Dear referees, dear editor,

Thank you for this opportunity to revise and improve our manuscript. Thank you for the many useful 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 tracking the changes in the text, underscored in blue. Please note that some typesetting issues can be found in this manuscript. These are only due to the limitations of the script "latexdiff" used to track changes in the text between LaTeX files. These typesetting problems are not found in the revised manuscript.

Regards

Nicolas Rodriguez, on behalf of all authors

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).

See title, lines 7, 14, 111, 465-470.

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.

See lines 13-17, 424-431, 454-470, 748-752.

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.

See Appendix B, Table B1 (at the very end), lines 14-15, 424-431, 454-455, 748-750.

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!

See lines 455-458.

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.

See lines 489-495.

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.

See lines 203-206.

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.

See lines 78-82.

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):

$$dM/dt = JC_P - QC_Q - ETC_{ET}$$

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.

See lines 716-721.

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.

See lines 246-249, 677-679, and appendix A2.

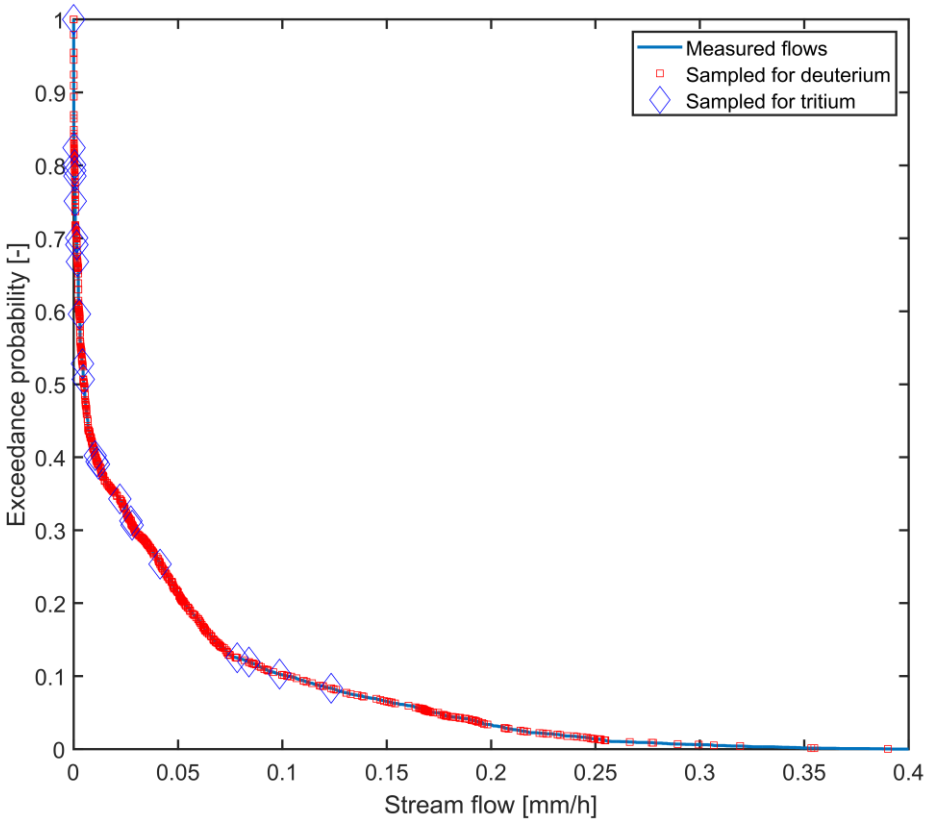
Moreover, we think that focusing on the flow-weighted isotopic signal is problematic for the goals of our study. The flow signal varies considerably more than the isotopic signals. The variations of the product signal (flux times isotope) therefore mostly depend on the flow variations. Although calibrating 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.

See lines 390, 630-639.

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).

See lines 170-171, 178-183.



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.

See Appendix B, Table B1 (at the very end), lines 14-15, 424-431, 454-455, 748-750.
See lines 455-458.

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.

See line 136.

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, “δ^2H”. We really meant δ^2H. We will rewrite this to avoid any confusion.

See lines 394-395.

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.

See lines 331-333.

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

Authors: Francesc Gallart (FG) reacted to our reply to his referee comments. He sent us some additional thoughts of improvement by email because the manuscript is currently in the "author comments only" phase. We obtained FG's approval to reply in HESSD by reproducing his comments below.

FG: The discussion through HESSd is over, but I wanted to shortly react to your kind response to my referee comment.

- Sampling.
*Thanks for the flow duration curve. It confirms my worries: tritium sampling is partial and deuterium sampling is biased. As usually, you plotted the curve of discharges respect to time although this not the relevant variable with such a skewed distribution, but the relative cumulated flow. I made some gross calculations of the area (flow*relative time) and found that the tritium sample for the highest discharge (exceeded in time only 0.09) was exceeded in flow about 0.45: Your tritium sampling discarded about 45% of the highest flows, so it is not only biased but really partial.*
The figures are fortunately much better for deuterium, but my gross estimate is that the 40 samples taken for the highest flows represent about 23% of the cumulated highest flows: 4% of samples represent 23% of highest flows: your sampling is much biased.
The implications are that: (i) my objections on the way you compare your deuterium results with tritium ones are highlighted (ii) your sampling is not representative of the stream flows. For deuterium you must flow-weight your samples in order to compensate the biased sampling.

- sample weighting.
I am very surprised by your answer. After Botter et al. (2011) "the residence time distribution describes the pdf of the ages of all ***water particles stored*** *inside a catchment/hillslope transport volume at a given time, and plays a key role in describing* ***the catchment storage of water and pollutants****". A water particle is a mass element. Concentrations cannot be stored. Your SAS functions select the ages in the catchment store to be output by runoff or ET and these (relative) mass outputs are updated in the catchment store. Your goal may not be the water mass balance, but you need the tracer mass balance to simulate the outputs of the system, and this cannot be made without mass weighting isotope inputs and outputs. Indeed, flow varies much more than concentrations, but this is the real hydrological world. One hour of high discharge may transport more water and tracer mass than several weeks of low flows.*
You may argue that your model should predict any unweighted stream water isotopic sampling. This might be true for a 'perfect' model if the precipitation isotopy was mass-weighted, but not for a model that has so much unexplained variance. For a non-perfect model, the result of the NSE depends on the samples you use, so you can try how diverse sets of samples give different NSE results and different behavioural parameters, but, frankly speaking, I would prefer to use precipitation and flow-weighted concentrations for a sound simulation.

I hope that these thougths will be useful for a better revision of your nice paper.
All the best
Francesc

Authors: We sincerely thank FG for the additional remarks. Regarding the sampling, we found similar numbers. The highest flows that were not sampled for tritium represent about 50% of the water that left the catchment via streamflow over 2015-2017. For deuterium, the highest flows associated with 40 samples (about 4% of the samples) represent about 20% of the water leaving via streamflow over 2015-2017.

In brief, this is what we will emphasize on in the revised manuscript (we nevertheless wrote more details below):

See lines 708-724.

a) The employed sampling technique is not designed to measure the tracer masses, but their concentrations. Only nearly-continuous sampling or time-integrated samples can measure the tracer masses.

b) The limited number of ^{3}H samples compared to $\delta^{2}H$ samples does not allow a comparison of the exported tracer masses for each isotope, but it still allows a comparison of the stream water ages for each isotope.
c) Flow-weighting the stream samples will not compensate for the potential lack of samples during high and/or low flows.
d) Simulating only the tracer concentrations is sufficient to validate the TTDs.
e) Time-varying TTDs already implicitly account for the catchment-scale mass balance, no additional flow-weighting of the input and/or output tracer signal is necessary.

Here are additional details on the reasoning:

a) Our sampling is based mostly on fixed time intervals generally larger than a few hours. Thus, it should not be a surprise that the water mass is not proportionally represented in the sampling scheme. For this, an adaptive sampling frequency based on accumulated flows needs to be implemented (e.g., one sample every dozen m^3). In our case this would nevertheless lead to a much larger number of samples, exceeding the available field and lab resources. With more frequent samples during higher flows and less frequent samples during low flows, the mass of water flowing out of the catchment would of course be better represented. However, this would imply that the samples are not evenly distributed in time (hence along the FDC), which could also be criticized for being unrepresentative of all hydrological times of the year (i.e., over-representation of wet and cold conditions). It appears that choosing a type of sampling scheme (i.e. flow-proportional vs. fixed time intervals) will not allow to have the samples evenly distributed in time AND representative of all the water mass leaving via streamflow, unless streamflow is constant. Only nearly-continuous in-situ measurements that are currently available for stable isotopes can avoid these limitations (e.g., von Freyberg et al., 2017). Alternatively, a time-integrative sampling technique (that implicitly uses flux-weigthing) should be used for streamflow if the goal of the work is to simulate the exported tracer mass and compare it to the observations (this is not our goal). Note that the precipitation tracer measurements are time-integrative by design.

b) Even with the time-based sampling scheme and the limited number of tritium samples, the good agreement between TTDs constrained by deuterium and the TTDs constrained by tritium shows that the large water mass not sampled for tritium is not creating a strong bias towards young or old water compared to deuterium. This was different in previous tritium studies that focused on baseflow, where perhaps 90% of the water mass leaving the catchment via streamflow was not sampled for tritium and contained all the young water fractions. Our tritium data set most likely contains a rather representative selection of young and old stream water, even if not all water mass was not sampled.

c) Our goal is to accurately estimate the streamflow TTD at all times of the year. Accurately simulating the tracer mass flux in streamflow will not help reach this goal better than accurately simulating the tracer concentrations only. This is for the reasons outlined below. To put it more quantitatively, our model errors take the form:

$$\varepsilon_C(t_{obs}) = C_{modelled}(t_{obs}) - C_{observed}(t_{obs})$$

where only the times corresponding to stream samples t_{obs} are used (this avoids interpolating $C_{observed}$). Minimizing ε_C at all times when we have observations allows us to constrain the TTDs to the most accurate estimate given our current tracer data set. If we were to flow-weight the tracer samples, the model errors would take the form:

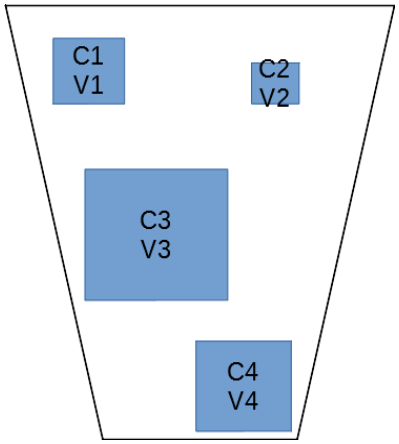
$$\varepsilon_{QC}(t_{obs}) = Q_{observed}(t_{obs})C_{modelled}(t_{obs}) - Q_{observed}(t_{obs})C_{observed}(t_{obs})$$

because measured streamflow is used as an input in our model (there is no $Q_{modelled}$). Note again that only the times t_{obs} when we have measurements $C_{observed}$ can be used. This is why flow-weighting the stream samples will not compensate for the lack of higher resolution tracer data over 2015-2017. There will still be some missing knowledge about the true variability of the tracer concentrations and the true tracer mass flux in streamflow. Furthermore:

$$\varepsilon_{QC} = Q_{observed}\varepsilon_C$$

This means that minimizing ε_{QC} by adjusting model parameters is similar to minimizing ε_C (because $Q_{observed}$ does not depend on parameters), and it yields the same TTDs. The NSE does not try to minimize each individual error but a squared sum of errors normalized by the observed variance. For ε_{QC} this would give much more weight to periods with high flows, and the TTDs during drier periods would not be accurate anymore. Now, the variance of QC is much bigger than that of C, which can also "artificially" allow higher NSE values. Therefore, with flow-weighting, the "performance" of the model would improve, but this would lead to less reliable constraints on the TTDs because NSE>x for ε_{QC} is clearly less strict than NSE>x for ε_C. The intuitive interpretation is that flows Q do not contain considerable information about the time scales of transport, only tracer concentrations do. Including the flows in the calibration can only reduce the information learned about stream water ages.

d) & e) Moreover, the convolution integral implicitly includes flow-weighting. We agree that "concentrations cannot be stored". Our approach does not store only concentrations, but also the associated particle volumes and thus mass. As written in section 2.3, Equation 1 expresses the fact that the stream concentration is the volume-weighted arithmetic mean of the concentrations of the water parcels with different travel times at the outlet. Let's imagine a streamflow grab sample represented below:



Each water particle k (there are n=4 particles represented here) has a given volume V_k and a given concentration C_k. The measured tracer concentration of the sample is:

$$C_{obs} = \frac{\sum_{k=1}^{n} C_k V_k}{\sum_{k=1}^{n} V_k}$$

which can be rewritten:

$$C_{obs} = \sum_{k=1}^{n} C_k \frac{V_k}{\sum_{k=1}^{n} V_k} = \sum_{k=1}^{n} C_k p_k$$

where p_k is the fraction of streamflow volume associated with particle k. Now, if we label each particle k with its age relative to the precipitation input, C_k and V_k simply become the corresponding past (time-varying) precipitation amounts and concentrations, and p_k simply becomes the backward TTD.

Equation 1 in the manuscript is simply the continuous version of the equation above, for n tending to infinity. Therefore, the backward TTD needs no additional flow-weighting with respect to precipitation because it already includes it (the time-varying V_k). Furthermore, if an unsteady TTD is used, the stream flow variations are already included in its definition (by the time-varying denominator$\sum_{k=1}^{n} V_k(t) = Q(t)\Delta t$), and no flow-weighting of C_{obs} is needed to correctly deduce the TTD from the convolution integral.

From this equation we now easily guess the data requirements of the approach, sufficient to estimate the TTDs and to respect the mass balance. In terms of tracer: a continuous tracer concentration input signal, and a time series of tracer concentrations in the outflow. The finer the resolution of the time series of the output concentration, the less uncertainty there should be about the TTD, because fewer weighted combinations of all the C_k will closely match all the C_{obs} simultaneously. In terms of hydrometric measurements: precipitation rates, and stream flows. In addition, to calculate the TTD from the Master Equation, storage needs to be deduced from the catchment-scale water balance equation. This requires actual ET to be calculated as well.

von Freyberg, J., Studer, B., and Kirchner, J. W.: A lab in the field: high-frequency analysis of water quality and stable isotopes in stream water and precipitation, Hydrol. Earth Syst. Sci., 21, 1721–1739, https://doi.org/10.5194/hess-21-1721-2017, 2017.

R2: The manuscript tested the hypothesis that ^{3}H tracer provides information over longer transit times than ^{2}H. The authors calibrated the StorAge Selection (SAS) function model for each tracer and examined information gain using the posterior distributions of the model parameters. They rejected the hypothesis based on their results. Nevertheless, they concluded that ^{3}H tracer is more informative and cost-efficient compared to ^{2}H.

The topic is timely and very interesting. However, the manuscript needs substantial revision. First, I do not think that the results presented in this study support most of their conclusions. Their SAS function-based model performed poorly even with 12 parameters, and it is not clear how much we can learn from the poorly performing and not well-constrained model. Second, I have several issues with their analysis and the hypothesis test. These points are described in more detail in what follows.

Authors: We thank the reviewer (R2) for the detailed assessment of the work and for suggestions of improvement. Regarding the hypothesis testing, we were not clear in our writing. We did not intend to test the statistical significance of the water age differences derived from different tracers, but rather wanted to prove that the age differences are much smaller than previously shown (Stewart et al., 2010) and assumed in most following tracer studies. As a consequence of the comments from Francesc Gallart (FG) and R2, also written in more detail in our reply to FG, we will now also include a statistical test in the revised manuscript.

We note R2's concerns about our model and data. We detailed below why we think that we can still derive robust conclusions from the modelling exercise. We will modify the manuscript to clarify this and to address R2's comments.

R2: The model has an unusually large number of parameters (12 parameters; e.g., Line 249) compared to the previous SAS function-based modeling studies. I believe that the authors illustrated the need for more parameters well in their previous study, which is now published in WRR. However, the model does not perform well even with the 12 parameters (with the maximum NSE 0.24 for ^{2}H), and I am not sure what we can learn from the poorly-performed model. The large number of parameters also causes several issues described below.

Authors: We understand R2's concern that the model does not perform sufficiently well despite the large number of parameters it has. We will rephrase parts of the discussion to stress that the model is of course not in perfect agreement with the observations, and that a better model may change the interpretation of the results to some extent.We already proposed some suggestions of improvement of the model for future studies (section 4.4 and our answer to a comment further below).

See lines 389-394, 625-639.

We agree with R2 that the NSE cannot be considered high, but we disagree with R2's interpretation that the model is performing poorly. In our previous study (Rodriguez and Klaus, 2019), we detailed why such a complex model structure is adequate for this catchment, even if the NSE appears unusually low. We also emphasized on the fact that 12 parameters is a small number to constrain the vast array of time-varying processes leading to the selection of particular water ages by Q and ET from anywhere in catchment storage (represented here in the Master Equation by ~10^{5} "age control volumes" and their associated age fluxes). We previously detailed the limitations of the NSE for evaluating model performance with such complex tracer time series (see also 4.4, the NSE assumes normally distributed, uncorrelated, and homoscedastic errors). Other performance measures have been proposed (e.g., Schoups and Vrugt, 2010; Ehret and Zehe, 2011), but they either require more parameters, or they are not designed for tracer time series but only for hydrographs.

See lines 267-271, 696-701.

Furthermore, the evaluation of model performance usually involves expert knowledge (Gharari et al., 2015; Hrachowitz et al., 2014) that cannot be expressed via the traditionally used objective functions (Seibert and McDonnell, 2002). The Weierbach $\delta^{2}H$ time series has unusually damped seasonal dynamics, while at the same time unusually strong flashy events occur. A close look at the behavioral simulations (see figure 4) reveals that some runs were actually able to match the flashy $\delta^{2}H$ dynamics quite well. A zoom on figure 4 allows to see the short-term simulation capabilities of the model (the very thin peaks of the simulation envelopes). We will add an inset with a zoom on particular peak in figure 4. We will add figures (see a few examples below) in the supplement showing more details about the behavioral simulations. In these figures, it is remarkable that only

several dozen data points among the more than 1000 were not captured by behavioral simulations in deuterium. These points are almost all during summer 2016 and summer 2017 (drier periods). The other interesting aspect is that behavioral simulations in tritium were able to match many of these extreme values. We believe that this is because the behavioral simulations in tritium were not penalized by the limitations imposed by the NSE, and were thus allowed to have more extreme variations.

See Figure 5, lines 395-398, and supplementary material.

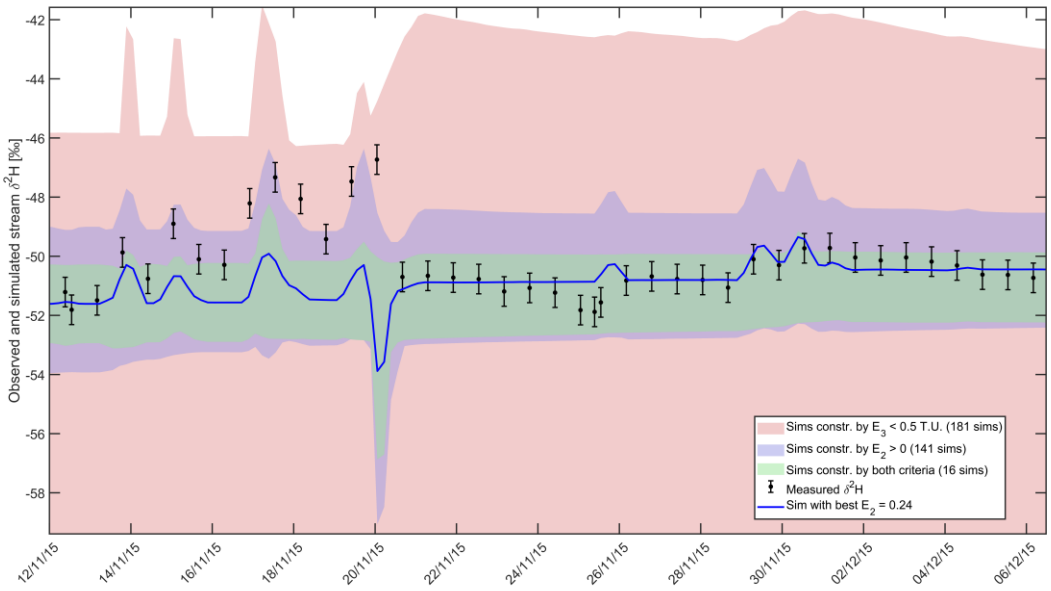


Figure: δ^2H simulations in Nov-Dec 2015

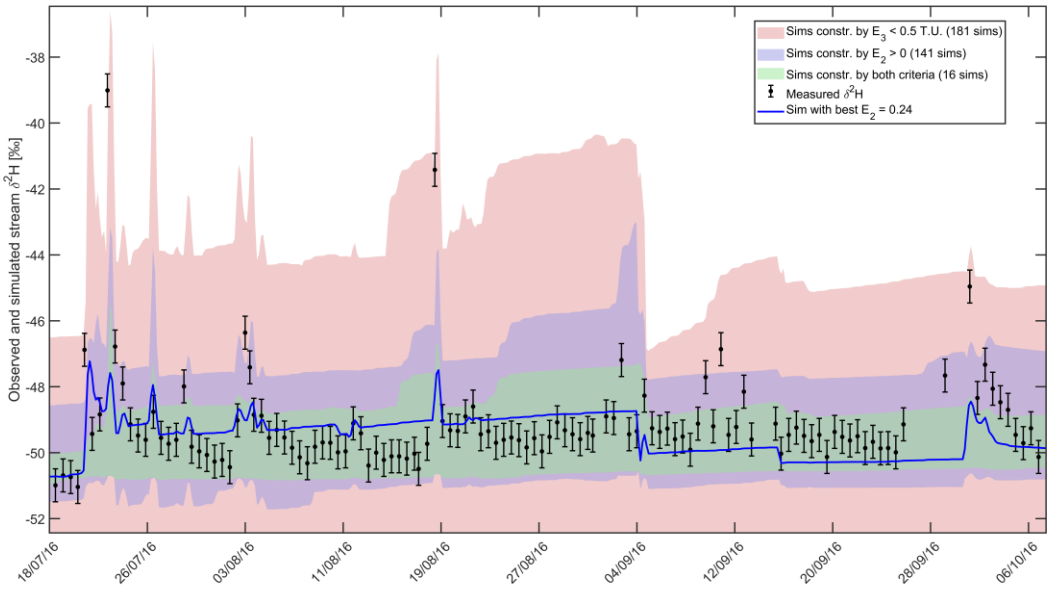


Figure: δ^2H simulations in Jul-Oct 2016

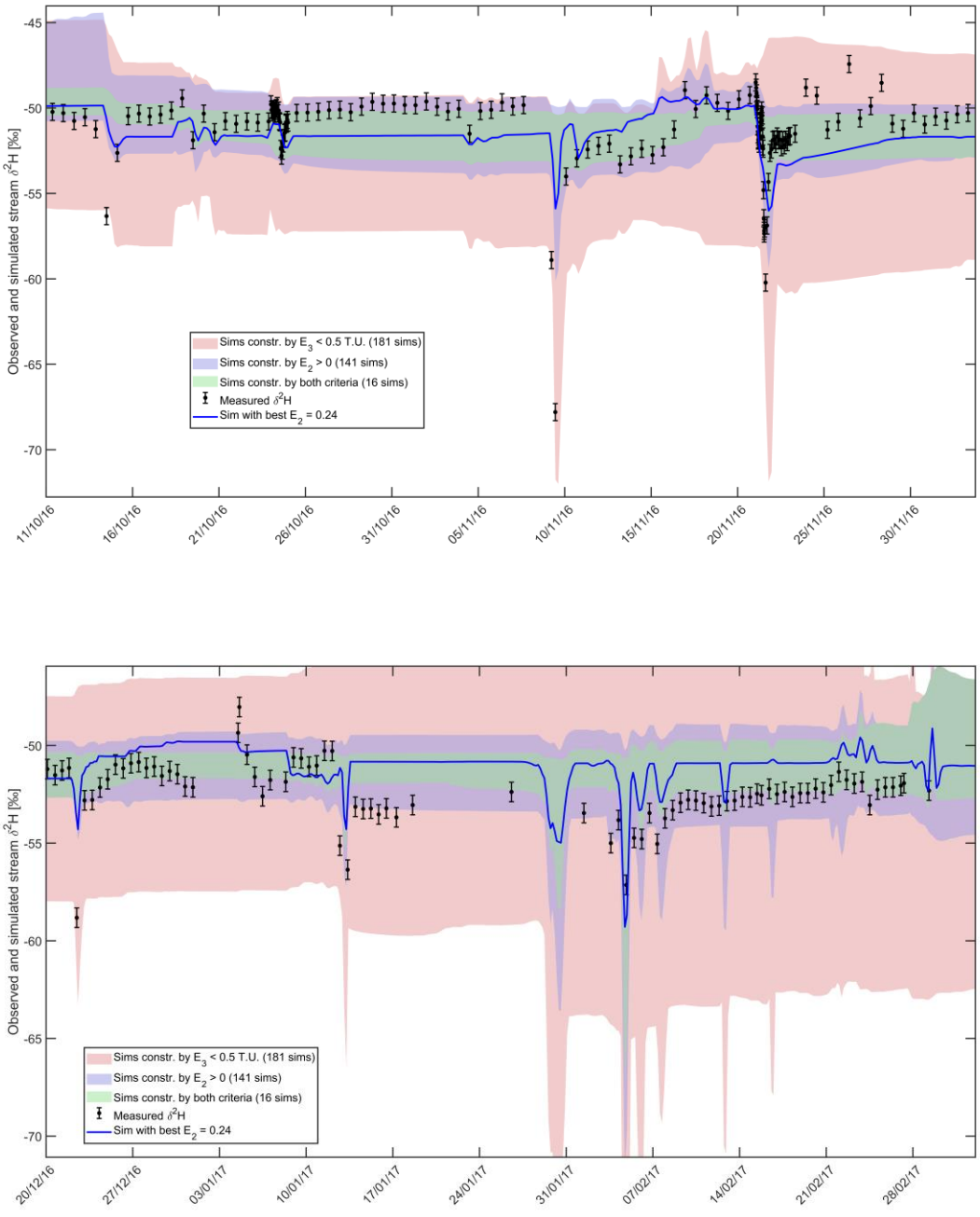


Figure: δ^2H simulations in winter 2016

Although higher NSE values were reported in the past for other δ^2H time series simulated with transient TTDs (e.g. NSE > 0.5; Benettin et al., 2017; Harman, 2015; van der Velde et al., 2015), we disagree to state that our model performs poorly simply because the NSE values are not as high. The NSE of the behavioral simulations is not closer to 1 partly because of the underlying assumptions about model residuals in the NSE (Rodriguez and Klaus, 2019). Care should be taken in interpreting the NSE values. The NSE does not allow a reliable performance comparison between different studies and it is not an absolute measure of model performance, because it implicitly uses the mean observed value as a benchmark model. This benchmark model is not always the best choice, as stressed in several studies (Seibert, 2001; Schaeffli and Gupta, 2007; Criss and Winston, 2008). In our particular case, the mean observed value is particularly penalizing because the δ^2H time series has many more points corresponding to very damped seasonal fluctuations than points corresponding to the large flashy fluctuations. Within tracer hydrology and modelling there is an urgent need for better ways of summarizing model efficiency. Yet, this is beyond the scope of this study, especially because it focuses the calibration task while our goal is to focus on what can be learned from the isotopic data set in terms of water ages. We will add these points to section 4.4 in the discussion.

See lines 631-639.

R2: Also, the dataset is very limited, and it is not clear if the limited number of samples and the limited sampling period support their conclusions. First, it is not clear if the ^3H dataset is enough. The number of samples is too limited to constraint 12 parameters.

Authors: The ^3H data set has, with the study of Visser et al. (2019), one of the highest number of stream samples analyzed for ^3H and used for travel time analysis. We understand that this may appear as a small number to constrain 12 parameters in the more general context of environmental modelling studies, but this is very common in travel time studies involving tritium. Many previous studies had about as many parameters as tritium samples or a just a few samples per parameter (Maloszewski and Zuber, 1993; Uhlenbrook et al., 2002; Stewart et al., 2007; Stewart and Thomas, 2008; Stewart and Fahey, 2010; Morgenstern et al., 2010; Cartwright and Morgenstern, 2016a, 2016b; Duvert et al., 2016; Gallart et al., 2016; Gusyev et al., 2016; Gabrielli et al., 2018). We will cite some of these studies and mention this point in sections 2.2, 2.6, and 4.4. Future studies may present a higher number of tritium samples if the analyses become more affordable.

See lines 182-183.

