Dear referees, dear editor,

Thank you for this second opportunity to revise and improve our manuscript and for the further suggestions of improvement. Please find below the referees' comments (in black), followed by our answers (in green) and the location of the changes we made in the manuscript (in blue). Unless stated otherwise, the line numbers we indicate are with respect to the manuscript with tracked changes, not the final revised manuscript. Our answers are followed by the manuscript showing the changes in the text, underscored in blue.

Regards

Nicolas Rodriguez, on behalf of all authors

Reviewer n°1 (Francesc Gallart, FG):

FG: The new version of the manuscript entitled 'A comparison of catchment travel times and storage deduced from deuterium and tritium tracers using StorAge Selection functions' by Nicolas B, Rodriguez et al. clearly improves the quality of the previous version in several aspects. From my point of view, this manuscript may be accepted for publication in HESS if minor changes are done for improving its clarity and scientific soundness.

We thank Francesc Gallart for taking again the time to review our work and for providing thoughtful suggestions of improvement.

FG: In the abstract it is resolutely claimed that deuterium and tritium provided similar aging of waters in Weierbach, but the time span of the results is not stated, so the reader may erroneously understand that this result is valid for any catchment with any MTT value. It is necessary to clearly state there that one of the conditions for this result is that "in catchments with limited residence times, radioactive decay may give information that is redundant with the natural variability of the tracer in precipitation" (line 481).

Good remark. However, we disagree with FG's above statement suggesting that this agreement between the tracers will be true only in catchments with "short" MTTs. We never proved this, and this is why we simply stated that there can be redundant information between the tracers when travel times are limited (which does not imply anything regarding catchment with longer travel times). Our opinion remains that if the true (unknown) catchment TTD contains a large fraction of very old water (e.g., > 10 years) and if the methods (tracer sampling, model, numerics…) are adequate (as in our study), then both deuterium and tritium should be able to suggest solutions (TTDs) with long tails which yield a good fit to the tracer output time series. Of course, in that case, tritium will likely result in smaller uncertainties than deuterium (i.e., likely more solutions with long TTD tails) because radioactive decay will more strongly discriminate TTD solutions with a short tail from those with a long tail. This is why tracers have different information contents on travel times, and not necessarily different travel times. Throughout the manuscript, we have been careful to clearly distinguish these two concepts. There is nothing that allows one to say that a priori, deuterium will absolutely not allow TTD solutions with long tails compared to tritium, which is what previous studies tended to suggest whereas it contradicts the physics underlying the transport of water towards an outlet.

We thus simply added:

"The streamflow mean travel time was estimated at 2.90±0.54 years using ^{2}H and 3.12±0.59 years using ^{3}H (mean ± one standard deviation). Both tracers consistently suggested that less than 10% of stream water in the Weierbach is older than 5 years."

and we finished the abstract with:

"In the future, it would be useful to similarly test the consistency of travel time estimates and the potential differences in travel time information contents between those tracers in catchments with other characteristics or with a considerable fraction of stream water older than 5 years, since this could emphasize the role of the radioactive decay of tritium for discriminating younger from older water."

FG: line 142: the sentence "The model's ability to simulate stream ^{2}H dynamics helped to further confirm that these flow processes are active in the Weierbach" is not acceptable. "Model performances measure the correctness of estimates of hydrological variables generated by the model and not the structural adequacy of the model vis-à-vis the processes being modelled, i.e. the hydrological soundness of the model" (Klemes, 1986).

We changed this sentence to:

"The model based on travel times presented in this study was developed in a step-wise manner based on this hypothesis of streamflow generation, and the consistency between simulated and observed δ^2H points toward a robust representation of the key processes."

The last sentences of that paragraph now read:

"Other studies carried out in the Colpach catchment (containing the Weierbach) suggested that first peaks are caused by lateral subsurface flow through a highly conductive soil layer and that second peaks are caused by groundwater flow in the bedrock (Angermann et al., 2017; Loritz et al., 2017). This is contrary to the conclusions from other studies in the Weierbach (Glaser et al., 2016; 2019), **showing that the key processes are still under debate**."

FG: Although my opinion is that input and output concentrations should be mass-proportional or-weighted to be processed in a mass balance model, I deem that the methods used in the paper may be acceptable if the way in which concentrations and masses are managed is fully explicit and the possible consequences of the methods used on the results is appropriately discussed.

