Referee #2

We would like to thank Referee #2 for the valuable comments and her/his time to review the manuscript. The questions/comments/concerns of Referee #2 (bold text) are addressed below. New text in the manuscript is given in italics.

Specific comments made by Referee #2:

1) Throughout the document – the WMO network compatibility goals should be referred to as "network compatibility" rather than just "compatibility" to distinguish the WMO usage from a strict metrological definition of compatibility. Or making the distinction upon first usage if that is preferred.

We changed to network compatibility goals throughout the entire document.

2) Page 2, line 11: Define Empa on first usage.

Done

3) Page 2, line 14: Define who is the WCC-N_2O on first usage.

Done

4) Page 4, line 10: "WCC-N_2O uses a set of TS traceable to a set of secondary standards …" Are these standards truly secondary (compared directly to the primary standards) or are they actually the normal tertiary standards distributed by the CCL? Or does this "secondary" label relate to the hierarchy internal to the WCC-N_2O?

Thank you for this comment. Yes, it was related to the internal hierarchy. The WCC-N_2O hosts a set of tertiary reference standards, which are regularly calibrated by the CCL. Those standards are used to transfer the scale to TS. Therefore, the laboratory standards hosted by WCC-N_2O are tertiary. To make that clear the corresponding sentence reads now: "WCC-N_2O uses a set of TS traceable to a set of tertiary standards, *which are regularly recalibrated against secondary standards at the CCL*."

5) Page 4, line 11-14: Have all of the comparisons been reprocessed onto the current CO_X2014A and N_2O_X2006A scales or are they presented on the current scale at the time of the comparison? If the latter is the case I would suggest making the scale explicit by adding a column to tables 1 and 2 showing which scale was actually used. As a follow up I would ask if the comparisons change significantly if reprocessed onto the current scale? This may not be possible within the scope of this paper but would be of interest to data users who would like to use this information to understand potential biases between data from various providers.

The data has not been reprocessed, and the comparison shows the results using the scales at the time of the audit. We added columns to table 1 and 2 with information on the calibration scales, as suggested by the referee. We added more information on the calibration scales in chapter 2.1, since we were using standards calibrated on the WMO-X2000 CO scale until 2011. However, we only used standards with an amount fraction larger than 185 nmol/mol, and comparisons showed that the two calibration scales agree well at these levels.

"*WCC-Empa continued using the WMO-X2000 calibration scale until 2011 but used only standards with an amount fraction larger than 185 nmol/mol. At these amount fractions, the difference between the WMO-X2000 and WMO-X2004 CO scales are very small and questionably significant within their uncertainties. We therefore consider these two scales as being identical for calibrations made at WCC-Empa.*"

We also realised that some of the comparisons involved different calibration scales. However, we decided to keep them since the contribution to the bias due to the scale difference is generally much smaller than the observed variability of all comparisons. The information about the calibration scales used at the stations and the WCCs is available from table 1 and 2. To further highlight a potential bias due to scales differences, we used different symbols in Figures 1 and 4 in the revised paper for the cases with differences in calibration scales.

We also agree that it would be interesting to see if reprocessing onto the current scale would improve the results. However, this is complicated because scale changes may not be linear, and individual reprocessing would have to be applied for each case. This is beyond the scope of the current paper.

6) Page 5, line 1: When did the parallel measurement approach begin?

The first parallel measurements were made in 2011. We added this in the revised version of the manuscript.

6) Page 6, line 28: WMO network compatibility goals are no longer listed as "±".

Correct. We changed it when we refer directly to the WMO network compatibility goals throughout the manuscript.

7) Page 7, line 3-6: Have any of the comparisons been reprocessed after the stations have had working standards re-calibrated and drift corrected? It would be very interesting to see if some of these larger offsets are improved with better calibrated standards. This would also provide a more accurate assessment of the bias in the station data but again may be beyond the scope of this paper.

In some cases, data was reprocessed during or immediately after the audit, but the first assessment was always done without changes to the system. In those cases, only the results of the second comparison are shown, since they better represent the current performance of a station. We added a sentence to clarify this in section 2.1:

"*Since the focus of the paper is on instrument performance, only comparisons involving fully functional instruments were considered. Furthermore, if data has been reprocessed due to any known biases e.g. in working standards, only the results of the final comparison were considered, since they best represent the performance of the measurement instruments at the time of the audit.*"

We further added a footnote in table 1 to clarify that the differences observed during the station audits at Lauder were due to an offset in a working standard and not related to the performance of the FTIR system.