R2: I can easily guess that the parameters are not well-constrained. Thus, it is obscure how much information we can extract from the time series, the posterior distributions of those parameters, the TTDs, and the SAS functions, which were used to test the hypothesis and examine if those tracers contain non-redundant information to each other.

Authors: We will include the parameter posterior distributions (see below) in a supplementary file. Most distributions are not flat (i.e. not uniform), indicating that the parameters are identifiable to some extent. We also note that all the parameters directly related to the shape of the SAS functions hence the TTDs (μ_2, θ_2, μ_3, θ_3, μ_{ET}, θ_{ET}) are visually clearly not uniform. We initially used Shannon's entropy H and the Kullback-Leibler Divergence DKL concepts for parameter identifiability instead of these figures to have a more objective and more quantifiable uncertainty assessment. We note that "how much information we can extract from time series, the posterior distributions of those parameters..." is exactly quantified via equations 8 and 9. We will explain these concepts in more detail in section 2.7 and add a line in table 2 corresponding to the DKL between prior and posterior distributions for each parameter.

See supplementary information, lines 375-376, 696-697, 340-341, 344-346, 349-353, table 2.

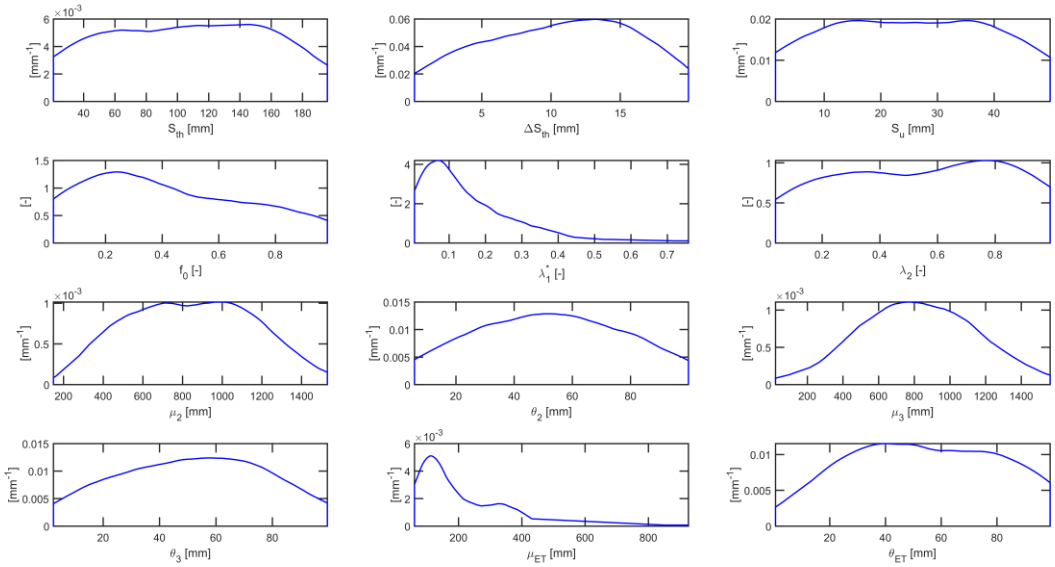


Figure: Posterior distributions constrained by deuterium

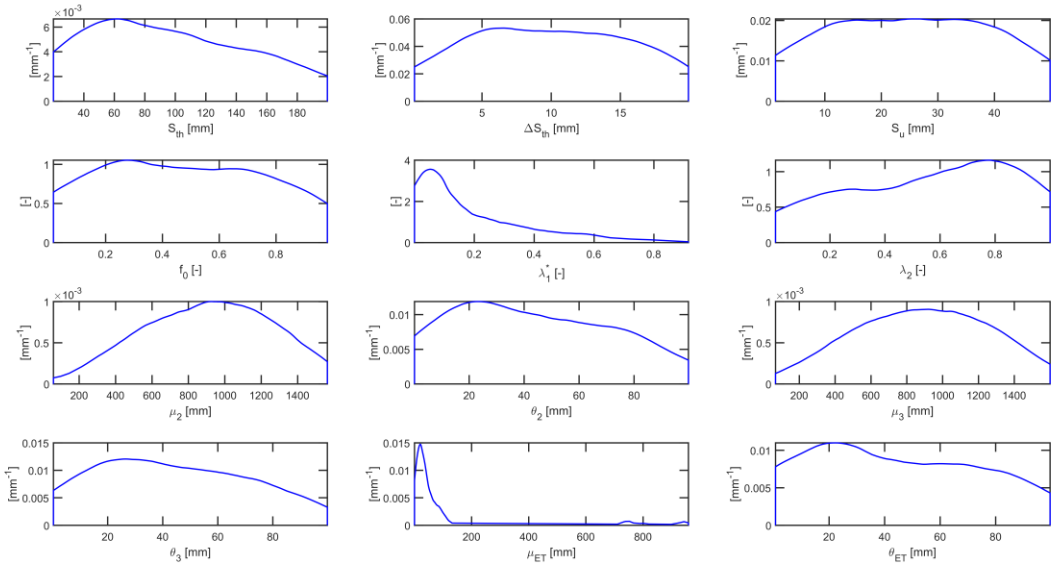


Figure: Posterior distributions constrained by tritium

R2: For example, the authors stated that “stable and radioactive isotopes have information in common about young water” in Lines 472-475. However, the argument cannot be supported by those 24 samples. Furthermore, how much information we can extract from the 2-years of ^{2}H data set? Can we talk about transit time longer than 2 years (at the maximum) based on the model results?

Authors: We are not sure what is meant exactly by “the argument cannot be supported by those 24 samples” and thus how to cope with this comment. As indicated in the following sentences (lines 472-475) we believe that the high variability of stream tritium concentrations, that follow the variations of precipitation concentrations, indicates that it is very likely the effect of young water contributions. This was unobserved before due to a focus on baseflow sampling, except for rare studies showing high tritium variability during short-term hydrological events (Hubert et al., 1969; Crouzet et al., 1970; Dinçer et al., 1970). Tritium has therefore been generally considered to be informative only about old water (we will emphasize on this detail in the corresponding paragraph). However, tritium can be used and has been used to detect young water contributions, for example in the first studies using hydrograph separation (Klaus and McDonnell, 2013).

See lines 21, 397-398, 587-594, 648-651

Moreover, as it can be seen in table 3, we have travel times above 2 years (e.g. mean > 2). We have travel times up to about 100 years (see figure 6). This is possible due to the 100 year spin-up period (1915-2015) that we systematically used before evaluating each simulation over 2015-2017. We will add a sentence to clarify this in section 2.5.

See lines 283-284.

R2: Second, I think that their Latin hypercube sampling (Line 262) suffers the curse of dimensionality. They sampled 12,096 parameter sets from the 12-dimensional parameter space. It can be easily guessed that those samples are very sparsely distributed in the 12-dimensional parameter space (i.e., $12^4 > 12{,}096$), and the sparse sampling can potentially limit their ability to construct well-constrained posterior distributions of those parameters.

Authors: We understand that 12,096 parameter samples for a 12 dimensional space can be less than one may hope for. We also understand that it would be ideal if we had several more orders of magnitudes in the number of samples. However, we are currently limited by computational time (more than 1 hour) to run the model with each parameter sample, despite the use of a highly parallelized code with a high performance computer. This computational time is so large because of the need to spin-up the model for 100 years (see above). Without this spin-up, a numerical truncation of the TTDs will occur.

As suggested by R2, the parameter sets are thus likely to be sparsely distributed. The LHS technique was thus employed to make sure that the samples are distributed as evenly as possible in this high-dimensional space (each parameter range is divided in 12,096 equal intervals that each contain at least one point). This technique has the advantages of a stratified sampling technique, while keeping the simplicity and objectivity of a pure random sampling technique (Helton and Davis, 2003). We will emphasize on this aspect in section 2.6.

See lines 306-309.

Finally, we want to point out that the posterior distributions from our approach using a simple Monte Carlo technique and a Latin Hypercube Sampling scheme are naturally more likely to appear less constrained than when using Markov-chain-based algorithms such as DREAM (Vrugt, 2016) or PEST (Doherty and Johnston, 2003). This is a visual effect. Our approach is similar to a global optimizer that tries to find the absolute optimum point by exploring the widest space as evenly as possible (especially when using LHS), say [0, 1] to make it simple. In contrast, Markov Chain Monte Carlo algorithms tend to quickly converge on “interesting areas” (say [0.05, 0.1]) and tend to stay confined there on several local optima. This means that the resulting posteriors appear naturally more constrained with MCMC algorithms because they only show values in the explored region of interest, say [0.05, 0.1], out of the total initial space ([0, 1]). We could not use MCMC algorithms for numerical reasons. For example, MCMC algorithms are poorly suited to systematically enforce parameter constraints (such as the sum of SAS function weights λ being 1).

See lines 696-706.

R2: Lastly, the poor performance of the model leads me to think that maybe their model structure is not adequate, and any discussion based on the model results should be conducted more carefully. It is clear that the model fails to reproduce short time-scale dynamics. Figure 4 shows that their ^{2}H-based model cannot capture the observed large fluctuation. It seems that the large fluctuation is, in part, due to the high correlation between $C_{p,2}$ and $C_{Q,2}$ especially when the system is dry, and It implies that short time-scale dynamics are not captured by the model (as they mentioned in Lines 512-513). The fluctuation seems much more pronounced in the ^{2}H time series. Thus, if we have a better model that captures the short time-scale dynamics, it may contradict the authors’ argument in Line 472: “stable and radioactive isotopes have information in common about young water.”

Authors: Please see our related answer about model performance above. We will stress in the discussion that a better model may change the interpretation of the results to some extent. We don’t think that a model performing better would change the conclusions of our study. Furthermore, in our model, the flashy events (that we assume to be young water contributions) are conceptualized in a novel way via λ_1 and its

parameterization depending both on storage and a proxy of storage variations. In the discussion of the original manuscript, we proposed suggestions for improvement in future studies regarding this part of the model (Lines 518-538). Yet, we disagree with R2. Behavioral simulations were able to match the flashy dynamics of δ^2H to a degree. We will supply figures as a supplement (see above) that will allow the readers to visually identify this aspect better (see one example below).

See supplementary material, and lines 390-394.

As R2 points out, these flashy events occur mostly during drier periods, but not only. During winter 2016, flashy variations in δ^2H can also be observed (figure 4 of the original manuscript). The flashy variations tend to follow the variations of precipitation δ^2H, and suggest the influence of young water contributions to the stream. However there is not a perfect correlation between $C_{P,2}$ and $C_{Q,2}$, even during dry periods (e.g. for Q < 0.02 mm/h) when the relationship seems visually clearer. This is most likely because of a strong annual groundwater contribution, conceptualized with the two gamma components in the SAS function (Rodriguez and Klaus, 2019). $C_{P,2}$ can thus explain only about 45% of the variations of $C_{Q,2}$ during dry periods. We will provide a figure showing this in the supplement of a revised manuscript, and include these comments in the discussion, section 4.4.

See supplementary material, and lines 655-657.

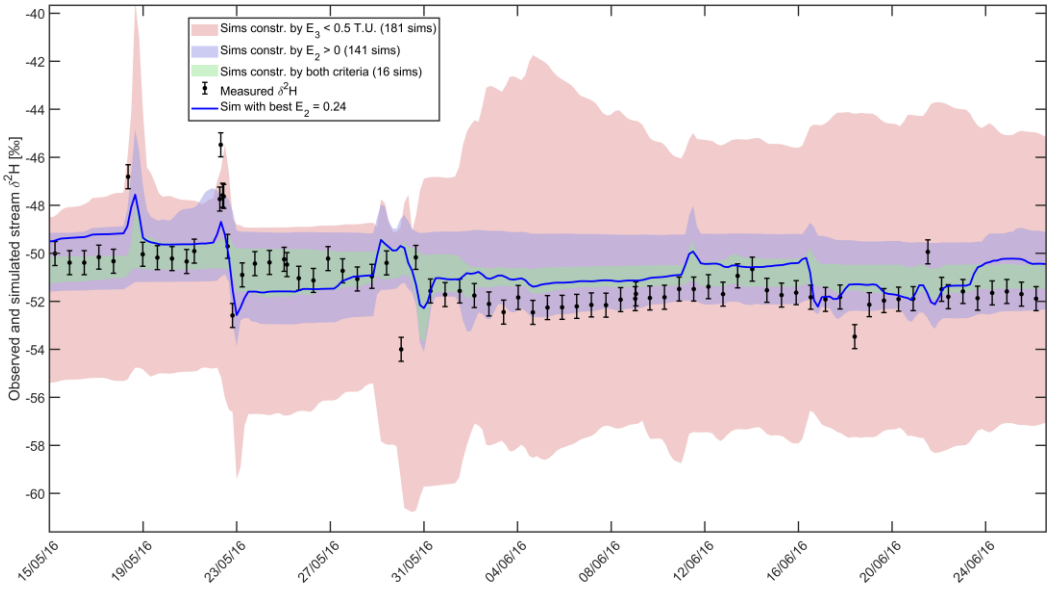


Figure: Simulations of δ^2H in May-June 2016

The flashy variations appear more pronounced for δ^2H, because there are many more samples compared to ^3H, and because the unit scaling is different. We think that these flashy events would be similar for tritium if we had more than 1000 samples. One of such flashy events was already captured with the 24 samples and can be observed in November 2016 for ^3H. Re-scaling the time series to be able to include the precipitation signal (as this was done for tritium in figure 5) makes the flashy events appear much less pronounced. For instance, compare the inset of figure 2 with figure 4 for δ^2H. The inset in figure 2 makes the tritium variations appear stronger than deuterium variations. Finally, as we detailed in the discussion (lines 515-517), a model passing through all observation points would still not allow to draw firm conclusions of the "own potentials" of each tracer in terms of water ages, because the number of samples for each tracer is not comparable. We think that high-frequency tritium observations would unambiguously show that young water contributions are as visible in tritium time series than in the δ^2H plot (e.g. Crouzet et al., 1970). The point of our work is to argue that there is only one streamflow TTD, and that an observed age difference between the tracers can be due to sampling limitations in one or the other tracer or to erroneous assumptions (e.g. steady-state). We will insist further on this point in section 4.4 of the discussion.

See lines 648-651.

R2: The use of the Kullback-Leibler Divergence DKL in the hypothesis test seems inappropriate. Throughout the manuscript, the authors stated that using both tracers together is valuable since DKL > 0 (e.g., in Lines 435-436 and Lines 468-470). However, the criterion DKL > 0 cannot determine whether the criterion is met because multiple tracers are used or because there is just any additional information. For example, DKL between the model constrained by, let's say, 100 ^{2}H data and the model constrained by the rest of the ^{2}H data will be greater than zero.

Authors: This is an interesting point. However, it is not only because DKL > 0 that we concluded that using both tracers together is valuable. As stated lines 436-437, using both tracers together reduced the entropy of the posterior distributions compared to prior distributions. Combining both tracers also allowed narrower groups of TTD curves in figure 6 and 7, and yielded lower standard deviations of the age and storage measures in table 3 and 4 despite having fewer samples. Second, DKL is strictly positive if and only if the compared probability distribution functions (pdfs) differ, meaning that they contain different information about the population(s) they describe. It does not matter for DKL whether the pdfs come from sampling different populations (in our case the posteriors constrained either by one tracer or two tracers) or from sampling the same population several times with different methods (e.g. using two distinct objective functions to constrain the parameters using only one tracer). In any case, DKL being strictly positive tells us that the posteriors are not equal, thus we learned something about the parameters and the water ages. The statement "DKL between the model constrained by, let's say, 100 ^{2}H data and the model constrained by the rest of the ^{2}H data will be greater than zero" may unfortunately be wrong. If the additional $\delta^{2}H$ data points do not visibly change the posterior pdfs compared to the initial 100 points, meaning that they do not bring considerably more information about the parameters hence the water ages, DKL can be close or equal to 0. We found DKL values about 10 times smaller than the maximum Shannon entropies corresponding to uniform prior distributions (table 2). This roughly 10% additional knowledge gained by adding one tracer is therefore not negligible. We will add these comments in section 4.3.

See lines 536-540, 583-585.

R2: Moreover, different performance measures were used for their models (Lines 265-270), and it makes the use of DKL even more inappropriate. The authors used the NSE for the ^{2}H-based model and used the MAE for the ^{3}H-based model. Thus, the difference between the posterior distributions estimated by those behavioral models can be, in part, explained by the choice of performance measure. For example, if the authors estimate the posterior distributions using the ^{2}H dataset based on the MAE, the posterior distributions would differ from those estimated based on the NSE. Then, DKL would be greater than zero. Thus, it is not hard to follow their argument that using both tracers together is valuable (e.g., in Lines 331-333, Lines 435-436, Lines 478-470, and Lines 580-581).

Authors: This is also an interesting remark. We therefore conducted additional analyses. Before we answer this comment, we want to mention that these additional analyses helped us realize that we mistakenly multiplied all the values in table 2 by $\log_2(10)$. This means that we will correct all the values shown in table 2 and mentioned in the text by dividing them by $\log_2(10)$. It is important to notice that this changes absolutely nothing to all the reasoning we applied and to what we wrote in the manuscript, since the values are all changed by exactly the same proportionality factor.

Following R2's suggestions, we recalculated table 2, using the criteria MAE < 1.3‰ for $\delta^{2}H$ and MAE < 0.5 T.U. for ^{3}H. We used the threshold 1.3‰ for deuterium to obtain a similar number of behavioral simulations (here, 149) than with NSE > 0 (148 solutions). We obtained similar results than for NSE > 0 and MAE < 0.5 T.U. Only minor differences can be observed for some parameters. We carefully checked and found that all our reasoning and our conclusions based on table 2 remain intact (lines 321-328 and discussion section 4.3). We will nevertheless include these additional results in the supplement. Following R2's comment that DKL would be greater than 0 if we used both MAE and NSE constraints on $\delta^{2}H$, we went further and calculated the DKL between posteriors constrained by NSE > 0 or by MAE < 1.3‰ and posteriors constrained by the combination {NSE > 0 and MAE < 1.3‰}. All the DKL values we found were below 0.02 bits. This information gain is negligible compared to what was learned by adding one tracer after another. It is not a surprise because the NSE and the MAE are both based on minimizing a sum of residuals (squared or not), making them almost equivalent. It would be very different if we used a measure based on residuals and another

based for example on a correlation measure (Legates and McCabe, 1999). Thus, in our case, the use of the DKL clearly shows that the information gain is not due specifically to the choice of distinct objective functions for ^{2}H and ^{3}H, but instead to the additional information contained in the other tracer.

See supplementary information, and lines 580-585.

R2: Furthermore, I disagree with their cost analysis (in Lines 445-451), which led them to conclude that ^{3}H tracer is more cost-effective (e.g., Line 17). As described in Lines 462-463, "The amount of information learned from the isotopic data probably scales nonlinearly and probably reaches a plateau as the number of observation points grows." However, they assumed "linearity" in their cost analysis. Thus, the analysis is not valid.

Authors: We thank R2 for this remark. The reviewer is right, that we would (most likely) not have concluded that tritium is more cost-effective, if we had more samples and if these samples did not bring more information about parameters and water age. The lines above the quoted statement (lines 458-462) and the conclusion (lines 574-575) also say that ^{2}H could have been more cost-effective with a smarter sampling, which could reduce the number of $\delta^{2}H$ samples hence the total analytical price. We will anyway remove the parts of the manuscript that mentioned cost-efficiency, to avoid misinterpretations.

See lines 24, 555-651, 572, 756-757.

Finally, we only hypothesized that "The amount of information learned from the isotopic data probably scales nonlinearly and probably reaches a plateau as the number of observation points grows". We will rewrite this sentence to make this clearer. We do not know if there is linearity or not. The only thing we know with our samples are the two points shown in the figure below (that we will include in the supplement). How information scales with the number of samples could be any of the dashed curves that represent very different scenarios. The other thing we are sure of is that the true curve can never decrease: there is no information lost by adding new samples. In the worst case, nothing is learned, and the information gain is 0. This means that no matter how many tritium samples we add in our case, tritium will always stay more informative in the absolute sense (14.85 > 13.55) than deuterium. We will thus only keep our statement that tritium was overall more informative.

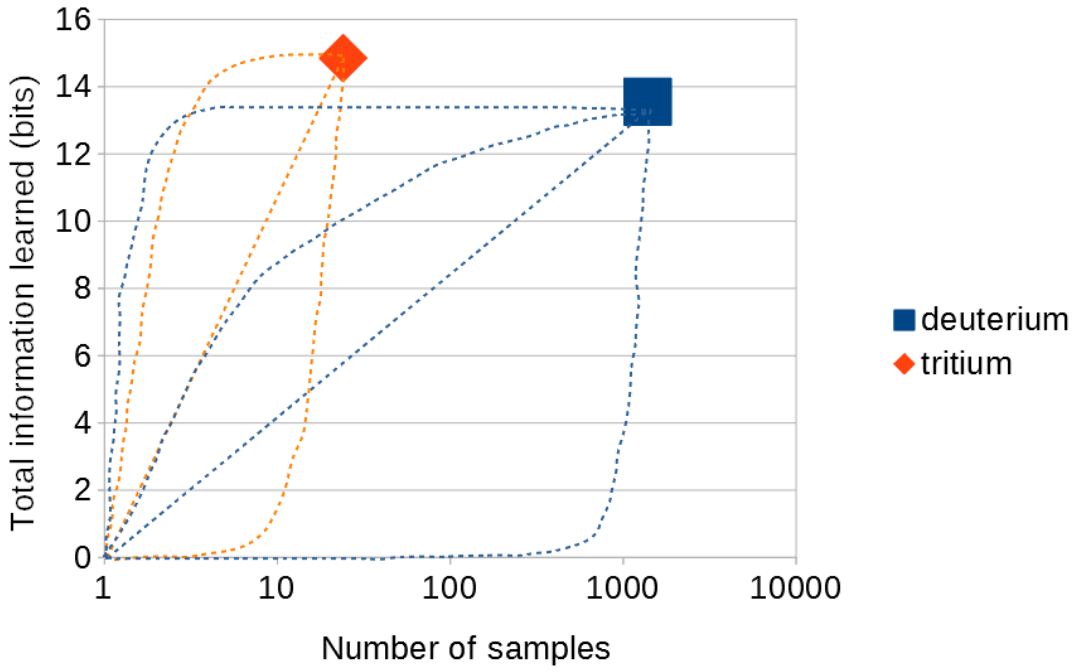


Figure: Information learned about water ages from each tracer (points) and potential relationships between the number of samples and the (necessarily) growing information content (dashed lines).

By simply dividing the total amount of information by the number of samples or by the total analytical price, we only applied some sort of normalization that does not assume linearity or nonlinearity. It would be different

if we used a normalized value (e.g. 0.619 bits per sample for tritium) to extrapolate how much information we could learn in the future by gathering more samples. This would correspond to drawing the unknown curves towards the right-hand side of the points in the figure above. We did not test in what way the amount of information grows with increasing number of samples, as detailed lines 463-467. We will rewrite this part to make sure this is clear, and so that future studies may look into this aspect. As we also detailed in our reply to FG, this test would introduce some subjectivity because not only the number of samples that is used would matter for this analysis, but also the way those samples would be selected among all that we have.

See supplementary information, and lines 572-579.

R2: Lastly, it seems that the ET SAS functions are very important in this study but rarely explained. One of its parameters, μ_{ET} is the most valuable parameter in terms of the information gain in this study (see Table 2). However, no explanation is provided why it is the most valuable and how it affects their interpretation of the results. For example, Figure 5 is one of the most important figures that clearly illustrates the difference between the ^{2}H-based model and the ^{3}H-based model. The simulated ^{3}H concentration using the ^{2}H-based model, in general, is higher than that simulated using the ^{3}H-based model. It means that tracer mass partitioned into discharge is smaller in the ^{3}H-based model during the period. Since there is no explanation on the difference, I had to guess that either more ^{3}H tracer mass is stored in the system in the ^{3}H-based model or more ^{3}H tracer was partitioned into evapotranspiration in the model. Overall, it seems that the partitioning is one of the most important differences between the two models. Thus, the partitioning should be explained in more detail

Authors: We thank R2 for this excellent remark. ET is critical for travel time studies. The water mass balance reads:

$$dS/dt = J - Q - ET$$

In the study we only have one water partitioning condition in the model and that is to decrease ET from PET to 0 when storage S drops below a certain threshold (S_{root}) (see appendix A2). This threshold conceptualizes the strongly increasing capillary forces that prevent water from being taken up by plant roots or directly evaporated at lower soil water contents (Rodriguez and Klaus, 2019). A similar strategy was employed for instance by Fenicia et al. (2016) and Pfister et al. (2017) in the Weierbach and neighboring Luxembourgish catchments. The choice of the SAS functions Ω_Q and Ω_{ET} has only an indirect link with the isotopic partitioning of J between Q and ET. The SAS functions represent only a preference of a given outflow for certain stored water ages. Since there is no one-to-one relationship between the stored water age at a given moment and the past tracer concentrations in the input (e.g. the age ambiguity of tritium, see figure 2), there is no explicit partitioning of isotopic concentrations in the model based on the SAS functions. We will add these details to the methods (2.4) and to appendix A2.

See lines 246-249, 412-414, 562-564, 675-681.

We did not focus on the parameters of the SAS function of ET because our study deals with streamflow travel times, and because we do not have tracer data representative of the ET flux that could be used to directly constrain its SAS function parameters. Instead, we indirectly constrained these parameters to the tracer data in streamflow. Similar to Van der Velde et al. (2015) and Visser et al. (2019), we found that the parameters of the ET SAS function have a non-negligible influence on the simulations of stream isotopic tracers. We agree with R2 that this relative importance of μ_{ET} was observed because of the long term isotopic partitioning of precipitation between streamflow and ET. We will include the figure below in the supplement. It shows, as suggested by R2, that the simulations constrained by ^{2}H generally yielded more tritium mass in streamflow over 2015-2017 than the simulations constrained by ^{3}H.

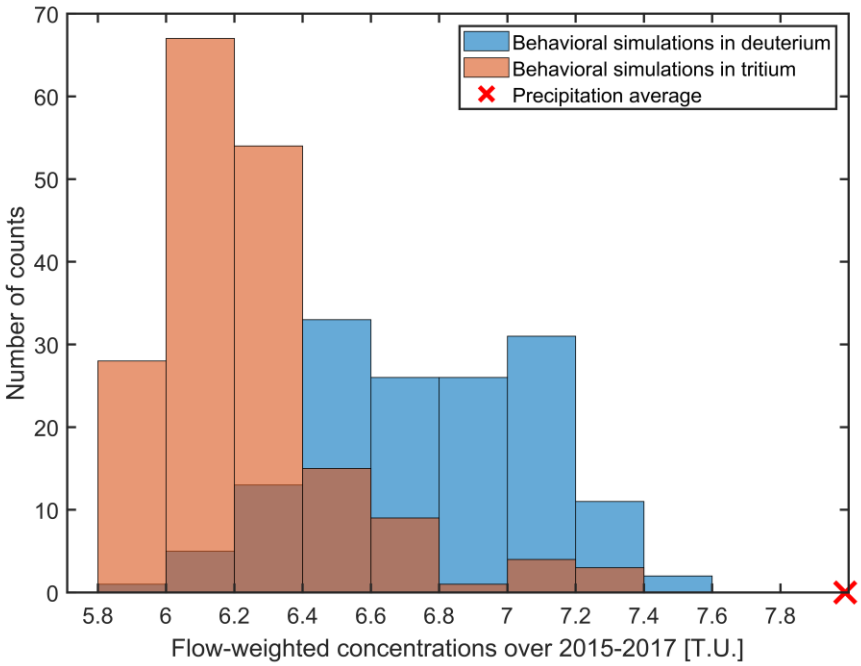


Figure: Simulated flow-weighted concentrations in the stream for the behavioral model runs constrained by deuterium samples or by tritium samples.

As R2 points out, this means that for the ^{3}H-based model, more tritium was stored, or ET removed more tritium from storage compared to the ^{2}H-based model (or both effects together). We do not have the necessary tracer observations (such as isotope samples in ET or isotope samples representative of storage) to say what mechanism happened in the catchment. In that instance, we cannot determine if the model used the correct mechanism or not. However, we can discriminate the solution based on the long term isotopic mass balance.

Tritium accumulation in modelled storage to momentarily decrease C_Q is only a short-term solution, because the stored tritium concentration cannot continuously increase in a physically realistic model. The only solution to reduce the stream tritium content in the long term (e.g. over 2 years like here, or longer) is to evacuate the excess tritium by ET. The posterior distribution of μ_{ET} constrained by tritium observations (see above) tends to lower values, indicating a stronger preference for younger water in ET compared to μ_{ET} constrained by deuterium observations. If we restrict our point of view from years 2000 to 2017, current precipitation generally has a higher tritium content than the water recharged before (see figure 2). Thus, by preferentially removing younger water, ET partly contributes to removing tritium from the system and to keeping the simulated stream tritium concentrations low over 2015-2017. This is why the information gain about μ_{ET} is so high with tritium data. It is interesting to see that the same mechanism must be occurring with stable isotopes because the information gain about μ_{ET} is also high with stable isotopes, and the figure above shows that behavioral solutions for deuterium also have a lower stream tritium content than current precipitation.

