Response to referee 1:
We thank the reviewer for his/her comments. Below are our responses in blue.

In the course of responding to the reviewers, two main changes occurred: (1) a trend discussion was added and (2) we emphasize the lack of MLS penetration below the upper troposphere throughout the paper.

Trend discussion

Before the conclusions the following paragraphs will be added:

As an example, Figure 8 shows the H_2O and O_3 trends in the tropics computed using monthly zonal mean deseasonalized anomalies of the raw model fields, as well as using all the available satellite-sampled measurement locations and only those passing the screening criteria in the tropics. As shown, when all available measurement locations are used, the MIPAS and MLS sampling allows accurate derivation of trends, with values matching those calculated from the raw model fields almost exactly. However, when only those measurements passing the screening criteria are used, both instruments have limitations: MIPAS trends are impacted because of the large percentage of measurements screened out below 100 hPa, which introduces non-negligible artifacts (up to 80% change for H_2O and up to 20% change for O_3); MLS trends are impacted because of the reduced vertical resolution, which limits its usefulness to the upper troposphere and above. Note that the impact of quality screening on MIPAS trends can be mitigated by using a regression model similar to the ones used by Bodecker et al. (2013) and Damadeo et al. (2014). These models have been shown to mitigate the effects of the non-uniform temporal, spatial and diurnal sampling of solar occultation satellite measurements. Furthermore, MIPAS trend analysis can be restricted to regions less affected by deep convection (for example, the mid tropical Pacific) to minimize the quality screening effects.

The estimated number of years required to definitively detect these trends is also shown in Figure 8. These estimates were computed assuming a trend model similar to the one described by Tiao et al. (1990), Weatherhead et al. (1998), and Millán et al. (2016), with a seasonal mean component represented by the monthly climatological means. As shown, with the MIPAS screened fields additional years are required for robust trend detection (up to ~150 years for H_2O and up to ~40 years for O_3 versus 50 years and 20 years, respectively, when all available measurements are used).

Similar analyses were performed for other latitude bands. Although the magnitude of the trends derived when using the MIPAS screened measurement locations was also impacted in these cases, no significant difference was found in the number of years required to detect such trends. In addition, no significant artifacts were found for HNO_3, CO or temperature for either the trend magnitude or the number of years required to detect such trends. Note that, when using real data, the effect of instrument noise upon trends will be negligible due to the vast number of MIPAS or MLS measurements associated with each monthly latitude bin. Drifts and long-term stability issues on these datasets [i.e., Eckert et al., 2014; Hubert et al., 2016; Hurst et al., 2016] will have to be corrected.

In the conclusion section, the trend discussion will be changed to:
These biases affect trends derived from these measurements using a simple regression upon monthly zonal mean data substantially affected by clouds. Further, the number of years required to detect such trends may increase due to the extra noise added to the time series by screening out measurements.

The following figure will be added (as figure 8 of the revised paper):

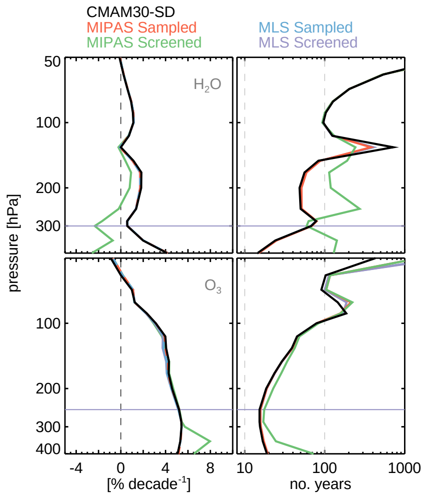


Figure 8: (left) H_2O and O_3 trends computed based on monthly zonal mean deseasonalized anomalies for the tropics (20S to 20N) using the raw model fields, all the available satellite measurements (MIPAS or MLS sampled) and only those measurements passing the screening criteria (MIPAS or MLS screened). Note that for O_3, we only use data starting from 2000 to capture the expected period of O_3 recovery. A purple line indicates the bottom (largest pressure) of the recommended range of the MLS retrievals. (right) Number of years required to detect such trends.

