Reply to Anonymous Referee #1:

First of all we'd like to thank the referee for his/her interest in our paper and the helpful comments, questions and suggestions. All comments will be carefully considered for the revised version of the manuscript.

Below we give point by point answers to each of the referee's comments.

General Comment:

Overall this paper is well written and is gives a detailed, statistics based analysis of the performance of multiple space based XCO2 and XCH4 retrieval algorithms. The main fault of the paper is the use of a simplistic distance time co location technique to match ground-based TCCON validation data to the satellite observations. This somewhat obscures the results. It has been shown (e.g. Guerlet et al., 2013) that methods that accurately take into account atmospheric dynamics (i.e., the advection of relatively high or low concentration air by prevailing winds) are both more robust and yield significantly better sample sizes. Doing so yields substantial improvements in statistics such as interstation bias. In fact, Guerlet et al. (2013) show that interstation bias is almost of no use when evaluated with the simple geometric colocation method, primarily due to the low sample sizes involved. In general I would prefer the authors redo their analysis with a more robust colocation technique, though I recognize their statement within the article that using a more robust colocation criteria "was not feasible due to practical considerations." I would think that if they truly want robust results, they would use the best methods available. Given that they have already selected the algorithms to proceed forward in their "round-robin" competition, I understand that redoing this analysis is unlikely, but I believe the authors should stress upfront the problems associated with the geometric colocation scheme and that there are other, better schemes available in general.

In defense of the methodology used, we do need to stress that indeed, a dynamic collocation method yields more data samples and thus more robust statistical parameters and that, had we included all our findings for all stations in our further analysis, parameters such as the (seasonal) relative accuracy could be somewhat crippled. However, we do not simply take all station results into account, instead we filter out any result which lacks in quality (either not enough data samples or with an elevated standard error). We also filter out the exact same data from the direct competitors' sample, regardless of whether it meets the quality criteria, so that under all circumstances the (seasonal) relative accuracy of competitors is obtained from an identical set of stations/seasons. This greatly increases the robustness of our analysis, nor do we shun from stating the confidence interval of our results. We thus do have confidence that the relative quality differences between competing algorithms are accurately portrayed using this method.

That said, we will expand the section which deals with the discussion of alternative collocation methods.

A second but more minor problem is that the authors make no mention of ocean versus land retrievals for any of the sensors & retrievals, or gain M vs. H retrievals (for GOSAT). Somewhere near the beginning of the paper, the authors need to make clear what type(s) of data for each satellite are included in the comparisons. If ocean data are included, error statistics may need to be evaluated separately as compared to those over land. It is well known that ocean vs. land retrievals each carry their own biases and error characteristics, partially because of the different cloud and aerosol scattering effects over ocean vs. land as well as the vastly different surface BRDFs (see e.g. Butz et al., 2013).

This is indeed an oversight. All analyzed data is over-land only, we will change the text accordingly.

Specific & Technical comments:

P8686,L17: This sentence is redundant as SRFC and OCFP have already been defined. Suggest removing.

This line introduces the SRFC and OCFC acronyms to differentiate them from their non-bias corrected SRFP and OCFP counterparts. We'll clarify the sentence

Section 3.3. Both SCIA XCH4 retrievals are proxy non-scattering algorithms, which implies they are nearly the same. It would be useful to state in a sentence or two how these two algorithms differ.

Apart from calibration, pre- and post- filtering differences , WFMD uses a method in which a linearized radiative transfer model (chosen from a look-up table) plus a low order polynomial is linear least squares fitted to the logarithm of the measured sun-normalized radiance. IMAP on the other hand uses an optimal estimation inversion method, which minimizes both the least squares difference between forward model and measurement as well as between the a priori and a posteriori state vector.

P8687,L5: Chapter 3.2 > Section 3.2

This will be corrected

P8691,L12: Regarding relative accuracy, this metric has been shown to be problematic to construct from TCCON alone, especially using a purely geometric colocation technique (Butz et al, 2013). It is even more suspect given the new information on station to station TCCON biases related to ghosting issues (also known as "laser sampling error") in the TCCON instruments (Dohe et al., 2013). It is important to mention these caveats surrounding RA.

