Dear Dr Janssen,

We hereby submit a revised version of our manuscript "Recent advances in measurement techniques for atmospheric carbon monoxide and nitrous oxide observations". We addressed all comments made by the two referees and would like to thank for their constructive and helpful feedback.

The replies to the referees as well as a revised manuscript with track changes can be found below.

With kind regards on behalf of all co-authors,

Christoph Zellweger

Referee #1

We would like to thank Referee #1 for the valuable comments and her/his time to review the manuscript. The questions/comments/concerns of Referee #1 (bold text) are addressed below. New text in the manuscript is given in italics.

General comment made by Referee #1:

The introduction section describes the background of the field and the aim of the paper shortly but still adequately. Results are dealt with high expertise and there is no doubt that authors have a deep understanding of the field. Since some graphical results are complicated, slightly longer explanations would help the reader, without a need to find a clarification from references. Tables and graphs itself are clear. Authors are referred to the key articles of this topic. The English of this paper is mainly sufficient. Altogether, my opinion is that the scientific quality of this manuscript is high, and it should be published with revisions.

Thank you for this comment. We agree that especially the bias / slope plots might be difficult to understand without further explanation. We refer to Zellweger et al. (2016) for a detailed description of the methodology, including an illustrative plot showing different bias / slope combinations. We believe that duplication of this explanation is not needed here. However, we added a sentence to clarify on page 6 / line 27: "*Perfect agreement would result in bias / slope pairs of (0 nmol/mol / 1)*".

Specific comments made by Referee #1:

-page 4, section 2.1: there is probably a SOP for the TS measurements at the stations? Like, how the TSs are prepared for the measurement? Is there used a similar sequence for TSs as the station uses for its own cylinder measurements or instructed by WCC? Is the protocol dependent on the instrumentation? Who is measuring the TSs in practice? I am not asking to answer all of these, but just shortly clarifying how the TS measurements done; with a certain fixed protocol or not; since this may affect to the results as well. Stations personnel may have bad habits, but they may also know by experience if their instrument needs special way to handle it.

Yes, there is a SOP for TS measurements at stations, which is available from the WCC-Empa webpage (https://www.empa.ch/documents/56101/250799/2.pdf/f5a8c0a5-884f-4e0c-b836-96d92dbd260c). However, there is no fixed protocol, since this depends on both the instrumentation and the capabilities of the stations. Usually, station operators are analysing the TS as unknown samples on their system, with at least three repetitions for each TS. It is clear that this approach has limitations, and additional parallel measurements with a fully independent analytical system are desirable and are implemented whenever possible during station audits. Brewer et al. (2019) recently showed that results of the audit approach by WCC-Empa are comparable to the WMO round robin experiments (NOAA, 2018). This confirms that the protocol followed by WCC-Empa is a valid approach to assess instrument performance. The limitations of the TS method are listed in section 2.2. We added the following sentence in section 2.1:

Usually, multiple analysis of a set of three or more TS is made and averaged for the final assignment of the TS value by the audited laboratory.

-page 5, lines 11-12: two questions concerning water vapour correction of WCC. 1) Since droplet method is not giving appropriate results, but in page 12 you are listing two - most probably superior - methods to determine water vapor correction function, why WCC is not using or even testing those

methods? 2) If the correction of the effect of water vapour to CO measurements is difficult, why WCC is not drying the sample?

1) It is true that the droplet method has limitations regarding the determination of the water vapour interference for CO. The methods described later in the paper were not yet published / available when the experiments done for this paper were made. Potentially, they would give better results, which would allow for a better compensation of the bias. However, a large part of the uncertainty is due to relatively large instrumental noise of the Near-IR-Cavity Ring Down Spectroscopy (NIR-CRDS) technique, which makes the determination of the appropriate correction function challenging. Furthermore, unpredictable short-term changes as shown in section 3.2 / Figure 10 further complicate appropriate correction.

2) As a consequence of the results, we are now drying the air, which is also one of the recommendations of this paper (see conclusions).

-page 8, lines 27-31/Figure 5: concept "relevant amount fraction" for N_2O remains a bit unclear. So, is the "relevant amount fraction range" a central 10 nmol/mol range representative for the unpolluted troposphere for the year of the audit and the "relevant amount fraction" a single representative value ±0.1 nmol/mol? Please try to state this more explicitly in the text.

Yes, this is correct. The amount fraction range is given in Table 2 for all audits. We agree that the description of the concept regarding the representative amount fraction was not clear in the manuscript. The relevant amount fraction is a single value and depends on the year of the audit. It was calculated using the value of the unpolluted air in 2016 and an annual growth rate of 0.8 nmol/mol. This value corresponds to the centre of the amount fraction range given in Table 2.

We made the following changes in the manuscript to better explain the concept:

Page 8, line 7: added "*range*".

Page 8, line 31: we added "*The relevant amount fraction corresponds to the value at the centre of the relevant range for the corresponding year.*"

-pages 8-9, section 3.1.2: it is clear how and why the compatibility goal for N_2O is determined. However, you clearly show in this manuscript that it will not happen in near future that atmospheric N_2O measurements will reach the goal. So, should the goal be revisited? I would like to see a statement in conclusions section regarding the necessity of these goals no-one is reaching.

We added the following sentences (from the report of the 19th WMO/IAEA Meeting on Carbon Dioxide, Other Greenhouse Gases and Related Tracers Measurement Techniques (GGMT-2017)) after the definition of the network compatibility goals in the introduction (page 2, line 20ff): "*These goals represent the maximum bias that can generally be tolerated in measurements of well-mixed background air used in global models to infer regional fluxes. Some network compatibility goals may not be currently achievable within current measurement and/or scale transfer uncertainties. However, they are targeted for applications which require the smallest possible bias among different datasets or data providers, such as for the detection of small trends and gradients (WMO, 2018).*"

We believe that it is beyond the scope of this manuscript to make recommendations regarding the compatibility goals, since this is done in the larger scientific community during the GGMT meetings.

-page 11, lines 1-2/Figure 10: if I understood correctly, the droplet tests are always made with the same CO amount fraction (300 nmol/mol), but is the water vapour response dependent on amount

fraction of CO, i.e. have you run the test with other CO amount fractions? It is a bit hard to see from the graph 10; there is no chronological trend, but the water vapour response is varying randomly?

We would like to thank the reviewer for this comment, since it made clear that our description of the experiment was not detailed enough. The water vapour experiments were made at different amount fractions ranging from 57 to 741 nmol/mol CO, but the limits shown in Figure 10 refer to an amount fraction of 300 nmol/mol. We found no dependency of the water vapour correction function on the amount fraction for both instruments, which is shown in Figure R1 where the ratios of CO(humid, corrected) / CO(dry) at a water level of 3% are plotted against the CO amount fractions used for the corresponding determination of the correction function. The limits for the WMO/GAW compatibility goal (green) and extended (yellow) compatibility goals are also shown.
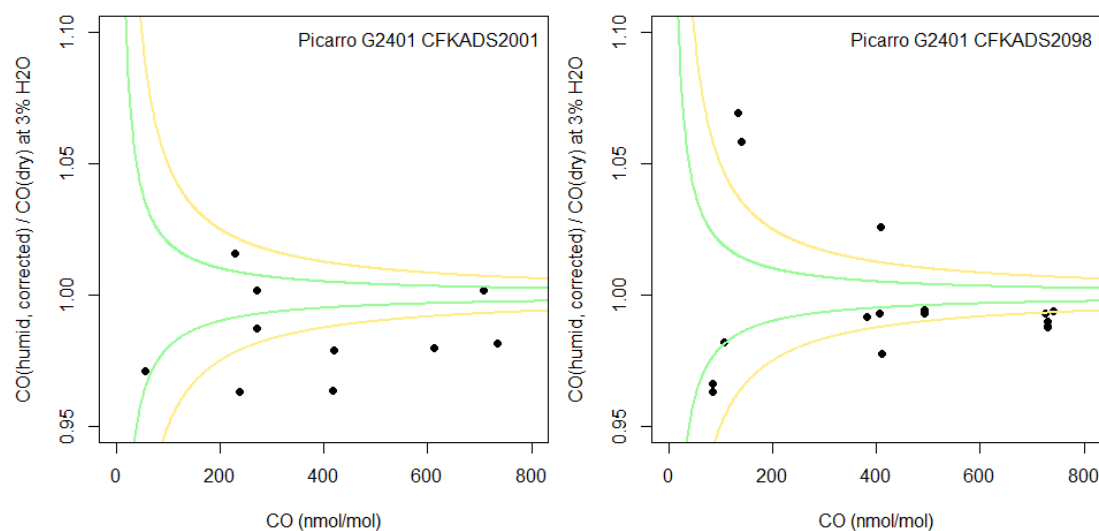
Figure R1. Black dots show the ratios of CO(humid, corrected) / CO(dry) amount fractions at a water level of 3% vs. the CO amount fraction of the CO standards used during the corresponding experiments of two different Picarro G2401 NIR-CRDS analysers. The coloured lines show the limits for the WMO/GAW compatibility goal (green) and extended (yellow) compatibility goal.

We also were not able to see a clear temporal trend in the change of the correction function, which is illustrated in Figure R2 below where the ratios of CO(humid, corrected) / CO(dry) at a water level of 3% are plotted against the time in weeks since the first determination of the correction function. The limits for the WMO/GAW compatibility goal (green) and extended (yellow) compatibility goals are also shown in Figure R2 as horizontal lines for an amount fraction of 300 nmol/mol CO. It looks like the Picarro G2401 CFKADS2098 gets less stable after approximately 150 weeks. However, both instruments were not able to fulfil the WMO/GAW network compatibility right from the beginning at a water level of 3%. This clearly supports the conclusion of our work that the internal water vapour compensation is inappropriate, and drying of the air sample is recommended.
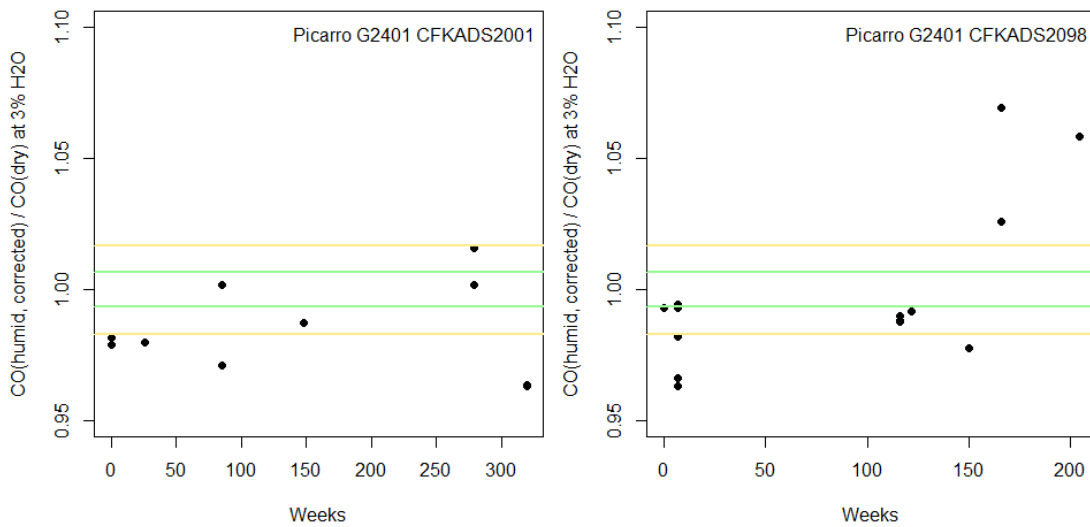
Figure R2. Black dots show the ratios of CO(humid, corrected) / CO(dry) amount fractions at a water level of 3% vs. time in weeks since the first determination of the water vapour correction function of the two different Picarro G2401 NIR-CRDS analysers. The coloured lines show the limits for the WMO/GAW compatibility goal (green) and extended (yellow) compatibility goal at an amount fraction of 300 nmol/mol.

To clarify the fact that different CO standards were used, we added a sentence in section 2.2: "*The CO amount fraction of the standards used for the determination of the water vapour interference ranged from 57 to 741 nmol/mol. No dependency of the water vapour interference on the CO amount fraction was observed.*"

-page 11, lines 4-9: you state that the clearly decreased bias between TI and PUY instrument was thanks to the drying of sample air for TI. However, you also changed the inlet line and never tested the other line with dryer. Please, mention this uncertainty more clearly in the text.

Thank you for this comment. It is correct that changing the inlet line causes additional uncertainty. However, there is evidence that the improvement was due to the drying, since parallel measurements were also made for CH_4 and CO_2 over the same inlet lines, and no change in the bias was observed for these two compounds with and without drying. In addition, we have seen similar results during parallel measurements at other stations of the GAW network. For example, a recent parallel measurement was made at the GAW station Bukit Koto Tabang. In this case, the WCC-Empa TI was equipped with a Nafion dryer for the entire comparison period, but the station analyser was measuring humid air for the first week. The bias was significantly larger during the period when the station analyser was measuring humid air, and decreased after the installation of a Nafion dryer in the same inlet line, as shown in Figure R3.
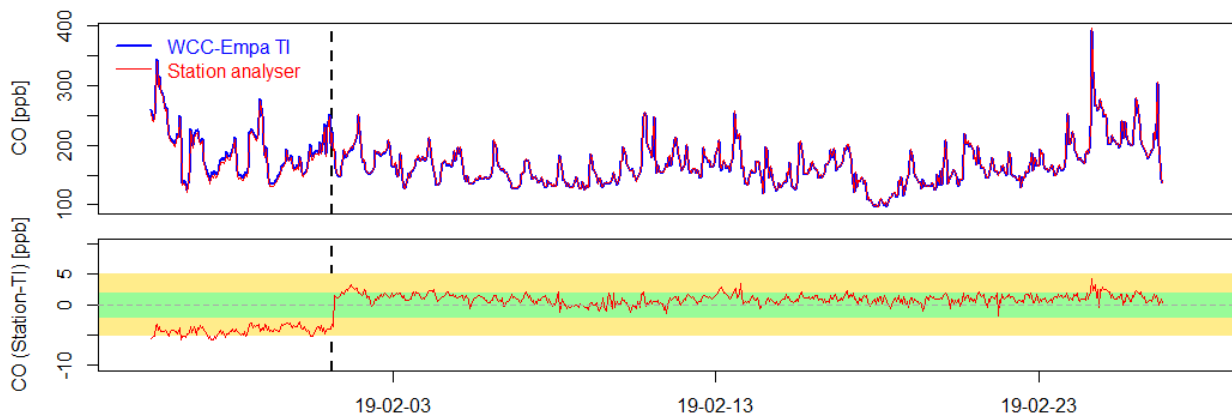
Figure R3. Comparison of the Bukit Koto Tabang Picarro G2401 analyser with the WCC-Empa travelling instrument for CO. Time series based on hourly data as well as the difference between the station instrument and the TI is shown. The coloured horizontal areas correspond to the WMO/GAW compatibility (green) and extended compatibility (yellow) goals. The dashed vertical line indicates the time of the installation of the Nafion dryer.

To clarify that the observed improvement is most likely due to the drying system but also to highlight the potential influence of the inlet system, we added the following sentences on page 11, line 8: "*Potentially, the change of the inlet system could also have been the reason for the reduction in the bias. However, this is unlikely because no change in the bias of CH_4 and CO_2 amount fraction, which were both measured simultaneously together with CO over the same inlet line, was observed.*"

-page 11, line 13: did you use the same individual instrument (TI) in PYU and in AMY? Differences between Picarro G2401 individuals, especially regarding CO performance, may be high.

Yes, the same instrument was used at PUY and AMY (Picarro G2401 CFKADS2098) (see Table 3, instrument serial number is given there).

-page 11, lines 16-19: since there was three clearly different periods; due the calibrations of AMY instrument; during the comparison measurements at AMY, please provide the biases for each period separately.

We added the bias for the individual periods. We also realised that the bias of 0.23±8.81 nmol/mol for the entire period given in the discussion paper was wrong, since it referred to a data set that included invalid data. We corrected to 0.10±3.20 nmol/mol.

-page 12, line 18: is the drying of the sample air the only option, as you state, or is it possible to reach the goal with well and frequently made water vapour tests as well?

Because sudden changes of the water vapour interference were observed, a correction based on frequent determination of the correction function might still be insufficient. Furthermore, determination of individual correction functions are associated with large uncertainties and drying therefore is recommended.

