Response to comments from Referee 1 (Dr. Ruediger Lang)

Black: Referee's comments; Blue: Authors' answers

We thank referee #1, Dr. Ruediger Lang for the review and for providing useful feedback.

Referee:
Accurate ground based observations of carbon dioxide and methane are becoming increasingly important in the context of continuous satellite remote sensing validation. While ground based in-situ and remote sensing networks measuring CO_2 and CH4 concentrations in the atmosphere exist, and are instrumental in measuring the continuous increase in greenhouse gases over the past decades, it has been recognised that the existing network of stations is still lacking with respect to its expected role in future monitoring and verification system-of-systems (MVS) of greenhouse gas emissions. Ground based remote sensing instruments like the Total Carbon Column Observing Network (TCONN) FTS instruments measuring total column dry air mole fractions of carbon dioxide and methane (XCO_2 and XCH4) are key for the validation and calibration of future operational satellite based observing systems, targeting greenhouse gas emissions. However, the current distribution of stations and their representability for the validation of global satellite based measurements, as well as their maintenance and operability for providing a continuous flow of quality-monitored data are providing big challenges ahead for the network to become the much aimed for fiducial greenhouse-gas reference measurements within an operational MVS context. One of the key-challenges is achieving and maintaining the very high accuracies needed – significantly below 0.5 ppm for XCO_2 – in order to become useful for the MVS and the Cal/Val of its satellite components.

The paper by Sha et al. is an important and significant contribution towards the latter aspect, by addressing the question on how to secure high accuracies across the network, e.g. through travelling instrument standards, while identifying, correcting and potentially even reducing instrument and measurement biases. The key finding of the paper is, from my point of view, the potential of at least one of the instruments taking part in the campaign exercise at Sodankyla, Finnland, functioning as a "travelling standard" in a potential future operational network of reference FTS instruments as currently operated under the TCCON umbrella. The paper also addresses important open questions with respect to remaining systematic biases and points at, maybe even more significant, remaining issues in measurement bias and precision. The paper is well written, although I think the paper could benefit from some restructuring of the results sections. I therefore recommend the paper for publishing in ACP, but would like to highlight in the following a couple of observations (apart from some additional minor comments), which the authors, from my point of view, should address.
Authors' response:
Thanks for the positive comments. As the paper focuses on the measurements of atmospheric components using ground-based instruments, we think it is better suited for publication in AMT rather than ACP and would like to keep it here.

Referee:

1) Ari-core comparisons and non-linearity effects

The discussion of the non-linearity effect found in the measurements of the Bruker IFS HR reference instruments (referred to as TCCON “reference”, 125HR), as well as the comparison to the AirCore measurements, as taken during the campaign at Sodankyla - which are considered to represent the “true atmospheric” state - are both presented only at the end of the major results section 5. This is first of all confusing, since the section on non-linearity effects implies that the non-linearity correction “has been applied to TCCON” data as a whole. At this stage, the reader wonders if this therefore now applies to all previous results, but the answer is probably not, since the results and different labelling then implies that the main comparisons results are not (yet) non-linearity corrected.

Second, the performance of the “reference instruments” with respect to what is considered the “true state of the atmosphere” represented by the AirCore measurements is an important result against which also the results of the measurements from the other systems have to be evaluated and interpreted. So both aspects have to be taken together. I would therefore recommend to present the results on the 125HR instrument non-linearity and the performance of the reference instruments against the “truth” (5.8 and 5.9) at the beginning of Section 5 (and ideally then present only the non-linearity corrected results for the TCONN reference – if this is considered a stable result – in the comparison against the other systems). This would help to interpret the results of the other systems when compared to the “reference” 125HR better with respect to the AirCore “truth”. In addition, establishing the TCCON 125HR as a “reference” and therefore then talking about a “bias” with respect to the reference for the other instruments (and not in terms of “differences”), requires, in my view, presenting the AirCore results first anyhow.

