

Response to Referee Sergey Assonov by Umezawa et al.

We thank the reviewer for his thorough reading and insightful suggestions.

Overall, we very much appreciate the reviewer's comments, in particular from the viewpoint of rigorous technical aspects. We emphasize that, although the issues are not finally solved, the attempt of an estimate of the overall measurement offset as presented in this study will help synthesis and modeling studies, at least as a temporal workaround. This is the original and main scope of this study; however, in the future we will continue efforts from participating laboratories for improving compatibility to maximize use of the available datasets. All of the reviewer's comments are helpful for the steps forward, but here in our response and in revising the manuscript, we focus on the original scope of this study.

Our responses are detailed below, in which **Comments from reviewers** and our responses are given in different styles.

Stable isotopes of methane, an important greenhouse gas, are used to understand the budget of methane sources and atmosphere oxidation processes. In particular, isotopes are needed to understand why the growth of atmospheric methane was levelled off in 2006-2007 and later renewed (Nisbet et al., 2016). From 2006 to 2016, $\delta^{13}\text{C}(\text{CH}_4)$ is discussed to be shifted to negative direction for ~0.3 ‰ (Nisbet et al., 2016). In order to interpret the isotope data, first of all one has to obtain reliable, high precision and high accuracy data, to be based on reliable calibrations. The problem is that offsets between labs are larger than the measurement reproducibility of each individual lab, in particular the intercomparison Round Robin in 2007–2016 demonstrated discrepancies up to ~0.6 ‰ in $\delta^{13}\text{C}$ and ~36 ‰ in $\delta^2\text{H}$ (Table 5). In the manuscript, the authors review the situation with calibrations, inter-comparisons and lab-to-lab discrepancies.

All in all, the authors have performed a great work to make a summary of numerous measurements done in several labs, their calibration histories and demonstrated discrepancies between data sets produced. However, it is evident that reliable data sets and even data synchronisation cannot be obtained at the

moment, and much more work is needed. The authors are in the position to make analysis of the situation and make suggestions on further steps necessary to obtain reliable data in the future; all that has to be listed in the abstract and stressed in conclusions. Suggestions and/or recommendations on calibration practices, use of reference materials and design of inter- comparisons would be valuable both for laboratories and also for Global Atmospheric Watch program at WMO. Biannual WMO-IAEA meetings on CO₂ and greenhouse gas measurement techniques are focused on improvements needed, including stable isotope measurements of atmospheric CO₂ and methane (e.g. GGMT-2015, GAW Report No 229).

The authors express the hope that CH₄-mixtures (Sperlich et al., 2016) to be made in the way similar to CO₂-in-air calibration mixtures being produced by MPI-BG (JRAS-mixtures) may be of great help. However, GGMT-2015 has recognised that after introducing JRAS-mixtures, lab-to-lab discrepancies in δ¹³C(air-CO₂) demonstrated by intercomparisons have not been decreased (GAW Report No 229) and much work is still needed. Thus, careful analysis of the current situation with methane isotopes as well as focused recommendations need to be given. If the authors are not in a position to come to concise recommendations, at least they should give a better summary of their work and analysis.

We thank the reviewer for his insightful suggestion. The scope of this study is to collect and present as much information on the currently available data of isotopic composition of atmospheric CH₄ as possible in a clear and concise manner. It is true that we need further efforts toward harmonization of the datasets, but we believe that our current best estimate of the inter-laboratory offset is the first important step. A new round robin comparison is being planned for starting next year, and further analysis of the situation and discussion on efforts in individual laboratories should be advanced. In the present study, we focus on a clear description of the current and historical measurements to help following steps to come. We have reformulated the manuscript to improve readability within the scope of this study, but further efforts in the path forward, for instance to

formulate recommendations for measurements is beyond our scope. Such efforts have to be coordinated independently e.g. at WMO-IAEA GGMT meetings and in line with their associated recommendations.

In general, the manuscript is extremely long, not easy to follow and understand causes of the problems. It can be further optimised by grouping some problems and then addressing these groups. Given that mostly $\delta^{13}\text{C}$ has been analysed, the reviewer focuses on $\delta^{13}\text{C}$ data, the same aspects are mostly valid for d2H data.

