Dear Christof Janssen,

Thank you for your time and effort as an editor on our manuscript entitled "Simple water vapor sampling for stable isotope analysis using affordable valves and bags". Attached please find our revised manuscript to be considered for publication.
This work has been previously submitted to AMT, with "major/minor revisions" being suggested to apply to our manuscript. The main recommendations from you and the two reviewers were to better present the results, limitations, and restrictions of our approach throughout the manuscript, especially in the abstract and discussion.

We have substantially revised our manuscript according to these suggestions, taking into account all criticisms and comments and we believe the implemented revisions improved the manuscript significantly.

Sincerely,

Adrian Dahlmann

Dear Authors,

Thank you for submitting your manuscript to EGUsphere/AMT. As evident from the referee reports, your article merits publication in AMT. However, and this has been also indicated in one of the reports, some of the limitations of the proposed method are not clearly stated and the interpretation of the provided data is not fully convincing. Therefore, the current presentation is not yet fully in line with AMT standards and still requires improvement before it can be published. I thus decide that the paper requires a major revision (with review). In preparing the revised version, you should

1/ Consider the recommendations of both of the referees.

1/ All the comments made by the reviewers are now included in the revised manuscript.

2/ Improve the abstract to better reflect the findings (& their limits) of your article.

2/ We have now added more information to the abstract to clarify the results, limitations and restrictions of our method:

- Information about the use of new / reused bags were added to the result description:
  L20: "The storage experiment with new bags demonstrated the ability to store water vapor samples for up to seven days while maintaining acceptable results for δ^2H, and acceptable to questionable results for $\delta^{18}O$. The memory experiment using new bags revealed that the influence of previous samples increased with duration of storage."
- The recommended use of similar samples was added:
  L27: "The reuse experiment confirmed that the bags can be filled repeatedly, provided they are used for similar sample lines and rinsed ten times with dry air."
- We point out that storing was tested mainly for 24 hours:
  L30: "Storing beyond 24 hours needs further investigation but appears promising. With new gas bags up to 24 hours of storage, we found accuracy of 0.2 ± 0.9 ‰ for $\delta^{18}O$ and 0.7 ± 2.3 ‰ for δ^2H. When the bags were reused and stored up to 24 hours, they yielded accuracy of 0.1 ± 0.8 ‰ for $\delta^{18}O$ and 1.4 ± 3.3 ‰ for δ^2H."
- The limitations are now highlighted:
  L33: "The proposed system is simple, cost-efficient, and versatile for both lab and field applications, however, case-specific testing is necessary given the remaining uncertainties."

3/ Do the same for the discussion and conclusions. I am worried about two particular issues here:

- First, the L22 results in experiment I seem to be completely ignored and are not mentioned in the discussion. In this test, only 2 measurements out of all 15 measurement points are in the 'accurate range'. These two points belong to the 7-day storage time. One interpretation is that this indicates that there is no clear temporal trend —– and this seems to be the chosen route. However, another possible interpretation is that due to some unknown reason low δ-values cannot be measured reliably by the bag method, which would invalidate your conclusions. It is likely that the Day 1 and Day 3 measurements suffer from contamination (a straight line through M22 and the lowest Day 1 measurement shows that all the incriminated (z>=2) day 1 and day 3 measurements are aligned along an isotopic mixing line), whose origin in the absence of further details is difficult to explain. For this reason, it is imperative that there is a discussion/evaluation of these particular results in the discussion section and that the impact on the validity of the method is discussed.

First, we explained and discussed the results in more detail. For this purpose, we have divided the discussion into two subsections, "4.1 Comparison to previous developments to store and measure water vapor " and "4.2 Limitations, future perspectives and cost classification". In particular, we will highlight the experimental results and provide an explanation of our findings and the uncertainties that have arisen.