Authors’ response:

Thanks for this suggestion.

Our original idea was to present the comparison results first with the official TCCON data and then show the results of the non-linearity corrected TCCON data, which was part of the lessons learned during the campaign. We agree to the referee’s viewpoint and have moved the (former) sections 5.8 and 5.9 to the beginning of Section 5. In addition, we have moved the other intercomparison results with the official TCCON data and the HR125LR data to the appendix. These results are useful as a reference comparison to the official TCCON data. They also provide useful analysis on the resolution dependent effects on the Xgas retrievals of the target gases.

Referee:

2) EM27 systematic biases with respect to the two references The EM27 (COCCON) instruments show a convincing performance with respect to the “reference” TCCON (125HR), both with respect to the systematic biases and precision (for the latter see below). This is alas true especially with respect to the non- corrected TCCON results. It seems peculiar that this bias is even lower than -0.2 ppm for the non-corrected comparisons, but then gets worse for the TCCON non-linearity corrected ones (-0.73 ppm). At the same time, the low-resolution “reference” measurements (125LR) show a high bias to EM27 of roughly the same order of magnitude than the effect of the non-linearity correction. Considering that the low resolution 125HR are more comparable in terms of information content (and probably AKs, although those have

unfortunately not been presented –see below), one wonders if the non-linearity correction should not also be applied to the 125LR measurements (or has this been done?), which then may lead to a consistency between EM27 and the 125LR. This would then also physically make a lot of sense considering the information content of both measurements. Also it would be important to rule out any link (e.g. in retrieval processing) between the standard 125HR measurements (not corrected) and the EM27 results, which potentially make them similar to the standard reference by default (and therefore more biased with respect to the corrected 125HR results and the AirCore "truth"). In this context, the processing algorithm ProFIT vs GFIT performances and their potentially relevant "peculiarities" are not much discussed at all. ProFIT is used both for the TCCON 125LR and the COCCON EM27s but not for the uncorrected TCCON "reference" it seems. Is the bias correction scheme used in ProFit maybe somehow related to TCCON measurements (or associated climatologies?).

I think it would be important to add a discussion of the (potential) relationship between the observed systematic biases between "reference" HR, LR, and non-linearity correction on one side, and the EM27 measurement results on the other side, which could shine some light on the underlying mechanisms.

Authors' response:

We have added a description of the PROFFAST code in the paper and the underlying differences with respect to the GFIT code. See lines 173 – 188: "PROFFAST is a code for retrieving trace gas amounts from low-resolution solar absorption spectra. It has been developed on behalf of ESA, in order to provide a source-open and freely available code (without any licensing restrictions) as required by the growing COCCON user community, e.g. for TROPOMI validation work. It is a least-squares fitting algorithm, which adjusts the trace gas amounts by scaling atmospheric a priori profiles. The retrievals are performed on spectra generated with the included PREPROCESS tool. This tool produces spectra out of the measured DC-coupled EM27/SUN interferograms. It includes a DC correction of the interferogram, a dedicated phase correction scheme for double-sided interferograms and several quality control tests (e.g. testing for the presence of out-of-band artefacts). The lookup table for cross-sections used by PROFFAST is created on the basis of HITRAN spectroscopic line lists: For H2O, CH4, N2O, HITRAN 2008 line lists are used (in case of H2O including some minor empirical adjustments), for CO2 and CO HITRAN 2012 line lists are used. PROFFAST uses the solar line list compiled by Geoff Toon, JPL, for GGG2014. In contrast to the TCCON GGG2014 processing, the empirical airmass-independent and airmass-dependent post-calibrations are applied species-wise including molecular oxygen. Thereby, the Xair equivalent provided by PROFFAST is on average normalised to unity, while it remains an un-calibrated intermediate result in GGG2014, which calibrates only the Xgas results. The PROFFAST approach of calibrating Xair is transparent for users, as the calibration factors can be directly related to deviations of the spectroscopic band intensities, and gives the user a more sensitive diagnostic tool at hand, as airmass-dependent artefacts in the reported quantity are also reduced."

