Response to Reviewer 1, re: "Potential effects of cryogenic extraction biases on inferences drawn from xylem water deuterium isotope ratios: case studies using stable isotopes to infer plant water sources", in review in HESSD.

Comments from the review are in normal font, and our responses are in **bold font**.

We thank Reviewer 1 for their time and effort in providing us with feedback. These comments can help strengthen the paper. It appears that several key aspects of our paper and our overall objectives did not come through clearly enough. The reviewer apparently wants us to provide exact numbers for correction values by species, and while this is a worthwhile objective – one that our paper also emphasizes, and one that may take the community some years to achieve – it is beyond the scope of the present manuscript.

It may help if we more clearly describe our reasons for writing this paper. We observed that many in the community were concerned that Chen et al.'s recent paper would *invalidate* the use of xylem water isotopes in many applications. Conversely, there were others who reacted by suggesting that these biases are already known to exist, but they're small and can be *ignored*. And still others suggested that the community needs to quantify the biases in every species of interest, and every range of conditions, before anything can be said on the subject. These perspectives are not new, but Chen et al.'s recent paper has highlighted the clash between them.

We thought it would be useful to aim for a middle ground that is both productive and statistically robust: identifying where cryogenic extraction biases could potentially be important (e.g., where differences among sources are similar in scale to the previously reported extraction artifacts). The clearest way to do this was by using previous studies as illustrations of a range of scenarios. Each of these illustrative studies shows some results that would be invalidated by cryogenic extraction biases as large as those reported by Chen et al., and other results that would be robust against such biases. Because it will be a long time before extraction artifacts are quantified in a wide range of different species and conditions, we provide a conceptual and quantitative framework for assessing the possible influence of these biases (with their currently uncertain magnitudes) on inferences derived from xylem water isotopes.

We are not trying to provide a definitive answer at this point -- nor do we claim to. We are trying to provide a framework for the community to think about these issues as we all learn more about these extraction artifacts. The scope of this study is intentionally limited to xylem water isotopic biases in order to avoid obscuring our message with discussion of other different issues (e.g., analytical errors or soil water extraction artifacts, which have been discussed more generally elsewhere).

The manuscript underlines existing issues with cryogenic vacuum extraction of plant water for isotope analysis. The authors demonstrate via five case studies how δ^2H biases resulting from this extraction method may lead to misleading conclusions when interpreting plant water sources.

The authors heavily rely on results of a study by Chen et al. (2020) without going much into detail about this study. This is often tricky for the reader (since details are not mentioned) and it requires going back and reading the study by Chen et al.

We will add more details about the by-species differences. Otherwise, our discussion of Chen et al. is matched to our use of Chen et al. We simply adopt their numbers, e.g., "we use -8.1 ‰ *for purposes of illustration*" (line 91) but we do not argue that these represent universal correction values, as frequently stated in our manuscript (including the abstract and conclusion). It is unclear what unmentioned details of Chen et al., are wanted, as our assumptions and uses of these values are clearly stated, and our language consistently reflects out treatment of these bias values as hypothetical or assumed. We will also now add discussion of the extraction artifact values described in other studies, of other species.

The authors apply a δ^2H bias correction to a handful of plant source water studies. But what are the selection criteria of these case studies? The manuscript does not go far beyond the discussion section of the Chen et al. (2020) study. The application of an average bias correction factor determined by Chen et al. to a limited selection of other studies is not enough. Are the tree species from the selected case studies even covered in the study by Chen et al (2020)? Otherwise, I highly question the overall applicability of an average bias correction factor to different tree species (from your case studies) that may react differently during cryogenic vacuum extraction. This would result in the necessity to calculate different bias correction factor for specific plant species. Chen et al. (2020) stress the fact that species-specific differences in the cryogenically obtained δ^2H values exist and need to be accounted for (together with data on stem relative water content).

Nowhere do we imply that an average bias correction factor is universally applicable. Instead, as clearly stated in our paper, we use the average of Chen et al.'s values as an example *for purposes of illustration*. The suggestion that we match species biases with studies that compare the same species elsewhere ignores the likelihood that these biases might also be a product of plant age, wood anatomy (e.g., ring vs diffuse porous), type of wood sample (e.g., twig or main-stem), extraction line set up, user, etc. Let's not forget that Chen et al. only used 3 samples per species, and did the extractions in one lab, and thus it is unlikely that their values are universally representative of even the species that they studied, let alone other species.

Researchers will always be limited by their inability to know the true error, bias, or uncertainty associated with any individual measurement. That reality underpins our motivation to write this paper – to show that one can make useful inferences even when the exact biases underlying any sample values are unknown. Some applications may warrant greater precision and thus call for developing lab-specific, species-specific, or user-specific correction factors. The unavailability of such detailed bias characterizations, however, should not preclude considering the potential effects of these biases and how they might be accounted for. We are not dictating what those assumed numbers should be in any general sense.

