatologies). Comparisons should focus on altitude levels above 20 hPa. HOCl shows in contrast strong diurnal variations and model data need to be compiled according to instrument measurement times for a more meaningful comparison.

- HO_2 shows a strong diurnal cycle with smaller variations during daytime than during night-time. Model data can be binned into daytime climatologies (comparable to SMILES and Aura-MLS daytime) and compared in the altitude region between 10 and 0.5 hPa. OH has a strong diurnal cycle and model output should be filtered in order to construct daytime 2pm climatologies (comparable to Aura-MLS). CH_2O and CH_3CN show small diurnal variations, thus allowing for a direct comparison of datasets even if they apply to different LSTs.

5.1.4 Suggestions for new diagnostics

The monthly zonal mean SPARC Data Initiative datasets provide a unique source of observational data for model evaluation diagnostics. Here, we present suggestions for new diagnostics covering different aspects of model validation. The new diagnostics use, in addition to the monthly zonal mean climatologies, parameters from the SPARC Data Initiative datasets that describe variability, location and timing of the underlying measurements.

CFC-11 mean profiles and standard deviations

Profiles of long-lived tracers (as also shown in *Section 5.1.2* for CH_4) have been used extensively over the past to analyse the effects of diabatic descent and mixing in the polar vortex [*SPARC,* 2010]. Here, we show CFC-11 profiles at the high SH latitudes (80°S-85°S) at the beginning (June) and end (September) of the Antarctic winter for MIPAS and the Whole Atmosphere Community Climate Model (WACCM) (**Figure 5.1.12**). The comparison of June and September CFC-11 profiles provides information on the combined effects of vortex descent, bringing lower CFC-11 mixing ratios downward, and of transport from lower latitudes, bringing higher CFC-11 mixing ratios towards the pole. Between 100 and 50 hPa, WACCM shows lower mixing ratios at the beginning of the austral winter but higher mixing ratios at the end of the winter when compared to MIPAS. Thus the CFC-11 decrease is not strong enough in the model suggesting that there is too little descent and/or too much mixing across the vortex edge.

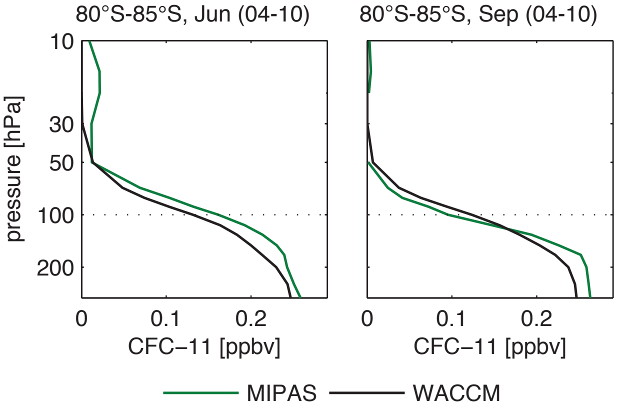

Figure 5.1.12: Vertical monthly zonal mean CFC-11 profiles for 80°S-85°S in June and September for MIPAS and WACCM.

Besides the monthly mean values, the SPARC Data Initiative datasets provide the standard deviations for each month, latitude bin and pressure level. **Figure 5.1.13** shows the standard deviation fields which describe the variability within each latitude band and month and are calculated over all given measurements in the respective bin. At 30 hPa (upper panels), elevated standard deviations at around 20°S/N indicate stronger variability in the trace gas field caused by breaking of planetary scale waves at the tropical pipe edge. The temporal extent (in the NH from December to March and in the SH from April to November) and magnitude of this event agree quite well between model and observations. Note that, at altitudes below 70 hPa, breaking synoptic scale waves cause more stirring and therefore prevent strong tracer gradients or any maxima in the standard deviation field. At 100 hPa in the tropics (~20°S-20°N), WACCM shows similar mean values but much lower standard deviations than MIPAS, which is very likely caused by the natural variability in this region being smaller than the MIPAS measurement error [*Toohey et al.*, 2010]. Most of the MIPAS variability is indeed explained by the MIPAS random error estimated to be around 17 pptv. Consequently, the standard deviation from observational fields should only be used for model evaluations in regions where the natural variability is larger than the measurement error. However, at 30 hPa the comparison reveals a striking absence of variability in the model in the SH high latitudes throughout the year, but in particular during SH winter, when the observations show high variability. This result implies a too low dynamical activity in the model, which may be related to the SH cold bias chemistry-climate models exhibit in this region [*Austin et al.*, 2003].

