Referee Response to Anonymous Reviewer #1.

We thank the reviewer for their thoughtful comments. The reviewer's comments are in **red bold text**, and our responses follow in black text.

MAJOR COMMENTS

(1) Switching OCO-2 data streams between Section 3 and Section 4: In Section 3, the authors used the full OCO-2 v7Br with a set of manual filters (loosely consistent with WL <=15). In Section 4, the authors use the 'bias-corrected' OCO-2 data from the lite files with WL <=11. Why? This was a huge disappointment - as a science user of the OCO-2 data, based on the analyses presented here, I cannot evaluate the relative quality of the nadir v. glint v. target data. What complicates matter further is that to determine the constant scaling parameter for bias correcting the nadir and glint data, the target data are used. This should be clearly stated in Section 4, i.e., point out the link with the discussion on Page 7. The science community recognizes that we do not have an infinite and unlimited number of validation data to work with. But the authors should keep the OCO-2 data set and the data filtering criteria same for both Section 3 and Section 4. Section 4 should have two parts – one with the dataset same as Section 3 and one with the bias-corrected, lower WL (i.e., higher quality) data that is currently discussed in Pages 9-10.

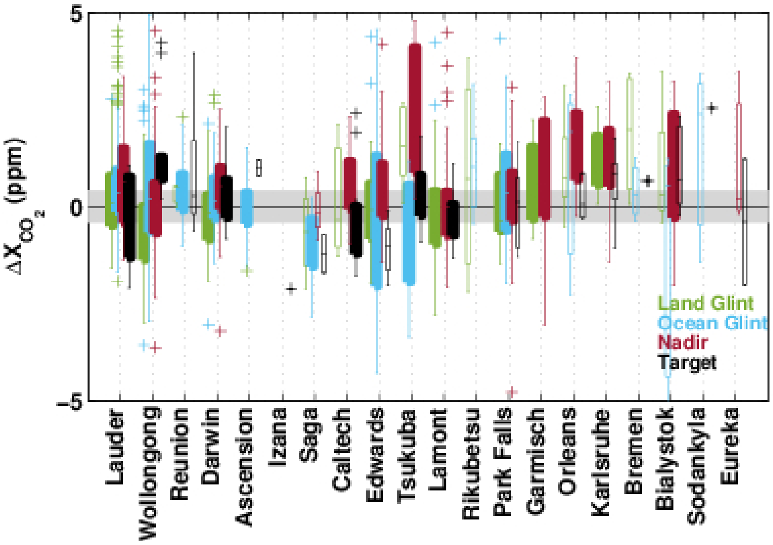
Apologies for not being clear about the filtering choices made in the paper. In order to compute the scaling between OCO-2 and TCCON for the B7 version of the OCO-2 data product (which is then used to create the lite files), we use the target-mode data from the full OCO-2 v7Br L2 dataset, which, at the time the scaling factor is computed, does not have a well-defined set of warn levels. Warn levels require a significant quantity of data with which to robustly run the Mandrake et al. (2013) algorithm, and we did not have sufficient data for this when defining the overall scaling for the B7 lite files. Thus, we used the manual filters described in the paper that successfully filtered the target-mode data, and are roughly consistent with what is now defined by target WL<=15.

Furthermore, as described in Mandrake et al. (2015), data filtered by warn levels in one measurement mode are not equal in quality to data in the same warn levels of another mode. For example, on P15 of Mandrake et al. (2015), it's stated that "Land Target has very small quantities of data available, which prevents a successful definition of WL 0-8. Thus, WL's from 0-8 should be grouped together and treated as an undifferentiated pool of equal quality data." We could, for example, have chosen to use a higher warn level filter for ocean glint data relative to nadir and land glint based on P21 of Mandrake et al. (2015), which states that WL>15 are likely contaminated for land glint and land nadir whereas for ocean glint data, the same is true for WL>=18. However, later in that document (on P29) the authors state that "Above WL12 errors well in excess of the stated a posteriori errors should be expected." Therefore, we chose to select WL<=11 in this paper for the land glint, ocean glint, and land nadir data.

In short, given the relative paucity of target-mode data and the characteristics of warn levels, it is not possible to determine a filter that would be exactly equivalent across all modes. We have done our best to present comparable data throughout the paper. We've also rearranged the manuscript to better describe that target mode is used to generate the scaling factor.

A discussion of our filtering choices has been added to the text. Figure 8 has been replaced by a figure with all modes included so that a comparison of the four observing modes can be made more readily:

Filled boxes indicate that there were at least 10 coincidences (or target-mode maneuvers) over the TCCON stations listed on the x-axis.