Indeed, as precipitation samples for ^2H are taken at fixed precipitation intervals, the resulting concentrations yield the same result than a mass-weighting. But nothing is said about how the bi-weekly bulk samples (time-proportional) are managed and merged with the mass proportional automatic samples. I do not mind if mass-proportional concentrations are interpolated to produce a 'continuous' signal because the mass is conserved. Furthermore, nothing is said about the ^3H sampling; were differences in monthly precipitation taken into account to weight input ^3H activities as usually done? Therefore, were ^2H and ^3H concentrations managed in the same or in different ways (precipitation weighting for ^2H and time weighting for ^3H)?

Respect to stream water sample concentrations, nothing is said in the manuscript but it should be clearly stated whether these were managed as unweighted discrete irregularly taken samples or were time- or flow-weighted. Furthermore, something about flow-weighting the available concentration samples should be included in the interesting discussion in lines 647-652 where the possible advantages of flow-proportional sampling are commented.

We added many more details in section 2.2 (see lines 167-170, 188-192, 195-203), and some details in sections 2.3 and 2.4 (see lines 213-214, 217-220, 242-245) to try and remove any remaining ambiguity about concentration weighting. Precipitation samples do not need to be weighted by precipitation amounts. They all are representative of the tracer mass flux in precipitation by design since they all are cumulative samples (and not based on fixed time intervals), and because we calculated the input concentration signals to make sure that the integral of [concentration times precipitation amounts] over a given period is always equal to the total tracer entering the catchment over that period. Simply stated: $C_{model\_input} \times \sum Precip = true\ \sum(C_{precip} \times Precip)$.

Stream samples are all instantaneous grab samples, and there is no need to weight them by streamflow because the time-varying TTDs already account for the fractions of precipitation water not reaching the stream due to ET or storage and for time-varying discharge rates, unlike the steady-state TTDs which require flow-weighting (see lines 217-220).

We also added these sentences in the discussion (lines 697-700):

"It is important to notice that weighting the available stream samples by streamflow in the calibration (i.e., calibrating on tracer loads instead of concentrations) would not compensate for this relative absence of samples during high flow conditions. In addition, it would bias the

calibrated TTDs towards high flow conditions, while our goal is to have TTDs which accurately represent the functioning of the catchment over all flow conditions (the whole 2015-2017 study period)"

FG: line 562: "Our conclusions are valid because the model captures accurately the travel times in the Weierbach" is really an inappropriate statement. This seems to claim that the model is hydrologically sound (in the sense of Klemes, 1986) because it reproduces well something that cannot be validated.

Good remark, we changed this to:

"The visually satisfactory tracer simulations enhance our confidence that the model accurately simulates travel times in the Weierbach."

FG: Line 295. I understand that model efficiency assessment and subsequently the efficiency thresholds for selection of behavioural models might be different for deuterium and tritium, but nothing is discussed afterwards on the possible role of this difference on some of the results obtained. For instance, more parameter sets are accepted as behavioural for tritium than for deuterium; this may be reasonable because sampling is much intensive for deuterium, so model rejection may be stricter for it, but some comment about this issue would be welcome.

We added this (lines 628-635):

"Our choice of performance measures (E_2=NSE and E_3=MAE) and selection criteria (L_2=0 and L_3=0.5) resulted in slightly more TTDs constrained by tritium than TTDs constrained by deuterium (148 curves for $E_2 > 0$ against 181 curves for $E_3 < 0.5$). These numbers are highly sensitive to performance thresholds, and our choices represent the closest match in the number of accepted solutions for each tracer, while considering only meaningful performance criteria variations (i.e., ≥ 0.1) and acceptable model performance. This guarantees a similar treatment of the two tracers (i.e. it avoids biases in travel times for a given tracer), while accepting only satisfying simulations for both tracers. Future work could assess the sensitivity of travel time differences between tracers for other performance measures and thresholds, and for contrasting numbers of accepted solutions."

FG: As written, the reader may understand that Stewart et al (2010) found or reported travel time differences up to 5 years at Weierbach.

We corrected this to:

"The TTDs obtained from each tracer were broadly consistent in shape, and the travel time differences were considerably smaller (i.e., <1 yr) **in the Weierbach** than in a previous comparison study **in four catchments from Germany and New Zealand** (up to 5 yr, Stewart et al., 2010)"

FG: line 440: "First, we treated 2H and 3H equally by calculating TTDs using a coherent mathematical framework for both tracers (i.e. same method and same functional form of TTD)" though sampling and model efficiency were differently managed (as discussed later).

Good suggestion. This now reads:

"First, we treated ^{2}H and ^{3}H equally by calculating TTDs using a coherent mathematical framework for both tracers (i.e. same method and same functional form of SAS function). However, sampling frequency and model efficiency criteria needed tracer-specific adaptations (see Sect. 4.4.2 and 4.4.3)."

