Response to Reviewer 6 on review of "A versatile water vapor generation module for vapor isotope calibration and liquid isotope measurements"

We are grateful for the comments and suggestions provided by Reviewer #6, which are truly valuable and insightful. The comments have made us reflect on how to improve future versions of both the vapor generator and the control software. We have below responded to the reviewer using red font. All 17Oexcess values have now been calibrated as suggested by the reviewer.

General Comments

The manuscript details a field-deployable custom vapor generation system that will facilitate field measurements of water isotopes; the analysis demonstrates the analytical advantages of the method relative to vaporization units that rely on discrete injection. The manuscript will benefit readers of AMT who utilize water isotope analysis systems in the field and laboratory. The authors claim that the custom unit will benefit field measurement campaigns because the vaporization unit is portable, easy to service, and can measure reference waters more quickly than commercially available units and with less isotope memory. The authors tested the system and show that it meets analytical precision targets. They also conduct a stability test of the vapor generation module by measuring the vapor stream with two Picarro instruments simultaneously to show that the noise from the vaporizer system is small relative the noise of the instruments. More information about operational protocols and maintenance of the vaporizer unit will help convince a reader of its field-worthiness.

Thank you for this advice. As we described in the manuscript we had partial success to unclog the capillaries using an ultrasonic bath. We have also tried to push citric acid through the capillaries, but it is not always working. We have since the submission of the manuscript found that using in-house prepared stands either from miliQ water from Bermuda or a mixture with Greenland snow allowed us to run the system in the laboratory every day for a 6 months period. We have updated the text about the length of operation in the manuscript such that it now reads: "We have used the calibration system in the laboratory and during field campaigns for about 2 years now and found that a stable performance of the vapor generation module is dependent on using clean standards. When using in-house generated standards consisting of a mixture of melted Greenland snow and milli-Q water from Bermuda we have successfully operated the vapor generation module daily (roughly 1-3 hours every day) in the laboratory for more than 6 months without changing the capillaries."

The apparent relationship between cavity temperature fluctuations and d18O on timescales of ~tens of minutes is an important finding of this analysis and should be highlighted; as the authors suggest, if the instrument temperature could be stabilized, this would improve precision for d18O, but this also has likely significant implications for the deuterium- and 17O- excess measurements as well.

Thank you. We agree that this is an important finding and have tried to stress this in the manuscript text and in the abstract:
We write in section 4.1
"We also further notice that this increase in Allan Deviation occurs at similar integration times as used for liquid measurements, which makes it even more important to improve on."

And in the Abstract:
"Using the vapor generation module, we document that an enhancement in the Allan Deviation above the white noise level for integration times between 10 minutes and 1 hour is caused by cyclic variations in the cavity temperature, which if improved upon could result in an improvement in liquid sample measurement precision of up to a factor 2"

Reorganization and reprioritization of some discussion points would improve this article. As it is written, the abstract does not summarize or highlight the most important outcomes of the study. Instead of focusing the abstract on theoretical future applications (i.e. instead of saying what it "could in principle" do), please revise the abstract to **document the novel advantages of the vapor generator and highlight the new results that are achieved by the tests described in the manuscript**.
We have since the submission of the manuscript been able to operate the instrument for more than 6 months without the need to do any maintenance. It has therefore been possible for us to update the abstract to now read:
"The vapor generation module as a calibration system has been documented to generate a constant water vapor stream for more than 90 hours showing the feasibility of being used to integrate measurements over much longer periods than achievable with syringe-based injections as well as allowing the analysis of instrument performance and noise. Using clean in-house standards, we have achieved to operate the vapor generation module daily for 1-3 hours for more than 6 months without the need for maintenance, illustrating the potential as a field-deployed autonomous vapor isotope calibration unit. "

We have also removed the sentence including the "could in principle" when dealing with performance, but not when it comes to number of standards/samples to be connected as it does not make sense to give an actual number as the restriction is physical space.

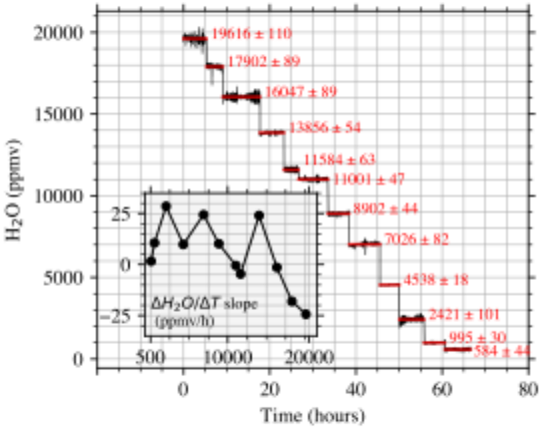
The vaporizer module is of course important for its potential to field-calibrate, but it is also useful for the analysis of instrumental noise, examining isotope-humidity dependence, and other performance metrics that are shown in the paper and deserve emphasis. For example, the abstract could perhaps highlight the vaporizer's ability to generate continuous water vapor for a wide range of humidity levels (which is an advantage for vapor measurements over CFA vaporizer configurations like Gkinis, Jones, or Davidge that target a single water vapor concentration) with the good stability at each level but increasing memory issues at low humidity.

