Response to reviewers of Steig et al, AMTD 6, 10191-10229, 2013

We appreciate the volunteered comments from E. Berman, and the formal reviews from Erik Kerstel and from an anonymous reviewer. We respond to all their comments, point by point, below. Review comments are in red italics; our response is in black.

1) Response to comments by E. Berman.

This manuscript describes the use of a new instrument to measure stable isotopes in liquid water and water vapor, including the 17O isotope. This is a timely and interesting topic which has seen much recent work. I, among many others I am sure, am looking forward to seeing what new questions are unlocked by the ability to more easily measure 17Oexcess. The authors' careful treatments of calibration and noise contributions are applauded. I have several specific concerns and questions regarding the manuscript.

We appreciate the positive comments about our work, particularly when considering that the lead author works for Los Gatos Instruments and so has a financial conflict of interest. Some of the queries here are vague, but all the points raised are addressed fully in our response to the formal reviews. We provide general responses here and refer the reader to our response to the formal reviews for the details.

General comments:

First, Figure 3 shows Allan deviation plots, with the minimum σAllan for 17O-excess of ~8 per meg, indicating that in the limit of long-term averaging, the best precision achievable by the instrument is 8 per meg. However, the reported standard error is lower than 8 per meg for every 17O-excess measurement in Table 2 and Figure 6 (as well as the Results section). The use of the standard error is statistically incorrect if, in fact, averaging of measurements does not reduce the noise associated with those measurements by the square root of the number of measurements averaged. The Allan deviation plot presented in Figure 3 shows that the data indeed do not average statistically below 8 per meg. The uncertainty reported with all measurements should reflect the actual limit of the instrumental averaging as demonstrated in Figure 3.

We appreciate the comment about "standard error" and we agree that strictly speaking, this is not the correct term, since we do not know that the measurements errors are Gaussian over long time periods. Where we do use the calculation σ/sqrt(N), one can easily back-calculate the raw standard deviation since we provide the number (N) of measurements made. Note that the purpose of Figure 6 and Table 2 is to compare results between the IRMS and CRDS methods, and in this context using this calculation (that is, σ/sqrt(N)) is conservative. At the suggestion of Erik Kerstel's review, we no longer refer to the results of this calculation as a "standard error", but instead use the term "standard deviation of the mean" and provide the equation used in the table and figure captions. We also make it clear that the errors may not be Gaussian.

In Figure 4, we simply report standard deviation – there is no ambiguity here, and the results are consistent with the Allan deviation plots. Indeed, they are slightly better, which is not surprising. As the review from Kerstel notes, the Allan deviation at very long integration times is not well characterized; this becomes increasingly difficult because of the possibility of fractionation of the water sample being used, problems that arise with clogging of the capillaries used in our vaporization system, etc. Because the experiment is not perfect, individual vials (e.g. as shown in Figure 4) would have better precision than implied by the Allan deviation plots, if indeed the real distribution were more Gaussian than those Allan deviation plots imply. This shows that such experiments provide a conservative estimate of the uncertainty.

Second, it is reported that a precision of 8 per meg for 17O-excess can be made in ~30 minutes. These precision numbers are indeed impressive, but a discussion of the accuracy of these measurements is also warranted. What is the accuracy? How are these measurements standardized? How much time is required for standardization of these measurements to reach said accuracy? How are instrumental memory effects accounted for?

Similarly, Figure 4 lists precisions for repeated liquid injections. Again, what is the accuracy? Since memory effects are generally far from negligible, the repeat precision (measured from repeated measurements of the same sample) is frequently not the same as the precision of measurements of unknown samples. Specifically, how long does it take (including standardization and accounting for memory effects) to measure the calibrated values presented in Table 2 and Figure 6? What precision and accuracy are obtained?

Accuracy is already discussed at length in the paper, as are our methods for normalization (or "standardization"). Figure 4 does not claim to – and cannot – address this. The only meaningful measure of accuracy is to compare measured, calibrated unknown samples with their known values, as we have done in Figure 6. Regarding the number of measurements and the question of instrument memory, we have added additional information in the paper to make this clear -- see our more detailed response to the formal review by Erik Kerstel.

Are the experimental uncertainties associated with the δ2H, δ18O, and δ17O values for the VSMOW2 and SLAP2 reference waters incorporated into the reported uncertainties for the measurements as reported vs. VSMOW-SLAP? Table 2 lists this for PW and VW calibration, but not for VSMOW2 and SLAP2 calibration.

All of these water standards have been compared with one another and found to be indistinguishable. Although one could in principle place uncertainties on the relationship between VSMOW and VSMOW2 (for example), there is insufficient information to demonstrate that they are different – and in particular, there is no data available with which to asses this the uncertainty for $^{17}O_{excess}$. SLAP2 is demonstrably different from SLAP only in its δD value. The *working definition* of $\delta^{18}O$, $\delta^{17}O$ and $^{17}O_{excess}$, normalized to SLAP and VSMOW are in relation to a deviation from the values of VSMOW2 and SLAP2, as defined in Schoenemann et al. (2013). Any statements about accuracy and precision are only meaningful with respect to those definitions. Errors in the measurements of the standards during a calibration run will of course result in errors in the resulting calibrated measurements of unknowns; that forms the basis of our estimate of the accuracy.

