We appreciate the many helpful comments of Anonymous Referee #2, and reply to the detailed comments below.

Ref #2: In their manuscript, Stewart et al. investigated tritium-based estimates of mean transit times (MTT) and the fraction of young water (Yf) in light of aggregation bias due to catchment heterogeneities. Furthermore, past studies are reinvestigated and evaluated in respect to aggregation bias. This topic is highly interesting, as most commonly the stable isotopes of water (Oxygen-18, Deuterium) are applied in tracer studies. In comparison to this, tritium is used more seldom, but it has the potential to elucidate longer transit times, where stable isotopes hit a boundary at about 4-5 years. I hope my comments and suggestions will be helpful to the authors and improve the manuscript.

Reply: We thank the Referee for this summary.

Ref #2: General Comments

1) Manuscript structure: I found the structuring of the text to be all over the place, making it hard to read for me, as I was expecting to have all the tools necessary to understand the paper after reading the Methods. However, the “Results” section basically starts with several paragraphs of new Methods. I would suggest to either changing the order of the text to properly divide Methods, Results and Discussion, or rename the header titles from “Results” and so on to something else to avoid confusion. Please see specific comments about my ideas which paragraph could be shifted to different sections.

Reply: We agree with this comment and will revise the Methods, Results and Discussion sections to improve the readability of the paper. The last paragraph of the Introduction section will also be re-written to reflect the updated structure, e.g. describing the young water fraction method. We will also look at whether it will help to separate the Results section into the Fig. 3 part (showing nonlinearity of apparent MTT with tritium concentrations) and Fig. 4 & 5 parts (showing aggregation errors).

Ref #2: 2) Methods: After introducing tritium (H3)-based TTD estimation, LPMs and their properties, “Results” starts and I am left with an unsure feeling of how the paper addresses the issues raised in the Introduction. I know you use the four GMs from Fig 2a, but in which combinations for the two virtual catchments? Only selected combinations, or all possible ones? How were the catchments mixed? (I know it is 50:50 because it says so later on, in Results: : :which links back to my comment about structure of the paper). Did you use the GM of each sub-catchment in Equation 1 and forward-propagated Northern and Southern hemisphere H3-data, then mixed it 50:50? All this information is missing, and young water fraction calculation or the literature re-evaluation is not even mentioned here. Furthermore, I think that the description of the individual LPM can be shortened without losing important information. Also, I think Table 1 and showing that the GM can mimic the shapes of other LPM is not essential for understanding of the paper. It is interesting to quickly summarize which LPM is useful for which application, however.

Reply: This is helpful comment for revising the Methods. We will update the Methods section with the requested information, and will add an extra subsection describing how the virtual calculations were carried out. (Your summary of how the tritium concentration of the mixture were calculated “Did you use the GM of each sub-catchment in Equation 1 and forward-propagate Northern and Southern

hemisphere H3-data, then mix it 50:50?" is essentially correct, although we might change a few words).

We agree that Table 1 is not essential for the paper, but it helps the reader to use our analyses without investigating all possible combinations of LPMs. The point is that by using the full range of the GM we have covered most of the other simple and frequently used LPMs (i.e. the EPM and DM) as well, since the GM can mimic their useful ranges of shapes. If the estimated errors between GM and other distributions are small in Table 1, GM can be a good approximation. From Table 1, EPM with exponential components (f) less than 0.44 do not give good matches with GM. This occurs when the piston flow component becomes dominant in the EPM, see the horizontal part of the EPM curve in Figure 2b. In such cases, the young water threshold analysis of GM is not applicable and is likely to cause large errors.

Ref #2: 3) Yf calculation: It is unclear to me how you calculated this. Yf is determined and calculated from the threshold age ty (Equ. 12), yet it seems the threshold age was calculated by comparison to apparent and true Yf already existing (page 9, first few lines)? True Yf comes from the individual Yf1 and Yf2 (Equ.13), but where are they and their ty1 and ty2 coming from? Also, ty should give good agreement with 10% of apparent and true Yf. 10% of what? What is the 100%? Maybe an explanatory figure would help here, and this should be also in Methods.