Also, please note that we repeatedly emphasize that more studies are needed to use any correction factors deterministically: see lines 21-22(abstract), 73-76 (including eq. 6), 89-92, 156-157, 245-250, 256-259. We also account for a range of by-species bias values throughout the paper: 187, 213-215, the entirety of Section 4, and Figure 2. Moreover, please note our word choice when we apply the assumed 8.1 ‰ correction ("a possible cryogenic extraction bias of -8.1 ‰" (99), "a presumed -8.1 ‰ bias" (112), "presumed cryogenic extraction biases" (118), "presumed cryogenic extraction bias of -8.1 ‰," (125), "if it were shifted by +8.1 ‰ to compensate for presumed cryogenic extraction biases" (135), "If these xylem water values were shifted by +8.1 ‰ to correct for presumed cryogenic extraction biases," (153), "correcting for the presumed extraction bias would not imply greater" (155). This is only a partial list, but it should be clear that our use of conditional, hypothetical language was not accidental.

Regarding the selection of studies, each was selected because it showcases a different scenario. Each also demonstrates findings that would be minimally affected by biases of ~8 ‰ as well as findings that would be greatly affected by biases of that magnitude. We clearly state that we selected these studies for their illustrative value: "The studies were selected because their analyses and inferences exhibit a range of sensitivity to potential biases, not because they are representative of the total body of literature". The suggestion of including a labelling study is appreciated; this is a great suggestion because, with labelling studies, the signal can be made strong enough to minimize the potential effects of extraction biases. We would gladly receive suggestions of other specific scenarios to include, and consider their additional value to the study. Adding more for the sake of having more does not sound useful for our concept-driven article, because, as stated, a holistic synthesis is not the scope of our work. Perhaps such a re-analysis should be pursued once more studies report extraction biases and thus we have a better grasp on their generality and variability.

Overall, the manuscript is a bit light. It would benefit from 1) case studies that include the tested species of the Chen et al. study 2) selecting more than five case studies to underline the statements and conclusions of the present study and 3) from including recommendations on how to perform bias correction on your own data and which factors to consider for that (species-specific differences, extraction system bias correction, isotope measurement accuracy etc.). The authors further ignore the well-known fact that cryogenic vacuum extraction from certain soil types causes isotope effects that need to be corrected for (such effects may also differ for the selected case studies). I admit that this is difficult when the raw data of the selected case studies is not available but this could be overcome by choosing newer studies where authors might be willing to share their data or where data is already publically available.

1) We do not believe this to be possible, and it is not clear how this would help advance the objectives of our research.

2) It would be helpful to know what points would be strengthened by doubling the number of studies (10 studies is suggested in a later comment). Each of our five studies support our conclusion that "conclusions drawn from cryogenically extracted xylem water δ2H are neither generally valid nor generally invalid; what matters is the size of the potential

extraction artifacts in relation to the isotopic signals in the data." Our goal is not to systematically evaluate the entirety of past papers on this topic (and we are clear about this). We also received suggestions (not through this review process) to use fewer studies because cases A, B, and E together are adequately illustrative. We do plan to include a labeling study in the revised paper.

3) The entirety of the paper demonstrates how one would assess the potential impact of xylem water extraction bias on their own data: e.g., assuming a value that is reasonable based on the literature, and carrying it through the analysis to evaluate how the findings could be influenced. A more detailed, specific correction would be ideal, but as we say in the paper, "Further work to quantify potential extraction artifacts will hopefully yield correction factors for a range of species and extraction protocols. These are urgently needed because evidence does not support using 8.1 ‰ δ2H as a universal correction factor, although it was adopted here for illustrative purposes."

4) As stated in the paper, "We do not address the potential consequences of cryogenic extraction biases in soilwater samples, because the current literature lacks a sufficiently clear consensus on the magnitude, or even the direction, of such potential biases in soilwater extractions". Soil extraction errors are important, and they've been extensively discussed elsewhere, but they have not been found to be as consistently directional as the plant extraction biases have been. The inconsistent biases observed in soils would require a different approach than the one that we outline for the more consistent biases observed in xylem water extractions.

5) As the reviewer suggests, in principle one could find studies where the raw data are available, and *if* the workflows are sufficiently well documented, the calculations could be reproduced in detail, with and without assumed extraction biases. The cost of this approach, however, would be that readers would need to wade through several pages of methods for each study that was analyzed in this way.