FG: line 524: “Performance measures E2 and E3 are” not identical but are “both based on minimizing a sum of...”

Changed as suggested.

Reviewer n°2 (R2):

R2: Thank the authors for taking the time to respond to the previous comments and revise the manuscript. However, there are several important points that I still can't agree with the authors. Also, I still think that the authors miss some critical points and rather rely too heavily on their model results to support their conclusions. Those points obscure what readers can learn from this study. I believe that this manuscript still has potential, but some issues need to be resolved before it can be considered for publication.

We thank R2 for the additional time spent on reviewing our work and for the additional comments. We agree with the reviewer that some of our statements were not nuanced enough and thus underrated some critical learning for the reader. We modified the manuscript accordingly at most instances or reasoned more clearly when we did not.

R2: THE LIMITATIONS OF THE DATA AND THE STUDY SITE

I noticed that the authors mentioned many limitations of this study in one section pretty well, but there are some crucial limitations that can obscure their major arguments significantly. I list some of those limitations here with my opinions.

Figure 3 shows that there are no tritium samples at high flow conditions, and thus, one cannot learn transport dynamics at the high flow conditions using ^{3}H no matter which model is used. Thus, I believe that any arguments based on the ^{3}H-based model results at high flow conditions (e.g., the TTDs at high flow conditions) are risky because such results are just based on "extrapolations" that the model did. In their result interpretation, the authors mostly used the TTD weighted by discharge, which is in part based on the TTDs estimated at high flow conditions and even gives more weights to those TTDs. Thus, many (and most) of the arguments in the abstract and the conclusion such as "Tritium and stable isotopes both had the ability to reveal short travel times in streamflow", "The travel time differences were small compared to previous studies, and contrary to prior expectations, we found that these differences were more pronounced for young water than for old water", "our results highlight that stable isotopes and tritium have different information contents on travel times but they can still result in similar TTDs.", and "We conclude that stable isotopes do not seem to systematically underestimate travel times or storage compared to tritium" are based on the extrapolation. Drawing scientific conclusions based on extrapolation is risky and not a good practice.

We understand R2's concerns, but we disagree with R2's above claims. Figure 3 shows several samples at high flow conditions (4 samples out of 24 correspond to flows exceeded around 10% of the time only, i.e. to Q > 0.1 mm/h, see Figure 5 and 6), even if they do not represent the highest flows recorded (as pointed to by Francesc Gallart in the last round of revisions, see reply to his previous comments). However, and more importantly, this comes down to a philosophical question whether we can use a model to derive some conclusions or if we need to fully rely on data sets that completely sample all occurring flow stages and scenarios. This of course does not only include sampling along the whole FDC, but all sorts of combinations between flow and antecedent conditions including hysteresis between storage and flow as well as the ET regime. For us it is evident from the literature that "extrapolating" between a limited number of data points using a model is a ubiquitous practice in time-varying travel time studies (see Benettin et al., 2015a; 2017a; Birkel et al., 2015; Harman, 2015; Heidbüchel et al., 2012; Hrachowitz et al., 2013; Klaus et al., 2015, Rodriguez et al., 2018, 2019, 2020). This is especially true for the previous (steady-state) travel times studies using tritium (e.g., Maloszewski & Zuber, 1982, 1993; Gallart et al., 2016), which worked with a much smaller number of tritium samples. In fact, some studies have less than 5 points over several years of observation. As argued by Francesc Gallart, not many

travel time studies tended to report their samples along streamflow values as we did in Figure 3, and we can expect that many also had to considerably "extrapolate" (for low or high flows), using their model in order to derive meaningful conclusions on travel times and related hydrological processes. This is somehow opposite to data-based methods (e.g. Kirchner, 2019), which tend to use only the available data to draw conclusions. As a result, data-based approaches as very data-hungry (typically needing tens of thousands of data points) and cannot be used to derive conclusions between data points or outside of observation periods. Calibrating a model to a tracer time series having less than 100 points as we did is helpful to round this issue of limited data availability. We would like to point out again that our tritium data set is one of the densest ever recorded, as stated in one of the sentences we added to the manuscript in the previous revision (see lines 179-181). However, additional data or different tritium data sets may confirm or challenge our current findings in the future.

