This paper has already been through a number of thoughtful reviews. Here I will focus on a few contentious elements rather than provide a complete review.

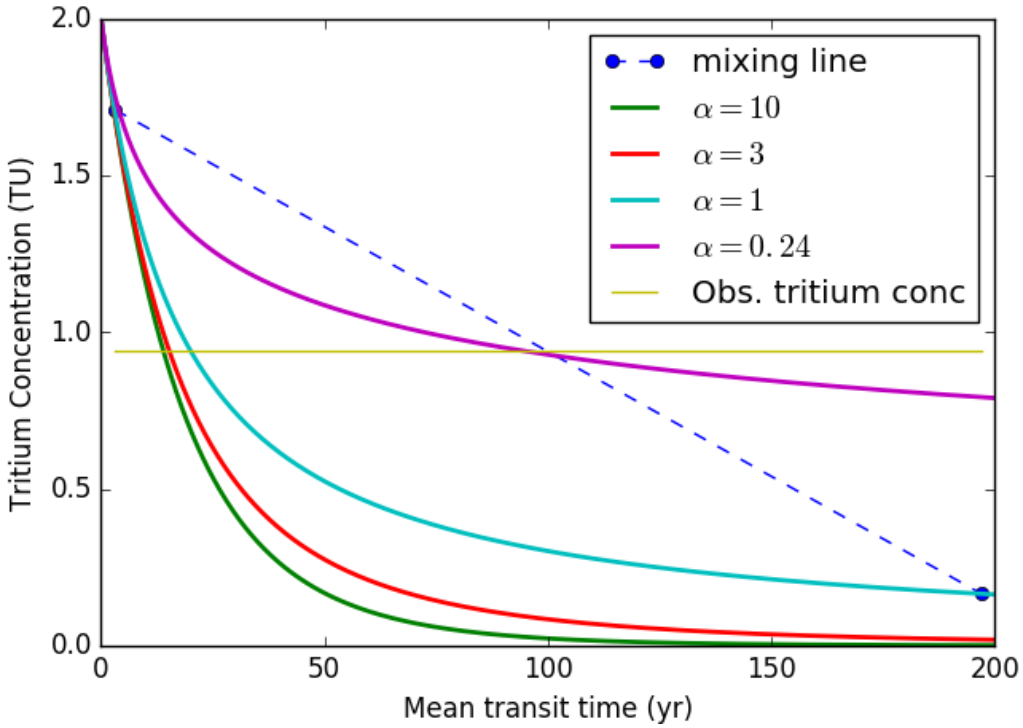
This paper aims first to extend the results from Kirchner (2016) on the effects of 'aggregation errors' on transit times estimated from seasonal cycles of stable water isotopes (or other passive tracer) to water ages estimated from tritium observations. I understand the "aggregation error" (as Kirchner used the term) as error in data interpretation arising from a poor choice of probability distribution for the transit time distribution (or 'Lumped Parameter Model' – LPM – as the groundwater community prefers) in certain circumstances. Specifically, it is the case where the sampled water contains contributions from sources whose individual transit time distributions have very different means. If the chosen distribution for the combined sample is unable to represent the breadth (variance and skewness) of the combined transit time distribution it will provide biased estimates of the mean transit time.

If the aims of the paper were simply to point out that spatially variable transit time distributions have similar effects on Tritium observations as Kirchner found they did on stable water isotopes, the paper could be a useful contribution – particularly if the robustness of the <18 year old fraction were more convincingly established. However the paper aims to have further-reaching conclusions regarding the superiority of 'compound' LPMs, and I have some slight issues with these as they currently stand – but I believe these can be remedied with some revision.

The heart of the analysis is the set of virtual experiments described in section 2.2. I think there is a limitation with the virtual experiments described, and which has perhaps led to some of the contentious reviews in the previous round. Specifically, the experiments are conducted by combining two 'simple' LPMs to predict a stream tritium concentration. This concentration is then analyzed by assuming a single 'simple' LPM again. The results show that this leads to significant bias. The authors conclude that 'simple' LPMs are susceptible to aggregation errors, and that a 'compound' LPM is required.

The problem with this is that their choices for the assumed 'simple' LPM are unnecessary constrained. As far as I can tell, the analysis is restricted to cases where both the 'true' LPM used to construct the data, and the 'assumed' LPM used to analyze it, are identical in form: both piston (Dirac delta distributions) or both gamma distributions with identical shape parameters (and always shape parameters greater than 1). The authors do not consider the case where (for example) the 'true' LPMs are exponential, but the 'assumed' LPM is a gamma distribution with a shape parameter less than 1. There is no reason not to do so that I can think of, since in practice we will not know the underlying distributions of the contributing parts. If we do not know the 'true' LPM, we are at liberty to adopt any physically reasonable distribution for the 'assumed' LPM.

In this case, it is possible to choose a shape parameter for which there is zero aggregation error (at least in terms of mean age). The figure below is identical to Figure 3a (where alpha is 1) but with the curves for alpha=3 and 10 (like figures 3 b and 3c) included, along with the curve for alpha=0.24, which was not considered by the authors. For the latter case there is zero aggregation error. The predicted tritium concentration and MTT (100 years) of the 'simple' assumed LPM are both almost exactly that of the 'compound' true LPM.



This seems to contradict the conclusion of the paper that suggests that 'simple' LPM should not be used because they have higher aggregation errors. Here a simple LPM perfectly reproduces the MTT of the aggregate. It is able to do so because the small alpha gives it a large variance and skewness, enabling it to capture the influence of the young and old components. Note that, of course, it is not an accurate representation of the true TTD – however the data (a single tritium observation) is insufficient to determine that. Also, I was only able to choose the 'right' value of alpha because I know what I was aiming for.

As the authors point out a compound distributions is a sensible choice where distinct sources can be identified, and the partitioning of flow between them is known. It is useful to be able to incorporate such information into the model used to interpret the data. However in the absence of such auxillary information there is nothing fundamentally different about 'compound' LPMs that makes them immune to aggregation errors in some way that other type of distribution are not. There are many other distributions that are also reasonable choices even in the case of large

heterogeneity, so long as they have enough flexibility to accommodate larger amounts of skew than an exponential or other type of distribution tested by the authors can.

Of course, in the absence of such auxillary information the compound LPMs and other flexible distributions will have multiple free parameters that must be estimated. As the number of free parameters increase, the ability of the model to reproduce calibration data inevitably increases, without necessarily increasing the physical realism of the calibrated parameters, or any metrics derived from the model (like the mean). This brute fact must always be acknowledged and dealt with by those urging us to adopt more complex models, and is not dealt with here (as several reviewers of the previous version pointed out, to no avail).

In conclusion, this paper makes a useful contribution, and I believe it warrants publication if the authors are able to clarify the issues raised above, and pay a little more attention to the trade-offs associated with increasing the number of free parameters.