First, the reviewer suggests grouping problems related to calibrations as following (some aspects are addressed below in more details):

1. Instrumental effects and raw data corrections. These include (i) cross-contamination (memory) in the mass-spectrometer ion sources; this effect shrinks the $\delta^{13}\text{C}$ values, namely the distance between NBS19-CO₂ used for many calibrations ($\delta^{13}\text{C}$ of 1.95 ‰) and $\delta^{13}\text{C}$ values of samples being around -47 ‰; (ii) consistent use of ^{17}O correction for raw CO₂ mass-spectrometry data, (iii) Kr-effect in continuous flow mode of mass-spectrometry runs (before optimisations) and its magnitude. Notably, the aspects (i)-(ii) are not addressed in the manuscript.

We mentioned the cross-contamination effect in section 5 of the original manuscript, but did not do so concerning the ^{17}O correction issue. We have collected information related to these possible problems from each laboratory and summarized them in an extended table. This table will be added in the revised manuscript. We have also added a new section to present a systematic description on these issues.

2. Consisted use of Reference Materials (RMs) and data management. Isotope trends can be understood in a reliable way if, and only if data are correctly positioned on the scale. Absence of any drifts can be demonstrated by re-calibrations vs. reliable RMs, to be repeated on regular basis. (The reviewer has not found much information on repeated calibrations.) First, one has to pay attention to the fact that $\delta^{13}\text{C}$ values of RMs have been revised in the past

(e.g. Coplen et al., 2006) and updated values have to be taken. This implies the need for data management & data archiving in the way allowing data reprocessing retroactively. Second, there is no reliable RMs aimed at $\delta^{13}\text{C}$ in methane, and such dedicated RM(s) is urgently needed. In particular, LSVEC, the high-level RM aimed at $\delta^{13}\text{C}=-46.60\text{‰}$ introduced for data normalisation by (Coplen et al., 2006) is found to be unstable. The scatter in $\delta^{13}\text{C}$ observed on different LSVEC aliquots is $\sim 0.35\text{‰}$, with the LSVEC value drifting in time due to adsorption of air-CO₂ (Assonov et al. in GGMT-2015, GAW Report 229). Calibrations based on NBS19-CO₂ & NBS18-CO₂ may be biased due to cross-contamination effect (see above) and dedicated instrumental tests have to be performed.

We agree with the reviewer regarding the importance of RMs and data management. We have added a description of repeated measurements of reference materials where available, and refer to the technical papers from each laboratory for more details. As for the revised $\delta^{13}\text{C}$ value of RMs, it is true there are laboratories that have not taken into account such corrections. Those laboratories that have ongoing measurement programs over years might consider retroactive data correction, but there are also laboratories that produced a significant amount of data but already closed these activities by now. In these cases, access to raw measurement data and re-evaluation may be laborious and in some cases even impossible. We agree that re-evaluation must become routine for future revisions of RM values. We, the greenhouse gas isotope community, can recommend proper corrections, but the datasets are made available by individual laboratories under their own responsibility. We consider that step-by-step efforts to spread such information broadly is the only way to improve compatibility with updated information for the RMs. We are also concerned for the lack of RMs for $\delta^{13}\text{C-CH}_4$ and $\delta\text{D-CH}_4$ and we have therefore extended discussion in the revised manuscript.

3. Practical calibration approaches and inconsistencies. Careful calibrations and regular re-calibrations is a must. Contrary, a common practice is to transfer calibrations from one lab to another (e.g. by transferring characterised CO₂ or CH₄ gas, or by performing calibration measurements for another lab, several examples are listed below). This practice cannot be recommended as it may,

and often does bring to unpredictable biases (e.g. due to inconsistent use of the 17O correction) and also precludes correct data management (corrections to be applied in a consistent way and/or corrections to be applied retrospectively). This and the fact that revision of $\delta^{13}\text{C}$ values for RMs took place in 2006 (Coplen et al., 2006), this demands for a careful revision of calibrations in each lab.

- 4. Routine protocols and operating procedures have to be established in each lab and followed, both for calibrations (more critical) and for sample analysis. Any deviation from established protocols (e.g. due to rotation of personnel) may bring to unrecognised bias.*

We thank the reviewer for these comments. We admit that the current situation is undesirable. However, we emphasize that the information that is presented in our paper is the result of time-consuming scientific efforts of individual laboratories. We acknowledge the historical measurements and hope to share the current status broadly, even if severe problems cannot be solved. The problems in scale transfer should be addressed through upcoming comparisons and relevant discussions, together with efforts in each laboratory for a better and accurate link to the VPDB and VSMOW scales. Regarding the routine protocols, an established management has been difficult in each laboratory, because measurement techniques for the isotopic composition of CH₄ have been in development over the last few decades. At the same time, even once a certain measurement technique was established, it has been hard for each laboratory to maintain the same measurement circumstance over the years, given that the measurement system is very complex. Therefore, the reviewer's point is still an issue for the path forward.