We agree that using flow-weighted TTDs will accentuate the role of TTDs at high flows. But time-averaged TTDs will also contain the instantaneous TTDs at high flows. Therefore, there is no obvious solution to avoid using the "extrapolations". The current travel time literature does not contain much guidance on what TTD to use in the analyses for what purpose (i.e. flow-weighted, or time-averaged). We believe that by default, flow-weighted TTDs should be used as they allow a more meaningful comparison between catchments with contrasting flow regimes. For example, the same time-averaged TTD for two contrasting catchments (say, ephemeral stream vs wetland) could hide very different catchment functionings which are better revealed by their different flow-weighted TTDs. If high frequency tritium data sets are available in the future, it would be very helpful to compare the TTDs from high-frequency data with the TTDs obtained by re-sampling a coarser time series (see relevant comments on this lines 556-560), and to see whether time-averaged TTDs yield smaller differences between a full data set and a limited data set than flow-weighted TTDs.

Finally, as explained in section 4.4.3 (lines 703-705): "the larger water mass not sampled for tritium is not leading to a strong bias towards young or old water compared to deuterium. The latter is shown by the good agreement between the TTDs constrained by deuterium and the TTDs constrained by tritium.", and it is worth keeping in mind that the deuterium data set is representative of the higher flow conditions (see Figure 3).

R2: Another problem is that travel time is relatively short in this catchment. It has been argued that the tritium is beneficial as it allows us to examine long time-scale transport dynamics (e.g.,> ~4 years in Stewart et al., 2010).

Again, we understand R2's concern. But first, we would like to stress again that that the statement of Stewart et al. (2010) is true only for very specific conditions (lines 572-576):

"The theoretical span of 0–4 years pointed out in Stewart et al. (2010) should however not be taken as the only range of travel times where ^{18}O, ^{2}H, and ^{3}H may have redundant information. As clearly written by Stewart et al. (2010), this limit corresponds to a steady-state exponential TTD only, while other TTD shapes (or unsteady TTDs) could yield much higher limits. More importantly, this limit can be lowered by the seasonality of the input function (see Stewart et al., 2010, p. 1647)."

However, in this catchment, the travel time is relatively short in general, and a considerable fraction of TTD (> ~90%) is defined over the travel time less than 5 years (based on Table 3). This short travel time obscures the relative importance of the use of ^{3}H to examine longer time scale transport dynamics because longer time scale transport is less important (or negligible) in reproducing the tracer dynamics in this catchment.

We agree that the travel times in our catchment are relatively short and that it may limit the potential of ^{3}H for discriminating younger from older water thanks to radioactive decay, as already clearly stated in the previous version of the manuscript (lines 515-520, see below):

"The travel times being below ~5 years in the Weierbach (Table 3) could be another reason for the limited information of ^{3}H on older water. ^{3}H decays by only about 25 % in 5 years, meaning that all the tritium activities of the water in the Weierbach have varied by at most ~2 T.U. since water entered the catchment. This is much lower than the 10 T.U. amplitude of tritium variations in precipitation. Thus, in catchments with relatively short residence times, radioactive decay may give information that is redundant with the natural variability of the tracer in precipitation".

We are not sure about the intention of this specific comment, however the reviewer seems to target defending the use of tritium. We also see great potential in the use of several tracers to derive catchment TTDs. However, as we argued throughout the manuscript -- and clearly supported by the calculated TTDs -- there should not be any physical difference in age derived from two tracers. Here, we are actually able to reconcile (with uncertainty) TTDs from tritium and deuterium when we are using a coherent methodological framework. Yes, this is done for rather short TTDs, but by no means would that be an argument for expecting different TTDs for the same water parcel when the TTDs become longer. However, data sets to study this in catchments with longer travel times are currently not available, and we are looking forward to see them and similar research for a wide range of TTDs.

It is also worth noting again that numerically, we have travel times up to 100 years in the catchment (Figure 7), but also that the fraction of stream water older than about 5 years (~10%) is not negligible, and that this range of travel times is in fact the one on which the tracers agreed the most (table 3, table B1).

Therefore, I worry if the study site is adequate, and I'm not sure about the worth of most of their arguments such as "The travel time differences were small compared to previous studies, and contrary to prior expectations, we found that these differences were more pronounced for young water than for old water", "we did not find that stable isotopes are blind to old water fractions as suggested by earlier travel time studies ", "Based on the results in our experimental catchment in Luxembourg, we conclude that the perception that stable isotopes systematically truncate the tails of TTDs is not valid", "our results highlight that stable isotopes and tritium have different information contents on travel times but they can still result in similar TTDs.", and "We conclude that stable isotopes do not seem to systematically underestimate travel times or storage compared to tritium". Is a similar conclusion expectable for a catchment that has a longer travel time? Or is it just because the studied catchment has short water travel time in general? If the latter is the case, what can be said about the well-known importance of tritium tracer by studying this catchment?

