Replies to the Comments:

The authors thank the reviewers for their insightful comments. In the following, the comments are included in black while our replies are given in blue.

General comments:

This is a very good paper and answers (negatively) the important question posed in the title, although given that the question is posed in the title, I do think the answer is surprisingly difficult to find in the text.

My primary concern with this manuscript is that I don't understand exactly how Figure 8, which is an extremely important figure, is produced. In response to a question in the quick review the authors now state: "One caveat is that the time periods for the de-seasonalisation inevitably vary among the satellite data sets and are different from that used for the FPH observations (and model simulations). While this affects the absolute differences, tests show that this has no decisive influence on the overall spread estimate nor the consistency of the temporal development of the differences shown in Fig. 8." This seems to imply that they have set the average difference equal to zero for each dataset. But if that is the case then, given that Figure 1 shows a trend of ~0.28 ppmv/decade between satellite and FPH, I do not see how the authors can make the statement that neglecting the fact that the satellite to FPH comparison changes with respect to time period "has no decisive influence on the ... temporal development of the differences shown". Based on that trend the difference between, e.g., the SAGE vs. FPH differences (average date ~1996) and the MLS vs. FPH differences (average date ~2010) must be ~0.4 ppmv, which is certainly not negligible on Figure 8. On the other hand, my interpretation that the average difference is equal to zero for each dataset is probably wrong (the MIPAS offset appears to be distinctly positive). In any case, the authors should explain how the offsets are calculated and not (absent a much better explanation) say that it doesn't matter.

General response #1:

What Fig. 8 shows can be expressed as follows:

$y_{difference}(t) = running\_average[\, y_{FPH}(t) - y_{other}(t) \,]$

$y_{FPH}(t)$ is the de-seasonalised time series observed with the FPH instrument at Boulder. $y_{other}(t)$ describes the de-seasonalised time series either from the model simulations or the satellite observations. For EMAC, WACCM, CMAM, CLaMS and MLS we considered what we defined as adapted Boulder time series. For the HALOE and MIPAS instruments the full Boulder time series were used. For SAGE-II the time series for the zonal mean between 35°N and 45°N was implemented. The resulting difference time series was smoothed using a running average of one year, requiring at least three valid data points during this period. If this criterion is not fulfilled the average is discarded. The smoothing is used because otherwise it is difficult to really extract any patterns from the differences.

What has been inconsistent so far was the de-seasonalisation period among the different data sets. For the FPH observations and the model simulations the time period from 1985 to 2010 was used, which is fine. For the satellite observations, however, inevitably the de-seasonalisation period had to be shorter and corresponded to the measurement period of the individual instruments, i.e. from 1992 to 2005 for HALOE, from 2002 to 2012 for MIPAS, from 2004 to 2016 for MLS and from 1986 to 2005 for SAGE-II. The difference time series $y_{difference}(t)$ is of course dependent on the de-seasonalisation periods of the data sets involved and differences in these periods are not optimal. That is why we added the caveat. In the revised version we have now eliminated this inconsistency. For the difference time series $y_{difference}(t)$ the FPH observations now always use the same de-seasonalisation period as the data set they are compared to, i.e. 1985 to 2010 for model simulations (as before), 1992 to 2005 for HALOE, 2002 to 2012 for MIPAS, 2004 to 2016 for MLS and 1988 (at 70 hPa to data just start in this year) to 2005 for SAGE-II.

Part of the caveat on the inconsistent de-seasonalisation periods focused also on the overall spread estimate and the consistency of the temporal development of $y_{difference}(t)$ among the different comparisons. As spread we defined the difference between the maximum and minimum of $y_{difference}(t)$ at given time. The overall spread is the average over all times. In terms of the consistency of the temporal development we referred to the dip in $y_{difference}(t)$ around 1993/1994, the subsequent increase until 2000, the relatively constant behaviour from 2001 to 2009 and so on seen in most comparisons. With the inconsistency of the de-seasonalisation periods now removed we can only

reiterate our caveat statements that this only marginally influenced the overall spread estimate and the consistency of the temporal development.

Abstract page 2 lines 2-4 “Overall, both the simulations and observations exhibit trend differences between Boulder and the zonal mean. The differences are dependent on altitude and the time period considered.” I’m not sure what information these lines add (of course there will be some differences) other than to confuse the reader, especially since the next 2 sentences then say that the differences are “not sufficient to explain the discrepancies”.

General response #2:

This is simply a summary. Even though this is trivial and presumably the expected behaviour, we think it is still worth to mention this.

Figure 1 – The error bars for the merged satellite dataset are very hard to see, but, more importantly, on the positive side they all seem to lie exactly on the zero line. Please check to make sure that this is indeed correct, and if it is, please explain why.

General response #3:

This is intentional! As described in the text we do not know the exact significance level of the trend estimates derived from merged satellite data set. What we know is that significance level is at least 2 and we assumed this level here for simplicity. What we absolutely wanted to avoid is any overestimation of the significance level. Thus, this conservative approach.