Thank you for this suggestion.. We have now updated the abstract to read:

"The vapor generation module can generate a stream of constant vapor at a wide variety of humidity levels spanning 300 ppmv to 30 000 ppmv and is fully scalable allowing in principle an unlimited number of standards or samples to be connected. This versatility opens up the possibility for calibrating with multiple standards during field deployment including examining instrument isotope-humidity dependence."

We do not find evidence that the vapor generation module shows an increasing memory effect at low humidity. Perhaps the reviewer is referring to decreased relative humidity stability at low humidity. This could perhaps be explained by the higher dilution flow rate, which could increase the pressure difference between the capillary and the open split. Should variations in this pressure gradient exists due to MFC instabilities there would be variations in the delivery of the liquid into the oven through the capillary leading to humidity variations.

We have further more include a figure (Figur 5 in revised manuscript):



Illustrating the stability of the humidity levels during a humidity-isotope calibration.

The comparisons with discrete injections (such as in Figure 4) are important documentation of the advantages of continuous over discrete vaporization, but do not uniquely reflect the contributions of this vaporization unit over other continuous vaporizers designed for CFA – the abstract (and manuscript) could better contextualize the new contributions made by the authors with this custom vaporizer through direct comparisons to other published calibration methods and continuous vaporizer units.

We have improved the abstract with the following sentence:

The vapor generation module can generate a stream of constant vapor at a wide variety of humidity levels spanning 300 ppmv to 30 000 ppmv

This versatility opens up the possibility for calibrating with multiple standards during field deployment including examining instrument isotope-humidity dependence. Utilizing the ability to generate an uninterrupted constant stream of vapor we document an Allan Deviation for ^{17}O-excess (Δ^{17}O) of less than 2 per meg for an approximate 3-hour averaging time

Using the vapor generation module, we document that an enhancement in the Allan Deviation above the white noise level for integration times between 10 minutes and 1 hour is caused by cyclic variations in the cavity temperature, which if improved upon could result in an improvement in liquid sample measurement precision of up to a factor 2.

The vapor generation module as a calibration system has been documented to generate a constant water vapor stream for more than 90 hours showing the feasibility of being used to integrate measurements over much longer periods than achievable with syringe-based injections as well as allowing the analysis of instrument performance and noise. Using clean in-house standards, we have achieved to operate the vapor generation module daily for 1-3 hours for more than 6 months without the need for maintenance, illustrating the potential as a field-deployed autonomous vapor isotope calibration unit. When operating the vapor generation module for laboratory-based liquid water isotope measurements we document a more than 2 times lower memory effect compared to a standard autosampler system.

Finally, the authors aim to demonstrate that the custom vaporizer has sufficient signal stability to measure 17O-excess, but **the authors need to be more cautious about the treatment of raw data to assess performance for 17O-excess**. It is misleading to assign units of per meg to the raw data since the raw signal variability is not equivalent to the calibrated signal variability. This treatment potentially affects data shown in figures 3, 5 and certainly in figures 8(i,j) and S2.

We agree with the reviewer that raw signal variability is not equivalent to calibrated signal variability. For this reason, we ran all the analysis and prepared new plots using calibrated data. The calibration lines were defined during the SP BER step change in the 90 hours long run. The system was primed with standard water vapor for 2 hours and the measured δ^{17}O δ^{18}O δD were defined by averaging 1 hour of data after the priming. The new Figures 3 and 5 have changed in the manuscript, accordingly. It should be noted that changes in the plots are extremely small because calibration affects only to a minor extent the signal variance (i.e. the slope of the calibration line is ~1 ‰/‰).
Following suggestions by reviewer #6, we also have edited Figure 8, which now reports the variability of the raw signal for each pulse around the mean calculated from all the pulses.

Previous work has established accuracy (not precision) as the dominant error in CRDS 17O-excess data, so while the precision seems great for these data, more work is needed to show that system operating conditions like automated, variable flow rates or secondary air dilution do not create calibration bias for 17O-excess. **Though small, the systematic offset in the calibrated 17O-excess data further suggest that there likely is a bias in the calibration, which should either be examined further to better characterize the limitations of this method or should be qualified appropriately in the text when claims are made about the quality of 17O-excess data.**

We agree with reviewer #6 that accuracy in 17O-excess is more problematic to handle rather than precision, since our manuscript and others highlight very high stability for 2140 models. Specific guidelines and *certified* 17O excess values to perform laboratory calibration should be provided by IAEA to this end. With the tools in our hand, we are only able to account for a potential error introduced by the weighting procedure for the preparation of the mixtures. This is described in the text as the maximum span obtained by mixing higher amount of the first standard (+0.01g) and smaller amount of the second standard (-0.01g), and vice versa. This conservative estimate of the sample uncertainty yields a 13 per meg variation of the final Δ17O value, which is actually larger than the inter-run precision we obtain.

We have further added the following line: "...but within the expected uncertainty due to the weighting uncertainty, which could produce an off set of up to 13 per meg."

Specific Comments

Some sections of the text need to be revised for clarity and completeness. In some sections, the claims that are made are not directly supported by the evidence provided in the tables/figures so it can be difficult to evaluate some statements. The whole manuscript would benefit from a careful reread and review by the authors.
We acknowledge that too many grammatical errors occurred.

