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.

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”.

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

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.

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.

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”?

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.

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.

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).

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