Reviewer 1 - Harro A.J. Meijer

This is a well-written, interesting paper about the 'next level' OF-CEAS instrumentation. I liked reading it, and is carries important new information. I recommend publication, but/and have a set of comments and questions that I invite the authors to respond to, preferably by adapting the paper where appropriate, or else in writing to the editor (and me).

We thank Harro Meijer for his very pertinent questions concerning the general form of the manuscript as well as the scientific content, which allow us to be exhaustive on the instrumental limits presented in the paper, and give us interesting perspectives for future studies. The answers are inserted in red and the citations from the manuscript are indicated in blue.

First a general question: how come it is so difficult to perform isotope measurements on water vapour at 10 ppm or even higher, while for example precise isotope measurements on methane (2 ppm) are routine nowadays? Is it because of the humidity range that has to be accomodated?

We thank the reviewer for this remark. What defines the minimal acceptable concentration for isotope ratio determination at a required level of precision, is not only the absolute concentration (expressed in ppm) but also the isotopologue abundance, which is of 1.11 % for $^{13}CH_4$ and of 0.2% for $H_2{}^{18}O$ and 0.03% for HDO (see Table 1, in The HITRAN2020 molecular spectroscopic database, 2022). This means that at equivalent concentrations, the number of molecules is 5 times lower for $H_2{}^{18}O$ and 40 times lower for HDO compared to CH_4 isotopologues.

Another point to consider could be the intensity of the targeted near-infrared transitions. The spectral area depends on isotopologues availability in a reduced frequency range (limited by the maximal frequency range of the laser source), and is also the result of a number of compromises, such as reducing the intensity of the interfering gasses and reducing spectral congestion. If we take the example of the wavelength range used by Picarro G2201-i, for CH4 isotopes measurement, we found (see in Rella et al. 2015, Fig. 1 and Defratyka et al. 2021) an absorption of ~1E-9/cm at the wavenumber 6029 cm-1 for $^{13}CH_4$, for 2 ppm of concentration. For water, an equivalent absorption would be obtained at a concentration of ~ 100 ppm for HDO and ~ 10 ppm for $H_2{}^{18}O$.

Finally, we are not familiar with the case of CH_4 isotopes, but as mentioned by the reviewer, it is true that working with low mixing ratios with water molecules instead of CH_4 adds biases in the calibration because of adsorption on the tubings (resulting in residual water mixing / memory effect, discussed further), which makes it more difficult to obtain a given level of accuracy.

Then in more detail, going through the paper:

Figure 1 The methane interference can potentially be severe for low humidity (1/100 to 1/1000 of the humidity in this plot). Apart from this plot, this interference has not been

discussed or even mentioned anywhere else. Is the methane spectrum always included in, and corrected for in the fits?

The methane spectrum is taken into account in the fits and calibrated with atmospheric air (we added few words on this in subsection 1.1). However, the humidity dependency calibrations performed in the lab do not contain CH4, as synthetic air bottles are used (for more details, see Zero Air and Zero Air Plus here https://www.airproducts.co.uk/gases/zero-air). Using a calibration gas containing a typical atmospheric proportion of 2 ppm of CH4 could bring possible effects at low humidity on the calibration curve. We are investigating how to set this up in the field, since this matrix effect could effectively affect the true value correction. We are developing this point in section 3.2, under the paragraph "Humidity and isotopic composition dependency".

We are confident that this will not change the main conclusion of the article, as considering a potential CH_4 effect should not impact the intrinsic linearity and precision of the instrument.

Figure 2 It is not clear to me how the residuals in cm-1 relate to the arbitrary normalized intensity (in arbitrary units). Even the x-axis is in free spectra range steps, not in cm-1. Please clarify.

Thank you for pointing out this inconsistency on the y-axis of the figure. The correct unit is now displayed in cm-1 for the absorption spectrum and for the residuals. We also changed the left side y-axis which is now labeled "Measured absorption".

The x-axis could be expressed in terms of cm-1 or in nm as shown in Figure 1, by knowing the absolute wavelength and that one FSR corresponds to 188 MHz. However, contrary to Figure 1, we indeed do not show the absolute wavelength but rather the relative frequency in units of FSR, as we want to highlight how the spectrum is constructed (i.e. that the sampling points are defined by the absorption value at optical resonances).

Compared to figure 1, there is a flat baseline here in figure 2. How can that be, as all lines in figure 1 will be proportionally lower?

The plotted intensity in Figure 2 corresponds to the measured absorption when pure nitrogen is injected. Fig 2 shows the absorption at 3 ppm H2O, for which the baseline is extremely close to the empty cavity losses corresponding to the visible baseline offset, which is around 0.342E-6/cm expressed in absorption per length unit.

Line 169 and elsewhere: suddenly decimal commas instead of points. Happens more often, with the declaration of the values for reference waters.

This has been corrected.

