Response to RC2:

Thank you for your thorough review and thoughtful comments on our manuscript. Our responses to each of your comments and questions are listed below in italic font

The authors evaluated AIRS MUSES CO profiles and retrieval error estimates with HIPPO, ATom, and NOAA GML aircraft observations. The following comments need to be addressed.

Figure 1: Please also add lines in addition to the dots so that the flight tracks are clearer.

It is more correct to say that Figure 1 shows the positions of all the aircraft profiles used in this study rather than the actual flight tracks. There are gaps where no usable aircraft and/or coincident satellite data were available. At Line 118 in the original manuscript, we have added the following sentence.

"The locations of all the aircraft profiles used in this study are shown in Fig. 1."

The first sentence of the figure caption for Fig. 1 has also been revised to "Locations of aircraft profiles used for HIPPO and ATom as colored dots and NOAA GML as black diamonds with 3-character string identifier. "

Line 183: Please change "Thus, each aircraft profile was evaluated against a set of AIRS profiles." to "Thus, each aircraft profile was compared to a set of AIRS profiles".

This line has been changed

Line 185: Could you specify how many levels are there in the AIRS MUSES forward model?

There are 67 levels, and that information has been added to the text.

Figures 3 and 7: "and the latitude bands are indicated in the upper left." The latitude bands are missing in the Figures. And the sentence "The red lines indicate the individual profiles, the black solid lines the mean difference or bias, and the dashed lines one standard deviation from the mean." needs to be corrected.

The latitude bands have been added to these figures.

Figure 4: Please add a legend to figure 4b, since you have a legend for black dots in figure 4a, and black dots represent different things in 4a and 4b.

Legends have been added to the bottom panels of Figures 4, 8 and 12 indicating that the blue triangles are the AIRS -AIRCRAFT differences and the black dots are the AVERAGE AIRS - AIRCRAFT differences in the 10-degree bins.

Section 4: I'm a little confused with the comparisons between "AIRS-aircraft standard deviation"

and “a priori–aircraft standard deviation” (Figures 5, 9, 13). Could you explain more on what is “AIRS-aircraft standard deviation” and “a priori–aircraft standard deviation”? If “AIRS-aircraft standard deviation” stands for “standard deviation of the difference between a priori and aircraft”, then this value only represent the variability of the bias of a priori from aircraft instead of magnitude of the bias. Therefore when AIRS-aircraft standard deviation is lower than a priori–aircraft standard deviation, it only means that the bias of AIRS has smaller variability than the bias of a priori. How could this indicate that AIRS perform better than a priori? Am I understanding it right? Do you mean “the square root mean of the difference between a priori and aircraft”?

For these large datasets RMS and standard deviation are equivalent. When the AIRS -aircraft standard deviation is smaller than the a priori – aircraft standard deviation it indicates that the retrieval is doing a better job at capturing the actual variability in the aircraft profiles than the a priori. Furthermore, when the AIRS – aircraft standard deviation is greater than the mean observation error it indicates that the observation error is underestimating the actual retrieval error.

Line 310: “……variability within the set of AIRS profiles collocated with an aircraft profile, which can be thought of as an empirical error.” To me the variability within the set of AIRS profiles collocated with an aircraft profile is representativeness error. I’m wondering if it is the same as what's discussed here and is it comparable to the theoretical error? As also shown by Figures 6 and 10, comparisons to plume obs show higher empirical error values than comparisons to background obs. This is because representativeness error is higher in plume (more heterogeneous) compared to background (less heterogeneous), and is not related to the theoretical errors. The comparisons to variability within the set of AIRS profiles collocated with an aircraft profile do not seem necessary to the main story of the manuscript. However, if the authors do include this part of the comparisons, please provide overall statistics of the empirical errors and theoretical errors in addition to the illustrative cases. And please also discuss what does it mean when the empirical errors and theoretical errors are close or far away.

The terms theoretical error and empirical error were adopted from Oetjen et al (2014) who performed a similar analysis on satellite and ozonesonde data. In their analysis Oetjen et al (2014) assumed that the coincident satellite profiles are basically seeing the same scene and therefore any variability in the retrieved profiles could be considered an empirical measure of the retrieval error. Yes, it is likely that the plume cases would probably feature more actual heterogeneous mixing ratios compared to the background cases and therefore empirical errors are more likely to be much greater from a statistical basis. However, our illustrative cases were selected to show the range of retrieval performance. For example, in Figure 6, the empirical errors for the HIPPO2 plume case are very large and much larger than the theoretical errors, but for the HIPPO3 case they are much smaller and similar to the theoretical errors.

For clarity and consistency, we have changed the names of the terms used for this analysis in the revised manuscript. The theoretical error is now called the mean observation error to be consistent with the terminology used in Figures, 5, 9 and 13. The empirical error has been renamed the AIRS profile variability.

The introductory information for this approach on Lines 310 – 312 of the original manuscript has been changed to the following.

"An alternative approach for evaluating the theoretical error is to compare it to the variability within the set of AIRS profiles collocated with an aircraft profile. If it is assumed that all satellite footprints in the collocated set are basically seeing the same scene then the variability in the retrieved profiles can be considered an empirical error (Oetjen et al., 2014). In this analysis this empirical error is referred to simply as the AIRS profile variability."

Figure 8 are different from Figure 4. In Figure 8b, the average differences are positive in -30S-10N band. And the average difference is negative at 30N, which is opposite to Figure 4b. Please add a brief discussion for this. And I was also wondering if the same a prior profiles were used for ATOM and HIPPO periods? Because the mean a priori error estimate for ATOM (Figure 9) is higher than that for HIPPO (Figure 5), which may partially contribute to the "better" retrieval performance relative to the prior for the ATom vs the HIPPO comparisons.

The HIPPO and ATom campaigns were conducted in different years and it appears that there were differences in the CO mixing ratios in the air masses sampled during the respective campaigns. For HIPPO the aircraft column average CO mixing ratios in the 30S – 10N band were all less than 100 ppbv, whereas for ATom they were much more variable and were as high as ~ 130 ppbv. Between 30N – 40N the HIPPO column mixing ratios ranged from ~70 ppbv to ~ 140 ppbv whereas for ATom they were lower ranging from ~60 ppbv to ~125 ppbv.

The a priori profiles used for the HIPPO and ATom retrievals were obtained from a climatology generated with the MOZART atmospheric chemistry model(Brasseur et al., 1998). A reference to this model has been added to the revised manuscript in Section 3.2. The differences in the a priori errors are due to the differences in the air masses (with different CO mixing ratios) sampled during the different years of the campaigns.

Brasseur, G. P., Hauglustaine, D. A., Walters, S., Rasch, P. J., Muller, J. F., Granier, C., and Tie, X. X.: MOZART, a global chemical transport model for ozone and related chemical tracers 1. Model description, J. Geophys. Res.-Atmos., 103, 28265–28289, 1998.

Line 349: "The distribution of errors in the 30N–60N latitude band is less skewed than for HIPPO (0.54 vs. 1.36) suggesting that a Gaussian distribution of errors is a reasonable assumption for this dataset." But outside the 30N-60N, distribution of errors seems more skewed for this dataset (e.g., -30S-10N).

We looked at the skew only in the latitude bands corresponding to those in Figures 5, 9 and 13 and found that the significant positive skew in the 30N -60N band for HIPPO was a plausible explanation for the failure of the diagnosed observation errors to represent actual retrieval errors in those profiles.

Line 560: "We also examined the relationship between retrieval bias and the CO mixing ratio." Please add a figure for this if possible.

We have added this figure as Fig. 16 and included and added the reference "(Fig. 16)" to the end of the sentence above.