The abstract claims that the vapor stream is constant for 90+ hours and that it therefore could be used in the field for more than three months, but it is unclear what "constant" means in this context, especially since the water vapor data that is shown in figure 3 exhibits some notable variations. It is also unclear (in the abstract) how operating for 90 hours in the lab translates to three months in the field – this is explained at the very end of the manuscript (in that it will theoretically be measuring standards for 1h/day) but is confusing in the abstract since it is not explained. **Further, the 90hr to 3 month relationship is speculative at best, since it has not been demonstrated that the unit will, for example, not clog with precipitates or encounter**

other operational setbacks over the three month window. More information about typical or intended operations of this system will help the reader better understand its advantages and limitations.

As described above we have updated the claim based on the operations since submission of manuscript:

"We have used the calibration system in the laboratory and during field campaigns for about 2 years now and found that a stable performance of the vapor generation module is dependent on using clean standards. When using in-house generated standards consisting of a mixture of melted Greenland snow and milli-Q water from Bermuda we have successfully operated the vapor generation module daily (roughly 1-3 hours every day) in the laboratory for more than 6 months without changing the capillaries."

We further have added the sentence at the end of section 4.3 illustrating the importance of using clean standards:

While our experience using our in-house clean standards shows stable performance by the vapor generation module, we have also used standards, which were provided to us by other labs. In those cases, we observed that the capillary would get clogged more frequently and had to be replaced every 2-4 weeks. To extend the lifetime of the capillaries it is possible to use a larger ID, but this can potentially results in less stable humidity values.

When using the vapor generation module for unknown liquid sample measurements it is hence likely that one need to include a change in capillary as part of operational procedure. While not discussed in the manuscript there is a price advantage of the capillaries (~10 EUR each) compared to syringes used in an autosampler (~100 EUR each)

There is a lot of confusing or vague language throughout the document and also many acronyms that have not been defined – please try to define acronyms before using them in the text or figure captions and make sure details of system components are explained the first time they are mentioned.

We have done this now for all acronyms.

Ln 53-55 suggests that excess values require a "relatively high output of individual number of samples measured" – can you clarify what you mean here?

What we meant was that the use of d-excess and 17O-excess still require a relative large number of samples to be measured in order to make robust conclusions, but we agree this is not clear in text and we have therefore removed that part of the sentence.

Ln 72-75: define "low uncertainty" and "large quantities".

Corrected:
(providing calibration pulses with an uncertainty of +/- 0.1/1.0 ‰ for d18O/dD)

And
(3-5 litre depending on deployment period)
Ln 91: Davidge et al 2022 utilizes a unique vapor generation system so “this system” should instead say something like “a similar system”.
Indeed - this is corrected now

ln 151: instead of noting the differences between this vaporizer and an earlier version of this vaporizer, this might be a good place to describe the differences and advantages of this system over other types of vaporization systems (e.g. it adopts the multi-channel selector valve of Jones et al. 2017, similar continuous vaporization setup to Gkinis but with vacuum pump, PID control for humidity, additional mixing tee, etc.).
Reviewer #5 have asked us to further expand on the differences between this vapor generation module and earlier versions hence we would like to keep this list drawing attention of the reviewer on the detailed improvements.
We believe that we describe in the section under the bullet point list the reasons for each addition as well as also references Jones et al. 2017 for the inspiration to connect a selector valve.
It is of course also clear, as we hope is illustrated by our referencing that we are building on the last decade of CFA and vapor calibration developments.

Many of the details noted on page 6 would be better left to the later sections of the paper since a general reader might not understand the specifics about salt deposits, number of ovens, valve circuitry, etc. at this point in the paper.
We would prefer to keep this detailed section under Methods and leave the Discussion section to performance.

Ln 190: if the pressure regulator resolution is 0.01psi, why is the regulated range of pressures that is listed so large (0.5-3.5psi)? How is this system typically operated for each humidity level, and how much does the pressure fluctuate for each humidity level during a typical measurement? More operational details are needed to help a reader duplicate this work.
0.01 is the resolution of the electronic pressure controller but to observe a significant variation of humidity (>100 ppmv) the headspace pressure needs to be changed in the order of 0.1 - 0.2 psi. Typical pressure applied using a new capillary for generating humidity levels around 5-10kppmv is around 1 psi. However, as the capillary becomes clogged the pressure can be increased to 2 psi. An experienced operator can also start out by a relative high pressure of 1.5-2 psi to initiate the flow of water through the tube and then reduce the pressure at the first sign of water being delivered to the oven. However, one need to be careful about pushing too much water into the oven and hence flooding it.

We haven't log the pressure fluctuation in the headspace, maybe we can show a step change in headspace pressure and step change pressure in humidity in this document?

All tables and figures should have sufficiently detailed captions so that the reader can easily understand what data is shown – please reread all figure and table captions and try to add more information.
We have tried to fix this

Ln ~265: perhaps it would be advantageous to highlight the regions of the plot that the author is describing in the text when talking about the slope of the allan deviation with time.
Due to the fact that white noise seems to be dominant for different periods for d18O compared to dD we think it will be difficult to illustrate a specific region, when it comes to white noise. We have, however indicated the expected slope for a decrease in Allan deviation when white noise being dominant.

