Author's response (HESS-2015-63)

My responses to the reviews were submitted as part of the open review process; they are reproduced below for reference.

Summary of changes in the revised manuscript:

As requested, I...:

1) moved the former sections 2.1 and 3.3 to Appendix A and Appendix B, respectively,

2) shortened section 3.6 (which is now 3.5 in the new numbering) by removing the former figure 15 and its associated discussion,

3) substantially shortened the summary and conclusions, eliminating over 300 words.

I also did my best to add the clarifications requested by the reviewers, without making the paper still longer and more complex.

I could not move the former section 2.2 to an appendix suggested by the editor, because this section defines the reference storage values $S_{l,ref}$ and $S_{u,ref}$, which are referred to extensively in sections 2.3 and 3.2 (in the new numbering). Readers would have no idea what these quantities mean unless they are defined and explained in the text.

Author's response to Anonymous Referee #1

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

Author's response to Referee #3 (Markus Weiler)

I really enjoyed reading this paper, however, I have to admit that it took me a while to find enough time to read through over 100 pages of the two papers combined. The paper nicely and very elegantly addresses the question how we should deal with hydrologic nonstationarity to estimate MTT and TTD. The paper is very well written, however, much too long (see comments below) and is certainly of high relevance to the readers of HESS. I have a couple of concerns, of which one could be a major factor changing the main outcome of the paper – if JK can resolve this I think the paper can be published in HESS.

I thank my colleague Markus Weiler (hereafter MW) for his thoughtful comments and suggestions. These will help in formulating the revisions to the manuscript.

There are two reasons that the papers appear somewhat long. First, I am trying to introduce a substantial analysis based on a new concept, so I have to tell the whole story. But secondly, the "over 100 pages" are an artifact of Copernicus Publications' policy of publishing discussion papers in what is effectively a half-page format, thus more than doubling the page count (and, perhaps not coincidentally, more than doubling the page charges that authors pay to Copernicus.)

I disagree that the paper is "much too long", particularly in relation to the substantial ground that it covers. For comparison, Seeger and Weiler (2014) was 51 pages (or actually half-pages) in HESSD; the present manuscript is 63 pages, or only 20% longer.

General comments: 1) I agree with the general approach to use any kind of model to test his approach if the sine wave fitting and calculated change in amplitude and phase shift using the MTT and Fyw will support the one or other approach. However, I miss a very relevant component in the model – evapotranspiration – which will either concentrate certain compounds in the catchments (Cl) or “remove” certain compounds (stable water isotopes). Several studies (e.g. Hrachowitz et al., 2014, Sprenger et al., 2015, HESS) have shown how relevant this part of the hydrological cycle can be including a very strong change in the sine wave of the precipitation signal. I hardly doubt, that a model including ET will come to the same results regarding the strong relation between sine wave dampening and young water fraction.

As stated on p. 3109, the analysis here ignores evapotranspiration in the interests of simplicity. Most analyses based on convolution methods do likewise. Including ET effects in a realistic way is a nontrivial exercise, and the details will be specific to individual tracers; to take just one example, evaporation and transpiration fractionate isotopes differently. Why does MW believe that including ET is important if, as he says, he doesn't think it will substantially affect the results? The point of this paper is to look at how heterogeneity and nonstationarity affect travel time estimates, not to look in general at all the things that could or should be included in a complete process model for tracers in catchments.

I am currently working on a separate paper on evapotranspiration effects on seasonal cycles of stable isotopes, and it will probably be nearly as long as the present manuscript. Including this material here would thus nearly double the length of the paper, which MW already says is too long.

2) The paper is very detailed with a lot of information and great thoughts of JK and many additional information that alone make the papers certainly worthwhile to read. Unfortunately, I fear, that a lot of these comments and thoughts will not be read, since they are hidden in this very long paper. I will make a couple of suggestion to shorten it, but JK may consider to move some parts of the paper to a new paper or an appendix with a clear message, where the reader is able to find his thoughts and ideas much better.

These are options that I already considered when I wrote the original manuscript, and for various reasons they create problems of their own. In any case, the current paper is not really so unusually long; it is only 20% longer than the HESSD version of Seeger and Weiler (2014).

Specific comments:

The introduction builds very much on paper 1 – however, it misses a more thorough literature review to this topic of non-stationarity and the need to develop other approaches and find solutions to infer catchments TTDs.

It is hard to know how to respond without a more specific idea of what MW thinks is missing here. Given that there is a lot of material to cover, I was trying to get down to the task without a long didactic introduction. I will consider whether there is an efficient way to provide more context here.

Section2.1 – Again, very informative and a lot of relevant and interesting ideas – however, it drifts a bit away from the main goal of the paper and makes the paper much longer to read and may result that less mathematically informed readers may stop here. Could JK not move much of this chapter in an Appendix – so modellers can have a more detailed look, but other readers can more easily grasp the main message.

I appreciate the point. Moving this material to the appendix will not save any length (since the appendix will also be part of the paper), and there are a number of logical links that might get broken. A smarter strategy is probably to provide a very clear textual signpost at the beginning of section 2.1, telling readers that if they are not interested in the details they can skip to section 2.3.