Page 7 – Here it says explicitly that: “observations before March 1992 were discarded”, yet in several plots data points are shown in 1991. Since what is shown are annual averages this might be mathematically okay, but I would strongly discourage showing anything before the first data included in the time series (at the earliest).

General response #4:

Nothing is actually shown before that date! Presumably, the confusion arises since the time ticks are placed in the middle of the year, but this is actually noted in the figure captions.

Page 11 – “We focus on the altitude range between 100 hPa and 20 hPa that is typically covered by the FPH observations and in almost all cases completely entirely in the stratosphere (Kunz et al., 2013).” Either “completely” or “entirely” will do, but not both.

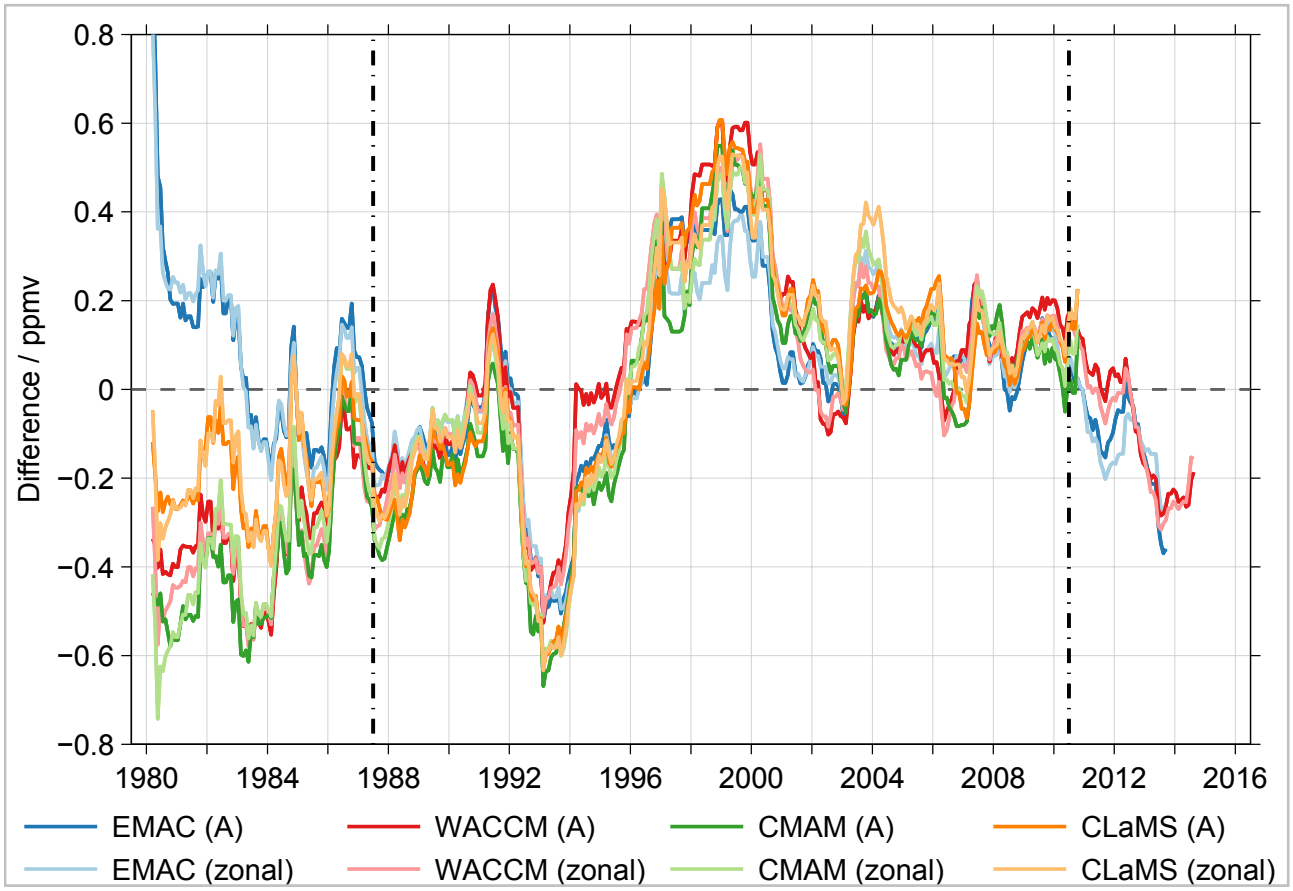
General response #5:

Sorry, this is our mistake. The word “entirely” has been removed.

Figure 8 – It seems to me that it would helpful to the reader, and would seemingly nicely summarize the main point of the paper, if the authors would add to this figure a line showing Boulder minus zonal mean for any one of the models taken from Figure 2 or 6.

General response #6:

We have tested this for the different model simulations as shown in the figure below. As expected the results are very similar for the Boulder and zonal mean time series. Because of this we decided not to include any zonal mean data from any of the simulations.


0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
Difference / ppmv
1980
1984
1988
1992
1996
2000
2004
2008
2012
2016
EMAC (A)
EMAC (zonal)
WACCM (A)
WACCM (zonal)
CMAM (A)
CMAM (zonal)
CLaMS (A)
CLaMS (zonal)