The comparison of the standard deviation fields from MIPAS and WACCM at 100 hPa (**Figure 5.1.13**, lower panels) reveals the absence of a mixing minimum during summer in the SH mid-latitudes in the model. The SH vortex edge region shows comparable variability during SH late winter, but higher variability in the model in early winter. The situation reverses at the very high SH latitudes, where the model has much lower variability over most of the year. In particular, the low standard deviations in the model during the winter from June to September suggest that the inner vortex south of 70°S in WACCM is less disturbed than implied by the MIPAS observations. Thus the missing decrease of the WACCM CFC-11 profiles in the vortex during winter (seen from the profile comparisons in **Figure 5.1.12**) is probably caused by too weak diabatic descent and not by too strong in-mixing. MIPAS on the other hand, shows elevated standard deviations during the winter related to zonal asymmetries in the CFC-11 field which can be either caused by asymmetric descent or by in-mixing. One to two months after the vortex breakdown the standard deviation of the CFC-11 field increases due to longitudinal asymmetries. This phenomenon can be observed earlier in MIPAS (December) than in WACCM

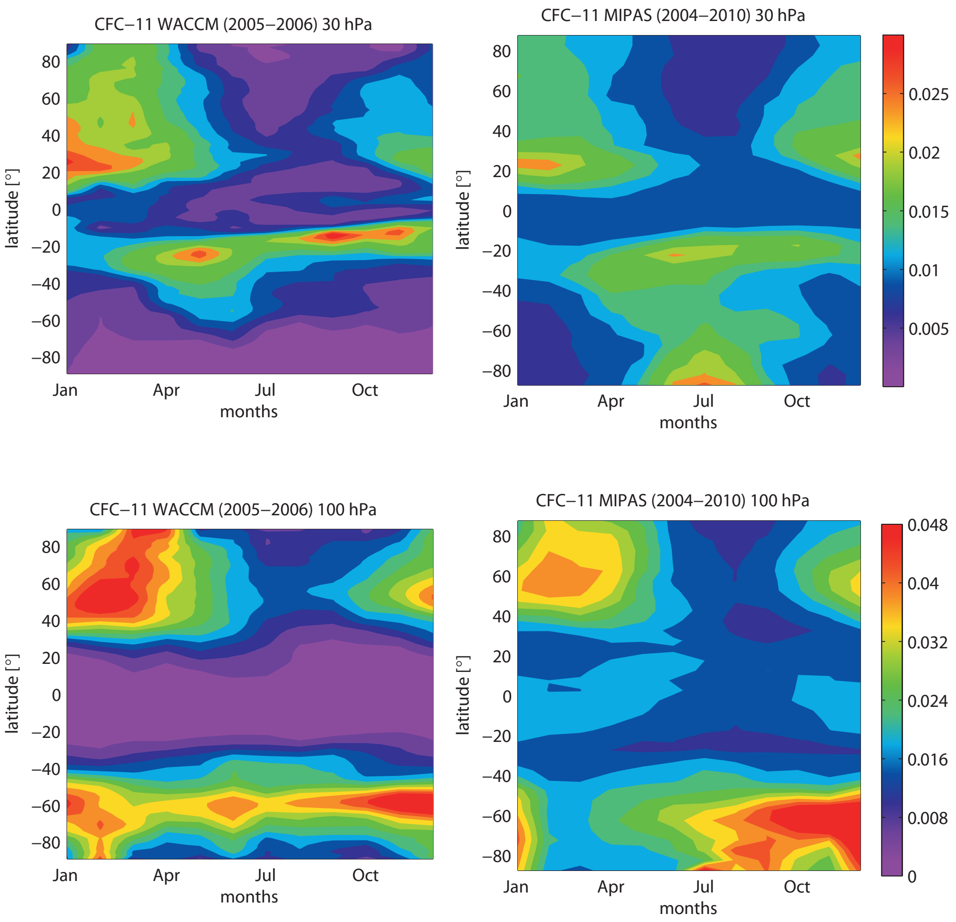

Figure 5.1.13: ***Time-latitude cross-sections of CFC-11 standard deviation fields for MIPAS and WACCM at 30 hPa (upper panels) and at 100 hPa (lower panels).*** *The standard deviation describes the variability within each latitude band and month and has been calculated over all given data points in the respective bin and month.*

(February) due to a late breakdown of the vortex in the model [*de laTorre et al.*, 2012]. At the NH high latitudes, the standard deviations show better agreement between observations and model suggesting more similarities in the dynamical situation.

CH_4 time-latitude evolution

While meridional and altitude profiles of CH_4 and N_2O have been extensively used in the past to test stratospheric transport in chemistry-climate models [*Eyring et al.*, 2006; *SPARC*, 2010; *Strahan et al.