Reply to reviewer 1 on manuscript amt-2024-43:

Review on manuscript amt-2024-43

Title: An easy-to-use water vapor sampling approach for stable isotope analysis using affordable membrane valve multi-foil bags

Author(s): Adrian Dahlmann et al.

General comments:

In the manuscript by Dahlmann et al. the authors present an alternative approach (in comparison to Herbstritt et al., 2023, Magh et al., 2022 and Havranek et al., 2021) of obtaining water vapor samples for analyzing the isotopic composition of soil water obtained with gas-permeable membranes (GPM) and storing them in multi-layer foil bags. They performed different experiments to test maximum storage time, potential memory effects and reusability as well as field applicability. The authors conclude that their approach is a simple, cost-effective, and versatile approach.

The paper is nicely written and well structured.

However, I have two main concerns:

As approach and concept in this manuscript are very similar to the paper by Herbstritt et al. (2023) - only a different type of GPM and a different (commercially available) bag type were used - the results should be compared point by point and discussed accordingly. Here some revisions and additional considerations are needed.

Moreover, the cited study by Jiménez-Rodríguez is only available as preprint in HESSD. It was under review in 2019 but not accepted due to substantial issues. No revision was provided by the authors afterwards, thus no accepted peer-reviewed version exists. This pre-print can therefore not be cited in your manuscript.

Thank you very much for the positive feedback on our work. We now revise the manuscript in detail to answer all your questions and those of the second reviewer. As part of this revision, we removed all statements/references to the preprint by Jiménez-Rodríguez. In addition, as recommended by reviewer 2, we separated the discussion/results and include a detailed comparison with Herbstritt et al. (2023) and other comparable methods, as well as a detailed recommendation about 1) the use of our method, 2) possible errors and how to avoid them, and 3) what should be tested for future use (in detail in the supplements "Handling Recommendations").

Specific Comments:

1. L28: what you call 'precision' is the deviation from the true value, isn't it?

   Yes, we changed it to "accuracy" for a better understanding and changed the wording according to the ISO 5725 definition throughout the manuscript.

2. L30: (and throughout the manuscript) 'water stable isotopes' not 'stable water isotopes'

Done.

3. L34: ‘suitable for many applications’ was already mentioned in L. 31. Please rephrase

   This sentence was deleted in the course of the revision.

4. L61: ’which can cost anywhere from’ sounds a bit sloppy

   It now reads: “To do so, primarily glass bottles or gas sampling bags with various fittings are used, which cost from ~1-200 euros per container.”

5. L67: ‘…provide new insights in research’ please be a bit more specific or delete.

   Done. It now reads: “These simplified and more affordable systems could therefore increase the number of studies on stable water isotopes and provide new insights in research by increasing the number of possible experimental sites and samples.”

6. L98: wrong! ‰ is not “million”

   Changed.

7. L112: P/N of bags? Fitting of bags? Volume of bags?

   We could not find a part number but we now added a link to the bags and the product name on the website (Multi Foil Bags with Stainless Steel Fitting, https://www.smelltest.eu/en/product/multi-foil-bags-with-stainless-steel-fitting/). All the information we received is now included here or in Table S1 in the supplement.

8. L115: use the manufacturer as reference for material properties such as ‘Water Vapor Transmission Rate’

   See comment above.

9. L125: Fig. 1: Where is the bag? On the right side in Fig. 1b?

   Yes, the bag is on the right side of Fig. 1b. In Fig. 1a, the connector is disconnected from the bag. We now explain this in the description of the Figure.

10. L130: how much standard water was added to the 100 mL bottle? And what was the size of the GPM inside the bottle? This is crucial for the isotopic equilibrium

    Thanks for your comment and question. The standard measurement is now explained (see section 2.1 Study area and basics of water stable isotope measurements).

11. L275: Jiménez-Rodríguez can’t be cited (see general comments)

    All statements/references to Jiménez-Rodríguez's paper have been removed.

12. L289: Fig. 3: would be nice if you could add “L22” and “M22” directly to the figure, there is enough white space. The Legend is also a bit small.

Changed.

13. L305: if you assume an error during the measurement, I'm wondering why do you present these data and didn't repeat the measurement?

    We did not repeat the measurement mainly because the results after 7 days looked promising and supported our reasoning against a potential storage effect. Following recommendations of reviewer 2, we will split the results and discussion and include all recommendations to avoid problems in the supplement ("Handling Recommendations").

14. L325 and L332: Jiménez-Rodríguez can't be cited (see general comments)

    All statements/references to Jiménez-Rodríguez's paper have been removed from the revised manuscript.

15. L324 to L342: please rephrase and take into account the findings of Herbstritt et al. (2023)

    We now discuss our results and how they differ from Herbstritt et al. (2023).

