

Review #1.

We thank the referee for the thoughtful review and comments that helped us to improve the manuscript. We respond to each comment below.

Referee: p2251, par1: Add a note on the required detection level for the O3 trend. This would help evaluate whether the SBUV instruments are sufficiently stable to detect such a trend. Right now, at the end of the paper, an SBUV instrument is called stable when it is drifting less than 1% per year. Which is quite a lot in my opinion with respect to the actually expected O3 trends.

Sec. 6 par 6., How stable does the instrument need to be in order to detect the expected stratospheric O3 trends?

Response: The required stability of the long term ozone data set for detecting ozone trends must be 1-3% per decade according to the User Required Document of the European Space Agency Ozone Climate Change Initiative. The short-term drifts presented here are for validation purposes only, as each individual record is too short to be used for long-term ozone trend analysis. For long-term trend computations, the relevant stability is the stability across the entire record of SBUV instruments. Work is ongoing to use the estimated drifts from each of the SBUV instruments presented in this study to estimate the potential long-term drift of a merged SBUV record (Frith et al., in preparation 2013).

Long periods of overlapping data and sufficient sampling are required to accurately estimate a relative drift and to reduce the standard deviation of the slope. We included Fig. S19 in the Supplement that shows a drift between N17 and MLO microwave as a function of the length of overlap between two instruments. Longer overlap significantly reduces the standard deviation of the derived slope. Also, over shorter time periods drifts can be larger (compared to the longer overlap periods) and significant at the 2σ -level due to short-term variability in the time series of differences. We included reference to Fig. S19 in section 2.4.4: "Long overlapping time periods and sufficient sampling are required to accurately estimate the relative drift and to reduce the standard deviation of the slope (see Fig. S19 in the Supplement)."

Referee: p2553, par3: Add the values of the 21 SBUV pressure levels. They are likely mentioned in (Bhartia et al., 2012), but it would certainly be useful to mention them here as well.

Response: We included two tables in the Supplement (tables S1 and S2) that show pressure scales for 21 partial ozone layers and 15 vmr levels. We put references to the tables in the text where appropriate (p. 2553 l. 19, p. 2555 l. 9 and p. 2559 l. 12).

Referee: p2554, par1: Add a reference with the background and motivation of the SBUV data screening.

p2554, par1: Clarify what happens if the mean latitude of O3 profiles is not within 1 degree of the band centre. I assume the monthly zonal bin is discarded.

Response: We screen data for mzm to reduce the noise and to ensure the mzm is adequately sampled. In general, 5% or fewer (in most cases 1% or fewer) individual profiles are rejected.

We slightly modify the text in this paragraph. More information about the screening procedure can be found in the corresponding readme file: http://disc.sci.gsfc.nasa.gov/daac-bin/DataHoldingsMEASURES.pl?PROGRAM_List=RichardMcPeters.

We added: " To create mzm profiles all level 2 ozone profiles are screened to ensure high quality, and only profiles with an error flag of 0 (no flag) or 1 (solar zenith angle in the 84-88° range) are accepted. In general, 5% or fewer (in most cases 1% or fewer) individual profiles are rejected. We also require that the mean latitude of measurements within each latitude band are within 1 degree of the center of the band, and similarly that the mean time of measurements within a given month is within 4 days of the center of the month (i.e., day 15) to ensure that the mzm is adequately sampled. The mzm is not computed if these criteria are not met. "

Referee: p2551, par1: Has the bias between ascending/descending O3 profiles in a (latitude, month) bin been investigated?

Response: Measurements from the same instrument for both ascending and descending modes are only possible in the polar regions during the boreal summer. In this study we limited our research by 50S-50N latitude range, and we did not asses biases between ascending/descending O3 profiles.

Referee: Sec 2.2: Add bias and precision estimates for each of the reference instruments, sometimes none or only one is mentioned. Given the importance of the study of SBUV stability in this work, it is imperative to mention drift estimates for the reference instruments as well.

Response: We added reported precision and accuracy for each type of instrument. We also added estimates of relative drifts where possible. However, drift estimates are often limited by specific time period or location and are not available for all instruments, especially ground-based instruments.

Referee: Sec 2.2.2: (Livesey et al., JGR 2003) recommend a) not to use UARS MLS V5 data after mid-1998 for trend analyses, and b) warn to be cautious with data from mid-1997 to mid-1998. MLS data from this period is used in this work, at least a comment on the possible instability of MLS data is in place.

