General comments

The manuscript presents a case study of tritium and oxygen-18 measurements at the mouth of the Neckar basin. Using long time series, the authors explore whether both tracers can yield similar estimates of the mean transit time, and if not, why. For this, three different approaches have been chosen: simple estimation of the mean transit time from the amplitude difference between the tracer signal in precipitation and in the stream, estimation using both isotopes with lumped parameter models, and estimation with a stream selection function (SAS) model coupled to a hydrological model. Based on modelling results, the authors advance far reaching conclusions on transit time estimation using a seasonal isotopic cycle, and on the superiority of the SAS approach compared to the lumped parameter modelling approach. The problem is that these conclusions are not substantiated by the authors' analysis, which is based on the one hand on a curiously simplistic implementation of the lumped parameter models and on the other hand on the use of an SAS model that is most probably overparameterized and does not really reproduce observations well. I see four essential shortcomings of the manuscript: (1) the central research question probably cannot be tested at all, (2) the chosen dataset is arguably not appropriate, (3) the authors rest all their analysis on an overparameterized model which they have not put to the test thoroughly enough, and (4) the comparison between the lumped parameter models and the storage selection function (SAS) approach, which the authors tried to make systematic and thorough, is still too superficial, biased and on some points downright false.

1. The main research question addresses the issue raised by Stewart et al. [1], who argued that mean transit time estimates obtained from stable isotope measurements were potentially biased towards lower values compared to tritium-based estimates because of the input function's shape, which more or less repeats itself on an annual basis. While Stewart et al.'s remark was certainly valuable as a *warning*, I think that trying to prove their explanation wrong or right *once and for all* is a fundamental mistake and a misunderstanding as to how Stewart et al.'s paper is to be understood. Firstly because for practitioners, the most simple answer is to measure both tritium and one of the stable isotopes and check whether both estimated mean transit times agree. Secondly because this will probably depend on the watershed and on the degree to which the different storage compartments contribute to total tracer fluxes, and thus cannot be generalized in any way. In some catchments, both estimates will probably be similar, and in other, they will deviate significantly from one another. And thirdly, because to test this observation as a hypothesis in the real world, one would need a large dataset comprising sufficiently long time series of both tritium and one stable isotope for many different catchments, and certainly not rely on a single case study for one catchment (the same can be said of Rodriguez et al. [2]). But as I have found out recently, such datasets are extremely rare. Speaking of dataset leads us to the second shortcoming.

2. The watershed selected by the authors to study the research question of point 1 is that of the Neckar in Germany, a large basin covering an area of 13,000 square kilometres (!). While the authors have split up the area in four regions for the input calculation, each of these is still very large, and importantly, **no gauging measurements at the outlet of each region was used to contrain the models and a single output site was available**. I find it difficult on the one hand to point out that lumped parameter estimates of the mean transit time from the damping of the seasonal input signal can be affected by the combined tracer fluxes from subcatchments with largely different mean transit times (forgetting that the SAS have to be in the same situation, since the very same seasonal signal is used for parameter estimation), and on the other hand to then proceed to calculate mean transit times for such a

large watershed, where precisely such differences are extremely probable. Even with a spatially distributed input, one still in the end calibrates a model on a single time series of the system's output. In all honesty, I do not think this data set is adequate for the task, as it aggregates the responses of so many subwatersheds with potentially very different responses to rainfall and tracer injection, not to mention the fact that the isotope signal in the stream network itself will be delayed simply due to the time streamwater needs to travel from the headwaters to the gauging station.

3. The implementation of the SAS models (both the lumped model and the "distributed" model) are presented by the authors as successfully reproducing observations. I think this affirmation does not stand a thorough analysis of the modelling results. Firstly, the SAS models are probably largely overparameterized, needing between 11 and 19 parameters, all of which are solely constrained by tracer and discharge measurements done at the outlet of the Neckar basin. Because no unique solution can obviously be found, the authors are forced to rely on ensemble solutions. But even then, observations are not reproduced well. Streamflow peaks seems to be generally overestimated, while the modelled seasonal variations of tritium in particular is much more attenuated than shown by the data. Given the importance the authors give to a correct representation of variable flow as prerequisite for a truthful calculation of tracer fluxes, they are too uncritical of model results indicating that the hydrological model might not to work properly. That the hydrological model does not work is not surprising, given the size of the chosen watershed and the absence of stream gauges upstream of its mouth, which is bound to be a problem for any hydrological model working on daily time steps.