16. L354: Fig. 4: see comments to Fig. 3

    Changed.

17. L378: "...has not been described in the literature before" please rephrase or delete. I would say it is comparable to Herbstritt et al. (2023), Fig. B1.

    Changed.

18. L378-400: Please discuss and compare the results of Herbstritt et al. (2023) on memory effects and conditioning here.

    We now discuss our results and how they differ from Herbstritt et al. (2023).

19. L504 ff: Please check the reference list thoroughly, e.g. Millar et al., (2018) or Orlowski et al., (2016b) are in the reference list but not cited in the text.

    Thank you very much. All references have been checked.

Technical correction:

1. L11: 'water stable' not 'water-stable'

   Done.

2. L16: ‘easy-to-perform’ instead of ‘easy to perform’

   Done.

3. L30: I suggest “…using GPM combined with…” instead of “…using GPM and…”

   Done.

4. L34: ‘suitable’ instead of ‘suited’

   Done.

5. L83: please use different bullet points for the field experiments e.g. i) ii) iii) or (a) (b) (c)

   Done.

6. L98: wrong wording: detailed in

   Done.

7. L108: Majoube is from 1971

   Done.

8. L215: ‘procedure similar to’ instead of ‘similar procedure to’

   Done.

9. L247: please delete line break

   Done.

10. L284: I suggest ‘On average’ instead of ‘In average’

    Done.

11. L396: ‘especially in’ instead of ‘in especially’

    Done.

12. L447: please add dates

    Done.

Reply to reviewer 2 on manuscript amt-2024-43:

R2: General comments

The authors present a new method of collecting discrete vapor samples for water vapor stable isotope analysis using inflatable multi-foil bags. The presented method contributes to a new, currently evolving field of stable isotope analysis still lacking an agreed-upon standard procedure suitable and attractive for many users interested in performing in situ isotope assays without field-access to an analyzer. Therefore, any reported experience in this regard is highly welcome and I recommend publication after proper revision.

The manuscript describes the use of bags, which differ only in valves (which do not seem to have an effect) from the ones used in a previous study (Herbstritt et al., 2023, doi: 10.5194/hess-27-3701-2023). I therefore suggest a more thorough discussion emphasizing how this work expands the findings of the previous study. Moreover, I don't understand how the proposed treatment of previously used bags would help to get meaningful results if reused for unknown samples. I have a feeling that the tested treatment to remove memory effects does not account for the potential conditions faced by researchers interested in using the proposed method regarding, e.g., feasible or necessary storage time and range of previously observed isotope values.

Formally, the authors decided to combine results and discussion. Unfortunately, this often leads to a limited description of the results. I believe the manuscript would benefit from a better distinction between description and interpretation of the presented findings. Also, some additional technical details (flow measurement devices, new or reused bags for the field test, rinsing atmosphere prior to reuse, etc.) should be added to the method section. Finally, a detailed SOP listing suggested settings and potential pitfalls may be helpful for future users of the proposed method.

Thank you very much for your detailed and constructive comments and for recommending publication. We have thoroughly revised the manuscript to explain in detail all questions raised in the comments or to clarify misunderstandings. In particular, we have highlighted the differences between our work compared to previous studies more clearly:

In situ measurements of water stable isotopes are usually performed with two different systems: the recently commercially available WIP system (as used in the study by Herbstritt et al., 2023) and originally developed by Volkmann et al., (2014; for soils) and (2016, for xylem of trees), and home-built systems with GPMs (following the original developments of Rothfuss et al., 2013 and as used in Kübert et al., 2020 or Kühnhammer et al., 2021). The main difference between these systems is that the WIP system dilutes the sample flow by reducing the water vapor concentration in the probe, hence enabling measurements with relatively constant water vapor concentrations. Home-built systems with GPMs usually measure the saturated airflow without dilution in the GPM. One of the main differences is that the water vapor concentration of a sample from the WIP system is usually lower than that of the self-built systems due to the dilution. This has the advantage of reducing the risk of condensation, but also leads to a lower water concentration and thus a reduction in sample volume. We believe that aside from the material used for the storage container (different types of bags, glass vials etc.), the in situ method itself is also an important part that can influence the method development of a new storage method and find it relevant to compare our approach using a self-constructed in situ system with the WIP system used in Herbstritt et al. 2023. We *have clarified more clearly on this and all other important differences between our work and that of Herbstritt et al, (2023) during the revision of the manuscript.*

Moreover, we agree that rinsing 10 times with dry air is not completely transferable, but our recommendation was based more on our results from the field experiment in February, where we followed exactly this principle (rinse used bags from October 10 times). However, we see that this was not fully explained in the first version of the manuscript and that the difference in the isotopic signal of the samples was also not nearly as strong as for the two standards (see old section 3.4 Field test), which limits the recommendation. For this reason, we have 1) changed the explanation of our field experiments to avoid misunderstandings and 2) performed an additional experiment following our field protocol where we will store one standard in new bags for one day, rinse the bags with dry air, and then fill them with the opposite standard. We then measured these samples one (and 3) days later and additionally discussed and presented these results in the updated manuscript (Discussion/Supplement). This gives more insight into the reusability of our method in different experimental settings (e.g. natural abundance vs. labelling approaches).

