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.

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.

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.

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.

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 (ee also several specific comments below particularly on pp. 5435-5436).

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).

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.

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.

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

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.

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.

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?

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.

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.

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.

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

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.

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?

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.

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.

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?

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.

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

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?

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

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.

20) p. 5436 L. 18 The weighting by $1/q^2$ 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.

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.

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).

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

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

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.

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

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.

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.

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

References:

Iannone, R. Q., Romanini, D., Kassi, S., Meijer, H. A. J., and Kerstel, E. R. T.: A microdrop generator for the calibration of a water vapor isotope ratio spectrometer, J. Atmos. Ocean Tech., 26, 1275–1288, doi:10.1175/2008JTECHA1218.1, 2009.

Kurita, N., Newman, B. D., Araguas-Araguas, L. J., and Aggarwal, P.: Evaluation of continuous water vapor δD and $\delta^{18}O$ measurements by off-axis integrated cavity output spectroscopy, Atmos. Meas. Tech., 5, 2069-2080, doi:10.5194/amt-5-2069-2012, 2012.

Penna, D., Stenni, B., Šanda, M., Wrede, S., Bogaard, T. A., Gobbi, A., Borga, M., Fischer, B. M. C., Bonazza, M., and Chárová, Z.: On the reproducibility and repeatability of laser absorption spectroscopy measurements for $\delta^{2}H$ and $\delta^{18}O$ isotopic analysis, Hydrol. Earth Syst. Sci., 14, 1551-1566, doi:10.5194/hess-14-1551-2010, 2010.

Sturm, P. and Knohl, A.: Water vapor δ 2 H and δ 18 O measurements using off-axis integrated cavity output spectroscopy, Atmos. Meas. Tech., 3, 67–77, doi:10.5194/amt-3-67-2010, 2010.