Ln 271-272: figure 3 is important both because it documents the ability of this vaporizer unit to generate consistent isotope data, but also because the authors have highlighted the difference in the allan variance analysis when truncating the data to account for memory effects within the system. The vaporizer unit exhibits excellent performance and this should be highlighted, but it is incorrect to say that the performance is better than that of other systems, since other systems show similar precision at these averaging times (e.g. Gkinis et al. 2010 for dD and d18O, Steig et al. 2021, or Davidge et al. 2022 for 17O-excess). It is therefore also incorrect to suggest that no previous work has managed system memory effectively.

We agree with the reviewer that to a large extend the performance for the vapor generation unit presented here show similar performace as Gkinis et al. 2010, Steig et al. 2021, and Davidge et al. 2022. However, there are some small but important differences.

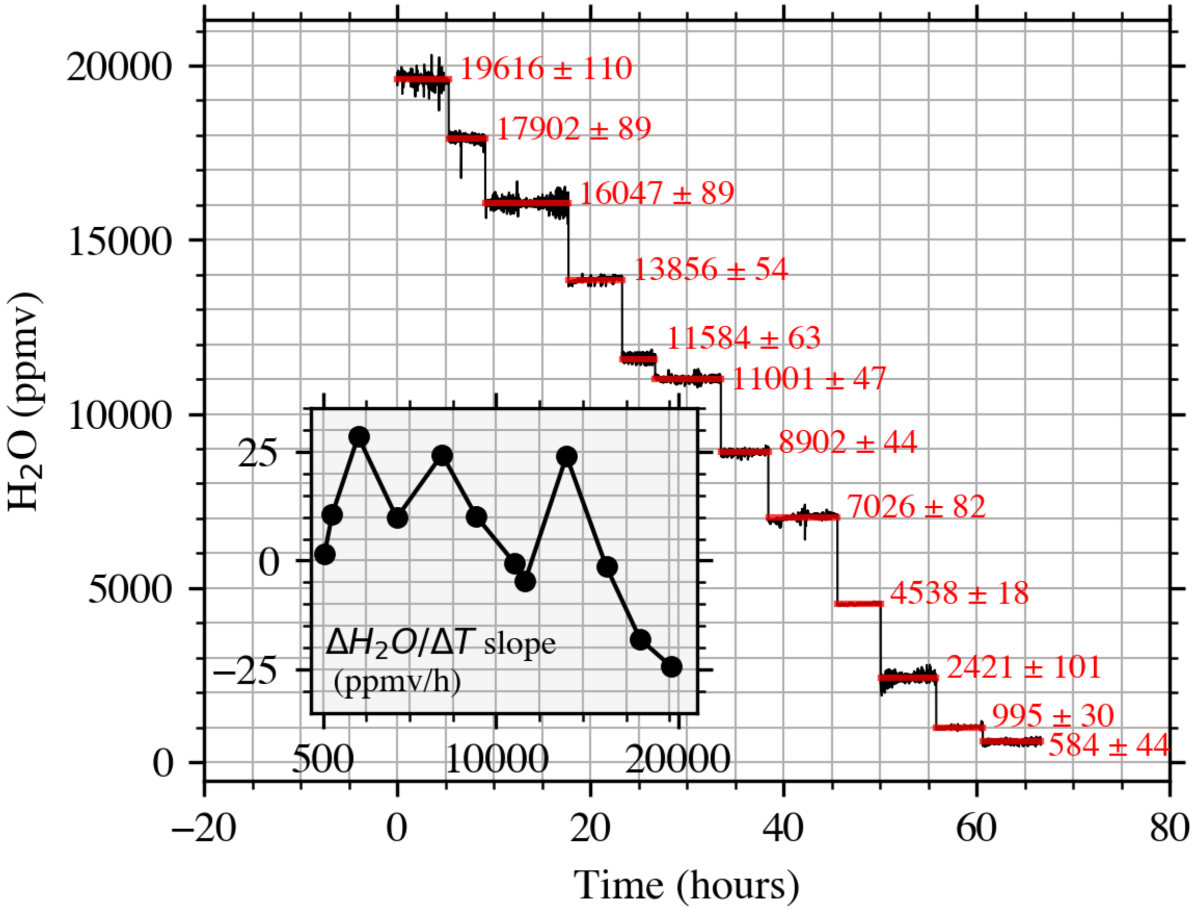
For dD we obtain a minimum in Allan Deviation at 1.5e4 s with an Allan Deviation of 1e-2 permil.
In Gkinis et al. 2010 the minimum is achieved at 3e3 s with an Allan Deviation of 3e-2 permil. After integration time 3e3s the Allan Deviation deviates from the -0.5 sigma_allan/Tau, white noise line, indicating either drift of the instrument of not completely removed memory effect. Comparing to our estimated Allan Deviation at 3e3s we obtain 2e-2 permil, which is interesting since it indicates that the difference between a 2140 and 1100 series picarro is a 30% reduction in noise for dD (For d18O the difference is 8e-3 vs 5e-3 at 7e3s integration time). As the d18O Allan deviation curve seems to continue decreasing after the dD curve has become constant seems to indicate an issue with memory effect.