As recommended by the reviewer, we have split the Results and Discussion sections to make the results more understandable and to discuss them in more detail. In the Discussion, we include all the information requested by the reviewers and compare our results with those of other studies (especially Herbstitt et al., 2023, but also Magh et al., 2022 and Havraneck et al., 2020). Throughout the whole manuscript, we have added information to make the differences more explicit (e.g. home-made GPM vs. WIP system or glass bottles vs. bags or lab vs. field experiments or different storage times) and to point out important advantages and disadvantages in a more comparative way. In addition to these comparisons, we have added a section on how to use our system and how to avoid potential problems ("Handling Recommendations").

I provide a list of specific comments below.

Specific comments

L10: "water stable", not "water-stable"

Changed.

L16: "easy-to-perform, in situ", not "easy to perform, in-situ" (also elsewhere in the MS: "in situ" without hyphen)

Changed.

L22: insert "spectrometer" or equivalent after "laser"

Changed.

L25: "can lead" seems too weak, as there will always be influence of previous samples. I suggest "does lead" or "will lead"

This sentence has been deleted.

L26: Consider rephrasing to: "…showed that the memory effect increases with duration of storage."

Changed.

L28: You state the precision, which describes the scattering of repeated or replicate measurements. What is the accuracy, i.e. the deviation from the target value?

We changed the wording according to the ISO 5725 definition throughout the manuscript.

L30 (and elsewhere): "Water stable", not "stable water"

Changed.

L38f: I do not see why hydrology and meteorology would focus on the biosphere. Consider rephrasing.

Changed.

L59 (and elsewhere): Do not cite preprints like Jimenez-Rodriguez et al. (2019). It is against AMT guidelines and it devalues the work of reviewers. Especially, do not call such work "successful" (L70) when in fact it has not successfully passed a peer-review process.

All statements/references to Jiménez-Rodríguez's paper will be removed from the revised manuscript.

L61: "less than 50 euros" is quite vague. Can you be more specific?

It now reads: "To do so, primarily glass bottles or gas sampling bags with various fittings are used, which cost from ~1-200 euros per container."

L63: in what aspect is the lab environment stable? – Temperature?

Yes. It now reads: "temperature-stable laboratory environment".

L63: What do you mean by "configuration"?

This can be misleading due to our wording. By "time-consuming configuration" we meant the time-consuming post-procession step (calculation) to obtain the true isotopic value. Using the glass bottle method requires the filling of missing sample volume with dry air to avoid suction inside the glass bottle. Naturally, this involves an additional correction step, due to the vapor concentration significantly declining over the measurement period (see section 2.1 VSVS lab test in Magh et al., 2022).

In our opinion and after testing the glass bottle method ourselves in collaboration with part of the co-author team of the Magh et al., paper, we concluded, that a storage method based on bags rather than glass vials is easier to handle. With the gas bag method, isotopic signatures of the gas inside the bag are directly recorded with a stable water vapor concentration.

Of course, this does not necessarily imply that one method is superior to the other, hence we rewrote this section.

L72: Again, be more specific about pricing. This helps other researchers considering using your method. In addition, the numbers never appeared in the manuscript again, i.e. they were not discussed. Nonetheless, you refer to them prominently in the manuscript's title. How do they compare to the per-sample costs of the containers used by Magh et al., (2022, doi: 10.5194/hess-26-3573-2022) and Herbstritt et al. (2023)?

We have added the approximate actual cost per bag in parentheses (costs have increased since we purchased our materials).

L95f: Please be more specific. How was vapor from standards produced? Was it in equilibrium with the liquid phase (resulting in temperature-dependent isotope fractionation) or flash-evaporated (with no fractionation)?

Thanks for your comment and question. The standard measurement is now explained (see section 2.1 Study area and basics of water stable isotope measurements).

L98: per mil, not parts per million (I wonder how this went unnoticed by five co-authors…)

Changed.

L108: The Majoube paper is from 1971, not 1961.

Changed.