We think that the lack of high-resolution tritium data explains why the simulations constrained by tritium observations tend to have a lower stream tritium content than simulations constrained by stable isotopes. On the one hand, with only monthly measurements of precipitation taken 60 km away from the study site, our knowledge of the true tritium content of local precipitation has some uncertainty. It is possible that we overestimate the flux-weighted tritium concentration of precipitation (see the red cross in the figure above). The same remark applies to the stream tritium content. The 24 samples probably do not fully represent the flux-weighted tritium concentration in the stream. It is thus possible that we underestimate the true value, and that more samples during hydrological events (such as flashy peaks) would increase the estimated value. We will condense and add this information to the results, section 3.1. We also think that this really points to a critical limitation in many hydrological studies: the lack of appropriate sampling schemes for tracers in ET in space and time.

See lines 708-723, 742-743.

R2: Line 375: Typo in “[0,∞[“

Authors: We think that R2 means that the open squared bracket “[“ should be a parenthesis “(“. If that is the case, we observed that both notations exist, and we prefer to keep the one already used. If that is not the case, we are sorry but we do not see the typo.

R2: Line 224: It is stated that $\lambda_1(t)$ is the smallest weight. However, it is not clear how that was constrained in the model calibration.

Authors: Essentially, $\lambda_1(t) < \lambda_1^*$ and λ_1^* is sampled between 0 and $1-\lambda_2$ (hence between 0 and 1) to have $\lambda_1+\lambda_2+\lambda_3=1$ (table 1 footnotes). This means that λ_1^* is sampled more often close to 0 than close to 1, and $\lambda_1(t)$ is generally the smallest weight. We did this because large values of $\lambda_1(t)$ generally corresponded to poor simulation fits in initial tests, and because it is necessary to impose at least one relationship between two λ coefficients to be able to randomly select three λ verifying $\lambda_1+\lambda_2+\lambda_3=1$. We will add more details about this and rephrase the sentence to avoid misinterpretations.

See lines 260-262.

R2: Lines 236-237: S_{ref} is chosen not calibrated, so probably introducing the chosen value here would be better, rather than introducing it in the next section, 2.6 Model calibration.

Authors: We will do as suggested.

See line 276.

R2: The initial condition for the SAS function model is not described. If there was a spin-up for the SAS function model (like the storage estimation), what tracer time series were used?

Authors: We will add this information to the paragraph.
We detailed in section 2.2 that we periodically looped back the 2010-2015 input data to create the spin-up time series (1915-2015). The initial condition corresponds to an exponential distribution of residence times (RTD) with a mean of 1.7 years. The initial SAS functions and TTDs are then calculated based on their chosen functional form and their parameters, using this initial RTD.

See lines 279-281.

R2: Lines 404-405: How this comparison between 2016 finding and 2017 finding helps readers to understand the higher age estimated using the ^{3}H-based model?

Authors: We will rewrite this sentence to make it clearer.

See lines 494-499.

R2: Lines 437-439: Those parameters are not independent. Thus, those were not independently constrained.

Authors: What we really meant is that the shapes of the components of the streamflow SAS function were constrained independently. The only imposed relationship between the parameters of three components is $\lambda_1+\lambda_2+\lambda_3=1$. This does not affect their shape, nor their location on the age axis (or age-ranked storage axis). We will rewrite the sentence to reflect this better.

See lines 541-554.

Authors: We noticed a typo in figures 4 and 6, where we wrote 141 behavioral simulations in deuterium instead of 148 (correct value, as stated in the text). We will correct this.

See figure 5, 7.

Cartwright, I., & Morgenstern, U. (2016a). Contrasting transit times of water from peatlands and eucalypt forests in the Australian Alps determined by tritium: implications for vulnerability and the source of water in upland catchments. Hydrology and Earth System Sciences, 20, 4757-4773. doi:10.5194/hess-20-4757-2016.

Cartwright, I., & Morgenstern, U. (2016b). Using tritium to document the mean transit time and sources of water contributing to a chain-of-ponds river system: Implications for resource protection. Applied Geochemistry, 75, 9-19, http://dx.doi.org/10.1016/j.apgeochem.2016.10.007.

Crouzet, E., Hubert, P., Olive, P., Siwertz, E., Marce, A., 1970. Le tritium dans les mesures d'hydrologie de surface. Determination experimentale du coefficient de ruissellement. J. Hydrol. 11 (3), 217–229.

Doherty, J. and Johnston, J.M. (2003), METHODOLOGIES FOR CALIBRATION AND PREDICTIVE ANALYSIS OF A WATERSHED MODEL. Journal of the American Water Resources Association, 39: 251-265. doi:10.1111/j.1752-1688.2003.tb04381.x

Duvert, C., Stewart, M. K., Cendón, D. I., and Raiber, M.: Time series of tritium, stable isotopes and chloride reveal short-term variations in groundwater contribution to a stream, Hydrology and Earth System Sciences, 20, 257–277, https://doi.org/10.5194/hess-20-257-2016, 2016

Ehret, U. and Zehe, E.: Series distance – an intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events, Hydrol. Earth Syst. Sci., 15, 877–896, https://doi.org/10.5194/hess-15-877-2011, 2011.

Fenicia, F., D. Kavetski, H. H. G. Savenije, and L. Pfister (2016), From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions, Water Resour. Res., 52, doi:10.1002/2015WR017398.

Gabrielli, C. P., Morgenstern, U., Stewart, M. K., & McDonnell, J. J. (2018). Contrasting groundwater and streamflow ages at the Maimai watershed. Water Resources Research, 54, 3937–3957. https://doi.org/10.1029/2017WR021825

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.

Gusyev, M. A., Morgenstern, U., Stewart, M. K., Yamazaki, Y., Kashiwaya, K., Nishihara, T., Kuribayashi, D., Sawano, H., and Iwami, Y.: Application of tritium in precipitation and baseflow in Japan: a case study of groundwater transit times and storage in Hokkaido watersheds, Hydrol. Earth Syst. Sci., 20, 3043–3058, https://doi.org/10.5194/hess-20-3043-2016, 2016.

Helton, J. and Davis, F.: Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems, Reliability Engineering & System Safety, 81, 23 – 69, https://doi.org/10.1016/S0951-8320(03)00058-9, 2003

Hubert, P., Marin, E., Meybeck, M., Ph. Olive, E.S., 1969. Aspects Hydrologique, Géochimique et Sédimentologique de la Crue Exceptionnelle de la Dranse du Chablais du 22 Septembre 1968. Arch. Sci (Genève) 3, 581–604.

Legates, D. R., and McCabe, G. J. (1999), Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation, *Water Resour. Res.*, 35(1), 233– 241, doi:10.1029/1998WR900018.

Maloszewski, P., and Zuber, A. (1993). Principles and practice of calibration and validation of mathematical models for the interpretation of environmental tracer data in aquifers. Advances in Water Resources, 16: 173-190

Morgenstern, U., Stewart, M. K., and Stenger, R.: Dating of streamwater using tritium in a post nuclear bomb pulse world: continuous variation of mean transit time with streamflow, Hydrol. Earth Syst. Sci., 14, 2289–2301, https://doi.org/10.5194/hess-14-2289-2010, 2010.

Pfister, L., Martínez-Carreras, N., Hissler, C., Klaus, J., Carrer, G. E., Stewart, M. K., and McDonnell, J. J.: Bedrock geology controls on catchment storage, mixing, and release: A comparative analysis of 16 nested catchments, Hydrological Processes, 31, 1828–1845, https://doi.org/10.1002/hyp.11134, 2017.

Schoups, G., and Vrugt, J. A. (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, *Water Resour. Res.*, 46, W10531, doi:10.1029/2009WR008933.

Stewart, M. K. and Fahey, B. D.: Runoff generating processes in adjacent tussock grassland and pine plantation catchments as indicated by mean transit time estimation using tritium, Hydrol. Earth Syst. Sci., 14, 1021–1032, https://doi.org/10.5194/hess-14-1021-2010, 2010.

Stewart, M. K., Mehlhorn, J., and Elliott, S.: Hydrometric and natural tracer (oxygen-18, silica, tritium and sulphur hexafluoride) 900 evidence for a dominant groundwater contribution to Pukemanga Stream, New Zealand, Hydrological Processes, 21, 3340–3356, https://doi.org/10.1002/hyp.6557, 2007.

Stewart, M. K. and Thomas, J. T.: A conceptual model of flow to the Waikoropupu Springs, NW Nelson, New Zealand, based on hydrometric and tracer (18O, Cl,3H and CFC) evidence, Hydrology and Earth System Sciences, 12, 1–19, https://doi.org/10.5194/hess-12-1-2008, 2008.

Uhlenbrook, S., Frey, M., Leibundgut, C., and Maloszewski, P.: Hydrograph separations in a mesoscale mountainous basin at event and seasonal timescales, Water Resources Research, 38, 31–1–31–14, https://doi.org/10.1029/2001WR000938, 2002.

Vrugt, J. A. (2016). Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. Environmental Modelling & Software, 75, 273-316. http://dx.doi.org/10.1016/j.envsoft.2015.08.013

~~Testing the truncation~~ A comparison of catchment travel times ~~with StorAge Selection functions using~~ and storage deduced from deuterium and tritium ~~as~~ tracers using StorAge Selection functions

Nicolas B. Rodriguez[1,2], Laurent Pfister[1], Erwin Zehe[2], and Julian Klaus[1]

[1]Catchment and Eco-Hydrology Research Group, Environmental Research and Innovation Department, Luxembourg Institute of Science and Technology, Belvaux, Luxembourg

[2]Institute of Water Resources and River Basin Management, Karlsruhe Institute of Technology, Karlsruhe, Germany

Correspondence: Nicolas Rodriguez (nicolas.bjorn.rodriguez@gmail.com)

Abstract. Catchment travel time distributions (TTDs) are an efficient concept to summarize the time-varying 3-dimensional transport of water and solutes ~~to~~ towards an outlet in a single function of water age and to estimate catchment storage by leveraging information contained in tracer data (e.g. ^{2}H and ^{3}H). It is argued that the ~~increasing~~ preferential use of the stable isotopes of O and H as tracers compared to tritium ~~as tracers~~ has truncated our vision of streamflow TTDs, meaning that the long tails of the distribution associated with old water ~~are~~ tend to be neglected. However, the reasons for the truncation of the TTD tails are still obscured by methodological and data limitations. In this study, we went beyond these limitations and ~~tested the hypothesis that~~ evaluated the differences between streamflow TTDs calculated using only deuterium (^{2}H) or only tritium (^{3}H)~~are different. We similarly tested if the~~ . We also compared mobile catchment storage (derived from the TTDs) associated with each tracer~~is different~~. For this we additionally constrained a model ~~successfully developed to simulate~~ that successfully simulated high-frequency stream deuterium measurements with ~~about 30~~ 24 stream tritium measurements over the same period (2015–2017). We used data from the forested headwater Weierbach catchment (42 ha) in Luxembourg. ~~The~~ Time-varying streamflow TTDs were estimated ~~in unsteady conditions by~~ by consistently using both tracers ~~coherently~~ within a framework based on StorAge Selection (SAS) functions. We found ~~equal TTDs and equal~~ similar TTDs and similar mobile storage between the ^{2}H- and ^{3}H-derived estimates~~. The truncation hypothesis was thus rejected. The small~~ , despite statistically significant differences for certain measures of TTDs and storage. 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. The differences we found could be explained by the calculation uncertainties and by a limited sampling frequency for tritium. We conclude that stable isotopes do not seem to systematically underestimate travel times or storage compared to tritium. Using both stable and radioactive isotopes of H as tracers reduced the ~~age and storage uncertainties. Although tritium~~ travel time and storage calculation uncertainties. Tritium and stable isotopes ~~had redundant information about younger water, using both tracers~~ both had the ability to reveal short travel times in streamflow. Using both tracers together better exploited the more specific information about longer ~~ages~~ travel times that ^{3}H inherently contains~~, and it could be even better in the next decades~~. The two tracers thus had ~~overall~~ different information contents overall. Tritium was ~~however~~ slightly more informative ~~and cost-effective~~ than stable isotopes for travel time analysis ~~. We thus reiterate the call of Stewart et al. (2012) to measure~~

tritium in the streams for travel time analysis, and emphasize the need for high-frequency tritium sampling in future studies to match the resolution in stable isotopesdespite a lower number of tracer samples.

Copyright statement.

1 Introduction

Sustainable water resource management is based upon a sound understanding of how much water is stored in catchments, and how it is released to the streams. Isotopic tracers such as deuterium (^{2}H), oxygen 18 (^{18}O), and tritium (^{3}H) have become the cornerstone of several approaches to tackle these two critical questions (Kendall and McDonnell, 1998). For instance, hydrograph separation using stable isotopes of O and H (Buttle, 1994; Klaus and McDonnell, 2013) has unfolded the difference between catchments hydraulic response (i.e. streamflow) and chemical response (e.g. solutes) (Kirchner, 2003) related to the different concepts of water celerity and water velocity (McDonnell and Beven, 2014). Isotopic tracers have also been the backbone to unravel water flow paths in soils (Sprenger et al., 2016), and to distinguish soil water going back to the atmosphere and flowing to the streams (Brooks et al., 2010; McDonnell, 2014; McCutcheon et al., 2017; Berry et al., 2018; Dubbert et al., 2019).

The travel time distribution (TTD) is nevertheless the concept determination of travel time distributions (TTDs) is the method relying the most on isotopic tracers (McGuire and McDonnell, 2006). TTDs provide a concise summary of water flow paths to an outlet by leveraging the information on storage and release contained in tracer input-output relationships. TTDs are essential to link water quantity to water quality (Hrachowitz et al., 2016), for example by allowing calculations of stream solute dynamics from a hydrological model (Rinaldo and Marani, 1987; Maher, 2011; Benettin et al., 2015a, 2017a). TTDs are commonly calculated from isotopic tracers in many sub-disciplines of hydrology and thus have the potential to link the individual studies focused on the various compartments of the critical zone (e.g. groundwater and surface water) (Sprenger et al., 2019). ^{3}H has been used as an environmental tracer since the late 1950s (Begemann and Libby, 1957; Eriksson, 1958; Dinçer et al., 1970; Hubert, 1971; Martin (Begemann and Libby, 1957; Eriksson, 1958; Dinçer et al., 1970; Hubert et al., 1969; Martinec, 1975) and it gained particular momentum in the eighties with its use in diverse TTD models (Małoszewski and Zuber, 1982; Stewart et al., 2010). It is argued that ^{3}H contains more information on the age of water travel times than stable isotopes due to its radioactive decay (Stewart et al., 2012). For example, low tritium content generally indicates old water in which most of the ^{3}H from nuclear tests has decayed. Despite its potential, ^{3}H is used only rarely in travel time studies nowadays (Stewart et al., 2010), most likely because high precision analyses are laborious (Morgenstern and Taylor, 2009) and rather expensive. In contrast, the use of stable isotopes in travel time studies has soared in the last three decades (Kendall and McDonnell, 1998; McGuire and McDonnell, 2006; Fenicia et al., 2010; Heidbuechel et al., 2012; Klaus et al., 2015a; Benettin et al., 2015a; Pfister et al., 2017; Rodriguez et al., 2018). This is notably due to the fast and low-cost analyses provided by recent advances in laser spectroscopy (e.g. Lis et al., 2008; Gupta et al., 2009; Keim et al., 2014) and the associated technological progress in sampling techniques of various

water sources (Berman et al., 2009; Koehler and Wassenaar, 2011; Herbstritt et al., 2012; Munksgaard et al., 2011; Pangle et al., 2013; Herbstritt et al., 2019). According to Stewart et al. (2012) and Stewart and Morgenstern (2016), the limited use of ^{3}H may have cause a biased or "truncated" vision of stream TTDs, in which the ~~the older ages associated with~~ long TTD tails remain mostly undetected by stable isotopes. Longer mean travel times (MTT) were inferred from ^{3}H than from stable isotopes in several studies employing both tracers (Stewart et al., 2010). Longer MTTs may have profound consequences for catchment storage, usually estimated from TTDs as $S = Q \times MTT$ (with Q the flux through the catchment), assuming steady-state flow conditions (i.e. $S(t) = \overline{S(t)} = S, Q(t) = \overline{Q(t)} = Q, MTT(t) = \overline{MTT(t)} = MTT$) (McGuire and McDonnell, 2006; Soulsby et al., 2009; Birkel et al., 2015; Pfister et al., 2017). Under this assumption, a truncated TTD would result in an underestimated MTT thus an underestimated catchment storage. A different perspective on catchment storage and on its relation with travel times may however be adopted by calculating storage from unsteady TTDs.

A water molecule that reached an outlet has only one ~~age~~travel time, defined as the ~~time it took to get there. This age is not affected by the isotopes (^{2}H, ^{3}H, and ^{18}O) carried by the molecule and used as tracers, because they do not influence its flow path or its advective velocity or its self diffusion in water (Devell, 1962)~~duration between entry and exit. The use of different methods of travel time analysis for stable isotopes of O and H and for ^{3}H (e.g. amplitudes of seasonal variations vs. radioactive decay) was ~~hence~~ first pointed out as a main reason for the discrepancies in MTT (Stewart et al., 2012). Further research is thus needed for developing mathematical frameworks that coherently incorporate stable isotopes of O and H and ^{3}H in travel time calculations. Moreover, several limiting assumptions were used in previous studies employing ^{3}H to derive ~~MTTs, which are in themselves insufficient statistics~~ the MTT, which is in itself an insufficient statistic to describe various aspects (e.g. shape, modes, percentiles) of the TTDs. For example, the steady-state flow assumption has been used in almost all ^{3}H travel time studies (McGuire and McDonnell, 2006; Stewart et al., 2010; Cartwright and Morgenstern, 2016; Duvert et al., 2016; Gallart et al., 2016). Yet, time variance is a fundamental characteristic of TTDs (Botter et al., 2011; Rinaldo et al., 2015), and it has been acknowledged in simulations of stream ^{3}H only very recently (Visser et al., 2019). ~~Recharge models are also often~~ Hydrological recharge models or tracer weighting functions have also been employed to account ~~only indirectly for the impact of evapotranspiration fluxes (ET)~~ for the influence of mixing of precipitation tracer values in the unsaturated zone and for the influence of the seasonal (hence time-varying) losses to atmosphere via $ET(t)$ (e.g., Małoszewski and Zuber, 1982) on the catchment inputs in ^{3}H (Stewart et al., 2007)~~and for the TTD of ET~~. However, these methods do not explicitly represent the influence of the TTD of ET on the age-labeled water balance and thus represent indirect approximations. In contrast for stable isotopes, explicit considerations of ~~ET~~ ET and of the influence of its TTD on the streamflow TTD are becoming common (van der Velde et al., 2015; Visser et al., 2019). Finally, more guidance on the calibration of the TTD models against ^{3}H measurements is needed (see e.g. Gallart et al., 2016). Especially, uncertainties of ^{3}H-inferred ~~age estimates~~ travel times may have been overlooked, while these could explain the differences with the stable isotope-inferred ~~age estimates.~~ travel time estimates.

Besides methodological problems, the reasons for the ~~age~~ travel time differences (hence apparent storage or mixing) are still not ~~well understood~~ understood well, because little is known about the ~~true age~~ difference in information content of ^{3}H

compared to stable isotopes when determining TTDs. First, ^{3}H sampling in catchments typically differs from stable isotope sampling in terms of frequency and flow conditions. Stable isotope records in precipitation and in the streams have lately shown increasing resolution, covering a wide range of flow conditions (McGuire et al., 2005; Benettin et al., 2015a; Birkel et al., 2015; Pfister et al., 2017; von Freyberg et al., 2017; Visser et al., 2019; Rodriguez and Klaus, 2019). Tritium records in precipitation and streams are on the other hand usually at a monthly resolution in many places around the globe (IAEA and WMO, 2019; IAEA, 2019; Halder et al., 2015). Only a handful of travel time studies employing ^{3}H report more than a dozen stream samples for a given site and for different conditions than baseflow (e.g. Małoszewski et al., 1983; Visser et al., 2019). This general focus on baseflow ^{3}H sampling introduces by ~~definition~~ design a bias towards older water. Second, the natural variability of ^{3}H compared to that of stable isotopes has rarely been documented. ^{3}H in precipitation has returned to the pre-bomb levels, and like stable isotopes it shows a clear yearly seasonality (e.g. Stamoulis et al., 2005; Bajjali, 2012). However, ambiguous ~~age~~ travel time estimates may still be obtained with ^{3}H in the northern hemisphere because the current precipitation has similar ^{3}H concentrations than water recharged in the 1980s (Stewart et al., 2012). Higher sampling frequencies of precipitation ^{3}H are almost nonexistent. Rank and Papesch (2005) revealed a short term variability of precipitation ^{3}H likely due to different air masses~~.~~. This variability was observed also during complex meteorological conditions such as hurricanes (Östlund, 2013). ^{3}H in streams also ~~show some~~ exhibits yearly seasonality (Różański et al., 2001; Rank et al., 2018), but short term dynamics are not ~~well understood~~ understood well because high frequency data sets are limited. Dinçer et al. (1970) showed that short-term stream tritium variations can be caused by the melting of the snowpack from the current and the previous winters. In addition, the ~~seasonally higher~~ seasonally-higher values of precipitation ^{3}H in ~~Spring~~ spring could explain some of the ^{3}H peaks observed in the large rivers (Rank et al., 2018). More studies employing both ^{3}H and stable isotopes and comparing their ~~age~~ travel time information content are therefore crucial to understand travel times in catchments from a multi-tracer perspective.

In this study, we go beyond previous work and ~~test the hypothesis that stream~~ assess the differences between streamflow TTDs and the associated catchment storage ~~are different~~ (considering their uncertainties) when those are inferred from stable isotopes or from ^{3}H measurements used in a coherent mathematical framework for both tracers. For this, we use high frequency isotopic tracer data from an experimental headwater catchment in Luxembourg. Here we focus on the stable isotope of H (deuterium ^{2}H) for which we have more precise measurements. A transport model based on TTDs was recently developed and successfully applied to simulate a two-year high frequency (sub-daily) record of $^{2}\delta^{2}$H in the stream (Rodriguez and Klaus, 2019). Here, we additionally constrain the same model within the same mathematical framework against ~~nearly 30~~ 24 stream samples of ^{3}H collected during highly varying flow conditions over the same period as for ^{2}H. We do not assume steady-state ~~and do not rely on a recharge model by employing~~ flow conditions and we employ StorAge Selection functions to account for the type and the variability of the TTDs of Q and ET that affect the water age balance in the catchment. The tracer input-output relationships and the ^{3}H radioactive decay are accounted for in the method, which reduces ^{3}~~H age~~ H-derived travel time ambiguities. We provide guidance on how to jointly calibrate the model to both tracers and on how to derive likely ranges of storage estimates and travel time measures other than the MTT. This work addresses the following related research questions:

– Are travel times and storage inferred from a common transport model for ^{2}H and ^{3}H in disagreement?

– Are the ~~water age~~ travel time information contents of ^{2}H and ^{3}H similar?

2 Methods

2.1 Study site description

This study is carried out in the Weierbach catchment, which has been the focus of an increasing number of investigations in the last few years about streamflow generation (Glaser et al., 2016, 2019; Scaini et al., 2017, 2018; Carrer et al., 2019; Rodriguez and Klaus, 2019), biogeochemistry (Moragues-Quiroga et al., 2017; Schwab et al., 2018), and pedology and geology (Juilleret et al., 2011; Gourdol et al., 2018).

The Weierbach is a forested headwater catchment of 42 ha located in northwestern Luxembourg (Fig. 1). The vegetation consists mostly of deciduous hardwood trees (European beech and Oak), and conifers (*Picea abies* and *Pseudotsuga menziesii*). Short vegetation covers a riparian area that is up to 3 m wide and that surrounds most of the stream. The catchment morphology is a deep V-shaped valley in a gently sloping plateau. The geology is essentially Devonian slate of the Ardennes massif, ~~phyllades~~phyllite, and quartzite (Juilleret et al., 2011). Pleistocene Periglacial Slope Deposits (PPSD) cover the bedrock and are oriented parallel to the slope (Juilleret et al., 2011). The upper part of the PPSD ($\sim$ 0–50 cm) has higher drainable porosity than the lower part of the PPSD ($\sim$ 50–140 cm) (Gourdol et al., 2018; Martínez-Carreras et al., 2016). Fractured and weathered bedrock lies from $\sim$ 140 cm depth to $\sim$ 5 m depth on average. Below $\sim$ 5 m depth lies the fresh bedrock that can be considered impervious. The climate is temperate and semi oceanic. The flow regime is governed by the interplay of seasonality between precipitation and evapotranspiration. Precipitation is fairly uniformly distributed over the year, and averages 953 mm/yr over 2006–2014 (Pfister et al., 2017). The runoff coefficient over the same period is 50 %. Streamflow (Q) is double-peaked during wetter periods (Martínez-Carreras et al., 2016), and ~~single peaked~~ single-peaked during drier periods occurring normally in summer when evapotranspiration (ET) is high.

Based on previous modeling (e.g. Fenicia et al., 2014; Glaser et al., 2019) and experimental studies (e.g. Martínez-Carreras et al., 2016; Juilleret et al., 2016; Scaini et al., 2017; Glaser et al., 2018), Rodriguez and Klaus (2019) proposed a perceptual model of streamflow generation in the Weierbach. In this model, the first and flashy peaks of double-peaked hydrographs are generated by precipitation falling directly into the stream, by saturation excess flow from the near-stream soils, and by infiltration excess overland flow in the riparian area. The second peaks are generated by delayed lateral subsurface flow. The lateral fluxes are assumed higher at the PPSD/bedrock interface due to the hydraulic conductivity contrasts ~~(Glaser et al., 2016, 2019)~~(Glaser et al., 2016, 2019; Loritz et al., 2017). Lateral subsurface flows are thus accelerated when groundwater rises after a rapid vertical infiltration through the soils (Rodriguez and Klaus, 2019). The model based on travel times presented in this study was developed in a step-wise manner based on this hypothesis of streamflow generation. The model's ability to simulate stream δ^2H dynamics helped to further confirm that these flow processes are active in the Weierbach (Rodriguez and Klaus, 2019). Water flow paths and streamflow generation processes in this catch-

ment are however not completely resolved. Other studies carried out in the Colpach catchment (containing the Weierbach) ~~highlighted the potential role of lateral preferential flow through macropores in the highly heterogeneous soils for the generation first peaks of the hydrographs (Angermann et al., 2017; Loritz et al., 2017),~~ 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 understanding from ~~various~~ other studies in the Weierbach (Glaser et al., 2016, 2020).

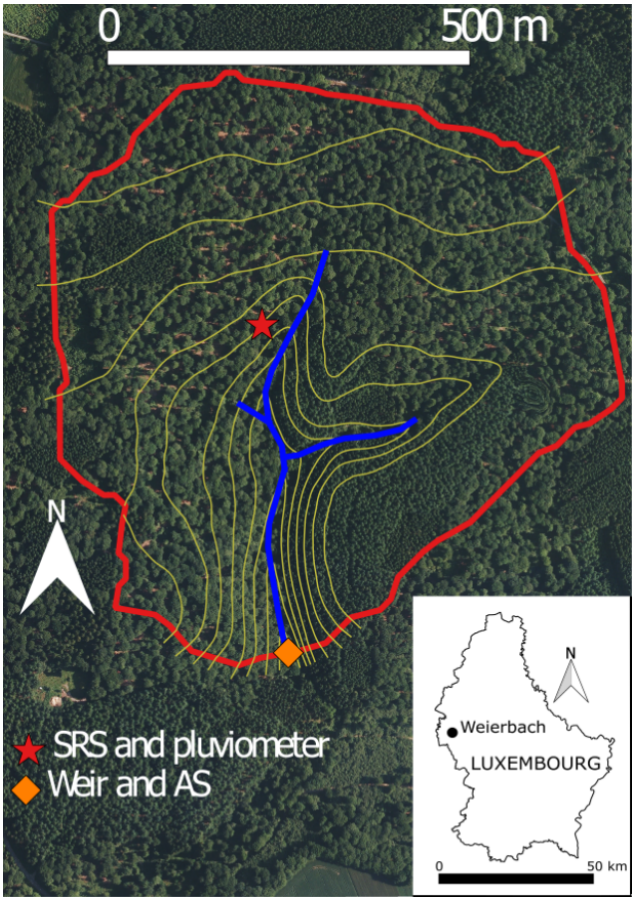


Figure 1. Map of the Weierbach catchment and its location in Luxembourg. The weir is located at coordinates (5°47'44" E, 49°49'38" N). SRS is the sequential rainfall sampler. AS is the stream autosampler. The elevation lines ~~go~~ increase by ~~increments of~~ 5 m from 460 m.a.s.l. downstream close to the weir location to 510 m.a.s.l. at the northern catchment divide.