There is nothing like a non-adequate study site for our posed questions. That said, similar studies in several catchments with a range of travel times would be highly appreciated in the future. No research group has tackled this question until today and thus other data sets are not at hand. However, that a water parcel cannot have two different mean travel times is independent of a catchment, and here we provided first insights that a consistent approach reveals similar TTDs and MTTs for both tracers. In the first review, Francesc Gallart had a similar remark about relatively short travel times, and we answered him:

"we disagree with FG's above statement suggesting that this agreement between the tracers will be true only in catchments with "short" MTTs. We never proved this, and this is why we simply stated that there can be redundant information between the tracers when travel times are limited

(which does not imply anything regarding catchment with longer travel times). Our opinion remains that if the true (unknown) catchment TTD contains a large fraction of very old water (e.g., > 10 years) and if the methods (tracer sampling, model, numerics…) are adequate (as in our study), then both deuterium and tritium should be able to suggest solutions (TTDs) with long tails which yield a good fit to the tracer output time series. Of course, in that case, tritium will result in smaller uncertainties than deuterium (i.e., likely more solutions with long TTD tails) because radioactive decay will more strongly discriminate TTD solutions with a short tail from those with a long tail. This is why tracers have different information contents on travel times, and not necessarily different travel times. Throughout the manuscript, we have been careful to clearly distinguish these two concepts. There is nothing that allows one to say that a priori, deuterium will absolutely not allow TTD solutions with long tails compared to tritium, which is what previous studies tended to suggest whereas it contradicts the physics underlying the transport of water towards an outlet."

This is why we added this to the abstract:

"The streamflow mean travel time was estimated at 2.90±0.54 years using ^{2}H and 3.12±0.59 years using ^{3}H (mean ± one standard deviation). Both tracers consistently suggested that less than 10% of stream water in the Weierbach is older than 5 years."

and we finished the abstract with:

"In the future, it would be useful to similarly test the consistency of travel time estimates and the potential differences in travel time information contents between those tracers in catchments with other characteristics or with a considerable fraction of stream water older than 5 years, since this could emphasize the role of the radioactive decay of tritium for discriminating younger from older water."

We believe that "the well-known importance of tritium tracer" has in fact been overstated in the previous studies, and mistakenly based on data and methodological limitations that we tried to overcome in this study (see the introduction in our manuscript). Starting from the assumption that tritium is the only tracer revealing old water necessarily leads to circular reasoning. For example, one common mistake from the past has been to focus tritium sampling on baseflow, based on the implicit assumption that it is more useful for old water. This of course naturally biased the results towards old water by sampling design (because baseflow does contain older water than hydrographs by definition). This mislead many to think that tritium reveals older water than deuterium. We could similarly declare by mistake that deuterium is more informative on short travel times than tritium if we sampled deuterium only during large hydrological events (as in isotope hydrograph separation) and not tritium (notice that it seems to be the case, currently). Conversely, assuming nothing a priori and treating the tracers as equally as possible resulted in similar TTDs based on deuterium and tritium despite the different tracer treatments imposed by sampling and calibration differences later on in the analysis.

Finally, we do support the idea that tritium is a very useful tracer, as it is more age-specific than deuterium. This is why we wrote (lines 520-526):

"In a few decades, water recharged in 1980–2000 may have completely left the catchments or may be a negligible part of storage, such that the log(^{3}H) of stored water may increase linearly with residence time (see the recent increasing trend in $C_{P,3}$* in Fig. 2). Thus in a few decades, tritium could be even more informative about old water contributions because there may be no travel time ambiguity anymore. Furthermore, the oscillations of tritium in precipitation over long time scales (>10 years) recently detected and related to cycles of solar magnetic activity (Palcsu et al., 2018) may give stream tritium concentrations even more age-specific meaning. Therefore, it is

important to re-iterate the call of Stewart et al. (2012) to start sampling tritium in streams now and for the next decades to use it in travel time analyses."

R2: THE MODEL AND THE RESULT INTERPRETATIONS

Again, their model cannot reproduce some short time scale transport dynamics (based on Figure 5 and the low NSE values).

We wonder if R2 saw the figures we added in the supplement, as we are convinced that they better show the ability of the model for reproducing the short time-scale transport dynamics (flashy peaks) for many events (blue envelopes, fig. S1-S9). We also already explained (also in more detail in Rodriguez and Klaus, 2019) that the NSE is not an absolute, universal, and perfectly objective way to estimate model performance (lines 608-617). This is especially true for tracer simulations, for which customized objective functions or visual inspections are sometimes preferred (Stadnyk et al., 2013; Gallart et al., 2016; Rodriguez and Klaus, 2019). It seems that we are simply reaching a difference of opinion about model performance, and we don't see how we can further improve the manuscript on this aspect.

