FG: The authors propose a very interesting piece of work that may shed light on the future joint use of deuterium and tritium isotopes on water age studies. The volume of the original analytical information is outstanding, the text is a little verbose but clear, the graphs are explicative and the rationale and methods are well explained although a relevant part is not described as it is under review in another journal.

Nevertheless, there are a few methodological issues that should be fixed or justified
before the manuscript is acceptable for publication.

Authors: We thank Francesc Gallart (FG) for the overall positive reception and constructive evaluation of our work. Please note that the mentioned study is now published (open access) in WRR as:

Rodriguez, N. B., & Klaus, J. (2019). Catchment travel times from composite StorAge Selection functions representing the superposition of streamflow generation processes. *Water Resources Research*, 55. https://doi.org/10.1029/2019WR024973

We are grateful for FG's relevant suggestions and we will provide appropriate modifications in order to improve to the manuscript accordingly.

FG: The procedure used by the authors to test the "truncation" hypothesis "that streamflow TTDs calculated using only deuterium ($^{2}H$) or only tritium ($^{3}H$) are different" does not follow the established methods for hypothesis testing. As a rule, for rejecting a null hypothesis it is necessary to verify that its probability is lower than a prefixed assumable error risk, typically $p<0.01$. High uncertainty of the results is not sufficient for rejecting a null hypothesis.

Authors: We understand that strong statements such as "We found equal TTDs and equal mobile storage between the $^{2}H$- and $^{3}H$-derived estimates" and our use of the words "hypothesis", "reject", or "testing" in the title could be interpreted as if we applied some statistical test in the traditional framework of hypothesis testing. Our intention was not to conclude on the statistical significance of the results, but rather to show that the potential water age differences obtained with the two tracers are not as drastic as generally expected since the study of Stewart et al. (2010). Our goal was thus to show a counterexample to the conjecture that the tails of the TTDs are systematically truncated when using seasonal tracers. We will thus revise the manuscript accordingly, to avoid misinterpretations. Notably, we will change the word "testing" in the title with "assessing". Moreover, the scientific method does not rely only on statistical hypothesis testing to move forward, for various reasons (Pfister and Kirchner, 2017). Important hydrological conjectures, such as the idea that streamflow is made only of overland flow, were proven wrong without a probability criterion because new experimental data (e.g. strong damping of stable isotopic signatures) provided clear evidence in favor of alternative explanations (Kirchner, 2003).

We did not mean to use the parameter uncertainties as a criterion to assess if the water age differences can be considered statistically significant or not. Instead, we simply pointed out that the observed differences are small. Since "small" is always subjective, we compared these age differences to what we had available, i.e. the parameter uncertainties. This comparison raised the question whether the age differences can be confidently interpreted as representative of a TTD truncation issue or not. We will revise this part of the discussion to make it clearer that the age differences are in fact smaller than what was expected based on the study of Stewart et al. (2010), and that this is actually the main reason why we doubt that the TTD tails are systematically truncated when using only deuterium as a tracer.

FG: The authors found that "differences between the various statistics of the TTDs were smaller than the uncertainties of the calculations when comparing the results obtained with $^{2}H$ alone and with $^{3}H$ alone". But the authors also state that "even though the uncertainties are sufficient to account for the differences between $^{2}H$- and $^{3}H$ -derived age and storage measures, it is worth noticing that $^{3}H$ systematically gave higher estimates". Therefore, even if the authors did not estimate the probability of the null (truncation) hypothesis, this last sentence suggests that its probability was not sufficiently low for rejecting it, so the result of this work is that the authors cannot reject the "truncation" hypothesis

Authors: We thank FG for pointing out this potential interpretation issue that can be addressed with a proper statistical analysis. We will therefore add a Wilcoxon rank sum test to the revised manuscript. The results

show that there is a statistically significant difference in most (but not all) of the age measures shown in table 3 (e.g. median age, mean age). We will include these results in the appendix and refer to them in the discussion.

