RC2: 'Comment on hess-2023-308', Anonymous Referee #2, 02 Feb 2024

Putman et al. use long term mean summer hydrology, gridded precipitation and groundwater isotope ratios, and in-stream water isotope ratio observations to evaluate the accuracy of the National Water Model, which is known to perform poorly at low elevations and in highly managed basins in the western United States. The authors use a water isotope mass balance approach to estimate river reach isotope ratios using National Water Model derived fluxes and compare 'modeled' isotope ratios with over 4,000 in-stream isotope observations. The differences between observed and modeled $\delta^{18}O$ and d-excess are used to evaluate statistical patterns in evapoconcentration of observations relative to modeled isotope values. Putman et al. conclude that offset between modeled and observed water isotope values are diagnostic of the lack of agricultural irrigation practices represented in the National Water Model and test this using 2015 Water Use Census data (annual water use by county) and developing an approach for estimating the amount of streamflow sourced from agricultural irrigation on a coarse catchment scale.

Overall, this manuscript is well written and provides a robust framework for evaluating model accuracy, considering different temporal and spatial patterns in the residuals, representing uncertainty in previously published datasets, and the utility of stable water isotope ratios in diagnosing misrepresentation of physical processes in hydrological models. This study is a solid contribution and worthy of publication in HESS; however, there are a few points I think need some clarification before the manuscript is finalized. Below is a discussion of my major comments, followed by more minor, line specific comments. Thank you for the opportunity to engage with this interesting and useful study.

Thank you. We appreciate your helpful review.

Major Comments

1. I'd like some additional discussion or clarity on why the interpretive framework in Figure 2 is applicable to the results of the d-excess(diff) and $\delta^{18}O$(diff) calculations presented. Typically, d-excess is calculated from the isotope ratios of a single/discrete water sample, but here d(diff) is calculated from isotope values that are a mix of spatially and temporally averaged modeled isotope ratios and point observations of in-stream isotope values. The framework in Figure 2 is based on dual isotope fractionation processes as units of water moves through the hydrological system. Given that the d-excess(diff) calculation here is based on spatially and temporally averaged modeled values, then d-excess(diff) and $\delta^{18}O$(diff), used as indices of evapoconcentration, are likely masking or missing a lot of variability in climate

conditions over time and space. Some additional reasoning would be useful for the reader. The discussion section describes how interannual variability is considered by looking at the regressions between the modeled values and the entire observational dataset versus the averaged observations, but I didn't clearly understand the lines being drawn between that interannual variability (due to variable climates?) and the consideration of how much variance is missing in the NWM that can be attributed to agricultural return flows.

I'm trying to parse the question of the reviewer, which I'm not totally sure if I understand, so forgive me if the explanation begins a little too simplistically. I don't want to make any logic leaps. Hopefully I satisfactorily address your concern in the paragraphs below.

The idea behind this framework is that it helps us interpret the meaning of the deviations of the observations from the estimates based on the mass balance approach. If the model and data inputs correctly capture all isotope-influencing sources and processes, then all points would cluster around (0,0). Instead, we see a spread along an 8:1 line as well as deviations from that 8:1 line. The structure of the deviations from the 8:1 line indicates that negative d18O(diff) tends to be associated with positive d(diff) if d(diff) is non-zero, whereas positive d18O(diff) tends to be associated with negative d(diff) if d(diff) is non-zero. This structure arises in this case because the mass balance approach tends to produce estimates with d-excess values of close to 10 (see Figure 3, gray dots), indicating no evidence of non-equilibrium processes influencing river processes (e.g., evapoconcentration (- d), mixed phase cloud processes (+d), snowmelt fractionation (+d)). On the other hand, observations have a wide range of d-excess values, though most tend to be close to or less than 10 (see Figure 3, blue dots), which are characteristic of evapoconcentration, though some plot above the GMWL, indicating potential for condensation-oriented non-equilibrium processes (e.g., snow processes). So, when the two datasets are compared / differenced, the non-equilibrium signals in the observations are highlighted.

So, the interpretation framework is predicated on the (unintentional) fulfillment of an equilibrium assumption by the mass balance approach and the deviation from that assumption by the observations. If the mass balance approach yielded some evidence of non-equilibrium signals (due to input data or process), then the interpretation framework would likely have different implications. This logic is already largely in place (in brief) in the methods:

"We can interpret combinations of d18O(diff) and d(diff) together, as well as d(diff) independently to infer the uncharacterized sources responsible for the observation-model difference. This framework is useful because the ratios of d2H to d18O of the isotopic inputs to the isotope mass balance tend to be close to 8, whereas those from the observations more often differ from 8. This means that all non-zero d(diff) values can be used to identify omitted water sources and where they are important to streamflow." However, I have added a caveat that the interpretations of the framework would change if the characteristics of the null hypothesis change (i.e., don't represent equilibrium conditions/an equilibrium assumption).

