# Response to Reviewer 2 comments on: “Validation of MOPITT carbon monoxide using ground-based Fourier transform infrared spectrometer data from NDACC”

by Rebecca R. Buchholz et al.

*Correspondence to:* buchholz@ucar.edu

We thank Reviewer 2 for their valuable comments. Below, reviewer comments (in italics) are addressed separately and the manuscript locations of changes are indicated. We have shortened some of the reviewer comments, while retaining the underlying message. A truncated comment is denoted as “[...]”. Please see reviewer response document for full comments.

***1. Averaging the MOPITT profiles and the MOPITT averaging kernels***

*[...] I have concerns regarding these averages. Are we allowed to average averaging kernel matrices? Geophysical representation problems may arise. What would M. Clive Rodgers say?*
*When validating satellite data with measurements with higher vertical resolution (FTIR, airborne profiles), one should smooth the high resolved profile by the satellite averaging kernels. This should be done pair by pair [...]*

In summary, Reviewer #2 supports this view by indicating that pair-by-pair is the methodology used in other papers (Kerzenmacher et al., 2012; de Laat et al., 2014; and Martínez-Alonso et al., 2014).

*The authors argue that it reduces the number of comparisons, that is for sure.*
*They argue that using this method of averaging profiles and AK has been tested, against comparisons by pairs. Page 6, lines 7-11, they quote results from the supplementary material of Martinez-Alonzo et al. (2014):[...]*
*a) It is worth noting that Martínez-Alonso et al. (2014) conducted the study using pairs comparisons, not averaging. The averaging method (called “approximation method”) is compared to the “full method” (described in the article) in a supplementary material. This method is presented to be relevant for the MOPITT Level 3 data analysis [...]”*

Martínez-Alonso et al. (2014) show the supplementary method so that their results can be applied to the MOPITT Level 3 research product. We extend the work of Martínez-Alonso et al. (2014) to include a comparison specific for the averaging technique used in this manuscript for total column values (details below).

*b) I’m puzzled but the results presented in these supplementary material (Table S2). Difference between two percentage values are expressed in percentage, which can be quite misleading. One should use percentage point (pp).*

We understand the distinction that is being made. For example 10% and 7% have 3 pp between them but are 30% different from each other, using the first value as a reference. We now use pp when mentioning a change in a % value.

*c) [...] If I understand Table S2 [of Martinez Alonso et al] well, one should read that the differences between the bias (in percentage) obtained by the two methods differ by 0.8 pp maximum for MOPITT/ACE-FTS comparison and by 4.2 pp maximum for MOPITT/HIPPO comparisons. The authors should correct this part of the paper, when quoting the supplementary material of Martínez-Alonso et al.*

Values in the parentheses of Table S2 [of Martínez-Alonso et al. (2014)] correspond with differences between smoothed values and MOPITT. These are the values we are interested in. Our value of 0.8 pp for MOPITT/ACE-FTS comparison was correct, but we have corrected the MOPITT/HIPPO value to 1.8 pp.

*The authors should use the proper method for this validation work, which is working by pair, i.e. smoothing each FTIR measurement by one collocated MOPITT AK, as many times as there are collocated MOPITT retrievals.*
*If the authors want to perform this "approximation method", they should clearly say this is not the usual way to perform validation and they should prove that we can rely on such biases with test-results for different stations [...]*

Our main reason for spatial averaging is to reduce random error in the satellite product (page 5, line 32 to page 6, line 1). We highlight this by reorganizing our statements about the benefits of spatial averaging. In our study, we are not interested in analyzing the random variability introduced by small geographical differences around a station. Instead we are interested in overall instrument bias and long-term drifts. We agree that there are no studies evaluating the effects of spatial averaging on validation results using total column. As such, we expand on the Martínez-Alonso et al. (2014) analysis to compare validation results from the point-wise technique with the 1° daily spatial averaging technique in a supplementary section, Appendix B.

In Appendix B, we show that averaging greatly reduces random error (R is improved), while bias and drift are generally equivalent. Our values are consistent with Martínez-Alonso et al. (2014). We conclude our method is sound and a valuable alternative to point-wise.

**List of changes from Reviewer 2, comment 1:**

- Page 5, line 28: to page 6, line 2: Re-organised
- Page 6, line 11: 1.2 pp -> 1.8 pp
- Page 6, lines 7-11: Reworked.
- New Appendix B: added to compare spatial averaging to point-wise validation.