Specific comments/questions

P10209 line 3: It is not clear how the 17O-excess of GISP comparison between ICOS, CRDS, and IRMS is calculated or why the CRDS value is averaged with IRMS values. A more meaningful comparison may be afforded by averaging only those measurements taken with the same technology. For IRMS and assuming that "current" is in the last decade, from Schoenemann et al.: Barkan and Luz = 11 per meg and Kusakabe and Matsuhisa = 41 per meg. From Berman et al.: LSCE = 32 per meg and JHU = 25 per meg. The current manuscript lists an IRMS value of 28 per meg. These five IRMS laboratories give an average 17O-excess of GISP of 27 ± 11 (1σ) per meg. A weighted average may be more appropriate, but I do not have all of the uncertainties *needed to calculate one. CRDS measurements are reported as 27 ± 4 per meg (see comment/question above about precision) and ICOS reports (Berman et al.) 23 ± 2 per meg. That these measurements all agree to well within the uncertainties is very exciting indeed and will certainly aid in the application of measurements of 17O-excess to hydrological and other applications.*

We are aware of no evidence that technological differences should matter, and indeed, the entire point of careful two-point calibration is to eliminate any such differences. The agreement within our own lab between the calibrated measurements on both types of systems supports this. We agree that it is very promising that so many different labs are in excellent agreement. Note that when we say "recent" we are referring to the same set of measurements considered by Schoenemann et al. (2013) to be the most reliable, plus the new measurements reported by us and by Berman et al. Note that we did not mean to imply that there is a problem with Berman et al.'s measurements – they agree very well with ours within the stated errors of both. We have modified the text to make these points clear.

P10209 line 14: . . . $\sigma18$ is the precision of $\ln(\delta18O+1)$. . . Also, in this paragraph you have neglected to reference Figure 7, where the data is displayed.

Thank you for noting this error, and the need for a Figure reference. We have corrected both.

Finally, the distinction between "laser-current tuned CRDS" and "wavelength-scanned CRDS" was a bit unclear. It appears that the latter uses laser-current scanning as well and that the difference lies not in the scanning but in the way that the frequency axis of the spectrum is determined. If this is correct, it might help readers to be more specific about this difference.

Erik Kerstel also brought up this point, and we refer to his review and our response to it.

2) Response to Erik Kerstel (reviewer #1)

The manuscript reports on modifications made to a commercial, laser-based, liquid water isotope analyzer and its subsequent characterization. The instrument achieves very impressive results that are equal to or surpass those of state-of-the-art mass spectrometric techniques in terms of precision and accuracy for all three water isotope ratios and the derived quantities of deuterium- and 17O-excess (d-xs and 17O-xs). It greatly improves on those conventional techniques in terms of measurement time (sample throughput) and ease of use. It also appears to deliver results comparable to those recently described by Berman et al. (2013) for another optical instrument, and it promises a great step forward when compared to its predecessors from the same manufacturer. The manuscript is generally well written. It gives a clear description of the major interest in the measurement of the derived isotopic quantities of d-xs and 17O-xs. Measurement precision, instrumental drift, water concentration dependence, and calibration to reference materials are all described with clarity and mostly with sufficient detail. I recommend publication of the manuscript, after the authors have had the opportunity to clarify the issues raised here below.

We thank the reviewer for the positive comments, and the thoroughness of the review. Below, we respond to the technical questions in detail.

P10192L9 and Article Title: Please use a more descriptive term than "current-tuned", as practically all CEAS and CRDS schemes use laser current tuning of the laser. See also comment with P10201L11.

We appreciate this point. It is true that current (and temperature) are used to tune laser frequency in many applications. The point here though is that a key technical challenge is to bring the frequencies of the cavity mode and the laser into coincidence, and to do this quickly on command. Here we exploit the advantages of tuning the laser frequency to resonance with one of a comb of longitudinal modes, and the way to do that rapidly is with current. This is different than other existing commercial systems (at least), and hence deserves highlighting in the title. To make this point clearer, we suggest changing the title to "Calibrated high-precision $^{17}O_{excess}$ measurements using cavity ring-down spectroscopy with laser-current-tuned cavity resonance"

P10195L22: Correct year of publication is 2004.

Thank you. Corrected.

P10196L2: It may be useful to refer the interested reader also to Aemisegger et al. (AMT 5, 1491, 2012).

We agree. This paper has now been cited.

P10196L12: With the exception of the very recent work by Berman et al. (2013), which is discussed later on, but could be mentioned here already.

We now cite Berman et al. in this introductory section of the paper.

P101999L8: Please explain what is meant with "external precision". There appears to be nothing external to the determination of the measurement precision (=repeatibility) on repeated samples. The same comment with your use of "external precision on P10205L13: What does "external precision" mean here? I am familiar with the terms internal accuracy (instrument compared to itself or to a copy of itself) and external accuracy (instrument compared to an external independent standard), but would like to see an explanation or reference to what you mean by external precision. My guess is you mean that the precision of the isotope measurement is determined by a (long) time series measurement of a single sample, as opposed to a measure of the baseline optical noise (NEAS) of the spectrometer, but I am not sure, and I am afraid this is so for most readers. Worse, some may mistakenly interpret is as a "calibrated precision" (sic) or (an estimator of) accuracy.