2.2 Hydrometric and tracer data

In this study we use precipitation (J, in mm/h), ET (mm/h), Q (mm/h), and δ^2H (‰) and ^{3}H (Tritium Units, T.U.) measurements in precipitation ($C_{P,2}$ and $C_{P,3}$ respectively) and ~~in the stream~~ streamflow ($C_{Q,2}$ and $C_{Q,3}$ respectively). Here ~~a~~ the

subscript 2 indicates deuterium (^{2}H) and ~~a~~ the subscript 3 indicates tritium (^{3}H). The analysis in this study focuses on the period October 2015–October 2017 ~~during which most samples were collected at higher frequencies than in the past~~ (Fig. 2). Details on the hydrometric data collection (J, ET, Q), and on the ^{2}H sample collection and analysis are given in Rodriguez and Klaus (2019).

The 1088 stream samples analyzed for ^{2}H were collected manually or automatically with an autosampler (AS, Fig. 1), resulting in samples every 15 hours on average over October 2015–October 2017. These samples represent most flow conditions in the catchment in terms of frequency of occurrence (Fig. 3). The 525 precipitation samples analyzed for ^{2}H were collected approximately every 2.5 mm rain increment (i.e. on average every 23 hours) with a sequential rainfall sampler (SRS) and in addition as bulk samples on a bi-weekly basis. The samples were analyzed at the Luxembourg Institute of Science and Technology (LIST) using an LGR Isotope Water Analyzer, yielding for ^{2}H an analytical accuracy of 0.5 ‰ (equal to the LGR standard accuracy), and a precision maintained <0.5 ‰ (quantified as one standard deviation of the measured samples and standards).

The 24 stream samples analyzed for ^{3}H were selected from manual bi-weekly sampling campaigns to cover various flow ranges. The ~~samples~~ manual selection was not based on flows ranked by exceedance probabilities but rather on the streamflow time series itself. The selected samples represent various hydrological conditions (e.g. beginning of a wet period after a long dry spell, small but flashy streamflow responses), based on data available for this catchment (see Sect. 2 and Rodriguez and Klaus, 2019) . The 24 tritium samples cover a wide portion of the flow frequencies (c.f. Fig. 3, all sampled flows conditions occurring more than 90% of the time). This number of ^{3}H samples is one of the highest used in travel time studies (c.f., Maloszewski and Zuber, 1993; Uhle , and it is limited by the analytical costs. The samples were analyzed by the GNS Science Water Dating Laboratory (Lower Hutt, New Zealand), which provides high precision tritium measurements using electrolytic enrichment and liquid scintillation counting (Morgenstern and Taylor, 2009). The precision of the stream samples varies from roughly 0.07 T.U. to roughly 0.3 T.U., but is usually around 0.1 T.U. Monthly values of ^{3}H in precipitation were obtained for the Trier station (60 km from the Weierbach) until 2016 from the WISER database of the International Atomic Energy Agency (IAEA) (IAEA and WMO, 2019; Stumpp et al., 2014). The 2017 values were obtained from the Radiologie group of Bundesanstalt für Gewässerkunde ~~(?)~~ (Schmidt et al., 2020). ^{3}H in precipitation before 1978 was calculated by regression with data from Vienna, Austria (Stewart et al., 2017).

For both ^{2}H and ^{3}H, the time series of tracer in precipitation was interpolated between two consecutive samples (e.g. A and B) as being equal to the value of the next sample (i.e. B). This was necessary to obtain a continuous tracer input time series (required for Eq. 1 to work). Since no measurements of J, Q, ET, and $C_{P,2}$ are available before 2010, we looped back their values of the period October 2010–October 2015 periodically before 2010 as a best estimate of their past values. We aggregated the input data (J, ET, Q, $C_{P,2}$, $C_{P,3}$) to a resolution $\Delta t = 4$ hours, which is small enough to capture the variability of flows and tracers in the input and simulate the variability of the flows and tracers in the output.

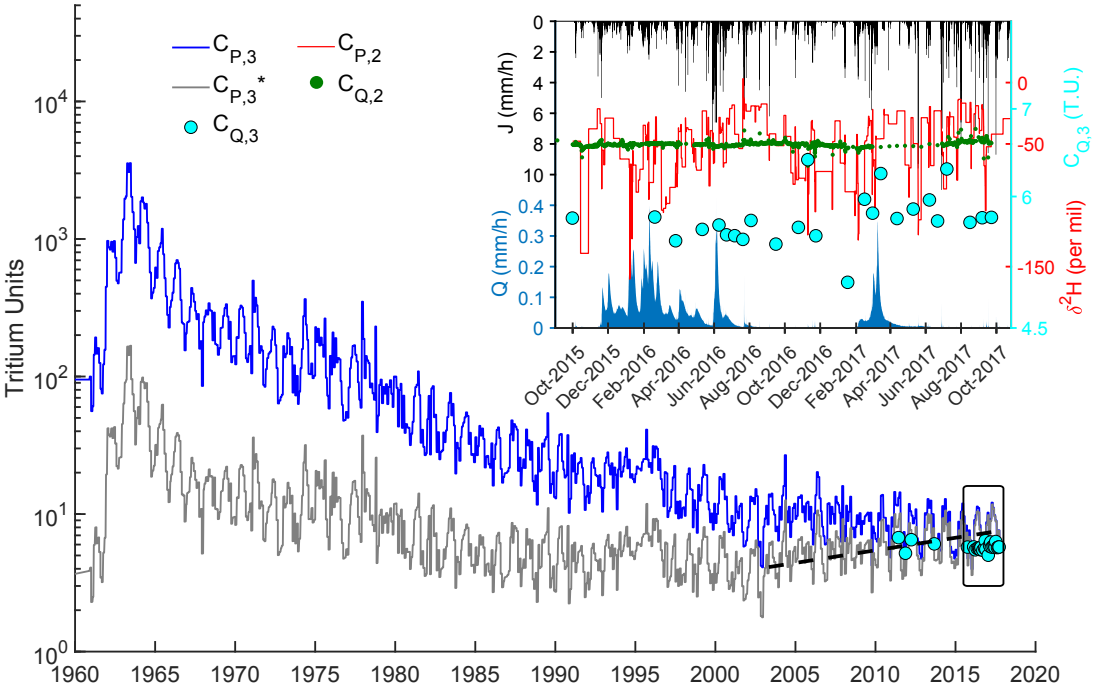


Figure 2. Data used in this study: ^{3}H in precipitation ($C_{P,3}$), the corresponding tritium activities accounting for radioactive decay until 2017 ($C^*_{P,3}$), δ^2H in precipitation $C_{P,2}$ (inset), precipitation J (inset), streamflow Q (inset), ^{3}H measurements in the stream ($C_{Q,3}$ both plots), and δ^2H in the stream ($C_{Q,2}$, inset). The period contained in the inset is represented as a rectangle in the bigger plot. The dashed line visually represents the increasing trend in $C^*_{P,3}$ that emerges as the effect of bomb peak tritium disappears (i.e. $C_{P,3}(t-T)$ stops decreasing around 2000 so $C^*_{P,3}(T,t) = C_{P,3}(t-T)\,e^{-\alpha T}$ starts decreasing with increasing T).

2.3 Mathematical framework

Mathematically, the streamflow TTD is related to the stream tracer concentrations $C_Q(t)$ according to the following Eq. (1):

$$C_Q(t) = \int_{T=0}^{+\infty} C^*_P(T,t)\,\overleftarrow{p_Q}(T,t)\,dT \tag{1}$$

where T is ~~water age~~ travel time (the age of water at the outlet), t is time of observation, $C_Q(t)$ is the stream tracer concentration, $\overleftarrow{p_Q}$ (probability distribution function, p.d.f.) is the stream backward TTD (Benettin et al., 2015b), and $C^*_P(T,t)$ is the tracer concentration of the water parcel reaching the outlet at time t with ~~age~~ travel time T (this parcel was in the inflow at time $t-T$~~, its travel time is thus T~~). This equation is always verified for the exact (usually unknown) TTD, because it

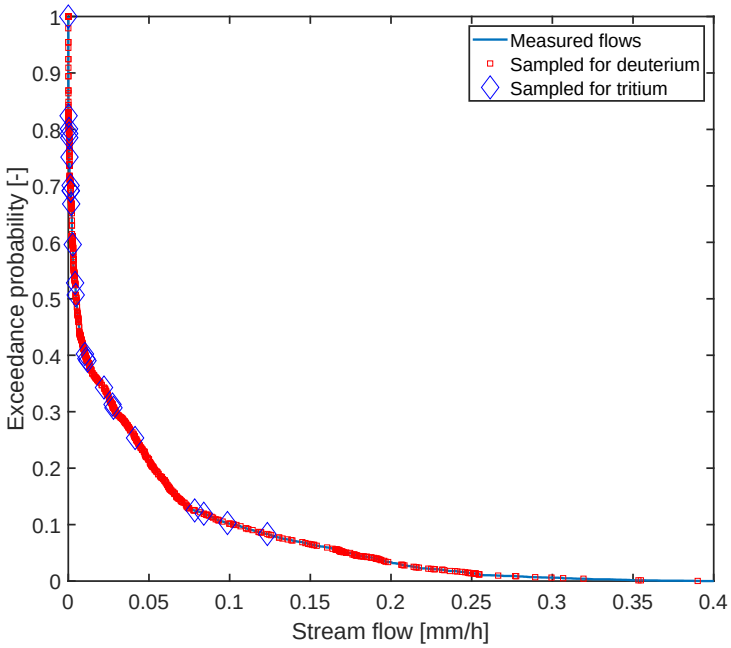


Figure 3. Distribution of stream samples (^{3}H and δ^2H) along the flow exceedance probability curve defined as the fraction of stream flows exceeding a given value over 2015–2017.

simply expresses the fact that the stream concentration is the volume-weighted arithmetic mean of the concentrations of the water parcels with different travel times at the outlet (the weighting of tracer concentrations by hydrological fluxes is thus implicit in $\overleftarrow{p_Q}(T,t)$). $C_P^*(T,t)$ depends on T and t as separate variables if the tracer concentration of a water parcel in the catchment changes between injection time $t-T$ and observation time t. For solutes like silicon and sodium, the concentration can increase with ~~age~~ travel time (Benettin et al., 2015a). For ^{3}H, radioactive decay with a constant $\alpha = 0.0563$ yr^{-1} implies $C^*_{P,3}(T,t) = C_{P,3}(t-T)\,e^{-\alpha T}$, where $C_{P,3}(t-T)$ is the concentration in precipitation measured at $t-T$. For ^{2}H, $C^*_{P,2}(T,t) = C_{P,2}(t-T)$. Thus, the streamflow TTD simultaneously verifies Eq. (2) and (3):

$$C_{Q,2}(t) = \int_{T=0}^{+\infty} C^*_{P,2}(T,t)\,\overleftarrow{p_Q}(T,t)\,dT = \int_{T=0}^{+\infty} C_{P,2}(t-T)\,\overleftarrow{p_Q}(T,t)\,dT \tag{2}$$

$$C_{Q,3}(t) = \int_{T=0}^{+\infty} C^*_{P,3}(T,t)\,\overleftarrow{p_Q}(T,t)\,dT = \int_{T=0}^{+\infty} C_{P,3}(t-T)\,e^{-\alpha T}\,\overleftarrow{p_Q}(T,t)\,dT \tag{3}$$

Practically, when measurements of ^{2}H and ^{3}H are used to inversely deduce the TTD by using Eq. 2 and 3, different TTDs may be found. These different TTDs may be called $\overleftarrow{p_{Q,2}}$ and $\overleftarrow{p_{Q,3}}$ for instance, referring to ^{2}H and ^{3}H, respectively. To avoid introducing more variables and to avoid confusion, we do not use the names $\overleftarrow{p_{Q,2}}$ and $\overleftarrow{p_{Q,3}}$ and we instead refer to

the TTDs "constrained" by a given tracer, using a common symbol $\overleftarrow{p_Q}$. We do this also to stress that the exact (true) TTD must simultaneously verify both Eq. (2) and (3), and that two different TTDs $\overleftarrow{p_{Q,2}}$ and $\overleftarrow{p_{Q,3}}$ cannot physically exist. This is a fundamental difference from previous work that assumed two different TTDs, using for example Eq. (3) for ^{3}H and another method for ^{2}H (the sine-wave approach) (e.g. Małoszewski et al., 1983). The framework in this study also uses the fact that the same functional form of streamflow TTD needs to simultaneously explain both tracers to be valid, unlike previous work that used different TTD models for different tracers (Stewart and Thomas, 2008).

2.4 Transport model based on TTDs

Most of the previous travel time studies using tritium assumed steady-state ~~and~~ flow conditions, an analytical shape for the ~~stream~~ streamflow TTD, and fitted the parameters of the analytical function using the framework described in 2.3. In this study, the TTDs are unsteady (i.e. time-varying or transient) and cannot be analytically described. Still, they can be calculated by numerically solving the "Master Equation" (Botter et al., 2011). This method has been applied in several recent studies (e.g. van der Velde et al., 2015; Harman, 2015; Benettin et al., 2017b), and is described in more details by Benettin and Bertuzzo (2018). The numerical method used to solve this equation in this study is described by Rodriguez and Klaus (2019).

Essentially, the Master Equation is a water balance equation where storage and fluxes are labeled with age categories. The Master Equation is thus a partial differential equation~~expressing~~. It expresses the fact that the amount of water in storage ~~having age T changes with time because of~~ with a given residence time changes with calendar time. This change is due to new water introduced by precipitation $J(t)$, ~~because of~~ to water aging, and ~~because of~~ to losses to catchment outflows $ET(t)$ and $Q(t)$. Solving the Master Equation requires knowledge (or an assumption about the shape) of the StorAge Selection (SAS) functions Ω_Q and Ω_{ET} of outflows Q and ET, which conceptually represent how likely water ages in storage (residence times) are to be present in the outflows at a given time. Solving the Master Equation yields the distribution of ~~ages~~ residence times in storage at every moment, that can be represented in a cumulative form with age-ranked storage S_T, defined as the amount of water in storage (e.g. 10 mm) younger than ~~T~~ T (e.g. 1 year) at time t. $T \rightarrow S_T$ is just a mathematical change of variable, and it has no meaning respective to the location or depth of a ~~certain~~ water parcel with ~~age T~~ a certain residence time in the catchment. By definition $\lim_{T \to +\infty} S_T = S(t)$, where $S(t)$ is catchment storage. Ω_Q and Ω_{ET} are functions of S_T and cumulative distributions functions (c.d.f.) for numerical convenience. SAS functions are closely linked to TTDs, such that one can be found from the other using the following expression (here for Q, but valid for other outflows):

$$\overleftarrow{p_Q}(T,t) = \frac{\partial}{\partial T}(\Omega_Q(S_T,t)) \tag{4}$$

The partial derivative with respect to ~~age~~ travel time T ensures the transition from c.d.f. to p.d.f. Assuming a parameterized form for Ω_Q and Ω_{ET} and calibrating their parameters using the framework defined in 2.3 yields time-varying TTDs constrained by the tracers in the outflows.

In this study, ~~we~~ the parameters of Ω_Q are directly calibrated by using Eq. 1 for C_Q. Since no tracer data C_{ET} is available, the parameters of Ω_{ET} are indirectly deduced from Eq. 1 using the tracer measurements in streamflow only. This is made

possible by the indirect influence of Ω_{ET} on the tracer partitioning between Q and ET and on the tracer mass balance (App. A2).

We assumed that Ω_{ET} is a function of only S_T and it is gamma distributed with a mean parameter μ_{ET} (mm) and a scale parameter θ_{ET} (mm). Rodriguez and Klaus (2019) showed that in the Weierbach, a weighted sum of three components in the streamflow SAS function is more consistent with the superposition of streamflow generation processes (i.e. saturation excess flow, saturation overland flow, lateral subsurface flow, see Sect. 2.1) than a single component. This means that Ω_Q is written as a weighted sum of three c.d.f.s (see appendix A1) (Rodriguez and Klaus, 2019):

$$\Omega_Q(S_T,t) = \lambda_1(t)\,\Omega_1(S_T) + \lambda_2(t)\,\Omega_2(S_T) + \lambda_3(t)\,\Omega_3(S_T) \tag{5}$$

$\lambda_1(t)$, $\lambda_2(t)$, and $\lambda_3(t)$ are time-varying weights summing to 1. ~~Essentially,~~ $\lambda_1(t)$ is ~~the smallest weight and it is parameterized to increase sharply~~ parameterized to sharply increase during flashy streamflow events, using parameters λ_1^*, f_0, S_{th} (mm), and ΔS_{th} (mm) ~~.~~ (App. A1). $\lambda_2(t) = \lambda_2$ is calibrated, and $\lambda_3(t)$ just deduced by difference. Ω_1 is a cumulative uniform distribution over S_T in $[0, S_u]$ (with S_u a parameter in mm). Ω_1 represents the young water contributions associated with short flow paths during flashy streamflow events. We chose rather low values of λ_1^* (see Table 1) such that $\lambda_1(t)$ is generally the smallest weight (because $\lambda_1(t) \leq \lambda_1^*$). The lower values of $\lambda_1(t)$ compared to other weights are consistent with tracer data suggesting limited contributions of event water to streamflow (Martínez-Carreras et al., 2015; Wrede et al., 2015). Ω_1 corresponds to processes in the near stream area: saturation excess flow, saturation overland flow, and rain on the stream (Rodriguez and Klaus, 2019). Ω_2 and Ω_3 are gamma-distributed with mean parameters μ_1 and μ_2 (mm), and scale parameters θ_1 and θ_2 (mm) respectively. Ω_2 and Ω_3 represent older water that is always contributing to the stream. This older water consists of groundwater stored in the weathered bedrock that flows laterally in the subsurface. Note that we used the same functional form of $\Omega_Q(S_T,t)$ for ^{2}H and ^{3}H to keep the functional form of the TTDs consistent between the tracers. Although composite SAS functions may considerably increase the complexity of the model compared to "traditional" SAS functions, they are necessary to account for different streamflow generation processes (Rodriguez and Klaus, 2019). These processes are potentially associated with contrasting flow path lengths and/or water velocities hence contrasting travel times. The accurate representation of these contrasting travel times is most likely vital for reliable simulations of stream chemistry (Rodriguez et al., 2020).

2.5 Model initialization and numerical details

Numerically solving the Master Equation requires an estimation of catchment mobile storage $S(t)$. ~~In this context~~Here, $S(t)$ represents the sum of "dynamic" (or "active") storage and "inactive" (or "passive") storage (Fenicia et al., 2010; Birkel et al., 2011; Soulsby et al., 2011; Hrachowitz et al., 2013). In this study the model is initialized ~~in October 1915 with storage $S(t=0) = S_{ref}$~~ with storage $S(t=0) = S_{ref} = 2000$ mm. This initial value is chosen large enough to sustain Q and ET during drier periods and to store water that is sufficiently old to satisfy Eq. (1). $S(t)$ is then simply deduced from the water balance as $S(t) = S_{ref} + \int_{x=0}^{t} (J(x) - Q(x) - ET(x))\,dx$. The initial ~~age~~ residence time distribution in storage $p_S(T,t)$ is exponential with a mean of 1.7 years, the estimated Mean Residence Time (MRT) by Pfister et al. (2017). Initial

conditions need not be specified for the SAS functions, since these are directly calculated from the initial state variables ($S_T(t=0) = S(t=0) \int_{x=0}^{+\infty} p_S(x, t=0)\, dx$) assuming a parametric form and a set of parameter values. The model is then run with time steps $\Delta t = 4$ hours and age resolution $\Delta T = 8$ hours. This way the computational cost is balanced with the resolution of the simulations in δ^2H. ~~The period October 1915–October 2015 serves as a long~~ A 100-year spin-up ~~period~~ is used to numerically allow the presence of water up to 100 years old in storage and to avoid a numerical truncation of the TTDs. This spin-up is also long enough to completely remove the impact of the initial conditions. This means that S_{ref} and the initial ~~age~~ residence time distribution in storage do not influence the results over October 2015–October 2017. $ET(t)$ is taken equal to potential evapotranspiration $PET(t)$ except that it tends non-linearly towards 0 (using a constant smoothing parameter n) when storage $S(t)$ decreases below $S_{ref} - S_{root}$ (mm), where S_{root} is a parameter accounting for the water amount accessible by ET (appendix A2).

2.6 Model calibration

The parameters of the SAS functions and the other model parameters were calibrated using a Monte Carlo technique. In total, 12 parameters were calibrated (Table 1). The initial ranges were selected based on parameter feasible values (e.g. f_0 between 0 and 1 by definition), on previous estimations (e.g. S_{th}), on hydrological data (e.g. S_u and ΔS_{th} deduced from average precipitation depths), and on initial tests on the parameter ranges (e.g. μ and θ). These ranges allow a wide range of shapes of SAS functions while minimizing numerical errors (occurring for example for $S_T > S(t)$).

Unlike our previous modeling work in this catchment (Rodriguez and Klaus, 2019), we fixed the initial storage in the model S_{ref} (to 2000 mm). We did this to reduce the degrees of freedom when sampling the parameter space in order to limit the impact of numerical errors on the calibration. These errors are due to numerical truncation of $\Omega_Q(S_T, t)$ when a considerable part (e.g. a few percent) of its tail extends above $S(t)$. This occurs when parameters μ_2, μ_3, θ_2, and θ_3 are too large compared to S_{ref} when the latter is also randomly sampled. Choosing a constant large value for S_{ref} thus guarantees the absence of truncation errors. S_{ref} has little influence on the storage deduced from travel times since the ages sampled from storage by streamflow are governed only by μ_2, μ_3, θ_2, and θ_3. These parameters are independent of S_{ref} as long at it allows sufficiently old water to reside in storage, which is ensured by its large value and by the long spin-up period we used (100 years).

The first step of the Monte Carlo procedure ~~we employed consists~~ consisted in randomly sampling parameters from the uniform prior distributions with ranges defined in Table 1. 12,096 sets of the 12 calibrated parameters were sampled as a Latin Hypercube (LHS, Helton and Davis, 2003). This sampling technique has the advantages of a stratified sampling technique and the simplicity and objectivity of a purely random sampling technique (Helton and Davis, 2003). It was chosen to make sure that the parameter samples are as evenly distributed as possible despite their relatively small number with respect to the high number of dimensions (due to computational constraints enhanced by the required long spin-up period). The model was then run over ~~October 1915–October~~ the 100-year spin-up followed by October 2015–October 2017, and its performance was

Table 1. Model parameters

Symbol	Type	Unit	Initial range	Description[a]
S_{th}	Calibrated	mm	$[20, 200]$	Storage threshold relative to S_{min} separating "dry" and "wet" periods
ΔS_{th}	Calibrated	mm	$[0.1, 20]$	Threshold in short term storage changes identifying "first" peaks in hydrographs
S_u	Calibrated	mm	$[1, 50]$	Range of the uniformly distributed Ω_1
f_0	Calibrated	–	$[0, 1]$	Young water coefficient for the dry periods
λ_1^*	Calibrated	–	$[0, 1]^b$	Maximum value of the weight $\lambda_1(t)$
λ_2	Calibrated	–	$[0, 1]$	Constant[c] value of the weight $\lambda_2(t)$
μ_2	Calibrated	mm	$[0, 1600]$	Mean parameter of the gamma distributed Ω_2
θ_2	Calibrated	mm	$[0, 100]$	Scale parameter of the gamma distributed Ω_2
μ_3	Calibrated	mm	$[0, 1600]$	Mean parameter of the gamma distributed Ω_3
θ_3	Calibrated	mm	$[0, 100]$	Scale parameter of the gamma distributed Ω_3
μ_{ET}	Calibrated	mm	$[0, 1600]$	Mean parameter of the gamma distributed Ω_{ET}
θ_{ET}	Calibrated	mm	$[0, 100]$	Scale parameter of the gamma distributed Ω_{ET}
S_{root}	Constant	mm	150	Water amount accessible by ET
m	Constant	–	1000	Smoothing parameter for the calculation of $\lambda_1(t)$
n	Constant	–	20	Smoothing parameter for the calculation of $ET(t)$ from $PET(t)$
Δt^*	Constant	hours	8	Width of the moving time window used to calculate short term storage variations $\overline{\Delta S(t)}$

[a] Details about the equations involving these parameters are given in appendix A1 and in Rodriguez and Klaus (2019)

[b] λ_1^* is in fact uniformly sampled between 0 and $1 - \lambda_2 \leq 1$ to ensure that $\sum_{n=1}^{3} \lambda_k(t) = 1$. This also ensures that values close to 0 are more often sampled than values close to 1 for λ_1^*.

[c] $\lambda_1(t)$ varies, λ_2 is constant, and $\lambda_3(t)$ varies and it is deduced using $\lambda_3(t) = 1 - \lambda_2 - \lambda_1(t)$

evaluated over October 2015–October 2017. We evaluated model performance in a multi-objective manner, by using separate objective functions for ^{2}H and ^{3}H. For deuterium, we used the Nash-Sutcliffe Efficiency (NSE):

$$E_2 = 1 - \frac{\sum_{k=1}^{N_2}(C_{Q,2}(t_k) - \delta^2 H(t_k))^2}{\sum_{k=1}^{N_2}(\delta^2 H(t_k) - \overline{\delta^2 H})^2} \tag{6}$$

where $N_2 = 1,016$ is the number of deuterium observations in the stream. For tritium, we used the Mean Absolute Error:

$$E_3 = \sum_{j=1}^{N_3} |C_{Q,3}(t_j) - {}^3H(t_j)| \tag{7}$$

where $N_3 = 24$ is the number of tritium observations in the stream. We used the MAE for tritium because it is common to report errors in T.U., and because of the limited variance of stream ^{3}H (due to the ~~low~~ limited number of samples and the low variability) making the NSE less appropriate (Gallart et al., 2016). The behavioral parameter sets that are used for

uncertainty calculations and further analysis were selected based on threshold values L_2 and L_3 for the performance measures E_2 and E_3 respectively (Beven and Binley, 1992). Parameter sets were considered behavioral for deuterium simulations if $E_2 > L_2 = 0$, and behavioral for tritium simulations if $E_3 < L_3 = 0.5$ T.U. We subsequently refer to these parameter sets and corresponding simulations as "constrained by deuterium", "constrained by tritium", and as "constrained by both" when both performance criteria were used. We chose these constraints to get reasonable model fits to the data, to obtain a comparable number of behavioral parameter sets for ^{2}H and ^{3}H, and to maximize the amount of information gained about the parameters when adding a constraint on the model performance for a tracer. This information gain was assessed with the Kullback-Leibler Divergence D_{KL} between the ~~posterior~~ parameter distributions inferred from various combinations of constraints L_2 and L_3 (Sect. 2.7).