Gallart, F., Roig-Planasdemunt, M., Stewart, M. K., Llorens, P., Morgenstern, U., Stichler, W., Pfister, L., and Latron, J.: A GLUE-based uncertainty assessment framework for tritium-inferred transit time estimations under baseflow conditions, Hydrological Processes, 30, 4741–4760, https://doi.org/10.1002/hyp.10991, 2016

Stadnyk, T.A., Delavau, C., Kouwen, N. and Edwards, T.W.D. (2013), Towards hydrological model calibration and validation: simulation of stable water isotopes using the isoWATFLOOD model, Hydrol. Process., 27: 3791-3810. doi:10.1002/hyp.9695

Such an inadequate model structure underestimates the information content in ^{2}H by resulting in not well constrained posterior parameter distributions for the behavioral models.

In the last round of revisions, we discussed that in our opinion, the parameter distributions are constrained to some extent because the posteriors are not flat. This is more precisely and objectively quantified by the fact that we had non-negligible reductions of entropy and information gains D_{KL} (lines 564-566). Thus, it is clear that there is a constraint from the data, however, we agree that uncertainty remains.

If they had a model that captures short time scale dynamics well, posterior distributions of the associated parameters of such a model could be more constrained than a ^{3}H based model.

We believe that this is a strong assumption and that it tends to contradict the philosophy underlying uncertainty analysis (e.g., GLUE, DREAM). In particular, this idea contradicts the "bias-variance trade-off" principle. In simple terms, parameter distributions generally tend to get flatter and flatter with increasing model performance past the optimum point defining the adequate trade-off between model performance (here, NSE) and model complexity (here, uncertainty, roughly proportional to the number of parameters) (James et al., 2013). Thus, a model performing better does not automatically imply that the behavioral parameter ranges will be narrower. For example, we could use a new model version with >100 parameters to reach a perfect model performance (e.g., NSE=1). Then, it is very likely that the parameter distributions will all be flat. This is because we would be going far beyond the optimum for the trade-off.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning with applications in R, (pp. 33–37). New York: Springer-Verlag. https://doi.org/10.1007/978-1-4614-7138-7

Thus, more information can be learned from ^{2}H, compared to a ^{3}H-based model, than what is described in this paper. Therefore, I disagree with the authors' argument that a better performing model would not change the conclusion of this study. For example, their argument "Tritium was slightly more informative than stable isotopes for travel time analysis despite a lower number of tracer samples" is susceptible to their model structure.

A priori it would appear that more information can be learned from using ^{2}H because of the larger number of stream measurements. However, our results show that the larger number of samples is not enough to tip the scales in favor of ^{2}H. This is because ^{3}H is inherently more informative on travel times due to radioactive decay (lines 21-22, 509-515, 808-810), and because we estimate that not all ^{2}H samples were equally informative on travel times (lines 550-553). We disagree that a model working better for simulating ^{2}H would necessarily increase the amount of information learned, because this would not necessarily narrow the posterior parameter distributions (see reply to previous comment). It is true that the information contents on travel times for each tracer are model-dependent in our analysis. Ideally, the travel time information contents might be compared using a shape-free (data-based) TTD estimation method (e.g., Kirchner, 2019). However, it is important to notice that there will always be underlying assumptions affecting the final result, similar to the choice of a model structure in our work. For instance, in the case of Kirchner's method, there is an implicit assumption (among others) of multilinearity between tracer inputs and outputs. This is also a model (a statistical model, precisely), which tends to be overlooked, because the method is deemed "model-free" or "data-based". Moreover, there are currently no data-based methods to estimate time-varying TTDs, while working in unsteady conditions is a requirement nowadays.

Following the reviewer's comments, we modified the manuscript accordingly (lines 560-563):

"Finally, the information contents on travel times that we have derived depend on our model structure (number of control volumes and SAS functional form). More work is needed in developing 'model-free' (e.g., data-based) unsteady TTD estimation methods in order to reduce the dependence of the results on modeling assumptions."

and this to the conclusion (lines 740-742):

"More work is also needed to compare the information contents of the tracers on travel times using data-based approaches in order to avoid a dependence on model structure"

Tritium was useless at wet conditions (because they have no samples at wet conditions), and the ^{2}H-based model was not constrained well at wet conditions (which underestimate the information content of ^{2}H in their analysis method) because of its structural problem.