Response: Thank you for the comment. We put comments about that in section 5.1: " N11 ascending and descending drifts relative to UARS MLS are not evaluated. N11 ascending data do not have sufficient overlap because of data loss after the eruption of Mt. Pinatubo (1991-1992) and limited spatial coverage as the N11 orbit approaches the terminator (1994-1995). N11 descending drift estimates are also not computed because the overlap is in the period after mid-1997 when UARS MLS data quality is reduced and should be used with caution in trend analyses (Livesey et al., 2003). However, we use UARS MLS data after 1997 when computing drifts for N14 ascending to increase the statistical significance of the results. "

Referee: Sec 2.2.3: I assume the vertical oscillations seen in Tropical UTLS Aura-MLS V3.3 data are not too relevant for your work, given that ozone is integrated over 250-16hPa. Is that correct?

Response: Yes, this is correct.

Referee: Sec 2.3: Add a short description of the ozonesonde data used to validate the tropospheric column (Sec. 4).

Sec 4, par3: There is no real description of the ozonesonde data, nor of the preprocessing + collocation criteria. These should be added in Sec 2.3, 2.4.1 and 2.4.3.

Response: We added section 2.3.4 where we briefly discuss ozone sonde data. Additionally, in sections 2.4.1-2.4.3 we described coincidence criteria applicable to comparisons with sondes.

“2.3.4 Ozone sondes

To validate SBUV measurements in the troposphere and lower stratosphere we chose 4 ozone sonde stations based on their long time records and their proximity to the northern midlatitude Umkehr stations. We use data from Boulder, Colorado, USA (40N), Hohenpeissenberg, Germany (48N), Lindenberg, Germany (52N) and Payerne, Switzerland (47N). Ozone sondes measure in-situ ozone concentration from the ground up to 30-35 km and report profiles of partial ozone pressures. Long-term ozone sonde measurements provide valuable information about ozone concentration in the troposphere. Two main types of ozone sondes have been in use at the stations used in this study: Brewer-Mast (Brewer and Milford, 1960) and electrochemical concentration cell (ECC) (Komhyr, 1969). Recent studies (e.g., Smit et al., 2007) demonstrates differences up to ± 5 -10% among different types of ozone sondes, that might affect the long-term stability of ozonesonde records if different types of sondes were used at a given station. Accuracy and precision of the ECC sonde measurements for altitudes below 30 km are ± 5 -10% and ± 3 -5% respectively (Smit et al., 2007).”

Referee: Sec. 2.3.2: Please add a (more recent) reference to a study of the bias and drift of 6 NDACC lidars (4 of which used in your work): Nair et al., 2012, AMT, doi:10.5194/amt-5-1301-2012.

Response: Thank you, we added in section 2.3.2 the following statement: "Recently, Nair et al. (2012) assess the performance of lidar measurements at 6 stations (including four stations considered in this study) relative to multiple satellite observations and demonstrate that biases and drifts are mostly within $\pm 5\%$ and $\pm 0.5\%$ per year."

Referee: p2558, 115: The resolution of lidars worsens with altitude, it reaches 3 km at 45 km altitude.

Response: Thank you for the notice, we corrected that in section 2.3.2: " The vertical resolution of lidars is about 0.3-0.5 km in the middle stratosphere, decreasing to 3-5 km around 45 km (Godin-Beekmann et al., 2003)."

Referee: P2558, 118: Mention that due to the 10 % error screening fewer measurements are available at bottom and top (above ~ 5 hPa) of the lidar profile. This information is useful in Sec 2.4.3.

Response: We added the following sentence: " The 10 % error screening significantly reduces the number of lidar measurements available at the bottom (below 40 hPa) and top (above ~ 5 hPa) layers. "

Referee: p2560, 11: SAGE-II and lidar data are used over long periods (>15 years) when considering the validation over all SBUV instruments. But not when single SBUV validation is done. Please clarify this, as the statement can be misleading.

Response: We re-phrased the sentence to make it clear that we are talking about validation of individual SBUV instruments: "Since we use SAGE II and lidar data for validation of individual SBUV instruments over comparatively short time periods,..."