- Second, the discussion section starts by comparing the current results with previous studies and claims a comparable performance (L388 and following "Our results are generally comparable in accuracy to previous studies of water vapor storage. For example, the Soil Water Isotope Storage System (SWISS) introduced by Havranek et al. (2020) showed a high accuracy during a 30-day storage period in a laboratory experiment (±0.5 ‰ δ18O and ± 2.4 ‰ δ2H). This result was followed by several experiments, which showed an actual precision of 0.9 ‰ and 3.7‰ for δ18O and δ2H in field applications with a storage time of 14 days (Havranek et al., 2023)." This is misleading, because the current study has only looked at storage times of not more than 7 days. How can these be compared to maximum storage times of 14 and 30 days of the two other studies? Moreover, the majority of the Havranek 2020 samples (9 for 18O and 10 for D) out of all 12 samples (6 overnight samples, 3 24-day storage samples and 3 30-day storage samples) with a target at δD = -122 ‰ and δ18O = -16.4 ‰ fall in the z <= 2 range (as expected from basic statistics). This is certainly not consistent and comparable with your results on the somewhat similar isotopic target L22 (with δD = -148.1

‰ and δ18O = -19.9 ‰) presented in Fig. 3c, where only a minority of samples is within the z <= 2 range.

Second, we agree that our results are not directly comparable to other methods such as Havreneck et al. (2020). We have now removed the first sentence ("Our results are generally comparable in accuracy to previous studies of water vapor storage.") to avoid misunderstanding. The brief introduction of the SWISS system by Havreneck et al. (2020) is still part of the discussion, but we now state that the accuracy is higher with their system and we strongly believe that it is necessary and fair to compare/classify similar methods when all information about differences is given:

L393: *"In general, it is difficult to compare the few different approaches to water vapor sampling for isotopic analysis because they vary in complexity and application (e.g., storage time or price per sample). However, our results for reused bag samples stored up to 24 hours are generally comparable in accuracy to previous studies of water vapor storage. For example, the Soil Water Isotope Storage System (SWISS) introduced by Havranek et al. (2020) showed a higher precision during a 30-day storage period in a laboratory experiment (± 0.5 ‰ $\delta^{18}O$ and ± 2.4 ‰ $\delta^{2}H$). …"*

In addition, we have now added the information about our recommended storage time of up to 24 hours throughout the manuscript, and regarding Havreneck et al. (2020), we highlight the fact that glass methods may be superior for long-term storage.

4/ Explain and justify why memory tests with storage (experiment III) have been restricted to measurement of the H standard after passing the L standard. In both previous experimental tests (storage (experiment I) and memory (experiment II)), L and H standards have been treated interchangeably and it has been shown that for some unknown reason, measurement of L is more critical and prone to contamination than measurement of M or H. Including the reverse measurement should provide a more realistic qualitative assessment of your approach and the results could have an impact on the required number of purges when utilising used bags.

Thank you for your comment. We understand your concern about the light standard and have now explained and discussed the issue throughout the manuscript. Below is some information to answer your questions:

- Experiment I only tested storage with new bags for up to 7 days, while Experiment III showed the memory effect with two very different standards (overnight storage with

the initial standard here resulted in acceptable values). Therefore, the results of experiment I, with a higher uncertainty for $\delta^{18}O$ compared to predominantly acceptable values for $\delta^{2}H$, cannot be related to experiment III.

- The larger uncertainty of the light standard of experiment II was indeed insufficiently discussed. Unfortunately, the explanation was deleted during the first round of revisions as part of the separation of the discussion. We have now added information that the stronger memory effect was caused by different handling of the samples (approx. 45 minutes of storage). No other results indicate that the heavy standard is easier to handle than the light standard.
- Experiment III was only carried out in one direction, as we first wanted to quantify the memory effect and in particular to show how long it could affect the subsequent samples.

However, by revising the manuscript, we have now clarified that the method can only be used for the range of isotopic signatures we tested. Further investigations for other ranges and directions, such as the experiment you recommended with the combined memory/storage from heavy to light standard, should be carried out. Unfortunately, we are unable to do this at this time due to construction in our labs.

5/ Introduce and use a consistent terminology around measurement uncertainty. It appears that you are following actual recommendations from BIPM as stated in the VIM ("JCGM 200:2012, International vocabulary of metrology – Basic and general concepts and associated terms (VIM),
3rd edition", electronic link provided below). Sometimes in the manuscript, however, the accuracy is specified which is contrary to these recommendations and it is particularly confusing to see that the accuracy is once defined as a range (± 0.5 ‰, L391 for example), and once as a measurement result with an associated uncertainty range 0.25 ± 0.41‰ (e.g., L408). I strongly recommend to use consistent terminology and notation (in preference in line with the recommendations from BIPM and ISO) and stick to that throughout the manuscript. Following VIM, it would be more correct to talk about a trueness range instead of the accurate range for z <= 2.