L112f: Would it be possible to state a part number for these bags as well? I am unable to find this product in a web query. In addition, how does a membrane-based valve work? Does the sample have to pass through a membrane?

We could not find a part number but we now added a link to the bags and the product name on the website (Multi Foil Bags with Stainless Steel Fitting, https://www.smelltest.eu/en/product/multi-foil-bags-with-stainless-steel-fitting/). All the information we received is now included here (section 2.2.1) or in Table S1 in the supplement.

It now reads: "The stainless steel 2-in-1 fitting combined the valve and septum, with the septum acting as a seal, allowing air to flow around it when the valve was open, and sealing when the valve was closed."

L115: This number seems to be huge! Assuming that the sample bags (front and back) have an area of roughly one tenth of a square meter, more than half of a sample (which comprises ~17µL or 17mg of water per 1 L air at room temperature when saturated) would be exchanged per day. Can this be true? Please, also state the conditions (temperature, relative humidity, vapor pressure gradient), under which the water vapor transmission rate was determined (without citing a preprint). Otherwise, this number is meaningless. Or disturbingly high.

Thank you for your comment. In fact, this number is not correct. We calculated the WVTR again and the correct value is 0.00465 gr/m2/24h. Here is the manufacturer's information and the calculation in metric system:

Water vapor transmission rate (FED 101): < 0.0003 gr / 100 in² / 24 hrs

- *100 in² = 0,064516 m² → 1550 in² = 1 m²*

- *< 0.0003 gr / 100 in² / 24 hrs * 15.5 = < 0.00465 gr / m² / 24h*

With a bag area of ~ 640 cm² it would be:

- *< 0.00465 gr / m² / 24h * 0.064 = 0.0002976 gr / bag area / 24h or*
- *< 0.2976 mg / 24h for a bag.*

With 15.3 mg of water sample in 0.9 L of air at room air temperature at saturation, this would be ~ 2% per day or ~ 14 % per week, but (as you already mentioned in your second question about the conditions) this is an extreme value tested with the "FED-STD-101 – Test Procedure for Packaging Materials" at high water concentrations (90%) on one side and low water concentrations (desiccant) on the other side at ~ 38°C (http://www.woodencrates.org/standards/FED-STD-101.pdf).

All statements have been corrected.

L121: Did you test a version without electrical isolation tape that did not work? I am wondering if the tape really makes a difference regarding proper sealing.

It is true that the electrical tape per se is not important for proper sealing. Initially, we tested the bags without tape, but the adhesive in combination with the PTFE tubing can break under tension, which (of course) leads to leakage. Therefore, we used the electrical tape to stabilize the connector (you could probably use any tape, but we had the electrical tape in abundant stock). We have explained this in the revised manuscript.

L127: What was the length of the GPM?

The length of the GPM is not as important here (< 5 cm), as the dry air passes the standard water, and it is more of a safety mechanism to prevent liquid water from entering the tube/analyzer. In the field experiment, we used approx. 12 cm GPM (comparable to soil GPM in e.g. Kühnhammer et al., 2021).

The measurements are now explained (see section 2.1 Study area and basics of water stable isotope measurements and 2.4 Experimental design).

L133: How was the flow rate measured? And what would have been the maximum possible flow ensuring equilibrium given the GPM length you selected?

Thanks for your comment and question. All information is now included in section 2.1.

It now reads: "A gas cylinder was used to induce dry gas at a low flow rate of 50 - 80 ml per minute (257-6409, RS Components GmbH, Germany). We ensured that the isotopic signature of the vapor would be at equilibrium with liquid water at this flow rate. We tested flows from the minimum required for Picarro operation (approximately 35 ml/min) to 300 ml/min and found accurate results to 100 ml/min."

L135: Under non-EQ conditions, the vapor isotopic composition would also depend on water isotopic composition and surrounding temperature. But not exclusively.

All information is now included in section 2.1.

L140: By “outgoing”, do you mean the flow going out of the sample vessel or the flow going out of the open outlet?

We are talking about the “open split”. We changed it for a better understanding:

“Since the laser spectrometer only has a flow rate of approx. 35 to 40 ml per minute, an open split was added to ensure a constant flow and to avoid pressure differences. The open split was continuously measured to ensure that no ambient air could flow back.”

L163: How dry was the air after passage through the desiccant? Was this value tested and constant over the course of the experiment?

It now reads: „The dry air supply box was tested prior to our experiments by measuring the outlet concentration of the dry box over the course of one day. However, the use of such a system should always be tested for the specific application, as a very high flow rate combined with very humid air could greatly affect the duration of possible use. During the experiments, we periodically tested the water concentration before and after the field campaigns and could not detect any increase after one day in the field. The water concentration of the dry air produced was approximately 200 ppm.”

