Dear Editor,

Thank you for the current handling of the manuscript, please find our responses to your and the reviewers criticisms below. We have kept your comments in black, our responses are in blue, and specific changes are in underlined blue.

Dear Authors,

The referees have given their feedback on your manuscript, with 2 referees raising substantial criticism and insisting that publication of the manuscript in the present form is not possible. They recommend major changes for reaching publication quality.

While your investigation and the results are certainly interesting for the readers of AMT and especially the remote sensing community, I agree with the more critical reviewers' views that there remain major methodological questions which need to be addressed before the science case can be made. Likewise, the manuscript lacks clarity and is sometimes difficult to follow. Significant changes to the paper are therefore required before publication in AMT can be considered. Consequently, I decide that the paper can be reconsidered after major revisions.

Pleaase take all referee suggestions into account when improving the paper. Scientific significance and scientific quality are two core criteria for publication in AMT. It is therefore important that the methodology is improved and well explained. As outlined by the referees, there are some major questions to be addressed:

We hope our latest changes answer the current criticism.

1) How representative are the data that you show and do they support the conclusions? Why has the investigation been restricted to only two days at two different sites? Why does one day only cover less than two hours of measurement? Would observations at other network sites lead to the same or similar conclusions? Your study would gain much from including more sites and extending the period of observation and this is urgently recommended. It is difficult to understand why the investigation does not include more sites and a more significant time span in the first place.

We agree with this assessment and have now expanded the amount of data used, specifically we have expanded the amount of data used at the Ascension Island and Tsukuba sites to capture variations over the course of a year, represented by >1000 individual spectra in each case. We have also added results from two additional TCCON sites, Ny-Ålesund and Darwin, both with significantly different local conditions.

2) It is claimed that 'significant' biases between databases are observed, but this conclusion is partly due to ignoring database uncertainties. If I am not mistaken, methane line intensities in

HITRAN already have 20% relative uncertainties, for example. Given these high values, it is rather surprising that the agreement between databases is better than a few percent. A comparison of databases, such as in this manuscript, certainly needs that related uncertainties are discussed. Moreover, a bias is significant when it exceeds 2 or 3 standard uncertainties (depending on the desired probability level). Given the TCCON sigma values in Tables 4 and 5 and assuming a similar sigma for the other databases, the observed differences are not statistically significant even if the intensity uncertainties in the databases are neglected. But this is not obvious without attributing actual uncertainties to the observed bias values.

Thank you for these comments, we have decided to overhaul how we represent our biases. We now use Absolute mean errors and Root mean square errors with respect to the standard TCCON methane retrieval window from GGG2014. These values are normalized by the retrieval errors in this standard window, which include noise and forward model errors (etc). We therefore define any biases as being significant when they are larger than 1, thus indicating a bias larger than the error associated with the standard TCCON retrieval. Any values lower than 1, we assume as insignificant.

3) While the discussion of the 13-CH4 isotopologue takes a lot of space despite the conclusion that meaningful isotopologie retrievals cannot presently be made, it is not discussed quantitatively how much the inclusion of the new TROPOMI window possibly improves the isotopic retrieval and how the very different averaging kernels impact on determining the delta.

Thank you for this comment, we have now generally reduced the amount of space given to 13CH4 retrievals, however we do retain this analysis and have expended upon it (please see the new section 3.4). We discuss how 13CH4 retrievals using the TROPOMI window are far more constant and show less bias away from the accepted tropospheric average than 13CH4 retrievals using the standard window.

4) Presentation and discussion of the CH4 bias sensitivity (towards other parameters, such as water column, SZA, ...) is lacking clarity.

We have fundamentally overhauled this section (please see updated section 3.3), it has been cut down and simplified and is now easier to follow.

- Table 8, for example, displays a coefficient of determination and not a linear relationship and I would expect that it is R^2 between the bias (y) and the vapour column (x) (and not vice versa). The use of R^2 as a figure of merit should be explained in more detail. While a value of R^2 > 0.5 indicates that a variability of a parameter y can primarily be explained by the variability of parameter x, a value of 0.2 still indicates that y somewhat depends on x and that there is an interference.

This table has now been removed, and we use Pearson's correlation coefficient as opposed to the R^2 values to determine the existence of a linear relationship.