*, 2011], the SPARC Data Initiative monthly zonal mean climatologies lend themselves to also study the time evolution of these profiles. **Figure 5.1.14** shows to this end a comparison of the time-latitude evolution of CH_4 at two different pressure levels (2 and 10 hPa) between the multi-instrument mean derived from the HALOE, MIPAS, and ACE-FTS instruments, and the Canadian Middle Atmosphere Model using a simulation nudged to observed meteorology (CMAM30). As explained in more detail in *Chapter 4.3*, the feature at 2 hPa has been attributed to the equatorial Semi-Annual Oscillation (SAO) [*Choi and Holton*, 1991], with the maxima in tropical CH_4 coinciding with maxima in upwelling. The 2 hPa and 10 hPa levels are furthermore distinct in the CH_4 variability seen in the polar region. At 10 hPa, the minima in polar regions during autumn and winter coincide with the maxima in downwelling within the Brewer-Dobson circulation [*Randel et al.*, 1998]. At 2 hPa, on the other hand, the minima show up in summer/autumn as the result of photochemistry, with CH_4 lifetimes decreasing to four months at these altitudes [*Randel et al.*, 1998; *Solomon*, 1986].

Comparison of CMAM30 with the observations yields overall encouraging results, with CMAM30 clearly indicating a SAO. Furthermore, the timing and extent of the low CH_4 in polar regions correspond well between observations and model at both levels. However, some differences can also be identified. For example in both hemispheres at 2 hPa, the photochemically induced minima during autumn are not quite as pronounced as in the observations. This could be due to a problem in the chemistry, but more likely results from too strong mixing between the tropics and the higher latitudes (partially due to numerical diffusion in the rather low model resolution). Likewise, the maxima seen in the tropics are not quite as pronounced as in the observations, along with the minima in polar regions at 10 hPa, indicating that CMAM30 exhibits too weak upwelling/downwelling or again too strong horizontal mixing. The overall good agreement between CMAM30 and the observations is partially due to using a model version that is driven by the observed meteorology. Note however that the influence of the nudging to the meteorological fields weakens towards higher altitudes above 10 hPa, so that the model seems to at least partially represent the right dynamical mechanisms that produce the SAO.