References:

Bodecker et al 2013: 10.5194/essd-5-31-2013
Damadeo et al 2014: 10.5194/acp-14-13455-2014
Tiao et al. (1990): 10.1029/JD095iD12p20507
Weatherhead et al. (1998): 10.1029/98JD00995
Millán et al (2016): 10.5194/acp-16-11521-2016

Lack of MLS penetration

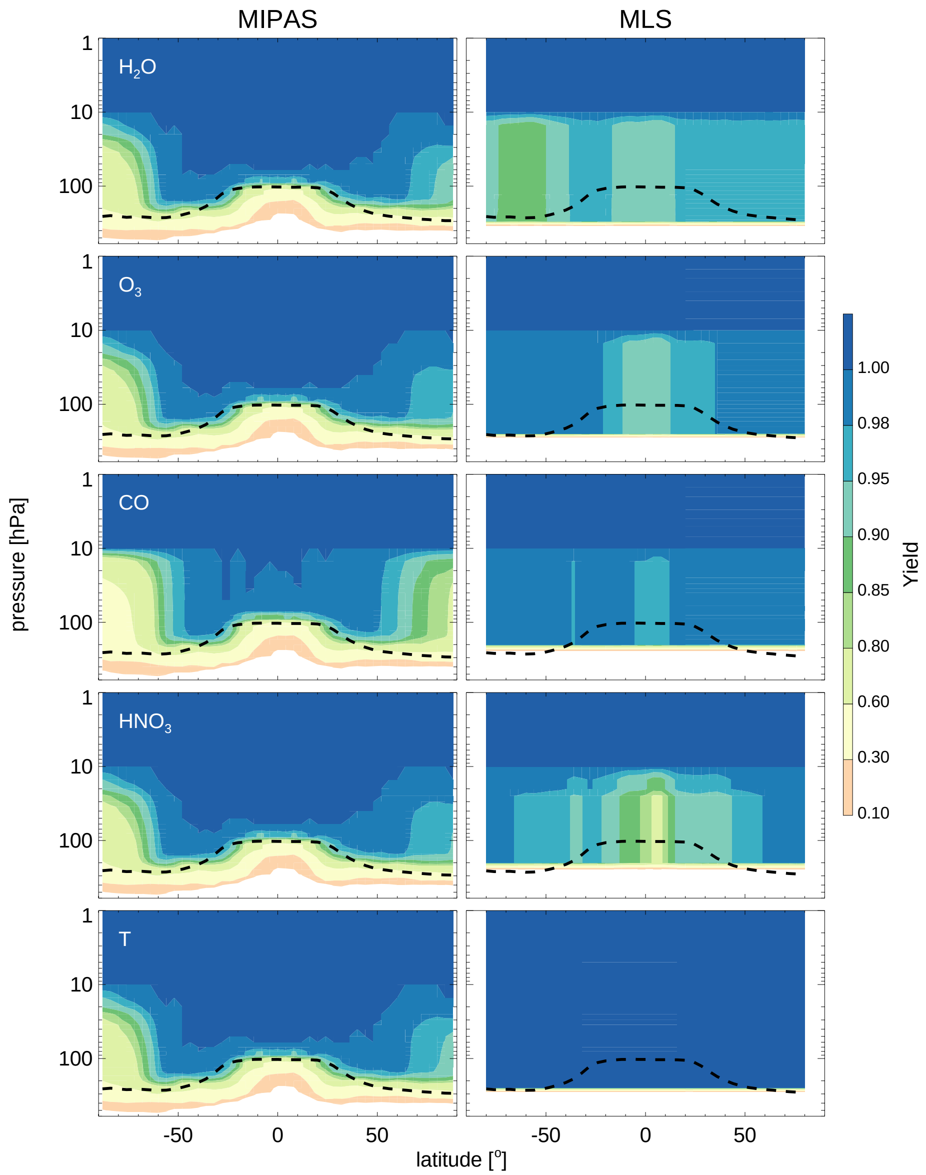
To emphasize more the MLS caveat, the sentence in P1L16 (in the abstract) will be changed to: In contrast, MLS data quality screening removes sufficiently few points that no additional bias is introduced, although its penetration is limited to the upper troposphere while MIPAS may cover well into the mid troposphere in cloud-free scenarios.