4. An important aspect of the manuscript is the comparison between three methods for estimating mean transit time from tracer data: the sine-wave method, the SAS approach and the lumped parameter models. The authors systematically stress the simplifications of the lumped parameter models and of the sine-wave method, which they present as so many weaknesses, but they often fail to recognize the similarities with, or indeed the downsides of the SAS approach. Also, the authors have sometimes missattributed important results, for instance concerning the role of the volume of tracer below the datum of the outlet, by ignoring publications from the eigthies and nineties. As for the implementation, the lumped parameter models have not been used to their full flexibility, which makes any comparison between them and the SAS questionable at best: (i) for the exponential model, a single exponential function was used, against the advice of Stewart et al. [3], who suggested to use a double exponential to reduce the "aggregation error" effect, (ii) only one of the two parameters of the gamma model was calibrated, and the other kept constant, (iii) variable flow was not simulated, although it would have been possible using the method proposed by Zuber [4]. The way the authors treat the issue of variable flow also ignores the argumentation of Zuber [4] and Zuber et al. [5] who suggest that if the total storage accessible to tracer is large compared to the variable storage, the steady state approach yields nearly the same mean transit time at the gain of one less fitting parameter. On the issue of fitting parameters, the authors have overlooked the potential advantage of using a lumped parameter model with three to four parameters compared to an 11 parameter model to fit two tracer time series. Potential problems due to model overparameterization should be taken much more seriously in the discussion. I think the comparison could be framed differently, and that the main difference lies not so much in SAS versus lumped parameter models, but rather in how they are implemented, and boils down to the number of fitting parameters one is willing to estimate from a measured tracer input and output. On the one hand, lumped parameter models are usually used in the most parsimonious way, requiring as few as two parameters (the mean transit time and the summer to winter infiltration ratio) to relate input and output (three in transient mode). Maloszewski, Zuber and colleagues in

particular have always insisted on the necessity to keep the number of fitting parameters as low as possible, because they believed in obtaining unique estimates (historically, this is what models had been used for in science since Galileo). On the other hand, SAS users sometimes rely heavily on confidence intervals of the estimated parameters, because the number of fitting parameter is larger (sometimes much larger) than can be constrained uniquely from the available data. This is the real difference, and it is one of culture rather than one of model choice. One could very well use the SAS parcimoniously. Consequently, a more telling comparison between the two approaches should (i) list out clearly the number of fitting parameters, (iii) use both approaches to their maximum potential and (ii) show the respective results of the fits. Recognizing and clearly listing out the *similarities* between both approaches would also be useful, instead of systematically presenting assumptions underlying the lumped parameter models as weaknesses, but without mentioning that the same holds for the SAS, or the other way around, claiming that the SAS is superior by pretending that something is not possible with the lumped parameter models. For instance, *both* approaches can be used in transient mode, and *both* approaches are lumped in the same way, as they relate input and output with one or more transfer functions that are not spatially distributed (different flow paths can be modelled though, by coupling models in series or in parallel, as shown in numerous publications by Maloszewski and colleagues).

To conclude, the manuscript in its present form misses out essential issues by overinflating the modelling results of the SAS approach while artificially curtailing the possibilities of lumped parameter modelling, and suffers from the use of a data set that is probably inappropriate for the task given the size of the watershed, which makes variable flow calculations doubtful and aggregates tracer response from largely different subregions.

Specific comments

L49: Basically, the problem lies in how to inject a tracer „instantaneously“ over the entire watershed, and not so much in the availability of„adequate observation technology“, since all one needs to do after the injection is to sample the output for long enough to reach near complete recovery. As a side note, the first catchment scale tracer experiment I know of is that by Rodhe et al. [6].

L50: The phrasing is too vague. Transit time distribution is EITHER inferred from input and output measurements ([7], [8]), OR assumed in order to calculate from input and output measurements useful catchment caracteristics (i.e. the mean transit time and the storage volume).

L52: Citations for the sine-wave method are missing, for instance, Maloszewski et al. [9].

L53: The lumped-parameter models may have been introduced for groundwater environments (see Eriksson [10]), but Piotr Maloszewski and Willibald Stichler in particular have used them early for surface water studies (see [11] and [12] for instance, but there are many more).

