Response to Reviewer #2

We wish to thank the reviewer for providing such detailed and critical comments of our work. Our responses are provided below the original comments, which are italicized for clarity.

Review of The stability and calibration of water vapor isotope ratio measurements during long-term deployments by Bailey et al., 2015

General comments
This paper presents an important and detailed analysis of the data quality of water vapour isotope measurements using laser spectroscopy over time periods of several years in remote locations with minimal manpower available for attendance. It focuses on the importance of the water vapour concentration dependency correction of laser spectroscopic measurements and compares different statistical fitting schemes to account for it. This study faces the challenge to condense lessons-learnt from completely different measurement setups and calibration approaches in a single study.

This paper is in the scope of Atmospheric Measurement Techniques and presents a valuable new perspective on the calibration of water isotope measurements, which have been booming since the arrival of novel commercial laser spectrometric measurement devices. The paper is generally well written, particularly the abstract and Section 6 (implication for long term deployments) are a pleasure to read. However, there are a few instances where it lacks clarity (especially in the methods and results discussion). I thus have the following major comments that the authors should address before final publication:

1) Several important terms ("concentration dependence", "instrument drift", "bias", "long-term stability", "prediction error", "reproducibility") used throughout the paper are not clearly defined and sometimes used in a diffuse way. It would help the reader a lot if a more consistent definition and also short literature review on these terms would be done in the introduction.

Throughout the paper, but particularly in the Introduction, the phrasing has been carefully refined to ensure that the different error terms are understood. For example, on pg. 6 of the Introduction, a new paragraph has been added that begins "Instrumental drift creates another source of isotopic bias…" In this paragraph, we explicitly define "repeatability"—which we now use consistently throughout the paper in place of "reproducibility"—and its connection to "instrumental drift." "Repeatability" is also contrasted with "precision" in this paragraph, to help the reader distinguish these terms later in Sect. 4, where an uncertainty analysis is presented. Numerous citations are provided of studies that have previously investigated "repeatability" and "precision."

"Concentration dependence" is defined and extensive citations are provided in the preceding paragraph.

Although we have not added definitions of "bias" and "stability," we have ensured that these terms are used in a manner consistent with general scientific usage. Specifically, "bias" is always meant to represent the difference between the measured and true value caused by systematic error, and "stability" indicates consistency with time.

"Prediction errors" are now defined explicitly in the 3rd paragraph of Sect. 2.4 as the "standard errors associated with the fitted values."

2) Since calibration is the central topic of this paper the state of the art of calibration methods (not only the commercial ones and the custom-ones made for their own measurements) published earlier in the literature (Iannone et al., 2009, Sturm and Knohl, 2010) should be mentioned and shortly reviewed, also addressing the challenges of finding a calibration method that is not biasing (i.e. strongly affected by memory effects or inducing unwanted fractionation effects) the isotope composition of the standard vapour production itself. There is one important paper that should be cited in this respect by Kurita et al., 2012. Furthermore, calibrating a water vapour instrument with an autosampler or a continuous water vapour source are two totally different approaches. This should be emphasised more clearly in the methods section.

The Introduction has been enhanced with a more detailed description of the three major types of calibration systems currently in use, and additional references, including those noted by the reviewer, have been added (see the 7th paragraph of the Introduction). Kurita et al.'s (2012) report on the difficulty of producing water vapor that is not biasing is now explicitly mentioned and cited. Furthermore, statements naming the two types of calibration systems used in the present analysis are now made in the Introduction (at the end of the 7th paragraph and in the concluding paragraph) and in the introductory paragraph of the Methods section (Sect 2.).

3) The methods section is very difficult to follow, lacks clarity and preciseness (particularly on pp. 5435-5436). A table summarising the measurement systems (version) used with their calibration system and calibration strategy including changes after a certain period of operation, frequency of normalisation to VSMOW-SLAP and how it was done, frequency of calibration runs, etc, would be extremely helpful to follow the discussion. The measurement statistics (humidity, isotope, temperature ranges outside, inside) could also be added to such a table to know in which conditions the instruments were operated.