We agree with the reviewer that our usage here is unclear. By external precision, we meant in this context the reproducibility of results from independent measurements – after calibration -- of independent samples that have the same composition. To avoid confusion here, we have replaced "external precision" at P101999L8 with: "The reproducibility of the calibrated $^{17}O_{excess}$ of repeated water samples ranges from 4 to 8 per meg…." For line P10205L13, we have replaced "external precision" with "reproducibility" to convey the same point without implying calibrated accuracy.

P10199L26: 1389 nm, considering the wavenumber is specified to four significant digits.

This is correct. We now use 1389 nm.

P10200L8: Please specify whether you use one (N2) or the other (air), and especially if the measurement is immune to changing the matrix (which I expect it is not without modification of the spectral fit).

The reviewer is correct in assuming that the measurement depends on the gas matrix. The commercial version of the instrument provide fitting routines tailored to air and nitrogen. Tests of the instrument as reported here were made with air. We have specified this.

P10200L9: Personally, I prefer seeing the SI unit of hPa (=mbar); 50.0 mbar = 66.7 hPa.

We agree and now use 66.7 hPa.

P10201L11: It is clear that the relatively small Picarro cavity results in a rather large FSR of 600 MHz. As Figure 1 shows, this results in only 2 to 3 datapoints per FWHM of the ro-vibrational transitions. For many readers this may be construed as a very poor sampling of the spectrum. You may want to point out here explicitly that this is still sufficient due to the precise frequency scale: there is very little noise on the horizontal scale of the spectrum (provided the stability of the cavity is not compromised by, e.g., thermal or electrical noise on the piezoelectric mirror translator – see last comment here below). In addition to the almost perfectly equidistant frequency scale, it is important to point out that each datapoint has a well determined frequency resolution corresponding to the width of a cavity longitudinal, TEM00 mode, equal to the FSR divided by the cavity finesse. It is thus interesting to know the finesse of the cavity. Related to this: what is the empty-cavity ring-down time, and thus the effective optical path length? In addition to this, please specify the NEAS (minimum detectable absorption normalized to bandwidth – measurement rate– and path length) for the three different instruments. This would allow a direct comparison of the spectrometer-only to other spectrometers that have been described in the literature. See, e.g., the discussion of different spectrometer figures of merit by Moyer et al. (APB 2008). It is for the time being of little use to be referred to the patent application as the text is not yet public. Although the authors make no such claim, the patent application suggests that the applied techniques are highly innovative. This may be true for the application in a Picarro spectrometer, but this is certainly not the case for the principles themselves. In fact, the "mode-by-mode" excitation of a cavity enhanced spectrometer is at the basis of FS-CRDS as introduced by Hodges in 2004, as well as of OF-CEAS, first published by Morville et al. in 2005. These works should be referenced in the current paper. The work of Hodges in particular may have implications for the defendability of the patent application presented here.

There are several points being raised here. The first is that sampling on an FSR comb looks sparse to the eye. The reviewer clearly appreciates that the sampling is still entirely sufficient. A more complete explanation than we originally gave in the paper is that there are three parts to our statement of a "well defined frequency axis": (1) the very precise horizontal axis, as appreciated by the reviewer; (2) the very precise measurement of cavity loss made possible by CRDS; and (3) the accurate spectral model derived from measurements made with higher frequency resolution. We therefore have a set of data points that are very precisely measured along both the horizontal and vertical axes and a very precisely defined model that fits these data points. The result is that the model parameters are tightly constrained. We have elaborated on these points in the revised paper.

A second set of questions raised in this paragraph concerns the finesse, ring-down time, and line width of the CRDS cavity. The commercial Picarro analyzers show considerable variability (by a factor of 1.5-2) in these parameters, due to variations in mirror coatings and optical alignment. We include numbers for the instrument we used in these experiments.

Ring-down time constant = 22 ms, corresponding to an "effective optical path length" of 6.7 km. Cavity finesse = 44,000, implying FWHM cavity mode line width of 14 kHz. In practice, however, the precision of the frequency axis is likely limited by mechanical stability, as the reviewer understands. Fortunately, 14 KHz is much better than we require so some degradation is acceptable. We cannot claim to know the absolute frequency precision of the FSR comb.

Finally, the reviewer asks us to specify the noise-equivalent absorption spectral density. This is 2.3×10^{-11} cm^{-1} $Hz^{-1/2}$, which is typical of our Picarro's analyzers. Note that since we can now benefit from 1/sqrt(t) averaging for measurement times at least as long as ~1000 s, this means that the noise equivalent absorption at the minimum of the Allan deviation plot is only 7×10^{-13} cm^{-1}. We will include all of these numbers in our revised manuscript.

P10201L17: Again, it should be noted that, even though this may present a "critical innovation" for the Picarro spectrometer, the OFCEAS H2O-isotope spectrometers built by us have been using this peak-area data analysis procedure since their inception in 2003 (mentioned in Kerstel 2004; first instrument publication in 2006 by Kerstel et al. in APB), as well as by the FS-CRDS instrument of Hodges and co-workers, and a number of other instruments. Please provide appropriate references.

We agree that the relationship between integrated absorption and absorber concentration is not fundamentally new (though our specific implementation of it is), and we now cite the suggested references and others.

P10202L11: Although the difference may not be large enough to be clearly identifiable, I am curious: Why the "soft-collision" Galatry model instead of the "hard-collision" Rautian-Sobelman or the Nelkin-Ghatak lineshape models? After all, the perturber (N2) has a higher mass than the spectroscopically active H2O molecule. Moretti et al., JMS 2000, argue for a "hard-collision" model for H20 lines at 1.4 micron.