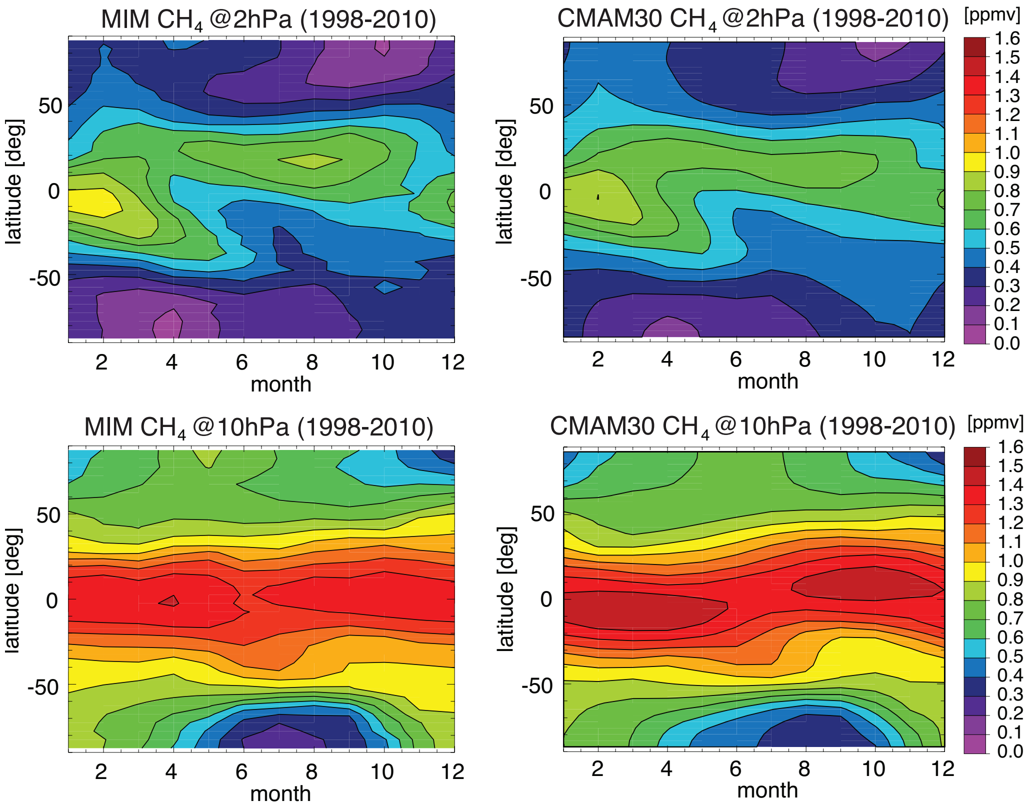

***Figure 5.1.14:** Time-latitude cross-sections of CH$_4$ mixing ratios at 2 hPa (upper panels) and 10 hPa (lower panels) from the multi-instrument mean (left) and CMAM30 (right).*

5.2 Implications for merging activities

With monthly zonal mean time series of stratospheric constituents available from all the SPARC Data Initiative instruments, the obvious question is why these have not been merged into one homogeneous data product which globally covers multiple decades. The reason is that such a project is a challenge in itself which requires solving a number of technical and methodological problems. One needs to try to eliminate outliers or even whole datasets if many problems are discovered (*e.g.*, after a careful multi-instrument comparison). Currently, there is not even full agreement about what the most appropriate merging techniques are. Techniques range from a simple merge of two single datasets by accounting for an inter-instrument bias that is calculated over some overlap time period [*Bourrassa et al.*, 2014] to merging of multiple datasets including detailed calculations of uncertainties [*Froidevaux et al.*, 2015], statistical methods to fill in observational gaps [*Bodeker et al.*, 2013], or the use of a nudged chemistry-climate model as transfer function between the instruments [*Hegglin et al.*, 2014].

Some of the problems arising in data merging can be solved by directly using the parent datasets instead of the merged dataset and using an analysis tool that is immune against one or the other of these problems. One example is the trend estimator by *von Clarmann et al.* [2010] which is immune against biases between subsets of data. An ideal solution for the general data merging problem, however, does not yet exist. The first important step towards optimal data merging is to develop a common language and to develop schemes to evaluate and report retrieval errors, altitude resolution and content of prior information in the data in an inter-comparable manner. Given that all different merging techniques have their weaknesses and strengths, it remains important that independent research teams approach data merging so that their results can be compared and used to identify not only instrument errors but also uncertainties in the merging techniques themselves.

In the following sections we discuss the most prominent problem areas that arise in data merging.