The Xgas values, which are calculated using GFIT are scaled to the WMO standards using calibration factors (airmass dependent and independent). The calibration factors were derived from dedicated campaigns at different TCCON sites. However, the XCO_2 and XCH4 products from the EM27/SUN are bias-corrected based on the scaling factors calculated from the extensive

COCCON development. The residual bias between the non-linearity corrected TCCON data and the EM27/SUN might be due to several reasons, such as (1) effect of non-linearity correction of the TCCON data on the retrieved Xgas values, (2) bias of the EM27/SUN due to the imperfect instrument specific scaling factor used which has been determined independently prior to this study from long-term intercomparison measurements performed at the KIT TCCON site. (3) bias of the Sodankylä TCCON station which may come from the imperfect use of the airmass independent calibration factor derived for the global TCCON. As more comparisons of the COCCON spectrometers with respect to the TCCON stations takes place in the future we need to verify the site-to-site bias of the TCCON with respect to the COCCON spectrometer. As mentioned by the referee as well, the use of the low-resolution spectrometer as a travelling standard will be very useful in this respect.

The HR125LR measurements (low-resolution measurements with the Bruker IFS 125HR) are also affected by the non-linearity of the InGaAs detector. As this is not our standard data set, no non-linearity correction has been done. It is used to check the resolution dependent effects of the measurements as discussed in the paper. For example, the intercomparison results help us to verify the bias in XCH4 during the springtime (as high as 0.01 ppm) which is due to the large difference between the a prior profile to the true atmospheric state and the influence of the different AKs between the low- and high-resolution spectrometers. The XCO_2 bias between the EM27SUN vs the HR125LR is 0.56 ppm. If we apply similar non-linearity correction to the HR125LR data, then the expected XCO_2 bias is about 0.06 ppm, XCH4 is about 0.006 ppm and XCO is about 4.5 ppb.

3) Averaging kernels, low res versus high res

The differences in performance and measurement information content for the low-resolution measurements (125LR and EM27) are of high importance, especially with respect to the fact that some of the forthcoming remote sensing satellite based systems may be operated at lower spectral resolutions, i.e. being more similar to the LR measurements in terms of spectroscopy. Also some of the differences observed between AirCore truth, high resolution and low resolution measurements, as discussed under 2), may be interpreted in this respect. The performance of these lower resolution systems are of importance also with respect to the knowledge of spectroscopy and potential future research needs there. But mainly the differences in performance between LR and HR (also with respect to AirCore "truth") may be better interpreted if the differences in their respective AK would be discussed and addressed - at least to some extend (and potentially also differences in GFIT and ProFIT a priori profiles if any – see also 2)). This could be done e.g., and ideally, in a corresponding figure to Fig. 17 (or adding such results to the latter).

Authors' response:

We completely agree to the referee's point of view. It is very important to identify differences caused by the resolution differences of the spectrometers and their related effects. As a result we have kept the HR125LR comparison results in the appendix. The AKs of the TCCON and the EM27/SUN for all SZA are shown in Figure 6 of Hedelius et al. (2016). The retrieval of the low and high-resolution measurement data sets were done using the TCCON a priori as the common a priori. As discussed in point 2 above, other differences between the PROFFAST and GFIT codes have been added to the paper.