You also leave open the questions how to dry and how low water levels are needed to reach. By looking the Figure 10, it looks that water levels below 0.3 %, roughly, would give CO values stable enough when Picarro G2401 instrument is used. That kind of a water levels are possible to reach by using Nafion dryer, for example. The drying issue is an endless story, but open this theme with a few sentences.

Yes, this is correct. It only is important that the water vapour level remains as constant as possible. Calibration gas and working standards also need to pass through the Nafion dryer to compensate for potential loss in the system.

We added the following sentences at the end of section 3.2:

"*However, this will most likely be insufficient to detect the sudden changes in the correction function that were observed in our experiments.* Consequently, drying of the sample air should be considered when measuring CO with a Picarro G2401 instrument. *Both cryogenic traps and Nafion dryers can be used. WCC-Empa now uses Nafion dryers for the parallel measurements during station audits. Both single tube (MD-070-48S-4) and multi tube (PD-50T-12MPS) Nafion dryers in reflux mode using the Picarro pump for the vacuum in the purge air were successfully used. This reduced the amount of water to approximately 0.06 – 0.22% (single tube) and 0.01-0.03% (multi tube), depending on ambient air humidity. In case of using Nafion dryers, the standard gases must also pass though the dryer to compensate for a potential loss over the dryer.*"

And in the conclusions:

"*This can be implemented through drying of the sample air with cryogenic traps or Nafion dryers.*"

-figure 7: since the water vapour was probably the main reason for bad agreement between TI and PUY instrument, please add a panel showing the H2O % reading of TI.

The reason for the bias is the inappropriate compensation of the water vapour interference made by the Picarro instrument. There is a dependency of the bias on the water vapour level, which we showed in Figure 8. Figure R4 shows the original Figure 7 of the paper with an additional panel showing the H_2O reading of the instrument, as requested by the referee. However, we feel that this provides only minor additional information compared to the dependency on the water level shown in Figure 8. We therefore prefer to keep the figure as it is.
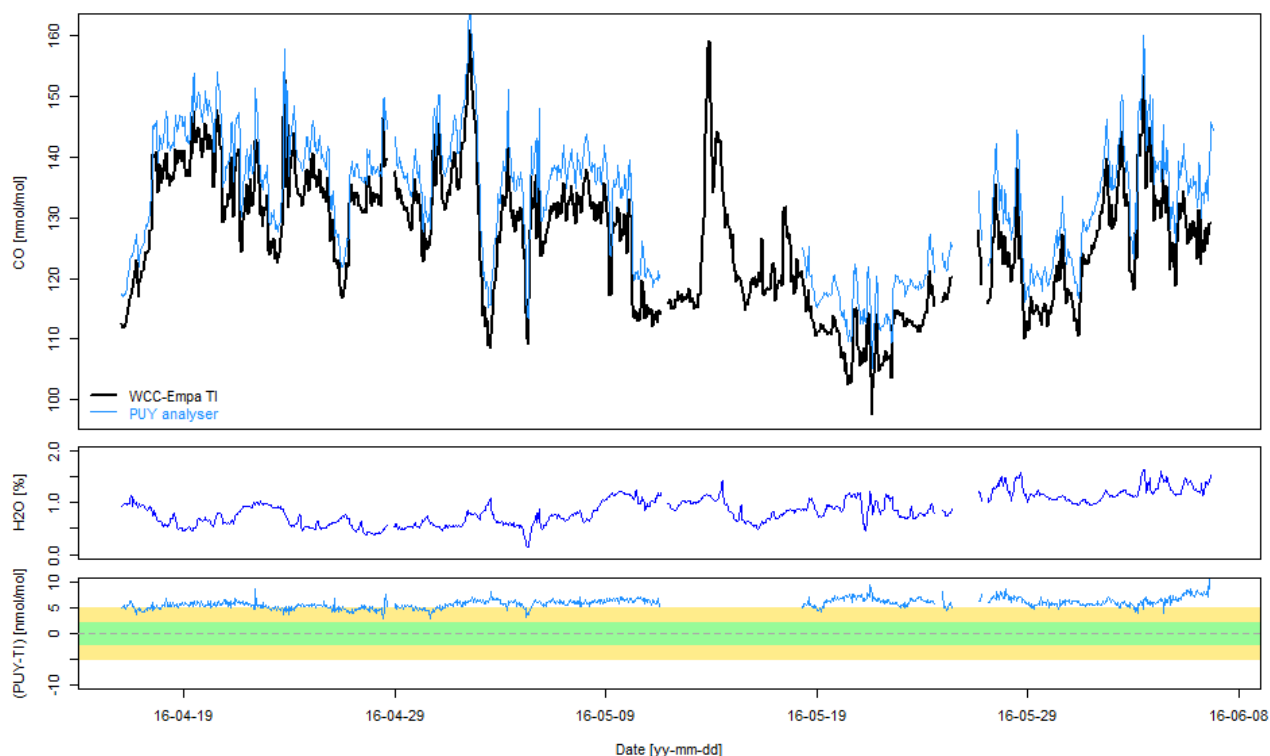

Figure R4. CO comparison at PUY between the WCC-Empa travelling instrument and the PUY Picarro G2401 for the period when the TI sampled humid air. Upper panel: CO time series (1 h data). Middle panel: H_2O measurements of the TI. Lower panel: CO bias of the station analyser vs time. The green and yellow areas correspond to the WMO compatibility and extended compatibility goals.

-figure 13: please, mark the calibrations of AMY instrument in lower panel with vertical lines, for example.

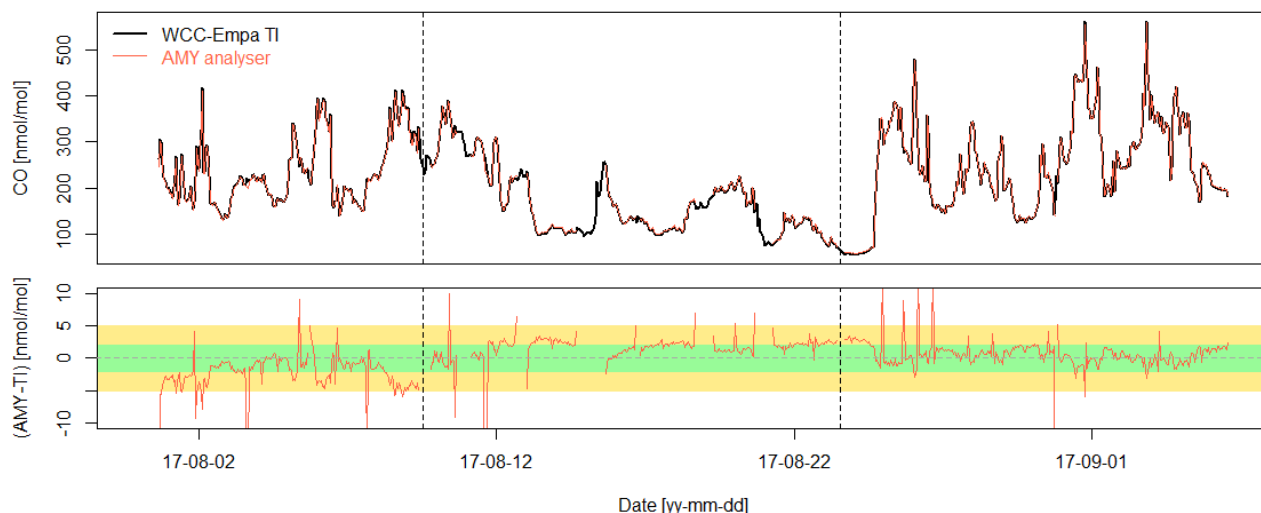We added vertical lines in the revised version of the paper.

Figure 1. CO comparison at AMY between the WCC-Empa travelling instrument and the AMY Los Gatos 30-EP QCL analyser. Both instruments sampled dry ambient air. Upper panel: CO time series (1 h data). Lower panel: CO bias of the station analyser vs time. The green and yellow areas correspond to the WMO compatibility and extended compatibility goals. The dashed vertical lines indicate the time of the calibration of the AMY instrument.

Technical corrections suggested by Referee #1:

-page 1, line 20: "analyse" > "analysed".

Accepted and changed.

-page 2, line 1: "GAW programme" is written with and without a capital "p" in this manuscript. Please uniform using lower case "p".

Accepted and changed throughout the paper.

-page 2, line 11: What is "Empa"? Please add institute full name and location.

Done.

-page 2, line 11: Use "WCC" instead of "World Calibration Centre". There is often used the full name instead of abbreviation, even though the abbreviation is introduced when first time mentioned. Like three lines below "World Calibration Centre" in written again, and with gas components and instrument techniques is the same issue. Uniform all these issues in the whole manuscript.

Done.

-page 2, line 14: WCC-N_2O, who operates? It is written on page 3, so move it to here.

Done.

-page 2, lines 28-30: introduction of the GC detectors for CO; now it is not fully clear that CG/FID with methaniser and CG/HgO are two separate instruments to measure CO. Please reword.

We reworded the sentence to "..., whereas flame ionization detection (GC/FID) in combination with a methaniser and GC with mercuric oxide reduction detector (GC/HgO) *were the two most commonly used techniques for CO measurements* (Zellweger et al., 2009)."

-page 4, lines 9-10: move the operator of CCL to where it is first time mentioned.

Done.

-page 5, line 6: add comma after "measurements".

Done.

-page 6, line 15: maybe "trend in the atmospheric concentrations" over "in the atmosphere".

Changed to "... due to the significant upward trend *of the N_2O mixing ratio* in the atmosphere..."

-page 7, lines 13-14: what is the meaning of "(see also Figure, right)"?

That was a mistake and we deleted it.

References

Brewer, P. J., Kim, J. S., Lee, S., Tarasova, O. A., Viallon, J., Flores, E., Wielgosz, R. I., Shimosaka, T., Assonov, S., Allison, C. E., van der Veen, A. M. H., Hall, B., Crotwell, A. M., Rhoderick, G. C., Hodges, J. T., Mohn, J., Zellweger, C., Moossen, H., Ebert, V., and Griffith, D. W. T.: Advances in reference materials and measurement techniques for greenhouse gas atmospheric observations, Metrologia, 56, 034006, 2019.

NOAA: WMO/IAEA Round Robin Comparison Experiment, https://www.esrl.noaa.gov/gmd/ccgg/wmorr/index.html, last access: 23 July, 2018.

WMO: 19th WMO/IAEA Meeting on Carbon Dioxide, Other Greenhouse Gases and Related Tracers Measurement Techniques (GGMT-2017), Dübendorf, Switzerland, 27-31 August 2017, GAW Report No. 242, World Meteorological Organization, Geneva, Switzerland, 2018.

Zellweger, C., Emmenegger, L., Firdaus, M., Hatakka, J., Heimann, M., Kozlova, E., Spain, T. G., Steinbacher, M., van der Schoot, M. V., and Buchmann, B.: Assessment of recent advances in measurement techniques for atmospheric carbon dioxide and methane observations, Atmos. Meas. Tech., 9, 4737-4757, 2016.

Zellweger, C., Hüglin, C., Klausen, J., Steinbacher, M., Vollmer, M., and Buchmann, B.: Inter-comparison of four different carbon monoxide measurement techniques and evaluation of the long-term carbon monoxide time series of Jungfraujoch, Atmos. Chem. Phys., 9, 3491-3503, 2009.

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.
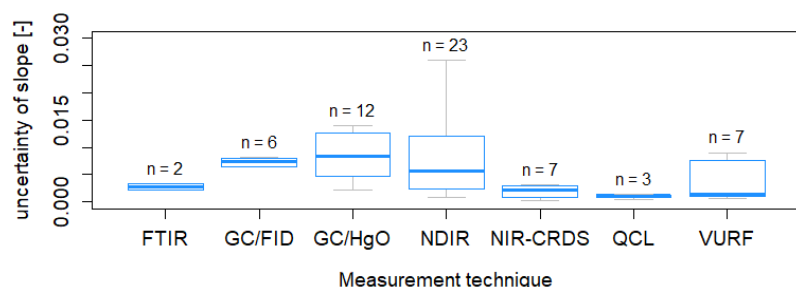

Figure 3. Boxplot of the slopes uncertainties of the of the regression analysis for the CO performance audits for different analytical techniques including the number of comparisons (n). The horizontal blue line denotes to the median, and the blue boxes show the inter-quartile range.

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.

Recent advances in measurement techniques for atmospheric carbon monoxide and nitrous oxide observations

Christoph Zellweger[1], Rainer Steinbrecher[2], Olivier Laurent[3], Haeyoung Lee[4], Sumin Kim[4], Lukas Emmenegger[1], Martin Steinbacher[1], Brigitte Buchmann[5]

[1] Empa, Swiss Federal Laboratories for Materials Science and Technology, Laboratory for Air Pollution / Environmental Technology, CH-8600 Dübendorf, Switzerland.

[2] Karlsruhe Institute of Technology (KIT), Institute of Meteorology and Climate Research (IMK-IFU), Garmisch-Partenkirchen, Germany.

[3] Laboratoire des Sciences du Climat et de l'Environnement (LSCE/IPSL), UMR CEA-CNRS-UVSQ, Gif-sur-Yvette, France

[4] National Institute of Meteorological Sciences (NIMS), Seogwipo-si, Jeju-do, Korea

[5] Empa, Swiss Federal Laboratories for Materials Science and Technology, Department Mobility, Energy and Environment, CH-8600 Dübendorf, Switzerland.

Correspondence to: C. Zellweger (christoph.zellweger@empa.ch)

Abstract. Carbon monoxide (CO) and nitrous oxide (N_2O) are two key parameters in the observation of the atmosphere, relevant for air quality and climate change, respectively. For CO, various analytical techniques have been in use over the last few decades. In contrast, N_2O was mainly measured using gas chromatography (GC) with electron capture detector (ECD). In recent years, new spectroscopic methods have become available which are suitable for both CO and N_2O. These include Infra-Red (IR) spectroscopic techniques such as Cavity Ring Down Spectroscopy (CRDS), Off-Axis Integrated Cavity Output Spectroscopy (OA-ICOS) and Fourier Transform Infrared Spectroscopy (FTIR). Corresponding instruments became recently commercially available and are increasingly used at atmospheric monitoring stations. We analysed results obtained through performance audits conducted within the framework of the Global Atmosphere Watch (GAW) quality management system of the World Meteorology Organisation (WMO). These results reveal that current spectroscopic measurement techniques have clear advantages with respect to data quality objectives compared to more traditional methods for measuring CO and N_2O. Further, they allow a smooth continuation of historic CO and N_2O time series. However, special care is required concerning potential water vapour interference on the CO amount fraction reported by Near-IR CRDS instruments. This is reflected in results of parallel measurement campaigns, which clearly indicate that drying of the sample air ~~is leading~~leads to an improved accuracy of CO measurements with such Near-IR CRDS instruments.

1   Introduction

The Global Atmosphere Watch (GAW) Programme of the World Meteorological Organization (WMO) coordinates a network of atmospheric composition observations comprising 31 global stations, more than 400 regional stations, and around 100 contributing stations operated by contributing networks (GAWSIS, 2018). These stations provide long-term observations of atmospheric greenhouse gases (GHGs) and reactive gases such as carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O) and carbon monoxide (CO), which are essential for understanding the GHG budget, both regionally and globally. To make full use~~To take full profit~~ of these observations, the uncertainty of these measurements must be minimised which implies consistent data series with traceability to common reference standards. Within the GAW ~~Programme~~programme, Central Calibration Laboratories (CCLs) provide reference standards that are linked to internationally accepted calibration scales (Rhoderick et al., 2016; Rhoderick et al., 2018). In addition, World Calibration Centres (WCCs) evaluate GAW stations through independent assessments by on-site system- and performance audits (Buchmann et al., 2009). The Laboratory for Air Pollution / Environmental Technology of the Swiss Federal Laboratories for Materials Science and Technology (Empa) ~~Empa~~ operates the ~~World Calibration Centre~~WCC for Carbon Monoxide (CO), Methane (CH_4), Carbon Dioxide (CO_2) and Surface Ozone (WCC-Empa) since 1996 as a Swiss contribution to the GAW ~~Programme~~ programme and has conducted over 90 system- and performance audits over the past 20 years. Furthermore, WCC-Empa collaborates closely with the ~~World Calibration Centre~~ WCC for nitrous oxide (WCC-N_2O) hosted by the Karlsruhe Institute of Technology (KIT), Institute of Meteorology and Climate Research (IMK-IFU) to increase the number of N_2O audits. In order to address scientific needs for interpreting regional or global scale atmospheric observations, the GAW ~~Programme~~ programme sets ambitious network compatibility goals, which are continuously reviewed and, if necessary, revised during biannual meetings of the WMO/GAW community (WMO, 2018). Network compatibility goals are set for amount fraction ranges observed in the unpolluted troposphere, while extended network compatibility goals reflect the less stringent requirements for urban and regional studies with larger local fluxes. The network compatibility goals currently stand at $\pm$2 nmol/mol for CO and $\pm$0.1 nmol/mol for N_2O, whilst the extended goals are set to $\pm$5 nmol/mol for CO and $\pm$0.3 nmol/mol for N_2O. These goals represent the maximum bias that can generally be tolerated in measurements of well-mixed background air used in global models to infer regional fluxes. Some network compatibility goals may not be currently achievable within current measurement and/or scale transfer uncertainties. However, they are targeted for applications which require the smallest possible bias among different datasets or data providers, such as for the detection of small trends and gradients (WMO, 2018).