5.2.1 Error characterisation of instruments

In the most straightforward scenario, multiple datasets are available for the same latitude bins and certain overlap time periods. In this case merging reduces to a weighted or unweighted mean of the data. The obvious advantage of weighting the data by their inverse estimated error, usually in terms of variance, is that reliable data dominate the merged product. Drawbacks and pitfalls, however, are:

- The error estimation schemes used for the different datasets may differ and different error types may be included. Thus, a better instrument can have larger error bars.
- For some instruments error covariances are reported, while for others only error bars are available. Optimal averaging, however, requires the covariance matrices.

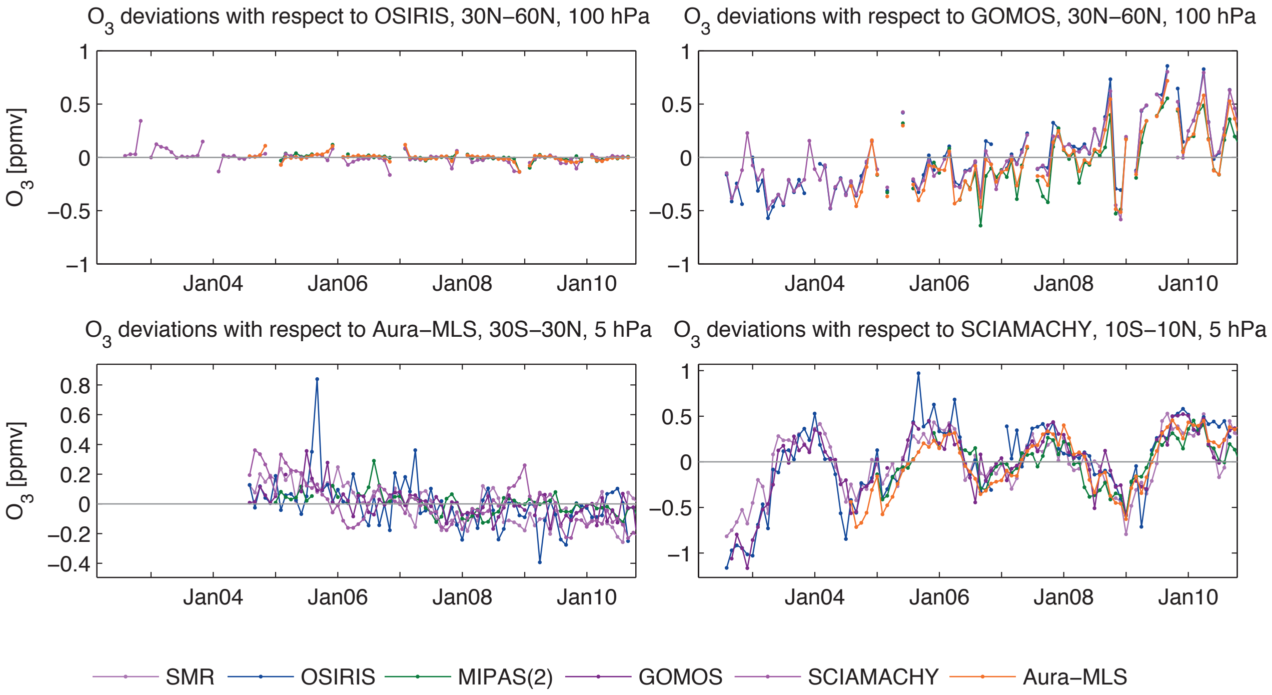

Figure 5.2.1: Time series of ozone deviations for 2002-2010. *Deviations of all instruments with respect to OSIRIS and GOMOS for 30°N-60°N at 100 hPa (upper panels), with respect to Aura-MLS for 30°S-30°N at 5 hPa (lower left panel) and with respect to SCIAMACHY for 10°S-10°N at 5 hPa (lower right panel) are shown.*

- For some instruments the error estimate includes by default the so-called smoothing error [*Rodgers*, 2000]. This quantity, however, does not follow the generalised Gaussian error propagation law [*von Clarmann*, 2014] and thus is not applicable to regridded data.