Section2.3 – I would propose to name the 3 input time series according to the Köppen climate zones they are in or any other short information. The catchment information is only relevant in this chapter, but not for the whole paper, so JK could avoid to always refer to the catchment names and the climate characteristic of the time series.

I recognize that the place names (Smith River, Broad River, Plynlimon) may not be directly relevant to the readers. It is relevant, however, that the three sites have varying degrees of seasonality, which is why I keep referring to the climate characteristics of the sites, wherever this information is important for the interpretation. Perhaps the primary designation should be the climate type rather than the place name (although this would imply that these particular places and time series were representative of those climatic zones, rather than just being examples of them). In any case, unless readers are well versed in the Köppen classification system, referring to the time series by only their Köppen codes (specifically, Csb in place of mentioning Smith River's

Mediterranean climate, or Cfb in place of Plynlimon's maritime temperate climate, or Cfa in place of Broad River's humid temperate climate) could be confusing.

Section 3.2 – there has been quite a bit of work done on this field – JK cites some of it, but it would help the chapter to include more of this work. In my opinion, the chapter is too close to the model and its assumptions and the parameters sets derived – if he could provide a more general conclusion, the chapter would certainly be helpful, but at the moment, it is mostly a very nice and interesting side way – I would move it to a Appendix chapter.

Again, it would help greatly to have more specific information about what MW feels is missing from the cited literature here. Section 3.2 is important for several reasons. First, it sheds important light on what controls long-term memory (vs. fast hydraulic response – the old water paradox) in the context of this simple model. Second, it provides a clear demonstration of how simple scaling arguments can yield useful insight into system behavior, including characteristic time scales. In the revision, I will try to make these implications more generally accessible to the reader.

Section 3.3 – Interesting, in particular for catchment modellers – however, it includes a new idea, which distracts from the main idea of the paper. If JK would have included the potential of tracer data or young water fraction to constrain the parameters, then the chapter may be helpful to support his ideas, but at the moment it is not. Maybe I also miss some things, as it is not completely clear to me how he derived Figure 8.

The results shown here are, MW and I agree, not essential for the analysis that follows. However, as we both also agree, they make an interesting point. In theory one could write a separate paper devoted to just that point, repeating the description of the model and so forth. As someone who tries to avoid such "salami tactics", I would prefer to keep this material here instead.

I believe MW's question about Figure 8 may be asking how I came up with the particular combinations of parameters. This was essentially trial and error, but informed by the scaling analysis of Section 3.2. I will try to make this clearer in the revision.

Section 3.6 – I would propose to shorten this section, in particular in relation to the figures related to the section. They do not show more detailed information than Figure 10-13. I think the main message of this approach can be summarized in a table or in the text and the figures could be moved to an appendix. I also believe that the illustration in Figure 14 is not necessary.

The point here is not to show "more detailed information than Figures 10-13", but rather to show *different* information than Figures 10-13. Figures 15-16 look superficially similar to 10-13, but they represent a fundamentally different situation (in which there is not just one nonstationary two-box model, but a whole collection of them, with different parameters. It is not obvious *a priori* that such a heterogeneous collection of nonstationary subcatchments would behave like the earlier much simpler system (in the sense that Fyw can still be reliably estimated, and MTT cannot). That is what these figures show. I also think Figure 14 is useful for explaining what has been done.

Remember that moving these figures to an appendix will not save any length; it will only mean that readers will have to flip back and forth between the appendix and the main text.

Section 3.7 – Interesting, but sometimes not clear how JK derived the data for Fyw of the different flow percentiles. This should be better explained in a method chapter, so the reader is able to follow the ways he calculates the data from the three catchments, which are not explained in the beginning in detail.

In principle a more tutorial explanation of the methods is possible, but at the cost of more text and possibly also another explanatory figure. In any case, I think any such explanation should be done here, not in a methods chapter that would not be understood without the necessary context (which readers only will understand once they have seen the results), and which probably would not be remembered by the time it is finally needed.

Summary and Conclusion: Since the paper is already very long, I would highly recommend to shorten the S&C. I think it is not necessary to repeat the main ideas and steps and relate them to the figures – which is a very uncommon format anyway. I would expect from JK the highlights and his vision for the future following his ideas.

I disagree with MW's assertion that the paper is "very long". It is, for comparison, only about 20% longer than the HESSD version of Seeger and Weiler, 2014.

I do agree that it is unconventional to refer to individual figures in the conclusions, but this is a deliberate strategy. Often when they encounter a particular statement in the conclusions at the end of a complex paper, readers often wonder, "Wait, did the authors really show that? *Where* did they show that?" Providing this information gives readers a thumbnail index showing where the main points of the paper are covered. This can save them from searching through pages of dense text. It is also a great help to many readers, who follow the "first-last-middle" strategy of reading the abstract first and the conclusions second, then scanning the figures, and then perhaps reading the text.

The figure captions are very long and often too detailed – I agree that a figure should be understood only with the figure caption, but JK includes already interpretation of the figure. In addition, shortening the names of the precipitation time series would help as well.