In-situ measurements of tropospheric CO and N_2O have been available since the late 1960s (Weiss, 1981; Khalil and Rasmussen, 1983, 1988). While early measurements were mainly analysis results based on flask samples, quasi-continuous measurements have been available since the early 1980s (Brunke et al., 1990). Although continuous measurements of CO and N_2O began approximately at the same time and were often collocated, challenges with respect to the measurement techniques for continuous measurements were completely different. Carbon monoxide shows ~~a~~ high temporal and spatial variability, whilst the detection of very small changes is needed for N_2O observations. In the past, atmospheric CO and N_2O measurements

at remote locations were almost exclusively made by gas chromatographic techniques (GC). GC with an electron capture detector (GC/ECD) was by far the most abundant measurement technique for N_2O, whereas flame ionization detection (GC/FID) in combination with a methaniser and GC with mercuric oxide reduction detector (GC/HgO) was were the two most commonly used techniques for CO measurements (Zellweger et al., 2009).

5 Recent years brought the rapid development of a variety of alternative CO measurement techniques, and a variety of methods is are now in use at atmospheric monitoring sites. Common methods include GC techniques (Gros et al., 1999; Novelli, 1999; van der Laan et al., 2009), non-dispersive infrared absorption (NDIR) (Parrish et al., 1994; Nedelec et al., 2003), vacuum ultra-violet resonance fluorescence (VURF) (Gerbig et al., 1999), Fourier Transform Infrared (FTIR) absorption (Griffith et al., 2012; Hammer et al., 2013a), Near-IR-Cavity Ring Down Spectroscopy (NIR-CRDS) (Chen et al., 2013; Yver Kwok et al.,

10 2015), and systems using Quantum Cascade Lasers (QCL) in the mid-infrared such as Mid-IR CRDS, Off-Axis Integrated Cavity Output Spectroscopy cavity enhanced off-axis Integrated Cavity Output Spectroscopy (OA-ICOS) (Baer et al., 2002; Provencal et al., 2005) and quantum cascade tuneable infrared laser direct absorption spectroscopy (QC-TILDAS) (McManus et al., 2015).

Alternatives to GC-ECD for N_2O are not as abundant, but several methods have been proposed in recent years. These include

15 instruments deploying optical techniques in the mid-IR, e.g. cavity ring down (CRDS) spectroscopy, FTIR, OA-ICOS, QC-TILDAS and difference frequency generation (DFG)-based systems. Lebegue et al. (2016) published a comprehensive overview of these techniques as well as their performance under controlled conditions.

The recently developed recent optical techniques for CO and N_2O have clear advantages concerning sensitivity, repeatability, linearity, time response, and temporal coverage, resulting in new measurement setups and calibration strategies. However,

20 only a few published studies comparing spectroscopic techniques with GC systems exist for CO (Zellweger et al., 2009; Zellweger et al., 2012; Ventrillard et al., 2017) and N_2O (Vardag et al., 2014; Lebegue et al., 2016). Such comparisons of traditional and upcoming new techniques are crucial for a smooth continuation of multi-decadal time series when introducing new analytical techniques.

In this paper, we analyse data collected during CO and N_2O performance audits made by WCC-Empa and WCC-N_2O from

25 2002 through 2017 from the perspective of the used measurement techniques. We further present ambient air CO comparisons made with a NIR-CRDS travelling instrument during WCC-Empa audits and show limitations of the NIR-CRDS technique with respect to water vapour interference. Assessment of atmospheric measurements through parallel measurements with a travelling instrument is complementary to performance audits with travelling standards and round robin experiments and is thus an essential, valuable quality control measure (Hammer et al., 2013b; Zellweger et al., 2016).

30 **2 Experimental**

System and performance audits (hereafter only called audit) by WCCs are part of the quality management framework of the GAW Programme programme (WMO, 2017a). WCC Empa is the designated WCC for CO (since 1997), and since 2009 a

3

collaboration between WCC-Empa and the WCC for N$_2$O ~~hosted by the Karlsruhe Institute of Technology (KIT), Institute of Meteorology and Climate Research (IMK-IFU), exists since~~ has allowed WCC-Empa ~~started including~~to include N$_2$O comparisons during station audits. The concept of station audits has been described elsewhere (Klausen et al., 2003; Buchmann et al., 2009; Zellweger et al., 2016). WCCs use two different approaches to conduct performance audits: (i) comparisons of travelling standards (TS), i.e. high-pressure cylinders with known nominal values of CO and N$_2$O amount fractions, and (ii) parallel measurements using a travelling instrument (TI). The TS method is widely applied, while the TI concept is used less frequently and limited by WCC-Empa to CO, CO$_2$ and CH$_4$.

2.1 Comparisons using travelling standards

The audit concept using TS supplies gases from high-pressure cylinders, usually dry natural air or synthetic air, on the instruments of the audited station. Usually, multiple analysis of a set of three or more TS is made and averaged for the final assignment of the TS value by the audited laboratory. Calibrations of TS against reference standards before and after the station audit ensure traceability to the CCL, which is run by the National Oceanic and Atmospheric Administration / Earth System Research Laboratory (NOAA/ESRL). The results are then analysed by linear regression of the values measured by the station vs. the reference values assigned by the WCC. At WCC-Empa, N$_2$O and CO amount fractions in the TS are calibrated since 2010 by an Aerodyne Quantum Cascade Laser spectrometer (QC-TILDAS-CS, Aerodyne Research Inc., MA, USA). Before that, an AL5001 Vacuum UV Resonance Fluorescence analyser (VURF) (AL5001, Aerolaser GmbH, Germany) was used for CO calibrations. Both instruments are described in more detail in Zellweger et al. (2012). Amount fractions are assigned to the TS using a set of several reference standards purchased from the CCL~~, which is run by the National Oceanic and Atmospheric Administration / Earth System Research Laboratory (NOAA/ESRL)~~. WCC-N$_2$O uses a set of TS traceable to a set of ~~secondary~~ tertiary standards, which are regularly recalibrated against secondary standards at the CCL ~~with direct amount fraction assignments by the CCL. Only comparisons involving instruments calibrated to the same calibration scale used by the WCCs are presented hereafter.~~ For N$_2$O, the calibration scales in use were WMO-X2000~~SIO-98 (Prinn et al., 2000)~~ for audits before ~~2005~~2006, and WMO-X2006 and X2006A (Hall et al., 2007; NOAA, 2018c) afterwards. CO refers to the WMO-X2000, WMO-X2004, X2014 and X2014A (NOAA, 2018a) calibration scales.

We analysed WCC-Empa performance audit results based on the TS method for carbon monoxide (2005-2017) and nitrous oxide (2009-2017), as well as results of N$_2$O audits conducted by WCC-N$_2$O (2002-2013). Details on analytical techniques, ~~and~~ instruments and calibration scales of these audits are summarised in Table 1 for CO and Table 2 for N$_2$O. 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. CO audits made by WCC-Empa before 2005 were not considered for the comparison due to the following reasons: (i) stations and WCC-Empa were often not referring to the same CO calibration scale. WCC-Empa was using the WMO-X2000 carbon monoxide scale, while many GAW stations were still reporting on the older WMO-X88 scale (Novelli et al., 2003) or

other scales. (ii) WCC-Empa at that time based its calibration of travelling standards only on CO standards above 185 nmol/mol; the WMO-X2000 calibration scale had linearity issues, which have been corrected by the succeeding WMO-X2004, X2014 and X2014A calibration scales. 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-

5   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. For CO, the assessment has been made in the same standardised way as for carbon dioxide (CO_2) and methane (CH_4) described in Zellweger et al. (2016), while a slightly different approach has been chosen for N_2O due to the fact that ambient air amount fractions increased significantly during the period of observation. The results section gives further details on the methodology.

10 ## 2.2   Ambient air comparisons

Assessments based on TS comparison, e.g. during station audits or round robin experiments, have limitations. They only cover the analytical system and exclude other aspects that might also be relevant, such as inlet or drying systems. The low water content of the TS may for example lead to a systematic bias, especially for analysers based on spectroscopic techniques with implemented water vapour correction algorithms. The assessment during on-site audits should therefore include parallel

15 measurements with a TI whenever feasible (WMO, 2011, 2012, 2014, 2016, 2018).

WCC-Empa implemented this additional approach for CO, CO_2 and CH_4 audits in 2011. Details of the set-up and procedure as well as results for CO_2 and CH_4 are published in Zellweger et al. (2016). Audits involving parallel measurements for CO were conducted using a NIR-CRDS analyser (G2401, Picarro Inc., USA) as a travelling instrument. The Picarro G2401 instrument has an internal water vapour correction for CO and reports dry air amount fraction only. However, these factory

20 based corrections are often not adequate (Chen et al., 2013). Due to the higher analytical noise compared to CO_2 and CH_4 measurements, corrections require a more comprehensive approach (Rella et al., 2013).

The internal water vapour correction of the TI was evaluated using the water droplet method (Zellweger et al., 2012; Rella et al., 2013). Approximately 0.8 ml of ultra-pure water is injected into a constant flow of about 500 ml min^{-1} of a dry working standard and delivered to the instrument using a bypass overflow. The CO amount fraction of the standards used for the

25 determination of the water vapour interference ranged from 57 to 741 nmol/mol. No dependency of the water vapour interference on the CO amount fraction was observed. For the WCC-Empa CO analyser, the water vapour influence on the CO amount fraction, which is already corrected by the internal water vapour compensation of the Picarro instrument, was then fitted by a quadratic function. Due to the relatively large uncertainties of individual experiments, we were not able to determine a reliable correction function and, therefore, relied on the factory settings for our experiments.

30 Parallel measurements with the TI of the following GAW stations are shown in this paper:

      (i)   Puy de Dôme (PUY), France, a global GAW station that is part of the European Integrated Carbon Observation System (ICOS). A separate inlet system leading to the same location as the air intake of the station analyser was in place for the comparison with the TI. An additional pump at a flow rate of approximately 2 L/min flushed this

WCC-Empa inlet line. For the last days of the comparison, the TI sampled from the station inlet using the same cryogenic dryer as the station instrument. During this period, the air was dried to a dew point of approximately -50°C.

(ii) Anmyeon-do (AMY), South Korea, a regional GAW station run and managed by the Environmental Meteorology Research Division of the National Institute of Meteorological Sciences (NIMS). Air was taken with both instruments from the AMY air inlet system, and the air was dried to a dew point of approximately -50°C using a cryogenic trap.

Table 3 gives an overview of the comparisons, including duration and instruments used. Detailed information about the stations is available from the GAW Station Information System (GAWSIS, 2018).

3 Results

3.1 Analysis of travelling standard comparison

One of the objectives of this work was to evaluate the performance of instruments for measuring CO and N_2O at remote atmospheric research observatories. Of particular interest is the question if modern spectroscopic techniques such as NIR-CRDS, TILDAS, OA-ICOS or FTIR have a significant advantage compared to traditional methods, and whether spectroscopic techniques improve the results of the performance audits carried out by the WCCs for the corresponding compounds with respect to precision and uncertainty. WCC-Empa made sixty comparisons during station audits using travelling standards for CO (2005-2017), and twenty for N_2O (2009-2017). In addition, WCC-N_2O conducted sixteen comparisons during station audits (2002-2013). Table 1 and 2 show details of analytical techniques and instruments of these comparisons for CO and N_2O, respectively. The three letter codes (GAW ID) refers to the different stations (GAWSIS, 2018). Results of audits at the central calibration facility run by the Centre for Atmosphere Watch & Services (CAWAS) and of the greenhouse gas analysis on-board the Civil Aircraft for the Regular Investigation of the atmosphere Based on an Instrument Container (CARIBIC) are also included in the comparisons.

Each of the audits shown in Table 1 and 2 involved the comparison of a set of travelling standards and was then evaluated by linear regression analysis of the measured values by the stations vs. the WCC assigned amount fractions, which are traceable to the CCL. To judge whether the combinations of the resulting slope and intercept meet the WMO/GAW network compatibility, respectively or extended network compatibility goals, the method described in Zellweger et al. (2016) was applied in analogy. For CO, the bias at 165 nmol/mol, which is the centre of the amount fraction range of 30-300 nmol/mol representing the unpolluted troposphere (WMO, 2018) was plotted against the slope of the individual travelling standard comparisons. This amount fraction range sufficiently covers the inter-hemispheric gradient, year-to-year variability, seasonal cycles as well as observed trends for the period of consideration at remote stations. For N_2O, using a fixed amount fraction range however might not be appropriate due to the significant upward trend of the N_2O mixing ratio in the atmosphere over the past decades. The range currently representing the unpolluted troposphere has been recently identified as 325-335 nmol/mol

6

(WMO, 2018), which corresponds well to the mean global atmospheric N_2O amount fraction of 328.9 ± 0.1 nmol/mol observed in 2016 (WMO, 2017b). A trend analysis made by Blunden and Arndt (2017) showed an annual increase of about 0.8 nmol/mol per year over the last decade, which is in agreement with a fairly constant annual growth rate of 0.81 nmol/mol per year from 1977 until today determined by the National Oceanic and Atmospheric Administration (NOAA, 2018b). Based on this, our

5   analysis of N_2O audit results was made using a variable amount fraction range covering 10 nmol/mol with the centre being representative for the unpolluted troposphere for the year of the audit. Table 2 gives the corresponding ranges used for the analysis. This method allows displaying the result of each individual CO and N_2O audit involving comparisons with travelling standards as a single dot in a bias vs. slope plot, similar to CO_2 and CH_4 results presented by Zellweger et al. (2016).

3.1.1   Evaluation of CO comparisons

10   Figure 1 shows the bias in the centre of the relevant amount fraction of the unpolluted troposphere of 30 – 300 nmol/mol CO vs. the slope for the CO audits listed in Table 1. Perfect agreement would result in bias / slope pairs of (0 nmol/mol / 1). The allowed bias / slope combinations meeting the network compatibility (green area) and extended network compatibility goals (yellow area) of ±2 nmol/mol and ±5 nmol/mol (WMO, 2018), respectively, are indicated. The distribution of the observed biases and slopes gives further information about potential systematic offsets, which could be present either at the WCC or at

15   the stations. If results are not systematically biased (e.g. by different calibration scales), a normal distribution of the observed bias and slope pairs around 0 nmol/mol (bias) and 1.0 (slope) is expected. This was the case at the 95% confidence level for the slope, which deviated with a mean value of 0.994 ± 0.068 (1σ) not significantly different from one (t-Test, $p = 0.47$). However, the mean bias of -2.6 ± 8.7 nmol/mol (1σ) was significantly different from zero ($p = 0.02$). A potential reason could be upward drift in standards, which is common for CO in air mixtures at ambient amount fractions (Novelli et al., 2003;

20   Gomez-Pelaez et al., 2013). Drift rates are usually on the order of up to one nmol/mol per year. To account for this, WCC-Empa frequently retrieves reference standards from the CCL. This might not always be the case at measurement sites. There, standards are often in use over long periods without re-calibration or acquisition of new standards. The use of standards having increased amount fractions due to drift for instrument calibration will then results in an underestimation of ambient CO, which potentially explains the observed mean bias.