Trying to avoid these problems by using the sample standard error of the zonal mean instead of the error estimates is not as simple as one might think, because (a) in the case of regular sampling patterns measurements cannot be regarded as independent random samples and thus the standard error is not the sample standard deviation divided by the square root of the sample size [*Toohey and von Clarmann*, 2013]; and (b) sophisticated schemes are needed to distinguish the two components of the standard error of zonal means, namely measurement errors and natural variability [*Laeng et al.*, 2014; *Sofieva et al.*, 2014].

A particular problem is that the quality of the measurement can depend on the atmospheric state itself, *e.g.*, in infrared emission spectroscopy the signal is larger and thus the precision is better when it is warmer. Weighting by the inverse error variance in such a case would introduce a representativeness bias towards warmer parts of the atmosphere.

Another problem arises from denotation ambiguities. Many terms used for error characterisation are not clearly defined, used in different contexts, and have ambiguous meanings. Accuracy characterises in some cases the total error, in other cases only the systematic part, precision excluded. The systematic error in some documents includes all error sources except noise, in other cases only error terms which are - in amount and sign - time-independent. Noise often is referred to as the random part of the error while equally often it is used for the pure measurement noise only. Some total errors are more comprehensive than others. Some error budgets refer to an ideal point measurement and include the so-called smoothing error which characterises the expected difference between the atmospheric state at one idealised atmospheric point and in a finite air volume. Other error budgets refer to the atmospheric state at finite resolution and do not include the so-called smoothing error.

5.2.2 Drifts and jumps between datasets

Drifts within datasets are often unknown because, contrary to the usual validation measurements, drift estimation requires availability of long-term datasets. Even if these are available, it is not always clear which of the instruments compared to each other has a drift. For trace gases where a large number of instruments are available, such as ozone, the long-term changes of the differences can provide information on possible drifts [*Tegtmeier et al.*, 2013]. Therefore, for each instrument an analysis of the temporal variations of the differences with respect to each of the other instruments has been performed. Such time series are characterised by seasonal patterns and month-to-month variability. After removing the seasonal cycle, longer-term changes can be the dominant signal. However, for nearly all ozone datasets and regions included in this study the differences display no apparent long-term changes. One example for this consistency is shown in **Figure 5.2.1** (upper left panel) in the form of the instrument difference with respect to OSIRIS in the NH mid-latitude LS. A few exceptions exist where clear changes of the differences over time can be identified (**Figure 5.2.1**). First, differences of all instruments with respect to GOMOS in the NH mid-latitude LS are mostly negative before 2008 and mostly positive afterwards indicating a change of GOMOS over time that is not seen by the other instruments. Note that GOMOS is excluded from the comparison to OSIRIS discussed above in order to present one example where the differences display no apparent long-term changes. For Aura-MLS, some discrepancies can be observed for the tropical US, with positive differences at the beginning and negative differences at the end of the time period, although not all instruments agree on this. SCIAMACHY differences in the tropics are dominated by the quasi-biennial oscillation (QBO) signal, while SMR (not shown here) displays larger values compared to the other datasets in 2003 but differences around zero after 2006. Note that here only drifts of a magnitude comparable to the deviations themselves have been identified; while for trend studies a more thorough analysis including possibly quite small long-term drifts is necessary.

Another option is the comparison of the instruments' time series with a model (used as a transfer function in the merging) which can yield additional evidence for which instrument is more likely to show a drift or a jump [*Hegglin et al.*, 2014]. An example for this is shown in **Figure 5.2.2**, where two data versions of SAGE II (v6.2 and v7.0) are compared to each other and a distinct difference in the beginning of the data record is revealed. The very good agreement between the maroon-coloured data version (v7.0) and the model (as seen in the bias-corrected differences fluctuating randomly around zero) provides the user with confidence that the red data version (v6.2) suffers from an inhomogeneity at the beginning of its record and therefore should not be used for merging during this time period.