L166f: What would happen, if the bags were filled to more than 90% capacity? And why isn’t a lower filling capacity stated in the first place? How about filling only to the minimum volume necessary to reach a plateau on the analyzer during analysis? Did you play with that variable as well? How would that impact feasible sample throughput? How would the reduced sample volume affect its vulnerability, e.g., regarding memory effects?

Thank you for these interesting questions. We now mention it in the beginning of the method section and give detailed information about bag handling in the supplement („Handling Recommendations“).

As recommended by the manufacturer, care was taken when filling the bags to ensure that the maximum volume did not exceed 90% of capacity, which could cause material damage.

Personally, we do not recommend a lower filling quantity, as this could change the volume to area ratio and increase the effect of the water vapor transmission rate. In turn, this could potentially increase storage and memory effects.

A reduced sample volume could potentially have a positive effect on sample throughput in the field, as the filling time would be significantly reduced. However, a higher sample throughput could also be achieved by simply using multiple dry air pumps, i.e. filling the bags simultaneously in the field, without having to reduce the sample volume.

L173ff: This statement is a repetition of L141f. Consider deleting.

Changed.

L180: 100 mL bottle volume minus 60 mL of water leaves 40 mL headspace volume which is exchanged in < 1 min(?). Is this sufficient for establishing equilibrium given the applied flow rates? Were the tubes submerged?

See comment on L95f / L133.

L193: Was this the observed temperature range during sampling? Then 25°C (L197) may not be enough to prevent condensation.

This was the temperature in the laboratory during storage. During the measurements, great care was taken to ensure that the temperature in the lab was higher than the temperature we measured during filling.

Detailed temperature information is now explained in section 2.1 for all experiments.

L218: Why did you test only the effect of one-day storage when you intended to store natural samples for up to seven days? Did you refill them with L22 before you "then proceeded" (L219) with H22? Why? Did you also assess the memory effect on samples stored in re-used bags for seven days after the previous samples had also been stored for that period? From your experience, what kind of preparation would be necessary in that case to still obtain meaningful isotope measurements from unknown samples stored in re-used bags?

For our applications, the one-day period is the most interesting because we usually spend a day in the field taking measurements and then have time to analyze the next day.

We now explain the experiments in more detail and provide information in the discussion about the memory effect after a longer storage period (3 days) and larger isotope differences (additional experiment). In addition, all the information we obtained from working with the bags is contained in the manuscript and in detail in the "Handling recommendations" (supplement).

L220: What do you mean by "usual steps"? Did you refill with H22 and measure/empty immediately? How are the obtained findings transferable to a setting where, e.g., L22 was the first sample collected with a new bag and H22 was the sample collected with the (now reused) bag – also stored for 1 day, or 3 days, or 7 days? I am afraid, this is the weak point of the entire reusability test. By emptying the bags overnight (L223), you avoided exactly the effects that samples in reused bags may be subjected to. The point of reusing bags for unknown samples collected remotely should be to NOT have to refill/empty them repeatedly with the sample of interest and then measure them immediately. Can you propose a preparation routine for to-be-reused bags that ensures the isotopic composition of any unknown sample to be reproducible with sufficient accuracy after typical storage times? If not, I am afraid, the combined storage and memory test is not very exhaustive. (Later, you suggest rinsing 10 times with dry air but you do not present data proving the usefulness of that procedure.)

Thank you for your comment. It is true that we first had L22 in a new bag for one day, and then H22 was filled, measured, and emptied directly. We agree that rinsing 10 times with dry air is not completely transferable, but our recommendation was based more on our results from the field experiment in February, where we followed exactly this procedure. However, we see that this was not fully explained in the first version of the manuscript and the difference in the isotopic signal of the samples is not as strong as for the two standards (see experiment III). For this reason, we have separated our field experiment from the field data and explained it in more detail (2.4.3). In addition, we performed an experiment following our field protocol, in which we stored one standard in new bags for one day, rinsed the bags with dry air, and then filled them with the opposite standard. We then measured these samples one (and 3) days later (see discussion for more details).

L229: Please state here already, if you used new or reused bags for this part of the study.

We now explain that we used new bags in October and reused bags in February.

L234f: This sentence sounds odd. Either insert "samples" after "45 cm" or delete "for" and change "taken" to "sampled"

Changed.

L239: Equilibrium is not indicated by stable values. Steady-state conditions are indicated by stable values. One way to test for equilibrium conditions is to vary the flow rate around the target value and see if this has an effect on readings of vapor mixing ratio and isotope signatures. Was this done?

As this method of sampling soil water isotopes is well established and has been used by us and several other studies, we did not carry out an additional test. However, when the system was set up, we found no effect on the signal with different airflow rates below 100 ml per minute.