The figure captions are written this way as part of a deliberate communication strategy. Minimalist figure captions often lead to unnecessary workload and confusion for the reader, who must jump back and forth between the figure and the text (perhaps several pages away) in order to understand what the figure says. Furthermore, readers often scan papers by looking at the figures without reading the text, meaning that the figures should be able to stand on their own.

Putting interpretations in figure captions can be a great help to readers, who can thereby get a sense of what the figures *mean* rather than just what they *are*. Experience has shown that authors often think that their figures will be self-evident (which of course they are *for the authors*, who already know what they are trying to say), and fail to comprehend how divergent a reader's understanding may be. Thus it is a smart

communication strategy to lean in the direction of over-explaining rather than under-explaining.

Aggregation in environmental systems: Catchment mean transit times and young water fractions under hydrologic nonstationarity

J. W. Kirchner[1,2]

[1]{ETH Zürich, Zürich, Switzerland}

[2]{Swiss Federal Research Institute WSL, Birmensdorf, Switzerland}

Correspondence to: J. W. Kirchner (kirchner@ethz.ch)

Abstract

Methods for estimating mean transit times from chemical or isotopic tracers (such as Cl^-, $\delta^{18}O$, or δ^2H) commonly assume that catchments are stationary (i.e., time-invariant) and homogeneous. Real catchments are neither. In a companion paper, I showed that catchment mean transit times estimated from seasonal tracer cycles are highly vulnerable to aggregation error, exhibiting strong bias and large scatter in spatially heterogeneous catchments. I proposed the young water fraction, which is virtually immune to aggregation error under spatial heterogeneity, as a better measure of transit times. Here I extend this analysis by exploring how nonstationarity affects mean transit times and young water fractions estimated from seasonal tracer cycles, using benchmark tests based on a simple two-box model. The model exhibits complex nonstationary behavior, with striking volatility in tracer concentrations, young water fractions, and mean transit times, driven by rapid shifts in the mixing ratios of fluxes from the upper and lower boxes. The transit-time distribution in streamflow becomes increasingly skewed at higher discharges, with marked increases in the young water fraction and decreases in the mean water age, reflecting the increased dominance of the upper box at higher flows. This simple two-box model exhibits strong equifinality, which can be partly resolved by simple parameter transformations. However, transit times

Style Definition: Normal: Left
Style Definition: Heading 1: Left
Style Definition: Heading 2: Left
Style Definition: Heading 3: Left
Style Definition: Heading 4: Left
Style Definition: Heading 5: Left
Style Definition: Heading 6: Left
Style Definition: Heading 7: Left
Style Definition: Heading 8: Left
Style Definition: Heading 9: Left
Style Definition: Footer: Left
Style Definition: Title: Left
Style Definition: Subtitle: Left
Style Definition: Comment Text: Left
Style Definition: Comment Subject: Left
Style Definition: Balloon Text: Left
Style Definition: equation: Left
Style Definition: Header: Left
Deleted: 2: catchment
Deleted: $\delta^{18}H$
Deleted: a different measure of transit times,
Deleted: and showed that it
Deleted: Even this
Deleted: ; hydrograph calibration cannot constrain four of its five parameters. This equifinality problem

are primarily determined by residual storage, which cannot be constrained through hydrograph calibration and must instead be estimated by tracer behavior.

Seasonal tracer cycles in the two-box model are very poor predictors of mean transit times, with typical errors of several hundred percent. However, the same tracer cycles predict **time-averaged** young water fractions (F_{yw}) within a few percent, even in model catchments that are both nonstationary and spatially heterogeneous (although they may be biased by roughly 0.1-0.2 at sites where strong precipitation seasonality is correlated with precipitation tracer concentrations). Flow-weighted fits to the seasonal tracer cycles accurately predict the flow-weighted average F_{yw} in streamflow, while unweighted fits to the seasonal tracer cycles accurately predict the unweighted average F_{yw}. Young water fractions can also be estimated separately for individual flow regimes, again with a precision of a few percent, allowing direct determination of how shifts in hydraulic regime alter the fraction of water reaching the stream by fast flowpaths. One can also estimate the chemical composition of idealized "young water" and "old water" end-members, using relationships between young water fractions and solute concentrations across different flow regimes. These results demonstrate that mean transit times cannot be estimated reliably from seasonal tracer cycles, and that, by contrast, the young water fraction is a robust and useful metric of transit times, even in catchments that exhibit strong nonstationarity and heterogeneity.

Formatted: English (U.K.)

Deleted: young water fraction

Deleted: young water fraction.

Keywords: transit time, travel time, residence time, isotope tracers, residence time, convolution, catchment hydrology, aggregation error, aggregation bias

1 Introduction

In a companion paper (Kirchner, 2015, hereafter referred to as Paper 1), I pointed out that although catchments are pervasively heterogeneous, we often model them, and interpret measurements from them, as if they were homogeneous. This makes our measurements and models vulnerable to so-called "aggregation error", meaning that they yield inconsistent results at different levels of aggregation. I illustrated this general problem with the specific example of mean transit times (MTT's) estimated from seasonal tracer cycles in precipitation and discharge. Using simple numerical experiments with synthetic data, I showed that these MTT estimates will typically exhibit strong bias and large scatter when they are derived from

spatially heterogeneous catchments. Given that spatial heterogeneity is ubiquitous in real-world catchments, these findings pose a fundamental challenge to the use of MTT's to characterize catchment behavior.