The Methods have been substantially re-written. Additional details have been added to Sections 2.2 and 2.3, following the reviewer's line-by-line comments, and a Table has been added that describes the calibration systems and approaches at each site. In particular, the revisions emphasize the motivation for different calibration approaches to help the reader understand why certain protocols were established or changed. Furthermore, Methods Sect. 2.4 has been substantially revised to improve the transparency of the local regression techniques (more on this below) and to explain how

the locally weighted polynomial regression was applied to correct the calibration data at each site.

4) I find the structure of the results section weak. After the methods it would be good to have a section "Results and discussion" with a short paragraph giving an outline of what is coming. Otherwise the reader gets lost in the different measurement setups and correction schemes.

As per the reviewer's suggestion, a short paragraph outlining the material that follows has been added.

5) The argumentation about the simultaneous correction of the concentration-dependence and the deviation from the VSMOW-SLAP scale providing better data accuracy is not consistent throughout the paper. The authors could not show a clear dependence of the water concentration correction on the isotope composition of the standard used for the characterisation in section 3.2. Furthermore they speak about a linear scaling to VSMOW-SLAP. If they assume a linear normalisation function then the two corrections (normalisation and water vapour mixing ratio dependency) are orthogonal and do not depend on each other. This aspect is confusing for me and not described clearly enough in the paper (see also several specific comments below particularly on pp. 5435-5436).

We have revised the argument supporting our use of the simultaneous correction in Sect. 2.4: the advantage of this approach is that it allows us to use all isotopic information from the Mauna Loa calibrations without requiring that we normalize the isotopic measurements to a reference humidity. This is particularly advantageous for long-term deployments (including in Hawaii) if the calibration system in use (in our case, the autosampler) does not provide a constant humidity output.

The reviewer is correct that if the concentration dependence is not dependent on isotope ratio, there should be no difference in error estimation between a joint prediction and summing individual errors in quadrature. Therefore we have rephrased the argument about joint prediction being advantageous so that it appears as a subjunctive statement in a parenthetical note. However, different error estimates will result if the isotopic information incorporated into the fitting is not the same in both cases. Because more of the Mauna Loa calibration data can be incorporated into the simultaneous correction, the errors from the joint estimation are slightly different than the sum of the individual errors produced when correcting the concentration dependence and normalizing to VSMOW-SLAP sequentially.

6) One of the major points in this study is to highlight the importance of the curve-fitting choice for minimising the uncertainties associated with the water vapour concentration-dependency correction. No figure however shows the actual data and the different curve fits (see also my specific comments below).

New panels in Figures 1 and 2 now show the concentration-dependence calibration data for both sites.

7) The fact that the uncertainties in isotope measurements can be very different depending on the timescale of interest (averaging of the data) is not discussed at all here but is an important aspect and needs to be addressed.

The 5th paragraph of the Introduction now clearly defines precision and discusses the importance of optimizing measurement averaging time in order to reduce related uncertainties. Also, an addition to the second paragraph in Section 4.1 emphasizes the fact that precision is often improved with longer averaging times.

The importance of averaging time for minimizing uncertainties related to instrumental precision at Summit is stated in the 2nd paragraph of Sect. 4.2.

Specific comments
Introduction: (Section 1)
1) p.5430, L. 4 Embed this parenthesis in the text or add it before the previous list of references.

The reference to Wen et al. (2012) has been moved out of the parentheses into the main text, as suggested. Moreover, the description of calibration systems in the Introduction has been bolstered with additional references.

2) p. 5430, L. 11 The Synflex-memory problem was already mentioned by Sturm and Knohl, 2010.

Sturm and Knohl (2010) have been added as a reference for the statement regarding Synflex-memory problems.

3) p. 5430, L. 15 I am not convinced that the wording "Additional biases" is adequate here. I find it confusing. Something like "Normalisation to VSMOW-SLAP is done by..." would be clearer.

The Introduction has been substantially re-structured so that two measurement biases are emphasized: 1) a concentration-dependent bias and 2) drift-induced changes in measurement repeatability. The phrasing "Additional biases" has, consequently, been eliminated.

4) p. 5430, L. 20 it should be "the slope and intercept", to me it is not a priori clear why the slope of the normalisation to VSMOW-SLAP would be constant.

To use terminology more consistently (following the reviewer's 1st major comment), we now refer to this bias as drift-induced changes in measurement repeatability throughout the text. This change avoids the possible insinuation that the VSMOW-SLAP slope is constant for laser isotopic analyzers.