In P5 L8 this sentence will be changed: In contrast, in general MLS yield values are better than 90%, although the measurements do not extend below the upper troposphere.

In a similar manner P7 L18 will be changed to: However, continuum absorption in the microwave suppresses signals from the mid and lower troposphere in a limb viewing geometry, limiting the MLS vertical range to the upper troposphere and above while MIPAS may cover well into the mid troposphere in cloud free scenes.

And in the new paragraph about trends we included: However, when only those measurements passing the screening criteria are used, both instruments have limitations: MIPAS trends are impacted because of the large percentage of measurements screened out below 100 hPa, which introduces non-negligible artifacts (up to 80% change for H_2O and up to 20% change for O_3); MLS trends are impacted because of the reduced vertical resolution, which limits its usefulness to the upper troposphere and above.

Furthermore, we noticed that the Figures were not displaying the correct MLS pressure cut off. The revised versions showcase much better the lack of MLS penetration (as an example, the updated Figure 2 is shown below). Also, we superimpose a mean thermal tropopause derived from MERRA2 to Figure 2, 3, and 4.



Updated Figure 2

Reviewer comments

This paper discusses the impact of non-uniform sampling of the MIPAS and MLS instruments on resulting zonally averaged data with particular emphasis on how data quality screening exacerbates biases. The authors describe how MIPAS and MLS have similar sampling biases but how MIPAS data is detrimentally affected by screening through the use of running MIPAS and MLS sampling through the CMAM30-SD CCM. While the methodology applied has some use and accurate in its basic conclusions, some of the more meaningful conclusions (e.g., impacts on ability to derive trends from different instruments in the UTLS) are not supported by analysis and thus speculative. Furthermore, the scope of this work is severely limited and the tone of the paper is seemingly self-serving. As such, I would recommend this work for publication only after additional work and major revision.

Major Comments:

Pg. 02, Line 29: "We emphasize that the results of this study refer only to the representativeness of the respective data, not to their intrinsic quality." This statement greatly detracts from the value of this work. This work makes very clear remarks regarding the impact of sampling biases on the ability to use certain kinds of data sets for trend analysis in the UTLS and how sampling biases will require longer data sets because of the added noise. However, the quality of the data sets that are used is a critical component to those kinds of analyses. Recent work in the 2014 Ozone Assessment or the SI2N effort (https://www.atmos-chem-phys.net/ special_issue284.html) have shown that sampling biases, while present in trend analyses, are not necessarily the greatest driver of trend uncertainty as data quality issues and instrument drifts are also present. Incorporating data quality into the calculation is necessary, as it is still possible for higher precision data with sampling biases to be more robust for trending work than lower precision data, though it ultimately depends on what those precisions and sampling impacts are.

This will be expanded to: We emphasize that the results of this study refer only to the representativeness of the respective data, not to their intrinsic quality. Their quality has been extensively evaluated in numerous data characterization and validation papers [i.e., Pumphrey et al., 2007; Read et al., 2007; Santee et al., 2007; Schwartz et al., 2008; Stiller et al, 2012; Hegglin et al., 2013; Raspollini et al, 2013; Neu et al 2014; Livesey et al., 2017; Sheese et al., 2017]. Furthermore, their long-term stability has also been studied [i.e., Nair et al., 2012; Eckert et al., 2014; Hubert et al., 2016; Hurst et al., 2016].

In addition, a new paragraph about trends will be added (see above) which will include the following sentence: Note that, when using real data, the effect of instrument noise upon trends will be negligible due to the vast number of MIPAS or MLS measurements associated with each monthly latitude bin. Drifts and long-term stability issues on these datasets [i.e., Eckert et al., 2014; Hubert et al., 2016; Hurst et al., 2016] will have to be corrected.