(2) Page 9, Lines 277-278: This to me is the most important summary line in the entire manuscript, and the authors need to justify it. The OCO-2 data used in this analysis is already bias-corrected. First this line should read – "differences between bias-corrected OCO-2 and TCCON are all less than : : :..". One may refer to this as the residual biases but the way this line is phrased is misleading. Second, somewhere in the text the authors also need to explain how they came up with the 0.5 ppm number – it is based off the last row in Table 3, which itself takes into account all the TCCON sites, i.e., in an average sense. But the range of differences is large across the TCCON sites. Can the authors provide an uncertainty bound on this "average" bias number, for e.g., 0.5 ppm ??? Once the authors address #1 above, then this number will be revised – I expect that with WL <=15, both N and the Bias will increase in Table 3. But this will be extremely valuable information for the science community – with varying WL cutoffs, how does the OCO-2 data compare to TCCON?

The sentence that was on Lines 277-278 has been rephrased to be clearer and more precise:

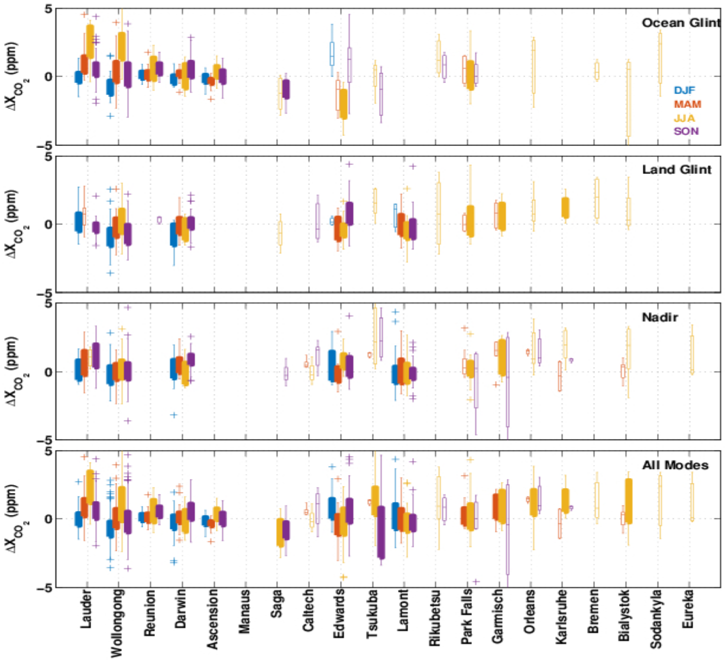
"The differences between aggregated, bias-corrected OCO-2 XCO2 data coincident with all available TCCON daily median measurements are -0.3 ppm, 0.2 ppm, 0.2 ppm for land glint, ocean glint, and nadir, respectively. The RMS values of these differences are 1.3 ppm, 1.4 ppm, and 1.3 ppm, respectively. The differences between the bias-corrected OCO-2 values and the TCCON medians differ from site to site; sites with more than 10 coincident measurements have differences in land glint mode ranging from -0.7 ppm (Wollongong) to 0.9 ppm (Karlsruhe); in ocean glint mode ranging from -1.1 ppm (Saga) to 0.4 ppm (Park Falls); in nadir mode ranging from -0.1 ppm (Lamont) to 1.6 ppm (Garmisch)."

A new table has been added mirroring table 3, but with warn levels <=15 to facilitate comparisons.

(3) Estimate OCO-2 errors/biases under varying surface properties: The discussion of TCCON sites in Sections 2.1 and 3.2 seems to indicate that different sites can be grouped together into specific "surface classes" (i.e., albedo). Have the authors attempted to generate statistics of OCO-2's performance based on the albedo around a TCCON site? For example, a figure similar to Figure 8 but broken up by albedo/surface topography and for different seasons may be highly informative. Such an assessment will allow the inverse modeling community to adjust the errors they specify on the OCO-2 data by season,

location, etc. Again note similar to #2 – the authors should make an attempt to address potential science questions that are of interest and relevance to the community. The manuscript in its current form does not do that.

We've added a discussion and the following figure showing the differences by season for the science mode data:



In the above figure, the filled boxes indicate seasons for which there are >10 comparison points between OCO-2 and TCCON; the open boxes contain at least 3 comparison points. Any site and season for which there were fewer than three comparison points were excluded from the plot. The different colours indicate the different seasons (blue = DJF, orange = MAM, yellow = JJA, purple = SON). The TCCON stations are ordered by latitude, where Lauder is 45S and Eureka is 80N. The equator is between Manaus (3S) and Saga (33N). The high southern latitude ocean glint bias is clear in the top plot. Otherwise, the main result from this analysis is that the OCO-2 XCO2 appear to have a bias that increases with increasing latitude, most clearly seen in the JJA data north of 45N (Park Falls). A seasonal bias is not clearly apparent, but the data volume is poor, especially in the northern hemisphere north of ~45N. These bias patterns are consistent with those seen in the target mode data (Fig 8 in the paper).