- What does it mean when you write (p 18) that relationships in Table 8 are unlikely to be purely down in water vapour?

This particular line has now been removed, however we do discuss in our updated section 3 how similar linear dependence is observed for water vapour and temperature variations. What we mean is that it may be difficult to completely disentangle the impact of water vapour and temperature dependences on the retrieval biases, since both water vapour and temperature and climb and descend during the course of a day.

- The conclusion at the end of the paragraph is also not clear and the claim that 'As with Table 8 the results from window 1 show the weakest relationships' is not entirely true. It seems that TCCON window 3 has almost always the lowest R^2, indicating that it is most consistent with TCCON window 4 - SEOM window 1 comes next.

This particular section has been removed.

How do you distinguish between systematic and undesirable correlations (due to spectroscopy, for example) and accidental correlations (spatial variation of the CH4 column that possibly correlates with SZA, etc.)?

We have tried to minimize such possibilities by performing the analysis using data over a whole year, and including 1000s of individual measurements. So that accidental correlations are averaged out over the long time period.

Finally, how do you explain that the bias depends on the water mixing ratio? Is there any hypothesis?

In our new section 2.4 we discuss how bias can depend on water vapour variability, largely due to differences in the apriori profile, the reality and the fact that GFIT is a scalar fit algorithm as opposed to a profile fit algorithm, potentially enhancing errors in the profile shape.

5) The overall presentation should be improved.

Besides these points that need thorough consideration, there are a few minor comments that should be looked at:

- The abstract talks about specific TROPOMI windows (plural form). In the text, there is only one such window, apart from the fact that the same window is given a different name when it concerns another isotopologue.

Thank you, this has been corrected.

- Bias values such as in Tables 4 and 5 should be given with sign and with uncertainties (standard deviation or error of the mean). Otherwise the numbers are difficult to interpret, especially since the numbers of digits shown falsely imply a very small uncertainty.

We found that the original set of metrics were confusing for the readers, and have removed these from the paper and have focused on two, identifying the bias w.r.t the standard TCCON methane window, normalized by the noise. These metrics are no-longer included in tables, but shown via bar charts which indicate relative differences more clearly.

- Why has there been only one hour of measurements been selected for Figure 3? Are there no other measurements on the same day? For studying a SZA dependency days with better coverage would be more favorable. Figure 3 also is the only instance in the paper, where windows 1 to 4 are designated by the centre wavenumber. This is confusing. Also why do many measurements come in pairs with gaps in between?

This figure has now been removed and replaced.

- Tables 4 and 5: it seems that GEISA yields the highest consistency in the inter-window comparison. This is an interesting point that requires diccussion.

This metric has now been replaced, and we do discuss instances of consistency in the databases in the new version of the paper.

- Section 3 lists a lot of results, but very little conclusions are drawn from the derived quantities. Often the message is missing. What does it mean when you write 'Windows 5 and 6 vary in their sensitivity, with the different spectroscopic databases showing similar sensitivity in window 5, but very different ones in window 6.'?

We have drawn further conclusions at the end of section 3, identifying that the variability of local conditions causes larger biases between spectral windows and spectroscopic databases.

- I am not convinced that there are 'significant' differences between different windows used for the retrieval of methane isotopologues.
Taking sigma_window of ~ 1.5 - 4 ppb for 13-CH4 as an uncertainty of measurement, an uncertainty on the order of 75 - 200 ‰ can be derived. This is not so much different from the observed variability in Table 7. Again, your definition of significance needs to be explained.

We have updated our definition of ‘significant’ differences, by normalizing biases by the retrieval noise. We therefore assume that bias values that exceed the

- One of the delta values in Table 7 is not compatible with the bias values shown in the Tables. Please check all values carefully.

This table has now been removed, and new values have been checked.

- Explain the choice of the 4th TCCON window as a reference. Wouldn't it be more consistent to use some average of TCCON windows as a reference? Or can the other windows be neglected as they are much narrower? The reader would likely be interested to know how each of the three windows contributes to the overall CH4 data product.

We agree with this comment, we now use the standard TCCON methane window from GGG2014 (a weighted average of the three standard windows) as our reference. We explain in the paper that this standard is a weighted average of the other windows, but TCCON literature does not elaborate on how these weighting are derived.