Finally, the choice of well-established *in-situ* measurements as reference instruments, which usually are trusted more than remote sensing instruments, leads to the problem of often lacking statistical significance and representativeness due to low data amounts. Despite this shortcoming, ground-based measurements of ozone from sonde and lidar networks have been shown to allow for comprehensive analysis of the long-term stability of satellite ozone datasets [*Hubert et al.*, 2015]. A complication of all these types of validation studies is that the reference instrument (or model) itself needs thorough validation.

For certain regions and/or time periods, available datasets do not overlap in time. In this case, it is not clear if any jumps in the data reflect natural variability or instrument biases. A model as a transfer standard again can help here [*Hegglin et al.*, 2014]. While this approach may be seen as contaminating an otherwise purely empirical product with model information, it capitalises on our physical knowledge of the atmosphere and provides at least a best estimate of what happened during a time period when observations were not available.

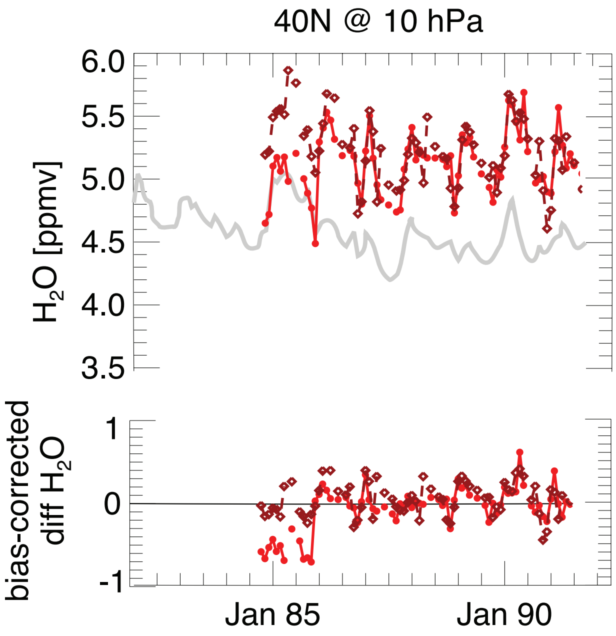

Figure 5.2.2: Model-based bias and drift estimates of observational data. *Time series of water vapour (upper panel) and bias-corrected difference to the model (lower panel) at 10 hPa and 40°N from two data versions of SAGE II water vapour (with red indicating v6.2 and maroon dots v7.0) and the Canadian Middle Atmosphere Model CMAM nudged to observed meteorology (grey line). Irregular behaviour in the bias-corrected differences reveals a problem in the red data version.*

5.2.3 Altitude resolution and a priori information

Key problems in any application where remote measurements of multiple instruments are considered are different altitude resolutions and different content of a priori information in the datasets. Some of the related problems can be solved by application of the averaging kernel matrix [*Rodgers*, 2000]; *e.g.*, the averaging kernel matrix can be used to degrade the altitude resolution of a high-resolution profile to make it comparable to a lower-resolution profile [*Rodgers and Connor*, 2003]. Such an approach has been applied in the SPARC Data Initiative when comparing the limb-viewing instruments with the nadir sounder TES in order to cross-validate ozone distributions in the UTLS with an independent dataset [*Section 4.27*; *Neu et al.*, 2014a]. TES measurements have been well-validated against ozonesondes in the UTLS and the dataset is frequently used for the evaluation of tropospheric ozone in chemistry-climate models. Observations of the higher vertical resolution limb sounders have been smoothed using the observational operator of TES. In the tropical UTLS, large positive biases of up to 50% have been identified for the limb-sounders with respect to the TES. While this study successfully provides a common basis for comparison of the large-scale ozone morphology in the UTLS, a couple of general problems remain unresolved for the general application of such comparisons:

- There exists a large number of datasets for which no averaging kernels are available, and can - due to the particular retrieval scheme used - not easily be produced.
- The application of the averaging kernel fails if the better resolved profile does not have sufficient altitude coverage to allow this operation for all relevant altitudes. There exist ad hoc solutions to this problem but these are not exact (see *Section 4.27* or *Neu et al.* [2014a]). In the SPARC Data Initiative evaluations, the TES a priori has been used to fill in the profiles below the lowest measurement level. To identify regions where the results are highly sensitive to this approach, virtual retrievals using two different filling methods have been calculated and compared.
- The situation is even worse if the altitude resolution of a measurement depends on the atmospheric state. This causes artefacts in estimated trends [*Yoon et al.*, 2013] or amplitudes of annual cycles (see *Section 4.2* or *Hegglin et al.* [2013]).