2.7 Information contents of ^{2}H and ^{3}H

~~Loritz et al. (2018) and Loritz et al. (2019)~~ Loritz et al. (2018, 2019) recently used information theory to detect hydrological similarity between hillslopes of the Colpach catchment, and to compare topographic indexes in the Attert catchment in Luxembourg. Thiesen et al. (2019) used information theory to build an efficient predictor of rainfall-runoff events. ~~Here~~ In this study we leverage information theory to evaluate our model parameter uncertainty ~~. For this~~ (Beven and Binley, 1992), and to assess the added value of δ^2H and ^{3}H tracers for information gains on travel times. First, we calculated the expected information content of the prior and posterior parameter distributions ~~constrained by deuterium or tritium~~ using the Shannon entropy $\mathcal{H}$:

$$\mathcal{H}(X|^iH) = -\sum_{k=1}^{n_I} f(I_k) \log_2 f(I_k) \tag{8}$$

In this equation, the parameter X (e.g. μ_1) takes values (e.g. 125 mm) falling in intervals I_k (e.g. $[100, 150]$ mm) that do not intersect each other and which union $\cup_{k=1}^{n_I} I_k$ equals I_X, the total interval of values on which X is defined (e.g. $[50, 500]$ mm). The definitions of the n_I intervals I_k for each parameter depend on the binning of the parameter values, given in Table 2. The ~~posterior probability~~ distribution f defines the probability of the parameter X to be in a certain state (i.e. to take a value falling in an interval I_k), when constrained by the criterion $E_2 > L_2\,(i = 2)$ or $E_3 < L_3\,(i = 3)$ ~~.~~ (posterior distribution) or none of those (prior distribution). f can also be calculated for a combination of these criteria ($\mathcal{H}(X|(^2H \cap {}^3H))$). When using the logarithm of base 2, $\mathcal{H}$ is expressed in bits of information contained in the ~~posterior~~ distribution f. The uniform distribution over I_X has the maximum possible entropy. Lower values of $\mathcal{H}$ thus indicate that the ~~posterior~~ distribution is not flat, hence less uncertain than the uniform prior distribution. In general, lower values of $\mathcal{H}$ indicate less uncertain parameters. Lower values of $\mathcal{H}$ for the posteriors also indicate that information on travel times was extracted from the tracer time series. We used the Kullback-Leibler Divergence D_{KL} to precisely evaluate the information gain from prior to posterior distributions:

~~We also use~~

$$D_{KL}(X|^iH, X) = \sum_{k=1}^{n_I} f(I_k) \log_2 \frac{f(I_k)}{g(I_k)} \tag{9}$$

where f is the posterior distribution constrained by $E_2 > L_2$ and/or $E_3 < L_3$, and g is the prior distribution. D_{KL} is expressed in bits of information gained when the knowledge about a parameter distribution is updated by using tracer data. Summing the $D_{KL}(X|^iH, X)$ for all the parameters and for a given tracer ($i = 2$ or $i = 3$) yields the total amount of information learned on travel times from that tracer. We also used the Kullback-Leibler Divergence D_{KL} to evaluate the gain of information when ^{3}H is used in addition to ^{2}H to constrain model predictions or vice versa:

$$D_{KL}(X|(^2H \cap {}^3H), X|^iH) = \sum_{k=1}^{n_I} f(I_k) \log_2 \frac{f(I_k)}{g(I_k)} \tag{10}$$

where f is the posterior distribution constrained by $E_2 > L_2$ and $E_3 < L_3$, and g is the posterior distribution constrained only by $E_2 > L_2$ ($i = 2$) or only by $E_3 < L_3$ ($i = 3$). D_{KL} is expressed in bits of information gained when the knowledge about a parameter posterior distribution is updated by adding another tracer. ~~D_{KL} can also be used to evaluate the gain of information from prior to posterior parameter distributions (by using g = prior and f = posterior).~~ Calculating D_{KL} also requires binning the parameter values to define the intervals I_k and calculate the distributions f and g. The binning for each parameter (Table 2) was chosen such that the resulting histograms visually reveal the underlying structure of the parameter values, while avoiding uneven features and irregularities (e.g. very spiky histograms).

3 Results

3.1 Calibration results

148 parameter sets were behavioral for deuterium simulations, with E_2 ranging from $L_2 = 0$ to 0.24. 181 parameter sets were behavioral for tritium simulations, with E_3 ranging from 0.24 T.U. to $L_3 = 0.5$ T.U. Additionally, 16 parameter sets were behavioral for both tritium and deuterium simulations, with E_2 ranging from $L_2 = 0$ to 0.19 and E_3 ranging from 0.36 T.U. to $L_3 = 0.5$ T.U. These solutions show that a reasonable agreement between the model fit to ^{2}H and the model fit to ^{3}H can be found.

The behavioral posterior parameter distributions constrained by deuterium or tritium or by both generally had similar ranges than their prior distributions, except notably for μ_2, θ_2, μ_3, and θ_3 (Table 2). To assess the reduction of parameter uncertainty, we calculated and compared the entropy of the prior and of the posterior distributions (Table 2). A visual inspection of the posterior distributions was also made, and we show here only the parameters μ_2, θ_2, μ_3, and θ_3 (Fig. 4) that directly control the range of ~~older water ages~~ longer travel times in streamflow, since they act mostly on the right-hand tail of the gamma components in Ω_Q. These parameters thus also have a direct influence on the catchment storage inferred via age-ranked storage S_T. The distributions of μ_2, θ_2, μ_3, and θ_3 are clearly not uniform. The distributions of the other parameters are provided as a supplement (Fig. S12-S13). Most distributions are not uniform, indicating that the parameters are identifiable.

Essentially, the results (Table 2 and Fig. 4) reveal that the parameter ranges decreased by adding information on ^{2}H or ^{3}H or both. This effect is particularly noticeable for f_0 and λ_1^*, which saw their upper boundary decrease, and for μ_2 and μ_3,

which saw their lower boundary increase considerably. These results also show that the posterior distributions depart from the uniform prior distributions when considering ^{2}H alone or ^{3}H alone (i.e. $\mathcal{H}(X|^iH) < \mathcal{H}(X)$ and $D_{KL}(X|^iH, X) > 0$ in Table 2). This effect is not very pronounced for most parameters, but clearly visible for λ_1^*, for μ_2 and μ_3 (e.g. uneven distributions of points in Fig. 4), and for μ_{ET}. The posterior distributions become considerably narrower when ~~considering both tracers~~ both tracers are considered, since $\mathcal{H}(X|(^2H \cap {}^3H))$ is much lower than $\mathcal{H}(X)$, which is visually represented by the distribution of points tending to cluster towards a corner in Fig. 4. Generally, more was learned about the likely parameter values by adding a constraint on ^{2}H simulations after constraining ^{3}H simulations than the opposite (i.e. generally $D_{KL}(X|(^2H \cap {}^3H), X|^3H) \geq D_{KL}(X|(^2H \cap {}^3H), X|^2H)$). Noticeable exceptions to this are the parameters μ_2, θ_2, and θ_3, which are more related to the ~~older ages~~ longer travel times in streamflow and to catchment storage than the other parameters.

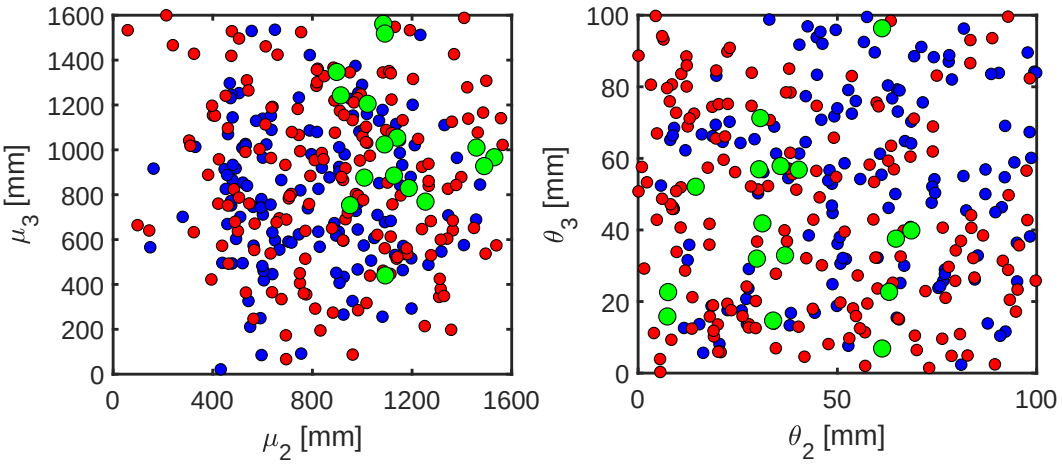


Figure 4. Distributions of SAS function mean (μ, left panel) and scale (θ, right panel) behavioral parameters directly controlling the selection of ~~older ages~~ longer travel times by streamflow, constrained by deuterium (148 blue dots), or tritium (181 red dots), or both (16 green dots).

Simulations of stream δ^2H captured both the slow and the fast dynamics of the observations when constrained by $E_2 > 0$ (blue bands and blue curve Fig. 5~~). This is not the case for~~ a), although some variability is not fully reproduced. The Nash Sutcliffe Efficiency (E_2) is limited to 0.24 despite visually satisfying simulations (Sect. 4.4.2). Most flashy responses in δ^2H ~~simulations constrained only by~~ (associated with flashy streamflow responses) were reproduced to some extent by the behavioral simulations (the very thin peaks of the blue bands in Fig. 5a, more visible in Fig. S1–S9). Nevertheless, about 3% of δ^2H data points were largely underestimated, pointing at a partial inability of the composite SAS functions to simulate the variability of the streamflow TTD (c.f. Sect. 4.4.2). Behavioral simulations selected using the other performance criterion instead ($E_3 < 0.5$ T.U.~~(red bands)~~, red bands in Fig. 5) did not match well the δ^2H observations overall. This shows that ^{3}H contains some information on travel times that is not in common with ^{2}H~~about the transport processes to the stream. Yet, ~~. Yet, these behavioral simulations are able to match all the observed δ^2H flashy responses in amplitude, suggesting that like

δ^2H, ^{3}H contains information on young water contributions to streamflow (Sect. 4.3). Also, δ^2H simulations constrained by both criteria (green bands) have a smaller variability than those constrained only by $E_2 > 0$, suggesting that ^{3}H ~~nevertheless~~ contains some information that is common with ^{2}H.

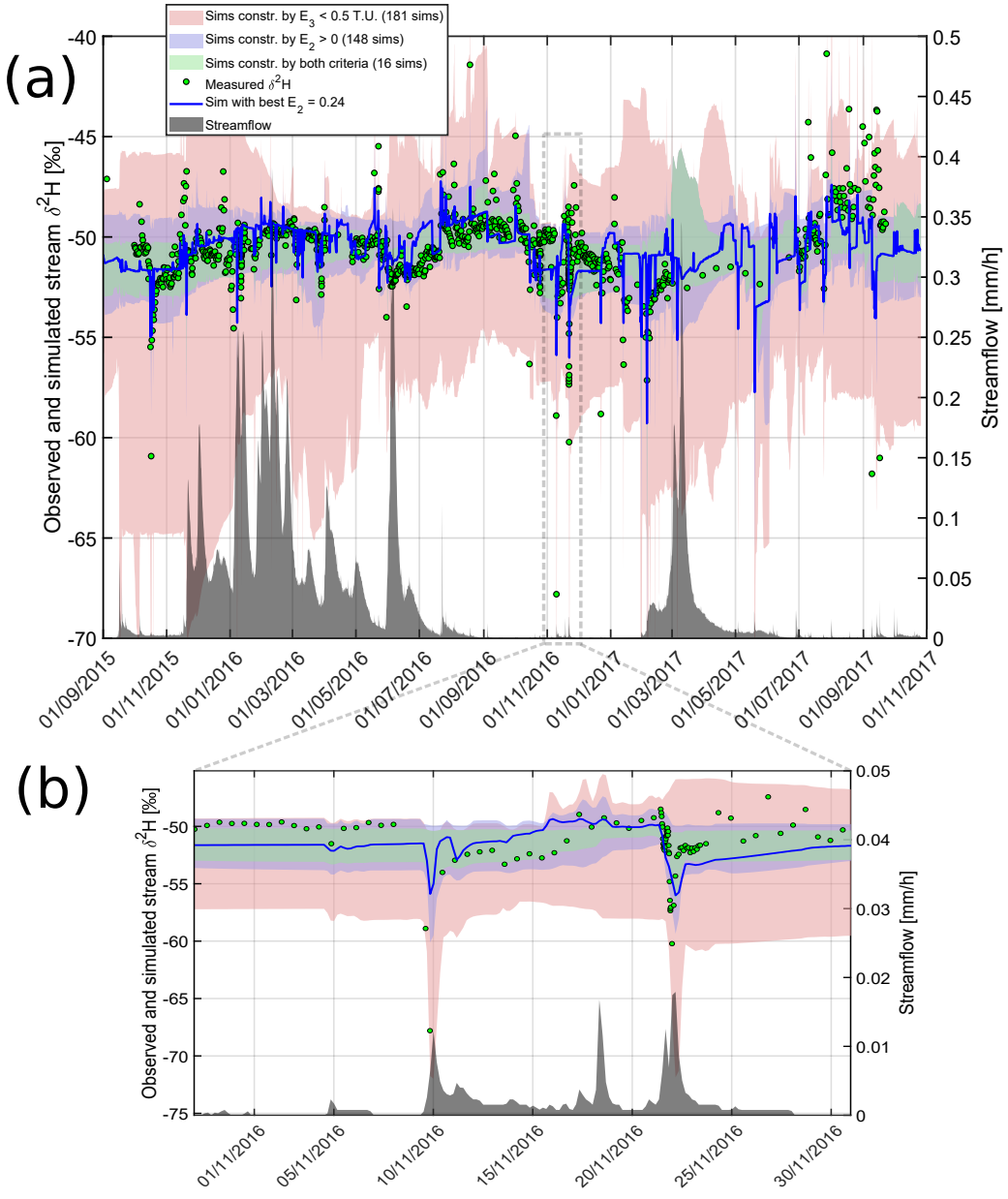


Figure 5. Simulations in deuterium. E_2 is the Nash-Sutcliffe efficiency in deuterium, and E_3 is the Mean Absolute Error in tritium units.

Simulations of stream ^{3}H generally matched the observations better in 2017 than before 2017 (red bands and red curve in Fig. 6). Some simulations (red bands) nevertheless matched the observations before 2017 relatively well. Similar to δ^2H simulations, both the slow and the fast simulation dynamics seemed necessary to reproduce the variability in ^{3}H observations (especially in 2017), although more stream samples would be needed to confirm that the model is accurate between the current measurement points. The higher stream ^{3}H values in 2017 that are better reproduced by the model correspond to an extended dry period during which streamflow responses are mostly flashy and short-lasting hydrographs. The ~~associated~~ ^{3}H values in 2017 are closer to precipitation ^{3}H, mostly around 10 T.U ~~.~~ (see also Fig. S15). The stream reaction to those higher values suggest a considerable influence of recent rainfall events on the stream, that steady-state TTD models relying only on tritium decay would probably struggle to simulate. This also suggests a stronger influence of old water in 2016 than in 2017 (see Sect. ~~4.4~~4.4.2). Simulations constrained by deuterium (blue bands) tended to overestimate stream ^{3}H. Simulations constrained by both criteria (green bands) worked well in 2017, but they overestimated stream ^{3}H before 2017. Similar to δ^2H simulations, this suggests that ^{2}H and ^{3}H have common but also distinct information contents ~~about~~ on transport processes to the stream. The tendency of the model constrained by deuterium and/or by tritium to overestimate the tritium content in streamflow suggests an non-negligible influence of the isotopic partitioning of inputs between Q and ET (Sect. 4.4.2, App. A2, and Fig. S15).

3.2 Storage and travel time results

For each behavioral parameter set, we calculated $\overleftarrow{P_Q}(T)$, the average ~~stream~~ streamflow TTD weighted by ~~streamflow~~ $Q(t)$ (over 2015–2017) in cumulative form (Fig. 7). Visually, there are no striking differences between $\overleftarrow{P_Q}(T)$ constrained by deuterium or by tritium, except a slightly wider spread for simulations constrained by tritium. The $\overleftarrow{P_Q}(T)$ constrained by both tracers clearly differ. The associated curves (Fig. 7c) show a much narrower spread. ~~They are~~ The travel time uncertainties are thus visually much lower than when using each tracer individually, highlighting the benefit of using both tracers together. The $\overleftarrow{P_Q}(T)$ constrained by both tracers are also slightly shifted towards higher ~~ages~~travel times. We calculated various statistics of the ~~distributions~~ $\overleftarrow{P_Q}(T)$ constrained by the different performance criteria to ~~compare them quantitatively~~ quantitatively compare the distributions (Table 3). This ~~shows~~ showed that the $\overleftarrow{P_Q}(T)$ constrained only by tritium systematically correspond to higher ~~ages~~ travel times (and lower young water fractions) than those constrained only by deuterium. ~~However, these age differences are small and could be explained by the uncertainties, which are larger for the younger age fractions, and systematically higher for tritium than for deuterium. The~~ A Wilcoxon rank sum test revealed that some statistically significant differences exist between the $\overleftarrow{P_Q}(T)$ ~~constrained by both tracers systematically correspond to the highest ages (and the lowest young water fractions). The corresponding uncertainties are much lower than when using individual tracers.~~ constrained by deuterium and the $\overleftarrow{P_Q}(T)$ constrained by tritium (App. B). Even if these differences are statistically significant, they remain lower than in previous studies (Sect. 4.1). In addition, the youngest water fractions and the oldest water fractions of $\overleftarrow{P_Q}(T)$ did not significantly differ according to the Wilcoxon rank sum test (App. B).

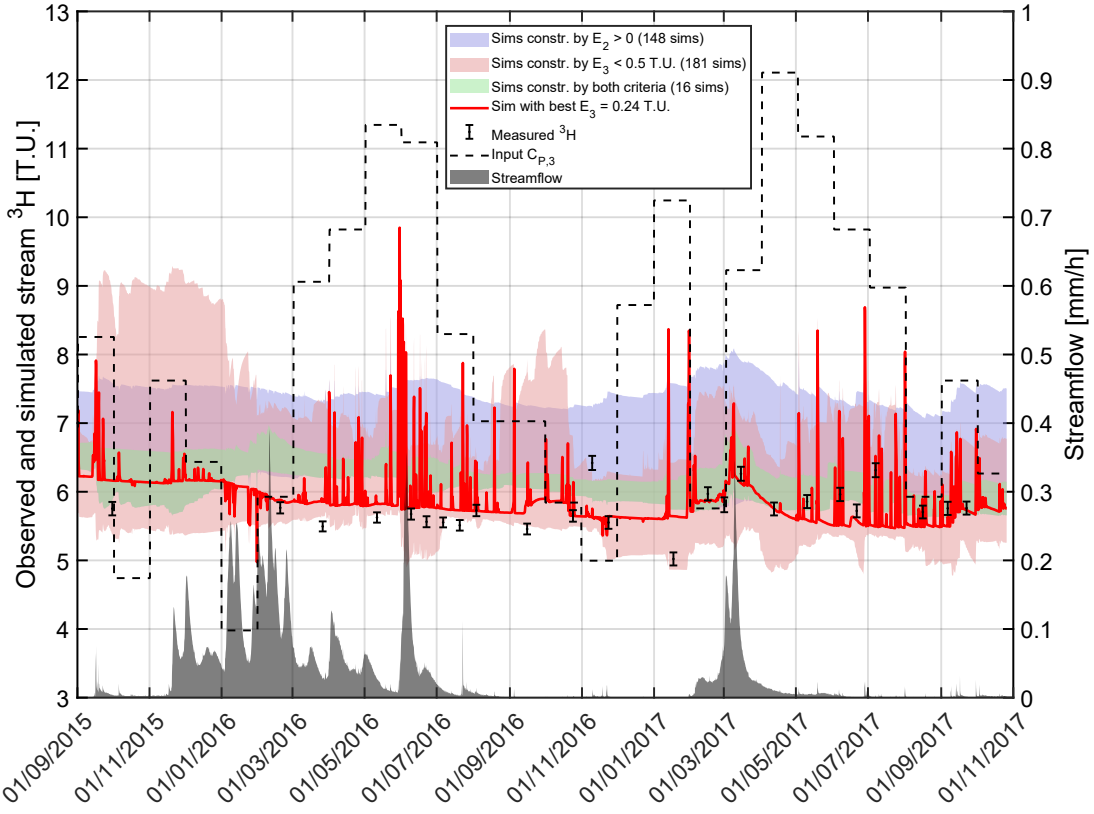


Figure 6. Simulations of stream concentrations in tritium compared to observations and to the variability in precipitation.

We defined the right-hand tail of the streamflow SAS function Ω_{tail} as the weighted sum of the two gamma components in Ω_Q:

$$\Omega_{tail}(S_T) = \frac{1}{\lambda_2 + \lambda_3^*}\left(\lambda_2\,\Omega_2(S_T) + \lambda_3^*\,\Omega_3(S_T)\right) \tag{11}$$

where $\lambda_3^* = 1 - \lambda_2 - \lambda_1^*$. Ω_{tail} thus ~~represents the right-hand tail of the SAS function Ω_Q, allowing~~ allows us to study in detail the asymptotic behavior of the function ~~in detail~~Ω_Q. In particular, this asymptotic behavior is time-invariant when plotted against S_T, because Ω_2 and Ω_3 are functions of S_T only. The behavioral parameter sets were thus directly used to calculate the curves $(S_T, \Omega_{tail}(S_T))$. These curves show similar differences for ^{2}H and ^{3}H than the curves $(T, \overleftarrow{P_Q}(T))$ (Fig. 8): a slightly wider spread is observed for Ω_{tail} constrained by tritium than deuterium (Fig. 8b), and the Ω_{tail} constrained by both tracers tend to converge to a narrow envelope of curves slightly shifted towards higher storage values (Fig. 8c).

To quantitatively study the implications of different Ω_{tail} for storage estimations, we computed statistics of a storage measure derived from these curves (Table 4). The 95th percentile of Ω_{tail}, called S_{95P} (black crosses in Fig. 8) allows for estimating

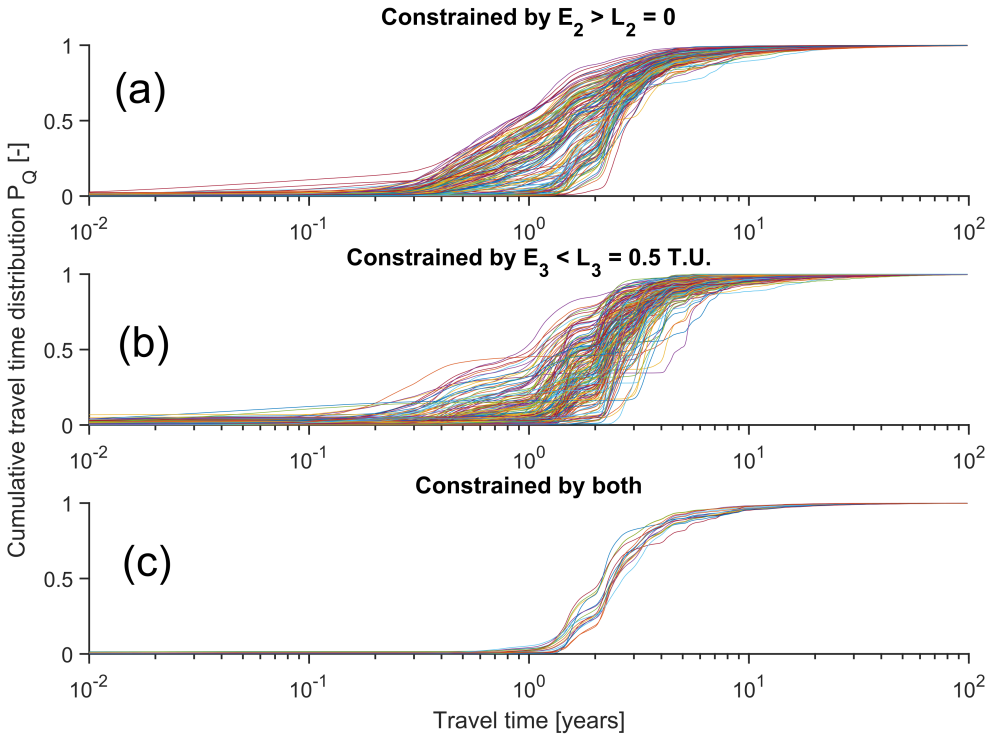


Figure 7. Flow weighted (2015–2017) cumulative stream TTDs for the behavioral parameter sets constrained by ^{2}H (a), by ^{3}H (b), and by both (c).

total mobile storage $S(t)$ from Ω_{tail}. In average, the Ω_{tail} constrained by tritium or by both tracers yielded significantly higher mobile storage $S(t)$ and smaller spread in $S(t)$ (Fig. 8~~and Table 4). Overall~~, Table 4, Table B1). Yet, the mobile storage $S(t)$ values estimated from the tracers are mutually consistent when considering the uncertainties.

4 Discussion

4.1 ~~Reconciliation of water ages~~ Consistency between TTDs derived from stable and radioactive isotopes of H

Our work shows that streamflow TTDs and the related catchment mobile storage $S(t)$ can ~~still~~ be estimated in unsteady conditions by using "ranked" SAS functions $\Omega(S_T,t)$ (Harman, 2015). Similar to Visser et al. (2019), we propose to coherently use the measurements of stream ^{2}H and ^{3}H to calibrate the parameters of the SAS functions, here defined in the age-ranked domain $S_T \in [0,+\infty[$ instead of the cumulative residence time domain $P_S \in [0,1]$. The calibrated tail of the streamflow SAS function Ω_Q (called here Ω_{tail}) could thus be used to approximate mobile storage $S(t)$ instead of defining the value a priori. The SAS functions also allowed us to estimate the unsteady TTDs defined in the ~~age~~ travel time domain T, and their statistics (mean, median, etc.). ~~Differences between~~ There were statistically significant differences between some TTD measures

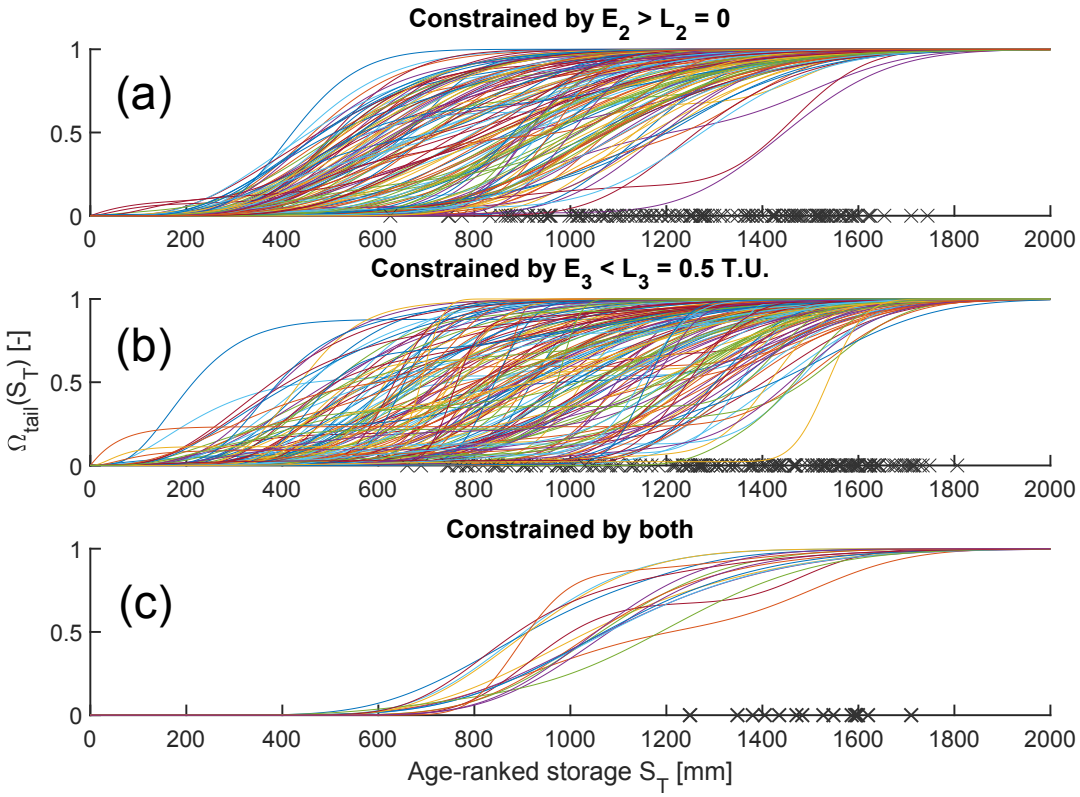


Figure 8. Cumulative right-hand tail Ω_{tail} of streamflow SAS functions for the behavioral parameter sets constrained by ^{2}H (a), by ^{3}H (b), and by both (c). Ω_{tail} is defined as the weighted sum of the two gamma components in Ω_Q. The black crosses indicate S_{95P} for each curve, i.e. the 95th percentile of Ω_{tail}

(e.g. mean, median) constrained by deuterium or by tritium (Wilcoxon rank sum test, App. B). Yet, the statistical significance may be questioned due to the contrasting number of ^{3}H samples (24) compared to δ^{2}H (> 1000), which is not accounted for in the statistical test. The Wilcoxon rank sum test only compares an equivalent number of accepted simulations (148 for deuterium against 181 for tritium) regardless of data considerations. The TTDs obtained from each tracer were broadly consistent in shape, and the travel time differences were considerably smaller (i.e., <1 yr) than in the previous comparison study (up to 5 yr, Stewart et al., 2010). This is particularly true for the MTT (only 8% difference in this study), which was the only travel time measure compared in the previous study (up to 200% difference in MTT for Stewart et al., 2010). In addition, our travel time differences were smaller for the 75th and 90th percentiles of the TTD than for the 10th and 25th percentiles. The 90th percentile differences were not statistically significant. This somehow contradicts previous statements (Stewart and Morgenstern, 2016) that tritium would reveal the long tails of the ~~various statistics of the TTDs were smaller than the uncertainties of the calculations when comparing the results obtained with~~ TTD which remain undetected by stable isotopes. Finally, our travel time differences

were smaller than the calculation uncertainties. The storage estimates derived from ^{2}H alone and with H or ^{3}H alone. Similarly, the derived storage estimates were consistent between ^{2}H and ^{3}H. The hypothesis of truncation of the TTD tails when using stable isotopesis therefore rejected at the present time based on the data from the Weierbach catchment. Moreover, stable isotopes do not seem to underestimate the related catchment storage compared to tritium. H were also statistically different but the differences were also small compared to the calculation uncertainties.