(4) Page 5, Lines 117-119: It is unclear what the authors mean that data from other target sites will help assess bias. Isn't bias already being addressed in this manuscript? Are data from any of these other target sites available? How do they compare to the OCO-2 nadir and glint mode data?

This sentence has been changed to read:
"There are several target locations that are not TCCON stations (Fig. 3, orange stars), and although data from those targets will not be analysed in this paper, the data will help assess the radiometric calibration of the instrument, its ability to measure large sources of CO2, validate its solar-induced fluorescence observations [Frankenberg2014], and its ability to measure vertically-resolved information about CO2."

(5) Given that O'Dell et al. [2016] is not yet available (on AMT, or elsewhere to the best of my knowledge),

the authors should provide a bit more description on the bias correction procedure or refer the reader to the OCO-2 technical documents. Mandrake et al. [2015] doesn't cover the recent version of bias correction algorithm that is in place. Again the authors need to be aware that this manuscript will be read by the bigger Earth Sciences community in general, and not just the core OCO-2 community.

The Mandrake et al. [2015] document does cover the recent version of the bias correction algorithm currently in place for the B7 version of the OCO-2 data. All references to O'Dell et al. [2016] have been removed from the manuscript.

MINOR COMMENTS

(1) Page 2, Lines 5-7: Specify that these numbers are valid for a selected subset of bias-corrected and screened OCO-2 data, i.e., even after bias correction, a WL filter of 11 was applied. Or report here the statistics for WL <=15 (or QF =0).

This now reads: "The OCO-2 XCO2 retrievals, after filtering and bias correction, agree well when aggregated around and coincident with TCCON data in nadir, glint, and target observation modes, with median differences less than 0.4 ppm and RMS differences less than 1.5 ppm."

(2) Page 2, Lines 25-26: The utility of XCO2 lies in the fact that we use it to infer surface fluxes of CO2 (a minor nitpicky point). Maybe just rephrase this sentence.

Changed to read: "... which is a useful product for carbon cycle science, as it is used to directly infer surface fluxes of CO2".

(3) Page 3, Line 39: Add the word data after TCCON

Done

(4) Page 3, Line 40: Replace the word "measurements" with observation modes

Done

(5) Page 3, Line 48: This is a rather loose statement. Many factors contribute to the CO2 seasonal cycle, one of which is the boreal forest. Kindly rephrase this statement or end it at the ": : : northern hemisphere".

Changed "the" driver to "a" driver.

(6) Page 3, Line 57: Replace "be measured" with "measure"

Done.

(7) Page 3, Lines 61: Replace "was" with present tense – kindly check the verb forms throughout the manuscript to make it more appealing to the reader.

Done

(8) Page 4, Line 74: Replace the word "measurement" with "region"

Done

(9) Page 4, Line 76-78: Unclear. Why can’t the variability be real? Especially if a weather front is passing through carrying dirty anthropogenic plumes. The authors need to specify caveats associated with this statement, and conditions under which variability in XCO2 can be considered an artifact.

Changed to: “As long as the target location is far from large emissions sources, XCO2 can be assumed constant spatially and temporally within a target region, because atmospheric XCO2 is very unlikely to change significantly over small geographic regions within 4.5 minutes.”

(10) Page 4, Line 104-106: Please be more descriptive of the exact surface properties or albedo conditions at these sites. Line 104-105 currently reads like a nursery rhyme.

Removed the offending sentence.

(11) Section 2.1: The authors can choose to add a column in Table 1 called ‘Notes’ or ‘Site Description’. By adding information in Table 1, they can and should cut out a lot of the details from this section. Given that the authors have not covered all the TCCON sites, or all seasons at all sites, it is unclear why specific sites have been discussed. This entire section should be rephrased and re-structured. Kindly understand that the readers’ time is valuable and provide information that is necessary and relevant. For example, how is the population of specific cities necessary to interpret any of the results in this manuscript (Page 5, Lines 113-116)?

The populations give a sense of the size of the urban region and thus its likely anthropogenic CO2 emissions. The section has been substantially shortened and the table has remained unchanged.