However we believe that these results do not change the core message of the study, for various reasons. First, as mentioned above, the age differences are small compared to those suggested by Stewart et al. (2010) and subsequently assumed by many researchers working with tritium. For example, the largest age difference we found (41%) was actually for the youngest water fractions, while our mean travel times differed only by <7%. In contrast, the mean travel times compared by Stewart et al. (2010) can for example differ by a factor of nearly 200%. Second, as written in the discussion, we think that these age differences can be mostly explained by the large difference in the number of tritium samples (24) compared to deuterium samples (>1000). Although the statistical analysis suggests a significant difference between $^{2}H$- and $^{3}H$- derived water ages, it is really important to remember that this analysis does not take into account the large difference in the number of tracer samples! Let's imagine the opposite situation: 24 samples for deuterium and >1000 for tritium, especially keeping in mind figures 6a and 6b. How would behavioral simulations look then? It is then difficult to say a priori whether the corresponding TTDs would be similar to those found now, and whether the TTDs would then be consistent between $^{2}H$ and $^{3}H$. We believe that currently, with only 24 tritium samples compared to >1000 deuterium samples, it is very unlikely that the consistency we found between the TTDs is a simple coincidence.

We will carefully reformulate the abstract, the discussion, and the conclusion, to include the statistical results, and to soften the claim that the TTDs are equal. Rather, we will present that the $^{2}H$- and $^{3}H$- derived TTDs are mostly consistent in terms of shape and percentiles (e.g. mean). We will also add in the discussion another potential physical interpretation about water age differences with respect to the self-diffusion of HDO and HTO in water.

FG: Furthermore, this hypothesis testing exercise had other issues. Indeed, although the authors "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 TTD)" they did not treat these isotopes with similar sampling strategies. Indeed, nearly 30 stream samples of $^{3}H$ collected during highly varying flow conditions cannot be compared with the 1088 stream samples of $^{2}H$ collected every 15 hours on average, even if the period was the same. Among the diverse causes that can explain the modest differences found between the results obtained with deuterium and tritium, the potential role of the different sampling strategy must be taken into account (differences respect to the sample number and flow representativeness, as also suggested by the authors in the discussion). The test performed by the authors compares the results and potentials of both isotopes when used under the current state of the art but not their own potentials. A rigorous test for comparing the own potentials of both isotopes would need to use an equal number of samples taken simultaneously for both.

Authors: Given the measurement techniques limitations and price, we are not sure that the concept of "own potential" can be clearly defined if the tracer signals are not continuous (i.e. with an infinite number of points). Indeed, each tracer will always be associated with a given (finite) number of samples, and this number of samples for $^{3}H$ will most likely be much lower than the number of $^{2}H$ samples unless the sampling for deuterium is voluntarily coarse. One may think that it could be useful to restrict the number of $\delta^{2}H$ samples to match the number and/or the timing of $^{3}H$ samples in order to define this "own potential". The first issue is that it would correspond to ignoring the facts (the measurements we already have), i.e. ignoring the true variability of $\delta^{2}H$ in favor of that of $^{3}H$, which appears conceptually wrong to us. We know that $\delta^{2}H$ varies in such a way and there is information (quantifiable, see section 2.7) to gain from it. Ignoring samples can only reduce the amount of information extracted from the tracer data, or worse, support the wrong interpretations. Moreover, in our case there are already more than $10^{48}$ ways to select 24 samples among 1088. It is nearly impossible to test all combinations. Even by being more strategic, for example by using a flow duration curve (FDC) to select 24 deuterium samples among 1088, there is still a lot of subjectivity involved. For instance, selecting samples distributed along the FDC implies a hidden assumption of a one-to-one relationship between a given flow value and streamflow generation processes or catchment state variables such as soil moisture, groundwater levels, or catchment storage. This means that one can never be sure that all "end members" or "wetness states" or "streamflow generation processes" are accurately represented in the selected tracer data set with such a method, and that there may always be a sampling bias. Finally, we did try to compare the "own potentials" in the discussion (4.3) by showing the amount of water

age information gained per deuterium/tritium sample or per €. This normalization per price or per number of samples allowed us to take some perspective on the results and to quantify to what extent tritium seems more age-informative than deuterium for our current number of samples, without having to ignore any deuterium measurement.

FG: This leads to another relevant issue on the sample treatment. The authors, as commonly made, weighted the isotopic signal of rainfall waters with the respective rainfall depths. But nothing is stated on the weighting of stream samples, as regrettably also recurrently made. So the reader has to assume that the raw (unweighted) isotopic signals of stream samples were used for constraining the model.