In Paper 1 I also showed that seasonal tracer cycles in precipitation and streamflow can be used to estimate the young water fraction F_{yw}, defined as the fraction of discharge that is younger than a threshold age of approximately 2-3 months. I further showed that F_{yw} estimates, unlike MTT estimates, are robust against extreme spatial heterogeneity. Thus Paper 1 demonstrates the feasibility of determining the proportions of "young" and "old" water (F_{yw} and 1-F_{yw}, respectively) in spatially heterogeneous catchments.

But real-world catchments are not only heterogeneous. They are also nonstationary; their travel-time distributions shift with changes in their flow regimes, due to shifts in the relative water fluxes and flow speeds of different flowpaths (e.g., Kirchner et al., 2001; Tetzlaff et al., 2007; Hrachowitz et al., 2010; Botter et al., 2010; Van der Velde et al., 2010; Birkel et al., 2012; Heidbüchel et al., 2012; Peters et al., 2014). This nonstationarity is more than simply a time-domain analogue to the heterogeneity problem explored in Paper 1, because variations in flow regime may alter both the transit-time distributions of individual flowpaths and the mixing ratios between them. Intuition suggests that catchment nonstationarity could play havoc with estimates of MTT's, and perhaps also with estimates of the young water fraction.

This paper explores three central questions. First, does nonstationarity lead to aggregation errors in MTT, and thus to bias or scatter in MTT estimates derived from seasonal tracer cycles? Second, is the young water fraction F_{yw} also vulnerable to aggregation errors under nonstationarity, or is it relatively immune, like it is to aggregation errors arising from spatial heterogentity? Third, can either MTT or F_{yw} be estimated reliably from seasonal tracer cycles, in catchments that are both nonstationary and heterogeneous, as real catchments are?

In keeping with the spirit of the approach developed in Paper 1, here I explore the consequences of catchment nonstationarity through simple thought experiments. These thought experiments are based on a simple two-compartment conceptual model (Fig. 1). This model greatly simplifies the complexities of real-world catchments, but it is sufficient to illustrate the key issues at hand. It is not intended to simulate the behavior of a specific real-world catchment, and thus its "goodness of fit" to any particular catchment time series is unimportant. Instead, its purpose is to simulate how nonstationary dynamics may influence

tracer concentrations across wide ranges of catchment behavior, and thus to serve as a numerical "test bed" for exploring how catchment nonstationarity affects our ability to infer catchment transit times from tracer concentrations. One can of course construct more complicated and (perhaps) realistic models, but that is not the point here. The point here is to explore the consequences of catchment nonstationarity, in the context of one of the simplest possible models which nonetheless exhibits a wide range of nonstationary behaviors.

2 A simple conceptual model for exploring nonstationarity

2.1 Structure and basic equations

The model catchment consists of two compartments, an upper box and a lower box (Fig. 1). In typical conceptual models the upper box might represent soil water storage and the lower box might represent groundwater, but for the present purposes it is unnecessary to assign the two boxes to specific domains in the catchment. The upper box storage S_u is filled by precipitation P, and drains at a leakage rate L that is a power function of storage; for simplicity, evapotranspiration is ignored. Thus storage in the upper box evolves according to

Deleted: (or, equivalently, P can be interpreted as "effective" precipitation)

$$\frac{dS_u}{dt} = P - L = P - k_u S_u^{b_u} \quad , \qquad (1)$$

where the coefficient k_u and the exponent b_u are parameters. A third parameter $0<\eta<1$ partitions the leakage L from the upper box into an amount ηL that flows directly to discharge and an amount $(1-\eta)L$ that recharges the lower box. The lower box storage S_l is recharged by leakage from the upper box and drains to streamflow at a discharge rate Q_l that is another power function of storage,

$$\frac{dS_l}{dt} = (1-\eta) L - Q_l = (1-\eta) L - k_l S_l^{b_l} \quad , \qquad (2)$$

where the coefficient k_l and the exponent b_l are the final two parameters. The stream discharge is the sum of the contributions from the upper and lower boxes, or

$$Q_S = \eta L + Q_l \quad . \qquad (3)$$

All storages are in mm of water equivalent depth, and all fluxes are in mm/day. The age distribution in each box is explicitly tracked at daily resolution for the youngest 90 days, and

by accounting for the aggregate "age mass" (Bethke and Johnson, 2008) of each box's water that is older than 90 days. The young water fraction F_{yw} is calculated as the fraction of water in each box that is up to (and including) 69 days old; this threshold age equals 0.189 years, which was shown in Paper 1 to be the theoretical young-water threshold age for seasonal cycles in systems with exponential transit time distributions.

Discharge from both boxes is assumed to be non-age-selective, meaning that discharge is taken proportionally from each part of the age distribution; thus the flow from each box will have the same tracer concentration, the same young water fraction F_{yw}, and the same mean age as the averages of those quantities in that box (at that moment in time). Tracer concentrations and mean ages are tracked under the assumption that both boxes are each well-mixed but also separate from one another, so their tracer concentrations and water ages will differ. The tracer concentrations, young water fractions, and mean water ages in streamflow are the flux-weighted averages of the contributions from the two boxes.

