Response to review by Reviewer #6

We thank Reviewer #6 for the detailed review. We have addressed the comments below in Red. We especially appreciate the in-depth questions, that allowed us to reflect on way to improve our system in future applications.

This manuscript has been greatly improved in scope, organization, and readability since the first submission. However, the details of each small experiment sometimes obscure the overall point of the paper, which is to demonstrate that this new vaporization system is a viable unit for field-calibration of atmosphere measurements and laboratory measurements of liquid standards or samples.

We acknowledge the point of the reviewer, but below provide a detailed argument why we suggest to keep the descriptions of the small experiments separate from the general message of the manuscript.

Development of the vaporizer allows for some interesting tests that provide important insights beyond that scope, and those insights are important outcomes of this paper which deserve to be highlighted, but can be distracting from the original point of the paper as they are currently presented. Consider restructuring a bit more to move all of those details to the discussion (e.g., showing that the allan variance performance is adequate for field calibration or lab measurements is a result, but showing how the analysis changes with long data truncation and implications for managing memory when processing data seems like more of a discussion point.) I encourage the authors to carefully reread each section and consider how the tests support use of the vaporization module in lab and field measurements so that they can streamline their message. The science is sound and the technical conclusions drawn about vaporization for CRDS will move these technologies forward, but the paper would be more compelling and easier to read if the message is clearer and the purpose of each of the tests is highlighted more effectively.

We have tried to improve the manuscript by making it more to the reader how the small experiments illustrate the versatility and operation of the instrument. We note in our answers below to the reviewer's comments that we believe that the system presented in the manuscript represents an improvement in the Allan Deviation compared to previously presented systems.

There are still several typos and other sections where clearer language would benefit the reader, many of which I have highlighted line-by-line in the attached document.

Thank you. We have noted this.

Finally, the conversation about system memory requires clarification in several places. More information about how the authors are quantifying the memory of the system and whether or not its performance is adequate for the field-calibration setup or the discrete lab measurements is

needed to show that this system can be used successfully for both functions. The authors in a few places talk about “handling memory” but seem to be referring to post-processing techniques to remove data from very long runs for the purpose of stability analyses, but it is not clear how the memory is treated in day-to-day operations. Can you include a section with general workflow and recommendations for making low-memory measurements and processing the data? The authors also (correctly) point out that the changing flow rates in the system can change the memory effect, but there is no analysis of this impact or best practices for operating a system like this with so many variables. How do you ensure that the memory is adequate for your applications?

We have so far only been using the system in field deployments and for laboratory experiments and not for routine operation of liquid sample measurements.
This means that no stable day-to-day operation has been developed or needed. Instead the way that we are dealing with memory effect is to simply measure the same standard as long time as possible given the constraints of the field deployment or laboratory experiments. For example for laboratory experiments, where we are trying to get 10^-2 order of magnitude d18O precision measurements, we would run the same standard for at least 12 hours.

The authors have not directly compared the data with published values from the continuous-flow systems, but will find that some conclusions drawn from the data truncation exercise are misleading when this is done (figures are overlain in the attachment to demonstrate this). While the allan deviation plots use longer analysis times than previously published data, the discussion of memory and stability improvements isn't very convincing since it seems that the performance is very comparable to published continuous-flow values. The purpose of this analysis seems to be to show that the vaporizer is very stable at long timescales -- and while this is true, it is also demonstrating longer memory length than other systems, which achieve similar values at long timescales without the significant data truncation. The very long allan variance data are important contributions to the literature and the system does show excellent long-term stability – and performance that is on the better end of what other systems have documented – but it seems like the authors want to claim that the performance is “new” or “better” when it both isn’t and also doesn’t need to be to show that the custom vaporizer unit designed here has excellent performance and unique advantages for its designed application (i.e. humidity-variable calibration). However, the increased memory (especially for dD) should be examined and it needs to be made clear to the reader that this is not an issue for the laboratory measurements or the field calibration that are intended for this unit.

We have address this very important comment in details below.

Please see more specific comments in the attached PDF. Looking forward to reading your updates that will come from this versatile new system!