- The comparison of Tables 8 and 9 should not start by mentioning a close agreement, because the Ascension data from August are showing a quite contrasting picture. It is striking that the discussion does not discuss this difference.

These tables have now been removed from the paper.

- The correlation plots C9 etc. show discontinuities, most prominently for the 13C containing isotopologue. This is not really discussed.

This section has been removed, we found this was a bit of an information overload.

- Why is the Tsukuba April 2016 the reference in section 3.5.2 and Appendix C ? It has the smallest number of data points.

We have significantly overhauled the Tsukuba data, so this comment no longer applies.

- One of the most remarkable results is that inclusion of window 5 seems to reduce the uncertainty of 13CH4 retrieval. Could this improvement somehow be quantified?

We found the retrieval errors generated by TCCON for 13CH4 in the 6076 cm-1 window were double those of the 4265 cm-1 window, which will go a long way towards explaining the improvement in the d13C retrievals.

- Insignificant digits in Table 4 should be removed.

We now show all results to 3.s.f

Reviewer 1

Dear Reviewer,

Thank you for your updated comments, the paper has now been significantly overhauled due to reviewer comments, but our responses are below in blue.

The re-submitted manuscript is now much shorter and more readable. I recommend publication subject to a few technical corrections:
L81: d13C metric which is based on a ratio of these isotopologues (it is not a simple ratio)

Thank you, we have updated this sentence and removed references to ratio.

L87: the abundance of 13C is 1.1% and D is only 0.01% so the abundance of 12CH4 is closer to 99% than 98%

Thank you, corrected.

L313: This referal to Appendix A is perhaps to brutal, could you add a setance summarising the expalnation referred to App A?

Appendix A has been updated in the paper overhaul.

Table 8 caption: as not at? The meaning behind this caption is not clear.

The original Table 8 has been removed in the updated paper.

L446: This would not be a not a significant .... please correct.

This line has been removed due to the paper update.

Reviewer 2

Dear Reviewer,

Thank you for your comments, we have kept your original comments below, and we have added inline responses in blue.

The authors did a thorough job considering my comments from the previous stage of review. I still have some substantial concerns that I recommend addressing before publication:

Major comments

P11, section 3.1: I think the averaging kernel discussion falls short of what the problem for delta-13CH4 retrievals is. The averaging kernels for 12CH4 and 13CH4 are substantially different. That implies that a 13CH4 “anomaly”, for example, close to the surface would be weighted differently in the total column than a proportionate 12CH4 anomaly. The difference in weighting would result in a spurious anomaly of the calculated delta-13CH4. My question from round 1 was what the magnitude of this averaging kernel effect on delta-13CH4 is and whether the effect is actually smaller than the targeted 1 permille (or how the effect is related to the huge variability in Table 7). I would think, that this presents a fundamental challenge in total column remote sensing of isotopologue ratios limiting the usefulness of delta-notations for that purpose. I recommend placing a caveat related to these considerations.

Thank you for this comment, we however have removed the Averaging Kernel discussion from this updated paper, since we were not sure as to the benefit of its inclusion and we have reduced the emphasis on 13CH4 retrievals in this update.

P17, Table 7: The consistency among the delta-13CH4 retrievals is awful and nowhere close to the required quality. What is the retrieval uncertainty of the delta-13CH4 values, what is the daily standard deviation? I actually recommend postponing a discussion of spectroscopic effects on delta-13CH4 to a stage where the retrieval procedure for 13CH4 will have been consolidated.

The retrieval uncertainty on 13CH4 retrievals is typically around 18-25% of the total column (so very large), this is a mean value over the entire available dataset in the new paper. However in the updated paper we provide d13C values averaged over thousands of retrievals, so the uncertainty on these values should be significantly reduced. We indicate this in section 3.4.

P18, Section 3.5: I have some concerns wrt. tables 8 and 9 and the related discussion:

Tables 8 and 9 have been removed in this current iteration of the paper.

- The coefficients of termination are substantial in many cases (I would consider r^2>0.2 substantial). So, I would object the statement that "Table 8 generally indicates that there is limited or no relationship …".

Thank you for this comment, however we have removed the use of coefficient of determination and use Pearson's coefficient instead, and we have overhauled this section of the paper.

- The coefficients have been calculated per day per station and it is the coefficients for the concentrations per se, not for concentration differences (e.g. wrt. a reference window). This probably implies that the parameter range (in particular for water vapor) is very small questioning the robustness of the calculated values.

