I appreciate Anonymous Referee #1's comments and suggestions. Where possible, these will be used to improve the manuscript during revision. Specific responses to individual comments are detailed below.

GENERAL COMMENTS

This manuscript extends the results of the companion paper ("Aggregation in environmental systems: seasonal tracer cycles quantify young water fractions, but not mean transit times, in spatially heterogeneous catchments") to the case of non-stationary hydrologic systems. Like in Paper 1, the author makes use of benchmark testing procedures based on a well-designed virtual experiment. Several results are presented from different system configurations, precipitation forcing, flow regimes, tracer data. Overall, the paper is well written and represents an important contribution to our understanding of catchment transport processes.

Many thanks for your kind remarks about the paper.

The first part of the manuscript (Sections 2-3.3) introduces and investigates the virtual hydrologic system. The author shows an interesting procedure to accurately solve the main transport equations and to reduce the equifinality of model parameters (which is typical of non-linear storage-discharge relationships). Although the author should add more reference to the existing literature (which in some cases already showed similar results with similar models – see Detailed Comments), the results are clear and of good scientific quality.

The second part (Sections 3.4-3.8) explores sine wave fitting methods applied to the virtual experiments. The results show that in a non-steady-state system, mean transit times (MTT) estimated from sine wave fitting methods generally do not match the "real" average MTT. Instead, such methods reliably estimate the average "young water fractions" (Fyw, introduced in Paper 1). This part is engaging and innovative, but, as such, it needs to be better framed. The central issues that, in my opinion, need to be solved are:

1) The reader may struggle with the definition of Fyw, because the definition of the threshold age is necessarily imprecise in real catchments (as shown in Paper 1). Hence, more effort could be put in explaining why the lack of a precise threshold age has minor importance.

There are three ways that this can be handled. The first approach is to note that, for distributions as widely varying as gamma distributions with alpha=0.2 to alpha=2 (see Figure 2 of Paper 1), the threshold age tau_yw (for which the young water fraction Fyw is close to the amplitude ratio As/Ap) only varies in the range of 1.4-3.1 months. For more common (or at least more commonly assumed) travel time distribution shapes (corresponding roughly to alpha=0.5 to alpha=1.5), the threshold age varies only from 1.7 to 2.7 months. Thus if As/Ap is measured at (for example) 0.3 for a particular catchment, this means that about 30% of discharge is younger than 1.7-2.7 months (or 1.4-3.1 months, if one wants to consider an even wider range of distributions). The key point here is that in practice, the difference between 1.4 and 3.1 months will not have a big effect on how this result would be interpreted (particularly in comparison to the mean transit time, which may be years).

The second approach is to quantify how much a wrong guess about the shape of the transit time distribution would affect the value of the young water fraction. The thought experiment goes like this: let's assume that (for example) we have an

exponential distribution (shape factor=1.0), for which the threshold age is 2.3 months. From tracer observations, we calculate that the amplitude ratio of the seasonal cycle is 0.3, and we infer that "30% of streamflow is younger than 2.3 months". Now, what if our assumption is wrong, and our transit time distribution actually has a shape factor of 0.5 instead of 1.0 (but damps the seasonal tracer cycle by the same amount that we have observed)? The key question is: what fraction of this alternative distribution is younger than 2.3 months? That is, how wrong will our inference that "30% of streamflow is younger than 2.3 months" actually be? The answer can be calculated from equation (10) of Paper 1 and the incomplete gamma distribution. For this alternative distribution, 36 percent of streamflow is younger than 2.3 months, rather than 30 percent. Our original estimate was wrong by about 6 percent (of the range of *a priori* uncertainty, which runs from 0 to 100 percent) or equivalently about 20 percent of our original estimate. If the true shape factor were 0.2 instead of 1.0, the error would only be about 4 percent.

Many readers and reviewers have been curious about this point, so I will include a systematic sensitivity analysis along these lines in the revised version of Paper 1. This will appear in Paper 1 so that it is handled when it first comes up, and because Paper 2 is already long and complex.

The imprecision in the threshold age also leads to a legitimate (though provoking) question: why bothering about young water fractions and not just studying tracer cycle amplitudes and shifts?

