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). 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.

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.

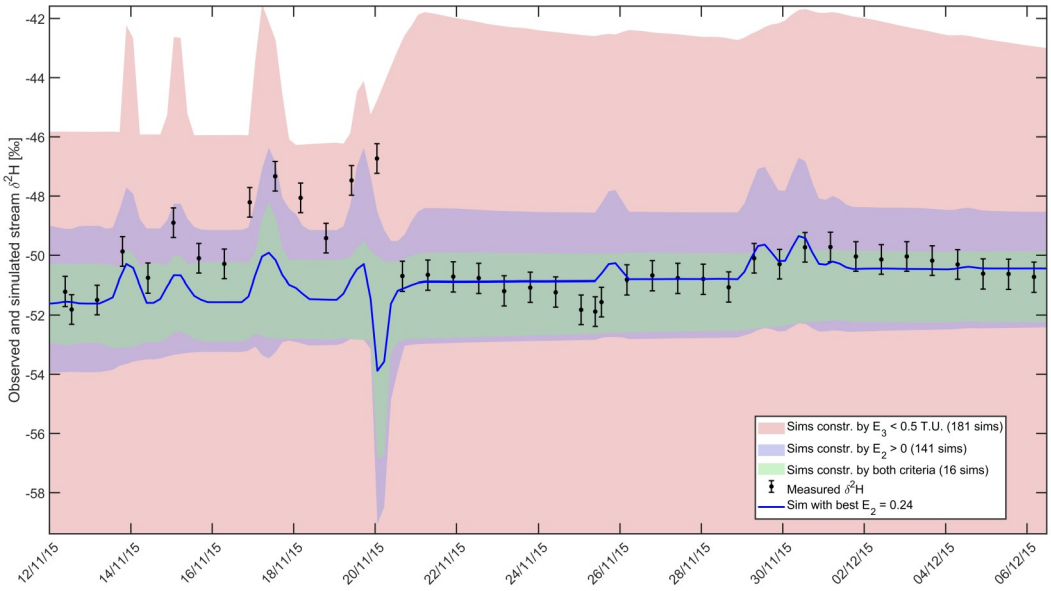

Figure: δ^2H simulations in Nov-Dec 2015

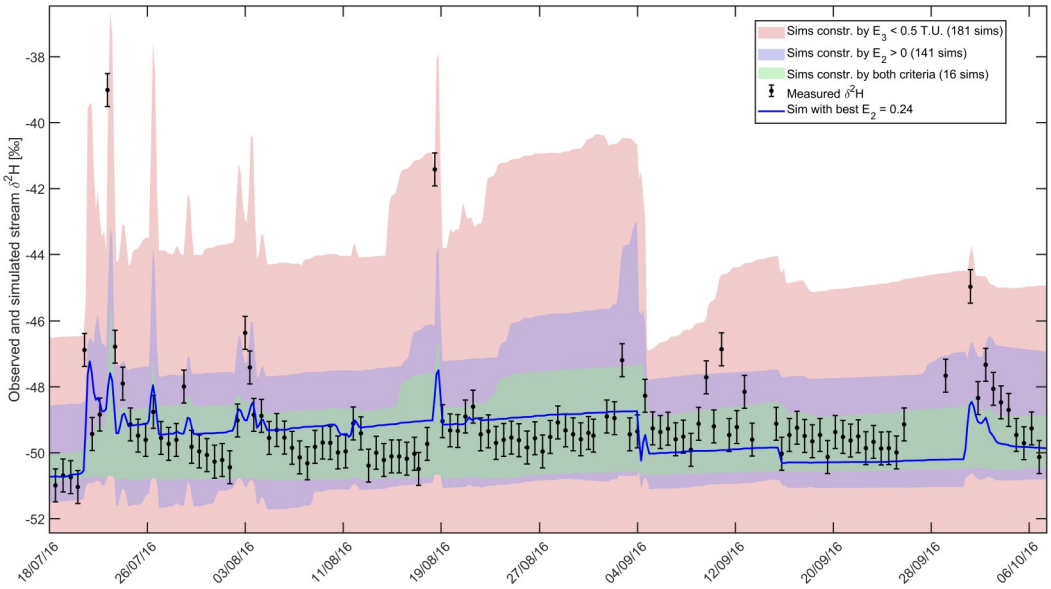

Figure: δ^2H simulations in Jul-Oct 2016

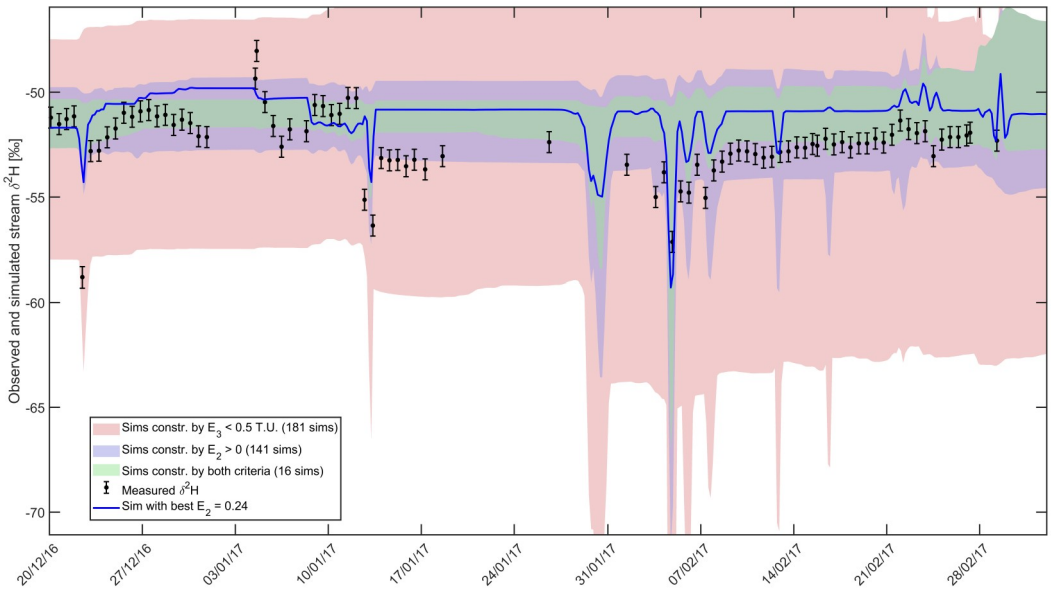

Figure: δ^2H simulations in winter 2016

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.

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 ^{3}H dataset is enough. The number of samples is too limited to constraint 12 parameters.

Authors: The ^{3}H data set has, with the study of Visser et al. (2019), one of the highest number of stream samples analyzed for ^{3}H 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.

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.