Figure 5 the number of points for the Picarro instrument are much higher than for the AP2E one. Why? DIfferent strategy?

Table 1 should include the number of measurements (so much more for the Picarro I presume)

Thank you for this comment. The calibration dataset presented in this article covers the period ranging from January 28th 2023 to January 28th 2024. It is composed for Picarro of 161 (AO1 standard) and 158 (FP5 standard) calibration points and for AP2E of 150 (AO1 standard) and 148 (FP5 standard) calibration points. From this dataset, we only keep the calibration points successfully performed on both standards and also add the two sigma filtering on the humidity signal. This gives for Picarro 146 calibration points and for AP2E 138 calibration points for both standards, which are now annotated in table 1.

We acknowledge that the choices we made to show the calibration points were not well suited. Indeed, it gave the impression that the number of Picarro points are much higher than for AP2E. This is due to the fact that the Picarro curves were placed on top, making many of the AP2E points invisible. In addition, the error bars reduced visibility and accentuated this impression, while providing no particular information. We have therefore decided to remove the error bars, indicate the average instantaneous noise over the whole series in the figure legend and improve the display of the curves.

Lines 190 further. This humidity / mixing ratio dependence is widely observed, not only for water vapour measurements, but also for isotopes in atmospheric CO2. In that field, two ways of dealing with it are in use: the 'ratio method', so building ratios first, and then correct for mixing ratio dependence (this is what you do here), the other method is the isotopologue method, which would first analyze the different isotopologues as different species, calibrating first them, and only then build ratios. See for instance papers:

Flores, et al. Calibration Strategies for FT-IR and Other Isotope Ratio Infrared Spectrometer Instruments for Accurate delta C-13 and delta O-18 Measurements of CO2 in Air. Analytical Chemistry {89}, {3648-3655} (2017).

Steur et al. Simultaneous measurement of δ 13C, δ 18O and δ 17O of atmospheric CO2 – performance assessment of a dual-laser absorption spectrometer. Atmos Meas Tech 14, 4279–4304 (2021).

Both have their pros and cons.

Would that be something to try out here? Or at least to mention and discuss. Or is it not applicable at all in this context?

Thank you for this useful comment. We now add citations, at the beginning of this section, of different application which use the mixing ratio dependency correction including the paper from Flores et al., and we specified that we are using the "ratio method" for our study:

We present in this section the characterisation referred in the litterature as the mixing ratio dependency, which is used in various atmospheric isotopic measurements such as O2 (Piel et al., 2024), CO2 (Flores et al., 2017) or H2O (Weng et al., 2020). We use in this study the

most common, "ratio method", which consists in calculating first isotopic ratios from the measured optical spectrum, and then correcting it from the mixing ratio dependency.

To the best of our knowledge, no similar comparison has been performed in the literature for water vapor isotopes. However, we think that the water mixing ratio calibration section is already long (2 pages out of a 13-page article) and will not develop this aspect further for the sake of readability.

FIgure 8 : The Allan deviation is only part of the final uncertainty: the humidity correction and the calibration error also contribute. While that is less important for tracking the diurnal cycle, the calibration error will influence the accuracy of the seasonal cycle.

This is true and very important to have in mind when looking at absolute isotopic values, although calibration errors are complicated to quantify. We've added a discussion mentioning this in section 3.2, in the paragraph "Humidity and isotopic composition dependency". What we want to emphasize with fig. 8 are the OF-CEAS limitations related to the precision of our signal. This characterization shows the minimal measurable water mixing ratio without taking into account error and biases introduced during the calibration procedure.

Line 300 see above: the humidity dependence of dD is relatively speaking larger than that of d18O, and also it calibration uncertainty (fig 7) is, relatively, larger. Would that influence your conclusion?

The uncertainties have to be compared to the diurnal (or seasonal) cycle, which has an 8-times large span for dD compared to d18O. This means that the absolute uncertainty on the dD signal can be 8 times larger to allow equivalent "results" in terms of diurnal cycle interpretation. If we look at the maximal range between the two standard calibration curves (fig. 6, left panel), ie at 50 ppm, we find a difference of 8 permil for d18O and of 80 permil for dD, resulting indeed in a slightly larger difference for dD even if we consider the factor of 8. Concerning the uncertainty from Fig.7, we stay below the factor of 8 for dD compared to d18O, which allows us to consider both isotopes with the same level of confidence.

Finally, we observe that the high humidity dependency is lower for dD, which makes it interesting for interpretation in the high humidity regime, and would provide less calibration errors.

Line 342 "applying a drying on the humidity generator" ?? unclear to me what you mean.

Thank you for highlighting this, this was indeed unclear. This has been changed by :

"sending dry air through the humidity generator chambers and tubings"

Lines 355-365 (and figure 6) I agree with your conclusion that while the gradual, linear dependence is indeed caused by a spectral 'misfit' (probably indeed the interference of the very strong lines further in the spectrum, or methane?), the low humidity part strongly indicates sample-to-sample memory effects. This is further supported by the fact

that the depleted ref goes up, and the 'enriched' one goes down.