L241: What was the time per in situ measurement (as compared to 15 min of bag filling)?

During this part of the experiment, we did at least 15 minutes of in situ measurements.

L242: The logger only records the readings from an attached sensor. What sensor was connected to the logger to obtain temperature measurements?

We added the sensor information.

L243: Please, also state here the durations of the individual steps. Most importantly, how long were samples stored in the reused(?) bags prior to measurements? How does this compare to the combined storage and memory test? And how is this transferable to a setting with no field-access to an analyzer? (I understood that bag measurements were conducted in the field shortly after filling.)

These questions arose from our misleading description of the experiments and should now be clarified.

L245: This statement is a repetition of L231f. Consider deleting.

Changed (Rewritten at top).

L282 (and elsewhere): For consistency, delete quotation marks for the names of the standards (here: L22 and M22).

Changed.

L290ff: This seems to be a repetition of the previous statement. Rephrase or delete

Changed.

L302f: "increased deviation" translates to high inaccuracy, not "imprecision". Accuracy describes the deviation from the target value and is not synonymous with precision, which describes the scatter of repeated or replicate measurements around a common mean.

We have changed this paragraph in response to your comment. It now reads:

"The second storage test using L22, showed a lower accuracy due to lower precision for δ²H, being – 0.1 ± 1.1 ‰ for δ¹⁸O and 2.8 ± 4.9 ‰. However, no time trend was observed. The decreased accuracy was mostly caused by the samples after three days, as all gas bags showed a significant enrichment (8.9 ± 2 ‰ on average). The higher inaccuracy after three days of storage must be due to an error during the measurement, as accuracy improved again after 7 days. The z scores show accurate values for δ²H (except after 3 days) and more questionable values for δ¹⁸O. The average z-score was 0.3 ± 2.7 for δ¹⁸O and 1.4 ± 2.5 for δ²H."

L303: insert "samples from" after "as".

Changed.

L305: please elaborate on the "error during measurement". What went wrong and how can users of your method avoid this error?

We did not repeat the measurement mainly because the results after 7 days looked promising and supported our reasoning against a potential storage effect. We have separated the results from the discussion and included all recommendations to avoid problems in the discussion and in detail in the supplement ("Handling Recommendations").

L312: I don't think it is fair to compare the accuracy of two methods that used totally different storage times (1-7 days vs. 30 days).

We now discuss this statement in a more balanced way.

L321: Given that Magh et al. (2022) used off-the-shelf components, I tend to say that their method is not more difficult to handle than yours. Further, the "static properties of the glass vials" (L322f) make overfilling impossible during sampling (as compared to a mandatory maximum of 90% in the case of the gas sampling bags used in this study) and allow for simply letting dry air flow in during measurement with no need of extra pumping. Apart from potential breaking, glass vials may also be more robust relative to the thin plastic and aluminum layers of sampling bags in many typical field settings (you report damaged bags yourself (L407)).

It is true that there are both advantages and disadvantages in handling, preparation and analysis compared to the system proposed by Magh et al. (2022), which we now discuss in a more balanced way. See also comments and replies above.

L329ff: Personally, I find it alarming when the standard closest to ambient air delivers the best results as it points to an unaccounted-for influence of ambient air. The question must be how you can ensure that your method delivers meaningful results regardless of the isotopic composition of standards or samples. And how does this impact the measurement of unknown field samples when collected using newly prepared, equally pre-treated bags?

This is of course true, but as we already wrote in L335-337: "The overall higher scatter (particularly for δ¹⁸O), which has a different isotopic signature than the ambient air, led to initial concern over potential exchange with ambient air. However, we do not think that is *likely as the visible scatter already appeared within one day of storage, was not directed towards isotopic signatures of ambient air and did not increase over time."*

L337: No. Flushing with dry air in the case of Herbstritt et al. (2023) did not cause the scattering. Rather, it was unsuitable to remove the scattering caused by previously collected, diverse samples as efficiently as flushing with moist air did.

This statement has been deleted with the separation of results and discussion.

L353: The connection between storage time and memory effect has already been shown in the Herbstritt study.

This statement has been deleted with the separation of results and discussion.

L356f: Insert "target" or equivalent before "standard deviation" (2x).

Changed.

L363: I don't know which part of the Herbstritt study you are referring to but as I understand they used ambient, non-saturated air of arbitrary isotopic composition to pre-condition their bags.

This statement has been deleted with the separation of results and discussion.

L377f: Clearly, the magnitude is a function of the isotopic spread between the standards used here. The exponential decrease – expressed in the standard deviation of an entire batch of to-be-reused bags – was also shown before (Herbstritt et al., 2022, Fig. 5b).