25   Figure 1 shows that reaching the network compatibility goals for CO is extremely challenging. The variety of measurement techniques is quite large with clear performance differences between methods. Newer spectroscopic techniques such as QCL based TILDAS or OA-ICOS spectroscopy (QCL hereafter), or CRDS generally show better performance compared to GC methods or NDIR. Moreover, they also yield higher data coverage due to the truly continuous observations in contrast to the semi-continuous GC measurements and the less frequently required application of reference gases compared to NDIR

30   measurements. Higher data coverage further reduces the uncertainty caused by incomplete sampling. Figure 2 summarises the percentage of comparisons that met the network compatibility and extended network compatibility goals for (a) all comparisons (see also Figure, right), (b) for GC/HgO and GC/FID systems only, (c) NDIR instruments only, (d) VURF instruments only, and (e) for NIR-CRDS and QCL instruments. FTIR is not shown separately, since only two comparisons of one instrument

were made. Out of the sixty comparisons, only thirteen (21.7%) met the network compatibility goal and an additional fourteen (23.3%) met the extended goal in the amount fraction range relevant for the troposphere. Good performance over the entire relevant amount fraction range is required, since atmospheric CO variability is large and pollution episodes, e.g. through long-range transport, are common even at remote locations. Calibration strategies therefore should cover the entire range, which is easier to implement for techniques with a linear response such as VURF, NIR-CRDS and QCL. The analysis of the performance audit results shows that 90% of the NIR-CRDS and QCL comparisons were meeting the network compatibility or extended network compatibility goal, while this was the case for less than 40% of the NDIR analysers or GC systems. From the total of ten travelling standard – NIR-CRDS/QCL comparisons, five (50%) were within ±2 nmol/mol, and additional four (40%) within ±5 nmol/mol. The corresponding numbers are significantly smaller for GC based methods (total 18 comparisons) and NDIR (total 23 comparisons), which clearly indicates an advantage of the recent methods compared to more traditional techniques.

However, these results also depend on calibration and potential issues or differences in the calibration scales. For example, an instrument with perfect repeatability and reproducibility but wrong calibration, e.g. by a bias in the calibration standard, can be outside the quality goals only because of calibration issues. In this case, the uncertainty of the linear regression of the travelling standard comparison is expected to be smaller compared to instruments with poorer repeatability and reproducibility. Therefore, the uncertainty of the linear regression analysis is another measure of the instrument performance. Figure 3 shows a boxplot of the standard uncertainty of the slopes of all CO performance audits grouped by different analytical techniques. The results also confirm the better performance of the QCL and NIR-CRDS instruments compared to GC techniques and NDIR. Interestingly, the performance of NDIR analysers and GC/HgO systems is similar but likely due to different reasons. While the repeatability of GC/HgO systems is generally superior compared to NDIR, appropriate compensation of the non-linearity remains obviously difficult compared to the normally linear but noisy NDIR analysers, resulting in similar performance of both techniques in the field for the amount fraction range from 30 to 300 nmol/mol.

Comparison with the the recent WMO/IAEA Round Robin Comparison Experiment, as done for N_2O (see below), is not straightforward. Changes in the calibration scale during the round robin experiment jeopardizes the direct comparison of the audit results with the round robin results.

3.1.2   Evaluation of N_2O comparisons

Figure 4 shows the bias in the centre of the relevant amount fraction range of the year of the comparison vs. the slope for the N_2O audits shown in Table 2 along with the allowed bias / slope combinations meeting the network compatibility (green area) and extended network compatibility goals (yellow area) of 0.1 nmol/mol and 0.3 nmol/mol (WMO, 2018), respectively. Only results of comparisons made on the same calibration scale andwith fully functional instruments were considered.

The results presented in Figure 4 show that reaching the WMO/GAW network compatibility goals remains difficult for N_2O. However, calibration ranges at stations can be intentionally limited to the ambient amount fraction typical for their location and time. These ranges are normally significantly smaller than those used in Figure 4 in the case of N_2O. Therefore, bias /

slope pairs outside the network compatibility goals do not necessarily imply that the measurements at a station are biased, but they are indicative of the performance of the instrument and its calibration over a given amount fraction range. The dashed green and yellow lines in Figure 4 denote the limits for meeting the network compatibility and extended network compatibility goals at the relevant amount fraction.

5   As discussed above for CO, the distribution of the observed biases and slopes is an indicator of potential systematic offsets, either at the WCCs or at the stations. No significant deviations at the 95% confidence level were observed for audits carried out by WCC-Empa, with a mean bias of 0.32 ± 1.09 nmol/mol (1σ), t-Test p value of 0.11, and a mean slope of 0.965 ± 0.093 (p = 0.21). WCC-N_2O comparisons also showed also no significant bias (-0.12 ± 0.89 nmol/mol, p = 0.35) but the deviation of the slope was significant (0.954 ± 0.067, p = 0.01). This result indicates that at the launch of the audits in 2002 the linearity problem

10 of the ECD was not fully considered in the data evaluation by the audited stations. The GC/ECD technique, which contributes most to the results, is known to be highly non-linear (Lebegue et al., 2016), and consequently, deviations are expected for amount fractions away from the relevant level if the non-linearity of the systems had not been determined accurately enough. With ongoing data quality assurance activities and the implementation of linearity corrections for the ECD response the slope now is close to one for more recent performance audits.

15 Fig. 5 presents the result of the above analysis as percentages of comparisons meeting the network compatibility and extended network compatibility goals. Until now, none of the performance audits conducted by either WCC-N_2O or WCC-Empa achieved the compatibility goal of 0.1 nmol/mol, and only one third of the results were within the extended goals of 0.3 nmol/mol when an amount fraction range of 10 nmol/mol is considered. This slightly improves if we consider only the bias at the relevant amount fraction. The relevant amount fraction corresponds to the value at the centre of the relevant range for the

20 corresponding year. Under these less stringent conditions, we find 19.4 % compliance with the network compatibility goal and 36.1% with the extended network compatibility goal. This is in line with the above mentioned small variations in N_2O at remote locations and the corresponding limited calibration range of many stations. Lebegue et al. (2016) recognised that measurements of small variations in the N_2O amount fractions using GC/ECD is very challenging, which is in agreement with the TS comparison results from the station audits of this work.

25 The results obtained during the performance audits by WCC-Empa and WCC-N_2O compare well with the recent WMO/IAEA Round Robin Comparison Experiment organised and coordinated by the CCL for N_2O hosted by NOAA. The sixth round robin experiment took place in 2014/15, and involved the comparison of two standards, one containing a lower (average 321.6 nmol/mol) and the other a higher (average 333.7 nmol/mol) N_2O amount fraction (NOAA, 2018d). A total of 25 laboratories participated in this exercise. With this data set, we made the same analysis as described above after the exclusion of two

30 laboratories using other calibration scales other than WMO-X2006A. The percentage of laboratories fulfilling the WMO network compatibility and extended network compatibility goal was very similar to the results from the station audits by WCC-Empa and WCC-N_2O, as shown in Figure 6.

Out of the 25 laboratories in the Round Robin Experiment, only two (8%) were entirely within the WMO/GAW network compatibility goal of 0.1 nmol/mol for the 10 nmol/mol range. At the relevant amount fraction, the percentage of laboratories that were not meeting the quality goals was very similar for the WCC audits (44%) and the round robin experiment (40%).

The above results, both for TS comparisons during audits and the round robin experiment, are clearly illustrating that it remains highly challenging to reach the network compatibility and extended network compatibility goals for N_2O. In contrast to advances made for the detection of CH_4, CO_2 (Zellweger et al., 2016) and CO, measurements of N_2O were in most cases still made based on gas chromatography, and only a few recent comparisons involved spectroscopic techniques. The data for N_2O clearly indicates advantages of the spectroscopic techniques compared to gas chromatography. The uncertainty of the observed intercepts and slopes of the linear regression gives information of on the linearity and repeatability of the system. The uncertainty of the slope of the linear regression was significantly smaller for QCL and FTIR analysers (median 0.0028, standard deviation 0.0031) compared to GC/ECD systems (median 0.0126, standard deviation 0.0284). Despite the better performance regarding linearity and repeatability of the spectroscopic techniques compared to GC/ECD, no clear advantage of the spectroscopic methods was observed during the performance audits. A potential reason could be the uncertainty of the calibration standards, which is in the case of N_2O in of the same order or even larger than the WMO/GAW network compatibility goal. The CCL determined a reproducibility of N_2O calibrations in the ambient range of ~0.22 nmol/mol (95% confidence level) (Hall et al., 2007; NOAA, 2018c), which is larger than the network compatibility goal. However, this uncertainty is low compared to uncertainties associated with gravimetric preparation of standards, which highlights the importance of maintaining and propagating calibration scales (Brewer et al., 2018) as implemented in the WMO/GAW programme. Therefore, it is yet too early to quantify this improved performance of spectroscopic techniques for N_2O and give a final statement with respect to the network compatibility goals.

10

3.2   Ambient air comparisons

The above results, as well as round robin experiments, are travelling standard comparisons and are therefore not covering all aspects of ambient air measurements. Other aspects include bias due to sampling procedures, drying or – related to it – insufficient accounting of spectral interferences, e.g. by water vapour. For example, Chen et al. (2013) demonstrated that
5   accurate measurements of CO in humid air is possible with the NIR-CRDS technique implemented by Picarro. Correction functions however are different for each individual instrument, and as a result of the work of Chen et al. (2013), these functions are now implemented in Picarro NIR-CRDS CO analysers after 2012.

WCC-Empa started with parallel measurements of ambient air for CO, CO_2 and CH_4 during station audits in 2011. The results of the greenhouse gas comparisons showed that additional information, e.g. related to air inlet systems, is obtained by these
10   comparisons (Zellweger et al., 2016). However, these comparisons were in many cases less conclusive for CO. Some parallel measurements showed differences that were not present in the travelling standard comparisons. Sampling issues were unlikely because the ambient air comparison of CH_4 and CO_2 agreed well. Therefore, other issues like interferences of ambient air constituents may cause an additional bias.

For example, the comparison made at the global GAW station Puy de Dôme (PUY) in 2016 showed significant deviations in
15   ambient CO measurements, as illustrated in Figure 7, while the TS comparison showed good agreement. During this period, the PUY analyser ~~TI~~ was measuring on average 5.85±0.94 nmol/mol (1σ) ~~lower~~ higher than the ~~TI~~ PUY analyser. Despite this bias, both instruments captured the temporal variation well. The WCC-Empa travelling instrument was sampling from the same air intake location but with a completely independent sampling line. In contrast to the PUY instrument, which sampled air dried to a dew point of -50°C, the air sampled by the travelling instrument was not dried. As discussed in the section 2.2,
20   the factory water vapour correction was used. The observed bias correlates with the measured water vapour, as shown in Figure 8, which indicates issues with the internal water vapour compensation of the TI. Water vapour correction functions of this instrument were determined three weeks before and three weeks after the comparison campaign with a droplet test, in analogy to the method described by Rella et al. (2013). Figure 9 shows the ratio of CO(humid, corrected) / CO(dry) against the measured water vapour content of the TI; CO(dry) is the amount fraction measured by the instrument in the absence of water, and
25   CO(humid, corrected) the water vapour corrected CO amount fraction reported by the Picarro G2401 during the humidification by the droplet test. Since the Picarro G2401 reports CO only as dry air amount fraction, the measured ratio should be equal to one and not depend on water vapour content. However, a significant change in the CO response in relation to water vapour was observed. The TI was underestimating the CO amount fraction in the experiment before the campaign (Figure 9a), and then changed to an overestimation after the campaign (Figure 9b). Possibly, this has been influenced by the upgrade to a new
30   software version of the TI between the two periods. Unlike for CO_2 and CH_4, individual water vapour correction functions for CO can currently not be determined with sufficient accuracy to achieve the WMO/GAW network compatibility goal of 2 nmol/mol. Individual experiments using the droplet test have a large uncertainty due to higher instrumental noise for CO compared to CH_4 or CO_2. Furthermore, CO correction functions seem to be less stable over time, and sudden changes are

possible. Figure 10 shows fitted ratios of CO(humid, corrected) / CO(dry) vs. the measured water vapour content for two different instruments over a period of several years. Both instruments show significant variation over time in the humidity corrected CO reported by the analyser. Consequently, drying of the sample air could improve CO measurements with Picarro G2401 instruments, and likely with Picarro G1302 and G2302 CO/CO_2/H_2O analysers. This has been confirmed by a period

5   of dry ambient air measurements of both instruments at PUY. Figure 11 shows the comparison of the two analysers during the audit collocation measurement. In this case, the TI was connected to the same sampling line as the PUY instrument after the cryogenic trap, and both instruments were measuring dry air. The bias of the ~~TI~~ PUY analyser significantly decreased to - 1.20±0.57 nmol/mol (1σ). This agrees well with the observed bias during the travelling standard comparison. Potentially, the change of the inlet system could also have been the reason for the reduction in the bias. However, this is unlikely because no

10   change in the bias of CH_4 and CO_2 amount fraction, which were both measured simultaneously together with CO over the same inlet line, was observed. Figure 12 is summarising the results of the performance audits at PUY with TS, as well as the bias observed during the comparison campaign with humid and dry measurement of the TI.

Figure 13 shows another example of a CO ambient air comparison made at the regional GAW station Anmyeon-do, South Korea, over a period of one month in 2017. The comparison was made between the AMY ~~cavity enhanced off-axis Integrated~~

15   ~~Cavity Output Spectroscopy analyser~~OA-ICOS analyser (LGR-30-EP, Los Gatos Research, USA) and the WCC-Empa Picarro G2401 travelling instrument. Both analysers were measuring ambient air dried to a dew point of -50°C using a cryogenic trap. Temporal variability at this site is significantly larger compared to PUY, and except for a few spikes, it was well captured by both instruments. The bias of the AMY analyser averaged to 0.~~23~~0.10±~~8.81~~3.20 nmol/mol (1σ) over the entire period of the campaign. However, during the first third of the campaign, the AMY instrument was slightly underestimating the CO amount

20   fraction compared to WCC-Empa (bias -2.28 ±2.91 nmol/mol), followed by a slight overestimation in the second third (bias 1.47±2.81 nmol/mol). The last third then showed good agreement between the two systems (bias 0.57±2.80 nmol/mol). These differences are likely due to different calibration strategies. The TI was measuring three standard gases to calibrate and compensate for drift of the instrument every 30 hours. In contrast, manual calibrations were made of the AMY analyser every 14 days with one calibration standard (dried ambient air traceable to the WMO-X2014A scale) applying as a step-wise change

25   fortnightly, and with no further corrections applied in the meantime. These manual calibrations coincide with the observed change in the bias. Consequently, more frequent calibrations or automated measurements of a working standard to compensate for drift would have further improved the agreement. The ambient air measurements made at AMY were also in agreement with the TS comparison, which is illustrated in Figure 14. The scatter in the bias is significantly larger for ambient air measurements compared to the TS comparison. Firstly, part of this may be explained by the calibration strategy, as discussed

30   above. Secondly, differences in the response time for both instrument types as well as residence time in the inlet might further add to the observed scatter, especially in case of rapid changes in the CO amount fraction, which frequently occurred at AMY. Both campaigns show that accurate measurements of CO are possible if the sample air is dried. So far, this has not yet been implemented at all measurement stations. The above case study at PUY as well as the experiments done involving the droplet tests only investigated the internally implemented water vapour correction of the Picarro G2401, which proved ~~proofed~~ to be

not sufficiently stable to achieve the network compatibility goals of the WMO/GAW ~~Programme~~programme. Alternatively, better determination of the remaining water vapour interference is needed. The droplet method might not be suitable due to the relatively fast drying process, which results in relative~~ly~~ high uncertainties due to the analyser's noise. Alternative methods, e.g. as described by Reum et al. (2019) or as implemented by the ICOS Metrology Lab, which ~~is using~~uses a Bronkhorst

5   Vapour Delivery Module (VDM) ~~able~~ to humidify a gas stream from a tank, might give better results. In addition to improvements of the droplet method, alternative ways to compensate for the water vapour dependent CO bias need to be explored. Chen et al. (2013) showed that the main uncertainty of the water vapour correction is due to the fact that the weak CO absorption line is bracketed by adjacent absorption~~s~~ lines of CO_2 and H_2O. Our results indicate that the compensation of the water vapour interference based on the work of Chen et al. (2013), which has been implemented in Picarro analysers newer

10 than 2012, does not correct appropriately all the bias and may change over time. Therefore, frequent determination of the water vapour interference will be needed to ensure long-term stability of the correction function or to characterise its change over time. However, this will most likely be insufficient to detect the sudden changes in the correction function that were observed in our experiments. Consequently, drying of the sample air should be considered when measuring CO with a Picarro G2401 instrument. Both cryogenic traps and Nafion dryers can be used. WCC-Empa now uses Nafion dryers for the parallel

15 measurements during station audits. Both single tube (MD-070-48S-4) and multi tube (PD-50T-12MPS) Nafion dryers in reflux mode using the Picarro pump for the vacuum in the purge air were successfully used. This reduced the amount of water to approximately 0.06 – 0.22% (single tube) and 0.01-0.03% (multi tube), depending on ambient air humidity. In case of using Nafion dryers, the standard gases must also pass though the dryer to compensate for a potential loss over the dryer.