Technical corrections and minor comments:

ln 12: I suggest rephrasing this to discuss the measurement of many samples instead of their "connection"
We are unfortunately not sure what the reviewer refers to

ln 17: "measuring unknown samples shows" is specific to 17O, right?
Indeed - corrected.

ln 18: the standard error is not provided -- perhaps this is a typo?
In fact - it was our poor sentence structure. It should be clear now.

ln 19: "enhancement" is not quite right here. Maybe rephrase to talk about the increase in deviation or noise level instead?
Corrected

ln 21: typo, should say "factor of 2"
Corrected

ln 27: "achieved to operate" could just say "operated"
Corrected

ln 41: typo after citation, should say "is classically"
Corrected

ln 46-47: please clean up citation formatting and italics. why not use the typical notation for the deuterium excess?
We presume that this will be corrected in the formating stage of the manuscript.

ln 70-73: language is very confusing and inefficient -- please rephrase the first few sentences. A suggestion for ln72 is to say "...deployment is the bubbler system, which has been used continuously..." ln 74: "and that there is minimal" could just say "and the minimal" ln 78: maybe "is not feasible for many campaigns" would be clearer
Corrected

ln 94: double-check the 17O-excess notation here and throughout for consistency
Corrected

ln 97-99: please rewrite this sentence as it is very confusing to follow
Corrected

ln 109: maybe "accuracy, we have further developed the patent application which was published in Steen-Larsen (2016)."
Corrected

ln 115: could say "sufficiently high accuracy for D17O"

Corrected

ln 115-130: it might be more intuitive to understand the purpose of the two case studies if this whole section is rewritten as a brief paragraph about objectives and how you tested them
We agree that we could write these bullet points as a paragraph. However, we prefer to keep the bullet points, since we believe that it allow a potential future reader to quickly get an overview of the improvements of the system.

ln 135 is a good example of how to refer to the patent for the first time -- consider revising ln 109 to be similar
We agree - corrected

ln 170: typo, should refer to 4.3. please check section references throughout.
Yes - indeed - corrected and checked.

ln 170 and section 4.3 talk about decline in humidity values over long timescales, but figure 5 shows that it sometimes increases. we find that depending on the air flow and water flow, the precipitates either clog the tee itself or the capillary, and depending on which flow is decreasing (air or water) the instability can cause an increase or a decrease in humidity. is this similar? it seems consistent with the values shown in figure 5.
The decline in humidity, which we refer to that are caused by clogging of the capillary is on much longer time-scales that the individual steps shown in figure 5. Typically, we would experience noticible clogging, when we would measure the same humidity level for more than 12 hour periods. We have not had reasons to suspect clogging of the tee itself.

ln 194: typo, "between 0.5 to 3.5 should" say "between 0.5 and 3.5"
Corrected

ln 195: why not just say "with a 1.59mm PEEK union"?
Corrected

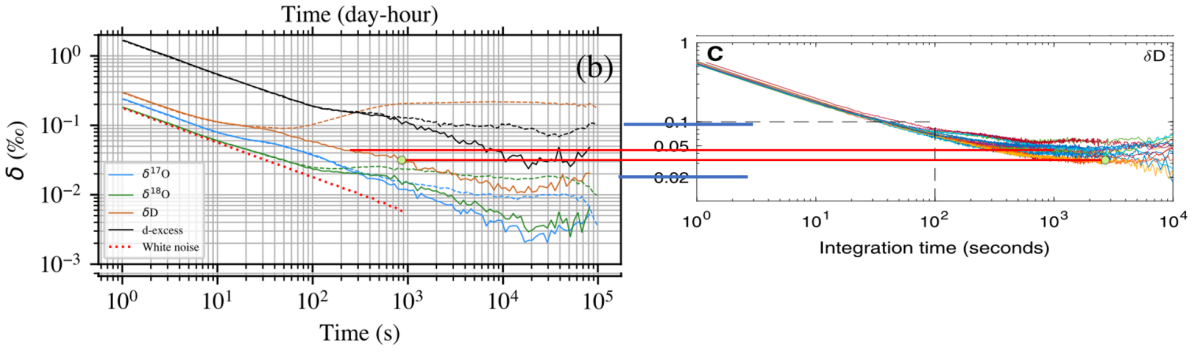
ln 220: "frequent" should be "frequently"
Corrected

ln 222: SW and WW haven't been defined yet -- please make sure these and other acronyms are defined before use
The acronyms are referring to the standard names in the Table. We will therefore make sure that there is a direct reference to the table. The names of the standards for examples SW and WW was the names, which we were informed about when receiving the standard water from our colleagues.

ln 265: please see comments above -- the performance shown here is excellent, but it is not significantly different from Steig et al 2021, which does not trim data as extensively as is shown

here or have nearly as much time for each run. this (and the comparison with fig 8) suggests an increase in memory for this system, which should be discussed in more detail, if only to demonstrate that it can be managed for the applications promised.