Further, it would be interesting to know if a bias correction is necessary and how a potential bias correction would look like for artificial isotope labeling studies. How would results be shifted?

Agreed, although it's not clear what type of labelling study is being suggested. The answer probably depends on the mechanism controlling the extraction biases. Given that there are active debates on that topic (lines 29-32), and weighing in on that debate is unnecessary for our paper, we will not speculate. We can, however, show how a +8 ‰ δ^2H bias correction may affect results; for example, it would have only minor effects on the interpretation of water uptake in Kulmatiski et al., 2010 (https://doi.org/10.1111/j.1469-8137.2010.03338.x), given the huge (10^2-10^4 ‰ δ^2H) signal introduced by the deuterated water.

My specific comments can be found in the attached pdf. (copied and pasted below).

Page: 1

Number: 1; Date: 2/8/2021 7:29:20 AM

You need to explain that this is an average over only nine tree species. Do the species of your selected case studies belong to either of the nine species? If not, do you not think that a species-specific bias correction would be more appropriate? Chen et al. found species-specific differences but no differences between habitats.

We will now state that this is an average over only nine tree species when we first discuss Chen et al. Otherwise, our response to this comment has been made above.

Page: 2

Number: 1; Date: 2/8/2021 7:29:20 AM
What about isotope labeling that artificially alters the input signal?

This will be discussed in our revision. Thank you for the suggestion.

Number: 2; Date: 2/8/2021 7:29:20 AM
I think the reader needs more information about the study by Chen et al.

We will provide earlier discussion of by-species extraction errors to mitigate potential reader confusion.

Number: 3; Date: 2/8/2021 7:29:20 AM
Do you mean delta here?

Yes, thank you.

Number: 4; Date: 2/8/2021 7:29:20 AM
same here: delta

Yes, thank you.

Number: 5; Date: 2/8/2021 7:29:20 AM
Again, a bit more information would be useful.

We will clarify that Kirchner and Allen previously reported equations for error propagation in mixing models.

Page: 3

Number: 1; Date: 2/4/2021 1:40:11 AM
We know about biases resulting from cryogenic vacuum extraction of soil water. Can you account for those too? Not only the 2H signal of plant xylem water would have to be shifted but also the soil water signal (depending on the extracted soil type and its physico-chemical properties). How would this affect your interpretations of the five case studies?

We *could* account for soil water extraction biases (see eqs. 1-6), but focusing on those would distract from the primary focus of our paper. We believe that it is more effective to focus

on one issue, instead of simultaneously analyzing multiple issues and consequently making their respective effects less clear. For the same reason we also do not intend to address the IRMS or CRDS analytical errors that have been extensively discussed elsewhere (although eqs. 1-6 are also relevant when accounting for those errors).

Number: 5; Date: 2/4/2021 1:43:48 AM
And with different extraction parameters (time, temperature, pressure). See results of interlab comparison for cryogenic soil water extraction (https://doi.org/10.5194/hess-22-3619-2018)
Do we need to introduce a lab-specific "extraction system bias"?

We will specify those details as examples of lab-specific or user-specific factors, although it would be a mis-citation to cite that study in this sentence as currently written. Answering this question about the "need" for a lab-specific bias is at odds with the primary purpose of our paper. We describe that the need to reduce uncertainty, e.g., by developing lab-specific corrections, will depend on the signal strength in the environment and the required inferential precision.

Number: 6 Date: 2/4/2021 1:43:48 AM
Could you include a slightly wider range of studies here (say ~10)? This would strengthen the manuscript, underline your statements and point to the urgent need of bias correction.

This is addressed above.

Number: 7; Date: 2/4/2021 1:43:54 AM
Again, I think you need to give the reader a bit more information and not expect them to read the study by Chen et al. first.

This is addressed above.

Number: 8; Date: 2/4/2021 1:43:48 AM
Did you ask the authors to share their data?

We frame our use of the data transparently. We describe the data and its exact sources transparently. Why should our study depend on data that can only be gained by having colleagues who are willing to share their data? We did not need to leverage any such connections to procure the data required to write this paper.

Number: 9; Date: 2/4/2021 1:43:48 AM
Can you please explain in more detail why you only focus on 2H? Simply because there are not so many studies reporting issues with 18O, does not mean they do not exist. I understand that biases in 18O reported by Chen et al. are rather small but the authors only considered a limited number of species.

Thank you for the suggestion. In a revised version, we will explain that much of the content of this paper could also apply to 18O if similar biases were suspected. In general (among the few studies available), extraction biases in 18O appear to be smaller, but so are

the environmental signals that need to be quantified. On balance inferences drawn from 18O appear to be somewhat more robust against extraction biases than those drawn from 2H.