Thank you for this important comment, we have now massively expanded the amount of data we are using, and make these comparisons against a reference, so these values are now much more robust.

- I also wonder about coefficients >0.5 for window 4. It essentially indicates that there is a very strong correlation which, however, could be due to spectroscopic errors or due to an actual correlation in the atmosphere on the particular day.

Thank you, similar to the comment above, we now make these comparisons using data over a whole year, so we have reduced the dependency on correlations on a specific day. However, w.r.t. the spectroscopic errors, these are the types of errors that we are trying to identify by making these window comparisons.

- I recommend evaluating a dataset that has substantial dynamic range in H2O (and SZA). Why did the authors actually choose Tsukuba and Ascension – there would be stations with more water variability (e.g. Darwin) and more seasonality (e.g. Park Falls)?

Agreed, we have now included two more stations, Darwin as suggested for water vapour, and we also included Ny-Ålesund as a contrast, with low water vapour and high SZAs.

- I would also recommend instead of showing tables of coefficients showing correlation plots that illustrate the parameter ranges and visually depict potential correlations. For example, there is a longstanding discussion of the airmass dependent bias in TCCON XCO2 (and related corrections, e.g. appendix E in Wunch et al., https://doi.org/10.1098/rsta.2010.0240, 2011). Is there a similar bias observed for XCH4 and does it dependent on the spectroscopic database and the retrieval window?

Thank you for this, we have included plots visualizing this (see Figure 5), but we found it difficult to determine correlations from this plot, so we have retained tables including r values for reference.

- It might also be useful to pick the window with the smallest correlations as a reference window and then, evaluate differences of concentrations.

We now use the standard methane retrieval window as our reference point for all calculations.

Minor comments

Throughout manuscript: It looks like a various places the revisions have not been carefully checked for correctness of replacements. Please double-check the entire manuscript for syntax and wording issues and make sure that there is a thorough proof-reading. (I stopped marking the most obvious syntax issues somewhere in the middle of the paper.)

Title: remove „isotopologue“. While technically the usage of the word is of course correct, it is misleading in a sense that the reader might expect a focus on the rare isotopologues since when discussing the main isotopologue 12CH4 the word “isotopologue” is commonly not used. Along my previous comments and also consistent with the authors’ revisions, the focus of the paper is clearly on the main isotopologue.

We agree, the title has now been modified to:

On the consistency of methane retrievals using TCCON and multiple spectroscopic databases

P1,L10: “We report …” - Sentence is very complicated and long (and maybe wrong grammar). Consider rephrasing and splitting.

Thank you, we have now changed the abstract and this sentence is no longer present.

P2,L21: “depending on the introduced error” – I do not really get the meaning of “introduced error”: interferences with SZA, water vapor und profile shape are not simple hypothetical errors. They come from the actual dataset.

Thank you, we have removed this section from the paper.

P2,L27: “is not as well understand as is the case for carbon dioxide” – What is the metric for this comparison, is there a reference to be cited? Considering that the year-to-year variability of carbon uptake by the land biosphere is on the order of 10-20% of the anthropogenic emission total and that this variability is not really understood even on the global scale, I would remove such statements. In fact, methane with a single sink process dominating might even be considered the much simpler case.

We accept the generic statement of “not as well understood as CO2” is not useful, we have replaced with:

but the processes via which it enters and is removed from the atmosphere are still not well understood.

P2,L35: GOSAT is not mentioned in the list. Why is that? GOSAT has been operating since the year 2009 delivering probably the best methane retrievals ever available. Due to its wide spectral coverage it allows for proxy as well as full-physics methane retrievals and, GOSAT-2 covers the 4300cm-1 band. So, GOSAT is the gold standard for methane. It needs to be mentioned.

We had not mentioned GOSAT because it does not uses the 4265cm^-1 spectral range, a particular point of focus of this paragraph. However, we have now identified both GOSAT and GOSAT-2 in the paragraph.

P2,L45: “with SCIAMACHY” -> “with SCIAMACHY and GOSAT”

Changed, thank you

P2,L46: “season, assess” – Syntax.

Sentence changed to read as:

Spectra are taken from four different TCCON sites in order to assess the impact of varying atmospheric conditions at different global locations.

