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):

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 ³H samples compared to δ²H 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 m³). 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:

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

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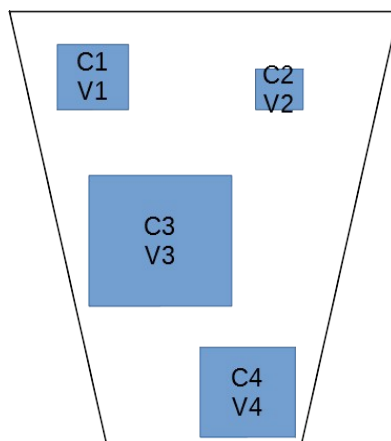
Each water particle k (there are n=4 particles represented here) has a given volume V_k and a given concentration C_k. The measured tracer concentration of the sample is:

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

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.