“Aggregation effects on tritium-based mean transit times and young water fractions in spatially heterogeneous catchments and groundwater systems, and implications for past and future applications of tritium” by Stewart, M.K., Morgenstern, U., Gusyev, M.A., Maloszewski, P., Hydrol. Earth Syst. Sci., doi:10.5194/hess-2016-532.

Authors’ Response to Referee #1

We appreciate the many helpful comments of Anonymous Referee #1

Ref #1: The paper builds on the publications by Kirchner (2016) analyzing the effects of aggregation of transit times distributions on estimating the MTT from 2 systems using tritium as a tracer. In addition, the paper also discusses the effects for a couple of examples from the literature. The paper is well written, however, it could be improved when better linking the examples with the analysis of the aggregation effects, considering from the beginning the use of compound LPMs and improving the readability and celerity of the figures. The list of comments and ideas below should help to improve the paper to be published in HESS.

Reply: We will aim to link the aggregation effects in the case studies better with the aggregation effect calculations, introduce the compound LPMs earlier and improve the readability and clarity of the figures.

Ref #1: General comments: 1) The definition of heterogeneity needs to be clearer in this paper. Most LPMs assume a certain heterogeneity in the system, otherwise there would not be a transit time distribution – the specific effect considered in this paper regarding the aggregation effects are two distinctly different systems, each heterogeneous, but usually separate flow systems resulting in the investigated effects – this could be – as nicely shown in the examples, a shallow and deep GW system, a confined or unconfined system etc. So, it would help to make this clear, heterogeneity is everywhere, but you are only analysing a specific set of heterogeneity with you analysis.

Reply: Simple LPMs are based on the assumption that the hydrological system is homogeneous – this gives rise to the transit time distribution described by the LPM. Our virtual experimental system assumes two different homogeneous subsystems each with their own transit time distributions (TTDs), which are allowed to range from being the same to being very different as shown by their mean transit times (MTTs). Having only two subsystems is the most extreme system, which will show the largest aggregation effects when the MTTs of the subsystems are most different. If there were more subsystems, aggregation effects would be smaller because of averaging over the larger number of subsystems. We agree that this needs to be pointed out.

Ref #1: 2) As you mention at some locations in the paper, a compound LPM usually addresses the possible effects of the aggregation error – at least this is how you define it in the example section. Hence, there is no aggregation effect if we would always use compound LPMs – you partly propose this in section 4, but in my opinion, this is not strong enough. If we apply a compound LPM and the two compounds are quite similar, we could also use a simple LPM (maybe simple is not a good phrase, I would prefer single – and there will be no aggregation effect. However, if the two compounds are different, a single LPM would result in a strong bias. Hence, we should propose to always use compound LPRs and then analyze the results in order to deal with the aggregation effect.

Reply: A compound LPM (which is a combination of two simple LPMs) is not likely to be a perfect choice for any particular system, but it will certainly give much less MTT aggregation error than a simple LPM in systems which are prone to aggregation error (i.e. have substantial contributions of water with very young MTT combining with water with old MTT). This topic is also addressed in the replies to Refs #2 and #3. A possible problem with compound LPMs is that there may not be enough data to constrain their parameters.

Ref #1: 3) Why are you only using the standard deviation as your objective function. This considers only part of the fit – maybe you should consider the paper of Kling and Gupta or others to better select an appropriate function to be used, in particular when comparing the different models in Table 1.

Reply: We have now used the Nash-Sutcliffe Efficiency (NSE) and will add this to Table 1. This gives fits at the same model parameter values as the standard deviation. The qualities of the fits (i.e. the NSE values) however, show different patterns.

Ref #1: 4) The definition of the young water fraction is problematic. Why should the threshold (ty) change with changing alpha. So, we would always need to define this ty and then apply it differently to the different sets or models etc. If we use something like the young water fraction, we should define it based on a fixed value that is related to the analysis and the question behind, However, for me the young water fraction is just another measure in order to avoid displaying the whole transit time distribution – this is what we need (as done nicely in Fig 10 – this should be repeated for the other examples as well) and not another factor in addition to the MTT. It would be helpful to discuss this and also to better clarify what the assumption of a non-constant ty would mean with respect to the whole analysis.

Reply: This is a good point. After submitting the manuscript, we experimented with keeping the threshold (t_y) constant and found that the value of 17.5 years gave consistent results for the young water fraction (YWF) over the whole range of α (0.1 – 10). This will be used in the revised version of the paper.

The larger story here is that we have now changed the method of finding the apparent TTDs by fitting gamma models (GMs) to the mixed tritium concentrations by varying both of the GM parameters (i.e. α and β, with MTT= α*β). Previously (as described in the Discussion paper) the apparent fit was found by varying β only. Aggregation effects on MTTs are similar using both methods, but the threshold (ty) for determining the YWF is now better kept constant at 17.5 years with the revised method.

However, we agree that getting the TTDs right is the important thing if it can be done, and using compound LPMs is an important step in this direction. Then the MTTs are likely to be more accurate. (The point of the YWF was that with the right threshold it would be approximately correct regardless of whether the TTD was correct or not. On the other hand, the MTT would be incorrect if the TTD was incorrect.)

Ref #1: 5) The whole section 3.3 is not necessary and should be removed including the figure. It is sufficient if you discuss without showing a similar analysis as already done in Kirchner (2016).

Reply: This section is short, and we feel it is better kept in to establish the link with Kirchner's (2016) work on seasonal tracer cycles. Although our methodology was the same as Kirchner's in that two components were combined, we followed a process of starting with the same MTTs and then allowing the second component to become older. In addition, this section shows the dependence of the aggregation error on the difference in MTTs more explicitly than the random sampling of non-similar MTT components method of Kirchner. We will consider expanding this section with the above text in the revised paper.