5.3 Implications for future planning of satellite limb-sounders

Past observations from limb satellite sounders have provided us with invaluable information on the chemistry (*e.g.*, *Waters et al.* [1993]; *Santee et al.* [1998]), transport (*e.g.*, *Park et al.* [2007]; *Stiller et al.* [2008]; *Hegglin et al.* [2009]; *Gille et al.* [2014]), and dynamics of the stratosphere (*e.g.*, *Randel et al.* [1993]; *Manney et al.* [2009]). This information has helped us understand many key aspects of the processes involved in stratospheric ozone depletion, the Antarctic ozone hole, and climate change. While we had a wealth of stratospheric limb observations during the past 30 years, it now has to be expected that there will be a lack of adequate limb measurements in the near future. This looming problem is due to an ageing fleet of currently still flying limb sounders (Aura-MLS, ACE-FTS, ACE-MAESTRO, OSIRIS and SMR) along with the lack of any concrete plans to launch new instruments except for SAGE III on the International Space Station (ISS) (which offers only limited spatial coverage) and the OMPS instruments (which only measure O_3, NO_2, and aerosol). These instruments may not be able to provide continuous temporal coverage, due to a nominal mission duration of Suomi NPP until 2016 and a replacement of the limb-viewing OMPS capacity on JPSS-2 in 2022 only.

The evaluations within the SPARC Data Initiative illustrate that there is no single best instrument that potentially covers all measurement needs, because instruments differ greatly in their measurement characteristics such as spatial and temporal sampling, viewing geometry, accuracy and precision, and measurement stability (*Chapters 2* and *3*). It is only through careful comparison between the instruments as done in this report that outliers can be detected, and weaknesses and strengths of instruments in measuring different species can be identified. An example is the realisation that SAGE II offers a valuable water vapour product that helps to extend the water vapour record from satellite observations (in particular HALOE) back to the late 1980s and also to improve this climate data record more generally [*Chapter 4.2*; *Hegglin et al.*, 2013; 2014].

Our evaluations also demonstrate clearly that there is no single instrument that can provide measurements of the full suite of atmospheric trace gas species with a high vertical and horizontal resolution, high accuracy and precision, and dense data coverage. Only a comprehensive set of high quality instruments that are complementary with respect to data coverage and target species allows development of a global picture of stratospheric composition. Such datasets enable among other things the analysis of temporal variations on different time scales and the quantification of important chemical budgets *e.g.*, of the chlorine family. As discussed in the previous *Section (5.2)*, data merging, even in the case of multiple overlapping instruments, poses a real challenge and complicates our understanding of long-term changes of the stratosphere in a changing climate. The future scenario we are currently facing with no overlap between instruments will render it impossible to derive reliable long-term changes of atmospheric trace constituents such as water vapor, ozone, and aerosol, and other important transport tracers.

Not only water vapour, but also other chemical trace gas species can be difficult to measure, especially when their atmospheric mixing ratios are close to the instruments' detection limits. Where agreement between instruments is found, the atmospheric mean state distributions and variability of trace gas species can be considered well-known (ozone [*Tegtmeier et al.*, 2013], water vapour [*Hegglin et al.*, 2013], N_2O, and CH_4 [*Hegglin et al.*, in prep.]). However, for other species that are measured by a few instruments only and for which not many ground-based validation measurements are available, our knowledge is still limited (many short-lived species such as HO_2, OH, BrO, ClO, *etc.*). It is key for the future planning of satellite limb sounders to design measurement systems that not only fit the purpose of covering specific measurement needs (in terms of scientific research question, region of interest, resolution, accuracy and precision, species list required), but also that offer redundancy between measurements, so that problems can be identified and adequately investigated.