For Steig et al. 2021 the Allan deviation for dD seems to deviate from the white noise line already around 1e2 seconds, while d18O continues to decrease down to 6e3 seconds. However, the “bump” hypothesized being driven by cavity temperature is occurring between 1e2 and 8e2, which also has an effect. However, it seems that memory effect is influencing the results.

For Davidge et al. 2022 no Allan deviation for d18O/dD is shown, but instead shows an allan deviation plot for 17O-excess. The Allan deviation plot hower does not extend past 6e3 s integration time. In our manuscript we present an Allan Devation plot extending to 8e4 second integration time - documenting a minimum at 2e4 second integration time at 1.3 per meg. Davidge et al. 2022 obtains a minimum of 3 per meg 3e3 second integration. This is comparable to our results.

We do believe based on the above argument that our system is presenting an improvement.
We have however removed the “suboptimal”-word here.
We have also added the following sentence: “Albeit for 10^3 seconds, we show similar Δ^{17}O $\boldsymbol{\sigma}_{\text{Allan}}$ as Davidge et al. (2022)”

The authors might consider showing the data from Table 2 as a figure, or perhaps showing the full sequence of measurements made over time to help the reader understand the tests that were conducted.
We agree and have added the following figure.

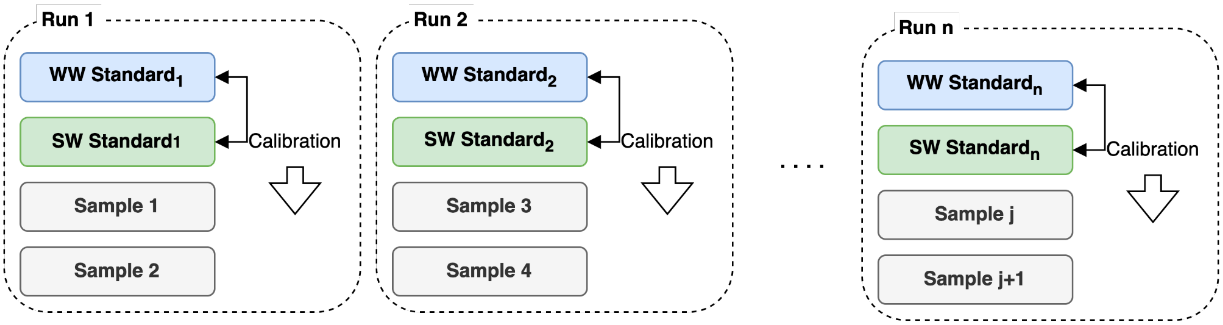


The 12500 ppmv threshold finding is very interesting, especially since the allan variance test shows the worst performance in this region. Have you tested whether this is a physical effect of the capillary diameter or some other design choice? Were all data generated for this study using the 127um capillary?

We have unfortunately not carried out a complete evaluation of all capillary diameters.

Ln 397: this replication looks promising, but more information about the calibration is needed here. This section should also include that this performance is comparable to other continuous vaporizers that have been developed for 17O-excess measurements (i.e. Steig 2021, Davidge 2022).

Following our answer to reviewer #3, a typical measurement session (run) is composed of analysis of two standards followed by analysis of two samples, as shown below. Independent calibration means that each run is calibrated using the two standards of that specific run. Average calibration means that the results of the analysis of the standards have been averaged and an average calibration line was built to calibrate the analysis of the samples.



This procedure is now described in the text as follows:

Since the standards were measured from 2 to 4 times for each measurement session, and the solutions were measured on different days, we applied both an average and independent calibration. The average calibration factors were estimated by averaging the raw observations of the standards injected for all the run. Instead, the independent calibration factors were calculated for each couple of reference standards and then applied only to the following couple of solutions (2 standards followed by 2 samples). The results of the experiment are reported in Table 3.

Also the following text has been edited:
Such reproducibility is comparable to precision achieved with IRMS (e.g. Barkan and Luz, 2005; Steig et al., 2014) and to the total error obtained within the latest development in continuous flow analysis of ice cores (Davidge et al., 2022) but better than analysis performed with optimal settings of vaporizer and autosampler from Picarro (which is 8 per meg following Schauer et al. 2016).

Ln 401: The offset is likely due to calibration, so additional details about how these data were calibrated would be useful. It seems unlikely that the age of the reference water has modified the 17Oexcess values if they have been in sealed containers and cold storage, but it is possible that the systematic bias is due to an error in the calibration standard assignments or that it is generated by the vaporizer unit itself, which must be carefully examined. Which standards were used for calibration? How frequently were they measured? How stable were the raw values in d18O/d17O between reference water measurements? Without this information it is impossible to speculate what might be the cause of these offsets.

We are not completely sure about the origin of the offset. We suggested an error in the weighting procedure as a potential source of uncertainty but also error in the calibration standard assignments is a potential uncertainty, as the referee suggests. In principle we should send our standards to another lab and perform a lab-intercomparison, which is however at the moment is out of the scope of this manuscript (here we focus on the precision and repeatability of the measurement).

We believe that a bias generated by the calibration system itself is less likely, since flash evaporator is a well-proven technique for water stable isotope analysis. Moreover, we performed internal laboratory calibration using VSMOW-SLAP2 using both the Picarro's liquid injection mode and the calibration system and we have obtained identical d18O, d17O, dD values for the two methods (the data is not reported in the study).