Page: 4
Number: 1; Date: 2/4/2021 1:43:48 AM
Gaj et al. 2016 (DOI: 10.1002/rcm.7787) suggest a post-correction of data based on the physicochemical soil properties.

We are aware, but the interpretation of those values is unclear, in part, because they depend on oven-drying and rewetting experiments (https://doi.org/10.1002/rcm.8348).

Number: 2; Date: 2/4/2021 1:41:44 AM
Does this account for instrumental measurement inaccuracy as well?
Which factors need to be considered when performing bias-correction on your own data?

The framework we describe can be used to evaluate any potential biases; any such biases should be assessed relative to the strength of the signal that one is trying to detect. We can clarify this further in the revision.

Number: 3; Date: 2/4/2021 1:59:58 AM
I know the average of 8.1 is based on tree water extractions but wouldn't that mean that we also have to correct other plant extracts' water? Can you further elaborate on that and be more clear about how your study goes beyond the Results and Discussion section of the Chen et al. paper?!

Chen et al. report the bias values. We present a framework to consider how and where such values matter, and then provide relevant equations, examples, and novel discussion of their importance in dual-isotope studies.

Number: 4; Date: 2/3/2021 3:42:42 AM
Is there a reason why two out of your five studies are kind of old? I do understand their relevance in the literature context but I think replacing those with "newer" studies (or adding newer studies) from which you could get the raw data would be beneficial to the manuscript.

We cannot agree or disagree with this because it is unclear why "newness" is beneficial. One benefit of using older studies is that many readers will already be familiar with them.

Number: 5; Date: 2/3/2021 2:56:08 AM
Early on, Meißner et al. (doi:10.1007/s11104-013-1970-z) showed that soil property effects lead to conflicting results for delta18O and delta2H in plant water uptake studies and Orlowski et al. 2013 (doi:10.5194/jsss-2-179-2013) showed with a simple greenhouse experiment that cryogenic extraction results may lead to misinterpretation of plant water sources.

Noted. We will discuss these examples to demonstrate past studies' considerations of soil water extraction errors, which, in part, further justifies our focus on plant water in this study.

Number: 6; Date: 2/4/2021 8:27:21 AM
Chen et al. recommend that stem relative water content should always be determined when performing stem water cryogenic extraction. Do you have access to this data for your selected case studies?

No, we do not have stem water content. We will discuss this in a revised version.

Page: 6
Number: 1; Date: 2/3/2021 3:41:49 AM
Elaborate which factors would contribute to a species-specific extraction bias.

We prefer to not speculate on that topic. We will change this to say "among these species".

Number: 2; Date: 2/4/2021 2:36:50 AM
I think a study-specific bias application would be more meaningful, since cryogenic extraction parameters differ for all studies and as we know from soil water extractions, such extraction parameters can have a significant effect on the resulting delta values. Further, a measurement artifact correction is then inevitable. This questions the approach the authors applied in their manuscript by simply applying the same 8.1 correction factor to a variety of different tree species without considering species differences as per Chen et al. or extraction system differences (that are difficult to account for without any additional tests on the respective system).

Regarding "questions the approach", please see our comments above. In principle a "study-specific bias application" would of course be interesting, but (as we outlined at the beginning, and discussed in the paper) we are not trying to draw definitive conclusions for specific studies. In the revision, we will further emphasize the need to account for the uncertainties in the assumed bias correction factors themselves, because those biases will vary among species (and laboratories, and investigators) in ways that will never be precisely known.

Number: 3; Date: 2/3/2021 3:16:12 AM
Please check Millar et al. (2018, doi:10.1002/rcm.8136) for that.

Noted and we will now mention this here. However, it should also be noted that our paper is focused on woody plants, and Millar et al. is not.

Page: 7
Number: 1 Author: Subject: Eingefügter Text Date: 2/3/2021 3:23:14 AM
in delta 2H direction

Yes, thank you.

Page: 8
Number: 1; Date: 2/3/2021 3:41:11 AM

The manuscript would benefit from including recommendations on what to include and how to perform bias correction on your own data.

We demonstrate how to perform a bias correction, using our own data (Section 3). If the reviewer means to suggest that we show how to estimate a bias correction factor for any specific study, we would suggest assuming that such biases exist and (in the absence of any other evidence) that they are similar in magnitude to those reported elsewhere. This is discussed in the conclusions, and elsewhere in the paper. Given how rapidly research is developing on this topic, perhaps a basis for more precise recommendations will be available soon; if soon enough, we could consider those findings.

Page: 9
Number: 1; Date: 2/3/2021 3:34:08 AM
Can you include an example of a labeling study to underline this?

Yes, that will be done in a next version. Thank you.