These claims remain unsubstantiated and we have to reject them. Tritium was useful at "wet conditions". We sampled streamflow during "wet conditions" (see reply to previous comment). Moreover, "the ^{2}H-based model was not constrained well at wet conditions" is also not a correct statement. We did not evaluate parameter distributions for "wet conditions", but for the whole study period (2015--2017). We think that perhaps, R2 confuses the concepts of model performance and parameter identifiability (as shown by the previous comments). What R2 meant is probably that the ^{2}H-based model struggled more to simulate "wet conditions", probably meaning the flashy events in R2's mind. To this, we can answer that some limitations of the model to simulate the flashy events for ^{2}H correspond mostly to drier conditions (see lines 636-639) during which these flashy events are visible (while during wetter conditions they are "drowned" in large flow volumes associated with a more damped isotopic signature).

R2: Also, the authors reported that the ^{3}H-based model learned more information (4.47 bits) compared to the ^{2}H-based model (which learned 4.08 bits of information). The authors argued that this is because ^{3}H informed the model about ET processes more, compared to ^{2}H, based on the posterior distributions of the model parameters (in line 504). However, it didn't come with any scientific reason why ^{3}H would inform more about the ET processes. I don't think that there is any literature on it, and I personally can't think of any reason. Without a scientific basis, the result seems just an artifact of their model structure and their method of analysis. Why would ^{3}H inform more about the ET processes than ^{2}H?

We thank R2 for this remark. However, it is important to not confuse ET travel times and ET processes. We never stated anywhere in the manuscript that tritium informed us more about ET processes. Rather, we wrote (line 542): "This is because tritium considerably informed us about the **travel times** in ET", and (lines 806-807): "Tritium was more informative on **travel times** than deuterium due to its stronger constraint on the parameter values of Ω_{ET}, μ_{ET} and θ_{ET}".

We hope R2 saw the detailed explanations we added on this in the last version (appendix A2). In brief, the parameters of the ET SAS function (μ_{ET} and θ_{ET}) have a non-negligible influence on the accuracy of streamflow isotopic simulations. This is even more pronounced for tritium simulations because radioactive decay implies that the age selection patterns for ET (i.e., its SAS function parameters) have a stronger influence on the accuracy of the long-term isotope balance in the catchment than for deuterium. We nevertheless realised that one additional sentence was necessary to strengthen the reasoning.

We modified the sentence lines 543-545 to:

"The particularly large information gains on μ_{ET} and θ_{ET} with tritium reveal a stronger influence of Ω_{ET} on the accuracy of stream tracer simulations than for deuterium, via an indirect influence on isotopic partitioning (App. A2)"

We added the following in appendix A2 (lines 806-812):

"Tritium was more informative on travel times than deuterium due to its stronger constraint on the parameter values of Ω_{ET}, μ_{ET} and θ_{ET}. Based on the reasoning above, this is simply due to the fact that the relationship between T and $C_{P^*}(T,t)$ is clearer for tritium due to its radioactive decay than for deuterium, for which there is essentially no relationship between travel time and tracer concentrations. In conclusion, information on the parameters of Ω_{ET} exists in the time series of $C_Q(t)$ and can be extracted by calibrating the model based on SAS functions, particularly from using tritium."

R2: MINOR COMMENTS

Line 424: Typo in $S_T \in [0,+\infty$ "[".

This was already stated in the first reviewer report; however, we do not see a typo. Perhaps due to the PDF reader R2 uses?

$$S_T \in [0, +\infty[$$

425 Ω_Q (called he

Figure: picture of line 424 (using Foxit Reader)

R2: Line 471: "The travel time and storage measures estimated from a joint use of ^{2}H and ^{3}H are the highest (tables 3 and 4)." This result is counter-intuitive. Why the joint use gives the highest travel time and storage, not something in the middle?

This is a good remark. Indeed, our intuition may suggest to use the arithmetic mean of the travel time and storage measures for ^{2}H and ^{3}H when combining both tracers, effectively resulting in a result "in the middle". However, this would be conceptually wrong, because this would correspond to the travel times and storage measures for the simulations constrained by ^{2}H or ^{3}H, while we are interested in those corresponding to simulations constrained by ^{2}H and ^{3}H. This can be seen by considering the various sets (populations) of simulations constrained by a given tracer or by a combination of the tracers. When selecting simulations constrained by ^{2}H and ^{3}H, we are in fact selecting a subset of all behavioral simulations (see Fig. 4). This subset is not a representative sample of the whole population (see Fig. 4), and it contains a particular selection of simulations which favor longer travel times and storage measures (this is called a sampling bias when making polls). This is because ^{2}H and ^{3}H have information in common about longer travel times.