**2.** ***Removing the lowest level of a MOPITT profile when different topography.***

*[...] Removing information of the MOPITT retrieved profile, as well as in the averaging kernel matrix raises a lot of questions. Are we allowed to remove one layer of information, i.e. to remove one column and one line of the averaging kernel matrix? It doesn't make any sens. Using a truncated matrice would introduce biases on every levels of the profile to be smoothed. The authors should extend the FTS profiles instead, like it is done in Kerzenmacher et al. (2012) and in de Laat et al. (2014). In both papers, satellite a priori information is used to extend the FTS/airborne profile.*

There are two reasons that our comparison procedure avoided using the scaling method to replace lower missing levels in the FTS profile. The first is that extrapolating below ground level of the FTS measurement site is not physically valid. The second is that using satellite a priori to fill in the "missing" information (as in Kerzenmacher et al., 2012) weights the “truth” toward the satellite retrievals, with the potential to artificially improve validation results. The lowest profile level can contribute over 20% to the total column amount. Therefore, estimating this value can produce a substantial influence on the “truth”. When more than one of the lowest levels needs to be estimated, influence on the “truth” will be larger.

However, we agree that applying averaging kernels on truncated profiles will also introduce bias at every level of the smoothed profile. We therefore asses the effect of both profile truncation and profile filling in a new Appendix C. In our version of profile filling, we use FTS a priori from WACCM (instead of satellite a priori) to calculate scaling factors. WACCM values are interpolated to the MOPITT profile coordinates.

In summary, Appendix C shows that while both effects are relatively small, profile filling is the superior method as it produces smaller changes in the bias calculations. Therefore, the methodology has been updated to use profile filling instead of profile truncation.

Further assessment of the profile filling (WACCM scaling) methodology against the original methodology is shown in the following Table 1 for the V6 TIR-NIR product. Results from comparing the two methodologies are consistent with the results in Appendix C, and show that generally the largest changes in relative bias are seen for those stations where a higher number of data points show altitude differences between instruments. Overall, validation values are only slightly altered by using the profile filling technique. For example, mean biases have changed from 2.8 to 2.4% for TIR-only; from 5.4 to 5.1% for TIR-NIR ; and 7.0 to 6.5% for NIR-only. Thus, the major conclusions of the manuscript have not changed.

**Table 1.** Comparing original results with replacing the lowest level (V6 TIR-NIR).

| Station | Original (with truncation) | | | WACCM scaling | | Difference‡ | | % data where |
|---|---|---|---|---|---|---|---|---|
| | r | bias (%) | bias (%) | r | bias (%) | r | bias (pp) | $P_{surf}$ differs |
| | | rel MOPITT† | rel FTS† | | rel FTS | | | FTS < MOP |
| Eureka | 0.67 | 3.36 | 3.48 | 0.88 | 3.06 | -0.21 | 0.42 | 100 |
| Ny Alesund | 0.66 | 12.15 | 13.84 | 0.61 | 12.86 | 0.05 | 0.98 | 9 |
| Thule | 0.66 | 3.58 | 3.71 | 0.64 | 3.56 | 0.02 | 0.15 | 37 |
| Kiruna | 0.83 | 4.50 | 4.71 | 0.80 | 3.99 | 0.03 | 0.72 | 19 |
| Bremen | 0.80 | 10.26 | 11.43 | 0.67 | 11.57 | 0.13 | -0.14 | 25 |
| Zugspitze | 0.87 | 0.03 | 0.03 | 0.90 | 1.67 | -0.03 | -1.64 | 100 |
| Jungfraujoch | 0.66 | 1.88 | 1.91 | 0.82 | -1.1 | -0.16 | 0.81* | 100 |
| Toronto | 0.77 | 8.78 | 9.62 | 0.76 | 9.09 | 0.01 | 0.53 | 2 |
| Izaña | 0.91 | -0.59 | -0.59 | 0.90 | 2.24 | 0.01 | -1.65* | 100 |
| Mauna Loa | 0.93 | -1.35 | -1.33 | 0.92 | -1.91 | 0.01 | -0.58 | 100 |
| Reunion-St Denis | 0.94 | 4.57 | 4.79 | 0.93 | 3.94 | 0.01 | 0.85 | 100 |
| Wollongong | 0.83 | 8.72 | 9.55 | 0.79 | 9.3 | 0.04 | 0.25 | 25 |
| Lauder | 0.93 | 11.09 | 12.47 | 0.93 | 11.24 | 0.00 | 1.23 | 8 |
| Arrival Heights | 0.73 | 8.90 | 9.17 | 0.79 | 6.65 | -0.06 | 2.52 | 20 |

† The relative bias is shown both relative to MOPITT, as presented in the original manuscript, as well as relative to FTS, following the new methodology suggested by Reviewer 2, comment #4.