These findings were made possible results emerged for a number of reasons. 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 TTD). Even though we clearly distinguished tritium from deuterium by accounting for the relationship between water ages and tritium activities $C_{Q,3}$ (term $\exp(-\alpha T)$ in Eq. (3))Second, we did not use it directly to calculate T to derive the travel times solely based on the radioactive decay of tritium in order to avoid biases due to mixing of various ages at the outlet (Bethke and Johnson, 2008) and in order to avoid the age travel time ambiguity caused by tritium from nuclear tests (Stewart et al., 2012). AlsoMoreover, we did not use multiple control volumes having different TTDs determined by tracer measurements in their input and output (Małoszewski et al., 1983; Uhlenbrook et al., 2002; Stewart et al., 2007; Stewart and Thomas, 2008). This way, we avoided adding large uncertainties related to difficulties in characterizing end members and gathering representative samples (Delsman et al., 2013). SecondThird, we explicitly accounted for unsteady flow conditions, which has been done in only one other previous study using tritium (Visser et al., 2019). This allowed us to estimate realistic average TTDs corresponding to the catchment inflows, outflows, and internal flows that are highly time variant. ThirdFourth, our tritium stream sampling was not focused solely on baseflow hence not biased towards old water. FourthFifth, we considered the entire TTDs by using various percentiles and statistics, and not only the MTT which is highly influenced by the improbable extreme values of T. This means that even if there is water older than e.g. 1,000 years in streamflow, it can be neglected if it represents less than e.g. 0.000001 % of the volume. Finally, we explicitly accounted for parameter uncertainty. This is important because absolute values without an uncertainty estimate cannot be reliably interpreted.

4.2 Yet, Does tritium seems to reveal help revealing the presence of older water!?

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 travel time and storage estimates (tables 3 and 4). The hypothesis of different transport velocities between Isotopic effects on the transport of water molecules containing deuterium and water molecules containing tritium can be rejected, because their self diffusion or tritium (i.e., on different isotopologues) seem insufficient to explain these travel time differences, because the self diffusion of these isotopologues in water are nearly equal (Devell, 1962), and their advective velocities are the same. On the one hand, However, flow paths in the relatively small Weierbach catchment are probably too short to allow travel time differences due to isotopic effects on self diffusion coefficients.

It seems likely that the higher storage, the slightly higher ages derived from tritium seem related to the apparent absence of responses of stream ^{3}H to the high precipitation ^{3}H in 2016, indicative of the dominance of old water compared to 2017. On

~~the other hand, tritium~~ higher travel times, and the larger uncertainties for tritium are related to the lack of high-resolution data. Tritium simulations included many small peaks corresponding to flashy streamflow responses associated with young water (Fig. 6). Only ~~some of those simulated~~ few simulated flashy peaks could be confirmed by the presence of stream ~~measurements at those times~~ ^{3}H measurements, especially in 2016. More stream ^{3}H samples during ~~these flashy~~ flashy ^{3}H events would probably ~~support even further~~ further validate these simulations of young water in streamflow and shift the TTDs constrained by tritium towards younger water. ~~We thus interpret the observed small age differences rather as the consequence of a~~ This is consistent with the larger travel time differences found for the 10th and 25th percentiles of the TTDs. The limited tritium sampling resolution (bi-weekly) ~~that~~ covered most of the flow probabilities (Fig. 3) but it may still be slightly biased towards hydrological recessions during which the youngest water fractions are absent by definition ~~.~~ (Sect. 4.4.3). Tritium and stable isotopes of O and H ~~sampled synchronously~~ synchronously sampled at high resolution would ~~thus~~ pave the way for further research on stream water ~~ages~~ travel times from a multi-tracer perspective.

~~It is also interesting that the age~~ The travel time and storage measures estimated from a joint use of ^{2}H and ^{3}H are the highest (tables 3 and 4). ~~In the end, tritium may~~ Tritium may thus have helped revealing the presence of old water in streamflow. However, it did so only when combined with deuterium. It is commonly assumed that ^{3}H ~~is more informative about~~ carries more information on old water because of radioactive decay that relates lower tritium activities to increasing ~~water ages~~ travel times (Stewart et al., 2010). However, as shown by Stewart et al. (2012) and in Fig. 2, current tritium values of the water recharged in 1980–2000 are similar to the tritium values of the water recharged today. Thus, the younger water disrupts the relationship between ~~water age~~ travel times and tritium values. ~~Adding supplementary information about the younger water in the calibration with the high frequency ^{2}H measurements may have partly helped "filtering" the currently complex relationship between water ages and tritium values, leveraging the potential of tritium for revealing the tail of the TTDs. The fact that water ages~~ It seems that using the high frequency δ^2H measurements reduced the ambiguity of tritium-derived travel times by helping to discriminate young and old water contributions to streamflow. The travel times being below ~5 years in the Weierbach ~~are limited to about 5 years~~ (Table 3) could be another reason for the limited information of ^{3}H ~~about~~ 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 limited residence times, radioactive decay may ~~only~~ give information that is redundant with the natural variability of the tracer in precipitation. 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 $\log(^3H)$ of stored water may increase linearly with ~~water age~~ 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 ~~age~~ 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.

4.3 ~~Age~~ Travel time information contents of stable and radioactive isotopes

The ~~fact that we found equal~~ similar travel time and storage measures when using ^{2}H alone or ^{3}H alone ~~does~~ do not mean that it is not worth sampling both. ~~Our results show that more information was learned about storage and travel times (all the $D_{KL} > 0$)by using both tracers together, which~~ Combining the tracers yielded a non-negligible information gain of ~10% of the initial $\mathcal{H}(X)$ for most parameters. In total, 12.7 bits of information on travel times were learned by combining the two tracers. This is more than twice the amount learned from each individual tracer (around 4 bits, see paragraph below). This amount of information can be calculated for a given tracer by summing $D_{KL}(X|(^2H \cap {}^3H), X)$ for all parameters (see Sect. 2.7, Table 2). Combining the tracers also resulted in lower uncertainties (lowest entropy $\mathcal{H}(X|\,(^2H \cap {}^3H))$ in Table 2, narrower groups of curves in Fig. 7 and 8, lower standard deviations in tables 3 and 4). This information gain on travel times was possible because ~~the~~ composite SAS functions (Eq. (5)) allowed us to ~~independently constrain different parts~~ constrain three nearly-independent components (Ω_1, Ω_2, Ω_3) of the same streamflow TTD with one tracer or the other~~, reducing~~. This reduced the potential trade-offs between the shapes suggested by one tracer or the other. ~~In addition, the streamflow TTD was constrained using only stream samples. On the contrary, Stewart et al. (Table I, 2010) showed three studies where multiple TTDs corresponding to different end members (e.g. surface runoff, groundwater) are constrained by tracers sampled in the associated outlets. Although reasonable fits were shown for the samples from the different end members, the fit of the combined TTD for the stream samples was not systematically checked (Uhlenbrook et al., 2002; Stewart et al., 2007; Stewart and Thomas, 2008)~~ .

~~For future studies it is worth mentioning the amount of information gained per isotopic sample or per euro invested in sample analysis. This amount of information can be calculated for a given tracer by summing for all parameters~~ These three components are formally related only by the ~~Kullback-Leibler divergences D_{KL} (see Sect. 2.7) between the prior and the posterior parameter distributions. With deuterium~~ requirement to have $\lambda_1(t) + \lambda_2(t) + \lambda_3(t) = 1$. Thus all their other parameters are independent.

With deuterium alone, we learned ~~13.55~~ 4.08 bits of information with 1385 samples~~, representing about 9.79×10^{-3} bits per sample or about 9.79×10^{-4} bits per euro~~. With tritium, we learned ~~14.85~~ 4.47 bits of information with only 24 samples. ~~We thus have a much higher relative information content of 0.619 bits per sample. However tritium analyses are more expensive, so the information content is only 1.44×10^{-3} bits per euro. It should be noted that for tritium the precipitation samples were not included in this cost as they were analyzed by the IAEA. Thus tritium~~ Thus, tritium was overall more informative than deuterium about ~~water ages, and it was also more cost-effective. One reason for this is that~~ travel times, even with a lower number of samples. This is because tritium considerably informed us about the travel times in ~~ET because it~~ ET. Tritium constrained the posterior of μ_{ET} ~~well~~ even better than deuterium (Table 2)~~that controls directly the ages in ET.~~ . The large information gains on μ_{ET} and θ_{ET} (especially with tritium) reveal a non-negligible influence of Ω_{ET} on the accuracy of stream ^{3}H simulations, via an indirect influence on isotopic partitioning (App. A2). This also highlights the importance of ~~considering explicitly ET~~ explicitly considering ET in streamflow travel time calculations (van der Velde et al., 2015; Visser et al., 2019). However, ^{2}H resulted in lower uncertainties for nearly all other parameters (e.g. lower Shannon entropy $\mathcal{H}(X|\,^2H)$, Table 2).

This is most likely due to the much higher sampling frequency for deuterium that allows for constraining the simulations better than with bi-weekly tritium measurements (see the simulation envelopes Fig. 5 and 6). From our experience in the Weierbach catchment, we estimate that for ^{2}H, a weekly sampling to cover the damped variations of δ^2H (i.e. about 100 samples over 2015–2017) complemented with an event-based high-frequency sampling (every 15 hours) of the flashy responses (i.e. about 300 samples over 2015–2017) could have given us as much information as the complete time series. This suggests that a more strategic sampling of ^{2}H ~~could~~ may outperform ^{3}H~~in terms of cost-efficiency~~. The amount of information learned from the isotopic data ~~probably scales~~ necessarily grows with an increasing number of samples. Yet, we do not know whether it scales linearly or non-linearly and ~~probably~~ whether it quickly reaches a plateau as the number of observation points grows ~~.~~ (Fig. S14). In the future, it would be useful to further use information theory (e.g. entropy conditional on sample size) to know how how this information scales and how many measurements are enough and when to sample isotopes for maximum information gain on ~~water ages~~travel times. This would imply artificially re-sampling a higher-frequency isotopic time series using various strategies (e.g. Pool et al., 2017; Etter et al., 2018) and re-calibrating the model many times, which would involve much subjectivity and come with an exorbitant computational price.

~~In the end~~Overall, stable and radioactive isotopes of H ~~have~~ had different information contents ~~. For example, they lead to different Shannon entropy $\mathcal{H}$ for the posteriors.Also, the Kullback-Leibler divergence~~ on travel times. The positive D_{KL} ~~was never 0, indicating that adding one tracer after the other still allowed us to learn something about parameter values. Finally~~values are not simply due to different performance measures for deuterium and for tritium (c.f. Table S1) but due to non-redundant information contents on travel times for each tracer. Performance measures E_2 and E_3 are both based on minimizing a sum of residuals and thus do not considerably influence what can be learned from tracer data (c.f. Table S2). Moreover, the parameters corresponding to the best simulations in ^{2}H did not correspond to those for ^{3}H and vice versa. ~~Our results suggest that ^{3}H is more informative about old water thanks to its radioactive decay.~~ Yet, stable and radioactive isotopes ~~have~~ had some information in common ~~about~~ on young water. ~~For example, both ^{2}H and ^{3}H stream samples showed reactions to precipitation ^{2}H and ^{3}H values during flashy streamflow events, revealing the role of young water during these events. This was previously unobserved for tritium~~ This is consistent with early tritium studies that tried to show its potential for detecting young water contributions to streamflow (Hubert et al., 1969; Crouzet et al., 1970; Dincer et al., 1970; Klaus and McDonnell, 2013). This has been overlooked in recent travel time studies because of the sampling focused on periods outside events (Stewart et al., 2010). The theoretical span of 0–4 years pointed out in Stewart et al. (2010) should however not be taken as the only range of ~~ages~~ travel times where ^{18}O, ^{2}H, and ^{3}H may have redundant information. As clearly written ~~in~~ 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). Finally, stable and radioactive isotopes had some information in common on old water as well. This is clearly shown by the increased travel time and storage measures when both tracers are used, which also highlights that they can give similar results.

4.4 Limitations and way forward

4.4.1 Hydrometric- versus tracer-inferred storage

The storage value derived from unsteady travel times constrained by tracer data (Table 4, ∼1200–1700 mm) is noticeably larger than the maximum storage (≃ 250 mm) estimated from point measurements of porosity and water content (Martínez-Carreras et al., 2016), from water balance analyses (Pfister et al., 2017), water balance analyses combined with recession techniques (Carrer et al., 2019), and from ~~the values used in~~ a distributed hydrological model ~~(≤ 700 mm, Glaser et al., 2016)~~ (≤ 700 mm, Glaser et al., 2016, 2020). Our storage value is more consistent with the ∼1600 mm derived from depth to bedrock and porosity data used for the Colpach catchment (containing the Weierbach) that was modeled with CATFLOW (Loritz et al., 2017). Large differences between hydrometrically-derived and tracer-derived storage estimates are not uncommon (Soulsby et al., 2009; Fenicia et al., 2010; Birkel et al., 2011) and in fact highlight the ability of tracers to reveal the existence of stored water that is not directly involved in streamflow generation (Dralle et al., 2018; Carrer et al., 2019). This "hydraulically disconnected" storage is nevertheless important to explain the long residence times in catchments (Zuber, 1986). More research is ~~thus~~ needed for improving the conceptualization of storage and unifying storage terminology and the various estimates obtained from tracers or other techniques. The storage value we found is not in complete contradiction with the previous estimates if we consider their uncertainties. Hydrological measurements (J, Q, and especially ET) are highly uncertain (Waichler et al., 2005; Graham et al., 2010; Buttafuoco et al., 2010; McMillan et al., 2012; McMahon et al., 2013) and their errors are accumulated in long term water balance calculations. An explicit consideration of those uncertainties in the future could reconcile the different storage estimates. Furthermore, it is worth remembering that simplifying storage from a complex spatially-distributed quantity to a simple compact 1D water column neglects the importance of subsurface heterogeneity, surface topography, and bedrock topography for the storage and release of water. As a result, upscaling local point measurements of storage capacity that are not representative of the whole subsurface is very likely to under or overestimate the true storage capacity of the whole catchment. This is even more true if the new techniques used to scan the subsurface over larger areas such as Electrical Resistivity Tomography (ERT) are themselves associated with uncertainties, requiring adaptations (Gourdol et al., 2018) and site-specific independent knowledge (Parsekian et al., 2015).

4.4.2 Model performance and uncertainty

Our conclusions ~~rest on the assumption that the model captures the water ages~~ are valid because the model captures accurately the travel times in the Weierbach~~accurately, which was validated~~. This was confirmed by the acceptable performance of the simulations, especially visually. Still, the performance in δ^2H or in ^{3}H could be improved in the future by testing other models of composite SAS functions. The best NSE for deuterium simulations (~~called~~ E_2) was 0.24, which is lower than ~~the values reported in a number of studies~~ several other using SAS functions (van der Velde et al., 2015; Harman, 2015; Benettin et al., 2017b). ~~It should be pointed out again the NSE may not be the most appropriate objective function to characterize performance against the~~ E_2 is penalizing for the δ^2H time series ~~from the Weierbach (see Rodriguez and Klaus, 2019). Future work could look for~~ in the Weierbach because the observed stream δ^2H has many more points corresponding to damped

seasonal fluctuations (Fig. 5a) compared to the large flashy fluctuations (Fig. 5b). E_2 also overemphasizes the timing errors, even if the shape of the simulation is perfect (Klaus and Zehe, 2010; Seibert et al., 2016). In addition, E_2 is not an absolute measure of model performance allowing comparisons between different studies (Seibert, 2001; Schaefli and Gupta, 2007; Criss and Winsto . Future work needs to develop more appropriate objective functions for δ^2H, especially with respect to the information gained from model calibration. This implies accounting for expert knowledge, intuition, and visual experience with simulations in a customized performance measure (Ehret and Zehe, 2011; Seibert et al., 2016), or finding an adequate benchmark model for δ^2H (Schaefli and Gupta, 2007), or correctly defining the statistical properties of the model errors (Schoups and Vrugt, 2010) . The best MAE for tritium simulations (called E_3) was 0.24. This is slightly higher than values of RMSE (close to 0.10) reported in a number of studies using tritium (Stewart et al., 2007; Stewart and Thomas, 2008; Duvert et al., 2016). However these studies had only a few stream samples, while Gusyev et al. (2013) report for instance a RMSE of 1.62 T.U. for 15 stream samples. Stream δ^2H seems to suggest larger fraction of young water than the simulations (c.f. underestimation of ~~many~~ flashy events in Fig. 5). Stream ^{3}H data seems to suggest larger fractions of old water than the simulations (c.f. overestimation of tritium activities over March–September 2016 in Fig. 6). A model passing through all observation points may thus show larger differences between the TTDs constrained by deuterium and the TTDs constrained by tritium. ~~However, there are not enough~~ It is important to recall that there are less ^{3}H stream samples compared to ^{2}H, ~~so~~ thus a comparison of the TTDs from this hypothetical ideal model could be misleading. Furthermore, the different scaling for the units for δ^2H and ^{3}H may also mislead the visual comparisons and interpretations on young water contributions based on the different amplitude of flashy tracer responses. We believe that a higher resolution of stream ^{3}H would unambiguously show the potential of tritium for revealing young water in the stream, as shown in the early tritium studies (Hubert et al., 1969; Crouzet et al., 1970; Dinçer et al., 1970).

The simulations in deuterium were better for decreasing δ^2H than for increasing δ^2H (better simulations of the flashy events in δ^2H pointing downwards, Fig. 5). This is probably because the increases in δ^2H generally correspond to drier periods, during which $C_{Q,2}$ starts reacting stronger to $C_{P,2}$ indicating that young water fractions (controlled by $\lambda_1(t)$ in the model) are higher than expected. $C_{P,2}$ can explain only about 30% of the variations of $C_{Q,2}$, but this can increase to 44% during drier periods (Fig. S10 and S11). The low explanatory power of $C_{P,2}$ is linked to the larger influence of groundwater for streamflow responses in the Weierbach (conceptualized with Ω_2 and Ω_3 having larger weights λ_2 and λ_3). During drier periods, we expect an increase in the non-linearity of the processes delivering young water to the stream. For example, the decreasing extent of the stream network and of saturated areas observed in the Weierbach during drier conditions ~~(??) is~~ (Antonelli et al., 2020a, b) is likely caused by decreasing groundwater levels (Glaser et al., 2020) and it could reduce the amounts of young water reaching the stream (c.f. van Meerveld et al., 2019). However, streamflow is lower during drier conditions, so the fractions of young water can still increase because of a less pronounced dilution of the young water in streamflow compared to wet periods. On the other hand, preferential flow observed in the soils of the Weierbach catchment and in the direct vicinity (Jackisch et al., 2017; Angermann et al., 2017; Scaini et al., 2017, 2018) may become more relevant during drier conditions and could increase the amount of young water contributing to streamflow, especially because precipitation intensities can be much higher in summer (due to thunderstorms) than in winter. The parameterization of ~~the~~ streamflow SAS functions via $\lambda_1(t)$ (Eq. (A5)) includes—to some extent—the effect of wet vs. dry conditions and the role of precipitation intensity, but it seems not to fully capture

how these factors influence young water fractions in the stream. Testing other parameterizations of $\lambda_1(t)$ or including ~~other~~ additional information such as soil moisture or groundwater levels in the current parameterization of $\lambda_1(t)$ may improve the simulations. Finally, the uncertainty of precipitation δ^2H could be higher during drier periods, because precipitation amounts can be too small (e.g. < 1 mm) over several weeks or because the precipitation intensities can be too high (e.g. > 5 mm/h) to be captured efficiently by the sequential rainfall sampler. This may lead to inaccuracies in the input data and thus to the inability of the model to simulate the corresponding flashy events in stream δ^2H. The representation of precipitation δ^2H ~~could thus~~ should be improved in the future by using more recent sampling techniques (e.g. Michelsen et al., 2019).

The ~~simulations~~ tendency of the model to yield higher average tritium values than the observations in streamflow over 2015–2017 (Fig. 6) and lower average tritium values than precipitation (see Fig. S15 where this is more visible) seems related to either not enough tritium residing in storage or removed by ET. The latter mechanism is only indirectly controlled by Ω_{ET} which loosely acts on the isotopic partitioning between Q and ET (App. A2). Unfortunately, no tracer data in ET can be used to close the tracer mass balance and to draw firm conclusions on the correct mechanism. In any case, an accumulation of tritium in storage to decrease the average stream tritium content is not a realistic behavior in the long term. The average stream ^{3}H is higher for the simulations constrained by $E_2 > L_2$ than $E_3 < L_3$ probably because of the lower resolution of ^{3}H measurements. The simulations overestimated ^{3}H in the stream particularly in 2015–2016 compared to 2017 (Fig. 6). In 2017 the simulations were better because the model used more ~~of the~~ young water (~~<~~ <7 days old, using Ω_1) to simulate the variability and the higher values of stream ^{3}H than in 2016. The lower ^{3}H in 2015–2016 could be caused by an increased ~~age~~ travel time in the older water components in 2015–2016 compared to 2017, due to changes in the importance of different subsurface flow paths in the Weierbach caused by a wetter period. The old water components Ω_2 and Ω_3 (Eq. (5)) represent subsurface ~~flows~~ flow paths likely occurring in the lower soils and following bedrock topography (Glaser et al., 2016; Rodriguez and Klaus, 2019) and potentially in weathered bedrock fractures (Scaini et al., 2018) or in the bedrock (Angermann et al., 2017; Loritz et al., 2017) . We used functions of S_T only for ~~these components~~Ω_2 and Ω_3, meaning that the ranges of ages they select do not change considerably with time (because the distribution of S_T is rather stable). Including ~~explicitly a~~ an explicit dependence on time for Ω_2 and Ω_3 could help to better represent ~~e.g. the fracture flows or deep groundwater flows~~ deeper flow paths in the catchment and improve ^{3}H simulations in 2015–2016. Eventually, the monthly resolution of ^{3}H in precipitation is coarser than the biweekly sampling in the stream, which can hinder accurate simulations. An increase in sampling resolution of tritium in ~~the stream to better constrain the TTDs in the future will~~ precipitation will be necessary in the future (Rank and Papesch, 2005) .

Finally, parameter distributions (Fig. 4 and S12-S13) and information measures (Table 2) suggest that some parameters are not strongly constrained by tracer data (but they are not unidentifiable either). This may result from the larger number of parameters than traditional SAS functions. Nevertheless, all these parameters are necessary to represent the array of non-linear and time-varying processes leading to the selection of particular ages from storage (numerically represented by $\sim 10^5$ control volumes) to generate both outflows Q and ET. This is essential to not neglect certain travel times that may become important for accurate water chemistry simulations (Rodriguez et al., 2020). Other methods to explore parameters

(using Markov Chains) such as DREAM (Vrugt, 2016) or PEST (Doherty and Johnston, 2003) could yield narrower posterior distributions. Nevertheless, these more advanced algorithms would need to be ~~followed by a considerable increase of sampling resolution in precipitation (Rank and Papesch, 2005).~~ adapted to allow parameter constraints, numerically-diverging solutions (typically for randomly selected combinations of parameters values that are incompatible), and multi-objective calibration.

~~Although we~~

4.4.3 Data constraints

The highest flows that were not sampled for tritium (Fig. 3) represent about 50% of the water that left the catchment via streamflow over 2015–2017. The high flows are mostly "second" delayed streamflow peaks in this catchment where double-peaked hydrographs occur in wet conditions (Sect. 2.1). Previous studies in the Weierbach using various tracers suggest that second peaks are likely composed of older water than first peaks (Wrede et al., 2015; Martínez-Carreras et al., 2015). Nevertheless the high flows in the second peaks may be associated with shorter travel times than low flows. Loritz et al. (2017) described the subsurface of the Weierbach catchment as highly permeable and hypothesized that it is able to rapidly transmit large amounts of young water during high streamflow events. This may explain the higher tritium-derived travel times due to the limited ^{3}H sampling in this study (e.g., 25% difference in median travel time). For deuterium, the highest flows are associated with 40 samples (about 4% of the samples) which represent about 20% of the water leaving via streamflow over 2015–2017 (Fig. 3). An adaptive sampling frequency based on accumulated flows (e.g., one sample every dozen m^3) could improve the representativity of the samples with respect to the flow volumes. This would not improve the results because the TTDs already account for the flow volumes by definition and because the larger water mass not sampled for tritium is not creating 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. Flow-proportional sampling would also lead to a much larger number of samples, rapidly exceeding the current field and laboratory capacities. This is why nearly-continuous in situ measurements would be preferable (e.g., Pangle et al., 2013; von Freyberg et al., 2017). Nevertheless, in situ measurements are currently not available for tritium.

We found much lower deviations for the ~~age~~ travel time and storage measures constrained by deuterium and tritium together (tables 3 and 4). However, it has to be acknowledged that ~~this is also because~~ there are only few accepted solutions (16), while there about 10 times more when using ^{2}H alone or ^{3}H alone. ~~Yet~~We should expect a higher standard deviation due to a lower number of accepted solutions to calculate this statistic using both tracers. On the contrary, the associated ~~curves~~ TTDs (Fig. 7~~and 8~~c and 8c) fall close to each other, ~~so the lower deviations have to be due also to~~ resulting in lower deviations that clearly point to lower uncertainties. A lower number of accepted solutions is in the end inevitable as it is an inherent consequence of using several performance measures independently as opposed to using a combined objective function (e.g. Hrachowitz et al., 2013; Rodriguez et al., 2018). Fewer accepted simulations are also an advantage to identify behavioral parameter sets (Klaus and Zehe, 2010). Less strict threshold criteria for behavioral solutions could increase the number of accepted solutions but they would accept less accurate simulations, which could lead to misleading conclusions. More stream ^{3}H measurements

would on the other hand allow the use of more advanced objective functions, which could lead to more accepted solutions. ~~Eventually, the~~ The input data measured over 2010–2017 and used to spin up the model from 1960 to 2010 (J, ET, Q, and $C_{P,2}$) could be unrepresentative of the real hydrometeorological and isotopic conditions of 1960–2015 due for instance to nonstationarity or climate change. These changing conditions could affect the modeled residence times in storage and thus the estimated streamflow travel times (Wilusz et al., 2017). Different methods to spin up the model could be tested in the future (Hrachowitz et al., 2011), especially to assess the effect the effect of changing hydrometeorological and isotopic conditions on the estimation of travel times. For this, isotope tracer records that span several decades like the ones that can be reconstructed from pearl mussels shells (Pfister et al., 2018, 2019) represent a crucial asset. Eventually, the precipitation tritium samples were taken about 60 km away from the catchment and may introduce some uncertainty.

5 Conclusions

Stable isotopes of O and H and tritium are indispensable tracers to infer the streamflow TTD and derive storage estimates in catchments. Our study addressed an emerging concern about the possible ~~deficiency~~ limitations of stable isotopes to infer the whole streamflow TTD compared to tritium. We went beyond previous data and methodological limitations and ~~thus~~ we did not find that stable isotopes are blind to old water fractions as suggested by earlier travel time studies. We found statistically significant differences between some travel times measures derived from each tracer, but these differences were considerably smaller than in previous studies. The differences we found can most likely be attributed to a higher number of stable isotope samples compared to tritium due to different analysis techniques. Based on the results in our experimental catchment in Luxembourg~~. However, we found that stable isotopes and tritium do~~ , we conclude that the perception that stable isotopes systematically truncate the tails of TTDs is not valid. Instead, our results highlight that stable isotopes and tritium have different information contents on ~~water ages~~travel times but they can still result in similar TTDs. In fact, inferring the streamflow TTD from a joint use of both tracers better exploits their ~~respective age~~ information contents, which results in lower uncertainties ~~. Even if~~ and higher information gains. Although ^{3}H appeared to be slightly more ~~cost-effective and~~ informative than ^{2}H ~~, a smart sampling~~ even with fewer samples, a different sampling strategy of the stable isotopes could outperform tritium. Future work could additionally compare streamflow TTD and storage from the two tracers in larger catchments where older water is expected, to give tritium more time to decay and better leverage its ability to point the presence of very old water out. We therefore recommend to: (1) keep sampling tritium in as many places as possible, as emphasized by Stewart et al. (2012); but also (2) to sample tritium at the highest frequency possible and synchronously with stable isotopes if possible. This is particularly important for the isotopic measurements in precipitation that drive all model simulations, regardless of functional forms of TTD and their parameter values. Overall this work shows that more tracer data is naturally better to gather more information about the catchments functions of storage and release.