Methods: (Section 2)
5) p.5433 L. 5 Here it is essential that the authors say exactly for which period 18 injections have been made and for which only 3. What is the percentage of the data where you have only 3 injections?

The time periods for the different calibration approaches are now provided in the 3rd paragraph of Sect. 2.2 and also listed in the new Table, which has been added following the reviewer's 3rd major comment.

6) p. 5433 L. 10 and L.16 (maybe also elsewhere) "manual" is confusing. Is it Picarro's standard delivery module that has been used? If yes mention it, if not more details about the setup should be provided.

The term "manual" has been replaced with "custom," and we have added references that describe similar systems to ours.

7) p. 5434 L. 1 The dry air characterisation should be stated in water vapour concentration units not in terms of dew point. Connect the next sentence ("Dry airflow bubbled") better to the previous one.

The water vapor concentration is now provided in mmol mol^{-1} and the text has been edited so that it is clear that it is the dry air produced by the dryer that bubbles through the liquid standard.

8) p. 5434 L. 14 It is important to mention here already (not only in the conclusions) that samples of the liquid water of the DPG could have been taken to monitor the drift.

At the end of the 4th paragraph of Sect. 2.3, we now explicitly state that this distillation could have been checked regularly had it not been for the fact that the Summit analyzers and custom dew point generator were inaccessible for 11-month periods.

9) p. 5434 L. 20 What means "spot checked"? How frequently?

The 5th paragraph of Sect. 2.3 has been substantially re-written to present more clearly the details of the Summit calibration. The approximate dates of the extended concentration calibrations are now given in the text and in the new Table.

10) p. 5434 L. 27 Simplify these two sentences by saying: "For all of the six-hourly calibrations the first 9 of 20 minutes... Longer sampling was prescribed at the lowest humidity (using the last 21 of 40 minutes)". Is this what is meant? At least remove the full sentence in parenthesis and explain better, if I got it wrong.

These last sentences of the 5th paragraph of Sect. 2.3 have been substantially revised to clarify the calibration approach.

11) p. 5435 L. 2-4 How frequently was the VSMOW-SLAP normalisation updated? What means "checked"? Were the slope and intercept of the normalisation updated or only the intercept? How strong was the change of the slope over time?

As now stated at the end of Sect. 2.3, the isotopic deviation from VSMOW-SLAP was checked in July 2013 by measuring three standard waters. Since only one isotopic standard was measured consistently during the Summit field deployment, we cannot comment on whether the slope of the deviation changed with time.

12) p. 5435 L. 24 – p. 5436 L. 6 Since the type of fitting procedure used is a key aspect of this paper, this section explaining the locally weighted regression is too short and vague.

Although the information necessary to replicate the nonparametric fit using the R LOCFIT package was included in the original text, we have substantially re-written Sect. 2.4, adding more details to the description of the local regression and clarifying how it is applied to correct the isotopic biases at each site.

13) p.5435 L. 6 The first sentence is confusing. I don't think that it is so trivial to attribute the "measurement bias" to different sources.

This phrasing has been eliminated since this was not the meaning intended.

14) p.5435 L. 8-15 What is meant by normalisation here? Be more explicit. Is it the normalisation to the isotope measurement in the mentioned humidity range or the normalisation to VSMOW-VSLAP that is meant, or both?

The beginning of Sect. 2.4 now clearly states that the data are normalized to a reference humidity, and the volume mixing ratio range of this reference humidity is provided for each site.

15) p.5435 L. 24 The chosen local fitting approach may have several advantages, which should be mentioned more clearly in the text. However the disadvantage is that it is totally intransparent and the whole dataset and fitting procedure is needed to be able to reproduce it. The authors should address this important aspect and provide more detail, since this is one of their key points of the paper.

We have added a sentence to the 2nd paragraph of Sect. 2.4: a major advantage of the local regression technique is that it ensures that variations in the isotope ratio across small subsets of the larger ambient *q* range are well calibrated, which ultimately results in a better characterization of the concentration dependence. This is verified by the lower RMSE presented in Sect. 3.1.1 and emphasized in the new panels of Figures 1 and 2. We disagree however about the fit being less transparent than fits derived with parametric functions, as all are based on least squares estimation. The revised text (2nd paragraph of Sect. 2.4) explains more clearly that the local regression simply performs this least squares estimation on small subsets of the data (e.g. where 1- or 2-degree polynomials are applied within fitting windows).