4   Conclusions

20 The different elements of the WMO/GAW quality management framework, including round robin experiments, performance audits with travelling standards and parallel measurements at stations provide complementary information which are essential for reducing the bias and uncertainty of time series measured by atmospheric research stations.

The assessment of performance audit results of CO and N_2O with respect to different measurement techniques showed clear advantages of newer spectroscopic techniques such as NIR-CRDS or QCL spectroscopy in the case of CO. However, parallel

25 measurements made using a Picarro NIR-CRDS analyser identified issues with the implemented water vapour compensation, and further improvement ~~are~~ is currently only possible by drying of the sample air. This can be implemented though drying of the sample air with cryogenic traps or Nafion dryers.

For N_2O, one of the limitations is the uncertainty of calibration standards. This highlights the importance of maintaining traceability to an internationally accepted calibration scale as implemented by the GAW programme.

30 By introducing modern spectroscopic measurement techniques such as CRDS or QCL, the number of GAW stations complying with the WMO/GAW network compatibility goals for CO and N_2O will increase. However, reaching the network compatibility

goal of 2 nmol/mol for CO and 0.1 nmol/mol for N$_2$O will remain challenging. Careful calibration strategies and appropriate water vapour corrections or drying of the sample air are required for both CO and N$_2$O.

Competing interests. The authors declare that they have no conflict of interest.

References

Baer, D. S., Paul, J. B., Gupta, M., and O'Keefe, A.: Sensitive absorption measurements in the near-infrared region using off-axis integrated-cavity-output spectroscopy, Applied Physics B: Lasers and Optics, 75, 261-265, 2002.

Blunden, J. and Arndt, D. S.: State of the Climate in 2016, Bull. Amer. Meteorol. Soc., 98, Si-S280, 2017.

5 Brewer, P. J., Brown, R. J. C., Tarasova, O., Hall, B., Rhoderick, G., and Wielgosz, R.: SI traceability and scales for underpinning atmospheric monitoring of greenhouse gases, Metrologia, doi: https://doi.org/10.1088/1681-7575/aad830, 2018. 2018.

Brunke, E. G., Scheel, H. E., and Seiler, W.: Trends of Tropospheric CO, N2O and CH4 as Observed at Cape Point, South-Africa, Atmospheric Environment Part a-General Topics, 24, 585-595, 1990.

10 Buchmann, B., Klausen, J., and Zellweger, C.: Traceability of Long-Term Atmospheric Composition Observations across Global Monitoring Networks, Chimia, 63, 657-660, 2009.

Chen, H., Karion, A., Rella, C. W., Winderlich, J., Gerbig, C., Filges, A., Newberger, T., Sweeney, C., and Tans, P. P.: Accurate measurements of carbon monoxide in humid air using the cavity ring-down spectroscopy (CRDS) technique, Atmos. Meas. Tech., 6, 1031-1040, 2013.

15 GAWSIS: GAW Station Information System, https://gawsis.meteoswiss.ch, last access: 23 July, 2018.

Gerbig, C., Schmitgen, S., Kley, D., Volz-Thomas, A., Dewey, K., and Haaks, D.: An improved fast-response vacuum-UV resonance fluorescence CO instrument, Journal of Geophysical Research-Atmospheres, 104, 1699-1704, 1999.

Gomez-Pelaez, A. J., Ramos, R., Gomez-Trueba, V., Novelli, P. C., and Campo-Hernandez, R.: A statistical approach to quantify uncertainty in carbon monoxide measurements at the Izaña global GAW station: 2008–2011, Atmos. Meas.
20 Tech., 6, 787-799, 2013.

Griffith, D. W. T., Deutscher, N. M., Caldow, C., Kettlewell, G., Riggenbach, M., and Hammer, S.: A Fourier transform infrared trace gas and isotope analyser for atmospheric applications, Atmos. Meas. Tech., 5, 2481-2498, 2012.

Gros, V., Bonsang, B., and Sarda Esteve, R.: Atmospheric carbon monoxide 'in situ' monitoring by automatic gas chromatography, Chemosphere - Global Change Science, 1, 153-161, 1999.

25 Hall, B. D., Dutton, G. S., and Elkins, J. W.: The NOAA nitrous oxide standard scale for atmospheric observations, Journal of Geophysical Research-Atmospheres, 112, 2007.

Hammer, S., Griffith, D. W. T., Konrad, G., Vardag, S., Caldow, C., and Levin, I.: Assessment of a multi-species in situ FTIR for precise atmospheric greenhouse gas observations, Atmos. Meas. Tech., 6, 1153-1170, 2013a.

Hammer, S., Konrad, G., Vermeulen, A. T., Laurent, O., Delmotte, M., Jordan, A., Hazan, L., Conil, S., and Levin, I.:
30 Feasibility study of using a "travelling" CO2 and CH4 instrument to validate continuous in situ measurement stations, Atmos. Meas. Tech., 6, 1201-1216, 2013b.

Khalil, M. A. K. and Rasmussen, R. A.: Carbon-Monoxide in the Earths Atmosphere - Indications of a Global Increase, Nature, 332, 242-245, 1988.

Khalil, M. A. K. and Rasmussen, R. A.: Increase and seasonal cycles of nitrous oxide in the earth's atmosphere, Tellus B, 35B,
35 161-169, 1983.

Klausen, J., Zellweger, C., Buchmann, B., and Hofer, P.: Uncertainty and bias of surface ozone measurements at selected Global Atmosphere Watch sites, Journal of Geophysical Research-Atmospheres, 108, 4622, doi:4610.1029/2003JD003710, 2003.

15

Lebegue, B., Schmidt, M., Ramonet, M., Wastine, B., Yver Kwok, C., Laurent, O., Belviso, S., Guemri, A., Philippon, C., Smith, J., and Conil, S.: Comparison of nitrous oxide (N2O) analyzers for high-precision measurements of atmospheric mole fractions, Atmos. Meas. Tech., 9, 1221-1238, 2016.

McManus, J. B., Zahniser, M. S., Nelson, D. D., Shorter, J. H., Herndon, S. C., Jervis, D., Agnese, M., McGovern, R., Yacovitch, T. I., and Roscioli, J. R.: Recent progress in laser-based trace gas instruments: performance and noise analysis, Applied Physics B-Lasers and Optics, 119, 203-218, 2015.

Nedelec, P., Cammas, J. P., Thouret, V., Athier, G., Cousin, J. M., Legrand, C., Abonnel, C., Lecoeur, F., Cayez, G., and Marizy, C.: An improved infrared carbon monoxide analyser for routine measurements aboard commercial Airbus aircraft: technical validation and first scientific results of the MOZAIC III programme, Atmospheric Chemistry And Physics, 3, 1551-1564, 2003.

NOAA: Carbon Monoxide (CO) WMO Scale, https://www.esrl.noaa.gov/gmd/ccl/co_scale.html, last access: 17 May, 2018a.

NOAA: Nitrous Oxide (N2O) — Combined Data Set, https://www.esrl.noaa.gov/gmd/hats/combined/N2O.html, last access: 28 May, 2018b.

NOAA: Nitrous Oxide (N2O) WMO Scale, https://www.esrl.noaa.gov/gmd/ccl/n2o_scale.html, last access: 17 May, 2018c.

NOAA: WMO/IAEA Round Robin Comparison Experiment, https://www.esrl.noaa.gov/gmd/ccgg/wmorr/index.html, last access: 23 July, 2018d.

Novelli, P. C.: CO in the atmosphere: measurement techniques and related issues, Chemosphere - Global Change Science, 1, 115-126, 1999.

Novelli, P. C., Masarie, K. A., Lang, P. M., Hall, B. D., Myers, R. C., and Elkins, J. W.: Re-analysis of tropospheric CO trends: Effects of the 1997-1998 wild fires, Journal of Geophysical Research-Atmospheres, 108, 4464, doi:4410.1029/2002JD003031, 2003.

Parrish, D. D., Holloway, J. S., and Fehsenfeld, F. C.: Routine, Continuous Measurement of Carbon-Monoxide with Parts-Per-Billion Precision, Environmental Science & Technology, 28, 1615-1618, 1994.

Prinn, R. G., Weiss, R. F., Fraser, P. J., Simmonds, P. G., Cunnold, D. M., Alyea, F. N., O'Doherty, S., Salameh, P., Miller, B. R., Huang, J., Wang, R. H. J., Hartely, D. E., Harth, C., Steele, L. P., Sturrock, G., Midgley, P. M., and McCulloch, A.: A history of chemically and radiatively important gases in air deduced from ALE/GAGE/AGAGE, Journal of Geophysical Research: Atmospheres, 105, 17751-17792, 2000.

Provencal, R., Gupta, M., Owano, T. G., Baer, D. S., Ricci, K. N., O'Keefe, A., and Podolske, J. R.: Cavity-enhanced quantum-cascade laser-based instrument for carbon monoxide measurements, Applied Optics, 44, 6712-6717, 2005.

Rella, C. W., Chen, H., Andrews, A. E., Filges, A., Gerbig, C., Hatakka, J., Karion, A., Miles, N. L., Richardson, S. J., Steinbacher, M., Sweeney, C., Wastine, B., and Zellweger, C.: High accuracy measurements of dry mole fractions of carbon dioxide and methane in humid air, Atmos. Meas. Tech., 6, 837-860, 2013.

Reum, F., Gerbig, C., Lavric, J. V., Rella, C. W., and Göckede, M.: Correcting atmospheric CO2 and CH4 mole fractions obtained with Picarro analyzers for sensitivity of cavity pressure to water vapor, Atmos. Meas. Tech., 12, 1013-1027, 2019.

Rhoderick, G. C., Kelley, M. E., Miller, W. R., Norris, J. E., Carney, J., Gameson, L., Cecelski, C. E., Harris, K. J., Goodman, C. A., Srivastava, A., and Hodges, J. T.: NIST Standards for Measurement, Instrument Calibration, and Quantification of Gaseous Atmospheric Compounds, Analytical Chemistry, 90, 4711-4718, 2018.

Rhoderick, G. C., Kitzis, D. R., Kelley, M. E., Miller, W. R., Hall, B. D., Dlugokencky, E. J., Tans, P. P., Possolo, A., and Carney, J.: Development of a Northern Continental Air Standard Reference Material, Analytical Chemistry, 88, 3376-3385, 2016.

van der Laan, S., Neubert, R. E. M., and Meijer, H. A. J.: A single gas chromatograph for accurate atmospheric mixing ratio measurements of CO2, CH4, N2O, SF6 and CO, Atmos. Meas. Tech., 2, 549-559, 2009.

16

Vardag, S. N., Hammer, S., O'Doherty, S., Spain, T. G., Wastine, B., Jordan, A., and Levin, I.: Comparisons of continuous atmospheric CH4, CO2 and N2O measurements - results from a travelling instrument campaign at Mace Head, Atmos. Chem. Phys., 14, 8403-8418, 2014.

Ventrillard, I., Xueref-Remy, I., Schmidt, M., Yver Kwok, C., Faïn, X., and Romanini, D.: Comparison of optical-feedback cavity-enhanced absorption spectroscopy and gas chromatography for ground-based and airborne measurements of atmospheric CO concentration, Atmos. Meas. Tech., 10, 1803-1812, 2017.

Weiss, R. F.: The temporal and spatial distribution of tropospheric nitrous oxide, Journal of Geophysical Research: Oceans, 86, 7185-7195, 1981.

WMO: 15th WMO/IAEA Meeting of Experts on Carbon Dioxide, Other Greenhouse Gases and Related Tracers Measurements Techniques, Jena, Germany, 7-10 September 2009 (WMO TD No. 1553), GAW Report No. 194, World Meteorological Organization, Geneva, Switzerland, 318 pp., 2011.

WMO: 16th WMO/IAEA Meeting on Carbon Dioxide, Other Greenhouse Gases, and Related Measurement Techniques (GGMT-2011), Wellington, New Zealand, 25-28 October 2011, GAW Report No. 206, World Meteorological Organization, Geneva, Switzerland, 2012.

WMO: 17th WMO/IAEA Meeting on Carbon Dioxide, Other Greenhouse Gases and Related Tracers Measurement Techniques (GGMT-2013), Beijing, China, 10-13 June 2013, GAW Report No. 213, World Meteorological Organization, Geneva, Switzerland, 2014.

WMO: 18th WMO/IAEA Meeting on Carbon Dioxide, Other Greenhouse Gases and Related Tracers Measurement Techniques (GGMT-2015), La Jolla, CA, USA, 13-17 September 2015, GAW Report No. 229, World Meteorological Organization, Geneva, Switzerland, 2016.

WMO: 19th WMO/IAEA Meeting on Carbon Dioxide, Other Greenhouse Gases and Related Tracers Measurement Techniques (GGMT-2017), Dübendorf, Switzerland, 27-31 August 2017, GAW Report No. 242, World Meteorological Organization, Geneva, Switzerland, 2018.

WMO: WMO Global Atmosphere Watch (GAW) Implementation Plan: 2016-2023 World Meteorological Organization, Geneva, Switzerland, 2017a.

WMO: WMO Greenhouse gas bulletin: the state of greenhouse gases in the atmosphere based on observations through 2016. https://library.wmo.int/doc_num.php?explnum_id=4022, last access: 28 August 2018, 2017b.

Yver Kwok, C., Laurent, O., Guemri, A., Philippon, C., Wastine, B., Rella, C. W., Vuillemin, C., Truong, F., Delmotte, M., Kazan, V., Darding, M., Lebègue, B., Kaiser, C., Xueref-Rémy, I., and Ramonet, M.: Comprehensive laboratory and field testing of cavity ring-down spectroscopy analyzers measuring H2O, CO2, CH4 and CO, Atmos. Meas. Tech., 8, 3867-3892, 2015.

Zellweger, C., Emmenegger, L., Firdaus, M., Hatakka, J., Heimann, M., Kozlova, E., Spain, T. G., Steinbacher, M., van der Schoot, M. V., and Buchmann, B.: Assessment of recent advances in measurement techniques for atmospheric carbon dioxide and methane observations, Atmos. Meas. Tech., 9, 4737-4757, 2016.

Zellweger, C., Hüglin, C., Klausen, J., Steinbacher, M., Vollmer, M., and Buchmann, B.: Inter-comparison of four different carbon monoxide measurement techniques and evaluation of the long-term carbon monoxide time series of Jungfraujoch, Atmos. Chem. Phys., 9, 3491-3503, 2009.

Zellweger, C., Steinbacher, M., and Buchmann, B.: Evaluation of new laser spectrometer techniques for in-situ carbon monoxide measurements, Atmos. Meas. Tech., 5, 2555-2567, 2012.