Answers to specific questions in the comment.

Which standards were used for calibration? The standard used were always SW and WW, the same standards used to produce the mixtures.

How frequently were they measured? The standards were analyzed every two samples (Please see figure attached for Reviewer #3. That means the standards were injected between 2 and 4 times for each run.

How stable were the raw values in d18O/d17O between reference water measurements? 1 standard deviations (SD) and the standard errors of the mean (SEM) of raw reference water measurements across all the runs are reported in the following table. Values in permil (‰).

	δ^{17}O (SD)	δ^{17}O (SEM)		δ^{18}O (SD)	δ^{18}O (SEM)
SW	0.026	0.009		0.037	0.013
WW	0.017	0.006		0.032	0.011

Throughout the paper and abstract there are speculations about what could be done in principle, but it is important to properly document what can be done in practice, especially since these systems do require maintenance and cannot run indefinitely or measure standards frequently enough for perfect calibrations. How often is it necessary to clean the capillary? What maintenance was required for this system during the study period and with what frequency? How can one best operate a system like this within those maintenance limitations to maximize the quantity and the quality of the data? Details about operational controls, conditions, and maintenance would help the reader better understand the performance of this vaporizer system.

We have improved the text based on our experiences over the last 6 months running the system daily in the laboratory. We have also updated the manuscript to illustrate the need for use of clean standards. Specifically:

We have used the calibration system in the laboratory and during field campaigns for about 2 years now and found that a stable performance of the vapor generation module is dependent on using clean standards. When using in-house generated standards consisting of a mixture of melted Greenland snow and milli-Q water from Bermuda we have successfully operated the vapor generation module daily (roughly 1-3 hours every day) in the laboratory for more than 6 months without changing the capillaries.

While our experience using our in-house clean standards shows stable performance by the vapor generation module, we have also used standards, which were provided to us by other labs. In those cases, we observed that the capillary would get clogged more frequently and had to be replaced every 2-4 weeks. To extend the lifetime of the capillaries it is possible to use a larger ID, but this can potentially result in less stable humidity values.

Ln 465 should also acknowledge the increase in the delta values.
Yes - we have updated the text

Ln 469 seems like a major limitation of this method – perhaps the authors should repeat this test with larger vials to eliminate this enrichment issue.
We agree and have learned from our mistake. By including this illustration in the manuscript we hope that anyone doing similar test will not make similar mistake.

Ln 480: please define “relatively clean”
Yes - have updated the discussion to say that using milliQ water as standards did not provide any clogging within 6 months.

Ln 483: is 1h per day sufficient for the calibration of all water isotopes? Why has this duration been chosen?
We have removed the reference to 1 hour per day calibration in the manuscript. However, from our experience to obtain high quality d18O and dD water vapor isotope measurements a rule of thumb is to calibrate for 1 hour per day.

The discussion in ln ~510 and data shown in Figure SM2 suggest that the performance of this vaporizer is changing over these 48h of analysis due to the automation of air/water ratios and necessary reduction in flow rate to accommodate the formation of salts in the capillary. Have you tested this system with milli-Q or other treated water to minimize this effect, and have you seen any improvement in this performance? A change of ~3 per mil dD over this short timescale should probably be investigated further. The way this is accounted for in CFA systems is by keeping the liquid/air injection rates as constant as possible, because otherwise it is impossible to know what the effect of the memory is at any point during the analysis when the flow rates are changing during analysis. What range of flow rates does this study utilize and can you attribute any of the changes in system performance to these variables?

We were in fact debating whether to include the figure SM2 and the discussion in the manuscript or not as we do not think that such a change in measured standard values provide an appropriate image of the performance of the vapor generation module that we wish to convey. However, we chose to include this illustration as it provide a good example of how memory effect can influence the measurements of the standards or samples. In other words: by illustration we hope people will be aware of the pit-falls.
We appreciate the suggestion on keeping the liquid/air rates constant and we are planning on logging the flow rate of the air in any follow up versions of the vapor generation module.
Not directly relevant to the discussion here, but we have been analyzing the long-term memory effect of the Picarro High Precision Vaporizer by logging the memory correction values provided by the Van Geldern method. We discovered a slow by increasing memory effect of the vaporizer, which potentially can have influence on long term measurement performance. Hence, the issue with drift in memory effect is not only an issue for our system.

Ln 521: define "relatively high measurement uncertainty" and other vague quantities throughout the paper.

The manuscript has been checked and edited accordingly using numerical quantities. Specifically for Ln 521, the text now is:

Due to the relatively high measurement uncertainty (e.g. σ_{Allan} = 9 per meg @ 15 minutes) and relatively small changes in $\Delta^{17}O$ observed in natural waters (~90 per meg), it is not the memory effect, which is the limiting factor influencing the $\Delta^{17}O$ measurements.