We have carefully revised the manuscript according to BIPM/ISO terminology and notation from and have redefined the z-score definitions, e.g.:

- *L31: "With new gas bags up to 24 hours of storage, we found accuracy of 0.2 ± 0.9 ‰ for $\delta^{18}O$ and 0.7 ± 2.3 ‰ for $\delta^{2}H$."*

- *L290: "A z-score < 2 represents an acceptable range, a z-score between 2 and 5 describes the questionable range, and a z-score > 5 representing an unacceptable range (Wassenaar et al., 2012; Orlowski et al., 2016a)."*
- *L310: "Consequently, z-scores were either within the acceptable range or close to it, again with no trend of decreasing accuracy over storage time."*

Minor issues:

- L291: There doesn't seem to be a gradient here. Replace by 'range'.

Changed.

- L204-211: The experimental description could be improved by directly indicating the number of bags at the first instance and the number of repeats (3), ie '5 gas bags' in L206 and 'twice with the opposite standard' in L208. The last phrase can then be deleted.

Thank you for your comment. We now added „*utilizing five newly prepared bags per standard.*" in L215 and the number of repeats „*We repeated the process three times (fill, measure, empty) with the opposite standard…"* in L218.

- L234/235: add space in front of 'cm'

Changed.

- L321: remove 'again', as it is the first time that the memory effect disappears.

Changed.

- L503: The doi of Havranek 2020 leads to a paper of Lee, Lee and Yoo entitled "Analysis of ceramides in cosmetics by reversed-phase liquid chromatography/electrospray ionization mass spectrometry with collision-induced dissociation". There is likely just a digit missing in the doi number.

Indeed! There was a 3 missing at the end. Thank you!

- Fig S2: Why is the z-score scale for δ18O flipped as compared to the direct δ18O scale? The data points seem to have undergone a reflection at the x = 0 axis when going from the left to the right box —— which should not be the case.

Thank you for your comment. We have changed the x-axis.

References:

VIM https://www.bipm.org/documents/20126/2071204/JCGM_200_2012.pdf/f0e1ad45-d337-bbeb-53a6-15fe649d0ff1

Report #1

General comments:

I appreciate the effort the authors have put into revising and improving the manuscript. However, I still have some concerns regarding the interpretation of the presented findings. In general, it appears to me that the obtained data quality requires certain circumstances to be achievable. Specifically, rather short sample storage times and a strict "identical bags for identical probes" procedure were necessary to achieve the obtained data quality.

Further, the seven-day storage test was not performed on unknown samples in reused bags which would make the method more applicable for potential users. Also, you state that your method is somewhat safe for samples in the natural abundance range but the memory test (3.3) was performed on standards representing this range and required numerous refills in order to get accurate isotope readings. Therefore, I think these restrictions need to be stated in the abstract already, not only hidden deep inside the manuscript.

You showed that flushing reused bags ten times with dry air yielded bag data in agreement with in situ measurements after one day of storage. Notwithstanding the fact that this procedure (or the number of necessary flushes) was not tested prior to field application, I think this is the main argument supporting your story. All other data are either not representative or show the limitations of your method and the importance of sticking to some narrow requirements that make your method less universal or practical. I strongly suggest to make this (short storage time, strict "identical bags for identical probes" procedure) clearer throughout the manuscript.

Thank you for reviewing our manuscript. We have now tried to present the findings, limitations, and restrictions of our method more clearly in the abstract (see the editor's general comment 2/), but also to discuss them in more detail (see the editor's general comments for specific citations).

In particular, we clearly state throughout the manuscript that storage with the given experiments can only be recommended for up to 24 hours and ideally for similar samples.

Regarding experiment III, we showed the memory effect only after one day of storage of two very different standards (light to heavy) without flushing of the bags. Here we have tried to quantify the memory effect without trying to remove it. But you are right, the additional experiment S2 revealed that memory effects still can occur even with 10 flushes. We have added this restriction to the abstract/discussion and note that this method should be further tested for different experimental designs.

Specific comments:

L10: insert “of soil or plant water isotopes” after “measurements” as you are talking about matrix-bound water that requires extraction.

Changed.