We appreciate the reviewer's interest in these details. When we collected the high-resolution spectra of the 7200 wavenumber water lines for the L2130 analyzer, we compared them to all the line shape functions discussed in the paper of Varghese and Hanson, including the Rautian-Sobelman hard collision model (the Nelkin-Ghatak model is essentially the same as Rautian-Sobelman). Empirically, we did not get better agreement between experiment and model with the Rautian-Sobelman line shape than with the Galatry function; in fact they both gave very similar residuals. We were able to improve on the Galatry function fits with a more general velocity-dependent model that introduces a term to the line shape that is odd in detuning (both the Galatry and Rautian-Sobelman functions are even in detuning) but the improvement was negligible and does not justify the computational expense of a much more complicated model function. The physical argument favoring the Rautian-Sobelman model over the Galatry model is not compelling. Both of these models treat the effects of collisions on motional line broadening with simplifying assumptions of how the motion of the absorbing species is changed in a collision: the Galatry model assumes diffusion in velocity space from "soft" collisions whereas the Rautian-Sobelman model assumes uncorrelated velocities before and after a collision, due to hard collisions. These extreme cases would seem plausible if the absorbing species is much heavier or lighter than the surrounding gas molecules, respectively. We have water molecules with 18 amu colliding with nitrogen molecules of mass 28, which is not an extreme ratio. Moreover, the shape of the inter-molecular potential is as important as the mass ratio of the colliding molecules in determining whether the collision is "soft" or "hard". It is also a slight overstatement to say that Moretti et al. "argue" for a hard-collision model. They chose to use the Nelkin-Ghatak profile without any further justification than that it includes collisional narrowing, and observed good agreement between their experimental spectra and the model, which is completely reasonable. They did not compare the Nelkin-Ghatak profile to the Galatry profile or any other function other than the Voigt profile and they did not present evidence that the hard collision model is more valid than the soft collision model. The short answer to this point is that we have looked at other line shape models and concluded that there is no significant advantage for our analysis in describing the absorption lines by a different function. We feel it is beyond the scope of our paper to go into these details: clearly, the Galatry profile is straightforward and entirely sufficient for our purposes.

P10205L13: see comment with P10199L8.

We agree with the reviewer here. See our response to P10199L8 above.

P10207L5: To me the fact that the deuterium concentration dependence has not changed is an indication that the concentration dependence is not primarily due to residual pressure broadening effects, but rather to an incomplete accounting for the dependence of the baseline structure on the water concentration (J. Landsberg, "Development of an OFCEAS laser spectrometer for water vapor isotope measurements at low water concentration", PhD thesis university of Grenoble and university of Groningen (in preparation).

We thank the reviewer for his suggestion, and we agree entirely that the deuterium concentration dependence is an indication that our spectral model is imperfect. We have added a line to this effect. The reviewer may be right, but it would be premature to speculate further at this stage; considerably more work would be required to further improve our already much-improved spectroscopy. We will be interested to read this thesis when it appears.

*P10207L17: For sake of clarity, it would probably help to point out that 2 * 5 * 10 injections are carried out per water sample, for a total measurement time of 5000 sec (1h23 > 30 minutes. . .) (whereas only 40 injections are retained to calculate a "memory-free" result). I understand that no specific calibration strategy (ordering of sample and reference measurements) was applied, but this becomes clear only after reading your comment on the LGR ICOS instrument in the Discussion section of the paper. One common strategy is to bracket the unknown sample measurements with those of reference samples. Such a strategy essentially doubles the required measurement time, as typically the same series needs to be repeated for the reference water, in this case requiring another 5000 sec. Thus, the total measurement time for one calibrated result would ten be equal to approximately 10,000 s, just within the stability time dictated by the AV curves of Fig. 3 for 17Oxs and dD (and only slightly beyond that for d17O and d18O). In this case of "the first calibration experiment" (with 5 samples, of which two are used for the VSMOW-SLAP scale correction and three are treated as unknowns) the total measurement time is 25,000 s. The AV curves then show that at this point the drift of 17O-xs is still negligible, but that the precision of the primary isotope ratios has degraded by a factor of two. The 17O-xs AV curve is actually ill-determined for longer times. Therefore, your experiment, which must have taken circa 7 hours, is also about the longest duration experiment that can be safely performed without having to worry that drift will start to degrade the measurement accuracy. In the calibrated error (accuracy) discussion I am missing a discussion of the propagation of the error of the reference water measurements (standards) on the VSMOW-SLAP scale correction, and* *thus on the accuracy of the isotope values of the unknown samples. The use of the standard error for each sample measurement series (5000 s) is correct in the assumption of normally distributed noise. With a measurement time that approaches or exceeds the optimum averaging time of the AV minimum, this is no longer true. This is in contrast to the situation of the measurements reported in Table 2, as I assume that, e.g., a VSMOW-SLAP calibration was carried out for each of the 6 determinations of GISP that were used to produce the averaged value and its standard error on the first row in Table 2. In this case the use of the standard error appears correct to me, as the mean value of the measurement does indeed become better known with each repeated calibrated measurement.*

We appreciate the reviewer raising this point regarding clarity. The critical issue here is with the calibration runs, in which we use four of five vials of the same water, 10 injections each, and use the data from the last three vials as our data – the first two vials being used to get rid of memory. Furthermore, we *also* use four or five vials of similar water priot to each standard (working standard VW in the case of SLAP, for example). Thus, we are confident that memory does not play a role in our results. We provide these details in the revised manuscript.

