This paper describes the extension of a previously published FTIR instrument and data analysis method (Hammer et al., AMT, 2013) to the measurement of $\delta^{18}O-CO_2$. High precision measurements of stable isotopes are arguably very useful to constrain the CO_2 sources and sinks, and FTIR with its multicomponent capability a very attractive approach. The progress that has been achieved with respect to the precision and stability using FTIR for atmospheric measurements are impressive, and the evaluation of the method for $\delta^{18}O-CO_2$ is, therefore, highly welcome. However, to date, $\delta^{18}O-CO_2$ is certainly the most challenging parameter to be measured in this FTIR setup, and it is not surprising that the original assessment by Esler et al. (2000) was very critical. While the paper by Vardag et al. is generally well written and shows promising results for $\delta^{18}O-CO_2$, it lacks rigor, detail and a more critical spirit to be published in its present form. I, therefore, suggest major changes and a final decision based on the reviewed paper.

Major Remarks

1. The approach of using CO_2 derived from two completely different ro-vibrational bands (asym. stretching and combination mode) for isotope ratio measurement is not an established method and should thus be discussed in much more detail with respect to temperature and pressure effects as well as spectral response of the FTIR. Fig.1 is not sufficient, because it doesn't reflect the real-world situation, where the FTIR is largely blind above 2320 cm^{-1}, thus limiting the observation of $^{18}O-CO_2$ to half of its P-branch at best. More detailed discussion on the fitting stability, reproducibility and residual noise characterization is definitively needed. The following are some (but not exhaustive) minimal suggestions:
(i) the range used for CO_2 (as a proxy for $^{12}CO_2$) should be shown in Fig 1, since it is an important part of the spectroscopic evaluation, (ii) give details on all spectral regions used to quantify the different species, and (iii) show a measured spectrum in the domain where $^{12}C^{16}O^{18}O$ can be observed.

2. Give more details on the cross- and interspecies correction. This is a highly critical part, well-illustrated in Fig 3 (e →f) where variations of 25 ‰ disappear yielding a surprising precision of < 0.1 ‰. For $\delta^{13}C-CO_2$, this correction is much less (< 1 ‰) and all critical parameters well discussed in the previous paper by Hammer (2013). Therefore, the authors should add a table including all sensitivities for the $\delta^{13}C-CO_2$ and $\delta^{18}O-CO_2$.

3. Reproducibility is used misleadingly throughout the paper.
There are several, slightly different definitions of this term (e.g. DIN ISO 21748, ASTM E-177, GUM Annex B). However, the main (and very useful) meaning is measurements in "a set of conditions that include *different* locations, operators… on the same or similar objects" (WMO/GAW Glossary of QA/QC-Related Terminology, chapter 2.24). In this work there is no data to determine reproducibility. More specifically, the whole chapter/title 3.2 must be adapted.

4. The mean and SD of the difference of a number of samples (gas cylinders) is not suited to determine whether two methods are compatible (or significantly different). Use an appropriate statistical test. Change this in the abstract and the text.

5. Given the importance of this new method, the description of the instrument and the laboratory setup should be extended, i.e. referencing to previous publications is not sufficient. Also give the most important spectroscopic parameters (e.g. resolution and optical path), figures such flow, temperature and pressure stabilities, as well as the typical laboratory conditions. Explicitly state if the samples where dried (how and at which level). I assume that this instrument is a slightly modified version of a commercially available instrument (Ecotec). Acknowledge the commercial supplier and give a summary of the changes with respect to the latest version which made the $\delta^{18}O-CO_2$ measurements

possible. It's important for a wide range of users to really understand why you are able to revise the original assessment by Esler et al. (2000).

6. Working gas measurements were done at a daily (24 h) basis. Since this is the time scale at which all measurements can be tight to a reference, the Allan plot (Fig. 4) should include data for 24 h.

7. In chapter 2.2.1 it is argued that the 3600 cm^{-1} range has lower temperature sensitivity. However, for isotopic ratio, it is mainly the difference of this sensitivity between the isotopic species that is of importance. The authors should elaborate on this.