Ln 523: change "error, which is" to "error that is" for clarity/correctness.
Corrected

Ln 524: without additional analysis of calibrated data it is a stretch to say that the module is "optimal for 17O-excess" but it is certainly promising to see such nice signal stability in the 17O-excess record. Previous work has established calibration as the major limitation on laser spectroscopy measurements for 17O-excess, so without additional analysis of calibrated data it is hard to accept these claims.
OK - we have changed the wording to "is highly capable of 17O-excess measurements"

Ln 525: similarly, while the unit can operate over long periods, unless it is possible to measure sufficient durations of the calibration standards for 17O-excess in between vapor samples, the resulting data could have large errors – this issue is discussed in both Steig et al. 2021 and Davidge et al. 2022. How long can the vaporizer operate before the capillary clogs or the flow rates change? This data would be important for understanding limitations around 17O-excess calibration.

OK we have updated the text to be more modest:

"Hence, the vapor generation module is highly capable of Δ^{17}O measurements as it can provide integration times over multiple hours of both standards and samples allowing optimal treatment of memory effects and measurement noise to be reduced to a minimum.

"

Though I look forward to following your updates in the future, Section 4.5 is not analysis of the data that is presented in this paper and could be better suited for a proposal.

OK - we have shorten the text to only discuss planned improvements for sample measurements.

Table and Figure Comments

Figure 1 – this will be difficult for many readers to follow. I suggest additional labels and defining acronyms and process control symbols.

DONE

Figure 2 – recommend additional labels to help the reader understand this figure

DONE

Figure 3 – clear figure. Can you comment on the variability in the water vapor concentration? Especially since earlier studies have linked water vaporization inconsistencies to isotope fractionation it seems important to better characterize these vapor fluctuations.

Figure 4 – great visualization of this relationship that shows a major advantage of continuous injection modes for laser spectroscopy. The legend is perhaps a little confusing because it is unclear what the legend means without also reading the caption.

DONE

Figure 5 – Why not show all water isotopes here? Also please consider revising the label on the y axis if the data used for this analysis has not been calibrated.

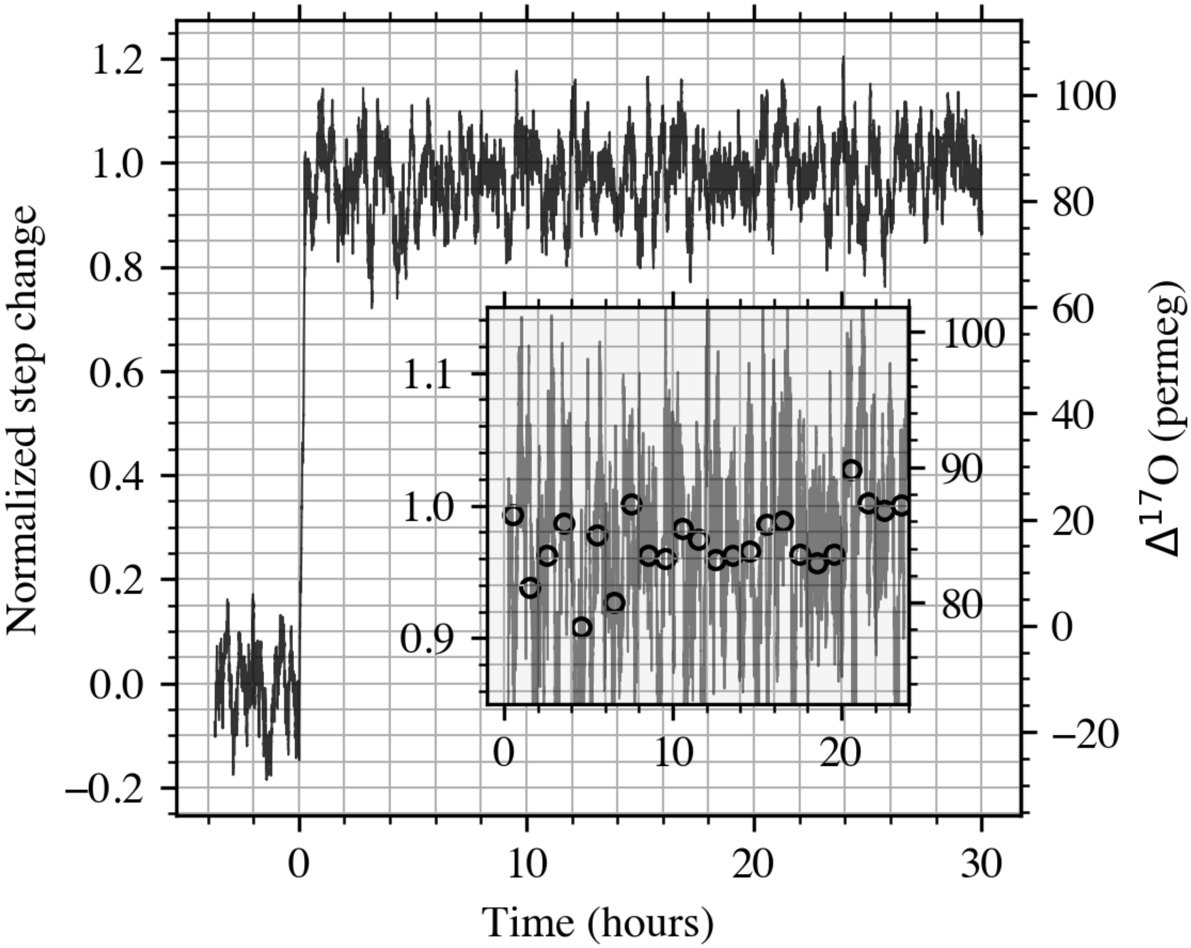
Additional plots in SM

Figure 6 – great figure with important implications for low-humidity measurements. Maybe instead of defining d18O_diff in the caption you could just label it d18O_3500ppmv – d18O or similar?
DONE