L56: The fact that the model representing the TTD must be chosen a priori has nothing to do with the steady-state approximation. These are two different things. A model must be chosen a priori in transient mode as well. And the authors could mention that the choice may be a priori, but is **not arbitrary at all**, and that model choice has to be guided by the boundary conditions and the sampling scheme. Additionally, SAS models are also based on an a priori choice for the selection functions.

L56: „While this assumption […]“. This sentence is much too vague and inaccurate. Firstly, Zuber [4] has clearly suggested in his paper presenting a transient approach for the lumped

parameter models that as long as the total storage accessible to tracer is large compared to the transient storage (what the authors refer to as „the temporal variability in the hydro-meteorological drivers“), then the steady-state approximation would yield nearly the same result as the transient fit. This hypothesis was then illustrated for a surface water case study of the Lange Bramke catchment [5], where this turned out to be indeed the case. This will of course depend on the local hydrogeological setting, but should be considered. Secondly, the variability in precipitation input is completely taken into account in lumped-parameter modelling, since the time steps of the input can be defined freely. So if daily data is available, nothing speaks against making calculations at that definition (whether this is such a good idea is another issue altogether). Thirdly, spatial heterogeneities in flow paths can **absolutely** be modelled using lumped parameter models by coupling them in parallel or in series, as was done routinely by Maloszewski and colleagues (starting with [13]). It is true that these potential heterogeneities are lumped together in a single measured output, but in that regard, the SAS face exactly the same limitation, namely that of extracting information from relating a single input to a single output. Fourthly, what do the authors mean by „misinterpretation“ ? Typically, the results of lumped-parameter modelling is a mean transit time of tracer and a storage volume which should be compared to the hydrogeological information available concerning porosity. It is not the model results that are misinterpreted, but rather, model results can be wrong if an inapropriate model has been chosen, for instance.

L59: Given the constant string of publications, in particular by Maloszewski and colleagues, over thirty years, exploring systematically the possibilities and limitations of such models, I think one cannot seriously argue that they lack a coherent framework.

L61: „without the need“. This is phrased as if the SAS approach could do away with a priori model choice. But then, in the next sentence, one learns the exact opposite. The SAS, just like the lumped parameter models, have at their core a series of functions necessary to relate input and output, so in that regard, they are the same, and trying to present the one approach as „freer“ from a priori choices as the other is incorrect.

L61: „change in water storage are considered“. So are they using the transient approach proposed by Zuber [4]…

L64-65: The explicit tracking is different from unsteady state, and should not be confused with it. By setting a constant storage, SAS can be used in steady state mode.

L75: „The second type […]“. This needs qualification. Dating can be done in two different ways using tritium. Either one takes advantage of the tritium peak resulting from the atmospheric bomb testing of the 1950s and 1960s, or now, in the post-peak era, from the shift between the mean annual input and output due to decay losses in the subsurface. Please note that other radioactive tracers used for dating such as krypton 85 display a steadily increasing trend since the 1960s, and as such, it is not so much the decay than the rate of increase that is used for dating. The same holds true for non-radioactive tracers such as the chlorofluorocarbons.

L85: The entire paragraph seems a bit out of place in an introduction. Why so many details concerning the upper limit of the sine-wave method?

L96: How is that back-of-the-envelope-calculation done ?

L97: The sensitivity of the sine-wave methods have nothing to do with potential aggregation biases, these are just two different issues.

L125: Is three years of measurements for a tracer that varies on an annual basis so bad ? This is three replicate. The handful of tritium measurements was enough for dating in the 80s when the decrease over time was still steep.

L127: What do the authors mean by „precluded“ ? The exponential model describes a continuous distribution of transit times from zero (for flow lines close to the outlet) to infinity (for flow lines near the watershed divide). How does that preclude longer transit times ? And since in the studies cited the same models have been calibrated for both tracers, the underlying distribution of transit times is also the same.

L128: „in a spatially lumped way“. Yes, but for the SAS, one also uses a „lumped“ input. And Maloszewski et al. [9] for instance modelled two separate reservoirs as well as quickflow „with a turnover time up to hours or days“, so not quite lumped. And how probable „aggregation problems“ are might depend quite significantly on the size of the watershed, and how smart the isotopic sampling was done (for instance by measuring the output at the outlet of different reservoirs within the watershed).

L135: Looking at the graphs showing modelling results in Rodriguez et al., I find it striking how bad the fit is. Sure, most measurements are within the confidence intervals, but this is masking the fact that the best solution misses most of the individual data points. Given this, how much credit should one give to the comparison of mean transit times done by Rodriguez et al. ?