Reply: We will describe this more clearly in the Methods. As explained in the reply to Ref #1, we have now changed our procedure for calculating the apparent MTTs and Yfs. The new procedure causes very little change to the apparent MTTs, but changes the Yfs in that the threshold age (ty) can now be kept constant for the whole range of α (0.3 – 10). This simplifies the procedure, but there is still the choice of ty to be made. Our present choice of 17.5 years was made essentially by trial and error by determining the biggest value of ty that will still produce minimal aggregation error in Yf for all values of α.

Ref #2: 4) Apparent Yf determined from LPM fitted to H3 of mixture: To my understanding, Kirchner 2016 showed that Yf of the mixture can retrieve the "true" Yf (calculated from our knowledge of the virtual system) using a gamma function, but only Yf is valid and the corresponding gamma function itself is not valid (otherwise we would have a valid gamma function and thus a correct MTT again, i.e., no aggregation bias). Equipped with this knowledge, how can we reliably trust the apparent Yf result if it comes from a LPM function that is fitted to the H3 mixture and will most likely not be e.g. gamma distributed anymore, but hyper-gamma distributed?

Reply: The Yf can be correct when the TTD is not, because the Yf only includes the part of the TTD with transit times between zero and the young threshold. The TTD becomes more and more flawed towards the long transit time end of its range (e.g. it contains no transit times greater than 4-5 years using seasonal tracer cycles). The MTT (which is an average over the whole range of transit times) is therefore also flawed.

Ref #2: 5) Chapter 3.3 seems unnecessary to me, and is very short in itself already. If you want to keep it, please elaborate on its importance.

Reply: We will elaborate this section a little, as commented in the reply to Ref #1.

Ref #2: 6) Chapter 3.4: I am unclear as to which results were already obtained by the cited studies and which results were calculated by the present manuscript.

Reply: So far the calculations are all from the cited studies, but we will make further calculations to explore the MTT and TTD interpretations further.

Ref #2: 7) I generally doubt the validity of Chapter 3.4, the literature review. To me there is suddenly a huge leap in logic/faith: that using compound LPM will give the true MTT. Or one that is "truer" than the versions of single LPM. It is assumed that just a good fit of tritium tracer data warrants to say that the model gives true results. I do not say they are wrong, I do not say they are true. I do say we cannot know, or I do not see any evidence here that would substantiate your assumption that the compound LPM would give the true MTT. Even if both parts that feed the mixed water in all described studies would be homogeneous in themselves: the virtual experiment catchments of Kirchner 2016 were also homogeneous in themselves, but different from each other, and still led to aggregation bias. We would need proof that the individual catchments are "homogeneous enough" (whatever that means) and that compound LPM, which are just simplifications of processes that we think occur in a catchment domain, correctly mix the two flows in a way that surely avoids aggregation bias. Just a good fit of observed tritium data is surely not a bad start, but not enough in my opinion. I am in favor or a) deleting Chapter 3.4 OR b) rewriting it much more cautiously, with discussing the considerations uttered here.

Reply: This is an important point, which we will address by rewriting this section more carefully and cautiously. We are not intending to say that using a compound LPM removes all possibility of aggregation error, only that using a "well-chosen" compound LPM will help to reduce aggregation error. There is then the problem of what is a "well-chosen" LPM? Our paper considered other information such as chemical and geological information as well as the observed tritium fits to answer this. Further aspects will be considered in the reply to Ref #3.

Specific Comments (page-line)

Ref #2: 2-24: "young water" appears here the first time. Maybe define it a bit more clearly. How young does it need to be to be considered young water?

Reply: This line refers to "younger water components", and it is a general rather than a specific reference at this point in the paper. It means water components younger than the average of the mixture of water making up the catchment or groundwater system outflow. We will add this explanation to the text.

Ref #2: 3-11: "[: : :] the one tracer". This makes it seem to me that two different tracers are used, but I rather get from this paragraph, that actually "when we only have tracer data of the mixture" is meant. Please clarify.