The model is solved on a daily time step, using a weighted combination of the partly implicit trapezoidal method (for greater accuracy) and the fully implicit backward Euler method (for guaranteed stability). Details of the solution scheme are outlined in Appendix A.

2.2 Parameters and initialization

The drainage coefficients k_u and k_l are problematic as model parameters, because their values and dimensions are strongly dependent on the exponents b_u and b_l. Therefore I instead parameterize the model drainage functions by the (dimensionless) exponents b_u and b_l and by the (dimensional) "reference" storage values, $S_{u,ref}$ and $S_{l,ref}$. These reference values represent the storage levels at which the drainage rates of each box will equal their long-term average input rates. That is, $S_{u,ref}$ is the level of upper-box storage at which the leakage rate L equals the long-term average input rate $\overline{P}$. Likewise, $S_{l,ref}$ is the level of lower-box storage at which the discharge rate Q_l equals the average rate of recharge $(1-\eta)\,\overline{L}$ (which, due to conservation of mass in the upper box, also equals $(1-\eta)\,\overline{P}$). The drainage function coefficients are calculated from the reference storage values as follows:

$$k_u\, S_{u,ref}^{b_u} = \overline{P} \quad , \quad k_u = \overline{P}\, S_{u,ref}^{-b_u} \qquad (4)$$

Deleted: <#>Solution scheme¶ For simplicity and efficiency, the hydrological

Deleted: fixed

Deleted: . This requires some care with the numerics, given the clear (though often overlooked) dangers in naive forward-stepping simulations of nonlinear equations (Clark and Kavetski, 2010; Kavetski and Clark, 2010, 2011). Here I use

Deleted: which is partly implicit,

Deleted: enhanced

Deleted: which is fully implicit,

Deleted: The hydrological

Deleted: is illustrated here for the upper box; the lower box is handled analogously. The storage in the upper box is updated using the following equation:¶

$$S_u(t_{i+1}) - S_u(t_i) = \Delta t \left(P - \rho k$$

. . . (4)¶

where $S_u(t_i)$ is the storage in the upper box at the beginning of the i^{th} time interval (with length Δt), $S_u(t_{i+1})$ is the storage at the end of that interval (and thus the beginning of the next), and P is the average precipitation rate over the interval. Equation (4) is implicit and nonlinear; there is no closed-form solution for the future storage $S_u(t_{i+1})$, which instead is found using Newton's method. The relative dominance of the trapezoidal and backward Euler solutions is determined by the weighting factor ρ, which takes on values between ρ=0.5 (trapezoidal method) and ρ=1 (backward Euler method). The value of ρ in Eq. (4) is determined for each time step using the simple stability criterion,¶

$$\rho = \min\left(0.5 + 0.5 \frac{\left(P - k_u S_u(\right.}{\left(P / k_u\right)^{1/b_u}} \right.$$

. . . . . (5)¶

where the numerator represents the amount that S_u would change during one time step if the instantaneous drainage rate L in Eq. (1) were projected forward in time, and the denominator represents the difference between S_u's current value and its equilibrium value at the precipitation rate P. Equation (5) says that if the trapezoidal method would move S_u ... [1]

Deleted: unnecessary and the more accurate trapezoidal method should dominate the solution instead ($\rho \approx 0.5$). On the other hand, if the trapezoidal method would overshoot the equilibrium v ... [2]

Deleted: above also guarantees exact consistency between stocks and fluxes (but note, not in the usual way by updating stocks with fluxes, but rather by calculating output fluxes from inpu ... [3]

Deleted: 1

$$k_l \, S_{l,ref}^{b_l} = (1-\eta)\,\overline{P} \quad , \quad k_l = (1-\eta)\,\overline{P}\; S_{l,ref}^{-b_l} \qquad . \qquad (5)$$

Expressing k_u and k_l in this way is equivalent to writing the drainage equations for the two boxes in dimensionless form, with the drainage rate expressed with reference to the long-term input rate as follows:

$$\frac{L}{\overline{P}} = \left(\frac{S_u}{S_{u,ref}} \right)^{b_u} \qquad (6)$$

$$\frac{Q_l}{(1-\eta)\,\overline{P}} = \left(\frac{S_l}{S_{l,ref}} \right)^{b_l} \qquad (7)$$

One advantage of this approach is that, whereas the drainage coefficients k_u and k_l have no clear meaning and their numerical values and dimensions can vary wildly, the reference storage values are measured in mm of water equivalent depth, and their interpretation is straightforward. **A further advantage of this approach is that it provides for varying degrees of residual storage without requiring any additional parameters to do so.** Because $S_{u,ref}$ and $S_{l,ref}$ are the storage levels at which long-term mass balance is achieved, they represent the equilibria **around which** S_u and S_l will tend to fluctuate, with the range of those fluctuations largely determined by the variability in precipitation rates and by the stiffness of the drainage functions, as specified by the exponents b_u and b_l **(see Sect. 3.2).**

The storages are initialized at the reference values $S_{u,ref}$ and $S_{l,ref}$. The tracer concentrations are initialized at equilibrium (that is, at the volume-weighted mean of the precipitation tracer concentration). Likewise the mean ages in each box are initialized at their steady-state equilibrium values, $S_{u,ref} / \overline{P}$ in the upper box and $S_{u,ref} / \overline{P} + S_{l,ref} / [\overline{P}(1-\eta)]$ in the lower box. After a one-year spin-up period, I run the model for ten more years, with results for those ten years reported here.