L139: I agree with Stewart et al.. Given the constant average value of tritium over the seven years of measurements in the Weierbach catchment, one has to conclude that the tritium peak has already been flushed out, which indicates mean tracer transit times of a few years at most, i.e. a negligible flux from flow lines with transit times longer than that.

L146: What do the authors mean by „integrated“ ? That both the tracer and water fluxes are modelled ? If one is interested in studying tracer storage and release dynamics, why try at the same time to reproduce measured discharge as well instead of using it as constraint ? Adding a hydrological model to the model describing tracer transport is bound to complicate the parameter estimation procedure and increase the overall „uncertainty“ by increasing the number of parameters needed fitting. And lumped parameter models are also „process-based“, since the transit time distribution should be chosen to reflect the hydrogeological situation, and in the case of variable flow, the tracer fluxes explicitly depend on storage volume, which controls discharge out of the system.

L147: Since lumped parameter models can also be used in variable flow situations, why did the authors not do it for a fair comparison ? It is a bit like comparing two racing cars, but with one of them forced to stay in first gear for the entire race.

L150: I do not think that Stewart et al. meant that the bias in estimated mean transit time is “systematic”. Rather, they warned that this might be the case more often than not, and that one should be aware of this, and if possible use both tracers simultaneously. Or to put it in a different light, if the actual transit time distribution does not deviate too much from the theoretical model, both estimates should be about the same. So maybe this is making much ado about nothing, and wanting to prove more than can actually be proven. Also, how do you generalize the acceptance or the rejection of this hypothesis for one catchment to all possible catchments ?

L153: Choosing an extremely large watershed, displaying an elevation difference of nearly a thousand metres and a precipitation difference of 900 mm per year, with an isotopic signal

potentially influenced by snow fractionation, may not be the best choice considering the limitations the authors have described before.

L174: Since the output was only available at the downstream end of the Neckar, near its confluent with the Rhine, only the input was roughly spatially distributed. For such a large watershed, I think this is a serious limitation of the data set, as the output lumps together so many different subwatersheds with different characteristics and hydrological responses. This seems contradictory to the warning higher in the text about “aggregation problems”.

L210: I understand the desire to take spatial variability of the input into account, but using kriging adds more parameters and more a priori decisions to the modelling.

L253: It is a pity that using lumped parameter models in transient mode was not considered in the step-wise approach adopted here.

L273: Another common lumped parameter model is the dispersion model. Given the size of the watershed and the large macrodispersion to be expected, using it too might have been useful.

L281: Why is “a priori” italicized here, but not even mentioned on line 365, where the authors chose *a priori* a uniform distribution for the SAS functions ? To be clear, one or more functions describing the storage of the tracer within the watershed are needed for LPM and SAS approaches, and they have to be chosen a priori. But the choice for the LPM is NOT arbitrary, as the transit time distributions can be derived from mass balance and groundwater hydraulics, be it the exponential, gamma or dispersion model. In that regard, the SAS approach is less process based, not more, as to my knowledge, there still is no physically-based justification for choosing uniform rather than gamma functions or anything else to describe how tracer is released from storage. Effectiveness (against which hard constraint ?) is too vague a reason, and numerical convenience as mentioned on line 371 is even a bad one.

L290: Since the authors kept alpha at 0.5, it is strictly speaking not a calibration parameter.

L303: I find the description of the hydrological model too superficial for an element that is essential for calculating variable tracer fluxes.

L320: For a watershed of the size of the Neckar, not all water entering the channel on day “t” will exit on the same day, so channel routing becomes necessary as well. Was this implemented here ? Judging from figure 2, it does not seem to be.

L374: This is seen from a modeller’s perspective, but could also be explained physically, as Zuber has done in his 1986 paper [4].

L375: This is Zuber’s [4] “minimum volume”. Please cite his paper.

L383 : The description of the sine-wave model is 11 lines long, that of the lumped parameter models 15 lines long, and that for the SAS model 88 lines long, which reflects well the difference in complexity. I wonder whether the data available warrants such a complex approach requiring so many fitting parameters.

L397: The implementation of the spatially distributed model requires many assumptions and additional parameters (8 compared to the “lumped” SAS model, which already has 11), all of which are solely constrained by a single measured output for discharge and two tracers at the outlet of the entire watershed with a total surface area of 13,000 square kilometres. Is this reasonable ?

