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 will thoroughly revise the manuscript to explain in detail all questions raised in the comments or to clarify misunderstandings. In particular, we will highlight 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 will elaborate 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 is not fully explained in the current version of the manuscript and that the difference in the isotopic signal of the samples is also not nearly as strong as for the two standards (see 3.4 Field test), which limits the recommendation. For this reason, we will 1) explain our field experiments in more detail to avoid misunderstanding and 2) perform an additional experiment following our field protocol in which 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 will then measure these samples one (and 3) days later and additionally present and discuss these results in the updated manuscript. (Unfortunately, we were not able to perform and present this test immediately for this reply, as we do not have all the necessary materials in stock). This will give more insights into the reusability of our method under different experimental settings (e.g. natural abundance vs. labelling approaches).

As recommended by the reviewer, we will split the Results and Discussion sections to make the results more understandable and to discuss them in more detail. In the discussion, we will include all 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). We will be more explicit about the differences (e.g. home-made GPM vs. WIP system or glass bottles vs. bags or lab vs. field experiments or different storage times) and point out important advantages and disadvantages in a more comparative way. In addition to these comparisons, we will add a section on how to use our system (SOP) and how to avoid potential problems.

I provide a list of specific comments below.

Specific comments

L10: “water stable”, not “water-stable”

Will be changed.

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

Will be changed.

L22: insert “spectrometer” or equivalent after “laser”

Will be changed.

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

Will be changed: “will lead”

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

Will be 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?

Thank you very much. We will add additional text on the accuracy so this wrong wording will be changed.

L30 (and elsewhere): “Water stable”, not “stable water”

Will be changed.

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

Will be 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?

We will change it to: “… , which can cost from ~1.2 euros to one to two hundred euros per container.”

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

Yes. We will change it to “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 clarify and present method comparisons in a more balanced way.

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 will include detailed prices (as an overview table) in the revised version and compare them with Magh et al. (2022) and Herbstritt et al. (2023) in more detail within a new part of the discussion.

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 water vapor was generated using a 100 ml glass bottle filled with approx. ~ 60 - 80 ml of standard water. Two semi-permeable membranes (GPM) were placed inside the bottle: 1) one for dry air supply, submerged in the standard water, and 2) one in the headspace for sampling of water vapor sampling and transport to the analyzer. Both GPMs were sealed with adhesive. We then continuously passed dry air at a low flow rate (equivalent to flow rates used in common in situ literature) through the water and through the GPM so that the collected vapor was in temperature-dependent water vapor equilibrium with the liquid phase (like e.g. Rothfuss et al. 2013 or Kühnhammer et al., 2021). The measured water vapor concentration was then compared to the saturated water vapor concentration at the given temperature (and pressure) to ensure saturation.

We will explain this in more detail with corresponding references in the revised manuscript.

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

Will be changed.

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

Will be 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 to find it on the website but we will include all information we have about the bags in a updated table in the supplement where we can additionally add 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/).

These multi-foil bags are equipped with a patented 2-in-1 stainless steel fitting. This fitting combines the valve and the septum in one. Simply put, the septum acts as a seal around which air flows out of the sample bag when the valve is open and seals the opening of the sample bag when the valve is 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).

We will explain/discuss this in more detail in the updated manuscript.

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 will explain 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).

For further details, see comment on L95f.

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

The flow was measured with a RS PRO air flow sensor (257-6409, RS Components GmbH, Germany). Here, we are talking about flow rates during our laboratory experiments with nearly unlimited water supply within the standard botte. We tested the standard bottles used starting with the minimum flow the picarro needs to operate (around 35 ml / min) and increased the flow up to 300 ml / min. Until around 100 ml / min (75 ml / min + picarro flow), it resulted in accurate results. With 100 ml / min + picarro flow the water concentration started to decrease slightly (with still acceptable results). A higher flow rate of 150, 200 and 300 ml/min + picarro flow then resulted a depletion of heavy ^{18}O and ^{2}H isotopes relative to the standard.

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

This is true, we will explain/discuss this in more detail in the updated manuscript. At equilibrium, the estimation of liquid isotopic composition is particularly straightforward, but we will also mention conditions under non-equilibrium conditions.

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?

Prior to our experiments, we measured the outlet concentration of the dry box over the course of one day. During the experiments, we regularly 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 about 200 ppm. 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.

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 will discuss it in more detail in the discussion section, but to answer them:

Overfilling can lead to damage to the bags and probably to a much higher stress on the material. At the beginning of our tests, for example, we found that the bags showed folds/creases after being overfilled, which were then repeatedly creased in the same way, leading to material damage. This is indeed very important especially when it comes to reusing the bags, so we now mention this in the updated manuscript at the beginning of chapter 2.2.

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.

Will be 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 and 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.

We will add this information and also discuss the implications for a wider use of the method.

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.

Yes, we measured L22 after one day of storage and then refilled and measured again to make sure there was no effect on the same standard after one day of storage.