It now reads: „However, when the water source was changed to H, there was a clear memory effect of a magnitude up to -4.9 ± 1 ‰ δ18O in and -37 ± 6.4 in ‰ δ2H (Fig. 5 and Tab. 2).“

There is no citation to Herbstritt et al. (2023) because the sentence is now in the results. However, we refer to the memory effect they found in the Discussion.

L379f (and elsewhere): I think it is not necessary to repeat the isotopic composition of the standards so often. Ideally, the outcome of your method should be independent of these values anyway.

Changed.

L382: Why did you stop at H7? It would also be important to confirm that the readings stay in that range.

The measurements during this experiment took a long time, which meant that we were only able to carry out 7 repetitions within two days. As H5 and H6 were already close to the accurate range, we decided not to carry out any further measurements.

L397ff: You advise to reuse bags but you did not show how the isotopic signature of unknown samples can be obtained in the foreseen application, i.e. remote sampling followed by lab-based analysis on a different day. In the storage and memory test you repeatedly flushed the reused bags with standard vapor until the readings were acceptable (after irrelevantly short "storage" times). The proposed procedure (filling and emptying at least seven times (L400) and promptly measuring) is certainly not desirable (or feasible) when collecting unknown samples in remote locations. What would be the achievable sampling frequency in that case? And would that still be an advantage compared to direct in situ measurements performed with an analyzer that has been brought to the field?

Thank you for your comment. We understand that with the explanations and results presented in the recent form, an unrestricted recommendation for reuse cannot be made. By splitting the October/February measurements with the additional explanation that rinsed and reused bags were used in February, we can currently only recommend this method for measurements in a narrow natural abundance range (and following strict guidelines). We performed an additional experiment (see comment above) to be able to make a statement about samples with larger differences in isotopic signature and discussed the results.

But to answer your questions for possible future experiments: Filling the sample bags ten times with the target sample in the field and then emptying them would make the system more complex, as one pump would be needed for filling and one for emptying. However, if a system were built for each sample bag that automatically fills (~15 minutes) and empties (~1 minute) the bags and collects the samples at the same time (you would need as many pumping systems as you have samples), such sampling could be done in about 3 hours with a theoretically unlimited number of samples.

L400: With what and for how long should re-used bags be filled? I am sure this has in impact on feasible sample storage time. Can you also comment on a quantitative relationship between the ranges in isotopic compositions of previous samples and the necessary number of pre-sampling filling cycles?

The bags were rinsed with dry air. These questions should now be clarified. We now explain it in the method section (2.4.4) and the "Handling Recommendations" (supplement).

L404: Did you compare in situ measurements and bag measurement only during two or during all 18 campaigns? If two, then how were conditions different, especially regarding elapsed time between sampling and measurement and relative to the sample storage time tested in the combined experiment? Please specify in the method section.

We have now rewritten this section with a separation of experiments and results/discussion for better understanding.

- *Yes, only two of the 18 campaigns compared in situ and bag measurements.*
- *In the first campaign, we first measured in situ and then the bags immediately after filling (resulting in a direct/bag measurement in ~30 minutes).*
- *In the second campaign, we measured in situ and filled the bags. The bags were then measured in the lab within 24 hours after filling.*

L407: To make life easier for potential users of your method, please specify "filling errors". In addition, how did you identify condensation? Where did you see it?

We have now added "Handling Recommendations" to the supplement for further details. Regarding condensation, we once measured a bag at a temperature that was too low (the AC flow was directed toward the bag), resulting in a small condensation peak during the bag measurement. Since we could not be sure that there was no effect on the rest of the sample, we *discarded this bag. Condensation during bags filling should be avoided by flushing the soil probes in the field with dry air prior to the measurement.*

L409: This is important and should appear in the method section already: What did you use for rinsing the bags and where was this step performed? Standard-derived vapor in the lab or the to-be-collected, unknown sample vapor in the field? If the latter, what was the required per-sample time required for this step? 10 x 15 min = 150 min?

This statement has been moved to the method section (2.4.4) for better understanding. Also, the experiments are now better explained and discussed.

It now reads: „To exclude any memory effects, as we saw in experiment III, the reused bags were rinsed 10 times with dry air (approx. 10 x 10 min).“

L432: On what kind of analyzers do co-extracted organic compounds interfere with water stable isotope measurements?

Laser based cavity ring down spectrometer like the CRDS we used (Picarro 2310-i).

It now reads: "The accuracy of CVE can vary greatly for soil samples and is associated with co-extraction of organic compounds, significantly interfering with the isotopic quantification using CRDS (Orlowski et al., 2016b)."

L444: After what?

Deleted.