L416: Why choose daily time steps, since the tracer data is available on a monthly basis and the stream gauge is situated at the outlet of a 13,000 square kilometres watershed ? Coarser time steps might also reduce the problems of overestimation of the discharge shown on figure 5.

L421: Rainfall-runoff modelling is a whole branch of hydrology in itself, and here, the authors have coupled it with a tracer storage and release routine. Isn't this adding up difficulties instead of reducing them ? And should not the authors be more critical of modelling results obtained with relatively little data with which to constrain the numerous model parameters?

L455: Why relegate the graphs showing the fits of the "base line models" in the supplementary material ? This does not help the reader to make a judgment for himself concerning the quality of the respective fits.

L460: That seasonal fluctuations are not reproduced without adjusting the fluxes to storage variation is not surprising. But for tritium dating, this is of no importance, because the passing of the tritium peak and the tritium decrease over time is what is used for fitting. See Zuber et al. [5] for a discussion of this.

L467: Obviously, an 11-parameter model will in many cases yield a better fit than a 1 parameter model. But avoiding overparameterization is also important in a sound scientific approach.

L471: It is not surprising that seasonal fluctuations are better reproduced by a model that takes seasonal variations in storage into consideration, compared to a model that does not. The same behaviour could most probably be obtained by using the lumped parameter models with a variable flow formulation (see Zuber et al. [5]).

L475: Same question as above. A hydrological model with 9 free parameters should reproduce relatively well any stream hydrograph, if only one stream gauge is considered, but an important question is whether the data is sufficient to constrain the model parameters in a way that is meaningful, and not artificial parameter tweaking.

L480: A couple of comparative graphics might do better than this long and rather tedious analysis of the respective model performances.

L485: It is not really surprising for a watershed of this size that the departure of the transit time distribution from an exponential model is large enough to lead to a discrepancy between estimated mean transit times. The authors could have taken up Stewart et al.'s [3] and Farlin and Maloszewski's [14] suggestion and used a double exponential to take this potential departure into account. Alternatively, varying the alpha parameter of the gamma model might have allowed a combined good fit to both oxygen-18 and tritium by increasing the weight of the very short transit times. A graph showing the fit is essential in the main text, rather than relegated in the supplemental information.

L486: Transit time distributions are not explicitly defined in the SAS approach, but since the selection functions are, transit time distributions are still implicitly defined.

L488: The importance of storage volume for the mean transit times was indeed shown, but by Maloszewski and Zuber in 1983 [13] and Zuber in 1986 [4].

L490: Zuber [4] was I think the first to clarify the importance of what the authors call "passive storage volume" in isotope hydrology.

L510: What do the authors mean by "broadly" ? The MTT range estimated using the lumped parameter models is *within* the larger range of the sine wave method.

L513: The fraction of younger water used to be applied loosely relatively to “older” water. But I suppose the authors refer here to Kirchner’s young water fraction [16], in which case, they might want to cite his paper, and correct the definition to between 2 and 3 months. Incidentally, the notation “F(T<3 m)” has not been defined previously (one has to guess the “m” stands for “months”, for instance).

L515: Nothing conclusive can be gained from this comparative analysis, since the setup of the lumped parameter modelling was artificially kept to a bare minimum, ignoring more complex possibilities such as taking into account variable flow rates or combining models (here for instance two exponential, or allowing the alpha parameter to vary, not to mention running the convolution with a variable storage volume).

L543: The equation relating mean transit time and storage volume can be found in Maloszewski and Zuber [13]. The phrasing is slightly misleading, as it implies that storage estimation is only possible with the SAS, which is not correct.

L566: But the authors have failed to follow up on Stewart et al.’s [3] suggestion to use a double exponential in combination with both tracers.

L576: This line of reasoning seems very biased to me. The point is that in order to simulate both tracer and water fluxes, the SAS need 11 to 19 parameters, all of which must be constrained solely by three time series (two tracers and discharge), all measured only at the outlet of a huge watershed. And one could very well (i) calibrate a lumped parameter model simultaneously for both tracers, as this only depends on the optimization procedure chosen, and (ii) estimate from the discharge measurements the additional parameter needed to add variable fluxes to the convolution. With the lumped parameter approach, this would be three to four parameters, depending on model choice.

L604: Before concluding that lumped parameters “are incapable of extracting meaningful information” from stable isotope measurements, the authors should first use lumped parameter models to their full potential.