Referee: Sec 2.4.2: The vertical resolution of the microwave instruments is worse than that of SBUV. Did you quantify the impact of comparing SBUV-MWR partial columns finer than the MWR-resolution?

Response: The vertical resolution of SBUV and ground based microwave instruments in the considered vertical range 25-1 hPa are very similar. The SBUV vertical resolution is about 6-7 km (Bhartia et al., 2012), and MWR vertical resolution is about 7 km between 22 and 44 km (Parrish, private communication). Thus measurements from these two instruments can be compared directly.

Referee: p2561, 128: SBUV profiles are weighted with distance from the correlative profiles. Is this spatial distance, or is a temporal component included as well? If yes, how? Do you have an idea of the horizontal smoothing error contribution?

Response: We weighted SBUV profiles only by the distance from the correlative profiles. We did not account for possible uncertainties due to difference in time of measurements within the established temporal coincidence criteria. We also did not specifically estimate the contribution of horizontal smoothing, but since we worked with the monthly mean profiles we assume that horizontal smoothing error is very small.

Referee: Sec 2.4.3: Clarify the space/time collocation criteria. Especially the spatial window is not clear to me. What I understood: - SAGE II: SBUV within ($\pm 1^\circ$ lat, $\pm 14^\circ$ lon, same day) - MWR: SBUV within ± 1.5 h at Mauna Loa and same day 9AM-5PM at Lauder - other: SBUV within ± 12 h

Response: We made appropriate changes in section 2.4.3 to clarify what exact criteria were used:

"Appropriate coincidence criteria in both time and space are very important for validation. Above 1 hPa diurnal ozone variation plays a significant role (e.g. Connor et al., 1994; Haefele et al., 2008), and time coincidence criteria should be stricter. For this reason we limit the vertical range of the validation to 1 hPa and below for all instruments except lidars, where we limit the upper range to 1.6 hPa due to the reduced number of lidar measurements above this altitude.

The spatial and temporal coincidence requirements vary depending on the spatial and temporal resolution of the external instruments. SBUV, UARS and Aura MLS all have good spatial resolution with sufficient sampling to produce representative monthly zonal mean values. Thus when comparing SBUV to MLS we simply compute mzm values for each instrument and compare them directly.

SAGE II data have comparatively poor spatial/time coverage, so in this case we subset the SBUV dataset to match the SAGE space/time coverage. For each SAGE profile we find all

SBUV profiles within ± 12 hours period and within ± 1 degree latitude and ± 14 degrees longitude. This is typically 1-3 SBUV profiles. When more than one SBUV profile match is found we average the profiles using a linear weighting by distance from the SAGE profile location. Then we construct monthly zonal means from the SAGE and sub-sampled SBUV for comparison.

We use the same procedure when comparing SBUV to ground-based instruments. We require at least five coincident profiles to calculate monthly means for ground-based microwave data and two profiles for lidar, sonde and Umkehr data. On average we typically have about 15 coincident microwave profiles, between 2 and 20 Umkehr profiles, between 2 and 15 lidar profiles and about 2-5 ozone sonde profiles each month. Measurements from the ground-based microwave spectrometer at Mauna Loa are available at high time resolution, so for these comparisons we restrict the time difference to ± 1.5 hours. The microwave instrument at Lauder also measures ozone profiles at high time resolution, but the number of profiles that satisfy ± 1.5 hour coincident criteria is too low for statistical significance. Instead we calculate the daily average using all measurements between 9 am and 5 pm local solar time to get a sufficient number of profiles.”

Referee: Sec 2.4.4: The bias is calculated as mean of absolute differences relative to a fixed x_a (I assume monthly zonal values?). While it could be calculated as mean of percent differences, where every term is relative to \widehat{X}_{ext} . Do you expect large differences? And is the a-priori reference more suitable in that case?

Response: Since we compared multiple pairs of instruments, we used the a priori as a reference to make it easy to cross-validate data. In addition normalization by the a priori in the drift computations eliminates the uncertainties related with the possible drift in the reference data set.

Referee: Sec 2.4.4: Eq. (2) represents the (biased) standard deviation of the absolute difference $\widehat{X}_{sbuv} - \widehat{X}_{ext}$, not the standard deviation for the relative bias as defined in Eq. (1). In general, I found it sometimes difficult to follow which standard deviation is referred to: is it sample standard deviation (from Eq.2), or bias standard deviation $\frac{\sigma}{\sqrt{N}}$ (with σ from Eq.2)? I assumed all results/figures refer to the standard deviation of the bias. In that case I would replace Eq.2 by the expression for bias standard deviation and mention in the text that this estimator will be used throughout the rest of the text.