2.3 Parameter ranges and precipitation drivers

Here I drive the model with three different real-world rainfall time series, representing a range of climatic regimes: a humid maritime climate with frequent rainfall and moderate seasonality (Plynlimon, Wales**; Köppen climate zone Cfb**), a Mediterranean climate marked by wet winters and very dry summers (Smith River, California, USA**; Köppen climate zone Csb**),

Deleted: 1

Deleted: 1

Deleted: 1

Deleted: it provides for varying degrees of residual storage without requiring any additional parameters to do so, as explained below. A further advantage of this approach is that

Deleted: that

Deleted: around

Deleted: .

and a humid temperate climate with very little seasonal variation in average rainfall (Broad River, Georgia, USA**; Köppen climate zone Cfa**). Figure 2 shows the contrasting frequency distributions and seasonalities of the three rainfall records. The Plynlimon rain gauge data were provided by the Centre for Ecology and Hydrology (UK), and the Smith River and Broad River precipitation data are reanalysis products from the MOPEX experiment [Duan, 2006 #2193; ftp://hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/]. The use of these real-world precipitation time series obviates the need to generate statistically realistic synthetic precipitation to drive the model.

The model used here shares a similar overall structure with many other conceptual models (e.g., Benettin et al., 2013), with several simplifications. But although the model used here is typical in many respects, I will use it in an unusual way. Typically one calibrates a model to reproduce the behavior of a real-world catchment, and then draws inferences about that catchment from the parameters and behavior of the calibrated model. Here, however, the model is not intended to represent any particular real-world system. Instead, the model itself is the system under study, across wide ranges of parameter values, because the goal is to gain insight into how nonstationarity affects general patterns of tracer behavior. Thus the fidelity of the model in representing any particular catchment is not a central issue.

For the simulations shown here, the drainage exponents b_u and b_l are randomly chosen from uniform distributions spanning the ranges of 1-20 and 1-50, respectively, the partitioning coefficient η is randomly chosen from a uniform distribution ranging from 0.1 to 0.9, and the reference storage levels $S_{u,ref}$ and $S_{l,ref}$ are randomly chosen from a uniform distribution of logarithms spanning the ranges of 20-500 mm and 500-10,000 mm, respectively. These parameter distributions are designed to encompass a wide range of possible behaviors, including both strong and damped response to rainfall inputs, and small and large residual storage. To illustrate the behavior of the model for one concrete case, I use a "reference" parameter set with values taken from roughly the middle of each of these parameter distributions (b_u=10, b_l=20, η=0.5, $S_{u,ref}$=100 mm, and $S_{l,ref}$=2000 mm). These parameter values are not "better" than any others in any particular sense; they are simply a point of reference (hence the name) for discussing the model's behavior.

3 Results and discussion

3.1 Nonstationarity in the two-box model

My main purpose is to use the simple two-box model to explore how catchment nonstationarity affects our ability to infer water ages from tracer time series. I will take up that issue beginning in Sect. 3.3 below. As background for that analysis, however, it is helpful to first characterize the nonstationary behavior of the simple model system.

Figure 3 shows excerpts from the time series generated by the model with the Smith River (Mediterranean climate) precipitation time series and the reference parameter set. One can immediately see that the upper and lower boxes have markedly different mean ages (Fig. 3e), young water fractions (Fig. 3d), and tracer concentrations (Fig. 3c), which also vary differently through time. Tracer concentrations in the upper box (the orange line in Fig. 3c) show a blocky, irregular pattern, remaining almost constant during periods of little rainfall, and then changing rapidly when the box is episodically flushed by large precipitation events. The lower box's tracer concentrations (the red line in Fig. 3c) are much more stable than the upper box's, because its mean residence time is roughly 40 times longer ($S_{l,ref}$ is 20 times $S_{u,ref}$, and with η=0.5, the flux through the lower box is only half of the flux through the upper box). Because much more rain falls during the winters than the summers, the mean tracer concentration in the lower box is closer to the winter concentrations than the summer concentrations. During the wet winter season, rapid flushing keeps the young water fraction near 100% in the upper box (the orange line in Fig. 3d), and can raise the young water fraction to 30-40% in the lower box (the red line in Fig. 3d). Conversely, during the late summer the young water fraction in the upper box temporarily dips to 50% or less, and the young water fraction in the lower box declines to nearly zero. The small volume in the upper box means that its water age (the orange line in Fig. 3e) is only a small fraction of a year. The mean water age in the lower box (the red line in Fig. 3e) is much older and exhibits both seasonal variation and inter-annual drift, reflecting year-to-year variations in total precipitation. Thus the two components of this simple system have strongly contrasting characteristics and behavior. These internal states of any real-world system would not be observable, except as they are reflected in the volume and composition of streamflow.