L605: In the scientific method, “anecdotal evidence” may be useful initially to recognize a problem, but has no place in the argumentation that should follow the first hunch.

L616: The basis for the authors’ argument is provided by using lumped parameter models inappropriately, and hence, cannot stand as solid evidence.

L619: Maybe, but then why haven’t the authors made use of the possibilities offered by lumped parameter models to consider transient flow and hydrological information ? The authors have arrived at the conclusion that the hypothesis can be rejected only by ignoring most possibilities offered by lumped parameter models.

L637: Actually, what James Kirchner meant was that estimating the mean transit time using the damping of the amplitude of a seasonal tracer measured at the outlet of a watershed where subwatersheds display dramatically different mean transit times can be completely erroneous, because the relationship between mean transit time and damping is not linear, whereas tracer mixing is. I see no reason why the SAS should not be just as prone to this kind of error, since the method also adopts a simple input-output approach. Splitting up the catchment into sub-regions does not change this if only done for the input. And given the size of the cathment, this problem might even be extreme. Or do the authors expect on the opposite that the size of the basin smoothes out subcatchment differences ? This is worthy of a much more thorough consideration in the discussion, and the authors should at least give solid qualitative reasons for neglecting aggregation problems.

L675: All conclusions reached in this paragraph are based on (i) a simplistic implementation of lumped parameter models that is far from the state of the art and (ii) the reliance on a overparameterized SAS model that fails to reproduce both tracer and discharge dynamics. All this should be redone from the ground up.

L1030: For both lumped parameter models used, the exponential and the gamma functions, the authors calibrated one parameter, and consequently ended up with a single best fit. Isn't that something like an advantage in a way ? Using more parameters that could be independently determined used to be a no go in hydrology up to the turn of the century. Also, the gamma model has actually two free parameters, not one, so keeping the alpha parameter constant at 0.5 is an a priori decision that seems strange after the authors' warning against a priori decisions concerning lumped parameter models further up in the text. And the authors have not considered the winter to summer infiltration ratio, which often shifts the mean annual isotope values towards the winter average [16]. Concerning the number of parameters, the SAS models used have between 11 and 19 parameters, compared to the one parameter for the lumped parameter models (two for the gamma, plus one if considering the winter to summer recharge ratio, plus one if making unsteady state calculations, which should have been done to exploit fully the possibilities of lumped parameter models in variable flow systems and allow a fair comparison with the SAS results). Given the data set used for parameter estimation is the same and consists only in measured inputs and outputs to two different tracers, are not the results of the lumped parameter models, being much more parsimonious, also much less uncertain ? Not trying to reproduce discharge, but only focusing on the isotopes, could help reduce the number of fitting parameters of the SAS models.

Figure 3:

There seems to be systematic offsets between observed and measured isotopic values, for instance in 2009, 2010 and 2011 for scenario 12. Generally speaking, the graphs composition makes it difficult to see how good the fit is. Please make the points smaller, and use lines instead of points for the predicted response.

Figure 4:

1. In the text, the authors write that the SAS models reproduce well, not only the general trend, but also the seasonal variability of tritium activity. This is not at all what plates b) and d) of figure 4 show. On the contrary, the observed seasonal variations is visibly larger than the modelled ones, which only show the slightest hints of an intra-annual amplitude. For instance, the hump observed in 2008 is missing, just like the peak in 2009. The same holds for the oxygen-18 time series shown on figure 3.

2. The 5^{th} to 95^{th} percentile of optimal solutions also clearly show that the optimal solutions often completely miss the observations, with a range that can be off by 10 TU and more. These are clearly not visually acceptable solutions, no matter what the measures of fit are.

Figure 5:

Plates b) and c). This type of representation does not allow at all to see whether the modelled discharge matches well the measurements or not, especially when events are all shoved together due to the lengths of the time series. A plot of the residuals against time, would be much better at this. Generally though, the model seems to overestimate peak discharge quite systematically. This overestimation will then necessarily be propagated to the calculations of the tracer fluxes, and seems incoherent with the authors' insistence on the importance of variable transit time distributions, since their analysis relies on a model

that systematically exaggerates peak discharge, and hence probably overemphasizes rapid tracer flushing. Furthermore, this again brings up the question whether trying to model both water and tracer fluxes is not a serious flaw, instead of taking the former as given and concentrating on the later, which would be the approach proposed by Zuber (1986). Why want to model discharge if it has been measured, as the aim is not to reproduce streamflow, but to study tracer export ?