Figure 7 – if possible, moving the legend away from the data would make it easier to read panel A. This is a very compelling and disturbing result!
DONE

Figure 8 – the authors should either calibrate the 17O-excess data or find another way to describe the spread of the raw spectroscopy response – showing values of 300 per meg is very misleading! Large variability in the 17O-excess raw data suggests that the errors in d17O and d18O are not perfectly correlated, which is likely to cause accuracy issues in the calibrated 17O-excess values. Because the data is not calibrated it is difficult to evaluate this data – please provide a record of calibrated 17O-excess over time in the revision of this manuscript.
DONE, normalized

Figure 9 – this figure would be more useful if calibrated 17O-excess values were shown.
Here is the 17O-excess step change (calibrated)..



.

Table 1 – define BER and SP. What is the uncertainty of the values of the excess measurements and the SP measurements? Where were the data measured? Please specify whether these values are measured relative the VSMOW-SLAP scale. It would be helpful to combine Tables 1 and SM1 and show all waters here since the reference waters from the supplement are referred to in the text.

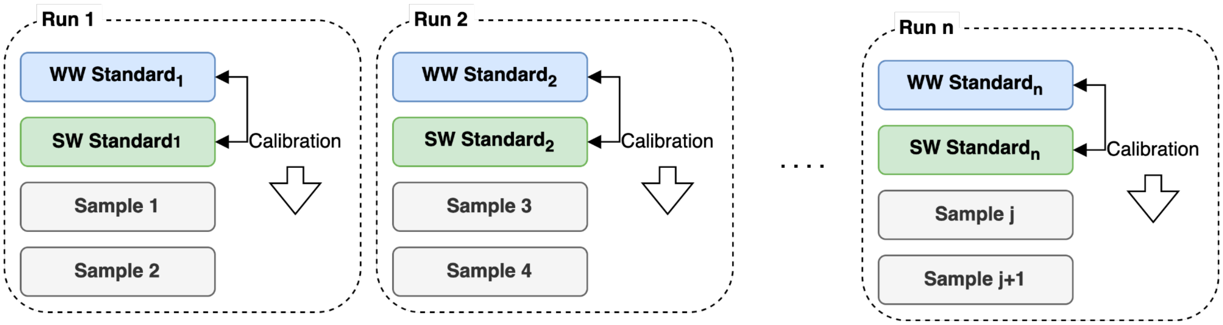
The standards used in this study have been provided by different stable isotopes laboratories (Laboratoire des Sciences du Climat et de l'Environnement, Centre for Ice and Climate at the Niels Bohr Institute, and the Stable Isotope Laboratory at the Institute of Arctic and Alpine Research, University of Colorado). Not all the reference values were provided with an uncertainty. We know that providing a reference value without any uncertainty is not best practice, but internal standards are subject to uncertainty and quality checks that are laboratory-specific. We treated all the standards as true values assuming no knowledge about their uncertainty. Hence, we removed the uncertainties from Table 1 for consistency.

Table 2 – as noted above, a plot of this data could be helpful for the reader to understand the different tests that were conducted. Is the standard deviation calculated for the raw data over the full duration of each test?

Also following the comment of the other reviewers, Table 2 has been converted into a figure. The table is now reported in the supplementary material.

Table 3 – please include details about how these measurements were calibrated.

Following our answer to reviewer #3: a typical measurement session (run) is composed of analysis of two standards followed by analysis of two samples, as shown below. Independent calibration means that each run is calibrated using the two standards of that specific run. Average calibration means that the results of the analysis of the standards have been averaged and an average calibration line was built to calibrate the analysis of the samples.



This procedure is described in the text as follows:

Since the standards were from 2 to 4 times for each measurement session, and the solutions were measured on different days, we applied both an average and independent calibration. The average calibration factors were estimated by averaging the raw observations of the standards injected for all the run. Instead, the independent calibration factors were calculated for each couple of reference standards and then applied only to the following couple of solutions (2 standards followed by 2 samples). The results of the experiment are reported in Table 3.

Table S2 – Please check for rounding error in the 17O-excess reference water assignments; from the isotope values and mixing ratios given I calculate 20, 11, and 17 per meg (not 21, 12, and 18), though this does not substantially change the result or interpretation.

There was a typo in the digits of d17O. However, the new values calculated are:
21,12,19 (only this one is different from the previous calculation).
This is the formula I have used:

(ln(**d17O**/1000+1)-0.528*LN(**d18O**/1000+1))*1000000
Also Table 3 in the manuscript must be changed.

Figure S1 - the memory effect for dD appears to be different during these two analysis windows; it could be interesting to examine the operating conditions during these runs and consider whether flow rates or other changed conditions could cause this difference.
Unfortunately we did not log the dry air flow rate, which will be key to be done in our future software updates.