We thank Anonymous Referee #2 for a positive and thoughtful review. In the following, we will respond to the Reviewers comments step by step.

Minor revisions: p5, l8-10: This enables ... denoted XCO2 I think this statement would be more clear to the reader when a line is added to indicate that an extension is needed above 14 km. Then you can indeed state that this extension is of limited consequence since most of the variability in XCO2 stems from the troposphere which is covered by the HIPPO profiles.

➔ Good point. We changed to: "This enables a comparison of individual sub-columns of air but also of column-averaged mixing ratios of CO$_2$, denoted XCO$_2$, if the profile can be reliably extended above 14\,km. As the troposphere dominates the variability in XCO$_2$, errors induced by extending profiles are expected to be small."

p7, l11-12: As most ... analysis. Add a line why SCIAMACHY does not provide data over oceans

➔ Added "…because it lacks a dedicated Glint measurement mode" and explained it better later as well, as requested by Rev. #1.

p7, l16: short-wave –> short-wave infrared

➔ done

p8, l11: How can averaging lead to the reduction of systematic errors?

➔ Removed systematic here.

p9, l9-11: Validation ... (Olsen and Licata, 2014). If Olsen and Licata already have compared IR/MW L2 and IR-Only L2 against HIPPO, then I would expect a sentence explaining how the current study differs and/or extends wrt. the cited paper.

➔ We rephrased and extended that sentence to reflect the main differences (using models to fill up the profile).: Olsen and Licata (2014) compare the IR/MW based and IR-Only based CO2 retrievals over the globe for 2010-2011 and for collocations with the deep-dip HIPPO-2, HIPPO-3, HIPPO-4 and HIPPO-5 profiles. Their global analysis reveals that the zonal monthly average difference rarely exceeds 0.5 ppm save at the high northern latitudes in January and October where fluctuations resulting from small number statistics dominate. Their analysis against HIPPO employs only the deep-dip measured profiles, i.e. those in which the aircraft reached the 190 hPa pressure level, to ensure good in situ measurement coverage of the AIRS sensitivity profile and to minimize the error introduced by their simple approximation of extending the aircraft profile into the stratosphere by replicating the highest altitude measurement. During the HIPPO-2 and HIPPO-3 campaigns, the AMSU channel 5 noise figure was acceptable, whereas during the HIPPO-4 and HIPPO-5 campaigns it progressively degraded at a rapid rate. For all campaigns, the two sets of collocations, averaging AIRS retrievals within ±24 hours and 500 km of the aircraft profile, exhibit the same bias and RMS to within 1 ppm for $|\text{lat}| \leq 60°$. The current study extends the in situ measurements to higher altitude by the means of CarbonTracker and MACC model output, thereby allowing use of all HIPPO profiles rather than only the deep-dip profiles. Our results are statistically consistent with the

latitude-dependent biases reported by Olsen and Licata (2014) and give a more detailed view of the scatter as a function of latitude.

p9, l14-18: For the differences ... should dominate If you first extend the HIPPO profile with model data, then integrate, and finally subtract the integrated model data, does the part above flight altitude not exactly cancel? HIPPO (< 14 km) + model (> 14 km) - model (0-TOA) = HIPPO (< 14 km) + model (> 14 km) - model (< 14 km) - model (> 14 km) = C2 HIPPO (< 14 km) - model (< 14 km). So, I do not see how the extension can contribute to the difference between HIPPO and model.

- ➔ About 20% of the total column is located above 14km and not all HIPPO profiles extended that far. If we use part of the model, these values indeed cancel and yield exactly 0 difference, making the agreement somewhat better. With the 80/20 weighting, it is similar to saying that delta-XCO2 is 0.8*(Model-HIPPO)+0.2*(model-model=0), thus potentially always dampening the differences. Or, if the profile extended only to 10km, dampening it even further.

p10, l14-23: Figure 4 ... potential biases. HIPPO 3 is nicely explained in this paragraph, but HIPPO 5 is depicted in the Figure but not mentioned. Any comment that the authors can make on the MACC and CT differences/similarities?

- ➔ Added “In HIPPO 5, at the end of the growing season, the situation is reversed as the profile slopes change sign after the large CO$_2$ uptake during summer.” And “For HIPPO 5, the deviations for CT2013B are somewhat smaller but it can be seen that most models suffer from these potential biases if large vertical gradients exist.”

p11, l2-10: Here, we look ... in the future. This alinea is mostly about measurements and campaigns that are not treated in the paper. I understand why the authors like to mention this, but maybe the conclusion, which includes a future outlook mentioning OCO-2, is the better spot for this.