References:

Pumphrey et al. 2007: 10.1029/2007JD008723
Read et al 2007: 10.1029/2007JD008752
Santee et al 2007: 10.1029/2007JD008721
Schwartz et al 2008: 10.1029/2007JD008783
Nair et al 2012: 10.5194/amt-5-1301-2012
Eckert et al 2014: 10.5194/acp-14-2571-2014
Neu et al 2014: 10.1002/2013JD020822
Stiller et al, 2012: 10.5194/amt-5-289-2012
Hegglin et al 2013: 10.1002/jgrd.50752
Raspollini et al 2013: 10.4401/ag-6338
Hubert et al 2016: 10.5194/amt-9-2497-2016
Hurst et al 2016: 10.5194/amt-9/4447-2016
Sheese et al 2017: 10.1016/j.jqsrt.2016.06.026

Pg. 06, Line 32 and Pg. 07, Line 25: "These poor metrics imply that any trends derived at these pressure levels will also be impacted by quality screening induced biases: the magnitude of the trends will be affected because of the change in the slope, and the number of years of observations required to conclusively detect trends will considerably increase due to the noise associated with the worsening of the coefficients of determination (e.g., Millán et al., 2016)."

Biases and trends can be completely independent and the slopes that are referred to here are not trend slopes but correlation slopes. For example, a seasonally dependent dry bias in H_2O as shown in Fig. 5 does not guarantee an induced bias in the trend values and so the impact of these sampling biases on trends is not directly addressed by this work. Furthermore, while it is true that increased noise in the data can result in requiring longer duration data records to determine significant trends, neither this study nor Millan et al. (2016) considers the attribution between data quality and non-uniform sampling. Without additional work to address the impacts of each of these factors, this entire statement is only speculative.

See discussion on trends above.

In general, the scope of this work is severely limited. The statement that an instrument's inability to see through clouds will result in less data is obvious while comparing two data sets to ascertain the impact on trends without considering data quality is neglectful. As it stands, this study has limited to no scientific value. However, the underlying concept of this work could be expanded albeit with simulated sampling and uncertainties. For example, the authors could simulate different sampling patterns from different orbit types and sampling frequencies for an expected limb sounder. Running these through model data would allow for comparisons of the impact of different potential sampling patterns. To further test potential data quality screening, utilizing model cloud fields or retrieval limitations of different observation techniques would create another variable to test rather than using the data quality screening of a specific instrument. Lastly, the authors could simulate differing data precisions or vertical resolutions from all of these sensitivity tests and incorporate the resulting uncertainties into trending analyses to determine the impact of all of these variables on the derived trends. This would provide a large solution space to test the impacts of potential measurement systems on their ability to be used for trending analyses in the UTLS region. While these suggestions are extensive and perhaps more comprehensive, some aspect(s) would be necessary to incorporate actual data quality into the authors' data quality screening to back up the claims regarding the impact on trend detection.

The main purpose of this study is to quantify the impact of screening biases using MIPAS and MLS as proxies for IR and MW limb sounder measurements, not to run a full Observing System Simulation Experiment (OSSE – the exercise the reviewer is essentially suggesting) to study different sampling patterns. That would make the paper much less focused. Furthermore, the MIPAS and MLS sampling patterns are so dense that the sampling bias is negligible for both (as shown here); that is, *the problem is not in the sampling pattern per se but rather in the inability to measure in cloudy scenes*, which leads to the "obvious" reduction in data. Hence no matter what sampling patterns are explored, the problem will still be there for IR based measurements. Furthermore, when dealing with real data, it will not only be a matter of the reduced number of measurements but also, the reduced representativeness of the average because clouds are correlated with the atmospheric state.

Moreover, an OSSE will require a threshold to determine which clouds affect the IR and the MW data. Such thresholds may be problematic to define and not actually useful in real-world scenarios. By using the screening applied to real data we circumvent this problem by using the criteria determined by the instrument teams to obtain the best data possible from their respective instruments.

