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 colocation 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 re-do 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 re-doing 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.

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).

Specific & Technical comments:

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

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.

P8687,L5: Chapter 3.2 -> Section 3.2

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.

P8691,L24: datapairs -> data pairs

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.

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.

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.

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?

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

P8695, L27: till -> through

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

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.

References:

Butz, A., Guerlet, S., Hasekamp, O. P., Kuze, A., and Suto, H.: Using ocean-glint scattered sunlight as a diagnostic tool for satellite remote sensing of greenhouse gases, *Atmos. Meas. Tech. Discuss*., 6, 4371-4400, doi:10.5194/amtd-6-4371-2013, 2013.

Dohe, S., Sherlock, V., Hase, F., Gisi, M., Robinson, J., Sepúlveda, E., Schneider, M., and Blumenstock, T.: A method to correct sampling ghosts in historic near-infrared Fourier transform spectrometer (FTS) measurements, *Atmos. Meas. Tech.*, 6, 1981-1992, doi:10.5194/amt-6-1981-2013, 2013.

Guerlet, S., Butz, A., Schepers, D., Basu, S., et al : Impact of aerosol and thin cirrus on retrieving and validating XCO2 from GOSAT shortwave infrared measurements. *Journal of Geophysical Research: Atmospheres*, 118 (10), 4887-4905, doi: 10.1002/jgrd.503322013, 2013.