L446: Please specify "wide"

It now reads: "Measurements of soil water isotope profiles over the full season (Fig. 7) revealed a wide range of isotopic signatures with 2.1 ‰ to -15.2 ‰ for $\delta^{18}O$ and 12.9 ‰ to -98.5 ‰ for $\delta^{2}H$."

L447: The period needs to be specified.

It now reads: "The isotopic signature of precipitation is represented by the local meteoric water line (LMWL), shown here for the period of September 2021 to September 2023."

L455: For additional plausibility, can you compare the nature of the scatter, e.g., by comparing the linearity (R^2) of the dataset, with that of precipitation data and other datasets of soil water isotopes? Is there a difference in linearity between the two campaigns with field-access to the analyzer and the other 16 without (if that was the difference)? How were standards produced and treated in these two different cases? How many validation standards were co-measured and what was their precision and accuracy?

We have changed the graphic to better show the different campaigns and have added a more detailed comparison/explanation of the different depth and seasonal development (Figure S1). Three laboratory standards were bagged and treated in the same manner as the samples.

L458: transpiration rather does not cause enrichment. Evaporation does. Please change "evapotranspiration" to "evaporative"

Changed.

L462f: Where do I find the seasonal variability you are referring to?

Will be changed. It now reads: "Overall, our findings from the field trial suggest a good agreement with GPM probe and bag-based soil water isotope measurements with the LMWL and are plausible in terms of seasonal variability (see Fig. 6c; e.g. compare offsets between cryogenically extracted bulk soil water isotope measurements and LMWL; e.g. Zhao and Wang, 2021).“

L465f: This seems to be a bit off. Usually, the lower boundary of the plow layer is around 20 cm, not 45 cm. Was it different in your case? Can you also comment on the large range of isotope values observed for 150 cm depth (yellow symbols in Fig. 6)? I would expect to see a less pronounced variation at that depth.

It‘s correct that the lower boundary of the plow layer is typically located around 20 cm but it depends on the soil conditions during plowing (high soil water contents can lead deeper plowing). We actually expected the lower plow boundary to be 20 cm and consequently the deeper probes to be unaffected by tillage. Hence, the probes at 45 cm and 150 cm were not recovered and reinstalled before and after tillage. In comparison, we routinely remove/reinstall the soil probes in the upper layers (5cm and 15cm) during/after tillage. After discovering the very low vapor concentrations in the probes in 45 cm depths, we suspected damage to the probes due to the tillage. Personal communications with our field manager revealed, that the tillage was indeed deeper than 20 cm and likely resulted in a compaction of the soil down to the 45 cm probes. We have repeatedly tried to measure these probes and could measure some of them in a vapor concentration matching the vapor saturation at the given temperature. Those measurements were deemed likely to be valid and were included in the manuscript.

L468: Why does soil compaction flaw the measurements? In situ measurements have been conducted successfully in boreholes of (I would say: rather compact) trees by one of the co-authors. So why wouldn't they work in compacted soil? And why would that be an issue at 45 cm but not at 150 cm depth?

See comment above. (The compacted soil is not the problem per se, only the fact that the probes in 45 cm were installed before tillage i.e. they were in the soil when the compaction occurred, which is the typical handling of sensors in many agricultural studies, e.g. only de-install sensors that are above the manipulation depth)

L475: I think, "appropriate" is inappropriate here. You did not test the effect on samples stored in reused bags for more than 1 hour. (Or you forgot to mention that.) Consequently, I do not see how reliable measurements of unknown samples stored for typical time periods in reused bags can be performed based on the findings of this study.

This statement should be clear with the rewritten method section, the separation of results / discussion and considering the field experiment and the additional experiment.

L476: rinsing with dry air does not match the procedure described in the combined memory and storage experiment. Please explain (before the conclusion), why rinsing with dry air – previously suspected to increase scatter – does (or should do) the same trick that flushing with moist air does.

This statement should be clear with the rewritten method section, the separation of results / discussion and considering the field experiment and the additional experiment.

L485: are these numbers based on two or on 18 campaigns?

It now reads: "Through the conducted field experiment (two campaigns with CRDS and bag measurements), we were able to show that the bags could be used in our case with an accuracy of 0.23 ± 0.84 δ18O [‰] and 0.94 ± 2.69 δ2H [‰], which allows a wide applicability."

L490: Not "can" but "will"

Changed.

S1: AMT is a European Journal. I suggest using the metric system and SI units.

Changed to SI units.

S2 & S3: What depths are you referring to? Weren't these measurements performed on standard vapor sampled in the lab?

"Depth" will be deleted. It now reads: "Differences during the storage experiment for M22 and L22 for each storage duration…"