In this regard, the most striking feature of Fig. 3 is the volatility of the tracer concentrations, young water fractions, and mean transit times in discharge (the dark blue lines in Figs. 3c-e),

Deleted: 4

Deleted: drainage bypasses the lower box entirely

Deleted: 3

as the mixing ratio between the two boxes (Fig. 3b) shifts in response to precipitation events. This mixing ratio is not a simple function of discharge (Fig. 4c); instead it is both hysteretic and nonstationary, varying in response both to precipitation forcing and to the antecedent moisture status of the two boxes (and thus to the prior history of precipitation). This dependence on prior precipitation reflects the fact that the boxes typically retain their water age and tracer signatures over time scales much longer than the timescale of hydraulic response, because their residual storage is large compared to their dynamic storage (see Sect. 3.2). As a result, both the young water fraction and mean age of discharge and storage are widely scattered functions of discharge (Figs. 4a and b). Likewise there is no simple relationship between either the young water fraction or mean age in storage and the corresponding quantities in discharge (Fig. 4d), although there is a strong overall bias toward water in discharge being much younger than the average water in storage.

Even though drainage from each box is non-age-selective (that is, the young water fraction and mean age in drainage from each box are identical to those in storage), this is emphatically not true at the level of the two-box system, because the two boxes account for different proportions of discharge than of storage. Furthermore, because the fractional contributions to streamflow from the (younger, smaller) upper box and the (older, larger) lower box are highly variable, the water age and young water fraction in discharge are not only strongly biased, but also highly scattered, indicators of the same quantities in storage (Fig. 4d).

The aggregate long-term implications of these dynamics are evident in the marginal (time-averaged) age distributions of storage and discharge (Fig. 5). From Fig. 5 it is immediately obvious that the age distributions in discharge are strongly skewed toward young ages, compared to the age distributions in storage, both for each box individually and for the catchment as a whole. This skew toward young ages arises for two main reasons. First, although drainage from each box is not age-selective, more outflow occurs during periods of stronger precipitation forcing, and thus shorter residence times. Thus the average ages of the outflow and the storage can differ greatly. Second, under high-flow conditions a larger proportion of discharge is derived from the upper box (which has a relatively short transit time), and at base flow more discharge is derived from the lower box (which has a larger volume and a relatively long transit time). Thus the short-transit-time components of the system dominate the discharge, while the long-transit-time components of the system dominate the storage. As a result, the mean age in discharge will generally be much younger

Deleted: below

Deleted: 4

Deleted: Whereas the short-term relationships between water age in storage and discharge can be visualized through scatterplots like Fig. 4, their

Deleted: in

than the mean age in whole-catchment storage, and likewise the young water fraction in discharge will be much larger than the young water fraction in storage. Note that this is the opposite of what one would expect from conceptual models **like those of Botter (2012)**, in which the mean water age in discharge either equals the mean age in storage (for well-mixed systems), or is older than the mean age in storage (for piston-flow systems).

Deleted: steady-state

More generally, and more importantly, these results imply that estimates of water age in streamflow cannot be translated straightforwardly into estimates of water age in storage. Instead, they may underestimate the age of water in storage by large factors, although in the particular example shown in Fig. 5, the difference is only about a factor of two. **Three closely related theoretical functions have recently been proposed to quantify the long-recognized (Kreft and Zuber, 1978) disconnect between the age distributions in storage and in discharge. These include the time-dependent StorAge Selection (SAS) function ω_Q of Botter et al. (2011), the Storage Outflow Probability (STOP) functios of van der Velde et al. (2012), and the rank StorAge Selection (rSAS) function of Harman (2015). While these functions are all grounded in elaborate theoretical frameworks, it remains to be seen whether they can be reliably estimated in practice using real-world data.**

A further implication **of the analysis above** is that the marginal age distributions are not exponential, even for individual boxes, **and** even though drainage from each box is not age-selective. In steady state, non-age-selective drainage (i.e., the well-mixed assumption) would yield an exponential distribution of ages in the upper box and in the short-time age distribution in streamflow. However, when the system is not in steady state and we aggregate its behavior over time, we are combining different age distributions from different moments in time with different precipitation forcing. This creates an aggregation error in the time domain, in the sense that the steady-state approximation will be a misleading guide to the non-steady-state behavior of the system, *even on average*. That is, even over time scales where inputs equal outputs and the long-term average fluxes are essentially constant – and thus the steady-state approximation, on average, holds – the average behavior of the non-steady-state system can differ significantly from the average behavior of an equivalent steady-state system.

One can further explore these issues by examining the marginal (time-averaged) age distributions for separate ranges of discharge (Fig. 6). Figure 6 shows that at higher

discharges, age distributions in streamflow are much more strongly skewed toward younger ages, reflecting the increased dominance of the upper box at higher flows. For the upper half of all discharges, the age distributions are more skewed than exponential; that is, they plot as upward-curving lines in Fig. 6b. For the top 25% of discharges, they are approximately power-law, plotting as nearly straight lines in Fig. 6c. The slopes of these lines are steeper than 1, however, implying that the distributions must deviate from this trend at very short ages; otherwise their integrals (i.e., their cumulative distributions) would become infinite. It is important to note the mean ages quoted in Fig. 6a imply that the tails of the distributions all extend far beyond the plot axes, which are truncated at 90 days. Note also that the distributions shown in Fig. 6 have different shapes in different flow regimes, suggesting that the model's high-flow behavior is not simply a re-scaled transform of its low-flow behavior.