Note that because there are as yet an insufficient number of well-defined working standards for $^{17}O_{excess}$, we cannot do the usual thing of bracketing the data with samples that are close in value to the target measurements. We note that the effect of memory appears to be smaller for $^{17}O_{excess}$ than for $\delta^{18}O$ and $\delta^{17}O$ individually, because their effective memory time appear to be identical (and thus cancels out). (This is not the case for δD vs. $\delta^{18}O$.) However, more work is needed to address this – we do not claim in this paper to have addressed the memory issue fully.

We do not determine errors through a propogation of uncertainty in the measurements of the standards. Note that there is no uncertainy in the "true" values of the VSMOW2 and SLAP2 standards, since we are *defining* $^{17}O_{excess}$ with reference to fixed values of these standards. That is, formally, $^{17}O_{excess}$ with respect to VSMOW-SLAP is, at least for now, defined by the calibration against the VSMOW2, SLAP2 standards (Schoenemann et al., 2013). There is no discernible difference between VSMOW, SLAP and VSMOW2, SLAP2.

We estimate the error of the unknowns by using the reproducibility of the final (calibrated) measurements themselves, which provides a measurement of all the combined errors. We do propagate uncertainty in the "true" values of the intermediate reference waters (VW, PW), and here the uncertainty in these measurements is Guassian (based on our original published IRMS data). As the reviewer points out, though, we cannot assume that the combined uncertainty in the unknowns is necessarily Guassian, because the overall measurement for the calibration runs is longer than the tme period for which the Allan variance is well characterized. Thus, the term "standard error" may not strictly apply. However, we emphasize that our calculations are *conservative* with respect to the question we are addressing, which is whether or not the IRMS and CRDS data agree. Clearly, if our estimates of the uncertainty are too small, then there is actually more overlap between the data. We have chosen to still use the standard deviation of the mean calculation ($\sigma/\mathrm{sqrt}(N)$) but we now make it clear that Gaussian statistics not may strictly apply. See also our response to the Berman comments. We have edited the text to make these distinctions clearer, and we have put a note in the table caption as well to avoid any confusion.

P10209L2: bracket at the right place: ". . . per meg (Schoenemann et al., 2013)."
Corrected.

P10209L14: sigma_18 instead of second mention of sigma_xs
Thank you for catching this! Corrected.

P10221FIG1 I find it highly surprising, in light of the cavity FSR being fixed, that the measurement points are distributed as they are, sampling each line (all three of them in the case of laser 1) symmetrically with a central datapoint at the maximum of the line. This cannot be accidental. Is there still a piezo on one of the cavity mirrors able to adjust the cavity length? Can the L2140-i thus switch between mode-by-mode scanning and the traditional Picarro WS-CRDS mode of operation (turning it into an L2130-i-C)? Is this piezo used to move the cavity frequency comb in between spectral scans? If so, doesn't the bias applied to the piezo lead to an even further reduction of the thermal stability of the cavity? Please explain . . .

We do use a PZT to move one mirror of the cavity in such a way as to align the comb of longitudinal lines to the absorption spectrum. The measured spectrum itself provides feedback to maintain this spectral alignment, essentially stabilizing the optical path length of the cavity despite changes in temperature, pressure, and above all humidity of the analyzed sample. The feedback does not degrade the stability of the cavity; it improves it, at least as far as the long-term constancy of optical path length is concerned. Since the mirror motion is controlled by spectroscopy of lines that have an invariant (to the precision we care about) frequency relationship, this complication does not affect the essential argument about the improved precision of the frequency axis. The reviewer noted that patent application is not yet a citable reference, so we have removed this and replace it with a more complete explanation as follows: "Note that, as in the L2130-I, a piezo is still used to physically move one mirror of the cavity in such a way as to align the comb of longitudinal lines to the absorption spectrum, and the wavelength monitor is still used for feedback to the laser frequency control electronics, thus allowing for rapid tuning to a frequency near the desired resonance. However, the fine frequency spacing for a given narrow spectral region (e.g., that for $H_2{}^{17}O$) is now determined only by the FSR."

---

3) Review by anonymous reviewer (reviewer #2).

The paper presents an improved version of a commercial optical water isotope analyzer produced by Picarro. This follows on to similar efforts for a competing commercial product made by Los Gatos Research and published in Berman et al. (2013). The new analyzer uses an improved laser stabilization method to achieve precisions similar to mass-spectrometric reference methods. The paper describes the method and results satisfactorily, but needs clarification and correction in the points listed below.

We very much appreciate the reviewer's very thorough review, and in particular their catching a couple of specific errors in the text. Detailed responses are given below.

General comments:
The term "accuracy" should not be used in the context of this paper because, from a metrological point of view, none of the measurements are traceable to the SI. They all depend on calibration against conventional delta values of the reference materials VSMOW and SLAP (i.e. delta values that have been fixed to a certain value for convenience, to eliminate scale variations between different laboratories). The term "re- producible" may however be appropriate: results give

numbers that reproduce results from other labs, after two-point normalization. To establish the accuracy of the results, a careful calibration of the instrument using gravimetric mixtures of pure (or at least enriched) isotopologues would be required. See also VIM3 - International Vocabulary of Metrology (http://www.bipm.org/en/publications/guides/vim.html).