First, we need to keep the "imprecision in the threshold age" in perspective. Mean transit time determinations also require assuming a shape for the transit time distribution, and the results are highly sensitive to that assumption. For the same conditions as the thought experiment outlined above (As/Ap=0.3), for example, the mean transit time *varies by a factor of 20* as the shape factor ranges between 0.2 and 1.0.

But to answer the question that was posed: one could of course compare tracer cycle amplitudes and phase shifts from one catchment to the next, but what would one learn by doing so? Without a theory to link these observable quantities to phenomena within the catchments themselves, why bother making such comparisons? The obvious advantage of Fyw over just amplitudes and phase shifts is that Fyw tells us something about transit times, which amplitude ratios and phase shifts don't, at least not directly.

2) The author sets the threshold age for the virtual experiment equal to that of a stationary exponential TTD, even if the system is non-stationary and its marginal TTD does not resemble an exponential pdf. Such a choice is not discussed and may look arbitrary. To what extent are the results affected by this choice?

This question can be answered through the sensitivity analysis presented in response to point (1) above.

3) The procedure to estimate Fyw (Section 4.3) shows two possible strategies. The first one is clear (as it was explained in Paper 1), but the second one, which includes phase shift information, is not described in paper 1 and is not critically discussed in this manuscript. Such a discussion is necessary (here, or in Paper 1) to understand how (and maybe why) this strategy works.

As indicated in the responses to the reviews of Paper 1, when I revise that paper I will be giving step-by-step instructions explaining how to include phase shifts in the estimates of Fyw.

A last general note is that the paper presents quite some results (18 figures, more than 15000 words), so it makes it difficult for the reader to get till the end. Any effort to reduce the manuscript length is welcome.

I will see what I can do. The paper is long because there are many interesting results to show.

DETAILED COMMENTS

3108 l. 29: "from tracer concentration" it should be specified that it regards sine wave fitting methods.

The sentence is correct as stated, because the model can potentially be used to test many other methods besides sine-wave fitting (although that is the only application that I have space for in the present paper).

3109 l. 12: avoid referring to "effective precipitation", even if many tracer studies do, because it implies that evapotranspiration only affects particles with age 0, which is unrealistic.

Good point!

3110 l. 4: this threshold age is not justified, nor it is checked a posteriori for the calibrated model. Indeed, depending on the parameter combination, the marginal distributions of the individual boxes and of the streamflow may resemble gamma distributions with shape parameter alpha quite different from 1.

Yes, but that is exactly the point. In the real world we *would not know* what the shape of the transit time distribution was. In the real world, all we have is the tracer behavior.

Therefore if, as part of these tests, we looked at the transit time distributions produced by the model and then chose the correct alpha value, we would be cheating, because in the real world we would not have this information. And in any case, my results show that the shape of the distributions generated by the model (and thus the "correct" alpha value) would be continually shifting.

The rationale for choosing alpha=1 is that this corresponds to the most commonly assumed distribution in many catchment studies (that is, exponential). The analysis presented here shows that we can get reasonable results, even if we assume the wrong alpha value, as also shown by the sensitivity analysis described in response to point (1) above.

3110 l. 9: this is also called "random sampling" scheme.

I deliberately avoided referring to this as random sampling for two reasons. First, it's a deterministic model, and I wouldn't want any readers to think that I was actually sampling particles from the boxes at random. Second, although random sampling is also an equal-probability sampling scheme, not all equal-probability schemes are random. Thus, calling this "random sampling" would be overly (and misleadingly) specific.

3110 l. 13: as the analytical solution exists for well-mixed volumes (e.g. Rinaldo et al., (2011)), why is the author tracking age numerically?

I assume the reviewer is referring to Rinaldo et al.'s equation A4 and A5. This so-called "analytical solution" is not in closed form. Instead it involves two integrals (one of which must somehow be solved over infinite time) for each slice of the age distribution and for each time step. Thus Rinaldo's approach appears to be vastly more computationally complex than the approach that I have taken.

3116 l. 18-27: this paragraph could be moved earlier in the text (Section 2), to make the use of the model clear from the beginning.

This is already made clear in the paragraph that comes right before the Section 2 heading. I will think about whether it makes sense to move the Section 2 heading one paragraph earlier in the text.

3119 l. 12-29: this paragraph should include reference to other papers that showed these findings in theoretical and applied contexts (e.g. van der Velde et al., (2012), Botter (2012), Hrachowitz et al., (2013), Harman (2015)).

