I appreciate Anonymous Referee #2's comments and suggestions. Where possible, these will be used to improve the manuscript on revision. Specific responses to individual comments are detailed below.

This paper explores how mean transit times (MTT) derived from seasonal tracer cycles aggregate when scaling up by adding small catchments to represent a larger catchment. Kirchner finds that the MTT does not scale well at all by performing thorough benchmark tests and he proposes a new metric: the young water fraction that by its definition scales as good as possible with spatial heterogeneity. Thus 2 main messages in this paper:1) From heron, do never use MTT again, 2) use Fyw instead. This new metric is interesting, but at the same time challenging to use as its definition contains uncertainty (i.e. the type of transit time distribution, which is always unknown).

Furthermore I fully agree with the call for thorough benchmarking of simple hydrological models in the face of spatial and temporal heterogeneity. The paper is well written and is an important contribution to hydrology.

Many thanks for your kind remarks about the paper. Regarding the point that the transit time distribution is always unknown: yes, but this is also true (and much more consequential) for mean transit time determinations, where it leads to order-of-magnitude uncertainties in MTT.

However, several questions remained after reading the paper:

How can I use the Fyw (fraction young water) with my data? Is the approach something like: First guess a range of alphas that are likely to represent my system, Let’s say 0.3- 1.5. Then derive the Thresholds Times with Eq 14,–> 0.12-0.22 years. Next derive from data the As/Ap. For example 0.3. This then means that around 30% of my stream water is younger than 0.12 to 0.22 years? Next we can refine this approach by including the phase shift? Between catchments we can now compare this Fyw. I think this could be explained more clear in this paper, for example with the data of figure 1.

The suggestion of a "worked example" is a good one. I will see whether I can fit it in, without making the manuscript too much longer.

What is the advantage of comparing Fyw over As/Ap between catchments?

The advantage is that Fyw tells you something about transit times, and As/Ap doesn't (at least not directly). In the second paper, for example, I show how these methods can be used to (for example) quantify how the young water fraction (and thus the fraction of water flowing by relatively fast flowpaths) varies between high flow and low flow. Furthermore the relative fractions of young and old water (and their variations with discharge regime) can be compared to stream chemistry, to define the chemical fingerprints of "young" and "old" end-members. Following Referee #1's suggestion, I will try to give the reader a taste of these potential applications already in the first paper, although they will not be spelled out in detail until the second paper.

What about evapotranspiration? Are the proposed methods valid when half of the water balance goes to evapotranspiration? You convincingly proofed that the MTT does not scale up well, but does FYW still scale well with evapotranspiration? Page 3069, line 9, seems to suggest it does not, but with amplitudes I can imagine it does work. Does this need further benchmarking?

Do you mean line 20 on page 3066 instead? In practice, convolution-based approaches (including those used to estimate MTT) ignore evapotranspiration (ET). Estimating how ET would affect Fyw determinations is not straightforward, because this will depend on how ET alters the concentrations of the conservative tracer. Thus this effect will differ, for example, between stable isotopes and chemical tracers. I am currently working on a manuscript that looks at this question for stable isotopes, but this is a rather complex topic that is well beyond the scope of the current paper.

Following your own reasoning on page 3070, line 20, a catchment consists of almost infinite number of flow routes, each with own travel times. All these flow routes are grouped to yield the catchment TTD. You showed that Fyw scales well for 8 subcatchments, but does it still scale well for 1.000.000 sub-flow routes?

By extension it should, but numerically demonstrating this would be computationally tedious, and well beyond the scope of the current paper.

Is there any chance that due to the central limit theorem an infinite number of weakly-related gamma distributions for each flow route (log-transforming them, adding them yielding a normal distribution, and back-transforming them to yield a log-normal distribution) yields a log-normal TTD distribution at the catchment outlet of which the MTT does scale well as long as we assume that the central limit theorem holds at all the subcatchments as well at the catchment? I dont think so, but Im also not entirely sure that the Fyw does much better.