Ref #1: 6) as already mention in some other points, you should make the examples better comparable in terms of defining the compounds LPMs, but also in respect to the figures and results shown. I think you should at least do the following: a) use the most common single LPMs (EPM, EM, GM, DM) and the best compound (or even better also a reference compound model) model and fit it to the observed TU showing the resulting fit (including a quantification of the fit) and the resulting TTD of each model. b) based on these figures, discuss the possible aggregation effects based on the models and the objective functions. c) define MTT, young water fraction for all model combination (in figure or table).

Reply: We will do further work on these case studies with these recommendations in mind. Changes will also be made in response to the comments of Refs #2 and #3.

Ref #1: 7) In the discussion, you should clearly differentiate what is necessary or problematic when using single LPMs or when using compound LPMs. For example, section 4.1 is only relevant when applying single LPMs.

Reply: We will carefully consider this in the rewriting.

Ref #1: 8) the statement in section 4.3 that individual tritium measurement can/should be used to estimate a series of TTDs for one system is in my opinion not helpful. We should rather use the series of tritium measurements to apply the best possible compound LPRs that is necessary to describe the heterogeneity of the system as nicely presented in several of the examples. A single tritium measurement is not helpful at all, as there is too much ambiguity. This could only be solved when combining the single tritium measurement with other tracers to identify the individual systems (multi components) and hence the complexity of the system studied. Hence, the whole issue of aggregation error will be solved to begin with.

Reply: We (or some of us) have tended to promote use of individual tritium measurements to determine TTDs (at least in the Southern Hemisphere), but we agree that there are often fishhooks especially in Northern Hemisphere systems, e.g. Japanese headwater catchments demonstrated by Gusyev et al. (2016). We support use of other tracers and other types of measurements on hydrological systems, e.g., time-series sampling, to resolve these ambiguities.

Ref #1: Specific comments: 1) The title is very long and awkward – it would be helpful to focus on the implications and less on the different ideas of aggregation – please shorten considerable.

Reply: Agreed, we will change this.

Ref #1: 2) P2/L32 – the tritium method does not only depend on the radioactive decay, but still also on the bomb peak: : :.

Reply: Bomb tritium is not required for tritium dating. Detecting the bomb peak was useful for dating in the past, but it is no longer useful in that way. Now the bomb peak just complicates the age dating based on radioactive decay of tritium and natural cosmogenic tritium is sufficient for dating.

Ref #1: 3) P5: The list of LPMs should be more consistent and maybe best in a table listing all used LPMs and possible compound LPMs. In the moment, this section is incomplete, as there are additional compound LPMs in the result section not introduced in the method. You should also consider a systematic use of the compound LPMs – 2P_EM, 2S_EM, or EM_P-EM, EM_S_EM, PFM_P_DM etc, I think this would generally help to get a better understanding how the compound LPMs are set up, in particular the one introduced in the example section.

Reply: We agree that the methods section should be tidied up, but feel that a comprehensive listing would be beyond the scope of the paper. We will re-write Equations (8 & 9) in more general form of LPM_1 and LPM_2 instead of EM allowing the use of any LPM in the equation. The TPLR will be given as an example.

Ref #1: 4) P7: why did you select 3 and 197 years for the analysis of the aggregation effect. It would be better the select a more realistic difference, which is commonly observed in the tritium analysis. In addition, you could also use TU/TU of Input as a relative measure in order to avoid defining the constant 2 TU for the analysis.

Reply: We selected MTTs of 3 and 197 years to display the relationships between mean transit time and tritium concentration in Fig. 3 because these (1) give a true MTT of 100 years, and (2) there is a good big difference between them. Any two MTTs would do as long as there was a sufficient difference between them.

Ref #1: 5) P8/L21: it is not clear to me how the "error of fitting" was determined – please define this in the method section and do not only show it for one selected combination but for all.

Reply: The "error of fitting" is the error of the apparent MTT resulting from fitting a simple LPM (a GM) to the tritium concentrations of the mixture of the two subsystem waters. It is one standard deviation. The error decreases as MTT1 increases (and the difference between MTT1 and MTT2 becomes less). Hence we have only shown it for the two lowest MTT1s. However, we will show it for all cases and add a brief explanation in the revised manuscript.

Ref #1: 6) P8/L25-30. The discussed small difference between the northern and southern Hemisphere are very difficult to see, as the results are shown in separate figures – either combine the figure to see the effects or use only the figure from one location and mention that the results are similar for the other site (maybe with some kind of quantification).

Reply: We will consider some quantification of the small differences.

Ref #1: 6) P9/L4: I assume you mean 197 years.

Reply: Yes, we fixed the typo.

Ref #1: 7) P9/L19: mention that alpha is from the Gamma model – not directly clear in this context.

Reply: We will clarify the meaning in the revised manuscript as suggested.

Ref #1: 8) P12/L3: The assumption that the old water component includes a zero tritium concentration is not clear for at all – is this possible at all?

Reply: Zero tritium concentration means less than the detection limit for tritium which depends on the measurement error at low tritium concentrations (usually taken as twice the standard deviation of measurement). Certainly 1200-year-old water would have far far less tritium than the detection limit.

Ref #1: 9) The figure should be improved considerably – the layout is changing, the borders are annoying, the arrangement is not consistent, points cannot be seen (e.g. Fig 6 and 7), style is changing in the paper (Fig 1-9 and 10-end).

Reply: The figures will be improved as suggested.

Ref #1: 10) Please describe the different lines in Fig 3 in a legend.

Reply: Agreed, we will add a legend in Fig 3.