- ➔ This is indeed better, we moved this to the Conclusions.

p11, l11-19: For the comparison ... were the truth). This my strongest comment on the paper: Since the requirements on XCO2 are so stringent, it matters for the comparisons in this paper how exactly 1) the HIPPO profiles are extended, 2) the averaging kernel is applied, and 3) the null-space is attributed. I would recommend to incorporate a small section/paragraph explaining the mathematical details. Questions that come to mind: Is the model information just attached to the HIPPO pro- file? If a jump would appear in such a profile, how is that treated? Is the smoothed (extended) HIPPO profile compared to the GOSAT profile without null-space contribution, or is there also a null-space contribution to the smoothed HIPPO profile? If the latter, which reference is used? The same as in the GOSAT retrievals, or the model?

➔ This is a good point even though we prefer to keep this short in the paper. Re 1). The HIPPO profiles are extended with the model data before applying the averaging kernel correction. 2). The AK corrected HIPPO values are computed as xa+A(xt-xa) with the a priori profile xa and the "true" profile xt (HIPPO + model). For GOSAT, the column averaging kernel was used, for TES and AIRS the averaging kernel for the respective retrieval layer.
We have not tested the impact of a jump in a profile; in the manuscript, a simple profile extension was performed without testing smoothness. In most cases, the impact should be relatively small. The null space contribution in GOSAT comparisons should be small as the column averaging kernels are relatively large throughout the entire column. In general, HIPPO data has always been filled in with model data, not satellite priors.
We added
For GOSAT: "For the HIPPO comparison against GOSAT data, we take the instrument sensitivity into account by applying the averaging kernel to the difference of the true profile (using the model-extended HIPPO dataset as truth) and the respective a priori profile. We perform this correction using both model extensions independently and then use the average of the two. "
For TES: "For the comparison with TES, we use the 510\,hPa retrieval layer and apply averaging kernel corrections using model-extended HIPPO data as {\em truth}, using both models indepdently and averaging results after averaging kernel correction."
For AIRS: "For the comparison with AIRS (Fig\,. \ref{fig:HIPPO_AIRS}), the sensitivity maximum varies around 300\,hP and we apply the averaging kernels similarly to TES."
We hope this will clarify the issue.

p11, l22-24: Even after ... for MACC. Please refer to Figs 5 and 6

➔ done

p11, l22: Even after normalization It is clear how the HIPPO data is corrected, but how is the other data corrected? With the HIPPO value, or with the average value of the particular model?

➔ With the HIPPO value. We added a sentence "For each campaign, we also normalize all data with the respective campaign average of the HIPPO dataset."

p13, l23: lower left quadrant Maybe the authors would like to note that these points are also outliers in the CT comparison. Not as strong as in the case of MACC, but still in C3 the same quadrant, which may be an indication that the transport errors in both models are roughly equal and/or the GFED data is somewhat off.

➔ We mentioned that "both models" show that feature.

p24, Fig 3: There are some strong excursions in the HIPPO profiles close to the surface; any explanation for these?

➔ These might be caused by dips close to the surface with HIPPO, potentially coming from the land data. It should not really affect XCO2 a lot as it only affects a small subcolumn.

HIPPO-1, 3, and 4 (and possibly 5), the differences between HIPPO and MACC resp. CT differs significantly for > 70N. Any explanation for this behaviour?

➔ We agree, there seem to be substantial differences but we don't have any explanation yet for this and would not like to speculate too much.

Please, reposition the legend box; CT-HIPPO 5 is barely visible.

➔ done

p26-p28, Fig 5-7: Mention the shift for both axes

➔ we now state "Scatterplot of normalized (with campaign average) CO2…"

p31,p32, Fig 10,11: Mention the shift for both axes

➔ see above

p4, l5: Greenhouse Gas Observing –> Greenhouse Gases Observing

➔ don

p4, l5: haven –> have

➔ fixed, thanks.

p4, l11: sensing measurement –> sensing measurements

➔ done

p5, l5-8: This sentence does not have a verb. Suggestion: The HIAPER Polo-to-Pole Observations (HIPPO) project consists of a sequence of ...

➔ replaced "sampling" with "sampled"

p6, l23: LSCE. To be on the safe side I would explicitly write out this acronym

➔ Done

p9, l21-22: consistent between model, –> consistent between the two models,

➔ done

p10, l6: usually 162 253 –> usually

➔ done

There are several places where ppb is used in stead of ppm: p11, l24 p11, l25 p21, Table 1 (2 instances)

➔ done

p16, l11: that –> than

➔ done, thanks