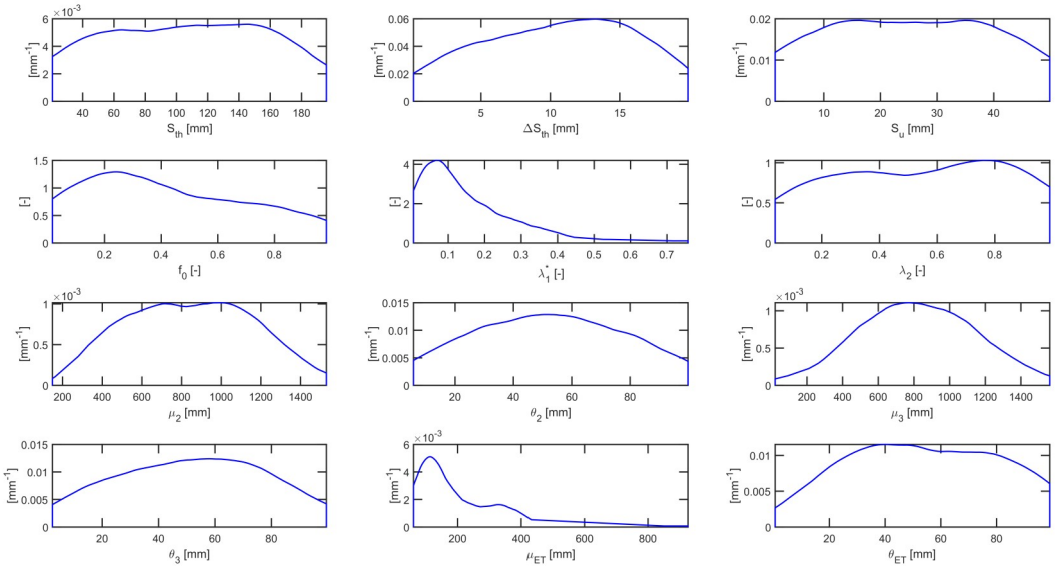

Figure: Posterior distributions constrained by deuterium

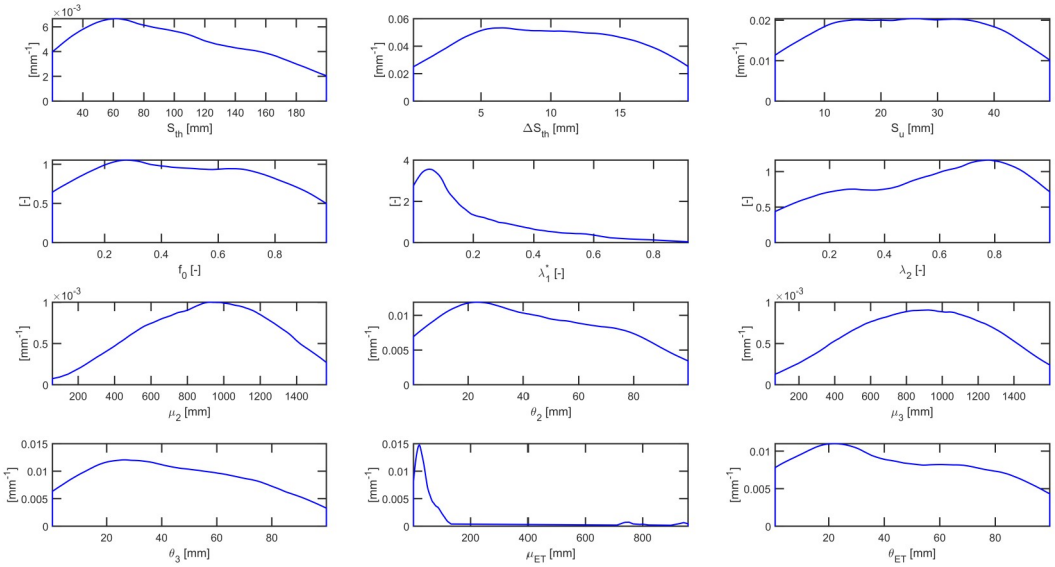

Figure: Posterior distributions constrained by tritium

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 ^2H 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).

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.

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.

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).

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). 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.

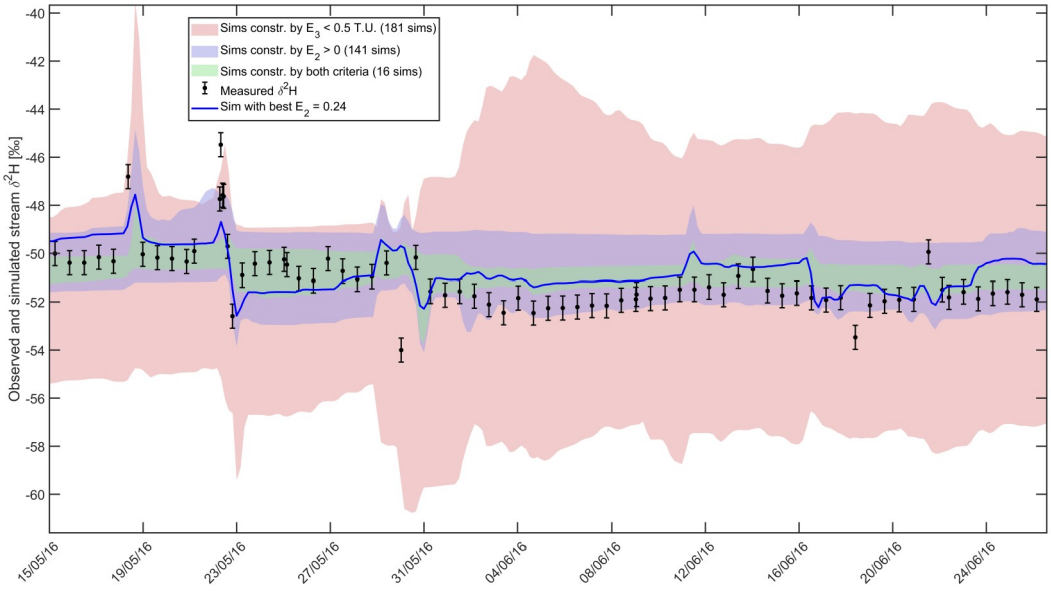

Figure: Simulations of δ^2H in May-June 2016

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.

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.

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.

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.

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.