Thank you for agreeing that the performance presented is excellent.
While we agree that for integration times up to 10^3 there are only a smaller improvement in our system compared to Steig et al. 2021. The average performance for dD at 10^3 corresponds to our integration at 2e2. However, while the Allan Deviation in Steig et al. 2021 does not improve after 1e3, ours continue to improve until 2e4. See figures below focusing on dD.
The reviewer is however correct in pointing out that the memory effect seems signficantly larger than in the system presented by Steig et al.. Having worked with the similar system system (Jones et al. 2017) as discussed by Steig et al., I do also believe that the system by Jones could have a smaller memory effect than ours. However, it comes at a cost of a 10-20 time higher flow rate.
However, a direct comparison in memory effect between the system used by Steig et al. 2021 and our system is not possible due to lack of information of which standards were measured before the run used to generate the Allan Deviation. We speculate that the Allan Deviation is generated based on the lab water run depicted in Figure 1 and that the run depicted is typical. This means that a idealized step change is equal to about 25 per mil in d18O. This should be compared to the step change of about 55 per mil in d18O applied in our setup.
It is furthermore not clear to us, how much of the data is removed after the step change in Steig et al. in order to remove the effect of memory in the system. In our figure 3 (shown below) we depict the influence of memory effect, when not removing the first 16 hours as dashed lines. Not removing any data after a step change is of course not realistic, but we do this to illustrate the influence of the memory effect on the Allan Deviation.
In short, it would be very interesting to 1) run the system of Jones et al. 2017 for 92 hours continuously and 2) generate a similar plot as Figure 4 based on a step-change of 55 per mil in d18O.



section 3.2 compares the memory for this new system to the memory of the autosampler, which demonstrates that this method could be useful for lab measurements. But how is memory handled for field-calibration? What is a typical field calibration workflow?

Often, for water vapor isotope field measurements, the signal measured is much larger than the measurement uncertainties. This means that less focus is placed on achiving calibrations with precisions achieved only after 1e3 second integrations. However, when measuring water vapor isotopes at low humidity levels, one has to be careful about achiving robust humidity-isotope-calibration of the instrument. To achieve optimal humidity-isotope calibration we would often run the same standard over-night before starting the humidity-isotope-calibration of the instrument.

ln 318: should specify that the short-term trend is for the water vapor concentration

Corrected - the text now reads:
As expected, the largest Allan Deviation with averaging time of 600 seconds is the one measured at 584 ppmv level (0.04, 0.05 and 0.17 ‰ for $\delta^{17}O$, $\delta^{18}O$ and δD, respectively). Between 5000 and 20000 ppmv, the 600 s Allan Deviation is characterized by small variability in general (0.013 ± 0.002, 0.014 ± 0.004 and 0.02 ± 0.01 ‰ for $\delta^{17}O$, $\delta^{18}O$ and δD, respectively). For unknown reasons the worst performances in terms of Allan Deviation are observed at 11584 ppmv (0.02, 0.03 and 0.06 ‰ for $\delta^{17}O$, $\delta^{18}O$ and δD, respectively for 600 s averaging time). This is in contrast with the analysis above, which shows the smallest short-term trend of the H_2O signal in the ~12500 ppmv region, which is ~0 ppmv/h.

ln 337: it seems like the variability in dD values at low-humidity ranges in Fig. 7 could also be influenced by the relatively longer memory of dD in the system – has this been investigated? Even if the liquid flow rate is the same at all levels because of the secondary mixing at TEE2, the retention time of water within the optical cavity itself should also worsen with decreased humidity which seems like it could contribute to this effect?