We also agree that a study of the temporal evolution due to reprocessing based on known biases and the impact of multiple audits at stations would be an interesting topic. However, as the referee recognised, this is beyond the scope of the current paper.

8) Page 8, line 21: Is this statement supported by meta data from the stations, i.e. is there a record of the number of standards used for those early audits that would support this conclusion?

A SOP for conducting audits is available at https://www.empa.ch/documents/56101/250799/2.pdf/f5a8c0a5-884f-4e0c-b836-96d92dbd260c. The WCCs strictly follow that SOP during audits conducted. Thus, for each audit the meta-information of the number and hierarchy of standards as well the amount fractions of the standards and the scale used for calculating measured ambient air amount fractions is available at the WCCs, also for the early audits.

9) Page 11, line 7: The figure plots the data as Station – TI, the text has the offset as TI – Station. I suggest changing either the sense of the comparison in this sentence or in the figure to be consistent.

We changed the sentence to "*The bias of the PUY analyser significantly decreased to 1.20±0.57 nmol/mol (1σ).*" To be consistent, we also changed the sentence on page 10, lines 15-16 to "*During this period, the PUY analyser was measuring on average 5.85±0.94 nmol/mol (1σ) higher than the TI.*"

10) Page 11, line 12: I think the description of the AMY offaxis_ICOS instrument should be "enhanced performance" off-axis integrated cavity output spectroscopy rather than the stated "cavity enhanced".

We changed it on page 3, lines 3-4 where it was also wrong, and used the abbreviation OA-ICOS on page 11.

11) Page 11, line 17-19: I suggest putting in the values for each third of the time period to show how different they are and how much of the variability is due to the calibrations.

Done

12) Page 17, Table 1: I suggest listing the CO scale for each comparison if they are not all the same.

We included this information in the revised manuscript.

13) Page 19, Table 2: I suggest listing the N_2O scale for each comparison if they are not all the same.

We included this information in the revised manuscript.

14) Page 22, Figure 3: As mentioned in the text there are only 2 comparisons of the same FTIR instrument. It might be good to show this by listing n values for each category or at least for the FTIR. I might also suggest keeping the same categories at shown in figure 2 (combining NIR-CRDS and QCL) to be consistent between the two figures but leave this to the author's discretion.

The number of comparisons is now listed in Figure 3. However, we prefer to keep the results for each single techniques in this figure, since FTIR for example does not appear in any category in Figure 2.

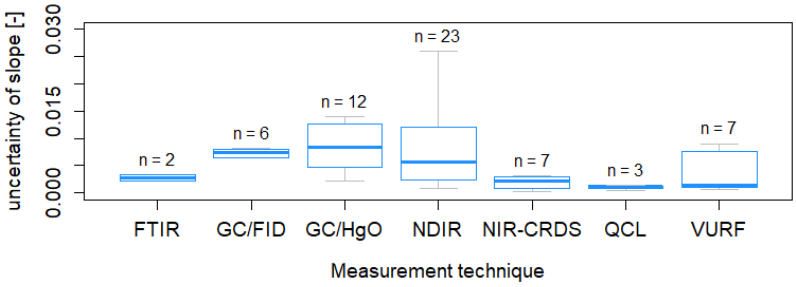

Figure 3. Boxplot of the slopes uncertainties of the of the regression analysis for the CO performance audits for different analytical techniques including the number of comparisons (n). The horizontal blue line denotes to the median, and the blue boxes show the inter-quartile range.

15) Page 24, Figure 7 caption (and other time series plots): The caption says "(1 h data)". I take this to mean the data from both instruments was averaged to hourly averages. If true I suggest making this point clearer.

Yes, this is correct. We changed the figure caption to "*Comparison of hourly averages of CO at PUY between the WCC-Empa travelling instrument and the PUY Picarro G2401 for the period when the TI sampled humid air. …*"

Minor technical corrections suggested by Referee #2:

All technical corrections suggested by Referee #2 were accepted and changed in the revised manuscript.