Reply: Agreed, we will change this.

Ref #2: 3-16: Choice of LPM based on hydrogeological situation: please give an example or reference at this point.

Reply: References are Maloszewski and Zuber, 1993 and Maloszewski et al. (2002).

Ref #2: 3-18: "water-bearing layers" to avoid confusion while reading (had to read three times)

Reply: Ok, we will make this change.

Ref #2: 4-6 "times. i.e. The water [: : :]" please correct the capital T and also the period after times seems strange.

Reply: We will change this.

Ref #2: 4-15: with “calender time” you refer to daily time steps? Monthly? Yearly?

Reply: We refer to yearly time steps, and will clarify in the text.

Ref #2: 5-23: Starting the sentence with a side-sentence in brackets “(Maloszweski [: : :]” looks weird, in my opinion.

Reply: We don’t think there is anything wrong with this.

Ref #2: 7-6: I would write 2.25 instead of 2.5, if you already use two digits after the comma for 0.05. Except for that I recommend not showing this information anymore, see General comment #2.

Reply: We used the expression “about 2.5” which 2.25 is.

Ref #2: 7-12 to 7-25: Methods

Reply: We think this would cause reader confusion.

Ref #2: 7-20: Please explain Fig. 3 a bit more in the manuscript to assist in fully understanding it. As far as I understand it, the black curves show TU that one would measure in streamflow. It seems to be the fitting result of the LPM to the mixed H3 signal (p7-L29f). How did you find the two catchments TU concentrations (necessary to find the mixed TU signal) based on the desired MTTs of 3 and 197 years? There must be some other MTT-TU function behind it, that is not shown? I’m basing my last assumption on Kirchner 2016, where the combination of e.g. two exponential distributions did not lead to another exponential distribution, but a hyper-exponential distribution. Thinking along these lines, this confuses me even more now: every red dot in Fig 3, that is, every TUMTT combination of the two individual catchments, lies exactly on the black curves that come from fitting the mixed runoff TU signal. But according to the logic here, the black curves should be wrong. How can the red dots lie on the black curve, if the shape of the distribution of the mixed runoff is not known and should be some hyper-something version?

Reply: We will explain this part better. We have applied a simple procedure which is explained in the paper. In Fig. 3, the red curve is the mixing relationship given by Equations 9 and 11, which have the fraction of the young component (b) as their parameter. The black curve shows the apparent MTT resulting from application of a simple LPM (a piston flow model in Fig. 3a or GM models in Figs. 3b-e) to the mixed tritium concentrations.

Ref #2: 7-24: With the assumption of a constant H3 input, are you not basically assuming that no groundwater much older than 50 years significantly contributes to runoff (that is, no groundwater which could possibly include the bomb peak). How realistic is this?

Reply: Our statement here is really beside the point, it doesn’t matter whether assuming constant tritium input is realistic (in the Southern Hemisphere now or in the Northern Hemisphere in the future) or not. The text from 7-20 to 8-11 describes a thought experiment which allows us to demonstrate the non-linearity of the apparent MTT with tritium concentration, and therefore the reason for the aggregation errors.

Ref #2: 7-28: I would get rid of the reference to Equ. 9 here

Reply: We don’t see the reasoning for this, but we will consider in the revised manuscript.

Ref #2: 7-30: Equ. 10 is the standard deviation and seems to be not fitting the text here. Do you mean Equ. 1?

Reply: Agreed. We actually meant Eqns 2 and 5.

Ref #2: 7-31: All the water in streamflow has the same age, not in soil/aquifer. I would specify that here.

Reply: Agreed, we will change this.

Ref #2: 8-12: Regarding Fig 4: Earlier it was mentioned that real input TU data would cause scrambled results in Fig 3. But you use real data now. Please clarify why we can suddenly use them, or if the scrambling would just have made analyzing the figure more difficult, but not prevent the data from being used. Also, the paragraph explains Methods (8-12 to 8-18). Additional information is needed: what were the MTT2 increments? Was the GM model used, as you talk about alpha parameter later? How was MTT2 changed then, by changing beta parameter in certain increments?