As for attribution of the source of the evapoconcentration signal, the reviewer is correct in that there are many factors that can influence the deviation of the observation from the model. We attempted to evaluate the potential influence of interannual and seasonal variability as explanations for the signal. Certainly, both of those modes of variability are responsible for some scatter in the results, as we demonstrate in the different sections of the manuscript. However, among the modes of variability we evaluated, the spatial variability was the most consistent across spatial domains and remained even when using average values. Unfortunately, due to the nature of our approach, it was not possible to evaluate all three modes of variability simultaneously, especially because of the sometimes small number of high leverage points, so the evaluation of the spatial mode certainly includes scatter from the interannual and seasonal modes of variability as well as other, unevaluated sources of variability. This likelihood of scatter from other sources of variability probably accounts for the lower predictive power of the statistical approach. However, to avoid overfitting our model, particularly in basins with fewer observations, we did not attempt to statistically evaluate all identified modes of variability simultaneously. I added a caveat specifically calling out the inability to directly address temporal aspects of variability in the methods section, and a nod to the contribution of temporal variability as a cause for scatter / low variance explained in the discussion section.

Evaluating all modes of variability at once might be possible in future studies that are able to resolve the temporal aspect of variability (i.e., producing estimates for each month and year) to match observations, and for smaller scale studies with higher temporal resolution sampling and input data. We hope that we, or others may be able to pursue this approach as an improvement to what we've put forth in this initial study.

2. The diagnosis of the National Water Model inaccurately representing agricultural return flows is well reasoned in the study and a conclusion that makes sense given the difficulty of many hydrological models in representing irrigation practices given that water use data is difficult to obtain (water users are often reluctant to share this information freely in highly managed areas). One concern I have is the practicality of reasonably incorporating agricultural return fluxes into the model and the approach for estimating this contribution taken in the manuscript. The simplified way of calculating ratios of water use contributions to stream flow in this study seems like a

reasonable first pass, but there are many unknowns. For one thing, water use can vary widely between water year types – so including some level of uncertainty or variability in the 2015 Water Use Census data would be helpful. For example, 2015 was a critically dry year in California, so water use data is likely reduced during that year compared to a wetter year on record and the ratio of groundwater to surface water use in the Central Valley is likely inflated for that year. Applying water use data that is from a specific year as a point of understanding contributions to streamflow should likely consider the isotope data from that specific year to match, since it's not representative of long-term mean conditions. I'd like to see some explicit discussion of what the water use data represents in the main text and whether it's representative of long-term conditions. One thing that could be considered (it may or may not be appropriate here) is the EPA's EnviroAtlas' dataset of different types of water use. They have estimated longer term datasets for agricultural water use, industrial, domestic, etc. https://www.epa.gov/enviroatlas

Thank you for this consideration. We agree – the use of annual-scale water use and the general approach taken could certainly lead to errors in our evaluations due to the potential for oversimplification, particularly because our observational data may come from anytime between 2000 and 2021.

Fortunately, the USGS has *just* released month and HUC12-scale estimates of water use (Haynes, et al., 2023, Monthly crop irrigation withdrawals and efficiencies by HUC12 watershed for years 2000-2020 within the conterminous United States: U.S. Geological Survey data release, https://doi.org/10.5066/P9LGISUM). We will evaluate the feasibility of replacing the analysis using the 2015 data with one using this improved dataset to address issues with uncertainty in this analysis.

If replacing the water use data product with the higher resolution version is not feasible, we will re-run the analysis using the mean of the 2000, 2005, 2010, and 2015 water use datasets. We would use these years since they are what is currently available at the same spatial scale as our original analysis.

Line Specific Comments

Line 89-92: d-excess is typically calculated for corresponding δ^{2}H and δ^{18}O values for a specific sample/observation. In this study, d(diff) is calculated from estimated average isotope values. Is d-excess still a reliable metric for evaporation when you are calculating from long term averages/values calculated using mass balance? It seems like the mass balance calculation step would not include all the non-equilibrium fractionation processes that could impact the d-excess value. Some additional reasoning somewhere in the text would be helpful for the reader.