In order better summarise all the aspects related to calibrations and make analysis, the reviewer suggests making a large table, listing all the labs and providing columns corresponding to each problem and how it has been addressed. This may be helpful to visualise the situation, adopted calibration approaches and inconsistencies. Without a careful analysis and guidance, numerous examples given in the manuscript may result in some misunderstanding and misleading.

We appreciate this suggestion. We have prepared an extended table so that readers can better understand the problems and possible causes of the offsets. The table has been attached at the end of this response.

Second, there is a common misbelief about inter-comparison activities and round robins. These cannot replace factual calibrations (and regular re-calibrations) and also cannot help bringing data on the VPDB $\delta^{13}\text{C}$ scale in reliable way. The inter-comparisons (e.g. round robin conducted during 2007–2016 demonstrate large discrepancies, up to 0.6 ‰ in $\delta^{13}\text{C}$ and 36 ‰ in δD (Table 5) and indicate problems only. As one cannot exclude calibration's biases and/or drifts, non-stable instrumental effects, change in lab 'operation procedures and other problems, lab-to-lab discrepancies in may differ over years. Systematic understanding of inter-comparison results is hardly possible without dedicated design of inter-comparisons; this has to include synchronised use of RMs, synchronised corrections for instrumental effects and synchronised ^{17}O correction. In particular, inter-comparisons have to be done on the same material (same samples or sample archives, or artificial mixtures) and preferentially addressing 1-2 effects (e.g. Kr-related baseline, calibration bias).

We thank the reviewer for bringing up the important point. As the reviewer says, inter-laboratory comparisons help to harmonize different datasets in a work-around fashion, but pinpointing the merged dataset explicitly on the VPDB/VSMOW scale is different. Accurate anchoring to the standard scale is of course critical in the measurement community to achieve high compatibility. However, our attempt in this study, assessment of the overall measurement offsets, may help data users to some degree since, for instance, a model can be tuned or offset to reproduce a harmonized dataset for a specific scientific topic. This study focuses more on this aspect, current best synchronization of the datasets.

Third, data management and uncertainty propagation. Similar to air-CO₂ isotope data, archiving all raw mass-spectrometry data as well as details of all calibrations can be recommended (e.g. see GAW- 229 Report, GGMT-2015). Then

uncertainties can be propagated, in order the uncertainty budget to be used as a tool aimed to demonstrate critical steps and improvements needed. However, the messy situation with calibrations, data management, corrections etc, all that precludes correct uncertainty propagation scheme.

We agree to the reviewer, and this issue should be addressed in upcoming discussions after we organize the current situation through this study and other opportunities.

Below some more details are given, starting with the problems related to calibrations:

1. *Cross-contamination (memory) in the mass-spectrometer ion source is known to shrink the $\delta^{13}\text{C}$ distance between sample gas and mass-spectrometer reference CO₂ gas. The magnitude of cross-contamination effect on MAT252 is reported to be up to a few 0.1 % (Verkouteren et al., 2003a, 2003b). In particular this is relevant to pre-2003 works (e.g. $\delta^{13}\text{C}$ measurements at MPI-C are done before 2003) when no specific measures have been taken such as Tantalum slits in the ion source and optimisation of ion source tuning.*
2. *All $\delta^{13}\text{C}$ data are obtained by correcting raw CO₂ mass-spectrometry data for 17O contribution, this is so-called the 17O correction. The 17O correction after Craig (1957) modified by Allison et al. (1995) and the one after Santrock et al (1985) has been used for many years; later the 17O correction was re-determined by Assonov & Brenninkmeijer (2003) and this correction has been recommended by IUPAC as the most accurate one and avoiding biases in $\delta^{13}\text{C}$ (Coplen et al., 2006; Brand et al., 2010). Inconsistent use of the 17O corrections (e.g. by transfer of calibrations from one lab to another) and/or comparing data sets obtained in different years and using different 17O corrections may bring to unpredictable bias(es). Notably, revision of $\delta^{13}\text{C}$ values for RMs (Coplen et al., 2006) is partly related to the use of 17O-correction after Assonov & Brenninkmeijer (2003).*
3. *One needs to stress that Kr affects the continuous flow mode of mass-spectrometry only; one may give an estimate of the magnitude and its direction.*

We thank for the detailed explanations.