Reply: The reason that the real input data were not used in the calculations shown in Fig. 3, but were used in those in Fig. 4 was because different quantities were plotted in the respective figures. Fig. 3 was used to demonstrate the non-linearity of the apparent MTTs with tritium concentrations; this would have been obscured if the real input data had been used. The real input data could easily have been used, but then we could not have demonstrated the non-linearity because the figures would have been a mess. Fig. 4 demonstrated the difference between the true and apparent MTTs and the real input data could be used (and was necessary) because tritium concentrations were not plotted as one of the axes.

The MTT2 increments were 50 years which causes 25 year increments in True MTT. The GM model was used to calculate the tritium concentrations of the two components. MTT2 was changed by changing β since $\tau_m=\alpha\beta$ (5-15). α was kept constant for the two components as shown in the headings in Figs. 4 and 5.

Ref #2: 8-20: How were the uncertainties for fitting young waters of MTT1 calculated? Might also be good to call it MTT1 here one time since it is only used in Fig. 4 and leads to some confusion initially when looking at Fig. 4.

Reply: The “fitting uncertainties” of the apparent MTTs are the standard deviations of the GM fits to the mixed tritium concentrations. They are only shown for the two youngest MTT1s (3 years and 25 years) because the uncertainties are the biggest for these. However, we will show them for all of the MTT1 cirves.

Ref #2: 8-22: The fitting errors are important because of more complex LPM in a good or bad way?

Reply: A good way, because big fitting errors should tell the researcher that their simple LPM is misrepresenting the data.

Ref #2: 9-4: I guess it is 197 instead of 397.

Reply: Yes, it should be 197 years.

Ref #2: 9-14: Please define a “reasonable” choice of young water threshold.

Reply: We don’t think we should define a reasonable choice of ty here, or at least this part of the text needs to be revised. I expect that a reasonable choice could be

inferred from Fig. 3a as being after the portion of the black curve which decreases nearly linearly from 2 TU. The black curve bends much more after 15-20 years.

Ref #2: 9-16f: I disagree that cutting-off of the long tail after ty and thus leaving only the short tail will ensure that the apparent Yf does not deviate from the true Yf. As the TTD sums to “1”, the long tail influences the short parts of the TTD and vice versa. If one is changed, the other changes too. If we know that the long tail is wrong, we can’t be certain that the short TTD part is correct if we basically just ignore the existence of the long tail by cutting it off. To use a metaphor, this is to me like healing a bleeding wound by just not looking at it. And if REALLY the part of the TTD model before ty is correct, how can it be that the part after it is NOT correct? The equation and the parameters to calculate the complete TTD do not suddenly change...we just cut off a certain section of it. If the part before the threshold is true, the part after it is true. If the part after it is wrong, the part before it is wrong. Maybe the question is: how wrong? Significantly? Probably not, looking at results from Kirchner 2016.

Reply: The fallacy behind this is that you are assuming that you know the correct TTD. If you have the correct TTD, then your argument is theoretically correct. Since you don’t, you can’t really say much about the long tail from the short tail and vice versa. Then the best you can do is make measurements, and Kirchner’s (2016) and our measurements show that the simple LPM short tails must be near enough to the correct short tails to minimise aggregation error. What you can do with the short tails is subtract them from 1 in order to calculate the approximate amount of water older than the threshold age (this is still not saying anything about the long tails).

Ref #2: 9-22: No explanation follows why the reason for this relationship is found in the gamma distribution.

Reply: We will remove this part of the text because ty will now be taken as constant. in the revised manuscript.

Ref #2: 9-26: 6 to 16 years

Reply: Also this text will be removed.

Ref #2: 9-28 to 9-31: Discussion

Reply: We will consider moving this text to the Discussion section.

Ref #2: 10-25: MRT, which I assume to be Mean Residence Time, is introduced for the first time and replaces the MTT without explanation. Please rectify. Also, it should be Fig 10c.

Reply: We thank the Referee for spotting these typos and will change the text.