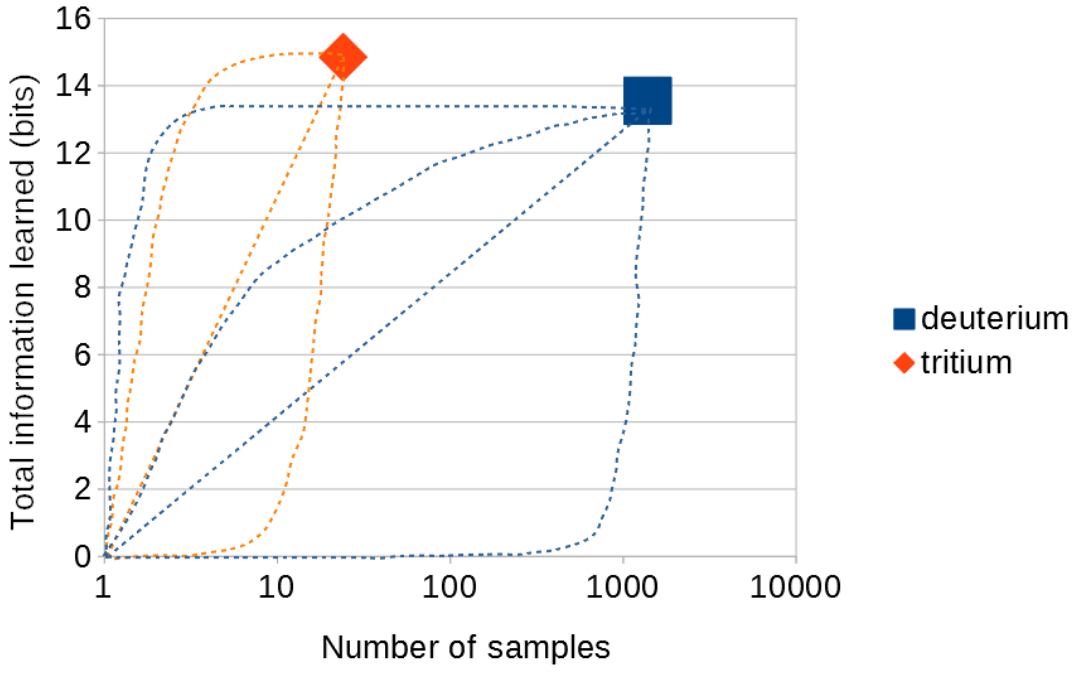

Figure: Information learned about water ages from each tracer (points) and potential relationships between the number of samples and the (necessarily) growing information content (dashed lines).

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.

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.

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.

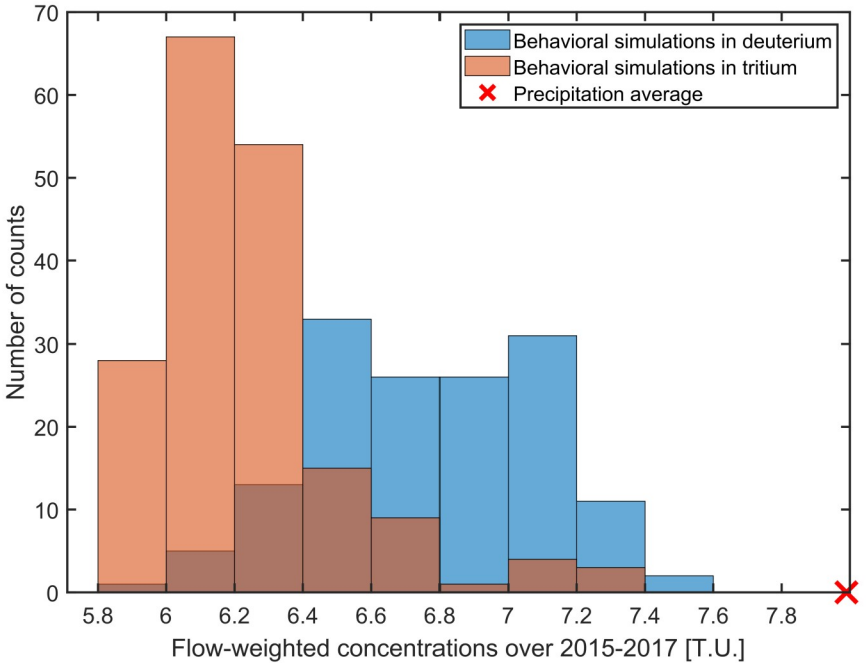

Figure: Simulated flow-weighted concentrations in the stream for the behavioral model runs constrained by deuterium samples or by tritium samples.

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.

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.

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.

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.

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.

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.

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.

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