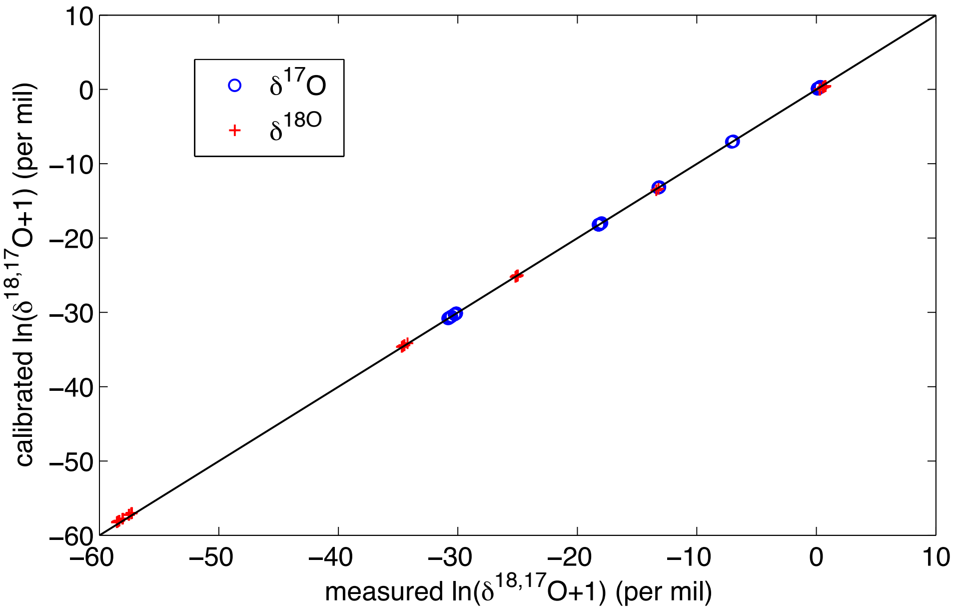
We respectively disagree with this statement. No claim is made with respect to absolute accuracy, which is virtually impossible to achieve with respect to isotope ratio measurements, particularly with $^{17}O_{excess}$. The isotope ratio terminology for $^{17}O_{excess}$ – see our paper on this (Schoenemann et al., 2013) – is *only* meaningfully defined in reference to a two-point calibration, and the term "accuracy" is therefore entirely appropriate. There is no ambiguity here. Our use of accuracy is identical to that used in the isotope ratio literature in general, and in laser spectroscopy techniques in particular. See for example: Wassenaar, L. I.; Coplen, T. B. & Aggarwal, P. K. (2013), 'Approaches for Achieving Long-Term Accuracy and Precision of δ18O and δ2H for Waters Analyzed using Laser Absorption Spectrometers', Environ. Sci. Technol. 48(2), 1123—1131.

Please explain whether the results in Table 2 include the calibration and normalization uncertainty from the analyses of SLAP vs. VSMOW. Results for both the commercial and the custom-built vaporizer should be shown and compared.

The fixed two point calibration means that we take the VSMOW2 and SLAP2 values to have zero uncertainty by definition of the VSMOW-SLAP scale (see our comments in response to Berman as well). We have added a note to the Table caption to make this clear. As far as the custom vaporizer is concerned its use was intended for the long Allan variance tests and thus a comparison would be irrelevant here. The custom vaporizer uses much more water for each analysis and cannot in practice be used for valuable standards.

Measured results for delta(18O), delta(17O) and delta(2H) for SLAP vs. VSMOW should be shown before normalization to demonstrate how large the scale compression is compared to the conventional values of -55.5 ‰ -29.7 ‰ and -428 ‰. A scatter plot of ln(1+delta(17O)) vs. ln(1+delta(18O)) prior to normalization should be shown for the CRDS data in Table 2 as well as the data from the corresponding IRMS measurements (currently not shown anywhere). What are the corresponding slopes of the plots? This would add important information about the nature of systematic instrumental errors, in addition to what has been included in the Discussion section and Figs. 7 & 8.

This request on the part of the reviewer is based, we assume, on an understanding deriving from IRMS measurements, in which the scale compression is well known. In the case of the laser spectroscopy measurements, the scale compression is not meaningful, because the "raw" values derive from a ratio of nominally absolute measurements, which are not made in reference a working gas as they would be with IRMS. Rather, the "reference gas" is simply a number in the software, whose value can only be determined from the calibration methods we have outlined. Any plot of "raw" data would thus be arbitrary, as it is entirely dependent on the (originally arbitrary) values that were originally put into the instrument software. For example, the $\delta^{18}O_x$ value for a sample (x) is given by our equation 15: $\delta^{18}O_x = {}^{18}R_x/{}^{18}R_{ref} - 1$, where $^{18}R_x$ is the measured ratio of the integrated absorption of the $H_2{}^{18}O$ molecule to that of the $H_2{}^{16}O$. R_{ref} is *not* measured, but is a nominal value adjusted such that "raw" measurements of VSMOW would result in a $\delta^{18}O$ and $\delta^{17}O$ for VSMOW = 0. The "raw" difference between sample x and another sample y is thus given by $({}^{18}R_x\text{-}R_y)/{}^{18}R_{ref}$, the magnitude of which scales directly with the value of $^{18}R_{ref}$. For completeness, we show below such a plot – the slope for both isotope ratios is indistinguishable from 1.