Ref #2: 12-21f: Following the reasoning about the different bias thresholds: Does this not mean that using tritium methods for streamflow, we would get bias-free estimates for transit times smaller than 6 years, which must include the seasonal cycle results if we would apply it to the stream, and ultimately agree with them?

Reply: We are not sure what is meant here. Using tritium, aggregation bias should be small for MTTs less than about 17.5 years (or present ty). Tritium concentrations show only a small seasonal variation (referred to as the “spring leak”) and it has not so-far been found useful for age-dating. Using seasonal tracer cycles (stable isotopes or chloride), aggregation bias should be small for MTTs less than 2-3 months according to Kirchner (2016).

Ref #2: Title of 4.2: Consider changing to “How much has aggregation affected tritium MTTs in past studies?”

Reply: We thank the Referee for this suggested title and will change the title accordigly.

Ref #2: 12-30: Conclusion #1: I disagree with ONLY affected by bias if “older than 6 years” and “if determined by simple LPM", for reasons already explained above.

Reply: It may be wise to soften this statement a little, but we basically believe it to be a fair statement.

Ref #2: 13-1f: If we take the variance into account in the given examples of 10 plusminus 8 and 10 plusminus 5, there seem to be quite a few catchments that have less than 6 years MTT.

Reply: Yes there are some, but the very large variances are mostly caused by outliers on the old side not the young side. Most of the MTTs are close to the mean or just below it.

Ref #2: 13-8 Conclusion #2: As mentioned above, I see no evidence for this statement.

Reply: The four case studies in Section 3.4 are examples of this statement. Also the Blavoux et al (2013) study described from 13-12 to 13-18. And four more key examples are given from 13-19 to 13-27.

Ref #2: 14-23: I must have missed the part in the manuscript that shows that simple LPM still work in case of long series of tritium measurements? Where is that shown?

Reply: If the simple LPM accurately represents the hydrological system, then the MTT derived from it will not have much aggregation bias. This is equivalent to the cases in the virtual experiments where MTT1 and MTT2 are the same and the true and apparent MTTs of the mixture plot on the 1:1 line in Figs. 4 and 5. An indication that the simple LPM provides a good description of the hydrological system is given if the simple LPM provides a good fit to a long series of tritium measurements. We will add a short sentence about this in the revised manuscript.

Ref #2: Table 1: In the description change “The shape parameter of the best-fitting versions of the other models [: : :]”, since it is not always the shape parameter for the other models, e.g. it is the dispersion parameter for DM.

Reply: The dispersion parameter is the shape parameter for the dispersion model since it controls the shape of its MTT.

Ref #2: Figure 2a: the scale parameter beta was fixed for each GM? Which value did it have?

Reply: β is fixed for each case depending on the value used for τ_m, since $\beta=\tau_m/\alpha$ (5-17). With $\tau_m=1$, β ranges from 3.3 to 0.1 for α ranging from 0.3 to 10.

Ref #2: Figure 4: In the legend: the orange MTT1 actually says “MTT!”

Reply: We fixed the typo and thank the Referee for spotting it.

Ref #2: Figure 7 is not mentioned in the manuscript.

Reply: We added a reference to Fig. 7 in the revised manuscript.

Ref #2: Generally the figures should be unified more in layout, e.g., get rid of the outer border.

Reply: Ok. We will revise these.

References

Blavoux, B., Lachassagne, P., Henriot, A., Ladouche, B., Marc, V., Beley, J.-J., Nicoud, G., and Olive, P.: A fifty-year chronicle of tritium data for characterising the functioning of the Evian and Thonon (France) glacial aquifers, J. Hydrol., 494, 116–133, doi.org/10.1016/j.jhydrol.2013.04.029, 2013

Kirchner, J.W.: Aggregation in environmental systems – Part 1: Seasonal tracer cycles quantify young water fractions, but not mean transit times, in spatiallyheterogeneous catchments, Hydrol. Earth Syst. Sci., 20, 279-297, doi:10.5194/hess-20-279-2016, 2016.