Furthermore, because the shape of the concentration-dependent bias is unique to each analyzer (which is supported by citations in the Introduction), one always needs the original calibration data to derive the appropriate curve. It would be inappropriate to take the parameters for a polynomial function derived from one analyzer and use them to calibrate another instrument.

16) p. 5435 L. 14 Remove the parenthesis, this is an important additional information.

The parentheses have been removed, as suggested.

17) p. 5436 L. 9 Is a linear correction for the VSMOW-SLAP scale normalisation assumed (as recommended by the IAEA)? What does the joint characterisation function look like?

A linear normalization is not assumed, since the locally weighted polynomial regression is modeled as a function of both humidity and the isotope ratio measured; however, the best fit to the data is nearly linear in the direction of the latter. This point is now made in the 5th paragraph of Sect. 2.4.

In a separate analysis, we evaluated whether there was a significant difference between A) correcting the concentration dependence and normalizing to VSMOW-SLAP simultaneously (as described here) and B) correcting the concentration dependence first and linearly normalizing to VSMOW-SLAP subsequently. The two different data-processing techniques had no significant bearing on our interpretation of the ambient data at Mauna Loa (see Bailey et al.: Precipitation efficiency derived from isotope ratios in water vapor distinguishes dynamical and microphysical influences on subtropical atmospheric constituents, J. Geophys. Res., 120, doi:10.1002/2015JD023403, 2015).

18) p. 5436 L. 10 Why is the natural log of humidity used here?

The motivation for characterizing the concentration dependence as a function of ln(q) is now stated in the second paragraph of Sect. 2.4.

19) p. 5436 L. 14-L. 17 "Different" from what? I am not sure that I understand the argument correctly. In the way I understand it, estimating the errors jointly has the advantage that one can account for the covariance in the measurement errors (if there is any). Explain better. Furthermore, I think that the approach of joint normalisation/water vapour mixing ratio correction is nice in theory but not very practicable. It requires a large amount of standards that are well distributed on the delta-scale. Furthermore repeated measurements at many water vapour mixing ratios are needed to achieve a precise correction, which I doubt is constant in time. Earlier studies have shown that the stability of the current laser spectroscopic instruments is not good enough to conduct measurements over more than a few hours to one day without calibration. These drawbacks are only mentioned in the conclusions. The trade-off between a precise

normalisation/water vapour mixing ratio correction, which should be ideally repeated regularly, while still optimising ambient measurement time should be shortly mentioned.

What was meant was that the estimates of error are different from one another. However, as this was unclear, the sentence has been restructured.

Regarding tradeoffs in calibration approaches, the 5th paragraph of Sect. 2.4 now includes a statement describing the balance between repeating calibrations and optimizing measurement time; however, as also noted, the reason a joint characterization was selected for Hawaii is because the autosampler did not produce volume mixing ratios reliably. Therefore, the joint characterization allowed us to use all of the calibration information without requiring that the isotope ratios be normalized to a single reference humidity level.

20) p. 5436 L. 18 The weighting by 1/q2 to account for the underrepresented low humidity samples seems arbitrary and is problematic in my view. Most of the measurement points are distributed around the centre of the humidity range and not at high humidities. 1/n would be statistically more meaningful. If no discrete number of points is available, 1/(ln(q)) or the number of points in water vapour concentration bins would be more adequate.

The motivation for this weighting is now described in the 3rd paragraph of Sect. 2.4, and the fact that $1/q$ and $1/\ln(q)$ do not perform as well is stated explicitly. Also, the new panel in Figure 1 shows more clearly that most of our measurement points are at higher humidities, which is why $1/q^2$ is particularly useful for characterizing the concentration dependence accurately at Mauna Loa.

Results and Discussion: (Sections 3-5)
21) p. 5437 L. 24 In the description of the filtered vs non-filtered data a comment on the trade-off between robust estimation using a maximum number of injections vs avoiding memory affected measurements should be more clearly made. I am not convinced by the analysis of the filtered vs non-filtered data. I would not expect to see a large effect if you remove only one of three injections. Several studies have shown that depending on the change in isotope signal associated with the injection the memory effect lasts for much more than 3 injections (e.g. Penna, et al. 2012). I would remove this aspect and simplify the (already complex and long enough) text accordingly. With the information I have from the text, I am not convinced that the authors have enough data (number of injections) to show that the chosen characterization function is more important than "any" filtering of autosampler data.