Response: We agree that Eq. 2 represents the standard deviations of the differences. We corrected that in section 2.4.4. Throughout the text whenever we mention "standard deviation" we meant σ calculated from Eq.2. We also use the term "standard error of the mean" or "standard error of the bias", that is equal to $\frac{\sigma}{\sqrt{N}}$ and we always note that in the text where applicable.

Referee: Sec 2.4.4: How do you calculate the standard deviation of the relative differences? Is it $\frac{\sigma}{x_a}$ (with σ from Eq.2)? In that case, the standard deviation of the relative bias would become $\frac{\sigma}{x_a\sqrt{N}}$?

Response: Yes, this is correct. We normalized the standard deviation by the a priori.

Referee: Sec 2.4.4: The explanation of the drift calculation should be clearer. Do you deseasonalize the SBUV and EXT timeseries separately, compute differences, and then regress? Or do you compute the differences, deseasonalize SBUV-EXT and then regress? I understood the 1st method, but the text is not entirely clear on that (the next phrase mentions that “the anomalies (i.e. SBUV-EXT) are deseasonalized”).

Response: Thank you for the comment. We indeed deseasonalized both data sets independently and then calculated differences and fit the regression. We re-arranged words in the corresponding paragraph to make it clear: "We deseasonalize anomalies to reduce persistence in the time series of residuals. This way we can assume that the residuals are random and normally distributed. Then we compute the time series of differences between the pair of deseasonalized anomalies and linearly regress the difference time series at all altitudes (see example in Fig. S14 in the Supplement)."

Referee: p2563, 118-20: Did you check whether the fit residuals are Gaussian?

Response: For long time periods (more than 5 years), distributions of residuals tend to look like a normal distribution. However, the overlapping time periods for N9 and N11 instruments are very short (~ 2-3 years). The 2σ error bars capture the uncertainties associated with the short sampling.

Referee: Sec 3.1, par2: The seasonal signature (1st four lines) is also discussed in Sec 3.1, par 8. Maybe move these lines to the end of the section?

Sec 3.1, par2: The “shorter overlap” sounded strange at the first reading, since SAGE and UARS-MLS have a similar overlap as Aura-MLS for individual SBUV (>~5 years). The poorer spatial and temporal sampling of SAGE and UARS-MLS leads to larger (>2x) standard deviations in the differences (evident from Fig. A1), which would make it more difficult to discover a seasonal cycle in the differences. Error bars are larger, but I did not really find that Fig. A9-10 are inconsistent with Fig. A8.

Sec 3.1, par2: In addition, I found it hard to see the seasonal cycle from Fig. 3, with its large temporal scale. Could you add a reference to the more useful figures A8-10?

Sec 3.1, par8: Partly mentioned in par2.

Response: We agree with the referee's remarks. We removed the first lines from section 3.1 par. 2 and move the discussion of the seasonal biases to par 8. We changed paragraph 8:

"We also estimate seasonal biases, defined as the difference in the seasonal cycles for the pair of instruments (see Supplement, Fig. S8-S10). Seasonal biases relative to Aura MLS are less than 2% in the tropics and mostly statistically insignificant (see Supplement, Fig. S8). Outside of the tropics we found a clear seasonal pattern relative to Aura MLS, though the amplitudes of seasonal variability are still mostly within 2-3%, increasing to 5-6% in the 10-6 hPa layer. There is an approximately 6-month lag between southern and northern midlatitudes. A clear seasonal signature can be also seen in Fig.3 in the time series of differences relative to Aura MLS in the extratropics of both hemispheres. We were not able to isolate clear seasonal structures from UARS MLS and SAGE comparisons, possibly due to the poorer spatial and temporal sampling. The amplitude of seasonal differences varies within $\pm 2-8\%$ and is mostly less than the 2σ standard deviations (see Supplement, Fig. S9-S10)."

Referee: p2565, 122-24: Fig. 3 suggests that the negative N11 bias for 4-2.5 hPa could be is mainly built up after ~1997, in the descending phase of the orbit. Is that correct?