I will look again at these papers and reference them where appropriate. Because the language and mathematical formalism in these papers is so different from the present manuscript, the same ideas may look quite different (or, conversely, different ideas might look quite similar).

3120 l. 8: this was also shown by Harman (2015).

I have read Harman (2015) – twice – and I cannot find an equivalent statement anywhere. Perhaps it is somehow implicit in the formalism and results (indeed it should be), but is it explicitly stated?

3125 l. 23-25: this is very well explained!

Thank you (although I think it's far from the best prose in the paper).

3129 l. 24: the definition provided by the author in Paper 1 is actually more complicated, because as one does not know the shape of the TTD, the threshold age cannot be specified.

The statement is correct as stated. By definition, for some threshold age tau_yw, Fyw is the fraction of water younger than that age.

It is an exaggeration to say that "as one does not know the shape of the TTD, the threshold age cannot be specified." Yes, the threshold age cannot be specified to

absolute arbitrary precision, and not without making any assumptions. But for a reasonable range of TTD shapes, one can obtain a reasonably well constrained range of threshold ages. If we are looking for assumption-free analyses, we will need to throw out basically all of modern hydrology.

3130 l. 1-11: this is quite unclear. The second of the two strategies is not described in paper 1, nor it is critically discussed prior to its use.

As indicated in the responses to the reviews of Paper 1, when I revise that paper I will be giving step-by-step instructions explaining how to include phase shifts in the estimates of Fyw.

3130 l. 14: here the author could be more explicit and specify that the method was proved reliable for compositions of gamma distributions with shape parameter ranging from 0.5 to 2.

Section 4.2 of Paper 1 (see also Figure 12 of Paper 1) shows that these methods work for combinations of gamma distributions with shape factors ranging from 0.2 (not 0.5) to 2, and mean transit times ranging from 0.1 to 20 years. The point is that once you combine two of these distributions you get a distribution that is not gamma-distributed at all. Thus, what I have shown is that this analysis also works for combinations of distributions that are not gamma-distributed. That's what this statement is meant to say.

3130 l. 16: I am not totally sure this virtual experiment can be considered representative of a “homogeneous” catchment. Or I don’t understand the author’s definition of homogeneity. Indeed, the system is made of two different sub-systems, characterized by markedly different time-scales.

OK, I get the point, and this needs to be clarified when the manuscript is revised. What I mean by "homogeneous" is that we have a single two-box model, and its parameters don't change throughout the catchment. One could say that this is vertically stratified but horizontally homogeneous. By contrast, what I call "heterogeneous" is the case where we have a different two-box model for each subcatchment. This is, in other words, both vertically stratified and horizontally heterogeneous. The terminology is intended to be analogous to the situation in Paper 1, where a "homogeneous" catchment was characterized by a single TTD, and a "heterogeneous" catchment was characterized by different TTD's in each subcatchment.

Obviously, one could term the two-box model as "heterogeneous" because it has two boxes. But I think it is useful to distinguish between such a two-box model, which exhibits one set of nonstationarity characteristics (depending on its parameter values), and an assemblage of such models (representing different subcatchments). This assemblage of models is "nonstationary and heterogeneous", and can potentially exhibit more complex nonstationarity, because not only do the individual subcatchments have different behaviors, they can also shift in dominance over time (because, for example, fast-responding catchments will dominate in early time, and slow responders will dominate in late time).

3130 l. 12-29 and Figure 10: when comparing the young water fractions (and MTT) derived from age tracking to those estimated from seasonal tracer cycle, please specify that the

formers are average values (time-average, flow-average, over the whole dataset, over a specific flow regime, etc), otherwise it is confusing.

OK, can do.

3131 l. 11: the author mentions “flow-weighted fits to seasonal tracer cycles”. How is this done?

This is done by weighted least squares, with weights proportional to flow or precipitation volume. If there are potential outliers one can use IRLS, Iteratively Reweighted Least Squares, with flow or precipitation weights, in combination with the usual iteratively updated point weights, which are downweighted for points with unusually large residuals.

3132 l. 13: as commented above, this virtual experiment does not look homogeneous to me. So how can the author separate the effect of non-stationarity from that of heterogeneity?