Authors: We did not state in the manuscript that we weighted the isotopic signal of precipitation with respect to precipitation amounts. We will clarify in sections 2.2 and 2.3 (especially equations 1, 2, and 3) that it is the unweighted signals (for stream and precipitation samples) that are used. Weighting functions for the input signal were introduced in travel time theory in early studies that considered only groundwater systems because these could reasonably be assumed to be at steady-state (Maloszewski and Zuber, 1982). In this case, the input function of groundwater systems is not described well by the precipitation signal because of mixing due to the complexity of flow paths in the unsaturated zone and because of water losses to the atmosphere via ET. It is not necessary to use an input weighting function with time-varying TTDs that consider the whole catchment and obtained with SAS functions, because the effect of ET is implicitly taken into account in the Master Equation (Botter et al. 2010), and because the effect of mixing in the unsaturated zone is included in the definition of the streamflow TTD. We will add this information in section 2.3.

FG: My point is that this approach, if actually used, will provide a set of model parameters adequate to describe the isotopic signal of the samples as they are in the record, but not to simulate the isotopic mass balances, i.e. the main rationale of the model. If the isotopic mass balances are sought, it is necessary to weight the isotopic signal of every sample with the associated flow (time span X discharge). Furthermore, looking to Figure 4, it seems that the most highly scattered $^{2}H$ samples were taken during low flows, so it could happen that the, really low, efficiency of the model would improve by flow-weighting the stream samples.

Authors: The isotopic mass balance takes the following form (Rodriguez and Klaus, 2019):

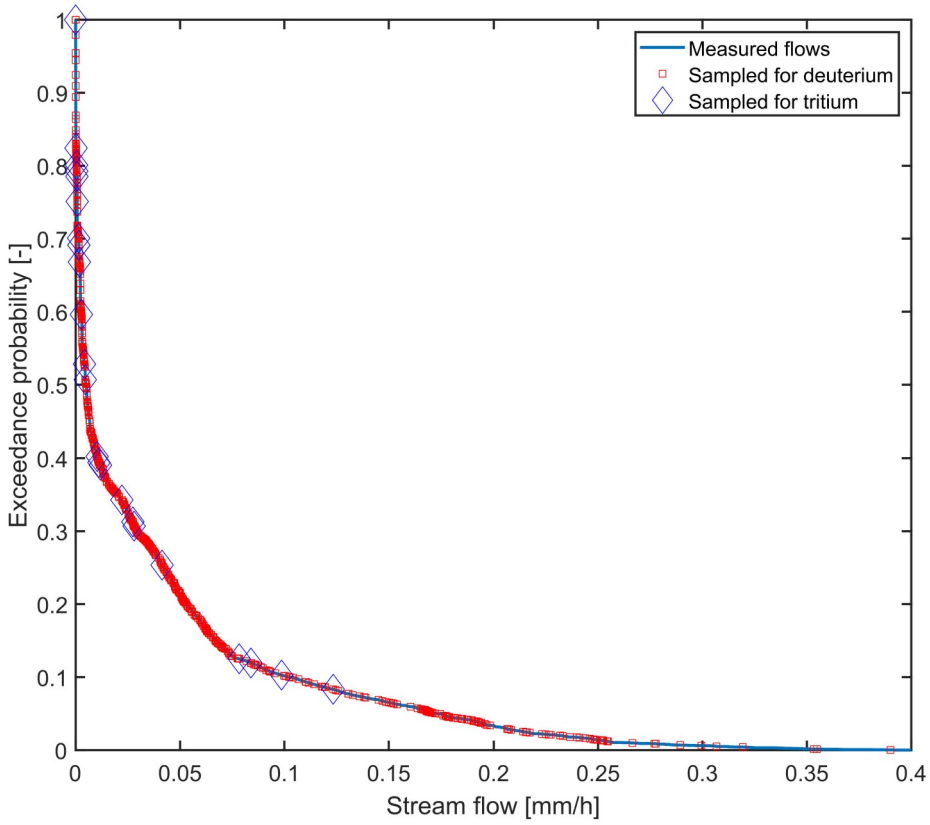
$$dM/dt = J\,C_P - Q\,C_Q - ET\,C_{ET}$$

With our model described in section 2.4, we are able to numerically calculate all the terms of the right hand side of the equation, hence the term on the left hand side as well. However, the main objective of the model is not to simulate the isotopic mass balance, but only to simulate the isotopic signal in a given outflow, here $C_Q$ for which we have tracer observations. This is sufficient to show that the transport from precipitation to the stream is correctly modelled and that the streamflow travel times are correct. Solving the isotopic mass balance is useful only to know in addition how the isotopic tracer mass in the catchment changes with time. We do not focus on this term because we do not have representative tracer data for the ET flux. This means that we are unable to compare our simulated $C_{ET}$ to any observation. Without appropriate tracer data for ET, both the flux term corresponding to ET (ET times $C_{ET}$) and the “mass change term” (dM/dt) cannot be verified against experimental data, and thus depend on each other. We will emphasize again on this point in section 2.4.