‡ Difference is calculated $|\text{original method}| - |\text{WACCM scaling}|$

* bias changes sign

## List of changes from Reviewer 2, comment 2:

- Page 6, line 32 to page 7, line 6: Altered methodology description for when FTS altitude is higher than MOPITT.
- Tables 2-5 and Figures 3-9, 11-15: Small changes in the validation results have slightly altered these tables and plots.
- Page 1, line 9; page 8, lines 25-28; page 9, line 1, 7, 9; page 11, line 31; page 12, lines 13-17; page 13, lines 5-6: Validation values have slightly changed.
- Appendix C: added to quantify the effect of profile truncation and filling.

**3.** ***Presenting averaged averaging kernels matrices (Fig. 3, 4, 12, 13).***

*One should be extremely careful when presenting an average averaging kernel matrix. It is rarely done. Some time it is but if one is sure that all the averaging kernel matrices look exactly the same. What if one or two matrices show completely different sensitivities?*

We are interested in the mean behavior of the AKs and how that differs between retrieval methodologies. Our aim in presenting mean averaging kernels is to describe the persistent differences between retrievals over different surface types (land versus water), using different spectral ranges (TIR, TIR-NIR, or NIR), or between different instruments (MOPITT versus FTS). Assessing the mean values is the least subjective way to assess differences. Our method was careful not to average AKs that could have vastly different shapes. Specifically, AKs from different surface-types were not combined and daytime-only retrievals were used.

Mean averaging kernels have been presented in numerous articles. For example, Vigouroux, C., et al. (doi:10.5194/acp-12-10367-2012, ACP, 2012) presents arithmetic mean VMR AKs, Martínez-Alonso et al. (2014) as well as Deeter et al. (2015 & 2014) present spatial average VMR AKs over a 4° to 5° box, Kerzenmacher et al. (2012) presents mean column AKs.

For these reasons we choose to continue showing mean AKs, but have altered the plots as listed below.

*I am not sure to have ever seen that for seasonal averages (Fig. 3 and 13) and I am not comfortable at all with this plots. It would be straightforward for the authors to pick one representative AK per season, e.g. the 15th of one month. Of course it is subjective, but we trust the authors to choose one representative AK. A least this AK matrix would be realistic, the result of one inversion. Authors could also present DFS time series.*

We now use the timeseries of DFS to indicate seasonality in measurement sensitivity, instead of column AKs.

**List of changes from Reviewer 2, comment 3:**

- Figures 4 and 12 are combined, now Figure 2.
- New Figure 2: Added column AKs.
- Figures 3, 12 and 13: removed.
- Page 7, lines 9-10: "(e.g. Toronto in **Fig. 2; bottom right plot**)"
- Page 7, line 23: "Sect. 4.1, 4.2 and 4.3." -> "Sect. 4.1 and 4.3."
- New Figure 3: DFS timeseries plot.
- Page 9, line 4-5: "Instrument sensitivity varies with season, which is reflected in the **DFS** seasonal variability. An example of the range of **seasonal** variability is shown by the **TIR-only DFS timeseries** at Toronto and Wollongong (Fig. 3)."

**4. Calculating relative bias**

*[...] The relative bias values should be calculated with respect to the FTS, not MOPITT, as FTS measurements are supposed to describe the "truth". This is how it is done in Kerzenmacher et al. (2012). Another way is to present only absolute biases, like in de Laat et al. (2014).*

We agree that biases relative to FTS "truth" makes more sense and have changed our calculations and descriptions to reflect this. Bias values generally change less than one percentage point (e.g. Table 1 in the response to Reviewer 2, Comment #2) and the overall conclusions of the manuscript have not changed.

**List of changes from Reviewer 2, comment 4:**

- Tables 3 & 4, Figures 7, 8, & 10 (now Figures 6, 7 & 9): Updated with new values.
- Page 1, line 9; page 8, lines 25-27; page 9, line 7, 9; page 11, line 31; page 12, lines 13-17; page 13, lines 5-6: Bias and bias drift values changed.

**General comments/technical corrections**

*The y labels of your plots should be "Pressure (hPa)" and not "Altitude (hPa)". The x labels "Averaging kernel value" should be "Averaging kernels".*

Axis labels have been changed. New Figure 2 is the only remaining plot where this was an issue.

*Page 1 line 10 : Replace "greater" by "higher" or "larger"*

Replaced with larger.