We do not believe that the depicted increase in variability of dD at low-humidity levels is a result of memory effect as proposed by the reviewer. We see the increase in variability in dD at low humidity level to be a result of the lower humidity, which gives less signal-noise ratio in cavity. However, we do agree that the hypothesis proposed by the reviewer, that there would be an increase in dD memory effect at low humidity levels due to the fact that there is less molecules to exchange with the molecules adhering to the side of the wall (which could maybe be assumed to a first approximation be independent of the number of molecules in the air).
The hypothesis could be investigated, but it would be useful if such experiment could reveil more information about molecular bonds and adhesion than just a qualitative description.

ln 364: many readers will have difficulty understanding the statement of D17O-->d17O conversion
Corrected and simplified.

ln 377: consider "values" or something more specific than "one" for clarity
Corrected

ln 379: it is unclear what you mean by "error in the scale"

Indeed - changed to "due to measurement error of the weighing scale"

ln 384+: this section is describing the experimental setup to examine the high-frequency noise – consider restructuring to include relevant details in the methods section of the paper instead
We agree with the reviewer, that it is possible to move the description of the experimental setup to the method section. In addition, this could also be done for the other application studies. However, below we argue for why we prefer to keep the structure of section as it is now:

1) Our main focus of the paper is to describe how the vapor generation module functions and to document its performance. This section is focused on the origin of the noise generated by the analyzer and is therefore an illustration of what the vapor generation module can be used for. By including the method description of this experimental setup into the method section of the article we feel that the manuscript would loose a bit focus.
2) We want to ensure that it is easy for the reader to build the vapor generation module. It is therefore important for us to keep this focus.

ln 392: typo, "exceeding" should be "excess"
corrected

ln 410-11: typo? – this sentence doesn't make sense to me
Corrected

ln 480-5: this does seem like a general problem for automated adjustments, though certainly not beyond characterization. Have you attempted to characterize the range of this effect for your system?
We agree with the reviewer that this can be characterized. We, however, have not done this yet, as we initially did not plan to log the flow rate of the dilution MFC, which is a requirement for doing so. As we are in the process of building a dedicated line for liquid measurements the characterization of the role of flow rate on memory effect is a natural action item. It is worth to remember that this characterization will be depending on the individual configuration of the system such as tube length.

Ln 484/S3: which standards are used for calibration in each of the tests? It is not clear if it is the same every time and which of the standards listed in the supplement are used.
We have updated the text to make it clear that it is the two standards BER and SP, which is being used.

Ln 491: what is "relatively longer"?
This sentence is focused on explaining the consequences of the different memory effect on dD compared to d18O. The consequence is that one need to have a *relative longer* measurement time for dD compared to the measurement time for d18O to achieve the same percentage of the isotope shift.

Ln 494: is this memory limitation practical? Is it possible to shorten the tubing, or remove other dead volume from the system?

We have not carried out an exhausted search for memory effect removal, but we suspect that the analyzer itself is one of the main drivers of the memory effect. One could in fact try different tube lengths and investigate if the memory effect would be linear and then calculate the inherent memory effect of the system for a zero tube length.

ln 520: relative discrete autosampler injections
We are not sure what the reviewer is referring to here.

ln 536: does not say what the standard error is (and looks copied from earlier section of the paper)
Our text was formulated poorly. This is corrected now.

ln 537-9: citations would be helpful here and elsewhere – this should also be discussed earlier in the paper
We do not have a reference for this speculation as this is something that we have discussed with colleagues. The thinking being that since there are less heavy isotopes in more depleted samples the uncertainty should also be larger. We have changed the text to a hypothesis.

table 1: how is 17Oexcess determined for BER? d17O does not have enough significant digits to determine these values.
The BER standard has been provided with only two significant digits, hence we assumed d17O to be equal to -0.0500 and calculated the 17O Excess.
We are aware that d17O ranging from -0.046 to -0.054 might introduce a 8 per meg variation in 17O excess. However, such a variability in d17O affects only to a lesser extent the slope of the calibration line. Therefore, the effect on the analysis of the Allan variance and liquid measurement reproducibility is not significantly affected.

table 2: I'm not sure there's an advantage to showing both sets of calibrated values -- perhaps just choose one calibration method, explain it in the text, and report the values in the table?

We decided to report the results obtained by the two calibration schemes to demonstrate the good reproducibility of the measurement. Indeed, the independent calibration scheme should be more sensitive to changes in the measurement conditions, while the average calibration scheme should smooth-out the variability more effectively. This effect is visible on the magnitude of the uncertainty associated with the mean of M85, which is larger for the independent calibration. However, the similarity of the results obtained with the two methods provide general evidence that the CRDS drift, the instability of the calibration system and the sample degradation are minimal.