P3,L61: The main reason why TCCON is better quality than satellite data is the fact that direct sun viewing safely allows for neglecting atmospheric scattering processes. Thus, the lightpath through the atmosphere is known with much better accuracy than for the satellite sensors.

We have added this additional clarification.

P3,L65: “generating using” – Syntax.

Corrected to generated.

P3,L67: “accurate/stable” – What does “stable” refer to?

Corrected to:

window proves to be as accurate as the standard TCCON windows,

P3,L69 “will may use” – Syntax.

Removed

P3,L75: “Our finding will inform … and are …” – Syntax.

Sentence has been removed.

P3,L83: “industrial” – What does industrial refer to? Is it the fossil fuel production chain?

Yes, we have changed industrial to fossil fuels to be clearer.

P4,L113: “This synthetic …” -> “These synthetic …”

Corrected.

P5,L124: “Note that all of the above are not …” -> “Note that not all of the abover are …”

Corrected.

P6,L145: “The introduction of 13CH4 …” – This reads like the 13CH4 retrievals were the cause for comparing the spectroscopic databases. I think the study SEOM-IAS is actually much more important for the purposes of this paper.

We have rephrased this paragraph to identify the recent addition of 13CH4 as a point of interest.

P10,L263: “scalar retrieval algorithm” -> “scaling retrieval algorithm”

Corrected.

P12,L302 “line mixing/overlapping” -> “overlapping”.

This section of the paper has been rewritten

P12, Fig.2: There is still substantial noise in the fit residual plots. Are these individual measurements or is it the co-added residuals from various measurements? If the former, would it make sense to co-add residuals?

These are individual retrieval plots, and while we do agree that co-adding would improve the noise, this is not typically done in TCCON retrievals, and we are trying as much as possible to replicated typical TCCON operations.

P13,Fig.3: The data at ~9.55 UTC (X12CH4 plot, upper panel) and ~9.75 UTC (error plot, lower panel) show substantial excursions. Why is that?

This figure has been changed in the new version of the paper.

P19, section 3.5.2: Following the recommendations in round 1 of review, the authors essentially moved all figures and tables related to that section into the appendix but kept a quite detailed discussion. If the discussion is really required and needs to stay in the manuscript, it would be good to have at least one summary figure or table in the main manuscript where the reader can refer to when going through the discussion.

These sections have been removed from this version of the paper.

P21,L450: The degrees of freedom for a profile retrieval of CH4 from TCCON will be small and thus, there will be a residual dependence on the a priori shape. I think a statement claiming that such an algorithm “would therefore not be subject to the methane profile errors” is too optimistic and hard evidence would be needed to convince me of that (quite general) statement.

We have changed this to read as:

Be less subject to methane profile errors

Reviewer 3

Dear Reviewer,

Thank you for your review, we have keep your original comments below in black and our responses in blue.

+ include more sites and motivate the choice of the sites, the authors should demonstrate that they encompass the range of possible situations for atmospheric methane, being therefore more useful to, e.g., the TCCON and satellite communities

We agree with this comment, we have included two more TCCON sites that represent a wide range of conditions such as SZA, temperature and water vapour. We elaborate on our choices in section 2.1.

+ do not limit the data sets to a few consecutive measurements, a day, or one or two months, especially if the underlying motivation is unclear or not provided. The resulting metrics will be more representative and robust

We agree, we have now included a significantly larger number of measurements from each site considered, typically >6000 over the course of a year.

+ try identifying more suitable spectral intervals for the 13CH4 retrievals; new spectroscopy has become available and more relevant windows might potentially be selected

While we agree that there is certainty room to improve on choosing spectral windows for 13CH4, we do not feel this paper is the correct study for determining the best windows for 13CH4 retrieval. This should be elaborated on in a separate study.

+ include model simulations or results (e.g., from Buzan et al., AMT, 9, 2016) in order to estimate a reasonable value for delta13C representative of a column measurement, this will be less speculative than the current approach relying on in situ surface measurements.

We agree that a more reasonable estimate of d13C for a column measurement could be used to make comparisons of d13C statistics against. However, we make the argument that most of our results indicate that we are not yet calculating consistent enough results to make this comparison worthwhile. We will therefore propose for future work, once more consistent results have been achieved across the spectroscopic databases, to make comparisons against a total column d13C value.