Response: Thank you for your observation. Yes biases indeed are more negative for the descending portion of N11. We added: "Between 2.5 and 10 hPa biases are more negative for the descending portion of N11 (after 1998)."

Referee: Sec 3.1, par7: Are the larger std. dev. for UARS MLS and SAGE II larger due to a smaller sample size? (see also comment on *Sec6, par2*).

Response: The standard deviations are larger for comparisons relative to UARS and SAGE II partially due to the poorer sampling, but also due to lower quality of SBUV instruments in 1990s (N9, N11 and N14). We noted: "Larger standard deviations for N9, N11 and N14 are partially due to the poorer sampling with SAGE II and UARS MLS, but are also due to the lower quality of these SBUV instruments."

Referee: Sec 6, par2: Are you discussing standard deviations of the differences, or the standard deviations of the bias? If it is the latter, is the larger bias std. dev. for SAGE and UARS-MLS mainly due to a smaller sample size? (see also comment on *Sec 3.1, par7*)

Response: In the conclusions we are talking about standard deviations of the differences for the wide latitude band 50S-50N. We specified that in the text. We believe larger standard deviations relative to UARS and SAGEII are due to both poorer sampling and lower quality of SBUV instruments in 1990s. We added: "Standard deviations of the differences with Aura MLS in the wide latitude band between 50S and 50N are about 1-2%, while for comparisons with SAGE II and UARS MLS standard deviations range from 3-4%, likely due to poorer sampling and lower quality of SBUV instruments in 1990s (N9, N11 and N14)."

In addition to avoid any confusion we added the following explanation in the section 3.1 par. 6: "Figure 5 shows the altitude dependence of the mean biases for individual SBUV instruments averaged over the wide latitude zone 50S and 50N relative to (a) SAGE II, (b) UARS MLS and (c) Aura MLS. Biases, standard deviations (see Fig. S1 in the Supplement) and drifts (see sect. 5.1 below) for 50S-50N are calculated by constructing the 50N-50S area-weighted mean time series and finally applying the equations presented in section 2.4.4. We use this approach rather than calculating mean values from the biases, standard deviations and drifts for individual latitude bands to reduce the noise associated with the limited sampling at some latitude bands and to isolate the robust patterns that help to characterize the performance of individual SBUV instruments."

We also added in section 3.1 par. 3: "The standard deviations of the differences relative to independent satellite instruments for individual latitude bands are mostly within 5% (not shown here)."

Referee: Sec 3.2: Add references to the relevant time series and standard deviation plots in the Supplement.

Response: Thank you for the comment, we have done that and put appropriate reference throughout the text.

Referee: Sec 3.2.2, par4: The vertical structure of the bias standard deviations is explained by the lower lidar precision above 2-5hPa, and the fact that fewer profiles enter the bias calculation due to the 10% lidar precision screening.

Response: We added: "This again might be a result of the reduced number of lidar measurements at higher altitudes due to the 10% lidar precision screening".

Referee: Sec 4, par3: The qualitative structure of SBUV (Fig.10) and sonde (Fig.11) is quite similar. Could you add a comment on that?

Response: The comparisons in Figure 10 and Figure 11 are SBUV compared to Umkehr and SBUV compared to ozone sonde, respectively. Qualitatively the amplitude range of the differences are similar, and possibly the time scale of variations, but otherwise we do not see notable similarities. We also note that comparing the similarities or differences between these two plots is equivalent to comparing Umkehr and ozone sonde data, which is beyond the scope of this paper.

Referee: Sec 5.1, par3: Drifts are larger relative to SAGE II, but these are insignificant due to the larger error bars. Please clarify this in the text.

Response: The error bars shown in Fig. 13 indicate two times the standard deviations of the slope. Even though error bars are larger, drifts for some instruments at some levels are still significant at 2σ level.

Referee: Sec 5.2, par2: Is the "mean drift" the mean of regression at individual stations, or the regression of mean time series?

Response: We initially calculated the mean drifts, shown in figure 14, as the simple mean of regressions for each individual SBUV instrument relative to the specific type of ground-based instrument. But we realized that it would be more appropriate to weight the mean drifts by the corresponding standard deviations of the slope. We updated Fig. 14, however the results remain almost the same.