Referee:
4) Precision of TCCON versus COCCON

The striking difference in measurement noise (Fig. 18) between TCCON and COCCON EM27 is not much discussed it seems. Is this feature a question of retrieval algorithm performance (or/and applied constraints therein), spectral resolution (like the more stable LR with respect to HR measurements), or is it solely related to instrument noise?
Authors' response:
The EM27/SUN shows a lower scatter as compared to the TCCONmod due to the low noise resulting from the averaging of the individual measurements. The individual measurement time for each instrument is provided in Table 2 in the paper. Within the period of five minutes, it is possible to average five measurements for the EM27/SUN data set. This is equivalent to 50 scans in total in 5 min with 10 scans for each measurement. Whereas a maximum of only two measurements are possible for the TCCONmod data set. This is equivalent to a maximum of 8 scans in 5 min with 4 scans for each measurement.
The scatter in the IRcube and Vertex70 is comparable to the TCCONmod due to the averaging of the similar number of measurements within the five minutes time interval.

Additional minor comments:

Referee:
p. 2, l.36ff: "Carbon dioxide (CO2) and methane (CH4) are the two main components of the carbon cycle of the earth's atmosphere. They absorb and retain the heat in the atmosphere causing the greenhouse effect and global warming" – I would remove "greenhouse effect" here. Otherwise, water vapour would have to be mentioned in this context.
Authors' response:
Done

Referee:
Section 5.3: I think it would be would be very helpful to introduce here the logic of the following sections (titles content) to guide the reader through what is coming, apart from potentially shifting the AirCore results and non-linearity sections first (see above).
Authors' response:
We agree. We now begin the results section with a description of the non-linearity issue with our reference measurement "TCCON". This is followed by a comparison of the standard and non-linearity corrected TCCON results with respect to the AirCore.

Referee:
Figure 1 to 4: It would be helpful to use the same axis limits in the scatter plots.
Authors' response:

Done. Note that these are now Figures 7 – 10 which are made for the TCCONmod case with the same axis for the scatter plots for each instrument except for the XCO_2 plot for the LHR instrument where the scatter is significantly higher as compared to the other instruments.

Referee:
Figure 16: A horizontal dashed line or similar at zero would really help to interpret the results. Especially since the range of the y-axis has to be quite large.
Authors' response:
Done

Referee:
Section 5.4.4. I think here it should be highlighted and maybe discussed in context of the discussion on biases between TCCON, TCCONmod, LR and EM27, why the EM27 is so much closer to one than TCCON (see also point 2) above).
Authors' response:
We have added an explanation in response to point 2 above and we included it in the last paragraph of section 5.5, see lines 631 – 642: "The Xgas biases between the low-resolution test instruments and the TCCONmod data sets as reference may be due to effects such as different responses to a priori profiles, interfering species in the retrieval windows or different averaging kernels. Furthermore it is important to note that, TCCON uses a network-wide constant scaling factor to scale its Xgas values to the WMO standards. The scaling factors specific to each gas for TCCON had been determined from several measurement campaigns where vertically distributed measurements of the gases were performed from airborne platforms using WMO calibrated instruments. The EM27/SUN uses species dependent scaling factors for XCO2 and XCH4 which had been calculated from long-term intercomparison measurements performed at the KIT TCCON site. However, no such instrument specific calibration factors were applied for the other instruments and also for the XCO results from the EM27/SUN measurements. This also contributes to the residual bias which is observed in this intercomparison result. The biases which are purely due to resolution differences are addressed by performing low-resolution measurements with the same TCCON instrument. These data are then used for an intercomparison relative to the TCCON as well as for the intercomparison with other low-resolution test instruments. Further details of the intercomparison results are given in appendix C and D, respectively."

Reference:

1) Hedelius, J. K., Viatte, C., Wunch, D., Roehl, C. M., Toon, G. C., Chen, J., Jones, T., Wofsy, S. C., Franklin, J. E., Parker, H., Dubey, M. K., and Wennberg, P. O.: Assessment of errors and biases in retrievals of X_{CO2}, X_{CH4}, X_{CO}, and X_{N2O} from a 0.5 cm^{-1} resolution solar-viewing spectrometer, Atmos. Meas. Tech., 9, 3527–3546, https://doi.org/10.5194/amt-9-3527-2016, 2016.