Data availability. The tritium input data until 2016 used in this study can be obtained from the WISER database portal of the International Atomic Energy Agency (values for 2017 will be accessible there too in the future, please ask Axel Schmidt from Bundesanstalt für Gewässerkunde in the meantime). The rest of the data used in this study is the property of the Luxembourg Institute of Science and Technology (LIST) and can be obtained by request to the corresponding author after approval by LIST.

Appendix A: Model equations

A1 Parameterization of the SAS functions

In this section we provide further details on the equations used in the model. The composite streamflow SAS function Ω_Q used in this study is:

$$\Omega_Q(S_T,t) = \lambda_1(t)\,\Omega_1(S_T) + \lambda_2(t)\,\Omega_3(S_T) + \lambda_3(t)\,\Omega_1(S_T) \tag{A1}$$

$\Omega_1(S_T)$ is a cumulative uniform distribution for S_T in $[0, S_u]$, where S_u (mm) is a calibrated parameter representing the amount of stored young water potentially contributing to flashy streamflow responses. Thus:

$$\Omega_1(S_T) = \begin{cases} \frac{S_T}{S(t)}, & S_T \in [0, S_u] \\ 1, & S_T > S_u \end{cases} \tag{A2}$$

$\Omega_2(S_T)$ and $\Omega_3(S_T)$ are direct functions of S_T and are gamma-distributed:

$$\Omega_2(S_T) = \frac{1}{\Gamma(\frac{\mu_2}{\theta_2})}\,\gamma(\frac{\mu_2}{\theta_2}, \frac{S_T}{\theta_2}) \tag{A3}$$

$$\Omega_3(S_T) = \frac{1}{\Gamma(\frac{\mu_3}{\theta_3})}\,\gamma(\frac{\mu_3}{\theta_3}, \frac{S_T}{\theta_3}) \tag{A4}$$

where Γ is the gamma function, γ is the lower incomplete gamma function, μ_2 and μ_3 (mm) are mean parameters (calibrated), and θ_2 and θ_3 (mm) are scale parameters (calibrated).

$\lambda_1(t)$, $\lambda_2(t)$, and $\lambda_3(t)$ sum to 1. These are simply time-varying weights giving each component (i.e. c.d.f. Ω) a dynamic contribution to streamflow generation. In particular, $\lambda_1(t)$ is made highly time-variant to represent the flashy hydrographs that have an on-off type of response to precipitation. $\lambda_2(t)$ is considered constant and calibrated to keep the parameterization parsimonious. $\lambda_3(t) = 1 - \lambda_2 - \lambda_1(t)$ is deduced by difference for parsimony as well. Since $\Omega_1(S_T)$ represents young water contributions and previous studies in the Weierbach showed that event water contributions depend on the catchment wetness

and on precipitation intensity (Wrede et al., 2015; Martínez-Carreras et al., 2015), $\lambda_1(t)$ was parameterized using storage $S(t)$ and a proxy storage variations $\overline{\Delta S(t)}$ (see Rodriguez and Klaus (2019) for more details):

$$\lambda_1(t) = \lambda_1^* \left[f(t) + (1 - f(t))\, g(t) \right] \tag{A5}$$

where $\lambda_1^* \in [0,1]$ (no units) is a calibrated parameter representing the maximum value of $\lambda_1(t)$, and $f(t) \in [0,1]$ and $g(t) \in [0,1]$ are given by:

$$f(t) = f_0 \left(1 - \tanh \left[\left(\frac{S(t)}{S_{min} + S_{th}} \right)^m \right] \right) \tag{A6}$$

$$g(t) = 1 - \exp\left(- \frac{\overline{\Delta S(t)}}{\Delta S_{th}} \right) \tag{A7}$$

$f_0 \in [0,1]$ (no units) is a calibrated parameter guaranteeing a minimum for $\lambda_1(t)$ during dry periods~~,~~; $S_{min} = \min(S(t))$~~,~~; and S_{th} (mm, calibrated parameter) is a storage threshold relative ~~the to~~ to the minimum storage S_{min} separating wet ($S(t) > S_{min} + S_{th}$) from dry periods ($S(t) < S_{min} + S_{th}$). $m = 1000$ is a fixed parameter used to smooth the function f with respect to $S(t)$. $\overline{\Delta S(t)}$ is a proxy of storage variations calculated as a moving average of storage variations over a time window $\Delta t^* = 2\,\Delta t$:

$$\overline{\Delta S(t)} = \max \left(\frac{1}{3} \sum_{j=0}^{2} \Delta S(t - j\Delta t), 0 \right) \tag{A8}$$

with $\Delta S(t) = \Delta t\,(J(t) - Q(t) - ET(t))$. $\overline{\Delta S(t)}$ essentially increases during precipitation events and decreases when $Q(t)$ or $ET(t)$ are high. ΔS_{th} is a threshold in $\overline{\Delta S(t)}$ above which $g(t)$ tends to 1, allowing $\lambda_1(t)$ to increase and decrease sharply during flashy streamflow events.

A2 Actual evapotranspiration and tracer partitioning between Q and ET

Actual evapotranspiration $ET(t)$ is calculated from potential evapotranspiration $PET(t)$ using the formula:

$$ET(t) = PET(t)\, \tanh \left[\left(\frac{S(t)}{S_{root}} \right)^n \right] \tag{A9}$$

where $S_{root} = S_{ref} - 150$ is a fixed parameter (mm) representing the storage threshold $S(t) = S_{root}$ below which $ET(t)$ starts decreasing from $PET(t)$ towards 0. A similar strategy was employed for instance by Fenicia et al. (2016) and Pfister et al. (2017) in the Weierbach and neighboring Luxembourgish catchments. This decrease is smoothed by the fixed coefficient $n = 20$. S_{root}

accounts for the water available for evaporation and plant transpiration until the capillary forces offer too much resistance. This formula thus represents the decrease in water losses to the atmosphere under water limited conditions.

In the model, this equation is the only explicit partitioning condition of the tracer influx $J \times C_P$ between evaporative losses $ET \times C_{ET}$ and streamflow $Q \times C_Q$. An implicit partitioning nevertheless exists for the following reason. The tracer mass balance equation is:

$$\frac{dM}{dt}(t) = J(t)\,C_P(t) - Q(t)\,C_Q(t) - ET(t)\,C_{ET}(t) \tag{A10}$$

where $M(t)$ is the tracer mass in the catchment and $C_P(t)$ is the tracer concentration in precipitation at time t. $J(t)\,C_P(t)$ is given by the input data, and $Q(t)\,C_Q(t)$ and $ET(t)\,C_{ET}(t)$ are partly determined by the SAS functions Ω_Q and Ω_{ET}. For $Q(t)\,C_Q(t)$, $Q(t)$ is measured data, and $C_Q(t)$ is directly related to Ω_Q through the related TTD $\overleftarrow{p_Q}$ (Eq. 1 and 4). The parameters of Ω_Q are thus directly determined by the fit of the simulations to observed $C_Q(t)$. Tracer data for $C_{ET}(t)$ is not available. Thus, the parameters of Ω_{ET} cannot be directly determined from data similarly to Ω_Q. Still, the parameters of Ω_{ET} need to yield C_{ET} values which satisfy the tracer mass balance (Eq. A10) in the long term (when $\frac{dM}{dt}(t)$ becomes negligible). If the parameters of Ω_{ET} do not allow the closure of the tracer mass balance, the simulations in $C_Q(t)$ will be affected and will not match the observations. Therefore, the fit between observed and simulated $C_Q(t)$ can be used also to indirectly deduce the parameters of Ω_{ET}, using the implicit tracer partitioning Ω_{ET} exerts. This partitioning is only indirect (or implicit) because there is no one-to-one relationship between T and $C_P^*(T,t)$ (Eq. 1), meaning that age selection patterns expressed by the SAS functions do not uniquely determine the average values of $Q(t)\,C_Q(t)$ and $ET(t)\,C_{ET}(t)$. 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.

Appendix B: Statistical significance of travel time and storage differences

The obtained differences in travel time and storage measures (Tables 3 and 4) were further compared to assess their statistical significance (Table B1). For this, we used a Wilcoxon rank sum test (also known as the Mann-Whitney U-test) for each of the time-averaged (flow-weighted over 2015–2017) statistics (e.g., the 10$^{\text{th}}$ percentile) of the distributions calculated from ^{2}H (148 distributions) or ^{3}H (181 distributions) and shown in Fig. 7(a,b) and 8(a,b). This tested the null hypothesis that the two underlying median TTDs or SAS functions obtained from each tracer are equal (i.e., the distribution obtained as the median of all the flow-weighted time-averaged distributions over 2015–2017 corresponding to the behavioral parameter sets for a given tracer). We chose this test because it is non-parametric, and because it allows taking into account the travel time and storage uncertainties by including all the behavioral distributions. All tests were made at the 5% significance level.

The results show significant differences (at the 5% level) between all measures except two. According to the statistical test, the youngest fractions of water (younger than ~2 months) and the oldest fractions of water (90$^{\text{th}}$ percentile, older than about 4 years) are most likely drawn from a common TTD, regardless of the tracer used. Despite significant differences of all other measures, this test suggests that the truncation of the long TTD tail when using only deuterium is not statistically plausible.

Author contributions. LP and JK designed the project and obtained the funding for this study. NR carried out the field and lab work. NR and JK did the modelling part. NR, LP, EZ, and JK jointly structured the manuscript, and NR wrote the manuscript with contributions from JK, EZ, and LP

Competing interests. The authors declare no competing interests

Acknowledgements. We thank Uwe Morgenstern from GNS Science and Axel Schmidt from Bundesanstalt für Gewässerkunde (BfG) for providing access to the 2017 precipitation tritium data. We thank Laurent Gourdol for his help with the preparation of the tritium input data, and for useful discussions about estimating TTDs with tritium measurements. We thank Uwe Ehret for providing Matlab scripts to compute information theory measures ($\mathcal{H}$ and D_{KL}). Nicolas Rodriguez, Julian Klaus, and Laurent Pfister (FNR/CORE/C14/SR/8353440/STORE-AGE) ~~, and Julian Klaus (FNR/CORE/C17/SR/11702136/EFFECT)~~ acknowledge funding for this study from the Luxembourg National Research Fund (FNR). This study also contributes to and benefited from the "Catchments as Organized Systems" (CAOS FOR 1598, INTER/DFG/14/9476192/CAOS2) research unit funded by DFG, FNR, FWF. We thank Jérôme Juilleret for his help to collect tritium samples in the field. The authors acknowledge support by the state of Baden-Württemberg through bwHPC. We thank Barbara Glaser for her help with the input data (PET).

References

Angermann, L., Jackisch, C., Allroggen, N., Sprenger, M., Zehe, E., Tronicke, J., Weiler, M., and Blume, T.: Form and function in hillslope hydrology: characterization of subsurface flow based on response observations, Hydrology and Earth System Sciences, 21, 3727–3748, https://doi.org/10.5194/hess-21-3727-2017, 2017.

Antonelli, M., Glaser, B., Teuling, A. J., Klaus, J., and Pfister, L.: Saturated areas through the lens: 1. Spatio-temporal variability of surface saturation documented through thermal infrared imagery, Hydrological Processes, n/a, https://doi.org/10.1002/hyp.13698, 2020a.

Antonelli, M., Glaser, B., Teuling, A. J., Klaus, J., and Pfister, L.: Saturated areas through the lens: 2. Spatio-temporal variability of streamflow generation and its relationship with surface saturation, Hydrological Processes, n/a, https://doi.org/10.1002/hyp.13607, 2020b.

Bajjali, W.: Spatial variability of environmental isotope and chemical content of precipitation in Jordan and evidence of slight change in climate, Applied Water Science, 2, 271–283, https://doi.org/10.1007/s13201-012-0046-1, 2012.

Begemann, F. and Libby, W.: Continental water balance, ground water inventory and storage times, surface ocean mixing rates and world-wide water circulation patterns from cosmic-ray and bomb tritium, Geochimica et Cosmochimica Acta, 12, 277 – 296, https://doi.org/10.1016/0016-7037(57)90040-6, 1957.

Benettin, P. and Bertuzzo, E.: *tran*-SAS v1.0: a numerical model to compute catchment-scale hydrologic transport using StorAge Selection functions, Geoscientific Model Development, 11, 1627–1639, https://doi.org/10.5194/gmd-11-1627-2018, https://www.geosci-model-dev.net/11/1627/2018/, 2018.

Benettin, P., Bailey, S. W., Campbell, J. L., Green, M. B., Rinaldo, A., Likens, G. E., McGuire, K. J., and Botter, G.: Linking water age and solute dynamics in streamflow at the Hubbard Brook Experimental Forest, NH, USA, Water Resources Research, 51, 9256–9272, https://doi.org/10.1002/2015WR017552, http://doi.wiley.com/10.1002/2015WR017552, 2015a.

Benettin, P., Rinaldo, A., and Botter, G.: Tracking residence times in hydrological systems: forward and backward formulations, Hydrological Processes, https://doi.org/10.1002/hyp.15034, 2015b.

Benettin, P., Bailey, S. W., Rinaldo, A., Likens, G. E., McGuire, K. J., and Botter, G.: Young runoff fractions control streamwater age and solute concentration dynamics, Hydrological Processes, 31, 2982–2986, https://doi.org/10.1002/hyp.11243, http://doi.wiley.com/10.1002/hyp.11243, 2017a.

Benettin, P., Soulsby, C., Birkel, C., Tetzlaff, D., Botter, G., and Rinaldo, A.: Using SAS functions and high-resolution isotope data to unravel travel time distributions in headwater catchments, Water Resources Research, 53, 1864–1878, https://doi.org/10.1002/2016WR020117, http://doi.wiley.com/10.1002/2016WR020117, 2017b.

Berman, E. S. F., Gupta, M., Gabrielli, C., Garland, T., and McDonnell, J. J.: High-frequency field-deployable isotope analyzer for hydrological applications, Water Resources Research, 45, https://doi.org/10.1029/2009WR008265, 2009.

Berry, Z. C., Evaristo, J., Moore, G., Poca, M., Steppe, K., Verrot, L., Asbjornsen, H., Borma, L. S., Bretfeld, M., Hervé-Fernández, P., Seyfried, M., Schwendenmann, L., Sinacore, K., De Wispelaere, L., and McDonnell, J.: The two water worlds hypothesis: Addressing multiple working hypotheses and proposing a way forward, Ecohydrology, 11, e1843, https://doi.org/10.1002/eco.1843, e1843 ECO-16-0180.R2, 2018.

Bethke, C. M. and Johnson, T. M.: Groundwater Age and Groundwater Age Dating, Annual Review of Earth and Planetary Sciences, 36, 121–152, https://doi.org/10.1146/annurev.earth.36.031207.124210, 2008.

Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction, Hydrological Processes, 6, 279–298, https://doi.org/10.1002/hyp.3360060305, 1992.

Birkel, C., Soulsby, C., and Tetzlaff, D.: Modelling catchment-scale water storage dynamics: reconciling dynamic storage with tracer-inferred passive storage, Hydrological Processes, 25, 3924–3936, https://doi.org/10.1002/hyp.8201, http://doi.wiley.com/10.1002/hyp.8201, 2011.

Birkel, C., Soulsby, C., and Tetzlaff, D.: Conceptual modelling to assess how the interplay of hydrological connectivity, catchment storage and tracer dynamics controls nonstationary water age estimates, Hydrological Processes, https://doi.org/10.1002/hyp.10414, http://dx.doi.org/10.1002/hyp.10414, 2015.

Botter, G., Bertuzzo, E., and Rinaldo, A.: Catchment residence and travel time distributions: The master equation, Geophysical Research Letters, 38, https://doi.org/10.1029/2011GL047666, 2011.

Brooks, J. R., Barnard, H. R., Coulombe, R., and McDonnell, J. J.: Ecohydrologic separation of water between trees and streams in a Mediterranean climate, Nature Geoscience, 3, 100–104, https://doi.org/10.1038/ngeo722, 2010.

Buttafuoco, G., Caloiero, T., and Coscarelli, R.: Spatial uncertainty assessment in modelling reference evapotranspiration at regional scale, Hydrology and Earth System Sciences, 14, 2319–2327, https://doi.org/10.5194/hess-14-2319-2010, 2010.

Buttle, J.: Isotope hydrograph separations and rapid delivery of pre-event water from drainage basins, Progress in Physical Geography: Earth and Environment, 18, 16–41, https://doi.org/10.1177/030913339401800102, 1994.

Carrer, G. E., Klaus, J., and Pfister, L.: Assessing the Catchment Storage Function Through a Dual-Storage Concept, Water Resources Research, 55, 476–494, https://doi.org/10.1029/2018WR022856, 2019.

Cartwright, I. and Morgenstern, U.: Contrasting transit times of water from peatlands and eucalypt forests in the Australian Alps determined by tritium: implications for vulnerability and the source of water in upland catchments, Hydrology and Earth System Sciences, 20, 4757–4773, https://doi.org/10.5194/hess-20-4757-2016, 2016.

Criss, R. E. and Winston, W. E.: Do Nash values have value? Discussion and alternate proposals, Hydrological Processes, 22, 2723–2725, https://doi.org/10.1002/hyp.7072, 2008.

Crouzet, E., Hubert, P., Olive, P., Siwertz, E., and Marce, A.: Le tritium dans les mesures d'hydrologie de surface. Determination experimentale du coefficient de ruissellement, Journal of Hydrology, 11, 217 – 229, https://doi.org/10.1016/0022-1694(70)90063-6, 1970.

Delsman, J. R., Essink, G. H. P. O., Beven, K. J., and Stuyfzand, P. J.: Uncertainty estimation of end-member mixing using generalized likelihood uncertainty estimation (GLUE), applied in a lowland catchment, Water Resources Research, 49, 4792–4806, https://doi.org/10.1002/wrcr.20341, 2013.

Devell, L.: Measurements of the Self-diffusion of Water in Pure Water, H2O-D2O Mixtures and Solutions of Electrolytes, Acta Chemica Scandinavica, 16, 2177 – 2188, https://doi.org/10.3891/acta.chem.scand.16-2177, 1962.

Dinçer, T., Payne, B. R., Florkowski, T., Martinec, J., and Tongiorgi, E.: Snowmelt runoff from measurements of tritium and oxygen-18, Water Resources Research, 6, 110–124, https://doi.org/10.1029/WR006i001p00110, 1970.

Doherty, J. and Johnston, J. M.: METHODOLOGIES FOR CALIBRATION AND PREDICTIVE ANALYSIS OF A WATERSHED MODEL1, JAWRA Journal of the American Water Resources Association, 39, 251–265, https://doi.org/10.1111/j.1752-1688.2003.tb04381.x, 2003.

Dralle, D. N., Hahm, W. J., Rempe, D. M., Karst, N. J., Thompson, S. E., and Dietrich, W. E.: Quantification of the seasonal hillslope water storage that does not drive streamflow, Hydrological Processes, 32, 1978–1992, https://doi.org/10.1002/hyp.11627, 2018.

Dubbert, M., Caldeira, M. C., Dubbert, D., and Werner, C.: A pool-weighted perspective on the two-water-worlds hypothesis, New Phytologist, 0, https://doi.org/10.1111/nph.15670, 2019.

Duvert, C., Stewart, M. K., Cendón, D. I., and Raiber, M.: Time series of tritium, stable isotopes and chloride reveal short-term variations in groundwater contribution to a stream, Hydrology and Earth System Sciences, 20, 257–277, https://doi.org/10.5194/hess-20-257-2016, 2016.

Ehret, U. and Zehe, E.: Series distance – an intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events, Hydrology and Earth System Sciences, 15, 877–896, https://doi.org/10.5194/hess-15-877-2011, 2011.

Eriksson, E.: The Possible Use of Tritium' for Estimating Groundwater Storage, Tellus, 10, 472–478, https://doi.org/10.1111/j.2153-3490.1958.tb02035.x, 1958.

Etter, S., Strobl, B., Seibert, J., and van Meerveld, H. J. I.: Value of uncertain streamflow observations for hydrological modelling, Hydrology and Earth System Sciences, 22, 5243–5257, https://doi.org/10.5194/hess-22-5243-2018, 2018.

Fenicia, F., Wrede, S., Kavetski, D., Pfister, L., Hoffmann, L., Savenije, H. H. G., and McDonnell, J. J.: Assessing the impact of mixing assumptions on the estimation of streamwater mean residence time, Hydrological Processes, 24, 1730–1741, https://doi.org/10.1002/hyp.7595, 2010.

Fenicia, F., Kavetski, D., Savenije, H. H. G., Clark, M. P., Schoups, G., Pfister, L., and Freer, J.: Catchment properties, function, and conceptual model representation: is there a correspondence?, Hydrological Processes, 28, 2451–2467, https://doi.org/10.1002/hyp.9726, 2014.

Fenicia, F., Kavetski, D., Savenije, H. H. G., and Pfister, L.: From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions, Water Resources Research, 52, 954–989, https://doi.org/10.1002/2015WR017398, 2016.

Gabrielli, C. P., Morgenstern, U., Stewart, M. K., and McDonnell, J. J.: Contrasting Groundwater and Streamflow Ages at the Maimai Watershed, Water Resources Research, 54, 3937–3957, https://doi.org/10.1029/2017WR021825, 2018.

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.

Glaser, B., Klaus, J., Frei, S., Frentress, J., Pfister, L., and Hopp, L.: On the value of surface saturated area dynamics mapped with thermal infrared imagery for modeling the hillslope-riparian-stream continuum, Water Resources Research, 52, 8317–8342, https://doi.org/10.1002/2015WR018414, 2016.

Glaser, B., Antonelli, M., Chini, M., Pfister, L., and Klaus, J.: Technical note: Mapping surface-saturation dynamics with thermal infrared imagery, Hydrology and Earth System Sciences, 22, 5987–6003, https://doi.org/10.5194/hess-22-5987-2018, 2018.

Glaser, B., Jackisch, C., Hopp, L., and Klaus, J.: How meaningful are plot-scale observations and simulations of preferential flow for catchment models?, Vadose Zone Journal, https://doi.org/10.2136/vzj2018.08.0146, 2019.

Glaser, B., Antonelli, M., Hopp, L., and Klaus, J.: Intra-catchment variability of surface saturation–insights from physically-based simulations in comparison with biweekly thermal infrared image observations, Hydrology and Earth System Sciences, https://doi.org/10.5194/hess-2019-203, 2020.

Gourdol, L., Clément, R., Juilleret, J., Pfister, L., and Hissler, C.: Large-scale ERT surveys for investigating shallow regolith properties and architecture, Hydrology and Earth System Sciences Discussions, 2018, 1–39, https://doi.org/10.5194/hess-2018-519, 2018.

Graham, C. B., van Verseveld, W., Barnard, H. R., and McDonnell, J. J.: Estimating the deep seepage component of the hillslope and catchment water balance within a measurement uncertainty framework, Hydrological Processes, 24, 3631–3647, https://doi.org/10.1002/hyp.7788, 2010.

Gupta, P., Noone, D., Galewsky, J., Sweeney, C., and Vaughn, B. H.: Demonstration of high-precision continuous measurements of water vapor isotopologues in laboratory and remote field deployments using wavelength-scanned cavity ring-down spectroscopy (WS-CRDS) technology, Rapid Communications in Mass Spectrometry, 23, 2534–2542, https://doi.org/10.1002/rcm.4100, 2009.

Gusyev, M. A., Toews, M., Morgenstern, U., Stewart, M., White, P., Daughney, C., and Hadfield, J.: Calibration of a transient transport model to tritium data in streams and simulation of groundwater ages in the western Lake Taupo catchment, New Zealand, Hydrology and Earth System Sciences, 17, 1217–1227, https://doi.org/10.5194/hess-17-1217-2013, 2013.

Halder, J., Terzer, S., Wassenaar, L. I., Araguás-Araguás, L. J., and Aggarwal, P. K.: The Global Network of Isotopes in Rivers (GNIR): integration of water isotopes in watershed observation and riverine research, Hydrology and Earth System Sciences, 19, 3419–3431, https://doi.org/10.5194/hess-19-3419-2015, 2015.

Harman, C. J.: Time-variable transit time distributions and transport: Theory and application to storage-dependent transport of chloride in a watershed, Water Resources Research, 51, 1–30, https://doi.org/10.1002/2014WR015707, http://dx.doi.org/10.1002/2014WR015707, 2015.

Heidbuechel, I., Troch, P. A., Lyon, S. W., and Weiler, M.: The master transit time distribution of variable flow systems, Water Resources Research, 48, https://doi.org/10.1029/2011WR011293, 2012.

Helton, J. and Davis, F.: Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems, Reliability Engineering & System Safety, 81, 23 – 69, https://doi.org/10.1016/S0951-8320(03)00058-9, 2003.

Herbstritt, B., Gralher, B., and Weiler, M.: Continuous in situ measurements of stable isotopes in liquid water, Water Resources Research, 48, https://doi.org/10.1029/2011WR011369, 2012.

Herbstritt, B., Gralher, B., and Weiler, M.: Continuous, near-real-time observations of water stable isotope ratios during rainfall and throughfall events, Hydrology and Earth System Sciences, 23, 3007–3019, https://doi.org/10.5194/hess-23-3007-2019, 2019.

Hrachowitz, M., Soulsby, C., Tetzlaff, D., and Malcolm, I. A.: Sensitivity of mean transit time estimates to model conditioning and data availability, Hydrological Processes, 25, 980–990, https://doi.org/10.1002/hyp.7922, 2011.

Hrachowitz, M., Benettin, P., Breukelen, B. M. V., Fovet, O., Howden, N. J. K., Ruiz, L., Velde, Y. V. D., and Wade, A. J.: Transit times — the link between hydrology and water quality at the catchment scale, Wiley Interdisciplinary Reviews: Water, https://doi.org/10.1002/wat2.1155, 2016.

Hrachowitz, M., Savenije, H., Bogaard, T. A., Tetzlaff, D., and Soulsby, C.: What can flux tracking teach us about water age distribution patterns and their temporal dynamics?, Hydrology and Earth System Sciences, 17, 533–564, https://doi.org/10.5194/hess-17-533-2013, 2013.

Hubert, P.: Etude par le tritium de la dynamique des eaux du Lac Léman. Apport du tritium à la limnologie physique, Ph.D. thesis, Hydrologie. Université de Paris, 1971.

Hubert, P., Marin, E., Meybeck, M., Olive, P., and Siwertz, E.: Aspects Hydrologiques, Géochimiques et Sédimentologique de la Crue Exceptionnelle de la Dranse du Chablais du 22 Septembre 1968, p. 581–604, 1969.

IAEA: Global Network of Isotopes in Rivers. The GNIR Database, https://nucleus.iaea.org/wiser, 2019.

IAEA and WMO: Global Network of Isotopes in Precipitation. The GNIP Database, https://nucleus.iaea.org/wiser, 2019.

Jackisch, C., Angermann, L., Allroggen, N., Sprenger, M., Blume, T., Tronicke, J., and Zehe, E.: Form and function in hillslope hydrology: in situ imaging and characterization of flow-relevant structures, Hydrology and Earth System Sciences, 21, 3749–3775, https://doi.org/10.5194/hess-21-3749-2017, https://www.hydrol-earth-syst-sci.net/21/3749/2017/, 2017.

Juilleret, J., Iffly, J., Pfister, L., and Hissler, C.: Remarkable Pleistocene periglacial slope deposits in Luxembourg (Oesling): pedological implication and geosite potential, Bulletin de la Société des naturalistes luxembourgeois, 112, 125–130, 2011.

Juilleret, J., Dondeyne, S., Vancampenhout, K., Deckers, J., and Hissler, C.: Mind the gap: A classification system for integrating the subsolum into soil surveys, Geoderma, 264, 332 – 339, https://doi.org/10.1016/j.geoderma.2015.08.031, soil mapping, classification, and modelling: history and future directions, 2016.

Keim, R. F., Kendall, C., and Jefferson, A.: The Expanding Utility of Laser Spectroscopy, Eos, Transactions American Geophysical Union, 95, 144–144, https://doi.org/10.1002/2014EO170007, 2014.

Kendall, C. and McDonnell, J. J.: Isotope tracers in catchment hydrology, Elsevier, Amsterdam, https://doi.org/10.1016/B978-0-444-81546-0.50001-X, 1998.

Kirchner, J. W.: A double paradox in catchment hydrology and geochemistry, Hydrological Processes, 17, 871–874, https://doi.org/10.1002/hyp.5108, http://doi.wiley.com/10.1002/hyp.5108, 2003.

Kirchner, J. W.: Aggregation in environmental systems - Part 1: Seasonal tracer cycles quantify young water fractions, but not mean transit times, in spatially heterogeneous catchments, Hydrology and Earth System Sciences, 20, 279–297, https://doi.org/10.5194/hess-20-279-2016, 2016.

Klaus, J. and McDonnell, J.: Hydrograph separation using stable isotopes: Review and evaluation, Journal of Hydrology, 505, 47 – 64, https://doi.org/10.1016/j.jhydrol.2013.09.006, 2013.