I don't understand the reasoning here. The point is not whether MTT scales well (by definition, the mean will always aggregate linearly), but whether *a procedure for estimating* MTT will work correctly, when the only inputs are observable behaviors (like tracer concentrations) that come from heterogeneous aggregates of subcatchments.

Minor comments: Title: As the authors refers in both papers to “paper 1” and “paper 2”, it would be good to include this number somewhere in the title of the paper as the papers are likely to end up in reverse order on a website (like now on HESSD).

The original manuscripts had such numbers in the title, but these had to be removed because of problems that would be created for any future papers in this series, which may appear separately. In the revision, I will try to clarify how each paper refers to the other one.

Page 3066, line 16: one can relax... flow-equivalent time. I dont think it is possible to express time as flow-equivalent time when sine wave fitting. Thus this statement is confusing to me in the context of this paper.

Obviously, a mathematical sine wave will no longer be a pure sine wave if the time base is locally stretched and shrunk in a non-uniform way. But in practice sine waves are fit to rather noisy tracer data, so it's not clear how much this will affect the fitted sine wave. In any case, the statement is a general one about convolution methods, and is not specific to sine-wave fitting (which is introduced three pages later).

Page 3066, line 13: “However in practical applications”: this statement renders all the above references impractical, while the objective of using time variant TTD actually is to be a bit more practical. To me considering a catchment as a stationary flux field is totally theoretical and only suited for catchment intercomparison studies. These stationary studies hardly have any practical relevance in helping to understand how to lower or mitigate solute fluxes.

What I meant was, "in applications using real-world data". I did not intend to label time-variant TTD approaches as impractical, since that is not an issue one needs to get into here. Your comment does, however, point to an important issue. There is a rich theoretical literature on time-varying TTD's, but it is only now starting to come to grips with the important problem of how we can determine what these TTD's actually are, in the real world, based on real-world data. This is not at all a simple problem. Since the first requirement of any approach to practical problems is that we must be able to use it reliably with real-world data, this represents a substantial challenge for time-variant TTD methods. Although the present paper deals with stationary (but heterogeneous) systems, the second paper shows that these methods can also be applied to data from nonstationary (and heterogeneous) systems.

Page 3078 line 25. Following your reasoning on page 3071, line20, each subcatchment consists of an almost infinite number of independent flow paths that contribute to stream discharge. Do you think you still get the results of figure 12 for an infinite number of subcatchments? Is this what you are saying on page 3079, line 3?

Yes, that is what I am saying.

Page 3080, line 25 MTT values derived from seasonal tracer cycles

Correct. But note that (as explained on pp. 3081-3082), there is little reason for optimism that other methods of estimating MTT from tracer data will be any more reliable. There are several reasons for this. First of all, MTT depends strongly on how long the long-time tail is, but conservative tracers are insensitive to such long-term behavior (either because the tracers themselves don't exhibit much decadal-scale variation, or because we don't have measurements that run that long). Secondly, to the extent that the seasonal cycle is the dominant feature of many tracer time series, that cycle will largely control the results obtained from those time series, no matter what methods are used to fit or interpret the data. Sine-wave fitting just happens to be the simplest and most analytically tractable of those methods, which is why I have studied it here.

Page 3082, line 10. Im not entirely sure what you mean with the time series convolution approach. If it refers to methods that solely use the waterbalance (water storage and water fluxes time series) to calculate the MTT, this aggregation bias is likely to be absent. At a larger or smaller scales this approach leads to a new water balance with new storage and water fluxes, which lead to a new MTT independent of the aggregation (close to [average Storage] / [average precip].

I mean time-domain convolution of tracer time series. Note that water balance methods lead to highly biased estimates of MTT, since at best they only measure dynamic storage and not passive storage.

However, I fully agree that MTT is an awful and often meaningless metric to use. Median traveltimes or indeed Fyw are much more meaningful.

Page 8030, line 19. You mean to say that Fyw is more useful than MTT?

I think you mean page 3083, and yes, that's what I meant to say.