The paper by Berman et al. (2013) should be referenced already in the Introduction, not only in the discussion.
We now reference this early in the paper.

The differences to the analyzer used by Berman et al. (2013) should be explained.
This paper does not claim to be a comparison study. A comparative study in the future would be valuable but is very far beyond the scope of our work. We do not have information on the Los Gatos system discussed by Berman et al. (2013) beyond what is published, and do not have access to one of these instruments. The spectroscopic methods between these two instruments are very different, and those differences have been addressed previously e.g. by Aemisegger et al. (2012), which we now cite.

SI units should be used. Replace Torr with Pa or mbar. sccm (p. 10200) with cm3 min–1 and give the relevant conditions (273 K, 1 bar?).
We have made those changes.

17O excess is a compound noun, so "excess" should be written as subscript index to "17O".
We agree and have done so throughout.

Specific comments:
10192/12: According to fig. 4, 40 measurements took 60 h, so 1 measurement should take 1.5 h, not 30 min, which is not significantly faster than the mass-spectrometric method, which takes 2.2 h (Barkan and Luz, 2005).
We agree that in practice, these methods may not be significantly faster (though they are much less labor intensive). But distinction needs to be made between the integration time (stated) and the actual analysis time including instrument pump down, etc. Note that for a fair comparison, the IRMS method then needs to include the sample preparation time, which is at least 20 minutes

per sample in our lab (not counting the necessary triplicate repeats because of memory). The laser spectroscopy method thus remains indeed faster. We do not claim "significantly faster" in the paper.

10192/15: It is not clear how "precision" is defined - repeatability of multiple injections of the same sample, continuous measurement of a single sample over a long time, reproducibility of an isotope delta on different days (i.e. including calibration).
We have made it clear that the term "precision" here refer to the reproducibility of repeated measurements, after calibration. This should be sufficient for the abstract, and we expect that interested readers will look at the rest of the paper for the details.

10197/9: Eq. 6 is wrong. The right hand side needs to be divided by (1+delta18O{^raw}{_VSMOW}). However, this error is inconsequential because it cancels out in the next step (Eq. 7).
We have corrected this. Thank you for catching this oversight.

10198/1: This statement is only correct if the measured values are normalized according to their measured delta18O values and a corresponding clarification should be added.
We agree and we have added this clarification.

10200/1: Give the values for sigma here. The stated precision (standard error) is the expected precision based on Poisson statistics, not necessarily the standard error that can be achieved under realistic conditions.
It is unclear which part of the text the reviewer is referring to. In each case where we report the standard deviation of the mean (no longer called "standard error"), we also provide the number of samples, so the sigma values can be readily calculated. In those cases (in particular, for averages of repeated measurements), we feel that the standard deviation of the mean is a more meaningful number than the simple standard deviation, since the latter is independent of the number of samples.

10200/10: This is not a concentration, but a mole fraction. SI units should be used, i.e. 20 mmol mol–1 or similar. How close does the actual concentration need to be to its nominal value of 20 mmol mol–1?
We now use mole fraction and SI units as suggested.

Note that we have not fully tested the instrument at multiple water concentrations, other than to determine the effective concentration dependence. All our measurements (other than the concentration dependence data) were analyzed between 18,000 and 20,000 mmol mol^{-1}, following the same guidelines that Picarro has provided previously for their L2130 instrument. We now make it clear in our paper that our results only apply (as far as we have established in this paper) to that measurement range.

10202/5: Compare and contrast this measurement principle to that used by Los Gatos Research in their instrument (Berman et al. 2013), which also uses a peak integration method.
As noted above, a comparison with the Berman et al. methods is not within the scope of this paper, and nor do we have sufficient information to do so.

10208/1: Please clarify whether seawater samples can be analyzed with the analyzer and if so, what precautions need to be taken.

Salt water samples would likely clog the capillary, so we have not tested this. It is likely that the instrument can be used for $^{17}O_{excess}$ of salt samples (as the L2130-*i* has been used for $\delta^{18}O$) but considerably more work would be needed to test whether the results for $^{17}O_{excess}$ are reliable. We have added a note to make it clear that we have not tested the instrument with seawater samples. Sample KD is a marine sample but the salt has been removed by a reverse osmosis system.

10210/16 & Figs. 7 and 8: Please explain what you mean by "residuals" and how the residuals have been calculated.
This is simply the difference of each of the measurements from the mean of all the measurements for a given sample, as is now stated.

10211/15 & Fig. 5: Please extend Fig. 5 to 5000 ppm to demonstrate the claimed linearity of the correction.
We do not have sufficient data to demonstrate this, nor did we intend to claim the correction is linear to low water vapor concentrations. In fact, we have since tested the instrument at 5000 ppm and it is still very linear. But we do not have the relevant comparison data for Fig. 5, from the old prototype instrument. We therefore do not report on low concentrations here, and have corrected the text accordingly so that it is consistent with the figure.

10211/22: Please cite Barkan and Luz (2005) as reference for the equilibrium value of 0.529.
Done.