Table 1: CO performance audits using travelling standards from 2005 to 2017

Station / Laboratory	GAW ID	Year	Instrument	Method	Intercept (nmol/mol)	Slope (-)	Bias at 165 nmol/mol CO (nmol/mol)	Calibration scale station	Calibration scale WCC
Ryori	RYO	2005	Horiba GA-360	NDIR	7.0	0.989	5.1	CERI*	WMO-X2000
Mt. Kenya	MKN	2006	TEI 48C-TL	NDIR	-4.2	0.965	-10.0	WMO-X2000	WMO-X2000
Zugspitze-~~Schneefernerhaus~~	ZSF	2006	AL5001	VURF	2.0	0.957	-5.1	WMO-X2000	WMO-X2000
Zugspitze-~~Schneefernerhaus~~	ZSF	2006	AL5002	VURF	1.3	0.952	-6.5	WMO-X2000	WMO-X2000
Zugspitze-~~Schneefernerhaus~~	ZSF	2006	TEI 48C-TL	NDIR	-2.0	0.988	-4.0	WMO-X2000	WMO-X2000
Hohenpeissenberg	HPB	2006	AL5001	VURF	0.6	0.995	-0.2	WMO-X2000	WMO-X2000
Hohenpeissenberg	HPB	2006	TEI 48S	NDIR	-1.3	1.000	-1.3	WMO-X2000	WMO-X2000
Jungfraujoch	JFJ	2006	Agilent 6890	GC/FID	1.9	0.987	-0.2	WMO-X2000	WMO-X2000
Jungfraujoch	JFJ	2006	Horiba APMA360	NDIR	1.2	1.012	3.2	WMO-X2000	WMO-X2000
Cape Point	CPT	2006	RGA-3	GC/HgO	2.2	0.980	-1.0	WMO-X2000	WMO-X2000
Bukit Kototabang	BKT	2007	TEI 48C-TL	NDIR	-1.7	0.980	-5.0	WMO-X2000	WMO-X2000
Assekrem	ASK	2007	Horiba APMA360	NDIR	4.0	0.995	3.2	WMO-X2000	WMO-X2000
Pallas	PAL	2007	Agilent 6890N	GC/HgO	0.2	0.979	-3.3	WMO-X2000	WMO-X2000
Barrow	BRW	2008	RGA-3	GC/HgO	-1.7	1.006	-0.7	WMO-X2000	WMO-X2000
Bukit Kototabang	BKT	2008	Horiba APMA360	NDIR	0.6	0.932	-10.7	WMO-X2000	WMO-X2000
Mt. Kenya	MKN	2008	Horiba APMA360	NDIR	-4.9	1.006	-4.0	WMO-X2000	WMO-X2000
Mt. Kenya	MKN	2008	TEI 48C-TL	NDIR	-10.0	1.032	-4.6	WMO-X2000	WMO-X2000
Ushuaia	USH	2008	TEI 48	NDIR	-1.7	0.957	-8.8	WMO-X2000	WMO-X2000
Ushuaia	USH	2008	Horiba APMA360	NDIR	0.9	0.997	0.4	WMO-X2000	WMO-X2000
Amsterdam Island	AMS	2008	RGA-3	GC/HgO	10.3	0.834	-17.0	WMO-X2000	WMO-X2000
Izaña	IZO	2009	RGA-3	GC/HgO	-5.6	1.032	-0.4	WMO-X2004	WMO-X2000

18

Izaña	IZO	2009	TEI 48C-TL	NDIR	-6.3	0.922	-19.2	WMO-X2004	WMO-X2000
Santa Cruz	SCO	2009	TEI 48C-TL	NDIR	0.8	0.897	-16.2	WMO-X2004	WMO-X2000
Mt. Waliguan	WLG	2009	Agilent 6890	GC/FID	8.2	0.904	-7.6	WMO-X2004	WMO-X2000
CAWAS	NA	2009	Agilent 6890	GC/FID	6.9	0.910	-7.9	WMO-X2004	WMO-X2000
CAWAS	NA	2009	Ametek ta500R	GC/HgO	-24.6	1.358	34.5	WMO-X2004	WMO-X2000
Bukit Kototabang	BKT	2009	Horiba APMA360	NDIR	-0.2	0.989	-2.0	WMO-X2000	WMO-X2000
Mace Head	MHD	2009	RGA-3	GC/HgO	1.2	1.006	2.2	CSIRO94	WMO-X2000
Lauder	LAU	2010	FTIR	FTIR	-0.6	0.979	-4.0	WMO-X2004	WMO-X2000
Mt. Kenya	MKN	2010	Horiba APMA360	NDIR	-8.1	0.978	-11.8	WMO-X2000	WMO-X2000
Mt. Kenya	MKN	2010	TEI 48C-TL	NDIR	-6.2	1.054	2.8	WMO-X2000	WMO-X2000
Cape Point	CPT	2011	RGA-3	GC/HgO	2.2	0.953	-5.6	WMO-X2004	WMO-X2000
Zugspitze-Schneefernerhaus	ZSF	2011	AL5001	VURF	1.7	0.977	-2.1	WMO-X2004	WMO-X2000
Zugspitze-Schneefernerhaus	ZSF	2011	AL5002	VURF	1.4	0.987	-0.7	WMO-X2004	WMO-X2000
Hohenpeissenberg	HPB	2011	AL5001	VURF	0.9	0.999	0.8	WMO-X2004	WMO-X2000
Bukit Kototabang	BKT	2011	Horiba APMA360	NDIR	4.5	0.909	-10.5	WMO-X2000	WMO-X2000
Pallas	PAL	2012	PeakPerformer 1	GC/HgO	2.8	1.042	9.8	WMO-X2004	WMO-X2004
Pallas	PAL	2012	Picarro G2401	NIR-CRDS	1.1	1.001	1.3	WMO-X2004	WMO-X2004
Zeppelin Mountain	ZEP	2012	Picarro G2401	NIR-CRDS	-2.5	1.015	0.0	WMO-X2004	WMO-X2004
Zeppelin Mountain	ZEP	2012	RGA-3	GC/HgO	-3.7	1.036	2.2	WMO-X2004	WMO-X2004
Mt. Cimone	CMN	2012	Agilent 6890N	GC/FID	-9.7	1.048	-1.8	WMO-X2004	WMO-X2004
Mt. Cimone	CMN	2012	TEI 48C-TL	NDIR	4.4	0.945	-4.7	WMO-X2004	WMO-X2004
Capo Verde	CVO	2012	AL5001	VURF	0.0	0.991	-1.4	WMO-X2004	WMO-X2004
Capo Verde	CVO	2012	LGR-23d	QCL	-3.8	1.023	0.1	WMO-X2004	WMO-X2004
Mace Head	MHD	2013	RGA-3	GC/HgO	-2.9	1.101	13.8	CSIRO94	WMO-X2004
Izaña	IZO	2013	RGA-3	GC/HgO	-2.3	1.010	-0.7	WMO-X2004	WMO-X2004
Bukit Kototabang	BKT	2014	Horiba APMA360	NDIR	-0.4	0.873	-21.4	WMO-X2004	WMO-X2004
Jungfraujoch	JFJ	2015	Picarro G2401	NIR-CRDS	-3.9	1.008	-2.6	WMO-X2014	WMO-X2014
Jungfraujoch	JFJ	2015	LGR-23r	QCL	-1.4	1.008	0.0	WMO-X2014	WMO-X2014
Ushuaia	USH	2016	Horiba APMA360	NDIR	-1.0	1.001	-0.9	WMO-X2014A	WMO-X2014A

Puy de Dôme	PUY	2016	Picarro G2401	NIR-CRDS	-0.1	1.001	0.0	WMO-X2014A	WMO-X2014A
CAWAS	NA	2016	Picarro G2302	NIR-CRDS	-4.2	1.010	-2.5	WMO-X2014A	WMO-X2014A
Mt. Waliguan	WLG	2016	Picarro G2401	NIR-CRDS	-6.2	1.034	-0.6	WMO-X2014A	WMO-X2014A
Linan	LAN	2016	Picarro G2401	NIR-CRDS	-2.0	1.008	-0.7	WMO-X2014A	WMO-X2014A
Mt. Waliguan	WLG	2016	Agilent 6890N	GC/FID	0.4	0.990	-1.2	WMO-X2014A	WMO-X2014A
Linan	LAN	2016	Agilent 7890A	GC/FID	-17.4	1.076	-4.9	WMO-X2014A	WMO-X2014A
Lauder	LAU	2016	FTIR	FTIR	1.6	0.955	-5.7	WMO-X2014A	WMO-X2014A
Anmyeon-do	AMY	2017	TEI48i-TLE	NDIR	-42.2	1.080	-29.1	WMO-X2014A	WMO-X2014A
Anmyeon-do	AMY	2017	LGR 30-EP	QCL	1.3	0.996	0.7	WMO-X2014A	WMO-X2014A
Jeju Gosan	JGS	2017	TEI48i-TLE	NDIR	16.8	0.969	11.7	WMO-X2014A	WMO-X2014A

*Chemicals Evaluation and Research Institute, Japan

Table 2: N$_2$O audits with travelling standards performed by WCC-N$_2$O (2002 - 2013) and WCC-Empa (2009 - 2017)

Station / Laboratory	GAW ID	Year	Instrument	Method	Intercept (nmol/mol)	Slope (-)	Ambient bias (nmol/mol)	Range (nmol/mol)	Calibration scale station	Calibration scale WCC	Audit made by
Zugspitze	ZFS	2002	HP6890	GC/ECD	56.97	0.8109	-3.11	312.7-322.7	SIO-1993	WMO-X2000	WCC-N$_2$O
Schauinsland	SSL	2002	HP6890	GC/ECD	55.18	0.8265	0.06	312.7-322.7	SIO-1993	WMO-X2000	WCC-N$_2$O
Cape Point	CPT	2003	Shimadzu GC-8A	GC/ECD	-3.73	1.0113	-0.12	313.5-323.5	WMO-X2000	WMO-X2000	WCC-N$_2$O
Zugspitze-Gipfel	ZUG	2005	HP6890	GC/ECD	0.04	1.0017	0.57	315.1-325.1	SIO-1998	WMO-X2000	WCC-N$_2$O
Jungfraujoch	JFJ	2006	Agilent 6890	GC/ECD	27.26	0.9144	-0.22	315.9-325.9	SIO-1998	WMO-X2006	WCC-N$_2$O
Pallas	PAL	2007	HP6890	GC/ECD	-3.20	1.0110	0.32	316.7-326.7	WMO-X2000	WMO-X2006	WCC-N$_2$O
Izaña	IZO	2008	Varian-CP-3800	GC/ECD	2.67	0.9915	-0.08	317.5-327.5	WMO-X2006	WMO-X2006	WCC-N$_2$O
CARIBIC	NA	2008	HP6890	GC/ECD	12.49	0.9600	-0.39	317.5-327.5	WMO-X2006	WMO-X2006	WCC-N$_2$O
Mt. Cimone	CMN	2008	Agilent 6890N	GC/ECD	43.86	0.8663	0.74	317.5-327.5	WMO-X2006	WMO-X2006	WCC-N$_2$O
Izaña	IZO	2009	Varian-3800	GC/ECD	7.81	0.9761	0.03	318.3-328.3	WMO-X2006	WMO-X2006	WCC-Empa
Mt. Waliguan	WLG	2009	Aglient 6890	GC/ECD	50.84	0.8569	4.60	318.3-328.3	WMO-X2006	WMO-X2006	WCC-Empa
CAWAS	NA	2009	Agilent 6890N	GC/ECD	4.52	0.9863	0.33	318.3-328.3	WMO-X2006	WMO-X2006	WCC-Empa
CAWAS	NA	2009	Agilent 6890N	GC/ECD	-4.27	1.0141	0.27	318.3-328.3	WMO-X2006	WMO-X2006	WCC-Empa
Mace Head	MHD	2009	HP5890	GC/ECD	-8.58	1.0278	0.42	318.3-328.3	SIO-2005	WMO-X2006	WCC-Empa
Lauder*	LAU	2010	FTIR	FTIR	-0.72	1.0026	0.12	319.1-329.1	WMO-X2006	WMO-X2006	WCC-Empa
Schauinsland	SSL	2010	HP6890	GC/ECD	9.66	0.9710	0.28	319.1-329.1	WMO-X2006	WMO-X2006	WCC-N$_2$O
Cape Point	CPT	2011	Agilent 6890N	GC/ECD	-4.52	1.0140	0.03	319.9-329.9	WMO-X2006	WMO-X2006	WCC-N$_2$O
CARIBIC	NA	2011	HP6890	GC/ECD	6.69	0.9784	-0.32	319.9-329.9	WMO-X2006	WMO-X2006	WCC-N$_2$O
Baring Head	BAR	2011	Agilent 6890	GC/ECD	0.36	1.0009	0.65	319.9-329.9	WMO-X2006A	WMO-X2006A	WCC-N$_2$O
Mace Head	MHD	2012	~~HP8590~~HP5890	GC/ECD	0.81	0.9989	0.44	320.7-330.7	SIO-2005	WMO-X2006A	WCC-N$_2$O
Mt. Cimone	CMN	2012	Agilent 6890N	GC/ECD	125.50	0.6166	0.62	320.7-330.7	WMO-X2006A	WMO-X2006A	WCC-Empa
Capo Verde	CVO	2012	LGR 23-r	QCL	14.68	0.9545	-0.13	320.7-330.7	WMO-X2006A	WMO-X2006A	WCC-Empa
Mace Head	MHD	2013	HP 5800 II	GC/ECD	-0.38	1.0004	-0.25	321.5-331.5	SIO-2005	WMO-X2006A	WCC-Empa
Izaña	IZO	2013	Varian 3800	GC/ECD	3.35	0.9889	-0.27	321.5-331.5	WMO-X2006A	WMO-X2006A	WCC-Empa

Anmyeon-do	AMY	2013	Agilent 7890II	GC/ECD	6.03	0.9802	-0.44	321.5-331.5	WMO-X2006A	WMO-X2006A	WCC-N2O
Jeju Gosan	JGS	2013	Agilent 7890II	GC/ECD	25.49	0.9211	-0.27	321.5-331.5	WMO-X2006A	WMO-X2006A	WCC-N2O
Bukit Kototabang	BKT	2014	Thermo IRIS 4600	DFG	-8.86	1.0259	-0.39	322.3-332.3	WMO-X2006A	WMO-X2006A	WCC-Empa
Jungfraujoch	JFJ	2015	LGR-23r	QCL	0.87	0.9965	-0.26	323.1-333.1	WMO-X2006A	WMO-X2006A	WCC-Empa
Cape Point	CPT	2015	Agilent 6890N	GC/ECD	-16.08	1.0519	0.96	323.1-333.1	WMO-X2006A	WMO-X2006A	WCC-Empa
CAWAS	NA	2016	AGILENT 7890A	GC/ECD	11.84	0.9645	0.16	323.9-333.9	WMO-X2006A	WMO-X2006A	WCC-Empa
Mt. Waliguan	WLG	2016	AGILENT 6890N	GC/ECD	11.86	0.9638	-0.06	323.9-333.9	WMO-X2006A	WMO-X2006A	WCC-Empa
Linan	LAN	2016	AGILENT 7890A	GC/ECD	7.26	0.9781	0.05	323.9-333.9	WMO-X2006A	WMO-X2006A	WCC-Empa
Lauder*	LAU	2016	FTIR	FTIR	-0.07	1.0036	1.12	323.9-333.9	WMO-X2006A	WMO-X2006A	WCC-Empa
Anmyeon-do	AMY	2017	Los Gatos 30-EP	QCL	6.43	0.9809	0.14	324.7-334.7	WMO-X2006A	WMO-X2006A	WCC-Empa
Anmyeon-do	AMY	2017	Agilent 7890II	GC/ECD	30.37	0.9059	-0.65	324.7-334.7	WMO-X2006A	WMO-X2006A	WCC-Empa
Jeju Gosan	JGS	2017	Agilent 7890II	GC/ECD	-2.95	1.0083	-0.22	324.7-334.7	WMO-X2006A	WMO-X2006A	WCC-Empa

* The difference between the two FTIR comparisons is due to an offset of the working standard used in 2016 and is not related to the performance of the analyser.

Table 3: Overview of ambient air CO comparison campaigns.