Figure 7:

1. Average transit time distributions for the SAS models were estimated by fitting a gamma model to the SAS modelling results, which according to the authors reproduce observations best (this is another point which needs to be discussed separately, see comments for figure 4). Since the gamma function is one possible model for the lumped parameter approach, then an appropriate parameter estimation procedure should be able to find this solution using lumped parameters directly (probably modelling variable discharge). So the fact that this solution was not found using lumped parameters, but via the SAS modelling, does not mean lumped parameters are inadequate, but rather that the modellers have not used them to their full potential.

2. There is much to say regarding the distribution of transit times that has been missed by the authors. Taking the best fit from the sine wave approach as one extreme (most of the weight at shorter transit times), and the best fit of the tritium fit to an exponential model as the other extreme (most of the weight at larger transit times), one can clearly see that the other distributions' weighting of transit times lies in between. Now using the transit time distribution concept, how can one explain both a relatively small damping of the seasonal signal for the stable isotope and a decrease in tritium activity over time that indicates a relatively large storage ? By placing a significant weight at shorter transit times AND at large transit times, and relatively less in the middle range (here between half a year and three years). The exponential model cannot do this, but the gamma model can, if appropriately parameterized (i.e. allowing BOTH the alpha and beta parameters to vary). Alternatively, a double exponential (as suggested by Stewart et al.) could also simulate this distribution. Not surprisingly, this is the kind of distribution that the best solution obtained from the SAS results show.

References

[1] 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, Hydrol. Process., 24, 1646-1659, 2010.

[2] Rodriguez, N. B., Pfister, L., Zehe, E., and Klaus, J.: A comparison of catchment travel times and storage deduced from deuterium and tritium tracers using StorAge Selection functions, Hydrol. Earth Syst. Sci., 25, 401-428, 2021.

[3] 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 an groundwater systems, Hydrol. Earth Syst. Sci., 21, 4615–4627, 2017

[4] Zuber, A.: On the interpretation of tracer data in variable flow systems, Journal of Hydrology, 86 (1-2), 45-57, 1986

[5] Zuber, A., Małoszewski, P., Stichler, W., and Herrmann, A.: Tracer relations in variable flow, 5th International Symposium on Underground Water Tracing, IGME (Institute of Geology and Mineral Exploration), Athens, 355-360, 1986

[6] Rodhe, A., Nyberg, L., and Bishop, K.: Transit times for water in a small till catchment from a step shift in the oxygen 18 content of the water input. Water Resources Research 32 (12), 3497–3511, 1996.

[7] Visser, A., Broers, H.P., Purtschert, R., Sültenfuß, J., and de Jonge, M.: Groundwater age distribution at a public drinking water supply well field derived from multiple age tracer (^{85}Kr, ^{3}H/^{3}He, and ^{39}Ar), Water Resources Research 49 (11), 7778-7796, 2013

[8] Massoudieh, A., Sharifi, S., and D.K. Solomon: Bayesian evaluation of groundwater age distribution using radioactive tracers and anthropogenic chemicals, Water Resources Research 48 (9), 2012

[9] 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 (1-4), 319-330, 1983

[10] Eriksson, E.: The possible use of tritium for estimating groundwater storage, Tellus 10, 472-478, 1958

[11] Małoszewski, P., Rauert, W., Trimborn, P., Herrmann, A., and Rau, R.: Isotope hydrological study of mean transit time in an alpine basin, Journal of Hydrology 140, 343-360, 1992

[12] Stichler, W., Małoszewski, P., and Moser, H.: Modelling of river water infiltration using oxygen-18 data, Journal of Hydrology (83), 1986, 355-365

[13] Małoszewski, P., and Zuber, A.: Determining the turnover time of groundwater systems with the aid of environmental tracers, I. Models and their applicability, Journal of Hydrology (57) 207-231, 1982

[14] Farlin, J., and Małoszewski, P.: On using lumped parameter models and temperature cycles in heterogeneous aquifers, Groundwater 56 (6), 969-977, 2018

[15] 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, Hydrol. Earth Syst. Sci. 20 (1), 279-297, 2016

[16] Grabczak, J., Małoszewski, P., Rożanski, K., and Zuber, A.: Estimation of the tritium input function with the aid of stable isotopes, Catena (11), 105-114, 1984