Moreover, we think that focusing on the flow-weighted isotopic signal is problematic for the goals of our study. The flow signal varies considerably more than the isotopic signals. The variations of the product signal (flux times isotope) therefore mostly depend on the flow variations. Although calibrating a model to such flux-weighted signal could improve the performance measures thanks to this, it would also overlook the isotopic variations. Our goal is not only to improve performance measures, but to accurately simulate the variable of interest, here the unweighted tracer signal, that carries most of the information about travel times (while water fluxes in themselves do not). We discussed in our related paper (Rodriguez and Klaus, 2019) the relevance of these unusually low values of NSE for deuterium and the issues with this objective function in our particular case. To avoid overlap with this study, we will refer the reader to this paper for more details on the choice of objective function.

FG: Another associated question is the representativeness of the stream samples of the diverse flow ranges in the catchment. In the discussion, the authors sensibly wonder if “tritium... may still be biased towards hydrological recessions” and “how many measurements are enough and when to sample isotopes for maximum information gain on water ages”. If the stream samples must represent the mass flow of water and tracers and a detailed flow record is available, it is possible to compare the distribution functions of both flow records (only measured versus measured and sampled) for assessing the degree of representativeness of the sampling designs. This kind of analysis should be customary in all catchment environmental tracing studies, particularly for small catchments where the flow duration curve is usually highly skewed.

Authors: This is a good remark. We will include the following figure showing the distribution of isotopic samples along a flow exceedance probability curve in section 2.2. Our sampling scheme covered flows with exceedance probabilities going down to 2e-4 for deuterium and down to 0.09 for tritium. This makes the sampling scheme rather representative of all flow conditions. Note however that we did not select the 24 tritium samples based on this curve, but based on the streamflow time series. We selected samples at different flow conditions representing interesting hydrological events (e.g. beginning of a wet period after a long dry period, small but flashy streamflow responses), based on our experimental knowledge of this catchment and on our previous experience with deuterium data (Rodriguez and Klaus, 2019). We will add this detail to section 2.2. Comparing the histograms of measured vs sampled flow records is not very meaningful for tritium because there are only 24 measured values (against more than 4000 for flow alone).




FG: Lines 12-13: The truncation (null) hypothesis cannot be rejected from the work results.

Authors: See the answer to the general comments. This is correct, the statistical hypothesis cannot be rejected. However, one has to keep in mind that the point of our work was not to conclude on the statistical significance of the age differences we found. Our point was rather to show that the TTDs are not so drastically different, which acts as a counterexample to the conjecture of Stewart et al. (2010) that seasonal tracers systematically truncate the long tails of the TTDs. Moreover, the current lack of high-resolution

tritium data means that it cannot be safely concluded from the simple statistical analysis of these results that the TTDs are truly different. We will revise the manuscript to make this aspect clearer.

FG: Line 122: “phyllade” is a French geological term. The closest English term, as far as I know, is “phyllite”

Authors: We thank FG for pointing this out. We will change it as suggested.

FG: Line 330: ... This is not the case for d3H...

Authors: We suppose FG thought that we meant “$^{3}H$” and not what is currently written, “$\delta^{2}H$”. We really meant $\delta^{2}H$. We will rewrite this to avoid any confusion.

FG: The model calibration method that consists of using a range of parameter sets instead of an ‘optimal’ parameter set was developed by Beven & Binley (1992). I suggest to cite this work also because it, as far as I know, was the first using the Shannon entropy for analysing the value of additional data in the calibration of a model.

Authors: We thank FG for the relevant suggestion, and we will add this reference.

Authors: We will also modify figure 5 to better represent the standard error (1 standard deviation of measurements) above and below the points. The current figure shows only half a standard deviation above and below the points.

Kirchner, J. W. (2003). A double paradox in catchment hydrology and geochemistry. Hydrological Processes, 17, 871-874. https://doi.org/10.1002/hyp.5108

Pfister, L., & Kirchner, J. W. (2017). Debates—Hypothesis testing in hydrology: Theory and practice. Water Resources Research, 53, 1792–1798. https://doi.org/10.1002/2016WR020116