Perhaps what is most troubling is the tone of the paper. It seems that the purpose of this paper is to act as a published reference for why a future MLS instrument must be used to ascertain long-term trends of trace gas species in the UTLS. The authors appear to go to great lengths to phrase the message as to assert the “superiority” of MLS measurements over MIPAS at every turn (i.e., pointing out all potential deficiencies in MIPAS data without mentioning any for MLS), and even go as far as attempting to undermine the usefulness of other current or future measurement systems as evidenced by the last two sentences in the paper. While this type of rhetoric is expected in a proposal, it is not suited for a scientific publication

We recognize the reviewer's concern, and hope that our changes detailed above (the inclusion of a discussion on the lack of MLS penetration), combined with our decision to delete the final sentence of the conclusion, alleviate them.

In addition, we will add in P4L5 (in the MIPAS IMK introduction): MIPAS IMK/IAA algorithm retrieves temperature and more than 30 species including O_3, H_2O, CO, CFCs, PAN, among many others.

Minor Comments:

Pg. 01, Line 23: “Further, satellite missions such as . . . have records that span more than a decade.” This statement is phrased in such a way as to suggest that no ground station data have records that long. I would suggest revising. The sentence will be changed to: Further, like many ground-based data sets, satellite mission such as …

Pg. 02, Line 09: “They concluded that coarse non-uniform sampling leads to nonnegligible biases . . .” Biases in what? Is that data biased from the sampling or are the analysis methods not conducive to using data with non-uniform sampling? The sentence will be changed to: They concluded that coarse non-uniform sampling leads to non-negligible zonal mean biases.

Pg. 02, Line 13: “They found that coarse non-uniform sampling patterns can induce significant errors in the magnitudes of inferred trends . . .” Again, is this a flaw in the data or the analysis method? This will be changed to: They found that coarse non-uniform sampling patterns can induce significant errors in the magnitudes of trends inferred directly from monthly zonal means, …

Pg. 04, Line 12: “MLS measures around 3500 vertical scans daily, providing nearglobal (82S to 82N) observations.” Please also include the geospatial sampling characteristics of the MIPAS instrument in its description. It is in P3 L22 of the original manuscript “MIPAS measured around 1350 vertical scans daily, providing global observations.” Thus, no changes have been made in the revised manuscript to address this comment.

Pg. 04, Line 20: "Further, we used the vertical grid of the CMAM30-SD fields; that is, we assume that MIPAS and MLS vertical resolution is good enough to resolve these model fields, at least in the upper troposphere / lower stratosphere (UTLS)." What are the vertical resolutions of the model and instruments in the UTLS?
The vertical resolution of the model is in the original manuscript (P3 L15). For MIPAS and MLS, the vertical resolution will vary by specie but overall, they are between 3 and 4 km. Taking this into account, the sentence will be changed to: Further, we used the vertical grid of the CMAM30-SD fields; that is, the impact of the vertical resolution of the measurements is not taken into account. However, note that, in this case, both instruments have similar vertical resolutions in the upper troposphere / lower stratosphere (UTLS), varying overall from 3 to 4 km.

Pg. 06, Line 31 and Pg. 07, Line 24: "The biases for O_3 and HNO_3 oscillate between -10. Unfortunately, we couldn't address this comment as it seems to us that it is incomplete.

Figure 5: It would be better to change the X-axis label interval on the time-series plot to every 5 years.
We changed it to every two years and increased the font size.

Figures 6 and 7: There is a lot of unnecessary white space in some of these plots, though given the desire to maintain consistent axes ranges between the two I can see why Figure 6 has so much white space. That having been said, I still think some reductions can be made to make the results easier to see.
The x-ranges have been reduced as much as they can be without cutting out any of the lines shown.

Additionally, for whatever reason the line thicknesses appear the same between Figs. 6 and 7 when zoomed out but are much thinner in Fig. 6 when zoomed in. They will be the same thickness in the new version.

Lastly, what is the bottommost pressure level on the Y-axis?
The bottommost level and the top pressure level are 400 and 50 hPa, these levels will be added to the figures.