We did not perform a test with standards where we tested the memory effect after very long storage times (< 7d), as these long storage times were beyond the scope of the current experiment (but this could of course be explored in the future). However, we used the same bags for the field measurements in October and February (the field campaigns where we compared bag to in situ measurements) and were able to obtain good results after rinsing with dry air 10 times. In the natural abundance range, we therefore assume that this treatment works reliably for sampling.

We will discuss this in more detail in the revised manuscript.

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 is not fully explained in the current version of the manuscript and the difference in the isotopic signal of the samples is not as strong as for the two standards (see section 3.3 combined storage and memory test). For this reason, we will 1) explain our field experiments in more detail and 2) perform an additional experiment following our field protocol in which 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 will then measure these samples one (and 3) days later and additionally present and discuss these results in the updated manuscript. (Unfortunately, we were not able to perform and present this test immediately, as we do not have all the necessary materials in stock). This will give more insights into the reusability of our method under different experimental settings (e.g. natural abundance vs. labelling approaches).

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

We will explain that we used new bags in October and reused bags in February. We will also adapt the graphics for a better understanding.

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

Will be 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?

We tested the standard bottles used starting with the minimum flow the picarro needs to operate (around 35 ml / min) and increased the flow up to 300 ml / min. Until around 100 ml / min (75 ml / min + picarro flow), it resulted in accurate results. With 100 ml / min + picarro flow the water concentration started to decrease slightly (with still acceptable results). A higher flow rate of 150, 200 and 300 ml/min + picarro flow then resulted a depletion of heavy ^{18}O and ^{2}H isotopes relative to the standard.

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?

The sensor information will be added.

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.)

In October, we first measured in-situ, then filled and measured the bags in the field, and remeasured them 1 day later in the lab. In February, in situ measurements were made in the field before filling and bags were measured in the laboratory the next day. We see that this part is not well explained here. We will rewrite this section (as well as the results for this experiment).

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

Will be changed.

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

Will be changed.

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

Will be 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 (which was - 19.9 ‰ $\delta^{18}O$ and - 148.1 ‰ $\delta^{2}H$) being – 0.1 ± 1.1 ‰ for $\delta^{18}O$ and 2.8 ± 4.9 ‰ for $\delta^{2}H$. No trend could be observed, similar to the previous experiment. The lower accuracy was mostly caused by the increased inaccuracy after three days, as all gas bags showed a significant enrichment (8.9 ± 2 ‰ on average). The z scores show the same result with accurate values for $\delta^{2}H$ (except after 3 days) and a lower precision with questionable values for $\delta^{18}O$. The average z-score was 0.3 ± 2.7 for $\delta^{18}O$ and 1.4 ± 2.5 for $\delta^{2}H$ (see Table 3 for detailed values)."

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

Will be changed.

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

We will discuss the error in the revised manuscript and recommend ways to avoid errors.

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).

This is correct and we will balance the comparison in the revised version.

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 will discuss in 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 $\delta^{18}O$) visible in the experiment using standard L22, 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 will be adapted in the revised manuscript 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 will be adapted in the revised manuscript with the separation of results and discussion.

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

Will be 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.

That's correct. It now reads: "In their study, the bags were additionally pre-flushed with ambient air of a known isotopic signature."

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).

We will include/clarify this in more detail in the revised 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.

Will be 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 this 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, see above). We will also perform an additional experiment (see comment above) to be able to make a statement about samples with larger differences in isotopic signature.

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.

We will include detailed suggested sampling protocols in the revised version.

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. This statement will be adapted in the revised manuscript with the separation of results and discussion.

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.

This statement will be clarified in the revised manuscript with the separation of results and discussion. In addition, we will add a more detailed explanation of the experiment in the methods section for a 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) as well as one day later in the lab.*
- *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 will add a section to the discussion that explains/discusses filling errors and how to handle them. 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?

We used dry air to rinse the bags. We will explain our handling in more detail in the Methods section and later in the Results/discussion section.

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). We will clarify this statement.

L444: After what?

Will be deleted.

L446: Please specify "wide"

We will add the "wide range" in a bracket.

L447: The period needs to be specified.

Will be changed. 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 will change the graphic to better show the different campaigns and add a more detailed comparison/explanation of the different depth and seasonal development. 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"

Will be 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.

We will clarify this statement in the revised manuscript and add a more detailed discussion on the implications of soil manipulation for long-term use of the in situ systems.

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 in itself only the fact that probes in 45 cm were installed before tillage i.e. 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 will be adapted in the revised manuscript with 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 will be adapted in the revised manuscript with the separation of results and discussion. See also comments and replies above for specifics.

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

These numbers are based on the two campaigns of in situ and bag measurements. We will adjust this statement in the revised manuscript with the separation of results and discussion.

L490: Not “can” but “will”

Will be changed.

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

Will be 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…”