Please see the discussion above (for 3130 l. 16).

3134 l. 12: please specify that the young water fraction is an average value (in this case, over a specific flow regime), because the real Fyw changes in time. Is this a time- of flow-average? Figure 12: same comment as above

I don't know what is meant by "flow-average". These are Fyw values that are estimated from the behavior observed in individual "slices" of the discharge distribution. Thus they are time-averaged and flow-specific.

I would have thought that it would be obvious that these Fyw values are time-averaged, since no time point is specified. Nonetheless, it can of course be stated explicitly.

3134 l. 20 and figure 13: please specify that the “real” young water fraction is an average value

The caption to Figure 13 already says, "Upper panels compare the TIME-AVERAGED Fyw in each discharge range..." (emphasis added).

3134 l. 21: this is an interesting results, it would be worth adding further comments.

OK, I will think about whether there is more that is worth saying here (although it's hard to know how much one can generalize from the one case of the Smith River data).

3137 l. 27: the young water fraction has a rather specific meaning, so it does not just estimate the fraction of “relatively” fast flowpaths. It estimates the fraction of flowpaths that supply the stream with water younger than about 2-3 months.

Yes, and by any measure these are relatively fast compared to the mean transit time. But yes, the 2-3 month time frame could be specified.

3141 l. 21 to 3142 l. 27: these paragraphs are rather long and could be condensed

But I think the point that is being made here is worth emphasizing (and explaining at sufficient length so that hopefully nobody misses the point).

3142 l. 15: this “inductive leap” is important. What does one learn from the virtual experiment for applying the method to real catchments?

This inductive leap is a very small step, compared to the (usually unspecified and possibly unrecognized) inductive leaps required to apply typical predictive models.

3142 l. 21-27: yes, but then one wants to apply the method to real-world catchments. So the model structure plays a role in suggesting whether the method is applicable to a real catchment at hand.

It only plays a role if one thinks that the model creates a particular kind of complexity in the simulated time series, for which the inferential method somehow magically works well, clear across the rather wide ranges of structures and parameter values tested here... whereas the real-world catchment produces some other kind of complexity in the simulated time series, for which the inferential method spectacularly fails.

I am not arguing that there are no conceivable circumstances in which the inferential method would fail. I *am* arguing, however, that the results of my analysis do not strongly depend on the realism of the simulation model, which is only used to generate the benchmark time series (rather than generate the inferences that are being tested). The model is just being used as a fancy random number generator, which must only produce benchmark time series that have realistic degrees of complexity. I am specifically contrasting this situation to typical catchment modeling studies, where models are intended to draw conclusions about real-world catchments, and therefore the realism of the model is of first-order importance.

Let's keep this in perspective. I've just spent a lot of time and energy to test an inferential method across a rather wide range of conditions, and to demonstrate its potential utility. Benchmark testing at this level of rigor is rare in hydrology.

Others may have different benchmark tests that they would like to try, and that would be great. But let's remember that in our field, conclusions are drawn every day from models that are highly sensitive to unverified assumptions, and that have undergone almost no rigorous testing at all. Poke holes in my approach all you want... but also poke holes elsewhere, where they are urgently needed.

TECHNICAL CORRECTIONS

3132 l.15 “likely to be underestimated”

Figure 18 caption: there is some mismatch between the brackets at lines 8-9

CITED LITERATURE

Botter, G. (2012). Catchment mixing processes and travel time distributions. Water Resources Research, 48(5), http://doi.org/10.1029/2011WR011160.

Harman, C. J. (2015). Time-variable transit time distributions and transport: Theory and application to storage-dependent transport of chloride in a watershed. Water Resources Research, 51(1), 1–30. http://doi.org/10.1002/2014WR015707.

Hrachowitz, M., Savenije, H., Bogaard, T. a., Tetzlaff, D., & Soulsby, C. (2013). What can flux tracking teach us about water age distribution patterns and their temporal dynamics? Hydrology and Earth System Sciences, 17(2), 533–564. http://doi.org/10.5194/hess-17-533-2013.

Rinaldo, A., Beven, K. J., Bertuzzo, E., Nicotina, L., Davies, J., Fiori, A., Botter, G. (2011). Catchment travel time distributions and water flow in soils. Water Resources Research, 47(7), http://doi.org/10.1029/2011WR010478