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, δ13C(CH4) 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 δ13C and ~36 ‰ in δ2H (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 be 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 give. 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.

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 δ13C has been analysed, the reviewer focuses on δ13C 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 δ13C values, namely the distance between NBS19-CO2 used for many calibrations (δ13C of 1.95 ‰) and δ13C values of samples being around -47 ‰; (ii) consistent use of 17O 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.
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 δ13C 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 δ13C in methane, and such dedicated RM(s) is urgently needed. In

particular, LSVEC, the high-level RM aimed at δ13C=-46.60 ‰ introduced for data normalisation by (Coplen et al., 2006) is found to be unstable. The scatter in δ13C observed on different LSVEC aliquots is ~0.35 ‰, 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.

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 δ13C 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.

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.

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 δ13C scale in reliable way. The inter-comparisons (e.g. round robin conducted during 2007–2016 demonstrate large discrepancies, up to 0.6 ‰ in δ13C and 36 ‰ in δ 2H (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 17O 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).

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.

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 δ13C 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. δ13C 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 δ13C 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 δ13C (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 δ13C 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.

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 δD-CH4 scale is based on measurements of standard gases at the Bundesanstalt für Geowissenschaften und Rohstoffe, Hannover, Germany";
- from UEHI to AWI, I cite: "The δ13C-CH4 measurements follow the UHEI calibration via comparison of measurements of an Antarctic air sample (Möller et al., 2013)."
- from INSTAAR to UB, I cite: "δ13C-CH4 value of −47.34±0.02 ‰ …… is anchored to the INSTAAR calibration"

There are examples of incorrect use of δ13C 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 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."

S.Assonov,

17-10-2017