We added the following after the cited sentence (lines 507-509):

"These measures are not intermediate values (i.e., the average of the results from the individual tracers) because deuterium and tritium have information in common about longer travel times (i.e., the simulations constrained by both tracers are a specific selection among all accepted simulations, see Fig. 4)."

R2: The 2010-2015 data that was used in the spin-up period need to be presented.

Good remark, we added 2 figures in the supplement.

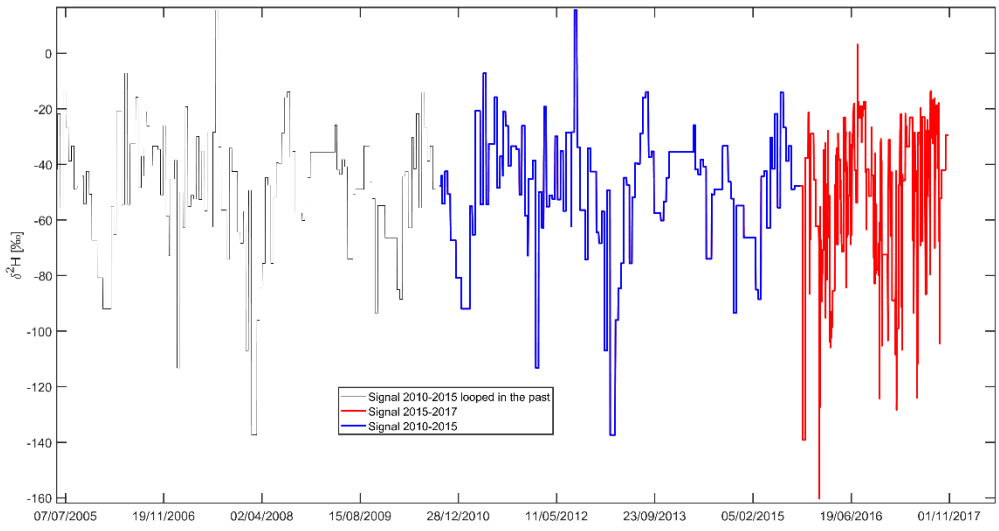

Figure S16. Spin-up data used in the model for deuterium ($\delta^{2}H$). The 2010-2015 measured data is looped back many times over the 1915-2015 period (black curve, only one repetition is shown). The 2015-2017 data is not used in the spin-up.

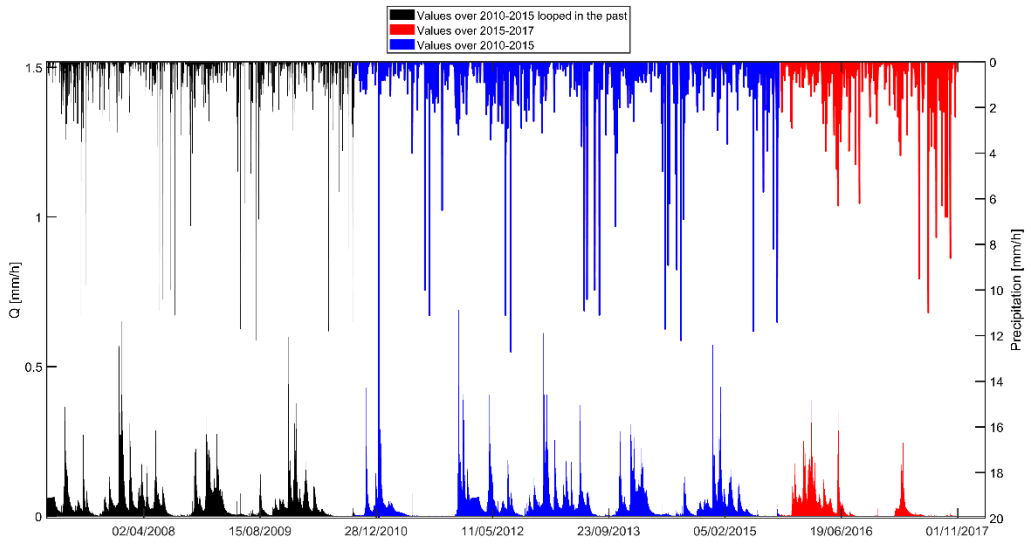

Figure S17. Spin-up data used in the model for streamflow (Q) and precipitation (J). The 2010-2015 measured data is looped back many times over the 1915-2015 period (black curve, only ~1 repetition is shown). The 2015-2017 data is not used in the spin-up.