Klaus, J. and Zehe, E.: Modelling rapid flow response of a tile-drained field site using a 2D physically based model: assessment of 'equifinal' model setups, Hydrological Processes, 24, 1595–1609, https://doi.org/10.1002/hyp.7687, 2010.

Klaus, J., Chun, K. P., McGuire, K. J., and McDonnell, J. J.: Temporal dynamics of catchment transit times from stable isotope data, Water Resources Research, 51, 4208–4223, https://doi.org/10.1002/2014WR016247, 2015a.

Koehler, G. and Wassenaar, L. I.: Realtime Stable Isotope Monitoring of Natural Waters by Parallel-Flow Laser Spectroscopy, Analytical Chemistry, 83, 913–919, https://doi.org/10.1021/ac102584q, pMID: 21214188, 2011.

Lis, G., Wassenaar, L. I., and Hendry, M. J.: High-Precision Laser Spectroscopy D/H and 18O/16O Measurements of Microliter Natural Water Samples, Analytical Chemistry, 80, 287–293, https://doi.org/10.1021/ac701716q, pMID: 18031060, 2008.

Loritz, R., Hassler, S. K., Jackisch, C., Allroggen, N., van Schaik, L., Wienhöfer, J., and Zehe, E.: Picturing and modeling catchments by representative hillslopes, Hydrology and Earth System Sciences, 21, 1225–1249, https://doi.org/10.5194/hess-21-1225-2017, 2017.

Loritz, R., Gupta, H., Jackisch, C., Westhoff, M., Kleidon, A., Ehret, U., and Zehe, E.: On the dynamic nature of hydrological similarity, Hydrology and Earth System Sciences, 22, 3663–3684, https://doi.org/10.5194/hess-22-3663-2018, 2018.

Loritz, R., Kleidon, A., Jackisch, C., Westhoff, M., Ehret, U., Gupta, H., and Zehe, E.: A topographic index explaining hydrological similarity by accounting for the joint controls of runoff formation, Hydrology and Earth System Sciences, 23, 3807–3821, https://doi.org/10.5194/hess-23-3807-2019, 2019.

Maher, K.: The role of fluid residence time and topographic scales in determining chemical fluxes from landscapes, Earth and Planetary Science Letters, 312, 48–58, https://doi.org/10.1016/j.epsl.2011.09.040, http://linkinghub.elsevier.com/retrieve/pii/S0012821X11005607, 2011.

Maloszewski, P. and Zuber, A.: Principles and practice of calibration and validation of mathematical models for the interpretation of environmental tracer data in aquifers, Advances in Water Resources, 16, 173 – 190, https://doi.org/10.1016/0309-1708(93)90036-F, 1993.

Małoszewski, P. and Zuber, A.: Determining the turnover time of groundwater systems with the aid of environmental tracers: 1. Models and their applicability, Journal of Hydrology, 57, 207 – 231, https://doi.org/10.1016/0022-1694(82)90147-0, 1982.

Małoszewski, P., Rauert, W., Stichler, W., and Herrmann, A.: Application of flow models in an alpine catchment area using tritium and deuterium data, Journal of Hydrology, 66, 319 – 330, https://doi.org/10.1016/0022-1694(83)90193-2, 1983.

Martinec, J.: Subsurface flow from snowmelt traced by tritium, Water Resources Research, 11, 496–498, https://doi.org/10.1029/WR011i003p00496, 1975.

Martínez-Carreras, N., Wetzel, C. E., Frentress, J., Ector, L., McDonnell, J. J., Hoffmann, L., and Pfister, L.: Hydrological connectivity inferred from diatom transport through the riparian-stream system, Hydrology and Earth System Sciences, 19, 3133–3151, https://doi.org/10.5194/hess-19-3133-2015, 2015.

Martínez-Carreras, N., Hissler, C., Gourdol, L., Klaus, J., Juilleret, J., Iffly, J. F., and Pfister, L.: Storage controls on the generation of double peak hydrographs in a forested headwater catchment, Journal of Hydrology, 543, 255 – 269, https://doi.org/10.1016/j.jhydrol.2016.10.004, 2016.

McCutcheon, R. J., McNamara, J. P., Kohn, M. J., and Evans, S. L.: An evaluation of the ecohydrological separation hypothesis in a semiarid catchment, Hydrological Processes, 31, 783–799, https://doi.org/10.1002/hyp.11052, 2017.

McDonnell, J. J.: The two water worlds hypothesis: ecohydrological separation of water between streams and trees?, Wiley Interdisciplinary Reviews: Water, 1, 323–329, https://doi.org/10.1002/wat2.1027, 2014.

McDonnell, J. J. and Beven, K. J.: Debates on Water Resources: The future of hydrological sciences: A (common) path forward? A call to action aimed at understanding velocities, celerities and residence time distributions of the headwater hydrograph, Water Resources Research, 50, 5342–5350, https://doi.org/10.1002/2013WR015141, 2014.

McGuire, K. J. and McDonnell, J. J.: A review and evaluation of catchment transit time modeling, Journal of Hydrology, 330, 543–563, https://doi.org/10.1016/j.jhydrol.2006.04.020, 2006.

McGuire, K. J., McDonnell, J. J., Weiler, M., Kendall, C., McGlynn, B. L., Welker, J. M., and Seibert, J.: The role of topography on catchment-scale water residence time, Water Resources Research, 41, https://doi.org/10.1029/2004WR003657, 2005.

McMahon, T. A., Peel, M. C., Lowe, L., Srikanthan, R., and McVicar, T. R.: Estimating actual, potential, reference crop and pan evaporation using standard meteorological data: a pragmatic synthesis, Hydrology and Earth System Sciences, 17, 1331–1363, https://doi.org/10.5194/hess-17-1331-2013, 2013.

McMillan, H., Krueger, T., and Freer, J.: Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality, Hydrological Processes, 26, 4078–4111, https://doi.org/10.1002/hyp.9384, 2012.

Michelsen, N., Laube, G., Friesen, J., Weise, S. M., Bait Said, A. B. A., and Müller, T.: Technical note: A microcontroller-based automatic rain sampler for stable isotope studies, Hydrology and Earth System Sciences, 23, 2637–2645, https://doi.org/10.5194/hess-23-2637-2019, 2019.

Moragues-Quiroga, C., Juilleret, J., Gourdol, L., Pelt, E., Perrone, T., Aubert, A., Morvan, G., Chabaux, F., Legout, A., Stille, P., and Hissler, C.: Genesis and evolution of regoliths: Evidence from trace and major elements and Sr-Nd-Pb-U isotopes, CATENA, 149, 185 – 198, https://doi.org/10.1016/j.catena.2016.09.015, 2017.

Morgenstern, U. and Taylor, C. B.: Ultra low-level tritium measurement using electrolytic enrichment and LSC, Isotopes in Environmental and Health Studies, 45, 96–117, https://doi.org/10.1080/10256010902931194, pMID: 20183224, 2009.

Munksgaard, N. C., Wurster, C. M., and Bird, M. I.: Continuous analysis of δ18O and δD values of water by diffusion sampling cavity ring-down spectrometry: a novel sampling device for unattended field monitoring of precipitation, ground and surface waters, Rapid Communications in Mass Spectrometry, 25, 3706–3712, https://doi.org/10.1002/rcm.5282, 2011.

Palcsu, L., Morgenstern, U., Sültenfuss, J., Koltai, G., László, E., Temovski, M., Major, Z., Nagy, J. T., Papp, L., Varlam, C., Faurescu, I., Túri, M., Rinyu, L., Czuppon, G., Bottyán, E., and Jull, A. J. T.: Modulation of Cosmogenic Tritium in Meteoric Precipitation by the 11-year Cycle of Solar Magnetic Field Activity, Scientific Reports, 8, 12 813, https://doi.org/10.1038/s41598-018-31208-9, 2018.

Pangle, L. A., Klaus, J., Berman, E. S. F., Gupta, M., and McDonnell, J. J.: A new multisource and high-frequency approach to measuring δ2H and δ18O in hydrological field studies, Water Resources Research, 49, 7797–7803, https://doi.org/10.1002/2013WR013743, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2013WR013743, 2013.

Parsekian, A. D., Singha, K., Minsley, B. J., Holbrook, W. S., and Slater, L.: Multiscale geophysical imaging of the critical zone, Reviews of Geophysics, 53, 1–26, https://doi.org/10.1002/2014RG000465, 2015.

Pfister, L., Martínez-Carreras, N., Hissler, C., Klaus, J., Carrer, G. E., Stewart, M. K., and McDonnell, J. J.: Bedrock geology controls on catchment storage, mixing, and release: A comparative analysis of 16 nested catchments, Hydrological Processes, 31, 1828–1845, https://doi.org/10.1002/hyp.11134, 2017.

Pfister, L., Thielen, F., Deloule, E., Valle, N., Lentzen, E., Grave, C., Beisel, J.-N., and McDonnell, J. J.: Freshwater pearl mussels as a stream water stable isotope recorder, Ecohydrology, 11, e2007, https://doi.org/10.1002/eco.2007, 2018.

Pfister, L., Grave, C., Beisel, J.-N., and McDonnell, J. J.: A global assessment of freshwater mollusk shell oxygen isotope signatures and their relation to precipitation and stream water, Scientific Reports, 9, 4312, https://doi.org/10.1038/s41598-019-40369-0, 2019.

Pool, S., Viviroli, D., and Seibert, J.: Prediction of hydrographs and flow-duration curves in almost ungauged catchments: Which runoff measurements are most informative for model calibration?, Journal of Hydrology, 554, 613 – 622, https://doi.org/10.1016/j.jhydrol.2017.09.037, 2017.

Rank, D. and Papesch, W.: Isotopic composition of precipitation in Austria in relation to air circulation patterns and climate, chap. 2, pp. 19–35, International Atomic Energy Agency (IAEA), 2005.

Rank, D., Wyhlidal, S., Schott, K., Weigand, S., and Oblin, A.: Temporal and spatial distribution of isotopes in river water in Central Europe: 50 years experience with the Austrian network of isotopes in rivers, Isotopes in Environmental and Health Studies, 54, 115–136, https://doi.org/10.1080/10256016.2017.1383906, pMID: 29082751, 2018.

Różański, K., Froehlich, K., and Mook, W. G.: Environmental Isotopes in the Hydrological Cycle, Principles and Applications. VOLUME III: Surface water, IAEA and UNESCO, 2001.

Rinaldo, A. and Marani, A.: Basin scale-model of solute transport, Water Resources Research, 23, 2107–2118, https://doi.org/10.1029/WR023i011p02107, 1987.

Rinaldo, A., Benettin, P., Harman, C. J., Hrachowitz, M., McGuire, K. J., van der Velde, Y., Bertuzzo, E., and Botter, G.: Storage selection functions: A coherent framework for quantifying how catchments store and release water and solutes, Water Resources Research, 51, 4840–4847, https://doi.org/10.1002/2015WR017273, http://dx.doi.org/10.1002/2015WR017273, 2015.

Rodriguez, N. B. and Klaus, J.: Catchment Travel Times From Composite StorAge Selection Functions Representing the Superposition of Streamflow Generation Processes, Water Resources Research, 55, 9292–9314, https://doi.org/10.1029/2019WR024973, 2019.

Rodriguez, N. B., McGuire, K. J., and Klaus, J.: Time-Varying Storage-Water Age Relationships in a Catchment With a Mediterranean Climate, Water Resources Research, 54, 3988–4008, https://doi.org/10.1029/2017wr021964, 2018.

Rodriguez, N. B., Benettin, P., and Klaus, J.: Multimodal water age distributions and the challenge of complex hydrological landscapes, Hydrological Processes, https://doi.org/10.1002/hyp.13770, 2020.

Scaini, A., Audebert, M., Hissler, C., Fenicia, F., Gourdol, L., Pfister, L., and Beven, K. J.: Velocity and celerity dynamics at plot scale inferred from artificial tracing experiments and time-lapse ERT, Journal of Hydrology, 546, 28 – 43, https://doi.org/10.1016/j.jhydrol.2016.12.035, 2017.

Scaini, A., Hissler, C., Fenicia, F., Juilleret, J., Iffly, J. F., Pfister, L., and Beven, K.: Hillslope response to sprinkling and natural rainfall using velocity and celerity estimates in a slate-bedrock catchment, Journal of Hydrology, 558, 366 – 379, https://doi.org/10.1016/j.jhydrol.2017.12.011, 2018.

Schaefli, B. and Gupta, H. V.: Do Nash values have value?, Hydrological Processes, 21, 2075–2080, https://doi.org/10.1002/hyp.6825, 2007.

Schmidt, A., Frank, G., Stichler, W., Duester, L., Steinkopff, T., and Stumpp, C.: Overview of tritium records from precipitation and surface waters in Germany, Hydrological Processes, n/a, https://doi.org/10.1002/hyp.13691, 2020.

Schoups, G. and Vrugt, J. A.: A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, Water Resources Research, 46, https://doi.org/10.1029/2009WR008933, 2010.

Schwab, M. P., Klaus, J., Pfister, L., and Weiler, M.: Diel fluctuations of viscosity-driven riparian inflow affect streamflow DOC concentration, Biogeosciences, 15, 2177–2188, https://doi.org/10.5194/bg-15-2177-2018, 2018.

Seibert, J.: On the need for benchmarks in hydrological modelling, Hydrological Processes, 15, 1063–1064, https://doi.org/10.1002/hyp.446, 2001.

Seibert, S. P., Ehret, U., and Zehe, E.: Disentangling timing and amplitude errors in streamflow simulations, Hydrology and Earth System Sciences, 20, 3745–3763, https://doi.org/10.5194/hess-20-3745-2016, 2016.

Soulsby, C., Piegat, K., Seibert, J., and Tetzlaff, D.: Catchment-scale estimates of flow path partitioning and water storage based on transit time and runoff modelling, Hydrological Processes, 25, 3960–3976, https://doi.org/10.1002/hyp.8324, 2011.

Soulsby, C., Tetzlaff, D., and Hrachowitz, M.: Tracers and transit times: windows for viewing catchment scale storage?, Hydrological Processes, 23, 3503–3507, https://doi.org/10.1002/hyp.7501, 2009.

Sprenger, M., Leistert, H., Gimbel, K., and Weiler, M.: Illuminating hydrological processes at the soil-vegetation-atmosphere interface with water stable isotopes, Reviews of Geophysics, 54, 674–704, https://doi.org/10.1002/2015RG000515, 2016.

Sprenger, M., Stumpp, C., Weiler, M., Aeschbach, W., Allen, S. T., Benettin, P., Dubbert, M., Hartmann, A., Hrachowitz, M., Kirchner, J. W., McDonnell, J. J., Orlowski, N., Penna, D., Pfahl, S., Rinderer, M., Rodriguez, N., Schmidt, M., and Werner, C.: The demographics of water: A review of water ages in the critical zone, Reviews of Geophysics, 0, https://doi.org/10.1029/2018RG000633, 2019.

Stamoulis, K., Ioannides, K., Kassomenos, P., and Vlachogianni, A.: Tritium Concentration in Rainwater Samples in Northwestern Greece, Fusion Science and Technology, 48, 512–515, https://doi.org/10.13182/FST05-A978, 2005.

Stewart, M. K. and Morgenstern, U.: Importance of tritium-based transit times in hydrological systems, Wiley Interdisciplinary Reviews: Water, 3, 145–154, https://doi.org/10.1002/wat2.1134, 2016.

Stewart, M. K. and Thomas, J. T.: A conceptual model of flow to the Waikoropupu Springs, NW Nelson, New Zealand, based on hydrometric and tracer (^{18}O, Cl,^{3}H and CFC) evidence, Hydrology and Earth System Sciences, 12, 1–19, https://doi.org/10.5194/hess-12-1-2008, 2008.

Stewart, M. K., Mehlhorn, J., and Elliott, S.: Hydrometric and natural tracer (oxygen-18, silica, tritium and sulphur hexafluoride) evidence for a dominant groundwater contribution to Pukemanga Stream, New Zealand, Hydrological Processes, 21, 3340–3356, https://doi.org/10.1002/hyp.6557, 2007.

Stewart, M. K., Morgenstern, U., and McDonnell, J. J.: Truncation of stream residence time: how the use of stable isotopes has skewed our concept of streamwater age and origin, Hydrological Processes, 24, 1646–1659, https://doi.org/10.1002/hyp.7576, 2010.

Stewart, M. K., Morgenstern, U., McDonnell, J. J., and Pfister, L.: The 'hidden streamflow' challenge in catchment hydrology: a call to action for stream water transit time analysis, Hydrological Processes, 26, 2061–2066, https://doi.org/10.1002/hyp.9262, 2012.

Stewart, M. K., Morgenstern, U., Gusyev, M. A., and Małoszewski, P.: Aggregation effects on tritium-based mean transit times and young water fractions in spatially heterogeneous catchments and groundwater systems, Hydrology and Earth System Sciences, 21, 4615–4627, https://doi.org/10.5194/hess-21-4615-2017, 2017.

Östlund, G. H.: Hurricane Tritium I: Preliminary Results on Hilda 1964 and Betsy 1965, pp. 58–60, American Geophysical Union (AGU), https://doi.org/10.1029/GM011p0058, 2013.

Stumpp, C., Klaus, J., and Stichler, W.: Analysis of long-term stable isotopic composition in German precipitation, Journal of Hydrology, 517, 351 – 361, https://doi.org/10.1016/j.jhydrol.2014.05.034, 2014.

Thiesen, S., Darscheid, P., and Ehret, U.: Identifying rainfall-runoff events in discharge time series: a data-driven method based on information theory, Hydrology and Earth System Sciences, 23, 1015–1034, https://doi.org/10.5194/hess-23-1015-2019, 2019.

Uhlenbrook, S., Frey, M., Leibundgut, C., and Maloszewski, P.: Hydrograph separations in a mesoscale mountainous basin at event and seasonal timescales, Water Resources Research, 38, 31–1–31–14, https://doi.org/10.1029/2001WR000938, 2002.

van der Velde, Y., Heidbüchel, I., Lyon, S. W., Nyberg, L., Rodhe, A., Bishop, K., and Troch, P. A.: Consequences of mixing assumptions for time-variable travel time distributions, Hydrological Processes, 29, 3460–3474, https://doi.org/10.1002/hyp.10372, 2015.

van Meerveld, H. J. I., Kirchner, J. W., Vis, M. J. P., Assendelft, R. S., and Seibert, J.: Expansion and contraction of the flowing stream network changes hillslope flowpath lengths and the shape of the travel time distribution, Hydrology and Earth System Sciences Discussions, 2019, 1–18, https://doi.org/10.5194/hess-2019-218, https://www.hydrol-earth-syst-sci-discuss.net/hess-2019-218/, 2019.

Visser, A., Thaw, M., Deinhart, A., Bibby, R., Safeeq, M., Conklin, M., Esser, B., and Van der Velde, Y.: Cosmogenic Isotopes Unravel the Hydrochronology and Water Storage Dynamics of the Southern Sierra Critical Zone, Water Resources Research, 0, https://doi.org/10.1029/2018WR023665, 2019.

von Freyberg, J., Studer, B., and Kirchner, J. W.: A lab in the field: high-frequency analysis of water quality and stable isotopes in stream water and precipitation, Hydrology and Earth System Sciences, 21, 1721–1739, https://doi.org/10.5194/hess-21-1721-2017, 2017.

Vrugt, J. A.: Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation, Environmental Modelling & Software, 75, 273 – 316, https://doi.org/https://doi.org/10.1016/j.envsoft.2015.08.013, 2016.

Waichler, S. R., Wemple, B. C., and Wigmosta, M. S.: Simulation of water balance and forest treatment effects at the H.J. Andrews Experimental Forest, Hydrological Processes, 19, 3177–3199, https://doi.org/10.1002/hyp.5841, 2005.

Wilusz, D. C., Harman, C. J., and Ball, W. P.: Sensitivity of Catchment Transit Times to Rainfall Variability Under Present and Future Climates, Water Resources Research, https://doi.org/10.1002/2017WR020894, http://dx.doi.org/10.1002/2017WR020894, 2017.

Wrede, S., Fenicia, F., Martínez-Carreras, N., Juilleret, J., Hissler, C., Krein, A., Savenije, H. H. G., Uhlenbrook, S., Kavetski, D., and Pfister, L.: Towards more systematic perceptual model development: a case study using 3 Luxembourgish catchments, Hydrological Processes, 29, 2731–2750, https://doi.org/10.1002/hyp.10393, 2015.

Zuber, A.: On the interpretation of tracer data in variable flow systems, Journal of Hydrology, 86, 45 – 57, https://doi.org/10.1016/0022-1694(86)90005-3, 1986.

Table 2. Parameter ranges and information measures before and after calibration to isotopic data

Parameter	S_{th}	ΔS_{th}	S_u	f_0	λ_1^*	λ_2	μ_2	θ_2	μ_3	θ_3	μ_{ET}	θ_{ET}
Unit	mm	mm	mm	–	–	–	mm	mm	mm	mm	mm	mm
Range for $E_2 > L_2$	$[21, 196]$	$[0.14, 20]$	$[1.4, 50]$	$[0, 1]$	$[0, 0.76]$	$[0, 1]$	$[149, 1530]$	$[6, 100]$	$[21, 1561]$	$[2, 100]$	$[51, 926]$	$[1, 9$
Range for $E_3 < L_3$	$[20, 200]$	$[0.1, 20]$	$[1, 50]$	$[0, 1]$	$[0, 0.92]$	$[0, 1]$	$[58, 1564]$	$[0, 100]$	$[67, 1600]$	$[0, 100]$	$[0, 959]$	$[1, 10$
Range for $E_2 > L_2$ and $E_3 < L_3$	$[25, 177]$	$[0.17, 20]$	$[11, 49]$	$[0, 0.82]$	$[0, 0.76]$	$[0.1, 1]$	$[897, 1530]$	$[7, 69]$	$[440, 1561]$	$[7, 96]$	$[51, 120]$	$[3, 9$
Binning[a]	[20:20:200]	[0:2:20]	[0:5:50]	[0:0.1:1]	[0:0.05:1]	[0:0.1:1]	[0:100:1600]	[0:10:100]	[0:100:1600]	[0:10:100]	[0:50:1600]	[0:10:1
$\mathcal{H}(X)$[b]	~~10.53~~ 3.17	~~11.03~~ 3.32	~~11.03~~ 3.32	~~11.03~~ 3.32	~~12.32~~ 3.71	~~11.03~~ 3.32	~~13.29~~ 4	~~11.03~~ 3.32	~~13.29~~ 4	~~11.03~~ 3.32	~~16.61~~ 5	~~11.03~~ 3
$\mathcal{H}(X\|^2H)$	~~10.36~~ 3.12	~~10.86~~ 3.27	~~10.93~~ 3.29	~~10.7~~ 3.22	~~9.87~~ 2.97	~~10.63~~ 3.2	~~11.43~~ 3.44	~~10.8~~ 3.25	~~11.93~~ 3.59	~~10.73~~ 3.23	~~9.2~~ 2.77	~~10.76~~ 3
$\mathcal{H}(X\|^3H)$	~~10.3~~ 3.1	~~10.96~~ 3.3	~~11~~ 3.31	~~10.96~~ 3.3	~~10.73~~ 3.23	~~10.76~~ 3.24	~~12.39~~ 3.73	~~10.8~~ 3.25	~~12.72~~ 3.83	~~10.8~~ 3.25	~~5.08~~ 1.53	~~10.7~~ 3
$\mathcal{H}(X\|(^2H \cap {}^3H))$	~~8.37~~ 2.52	~~9.8~~ 2.95	~~8.4~~ 2.53	~~9.67~~ 2.91	~~8.8~~ 2.65	~~9.67~~ 2.91	~~8.24~~ 2.48	~~7.24~~ 2.18	~~9.67~~ 2.91	~~9.14~~ 2.75	~~2.33~~ 0.7	~~8.24~~ 2
$D_{KL}(X\|^2H, X)$	0.05	0.05	0.04	0.10	0.27	0.12	0.56	0.07	0.41	0.10	2.22	0.09
$D_{KL}(X\|^3H, X)$	0.07	0.02	0.01	0.02	0.13	0.08	0.27	0.07	0.17	0.07	3.45	0.10
$D_{KL}(X\|(^2H \cap {}^3H), X)$	0.64	0.37	0.76	0.42	0.60	0.41	1.52	1.14	1.10	0.57	4.30	0.84
$D_{KL}(X\|(^2H \cap {}^3H), X\|^2H)$	~~1.3~~ 0.39	~~1.2~~ 0.36	~~2.49~~ 0.75	~~1~~ 0.3	~~1.2~~ 0.36	~~0.9~~ 0.27	~~4.09~~ 1.23	~~3.42~~ 1.03	~~2.49~~ 0.75	~~1.59~~ 0.48	~~4.29~~ 1.29	~~2.39~~ 0
$D_{KL}(X\|(^2H \cap {}^3H), X\|^3H)$	~~1.59~~ 0.48	~~1.2~~ 0.36	~~2.66~~ 0.8	~~1.4~~ 0.42	~~1.1~~ 0.33	~~1.13~~ 0.34	~~3.59~~ 1.08	~~3.02~~ 0.91	~~2.79~~ 0.84	~~1.23~~ 0.37	~~6.78~~ 2.04	~~2.49~~ 0

[a]Binning is indicated as $[a : b : c]$, where a is the left edge of the first bin, b is the bin width, and c is the right edge of the last bin.
For instance, $[7 : 2 : 11]$ indicates data sorted with the two bins $[7, 9]$ and $[9, 11]$

[b]$\mathcal{H}$ and D_{KL} are expressed in bits.

Table 3. Statistics of $\overleftarrow{P_Q}(T)$ constrained by deuterium or tritium

~~Age~~ Travel time statistics	^{2}H ($E_2 > 0$) [mean ± std]	^{3}H ($E_3 < 0.5$ T.U.) [mean ± std]	^{3}H–^{2}H differences Absolute difference	^{2}H and ^{3}H [mean ± std]
10th percentile [years]	0.78 ± 0.49	1.10 ± 0.57	0.32 years	1.44 ± 0.11
25th percentile [years]	1.16 ± 0.56	1.54 ± 0.59	0.38 years	1.85 ± 0.22
Median ~~age~~ [years]	1.77 ± 0.55	2.19 ± 0.64	0.42 years	2.38 ± 0.15
75th percentile [years]	2.78 ± 0.61	3.07 ± 0.74	0.29 years	3.26 ± 0.39
90th percentile [years]	4.64 ± 1.27	4.79 ± 1.41	0.15 years	5.19 ± 0.86
Mean ~~age~~ [years]	2.90 ± 0.54	3.12 ± 0.59	0.22 years	3.45 ± 0.28
F_{yw}[a] [%]	1.5 ± 1.6	1.8 ± 2.3	0.3%	0.61 ± 0.53
F(T < 6 months) [%]	10 ± 8.6	6.3 ± 8.2	-3.7%	0.75 ± 0.58
F(T < 1 year) [%]	24 ± 17	11 ± 12	-13%	2.1 ± 1.5
F(T < 3 years) [%]	77 ± 8.5	71 ± 16	-6%	70 ± 6.6

The mean and standard deviations are calculated from all retained behavioral solutions for a given criterion. [a] Fraction of "young water" (Kirchner, 2016), younger than 0.2 years

Table 4. Storage estimate S_{95P} constrained by deuterium or tritium

Statistics of S_{95P}	^{2}H ($E_2 > 0$)	^{3}H ($E_3 < 0.5$ T.U.)	^{2}H and ^{3}H
Mean ± st. dev. [mm]	1275 ± 245	1335 ± 279	1488 ± 135
Median ± st. dev. [mm]	1281 ± 245	1392 ± 279	1505 ± 135
Min [mm]	625	660	1249
Max [mm]	1744	1806	1710

S_{95P} is calculated as the 95th percentile of Ω_{tail} (eq. 11)

Table B1. Results from the Wilcoxon rank sum test comparing the travel time and storage measures between ^{2}H and ^{3}H behavioral solutions. The null hypothesis is that the measures are extracted from the same underlying distribution for both tracers.

Travel time or storage measure	Decision about the null hypothesis	p-value
10^{th} percentile	Rejected	3.3×10^{-6}
25^{th} percentile	Rejected	5.9×10^{-8}
Median	Rejected	1.5×10^{-8}
75^{th} percentile	Rejected	1.1×10^{-3}
90^{th} percentile	Accepted	0.30
Mean	Rejected	3.5×10^{-5}
F_{yw} [a]	Accepted	0.37
F(T < 6 months)	Rejected	5.3×10^{-6}
F(T < 1 year)	Rejected	2.7×10^{-10}
F(T < 3 years)	Rejected	2.5×10^{-3}
S_{95P}	Rejected	1.4×10^{-2}

All tests were made at the 5% significance level.

[a] Fraction of "young water" (Kirchner, 2016), younger than 0.2 years