We do agree that our filtering procedure will not eliminate all memory effects; however, please be aware that we are eliminating two injections (following Penna et al., 2010), not a single injection. We have added references that show that the first two injections capture most of the hysteresis, even though it may take several more for the measured isotope ratio to approach its "final" value. Therefore we believe the analysis remains

useful for discussing the relative importance of curve fitting and hysteresis. Nevertheless, we have removed the word "any" so as not to overgeneralize the results.

In addition, we have inserted a new sentence in the 1st paragraph of Sect. 3.1.1 that emphasizes the tradeoff between discarding injections to reduce memory effects and maximizing the calibration sample size in order to improve the accuracy of the curve fitting.

22) p. 5437 L. 24 – p. 5438 L.2 I would like to see the actual datapoints underlying this analysis with the different characterisation functions (see also my major comment 6 and the specific comment on Fig. 1).

The data points have been added to Figures 1 and 2 and the local regression curves overlain. The other curves are not shown, however, since the pre-existing panels already show the differences between the local regressions and parametric fits.

23) p. 5438 L. 11-13 This sentence is confusing. Mention that the comparison is done in Fig.2.

The sentence has been revised so that it is clearer that the comparison is done through Figure 2.

24) p. 5441 L. 2 "The signal-to-noise" at low humidities is an important point. The estimated curve fitting effects are smaller than the measurement precision at the low humidity end (<2mmol/mol, except for the quadratic fit, which seems very inadequate when measuring in low humidity environments). I think that the authors should avoid making a general statement of which fitting procedure is best. This strongly depends on the individual instruments and the measurement range in which they are operated.

The final sentence of Sect. 3.1.3 has been revised as suggested to avoid over-generalization.

25) p. 5445 L.25 How is precision calculated here? Based on what averaging time scale of the raw data?

Precision is defined in the overview of Sect. 4 as the standard deviations of the calibration points: one-minute averages at Summit and approximately two-minute-long injections at Mauna Loa. An additional comment has been added to the 2nd paragraph of Sect. 4.1 to clarify that the precision could be improved by optimizing the measurement time.

26) p. 5446 L. 1-5 and p. 5447 L. 5-10 I am not convinced by this argument. In the case of long-term deployments I would not be already satisfied if the diurnal to synoptic variability is larger than the measurement uncertainty. Here the authors could be more consistent with the stipulated aim of the paper and thus also a bit more ambitious in the target uncertainty values that should be achieved to make long-term water vapour

isotope measurements useful. The authors should also aim at showing that they can also resolve interannual variability with these measurement setups.

We have revised Sect. 4 in two places to address the reviewer's concern.

For Mauna Loa (Sect. 4.1), we now note that differences of about twice the total error might be desirable to distinguish any paired observations. However, long-term deployments are typically not so interested in distinguishing individual observations as they are in identifying long-term statistical trends. As the new sentences in Sect. 4.1 point out, isotopic changes smaller than twice the total error could be detectable if a long-term linear trend is identified whose slope differs significantly from the linear trend in the calibration data.

For Summit (Sect. 4.2), we discuss the changes in calibration strategy necessary to increase the measurement sensitivity so that a 0.8 permil difference in $\delta^{18}O$ (equivalent to a ~1°C temperature change) would be detectable between any paired observations.

27) p. 5447 L. 18-19 "beginning with... or..." This is confusing. Rephrase.

The awkward clause has simply been removed.

28) p. 5450 L. 3 I do not agree with the first point. An ideal calibration system should cover the ambient measurement range. A very precise water concentration correction at low water concentrations is not useful if the measured ambient humidity is at the high end. Reformulate.

We have changed the point to reflect the author's suggestion.

Figures and tables:
29) Fig. 1: As mentioned above it would be important to have a second subplot here with the datapoints and the shape of the water vapour mixing ratio dependency curves underlying the fits as well as the chosen fits. Otherwise the whole curve-fitting is very intransparent.

We have added subplots with data to both Figures 1 and 2.

30) Figs. 8 and 10: shortly explain what is meant by "prediction", "reproducibility" and "precision" in the legend.

These terms are described in more detail now in the captions. Additional detail would not have fit easily in the legend space.