1. We have added texts for cross-contamination effect.
2. We have added texts about the ^{17}O correction, which was not taken into account in the original manuscript.
3. The Kr effect on CF-IRMS measurements was discussed in the introduction, but we have made this more visible by assigning a new subsection. As has been described by Schmitt et al. (2013), the Kr effect varies with time and measurement settings and depends on CH_4 mole fraction (Kr/ CH_4 ratio). In addition, the effect has not been qualitatively identified on a routine basis at some laboratories. Therefore a systematic presentation of magnitude and direction of the Kr effect is difficult.

The reviewer suggests listing all the cases of calibration' transfer from one lab to another one, again this may be given in a table. Several examples of calibration transfer (not all cases) are listed below:

- ***from MPIC to IMAU (page 6, lines 5-8);***
- ***from IMAU to MPI-BC (p. 7, l. 5-6), I cite “Initially, the GC-IRMS measurements had been anchored to a working standard air calibrated by IMAU.” Note, IMAU has transferred calibration from MPIC, see above.***
- ***from Bundesanstalt für Geowissenschaften und Rohstoffe to MPIC (p. 6, l. 22-23), I cite: “The MPIC $\delta\text{D-CH}_4$ scale is based on measurement of standard gases at the Bundesanstalt für Geowissenschaften und Rohstoffe, Hannover, Germany”;***
- ***from UEHI to AWI, I cite: “The $\delta^{13}\text{C-CH}_4$ measurements follow the UHEI calibration via comparison of measurements of an Antarctic air sample (Möller et al., 2013).”***
- ***from INSTAAR to UB, I cite: “ $\delta^{13}\text{C-CH}_4$ value of $-47.34 \pm 0.02 \text{ ‰}$ is anchored to the INSTAAR calibration”***

These scale transfer information has been summarized in an extended table prepared in this revision.

There are examples of incorrect use of $\delta^{13}\text{C}$ values for RMs.

- ***For UCI' calibration, reference is given to (Rici et al., 2001), and here I cite from Rici : "values of 1.92 and -47.18‰ for NBS-19 and IAEA-CO-9, respectively." There are 2 problems, namely NBS19 has by definition the value of 1.95 ‰ (no revision followed) whereas the latest revision of δ13C values (Coplen at al., 2006) gave for IAEA-CO-9 the value of -47.32 ‰; the difference of 0.14 ‰ is not negligible.***
- ***For TU, Umezawa et al. (2009) takes NBS18' recommended δ13C value as 5:029 ‰ whereas the latest revision by Coplen et al. (2006) gives for value of -5.01 ‰ (the difference 0.02 ‰).***

There might be a misreading of Rice et al. (2001). They validated their scale, which had been maintained over years, by measuring CO₂ produced from NBS-19 and IAEA-CO-9 against their laboratory standards. The measured values against the UCI scale are those cited by the reviewers. Such validation measurements do not always show complete agreement but this case was in reasonable agreement. It is no wonder that the revised value for IAEA-CO-9 by Coplen et al. (2006) was not used at the time of publication of Rice et al. (2001).

As for the calibration at TU, TU no longer has NBS-18 based δ¹³C data open to data users. TU δ¹³C-CH₄ data since Umezawa et al. (2009) are all on the NBS-19 based scale, so there is no inconsistency related to the revision of NBS-18 value.

There may be cases of unexplained drifts and effects of large magnitude, such as one reported for TU. The manuscript refers to Umezawa et al. (2012) and, I cite from this work: "The measured 13C value of the test gas was stable to be (-47.12±0.10) ‰ from the start date of our measurement until April 2008. Afterward, the measured value suddenly shifted to (-46.85±0.09) ‰, keeping the same precision as before. The cause is still unclear, since we had not changed any measurement settings. To keep data consistency, we added -0.27‰ to the measured values after the gap."

Such shifts can happen in reality. Some laboratories have exchanged samples, co-operated co-located samplings and so on to help find these drifting issues. Relevant information has been given in the revised table.