Referee: Sec 5.2, par2: Do the (2 sigma?) error bars in Fig 13-14 represent std. dev. of mean drift or the std. dev. of the drift sample?

Fig 13-14: Are error bars 1 sigma or 2 sigma? Mention this in caption. See also comment Sec 5.2, par2.

Response: Thank you for the comments. In Fig. 13 we show 2-sigma standard deviations. We add this explanation into the figure caption: "The horizontal error bars indicate two times the standard deviation of the slope."

In Fig. 14 we mistakenly showed 1-sigma standard deviation instead of 2-sigma, we fixed that in the updated figure. We clarified that in the text and in the figure caption: "The mean drifts are calculated as the mean of regressions at all considered stations weighted by the corresponding standard deviations. The horizontal error bars indicate the 2σ standard deviation of the slope."

Sec 5.2: The magnitude and vertical structure N16-N18 drift results are in good agreement for the Aura-MLS and microwave comparisons (to slightly lesser extent for lidars). Could you add a comment on that?

Response: Thank you for the observation. We have added in sec 5.2: "The magnitude and vertical structure of drifts for N16-N18 relative to ground-based microwave and lidars are consistent with the drifts estimated relative to Aura-MLS".

Referee: Sec 6: Could you add a comment on the latitude dependence of the bias and drifts for some instruments?

Response: In conclusions we tried to emphasize only those results which were consistent from comparison with one instrument to another and from one location to another. These differences we believe characterize uncertainties of the SBUV algorithm.

Referee: References: Please update (if any) the status of the papers in preparation/discussion/review.

Response: Done.

Referee: Fig 1: Add short phrase in caption that SAGE and MLS instruments are shown as well.

Response: We added: " Periods of operation for SAGE II, UARS MLS and Aura MLS are denoted at the bottom of the figure."

Referee: Appendix: Figures have different labels (A.xx) than the references in text (S.xx), please fix this.

Response: Following the technical editor recommendation, we decided to have an on-line Supplement file instead of an Appendix. We made appropriate corrections in the text and in the Supplement to reflect these changes. A correct reference would be Fig. S1 etc.

Referee: Appendix: Add y-axis label for Figs 2-4 and 8-13.

Response: Done

Referee: Appendix, Fig 6: "Larger deviations were detected for the upper layer due to the reduced number of lidar observations." See comment *Sec 3.2.2, par4*: the increased std. dev. at the top is due to lower precision of lidar and the (subsequently) reduced number of observations.

Response: Fixed it.

Referee: Appendix, Fig14: Not discussed in the text. If this is not planned, it should be dropped from the Appendix.

Response: We put reference to this figure in section 2.4.4.

Referee: Sec 5.1, par2: It is difficult to quantify the drift and its error bar from Fig 12.

Fig 12: In general, it is difficult to read magnitude of drift and error from the plot, since so many lines are superimposed. Is there a possibility to improve this figure, e.g. by slightly offsetting (in X) the error markers for the different SBUV instruments?

Fig 12: Replace “Percent” → “Percent per year” in label on Y-axis.

Fig 12: I assume that drifts estimates are shown for each 5° latitude band, while the error bars are only shown for 5 particular bands. Is that correct? Please clarify this in the caption. Upon 1st reading one could think that the drift results are for 20° lat bands.

Response: We agreed with the referee's comments regarding Fig.12. We decided to show drifts at two selected levels only. We also added Fig. S18 in the Supplement, which shows vertical profiles of the drifts for three latitude bands: northern mid-latitudes 30N-50N, tropics 20S-20N and southern mid-latitudes 30S-50S. These two figures together clearly demonstrate the main points discussed in section 5.1. In the updated Figure 12 we shifted error markers and added an explanation into the figure caption: "The horizontal error bars indicate two times the standard deviation of the slope. Error bars were calculated for each 5-degree latitude bin but are shown only for 5 latitude bins: 40-45S, 20-25S, 0-5N, 20-25N, 40-45N). Error markers for different SBUV instruments are slightly shifted relative to each other for easier viewing."

Referee: Fig 14: Replace “Drift, %” → “Drift, % per year” in label on X-axis.

Appendix. Replace “Drift, %” → “Drift, % per year” for Fig 15-17.

Response: Done

Technical corrections

Response: Authors are thankful to the referee and have accepted all technical corrections pointed by the referee.