*Page 2, line 15: "Emmons et al., 2009, 2004)" > "Emmons et al., 2004, 2009)"*

Changed order of references. Also Deeter et al., 2014, 2010 > Deeter et al., 2010, 2014.

*Page 4, line 19 : "The most recent retrievals are used for this validation study" > "The most recent retrievals available in 2016 are used for this validation study"*

Added "available in 2016" to sentence page 4 line 19.

*Page 6, line 30: "Fig. 2 (a)" Please mention "inspired by Fig. 2 of Kerzenmacher et al. (2012)", in the text or in the caption.*

Figure 2 was removed because the updated methodology is very similar to Kerzenmacher et al. (2012) (in response to Reviewer 2, Major comment #2). Page 6 line 32 to page 7, line 6: reworked to describe altered methodology (see response to Major comment #2).

*Page 7, line 28: “had” > “as”?*

Now page 7, line 30 "..., **assuming** the FTS measurement **describes** the true atmosphere."

*Page 7, line 28: “described” > “describes”?*

see above

*Page 8, line 11: “DFS are provided for each MOPITT retrieval and are calculated for FTS measurements from the AK matrices”. It sounds like MOPITT DFS are not calculated from the AK matrices. Consider “For each MOPITT retrieval and FTS measurement, DFS are calculated from the AK matrices”.*

Changed to: “DFS are calculated from the AK matrices for each MOPITT and FTS retrieval.”

*Page 8, lines 12-15: These two sentences should be removed. Unclear. Yes, theorical DFS may be higher and also smaller depending on the a priori covariance matrix. We know that the authors don’t perform instrument comparison.*

Removed

*Page 9, line 1: “(apart from at Ny-Ålesund and Bremen)” > “(apart from at Ny Ålesund, Bremen and Lauder)”*

Now page 8, line 25: Added Lauder to the list.

*Page 9, line 2: “(except at Ny-Ålesund)” > “(except at Ny-Ålesund, Lauder and Arrival Heights)”*

Now page 8, line 26: Added Lauder to the list.

*Page 9, line 8: How do you perform the normalization? I see that some values are above 1.*

Normalization was completed relative to the complete dataset mean AK. However, the seasonal column AK figures (3 and 13) have been removed in response to Reviewer 2, Comment #3. Consequently, page 9, Line 7-8 (now line 4-5) now references DFS timeseries plots instead.

*Page 9, lines 22-30: This paragraph is not very convincing. Above 30 ° N, there are low and high correlations for the 3 MOPITT products. I don’t understand the first sentence (line 21).*

The new comparison shows even less dependence. This paragraph has been removed.

*Page 13, lines 4-5: What would the mean drift be if we remove the stations above 60 °N?*

Now page 12, line 30-31: Added in sentence “Mean trends below 60° N are -0.12% $\text{yr}^{-1}$ for TIR-only, -0.33% $\text{yr}^{-1}$ for TIR-NIR and -0.17% $\text{yr}^{-1}$ for NIR-only.”

*Page 18, line 15: “instrument.,” > “instrument,” (remove dot)*

corrected

*Page 20, footnote: “125 HR” > “125HR” (remove space)*

corrected

*Page 21: “where not enough pixels”: how many is enough? Are these median DFS calculated from the MOPITT retrievals used for the validation (tables 3a, 3b, 3c)? The numbers of retrievals are those indicated in Tables 3a, 3b and 3c on the ”# obs” line? It should be mentioned.*

Table 2: Altered table footnote to mention the threshold where the number of pixels of each surface type were no longer approximately equal. Updated table caption to mention that DFS are determined for overlapping times between instruments and that the number of retrievals correspond with numbers in Tables 3a, 3b, 3c and 4.

*Page 25: “Values for land validation is” > “Values for land validation are”*

corrected

*Page 32: In the caption with the symbols: Replace “pixels 2-4” by “pixels 2 to 4” or “pixels 2-3-4” or “pixels 2→4”*

(now page 31) replaced with “2 to 4”; also changed on page 10, line 20 and 21

*Page 33: “bias (second row)” > “bias ((MOPITT-FTS)/MOPITT, second row)”. But the method should be (MOPITT-FTS)/FTS, see specific comment #4.*

Now page 32: “bias (second row)” > “bias ((MOPITT-FTS)/FTS, second row)”

*Page 34: The bias (MOPITT-FTS)/MOPITT is not the same as the retrieval bias. Modify the caption.*

Now page 33: retrieval > relative; added (MOPITT-FTS)/FTS