Location	Coordinates	Start	End	Station instrument	Travelling Instrument
PUY	45.7723 N 2.9658 E	2016-04-11	2016-06-22	Picarro G2401 #CFKADS-2161 dry measurements	Picarro G2401 #CFKADS2098 humid and dry meas.
AMY	36.5383 N 126.3300 E	2017-07-31	2017-09-05	LGR N_2O/CO-30-EP #15-0213 dry measurements	Picarro G2401 #CFKADS2098 dry measurements.
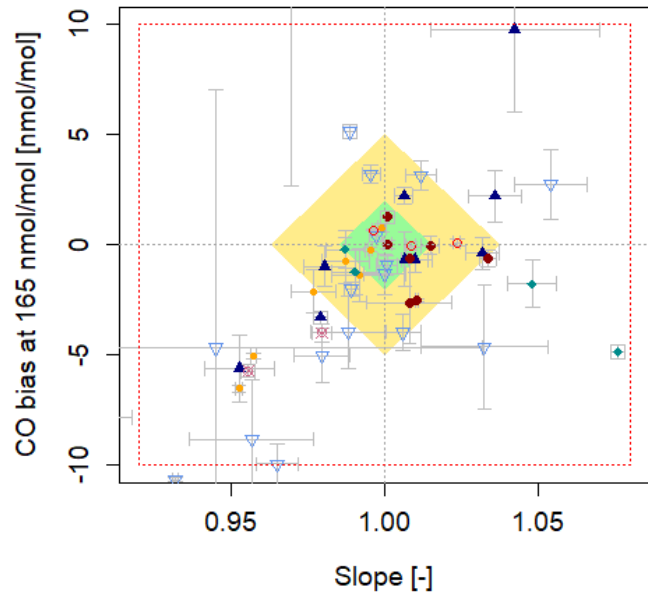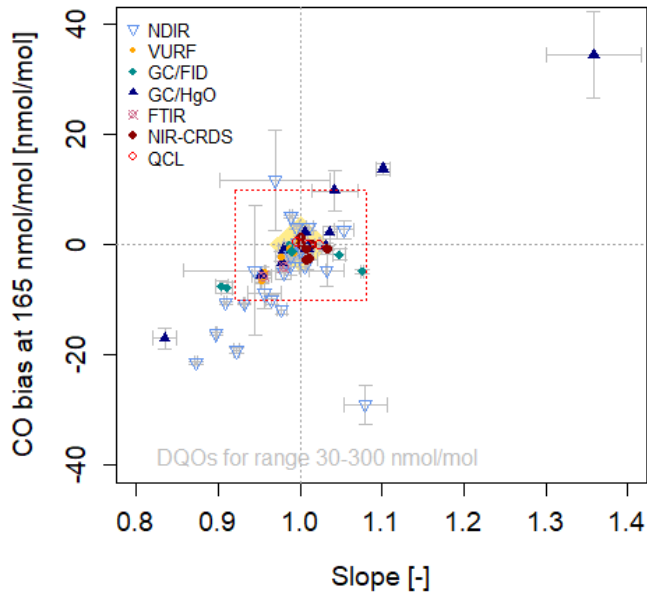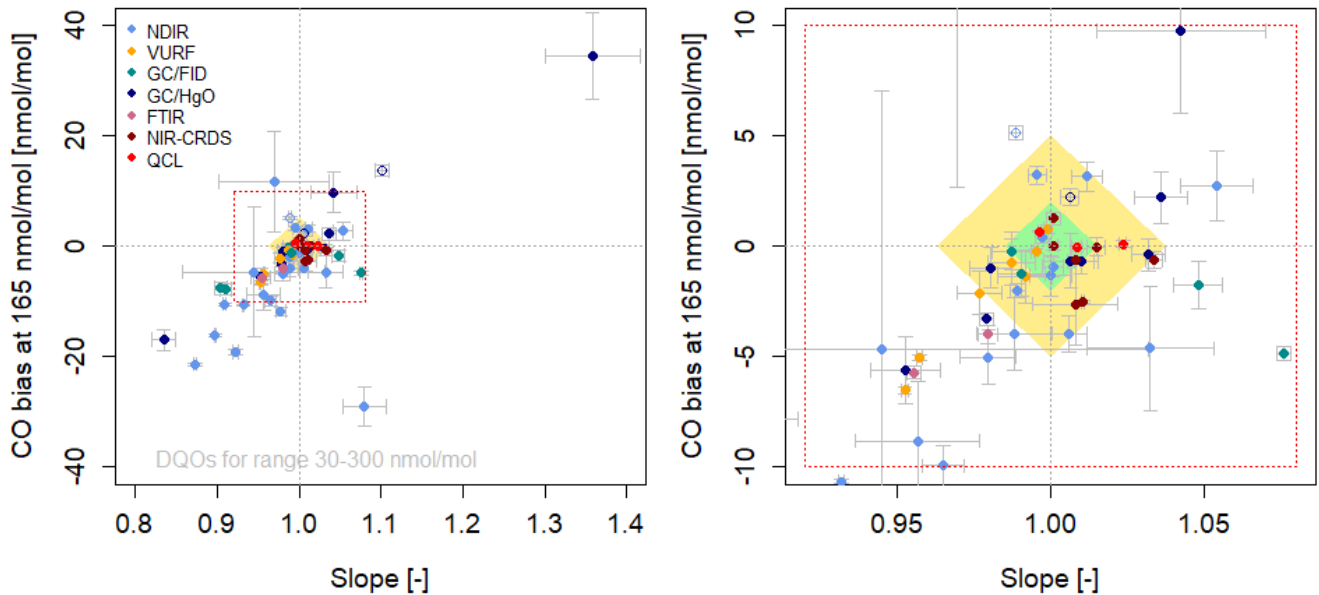
Figure 1. Left: CO bias at 165 nmol/mol vs. the slope of the audit for individual travelling standard comparisons. Different ~~symbols and~~ colours indicate different measurement techniques of the station analysers. <u>Filled symbols refer to comparison with the same calibration scale at the station and the WCC, while open symbols indicate a scale difference.</u> The error bars correspond to the uncertainty of the slope and the bias (1σ). The green and yellow areas correspond to the WMO/GAW <u>network</u> compatibility and extended <u>network</u> compatibility goals for the amount fraction range of 30 - 300 nmol/mol. Right: detail of the red dotted box of the left panel.
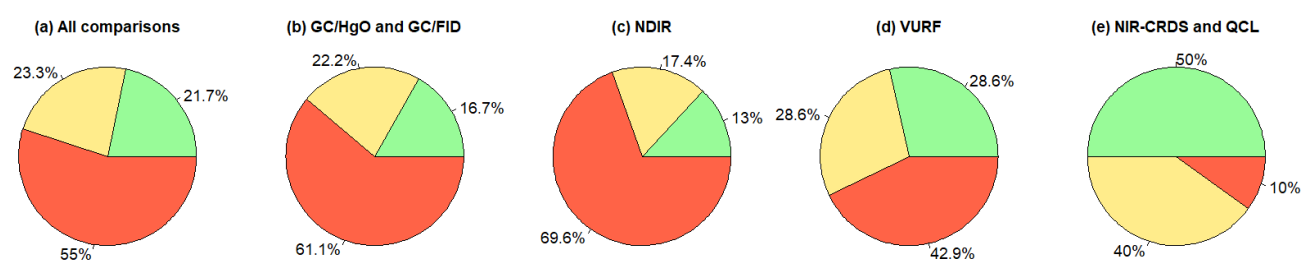
Figure 2.Percentage of CO performance audit results that were in the range of 30 - 300 nmol/mol within the WMO/GAW <u>network</u> compatibility goals (green), the extended <u>network</u> compatibility goals (yellow), or outside the <u>network</u> compatibility goals (red area) for (a) all comparisons, (b) GC systems, (c) NDIR analysers, (d) VURF analysers, and (e) NIR-CRDS and QCL systems.
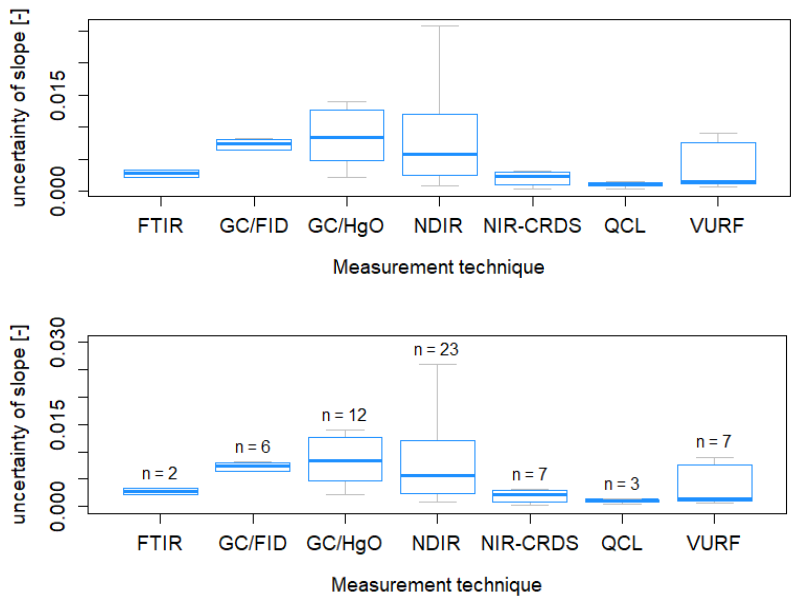
24

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.
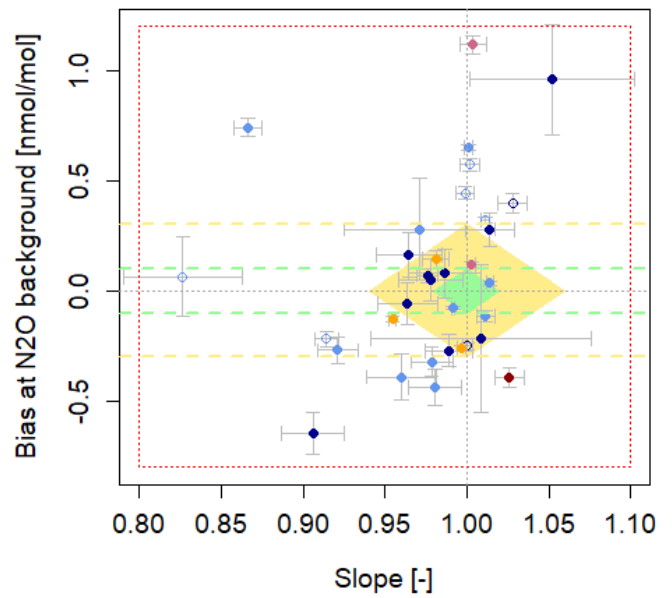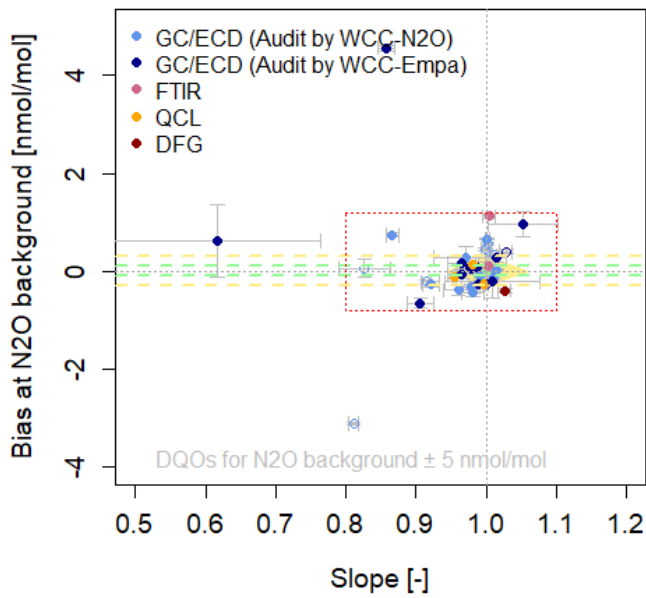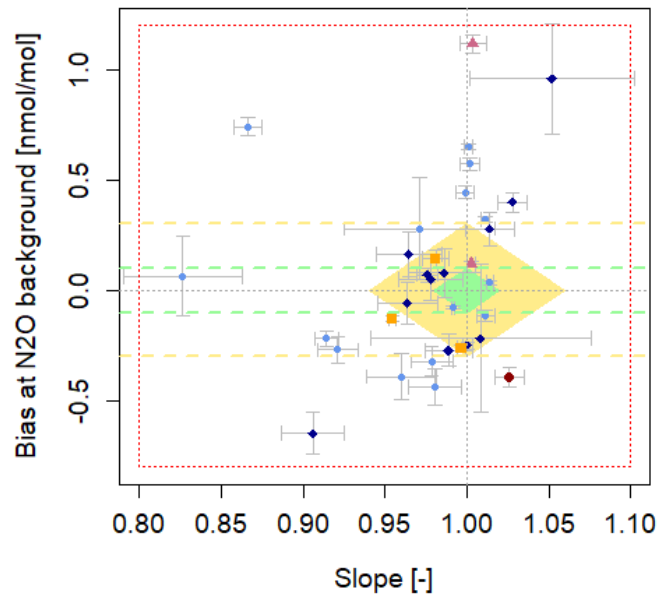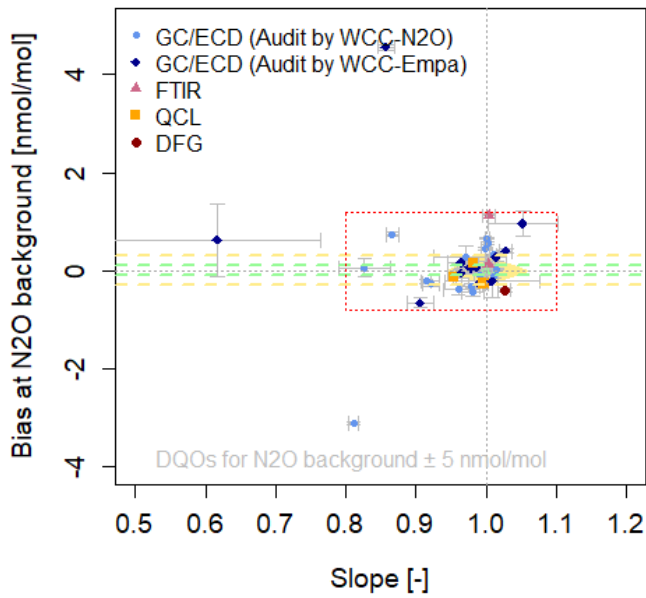
Figure 4. Left: N₂O bias in the centre of the relevant ambient air amount fraction (Table 2) vs. the slope of the audit for individual travelling standard comparisons. Different ~~symbols and~~ colours indicate different measurement techniques of the station analysers. <u>Filled symbols refer to comparison with the same calibration scale at the station and the WCC, while open symbols indicate a scale difference.</u> The error bars correspond to the uncertainty of the slope and the bias (1σ). The green and

yellow areas correspond to the WMO/GAW network compatibility and extended network compatibility goals for the range of ±-5 nmol/mol around the centre of the relevant amount fraction range, and the dashed green and yellow lines show the limits at the relevant amount fraction. Right: detail of the red dotted box of the left panel.
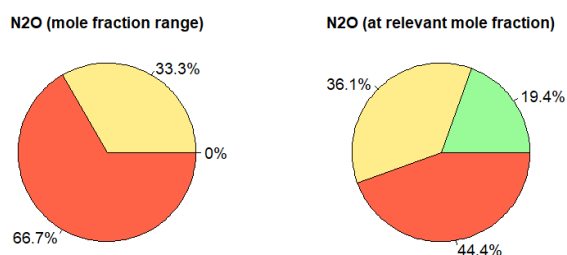
Figure 5. Left: Percentage of N₂O performance audit results that were for the range of the relevant amount fraction ±5 nmol/mol within the WMO/GAW network compatibility goals (green), the extended network compatibility goals (yellow), or outside the network compatibility goals (red area). Right: Same as on the left side but at the relevant amount fraction (see text for details).
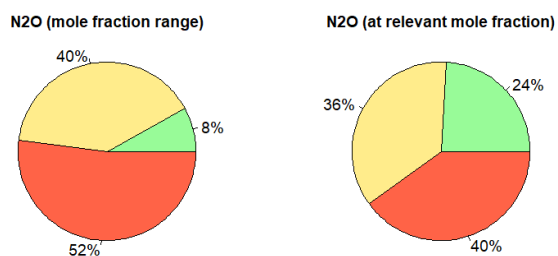
Figure 6. Left: Percentage of the results of the 6ᵗʰ round robin experiment that were for the range of the relevant amount fraction ±5 nmol/mol within the WMO/GAW network compatibility goals (green), the extended network compatibility goals (yellow), or outside the network compatibility goals (red area). Right: Same as on the left side but at the relevant amount fraction (see text for details).