“There are several TCCON stations that are located in regions with significant spatial variability in topography or ground cover. For example, the Lauder TCCON station is in the midst of rolling hills, the Wollongong TCCON station is between the ocean and a sharp escarpment, and the Edwards TCCON station is adjacent to a very bright playa, a land surface property previously identified from the Greenhouse Gases Observing Satellite[GOSAT, Kuze2009,Kuze2016] results as challenging for XCO2 retrievals [Wunch2011a]. With target-mode measurements, the impact that local surface variability has on the XCO2 retrievals becomes apparent.

Other TCCON stations (e.g., Park Falls, Lamont) have relatively uniform surface properties and are reasonably far from anthropogenic CO2 sources, but the ground cover can vary from season to season. The Sodankyla and Eureka sites, located at high northern latitudes, challenge the OCO-2 algorithm at very high solar zenith angles and airmasses, and with snowy scenes. Izana, Reunion and Ascension, all lower-latitude sites, are located on small islands remote from large land masses, but with significant topography. The Izana TCCON station (28.3N) is at 2.37 km altitude, whereas the Reunion (20.9S, 0.087 km) and Ascension Island (7.9S, 0.032 km) stations are closer to sea level.

Several TCCON target stations are near or in urban regions with varied topography and emissions sources: Pasadena (pop. ~17 million), Tsukuba (pop. ~228,000), Paris (pop. ~2.24 million), and Karlsruhe (pop. ~300,000).”

(12) Page 6, Line 144: This is the first time that the phrase “warn level” has been used. What does this mean? Please provide a description and point to appropriate references.

An explanation of warn levels has been added: “Warn levels determine sets of OCO-2 data with consistent quality data (as defined by the RMS scatter) within an observation mode [Mandrake2013,Mandrake2015]. A significant volume of data is required to generate warn levels which is difficult to achieve with the reasonably sparse target mode data. Furthermore, warn levels in one measurement mode are not equal in quality to another mode."

(13) Page 6, Line 148: It is unclear what the authors mean by – "limitations in the information content of the measurements" and how it causes systematic biases.

Added clarification: "... by limitations in the information content of the measurements (i.e., the spectra do not contain enough information to resolve multiple independent vertical pieces of information)..."

(14) Page 6, Line 152: Add the term algorithm or procedures after "bias corrections".

Done

(15) Page 6, Lines 152-157: The authors may want to number the three key biases as (a), (b), (c), and discuss them in the order they are numbered.

Changed the order to be consistent with the discussion.

(16) Page 6, Line 170: The authors should provide more details here for the average reader – how can examining data near coastlines provide an estimate of biases? Or refer to the OCO-2 technical documents.

Added a reference to Mandrake et al., 2015.

(17) Page 6, Lines 176-177: This is another example of a poorly written statement, which can be easily misinterpreted. Generating high-quality OCO-2 data is crucial for obtaining surface flux estimates with reasonable accuracy. I do not understand why the authors want to make this consistent with "the state of the art inversions of surface in situ data". I believe the authors are implying a statement about the quality of the XCO2 estimates and the need to bring them on an equal quality level with the surface in situ data – I don't see why inversions and fluxes are brought into the mix here. Kindly rephrase.

Rewritten: "Placing the OCO-2 data on the World Meteorological Organization's (WMO) trace-gas standard scale is crucial for obtaining accurate flux estimations that are consistent with the inversions of the surface in situ CO2 measurements that are carefully calibrated to the WMO scale [Zhao2006]."

(18) Page 7, Lines 191-204: Can the authors comment on the significance of their calculated correlation coefficients? In fact this is more relevant for Section 4 (i.e., Table 3), where in several cases N<10 at an individual TCCON site. The authors need to proceed with caution when calculating statistics using a low number of samples. And that should be acknowledged as a caveat.

The table now shows sites with N>10 in bold font to indicate that they are more statistically significant.

(19) Page 9, Lines 281-283: I do not understand why the correlation coefficient is impacted because the XCO2 variability is lower. It may impact the RMS but I do not understand the rationale behind the impact on R2. Can the authors clarify their reasoning here?

Removed statement.

(20) Page 10, Lines 294-295: This statement should come with a set of caveats about the way the data were selected and/or filtered. See earlier comments.

Done: "Aggregated OCO-2 XCO2 estimates filtered with warn level <=11 and outcome_flag=0 generally compare well with coincident TCCON data at global scales, with absolute mean biases less than 0.4 ppm and RMS differences less than 1.5 ppm.”

(21) Figure 8: The positive differences observed at Reunion Island stands out in this graph. Can the authors comment on why such large differences are being observed at this location? Such large differences do not show up in Figure A1(r) or the statistics reported in Table 3.

It is likely pulled up by the glint over ocean high bias in the southern hemisphere winter.