This effect will not fully vanish in the station, because (1) the residual water vapour is probably quite fractionated, and will thus still be different from outside humidity, and (2) you must calibrate your instrument regularly using two waters with very different isotope values.

I would like to see more discussion of these low-humidity part effects (after all, that is the truly innovative part of the instrument and your paper): for example, are the values (in fig 6) influenced by the 'sample history' before one of the reference waters and how would that be different in the station? Would you expect that sample water vapour with isotope values intermediate between the 'high' and 'low' ref waters scale linearly between them (as you suggest in lines 250-253)? What is the added uncertainty in this region?

We agree with you that the memory effect will not fully vanish, although it will be lower in the stations. In fact, we think that this effect is mainly driven by atmospheric, fractionated water vapor sticking to the tubings: calibration represents less than 5% of the measurement time and most of the time the tubings are exposed to lab air.

Indeed, the fact that the “low ref goes up” and “high ref goes down” indicates that an “enriched” residual water could be responsible for this low humidity divergence, possibly originating from the environmental isotopic composition in the lab. To check this hypothesis, we corrected the calibration curves by assuming a mix with a residual water with a concentration situated between 10 to 20 ppm and an isotopic composition estimated by sending dry air in the instrument through the same tubings and without any water injection. This correction effectively flattened both d18O curves (fig. 6, top left), but not the dD curves (fig 6, bottom left) which do not show any symmetry. This indicates for us that residual water can not be the only source of this low humidity divergence, and that other phenomena, like spectroscopic effects could be responsible for this. This would need a complete and more specific study which is beyond the scope of this paper. We thank the reviewer for having raised this question, and think that a next study focused on this particular question could bring a lot to the community of atmospheric water isotopes.

Here are the answers to your questions:

- concerning sample history: To reduce at maximum this history, during the lab humidity-isotope calibration we always start with the high humidity injections (1000 ppm) and finish with the low humidity step (50 ppm). This means that when the low humidity step is performed, the tubings have been flushed with the same standard for at least 10-15 hours. The 6 calibrations performed across 3 months show no particular trend, which gives us a good confidence on the repeatability of the calibration. No difference was observed by switching the order of the calibration standard.
- concerning the linearity and uncertainty between the low and high ref : in Figure 7, the isotopic compositions measured with TD3 (with corresponding humidities lying in the divergence area, from 67 to 700 ppm) have been corrected using a linear

combination of the highly and lightly depleted standard calibration curve, resulting in a linear dependency between the measured and true value, which supports (at least at first order) the hypothesis of a linear scaling between the calibration curves. The accuracy and precision are given in section 2.3: this uncertainty integrates the instrumental drift over a few months of calibration and the deviation from linearity that one could expect between the high and low ref (indeed, Weng et al. suggested a quadratic relation between the a, b and c parameters used for the mixing ratio dependency fitting)

We are currently including some of the most important elements of this discussion in the manuscript in section 3.2, under the second paragraph renamed "Humidity and isotopic composition dependency and calibration uncertainty".

Line 364-365: 'We insist thus on the importance of calibrating the instrument in the field to correct for those artefacts (Casado et al., 2016).' Indeed! May be change the word 'insist' into 'emphasize' ?

Thank you for this suggestion, the word has been changed.

line 371-372 "The internal architecture of these analysers therefore reduces the risk of breakdowns during the deployment, but requires an expertise to finely tune them. "

To me this feels like a blessing in disguise, or may be just the exact opposite. Deploying these instruments thus always requires expertise (and time and some equipment) in the field. Any more comments to that? For example negative field experience with fixed mounted equipment such as the Picarro's ?

Thank you for this question. The paragraph "interesting features for field operation" has been modified and Picarro was clearly mentioned to avoid any ambiguity (see answer to "Reviewer 3" for more information).

I agree that working with these instruments always requires qualified staff, and it's clear that Picarro devices have the advantage of being more user-friendly, enabling deployment without any specific expertise (e.g. skills for optical alignment). The downside of this ease of use is that in the event of a breakdown, the system is not sufficiently open, making it impossible for users to intervene "in depth" in the device. Since remote intervention via remote desktop is complicated in Antarctica, this would mean sending the instrumentation back to mainland France for repair, and thus abandoning the mission.

"Competing interests. The authors declare that they have no conflict of interest."

Two of the co-authors work with the (commercial) producer of the instrument. How do they avoid a conflict of interest?

Indeed, thank you for this remark, this work has been performed in collaboration with the instrument producer. We changed the competing interest section with this new paragraph:

This work was made possible by a collaboration with the company AP2E who produced the analysers presented in this manuscript. The characterisation of the analysers was done independently at LSCE, with no interference from AP2E.