We will certainly add a discussion on TCCON's laser sampling error. Currently the inter-station bias in the TCCON network is estimated to be 0.4 ppm for XCO_2 and 3.5 ppb for XCH_4. These estimates were drawn prior to the discovery of the laser-sampling error and thus were derived from the TCCON dataset used. It is yet unclear how the laser sampling error will effect (after correction) these numbers. The RA is a marker for the combined algorithm and TCCON interstation bias and

any interpretation on the algorithms' accuracy, needs to take into account the TCCON network uncertainty. However as we are interested in the relative quality of competitors, any impact of the laser sampling error will equally affect all competing algorithms.

P8691,L24: datapairs > data pairs

This will be corrected

Figs 4,7,10,13: Y-axis labels should be accurate. Bias (ppm), Scatter (ppm), Correlation, N. At the very least change the y-axis label for correlation, with is simply wrong and definitely does not have untis of ppm.

This will be corrected

P8693, top: Please state what we are trying to learn from the Quantile-Quantile plots. It is not clear that figure 3 is necessary, if it is just about the normality of the datasets. Suggest rewording this section for clarity, or removing the Q-Q discussion and/or plot altogether, unless they are deemed to be of high importance by the authors.

It is indeed just about the normality of the datasets. Section will be reworded

Section 6.1: Please state why the BESD data density is so much lower than WFMD. Is it simply due to algorithm speed, or is it due to necessary pre or post filtering that removes bad or questionable data? This is important, because the former deficiency could be solved with more computing power, while the latter is fundamental. The latter also suggests that WFMD could be improved by employer the BESD filters. Could you state if the WFMD goodness parameters are similar or worse when comparing match soundings? That would be extremely useful to know.

BESD is indeed computationally much more expensive than WFMD but this is not the cause of its lower data density. The reason is that BESD uses entirely different filters. Maybe most importantly BESD uses a very strict MERIS cloud mask but all other filters also differ. During the development of BESD these filters are continuously updated so as to increase the data density (without incurring quality loss) but the SCIAMACHY 30x60km pixel size does impose a fundamental limit to this pre-processing optimization as a significant portion of pixels will inevitably be contaminated with clouds.

Matching soundings is an interesting avenue but leads to rather sparse data samples. One would assume that data which passes a 'strict' filter will by default also pass the filtering process of an algorithm with a less restrictive filtering process. However, often this is not the case. Data filtering is an intricate part of the algorithm and cannot readily be transposed onto another algorithm. A preliminary validation experiment using more recent developments of the algorithm indicate that,

taking the overlapping data only, results in a 33% loss in BESD and 71% loss in WFMD collocations with TCCON. The single measurement precisions and biases were not altered using the overlapping data only, although it needs to be noted that (especially for the WFMD algorithm which features a lower single measurement precision) a definitive analysis requires more data points to reach robust conclusions.

Section 6.2: it is interesting that SRFP performs distinctly less well over the southern hemisphere TCCON sites than OCFP. This also happens in the methane retrievals. Is there any speculation as to why this happens?

Do you allude to the slight increase in the SRFP scatter for Wollongong and Lauder? Certainly for XCH4, this does not result in a distinctively inferior performance compared to OCFP. Even so, if you indeed point to the scatter, this was due to the SRFP filtering process. Analysis with future versions of SRFP no longer show this increase over both stations.

Figs 7,13: Please change GOSA to GOSAT for all the figure titles.

This will be corrected

P8695, L27: till > through

This will be corrected

P8696, L20-26: This could also be affected by the TCCON ghosting issues (laser sampling errors).

This is very unlikely. The CO_2 TCCON recommended corrections due to the laser sampling error are currently 0 ppm for Darwin, -1 ppm for Wollongong and 0 ppm for Lauder. There is currently no listing of correction factors for other species, but they are likely to be proportional. So for the Southern Hemisphere stations we are looking at a possible correction of approximately -4.4 ppb for Wollongong. The observed biases (26 ppb for IMAP, 13 for WFMD) exceed the proposed corrections. Secondly the GOSAT XCH_4 algorithms do not exhibit this behavior. We will address this issue in the text

P8698, L25: “This is of course a direct result of the sample data size.” I would suggest this is only partially result of the (small) sample sizes, but also due to potential station dependent TCCON biases such as discussed in Dohe et al. (2013). I suggest you mention that here again as it is not well known and of high importance in interpreting your results.

This line is there to differentiate between the results in Table 13, which are calculated from all the individual data points (thousands) and the parameters in Table 14 which are derived from (at best) 10 station bias or 40 seasonal station bias results. Both could be affected by the TCCON station dependent biases (which is acknowledged regardless of the issue raised by Dohe et al.), but the difference in the magnitude of their reported errors is due to their sample data size. We'll clarify the statement.