27
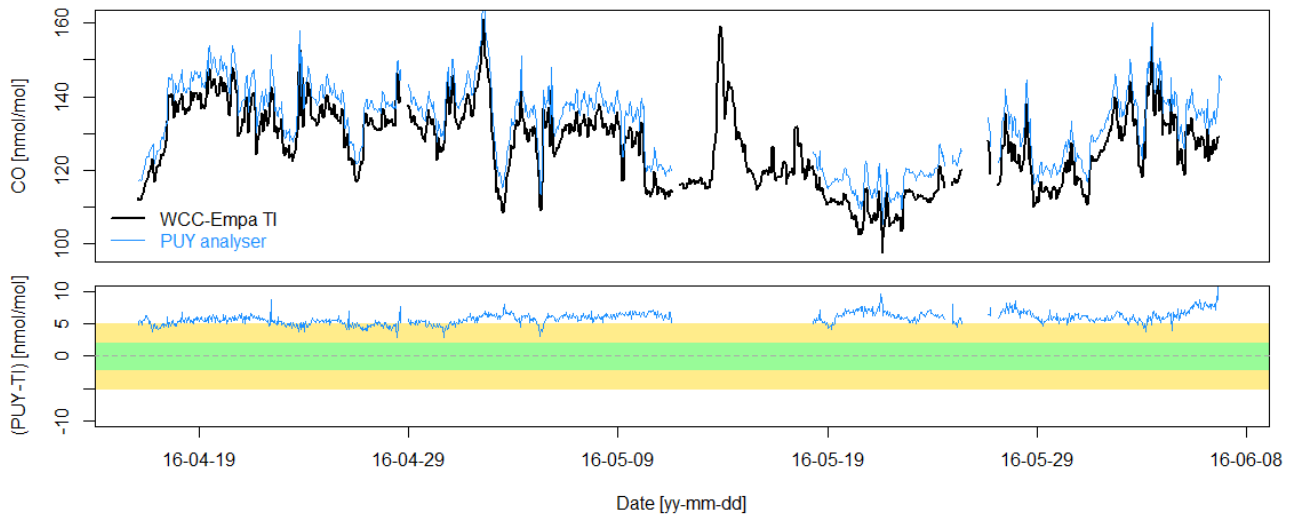
Figure 7. ~~CO comparison~~Comparison of hourly averages of CO at PUY between the WCC-Empa travelling instrument and the PUY Picarro G2401 for the period when the TI sampled humid air. Upper panel: CO time series ~~(1 h data)~~. Lower panel: CO bias of the station analyser vs time. The green and yellow areas correspond to the WMO network compatibility and extended network compatibility goals.
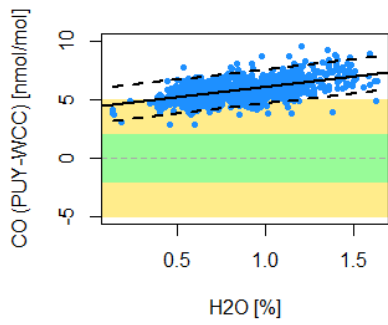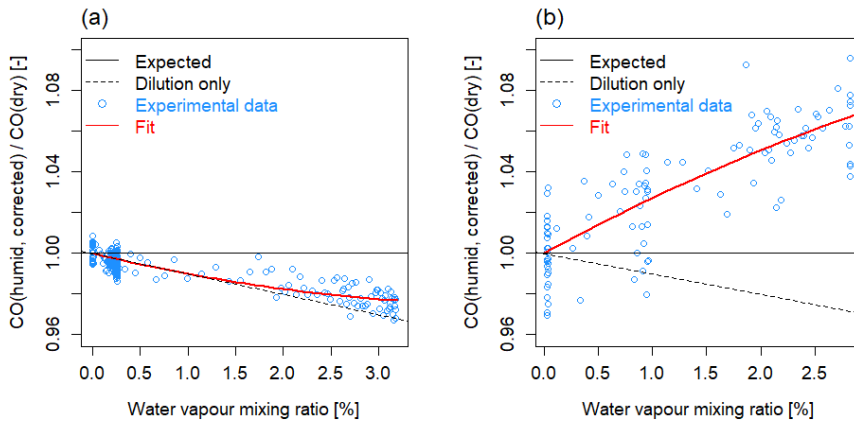
28

Figure 8. Bias of the PUY Picarro G2401 vs. the water vapour measured by the TI. The solid black line shows the linear regression with 95% confidence interval (dashed lines). The green and yellow areas correspond to the WMO network compatibility and extended network compatibility goals.

5

Figure 9. CO(humid, corrected) / CO(dry) vs. the reported water vapour for the experiment before (a) (2016-03-23) and after (b) (2016-07-14) the comparison at PUY.
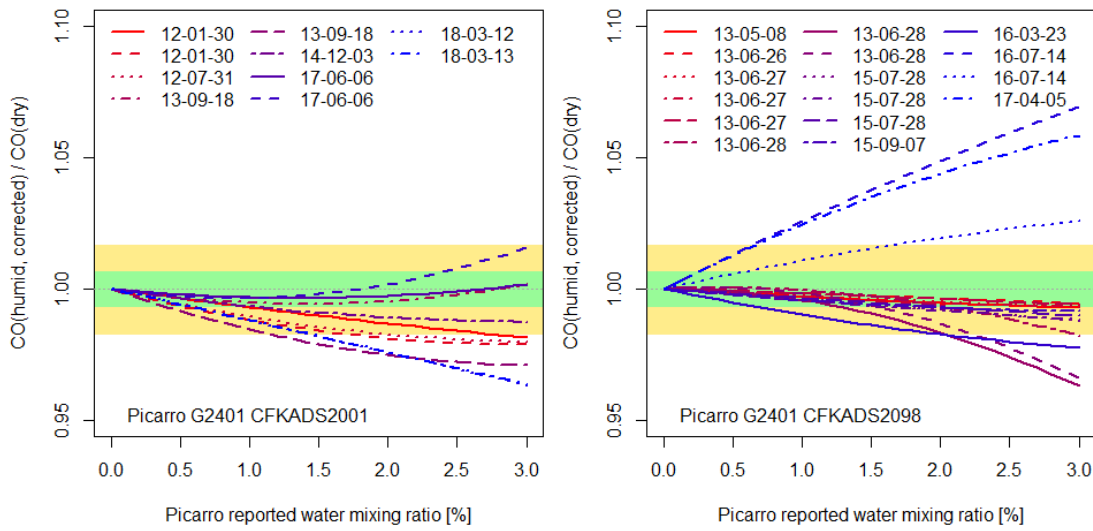
Figure 10. Ratios of CO(humid, corrected) / CO(dry) amount fractions vs. the water vapour mixing ratios of two different Picarro G2401 NIR-CRDS analysers over time. The legend shows the date (yy-mm-dd) of the experiment. The coloured areas are the limits for the WMO/GAW network compatibility goal (green) and extended (yellow) network compatibility goal at the amount fraction of 300 nmol/mol CO.
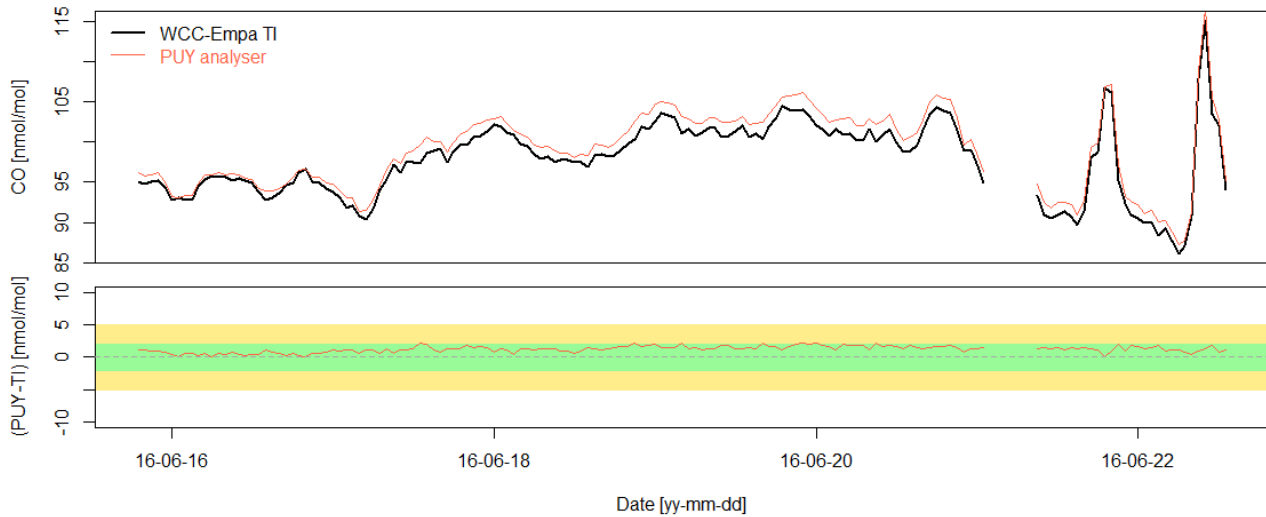

Figure 11. Comparison of hourly averages of CO CO comparison at PUY between the WCC-Empa travelling instrument and the PUY Picarro G2401 for the period when the TI sampled dry air. Upper panel: CO time series (1 h data). Lower panel: CO bias of the station analyser vs time. The green and yellow areas correspond to the WMO network compatibility and extended network compatibility goals.
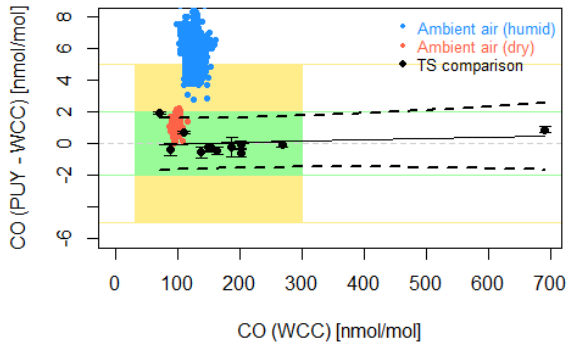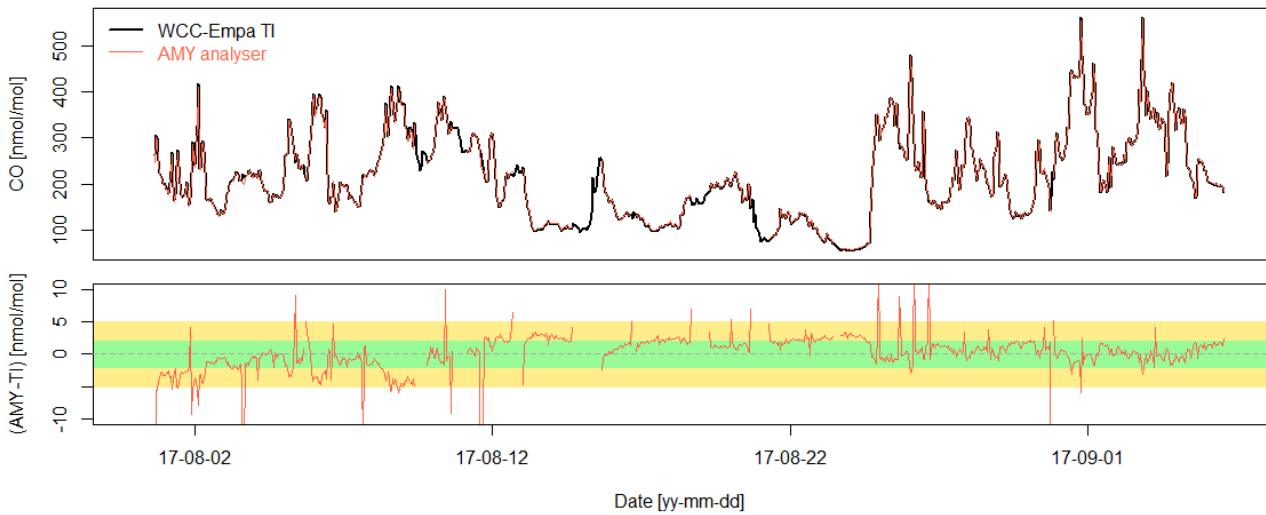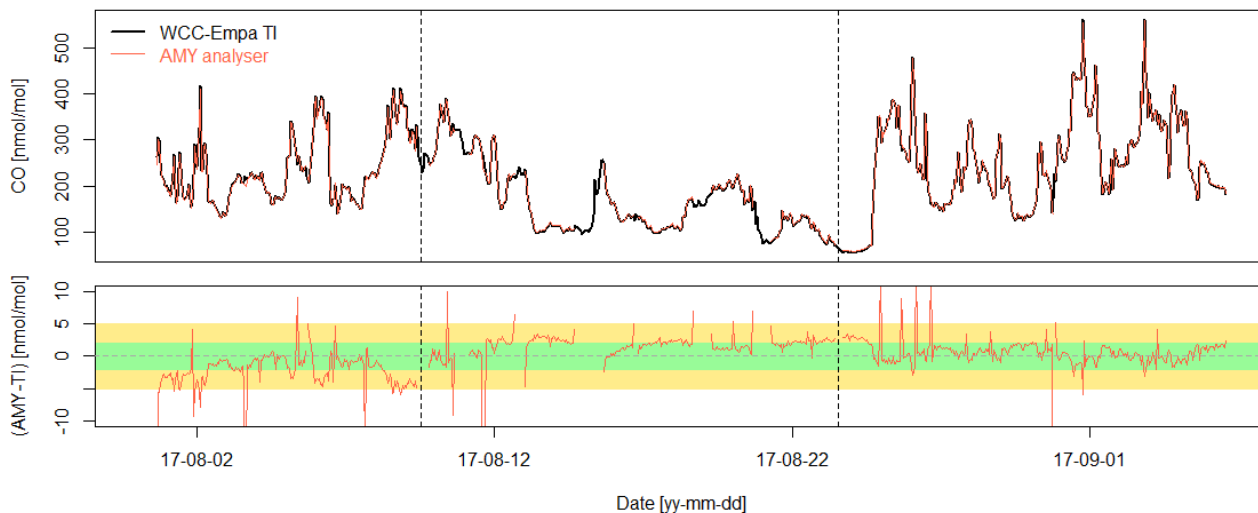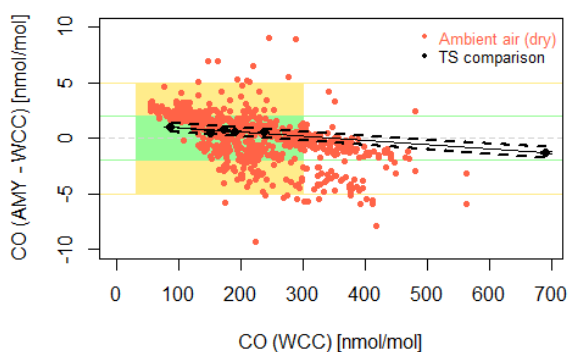
30

Figure 12. Bias of the PUY Picarro G2401 CO instrument vs. WCC-Empa assigned values. Black dots represent the average of data at a given level from a specific TS comparison. The error bars show the standard deviation of individual measurement points. The green and yellow lines correspond to the WMO network compatibility and extended network compatibility goals, and the green and yellow areas to the amount fraction range relevant for PUY. The dashed lines around the regression lines are the Working-Hotelling 95percentage confidence intervals. The coloured dots show the bias during the ambient air comparison without (blue) and with (red) drying of the air sampled by the TI.

Figure 13. Comparison of hourly averages of CO CO comparison at AMY between the WCC-Empa travelling instrument and the AMY Los Gatos 30-EP QCL analyser. Both instruments sampled dry ambient air. Upper panel: CO time series (1 h data). Lower panel: CO bias of the station analyser vs time. The green and yellow areas correspond to the WMO network compatibility and extended network compatibility goals. The dashed vertical lines indicate the time of the calibration of the AMY instrument.

Figure 14. Bias of the AMY Los Gatos 30-EP CO instrument vs. WCC-Empa assigned values. Black dots represent the average of data at a given level from a specific TS comparison. The error bars show the standard deviation of individual measurement points. The green and yellow lines correspond to the WMO network compatibility and extended network compatibility goals, and the green and yellow areas to the amount fraction range relevant for PUY. The dashed lines around the regression lines are the Working-Hotelling 95percentage confidence intervals. The red dots show the bias during the ambient air comparison.