L11: consider replacing “semi-permeable” by “vapor-permeable” or “hydrophobic”

Changed to „gas-permeable membranes“ for consistency.

L21: I find this statement somewhat misleading. “seven days” seems to suggest that unknown samples can be stored for such a long time. In reality, the performance was a function of isotope values with the standard close to ambient air performing best.

We understand that this could be confusing. We now changed it to:

L20: „*The storage experiment with new bags demonstrated the ability to store water vapor samples for up to seven days while maintaining mostly acceptable trueness for* δ^2H*, and acceptable to questionable trueness for* $\delta^{18}O$.“

In addition, we have added the storage time to the results at the end of the abstract to avoid misunderstandings:

L30:„*Storing beyond 24 hours needs further investigation but appears promising. With new gas bags up to 24 hours of storage, we found accuracy of 0.2 ± 0.9 ‰, respectively, for* $\delta^{18}O$ *and 0.7 ± 2.3 ‰ for* δ^2H*. When the bags were reused and stored up to 24 hours, they yielded accuracy of 0.1 ± 0.8 ‰ for* $\delta^{18}O$ *and 1.4 ± 3.3 ‰ for* δ^2H.“

L23: replace “samples” by “bags”

Changed.

L28: do you mean “replicate measurements”? “Repeated measurements” seems to suggest multiple analyses of the same bag(s).

Yes, we have changed it to “replicate measurements” now.

L31: “cost-effective” or rather “cost-efficient”?

Changed.

L49: To my knowledge, Marshall et al (2020) should not be on this list as they did not use membranes when drilling holes through tree stems.

Thank you for your comment. It is correct that Marshall et al. (2020) didn't used gas permeable membranes. We now cite the Paper "Xylem water in riparian willow trees (*Salix alba*) reveals shallow sources of root water uptake by in situ monitoring of stable water isotopes" by Landgraf et al. (2022). Here, the borehole equilibration method was used with gas permeable membranes.

L52: The (potential) difference in costs is mainly the power source as for subsequent, lab-based measurements a costly analyzer is needed as well.

It is true that an expensive analyzer is also needed, but not in the field. With our method, one lab analyzer could be used to measure numerous field sites. It now reads:

L53: "*However, direct, continuous in situ field setups are very cost-intensive, technically challenging and require a permanent power supply in the field as well as strong expertise to maintain. Moreover, direct in situ field setups require full-time operation of one laser spectrometer (e.g. a CRDS) each, whereas a vapor storage method could be operated with one CRDS for several field setups.*"

L53: "strong expertise" is also required (or at least favourable) for lab-based analyses. - Hail to the lab personnel! ;)

You are absolutely right!

L60: temperature stability is less important than consistently exceeding the sampling temperature during analysis

We now changed this sentence to clarify this statement:

L63: „*The advantages of these methods include the ability to quickly measure stored samples at elevated temperatures relative to the source in a temperature-stable laboratory environment.* "

L109: I think you mean "consistent" when you say "accurate". Also, in this paragraph can you explain how you get from raw analyzer readings to VSMOW-referenced vapor values before you calculate liquid water values using Majoube`s equation?

„Accurate" was changed to „consistent".

Regarding the calibration, we added a new section in the end:

L127: “In laboratory experiments, calibration was performed by measuring the described glass bottles before the start of the measurement and the used standard during and after the experiment for drift correction. In field experiments, the standards covering the expected sampled isotopic range were filled into bags and treated similarly to the samples. Calibration was then performed.”

L133: Please insert “nominal” before “capacity” as 90% is the actual, usable filling capacity.

Changed.

L148: Please insert “Flow at” before “the open split”.

Changed.

L219: It still puzzles me that you emptied the bags over night instead of testing the effect of this scenario which is probably representative for many occasion when analyses are not performed on the same day as sampling.

At this stage of our research, we only wanted to quantify the effect of the first standard stored in the bag. If we had left the samples filled over the break, we would not have been able to guarantee this.
The following experiment IV was conducted for potential effects that could occur overnight.

L245: How were “concurrently measured” raw vapour concentration readings of the Picarro calibrated?

The calibration is now explained in the end of section 2.1.

L406: Delete “known” as their method does not require knowledge of ambient vapour isotope values.