8. Chapter 2.2.2 introduces the term "absolute calibration". Do not call this absolute because it is easily confound with spectroscopy as an absolute method. Similarly, the concept of "empirical" calibration is confusing and arbitrary. Both procedures are simple calibrations of physically sound but approximate value given by the spectrum and the fit procedure. In one case you calibrate the ratio and in the other case the individual isotopologues.
   Remarks: (i) I'm aware that Griffith (2012) used these terms but they are still misleading; (ii) I somewhat insist here because in the related publication by Hammer et al. (2013) implies or suggests that the FTIR values are correct (absolute), but different from the international scale (thus not absolute) "As the raw absolute mole fraction determination of the in situ FTIR analyser differs from the internationally accepted WMO scales by up to a few percent". This is not correct and not meant by Griffith et al. (2012) who correctly state that "In reality, the raw FTIR determination of trace gas concentrations is highly precise, but typically uncertain to within a few percent…"; (iii) absolute calibration is a topic in itself, e.g. attempted in detail by Griffith et al. (Analytical Chemistry, 81 (6), pp. 2227-2234, 2009).

9. Chapter 3.1: rewrite this paragraph using precision as a term with respect to the Allan--deviation, and also consider less complicated phrasing to describe the corresponding results. The last sentence is misleading or at least not complete. Reaching a precision of e.g. 0.15 ‰ (30 min avg) is necessary but not sufficient to observe a diurnal cycle of 1 ‰ amplitude. What you really need is a system that is stable enough to obtain a repeatability ("…replicate measurements on the same or similar objects …", GAW glossary) over the observation period.

10. Chapter 3.3: These measurements are very important and as such convincing. However, it seems that the compatibility was assessed based on mean residuals of FTIR vs MS measurement results. As the target of this study is to resolve diurnal changes of delta values, the compatibility should be discussed in relation to individual flask measurements, and would be in the range of 0.5 ‰ for $\delta^{18}O$-CO_2.

11. In view of the above, the discussion section should be reviewed carefully.

Minor Remarks

12. Title: reconsider the explicit statement about the place (Heidelberg, Germany) that the measurements were made. It's largely irrelevant for the content of the paper. Similarly, check whether each of the 25 entries for Heidelberg throughout the text are really necessary.

13. Throughout the text (including the abstract), I would suggest using the term *precision* in the context of the two-sample variance instead of repeatability. This is more common and it leaves repeatability for other circumstances, e.g. repeated measurements of a target and the derived SD.

14. In the abstract, the 10 min reproducibility and the 30 min Allan deviation refer to the same two-sample variance test. Therefore, use a phrase that shows this accordingly, e.g. … the precision, as derived from an Allan variance test reaches x ‰ (y min), x ‰ (y min), …

15. Explicitly state that this is the third of a series of publications on the same instrument; Vardag et al., (2014, AMTD, doi:10.5194/acpd-14-10429-2014) is not yet cited.

16. Step 2: it's not clear what is meant by "measurement" – at what time averaging are the corrections done?

17. Step 2: I see no good statistical meaning of the "residual sum of squares…by the number of measurements". Even less do I see why this statistical value would be adequate to come to the conclusion "… no further concentration dependence".

18. Step 4: what algorithm was used to do the smoothing?

19. Chapter 3.1: cite literature for the Allan deviation; I would suggest P Werle et al., Appl. Phys. B, 57, 131–139, 1993.

20. Avoid "minutely"

21. Page 6505 Line 16: the use of δ-notation for isotopologue (and isotopomer) ratios is common for spectroscopic techniques, no need to introduce „Molec-d18OHitran"

22. Page 6506 Line 20: Give at least some basic information on the applied mass spectrometer.

23. Page 6508 Line 23: FTIR and MS measurements are not only on the same scale but also anchored via identical CO2 reference gases. Please rephrase. Is the MS an IRMS?

24. Page 6510 Line 12: Give the approximate flow applied for flushing the glass flasks.

25. Page 6510 Line 7: As no results are given in Sect. 2.3 the compatibility of FTIR and MS cylinder measurements cannot be assessed … or do the authors refer to Fig. 3?