This line-specific comment is related to the general comment on the framework discussed above. In this case, yes, it is. However, that's because the mass balance approach produces results that are effectively a null hypothesis that reflects only equilibrium conditions (i.e., d2H/d18O ratios of about 8) whereas observations vary more widely reflecting influences of non-equilibrium processes on different source waters. As mentioned in general comment 1, I have added a caveat in the description of the framework that in cases where a modeled value contains some non-equilibrium signal (d2H/d18O ratios different than 8), the interpretations of the d(diff) values may be different. I have also added "The model estimates reflect an assumption that water sources contributing to streamflow were subject only to equilibrium fraction, whereas observations indicate contributions of waters influenced by non-equilibrium processes." to the results section "Model-observation differences" to help clarify the applicability of d(diff) to diagnosing water sources with non-equilibrium conditions.

Line 108-110: remove "associated with irrigation"

Removed.

Figure 1 Caption: in text citation format typo for (Bowen 2022b)

Fixed.

Line 191: "Where available, we filled these data gaps using method outlined in Text S2." I would briefly explain that the authors used the gridded DJF precipitation isotope products to help fill the gaps, since they are listed in Figure 1, but not mentioned anywhere in the main text.

I have updated the sentence to "Where available, we filled these data gaps using either other groundwater depths or mean winter precipitation (DJF) as described in Text S2"

Line 220: typo, should be "This decision was made…"

Fixed.

Line 245: "We evaluated the results with all unaveraged observations and mean isotope ratio at river reaches with multiple observations." I'm not clear on what this means. The correlation/regression analyses would need to be done between monthly average isotope ratios for an apples-to-apples comparison, rather than mixing discrete observations with monthly average model values.

I believe this misunderstanding reflects some confusion about our approach. The mass balance results are available at the long-term average summer (JJA) season scale (not long term average monthly or year-month average scale). The comparisons are made to all data points (unaveraged) and with mean values for reaches with more than one observation (averaged).

We include the comparison results with all data (even though we acknowledge that the timescale of the observation and the modeled result are different) to evaluate the results at a greater number of reaches and thus covering more of the spatial domain. However, that approach leads to scatter due to mismatches in the time integration of the modeled vs observational data, so we also included a comparison using only reaches with more than one observation using the average observational value at that reach.

I haven't made a change to the text as I believe there is sufficient information available that the description is clear and the issue was not flagged by other reviewers.

Line 250: "(Text S3)" has an extra space after first parenthesis.

Fixed.

Figures 6: This is a really nice figure illustrating the different temporal evolution of δ^{18}O(diff) and d(diff) throughout the different major basins in the western US! Please list the distribution statistics in the caption (i.e., box represents what percentiles, what are the smaller, shaded boxes in the Great Basin boxes, diamonds are outliers?).

These kinds of plots are called 'letter-value plots' or in python, 'boxenplots' (Hofmann et al., 2017, https://doi.org/10.1080/10618600.2017.1305277). After further consideration, I think that this plot would be better off as a boxplot instead, so will update the plot and added the standard boxplot description to the caption so the meaning of the domain is clear.

Line 454: typo, remove "a" after due to

Removed.

Line 458: $p<0.001$ is listed for significance, whereas it has been $p<0.1$ or $p<0.01$ in other parts of the manuscript. I suggest staying consistent with listing the p-value in the text.

Figure 8: y-axis label I believe should be listed as ‰ instead of %.

You are correct. It has been changed.

Line 468: The meaning of the first sentence is unclear. The $\delta^{18}O$(diff) and d(diff) are statistically significant relative to what?

Yes, this sentence is not clear. We meant to say that the values of d18O(diff) and d(diff) values in headwater areas are statistically different from 0. The sentence has been revised for clarity.

Figure 9: Please explain the statistics of what the boxplots represent in the caption. Also, is not all data shown? Every land cover type looks like it has groundwater d less than -20 but that's where the plot stops.

These kinds of plots are called 'letter-value plots' or in python, 'boxenplots' (Hofmann et al., 2017, https://doi.org/10.1080/10618600.2017.1305277). These kinds of plots are useful for datasets with 1) a large amount of data and 2) show more detail about data distribution than boxplots. They are used in this case because the data are numerous and have a non-normal distribution. Letter-value plots (boxenplots) start with the median (Q2, 50th percentile) as the centerline. Each successive level outward contains half of the remaining data. So the first two sections out from the centerline contain 50% of the data. After that, the next two sections contain 25% of the data. This continues until the outlier level. The plot is cut off at -20‰ because more than 85% (and up to 95%) of the data is displayed on the plot, but the tails are quite long. The point of the figure is effectively made with the median and 75% of the data.

I have added more description of the plot type to the figure so it's clearer to the reader.

Line 556: typo, should be "stream"

Fixed.