10212/18: Delete sentence starting "Direct analysis of ambient water vapor ..." as this has not been demonstrated and cannot be a conclusion of the paper.
We agree that we have no demonstrated this, but do think it is valuable to point out that in principle this instrument can be used this way. What the analyzer measures is a gaseous sample at low pressure. This essentially means that a study based on liquid standards and samples has to solve the problem of producing a vapor stream. As long as this is done properly the calibration can be used for either liquid or vapor sample measurements assuming one accounts for the water concentration dependence in a proper way. Similarly all vapor studies we know that use CRDS perform a calibration based on liquid standards in a way very similar to what we do here. We have rewritten the sentence to make it clear that such applications are not yet demonstrated.

10212/22: Please tone down the language - if any paper has "established" CRDS measurements of 17O excess as alternative to IRMS, then the paper of Berman et al. (2013) should be cited.
The reviewer implies that we are overstating the significance and originality of our work. We respectfully disagree. Berman et al.'s work should be referenced in this paper, and we have done so. However, Berman et al.'s work makes use of off axis integrated output spectroscopy, not CRDS, so our statement is correct – our work does indeed establish CRDS as an alternative to IRMS. In any case, our work is independent of Berman et al. and cannot be said to "follow on" from the work of Berman et al. as the reviewer suggests in his introduction. Our work was complete prior to our awareness of the work by Berman et al., and was done entirely independently. Indeed, we presented the bulk of our results at the Goldschmidt symposium in early 2013, and first presented the essential concepts and prototype work at EGU in 2012; we now cite both the Goldschmidt summary and Berman et al. paper in the introduction to our paper to clarify the history of these independent developments. Note that one of us (grad. student Schoenemann) contributed to the Berman et al. paper by providing guidance on the methods for calibration (as is stated in their acknowledgements).

Fig. 1: Explain what fit function was used for the peaks.
This is explained in p 10201, 10202 in detail, in the discussion of Galtry functions.

Table 1: It is unclear what you mean by standard error. For example. for GISP, is this the standard deviation of the 20 individual (daily?) results, i.e. the standard deviation of the mean? Or the mean of the standard error of the 20 individual results? The former value would be more informative, but should be called "standard deviation of the mean", for the avoidance of doubt.
We have modified the text to make this clear (we mean the standard deviation of the 20 independent single-vial measurements, divided by 20). We now use the term "standard deviation of the mean". See also the comments to other reviewers regarding the "standard error" terminology.

Technical corrections:
In species such as H218O, the subscript 2 and superscript 18 should be separated to make clear which atom the indices belong to.
We defer to the editor on this. We believe it is clear as is.

10192/22: The element symbol is missing on the left hand side of the equation, e.g. delta {ˆi}E =
Thank you for noting this. We have corrected it.

10192/23: The physical quantity "2H", "1H", etc. refer to needs to be included, e.g. isotopic abundance x or concentration c, so that 2R = x(2H) / x(1H), etc.
We agree. We have added this clarification.

10193/3: The center dot should be deleted.
Done.

10193/19: Define n{ˆ0}.
This is the standard chemistry terminology for the neutron; we do not think it necessary to define it here, but we defer to the editor.

10193/21: The indices should be "sample" and "reference" to be consistent with equation 1. They should be in upright font.
Done.

10193/24: This applies only to equilibrium for T approaching infinity.
Agreed. We say "approximately" since this is an approximation for any real temperature.

10194/2: lambda = ln(Q17/Q16)/ln(Q18/Q16). Instead of nuclide numbers, exact atomic masses should be used so that lambda = 0.5305.
The reviewer is correct. We have modified the text accordingly. This is a good opportunity to correct the record – the approximation of 0.529 has appeared in some earlier work. The correct calculation (1/15.994914-1/16.999131)/(1/15.994914-1/17.999160) = 0.5305.

10194/14: The center dot denotes a scalar product of matrices and should be deleted.
Done.

10194/17: A reference needs to be given.
We have added a citation of Barkan and Luz, 2005.

10194/19: The crosses should be deleted or replaced by an equal sign.
Done.

10195/13: 30 min (Barkan and Luz, 2005)
Corrected. The value varies from lab to lab, but we use the Barkan and Luz value as our example, as suggested.

10197/9: The index should be "sample" for consistency (see above).
Done.
10197/11: The left hand size should show delta18O.
Corrected. Thank you for catching this error.

10200/2: Delete tilde sign (~) - the number of significant figures should be chosen so that the value reflects its uncertainty. 10200/6: Delete tilde sign.

Done.
10201/19: Add "molecular" before monochromatic and delete "molecule–1" from within the parentheses. The units of kappa should be cm2 Hz–1 (based on Eq. 10).

Done. We agree this is clearer. The correct units are just cm^2.
10201/22: The center dot should be deleted.

Done.
10202/13: The symbol "Re" should be used to designate the real part of the imaginary number. A single R with a left bar designates the set of real numbers.

Thank you – corrected.

10203: All equations need to indicate the measured quantity on the right side, e.g. 18R = A(H218O, 11) / A(H216O, 2).
We agree – corrected.

10204/11: Replace "Tee" with "tee union" (lower case, no quotation marks).
Done.

10204/16: Replace concentrations with mole fractions.
We have used concentrations throughout the document and feel this is clear, and consistent with previous work.

10210/6: Delete equal or approximately equal sign.
Done.

10211/5: What does "approximately 100" mean? 99? 100? 80? 120? Do you need the approximation sign?
Agreed. We do not need the approximation sign.

10212/12: Replace "deviation" with "difference".
Done.