Thank you for your comment. It now reads:

L413: „*To circumvent these memory effects, they explored preconditioning of the bags with moist, isotopically homogeneous air sample where the goal was not to eliminate the memory effect, but to make it predictable and remove it.*“

L415: Did Herbstritt et al. (2023) report leakages or why is this considered an important difference to their approach?

No, they did not report any leaks, but we wanted to emphasize that we are building an easy-to-use connection. We have now rewritten this sentence to focus on the simplification of gas transfer.

L422: „*Second, we have modified the valve inlets to the bags in a way that simplified gas transfers and may reduce leakage.*“

L429: By saying “we know the previous sample signature”, do you imply that bags can ONLY be reused for identical samples/probes? What if I have a limited set of bags and want to use it on different locations before repeat sampling at the location visited first? If this does not work – not even with flushing dry air ten times - you should say so.

We now clearly state throughout the manuscript that sampling can only be recommended for the given storage time of up to 24 hours and is also limited by the isotopic signature of consecutive samples. In particular, we have added the final sentence in this section:

L439: “*Concluding, our results suggest comparable accuracy to other methods for 24 hours, but the accuracy of long-term storage and high isotopic differences for consecutive samples should be further tested.*”

L438: Evaporation does not cause scatter “around” the LMWL but rather below as evaporation lines have slopes significantly smaller than the one of the LMWL.

We now changed it to:

L447: „*This results in a wide range of isotopic signatures throughout the complete cultivation season, as can be seen in the smaller slope compared to the LMWL in the upper soil layer (Fig. 7).*“

L455: “relatively low” is misleading as only on study (Havranek et al., 2020) reported higher per-container costs. Or do you mean “relative to overall expenses of a typical field campaign”?

Yes, it now reads:
L504: „*The cost of the commercial gas bags we used was relatively low compared to the total cost of a typical field campaign.*“

L467: “cost-effective” of “cost-efficient”?

Changed.

L472f: As the method should ideally be applicable for unknown samples, it is quite a limitation that differences between consecutive samples should be small. Reference to the natural abundance range is misleading in this context as Experiment III (using standards within that range) showed that numerous refills were required before good data were recorded.

See general comments above.

Report #2

Thank you for your effort in the revision of the manuscript.
However, I still have a few minor comments.

Specific comments:

L. 49: Marshall et al. didn't use gas permeable membranes in their borehole method.

Thank you for your comment. It is correct that Marshall et al. (2020) didn't used gas permeable membranes. We now cite the Paper "Xylem water in riparian willow trees (*Salix alba*) reveals shallow sources of root water uptake by in situ monitoring of stable water isotopes" by Landgraf et al. 2022. Here, the borehole equilibration method was used with gas permeable membranes.

L. 412 ff.: The first three points are the differences, whereas the fourth point is in agreement with the paper of Herbstritt et al. (2023). I therefore recommend rephrasing the beginning of the sentence in L. 417, e.g., 'In agreement with Herbstritt et al., (2023), we have identified…"

We totally agree and changed the beginning of the sentence:

L425: „*Aside from the differences, we likewise identified a time-dependent memory effect, which is consistent with the notion that some diffusion/adsorption process occurs over many hours within the walls of the bag, setting an isotopic signal that requires multiple flushes to remove.*"

L. 425 and L. 470: "…ten-times flushing with dry air (Fig. S2)…" I can't find any evidence in Fig. S2, why dry air flushing is more recommendable over moist air flushing. Fig. S2 seems to be about new bags vs. reused bags after dry air flushing. I don't see any dry vs. moist flushing data. Did you do this comparison? If not, you can't proof the statement in L. 470 and should delete it in the Conclusion section.

Thank you for your comment. It is correct that the additional experiment did not test a comparison between dry and moist air flushing. Based on your comment, we have removed the statement in the conclusion.

Technical correction:

L. 109: …results up to 100…

Changed.

Figure 3: you could perhaps add ‘M22’ to 3b and ‘L22’ to 3c, according to the labeling of 3a.

Changed (also for Figure 4 for consistency).

Figure 5, caption: Please begin with the description of (a), then (b). Where is the ‘arrow’?

Changed caption and added arrow.

L. 403: replace the “;” between ‘et al.’ and ‘2023’ by a “,”

Changed.

L. 459: delete ‘our’ after “…commercially available…”

Changed.