Deleted: top

Deleted: The age distributions

Deleted: for the top 25% of discharges

Deleted: However, the

Deleted: It is

Deleted: important to note

3.2 Residual storage and the disconnect between transit time and hydraulic response time scales

The model's complex, nonstationary water age and tracer dynamics arise from the disconnect between the timescales of hydraulic response and catchment storage in each box, and from the divergence in both these timescales between the two boxes. These contrasting timescales can be estimated through simple scaling and perturbation analyses, as outlined in this section.

Total catchment storage consists of two components: the dynamic storage that is linked to discharge fluctuations through storage-discharge relationships like Eqs. (6)-(7), plus the residual or "passive" storage that remains when discharge has declined to very slow rates. The range of dynamic storage exerts an important control on timescales of catchment hydrologic response, while the much larger residual (or "passive") storage has little effect on water fluxes but is an essential control on residence times (Kirchner, 2009; Birkel et al., 2011).

Deleted: (16)-(1

Deleted: Equations (16)-(17

In real-world catchments, sharply nonlinear storage-discharge relationships (Kirchner, 2009) guarantee that **dynamic storage will be small compared to residual storage. This behavior is mirrored in the model, where if Eqs. (6)-(7) are strongly nonlinear (i.e.,** if the drainage exponents b_u and b_l are much greater than 1**)**, the volumes in the upper and lower boxes will vary by only a small fraction of their reference storage values $S_{u,ref}$ and $S_{l,ref}$ (e.g., Fig. 3f). They will **remain** relatively **constant** because, when the drainage exponents b_u and

Deleted: vary

Deleted: little

b_l are large, the storage volumes cannot become much smaller than $S_{u,ref}$ and $S_{l,ref}$ without drainage rates falling to near zero (thus stopping further decreases in storage), and conversely, the storage volumes also cannot become much larger than $S_{u,ref}$ and $S_{l,ref}$ without drainage rates becoming very high (thus stopping further increases in storage). Thus $S_{u,ref}$ and $S_{l,ref}$ will be a good approximation to the residual storage volume, whenever the drainage exponents are much greater than 1.

One can express this concept more quantitatively (though only approximately) using a simple perturbation analysis. A first-order Taylor expansion of Eqs. (6) and (7) shows directly that the fractional variability in drainage rates and storage are related by the drainage exponents in the two boxes:

$$\frac{\Delta L}{\overline{P}} \approx b_u \frac{\Delta S_u}{S_{u,ref}} \quad (8)$$

$$\frac{\Delta Q_l}{(1-\eta)\overline{P}} \approx b_l \frac{\Delta S_l}{S_{l,ref}} \quad (9)$$

The variability in drainage rates from the upper and lower boxes, denoted as ΔL and ΔQ_l, will be controlled by the temporal variability in precipitation; thus for a given precipitation climatology, the dynamic variability in storage (denoted as ΔS_u and ΔS_l) will scale according to the ratios $S_{u,ref}/b_u$ and $S_{l,ref}/b_l$. For example, when the model is driven by Smith River precipitation and uses the reference parameters (Fig. 3), the variability in discharge from the lower box, as measured by its standard deviation, is 3.7 mm d^{-1}, nearly equal to the average lower box discharge of 3.8 mm d^{-1}. Because the reference value of b_l is 20, Eq. (9) implies that the standard deviation of lower box storage should be approximately $1/20^{th}$ of the reference storage $S_{l,ref}$, or roughly 100 mm. Consistent with this estimate, the actual standard deviation of S_l is 84 mm or about 4% of the total. Figure 3f shows that at least 90% of $S_{l,ref}$ is residual storage that never drains during the 10-year simulation, roughly consistent with the perturbation analysis.

The perturbation analysis also yields estimates for the time scale of hydraulic response (which controls how "flashy" the discharge will be), through a rearrangement of Eqs. (8) and (9) as follows:

Deleted: their reference storage values

Deleted: their reference storage values

Deleted: The

Deleted: 1

Deleted: 1

Deleted: 1

Deleted: 1

Deleted: 1

Deleted: ;

Deleted: , implying that most of the rest is residual storage

Deleted: 1

Deleted: 1

$$\frac{\Delta S_u}{\Delta L} \approx \frac{S_{u,ref}}{b_u \overline{P}} \quad \text{(hydraulic response time scale, upper box)} \quad (10)$$

$$\frac{\Delta S_l}{\Delta Q_l} \approx \frac{S_{l,ref}}{b_l (1-\eta) \overline{P}} \quad \text{(hydraulic response time scale, lower box)} \quad (11)$$

Again using the reference parameter values and Smith River precipitation (for which $\overline{P}$ is roughly 7.6 mm d^{-1}), Eqs. (10) and (11) imply a hydraulic response time of roughly 1.3 days (for b_u=10) in the upper box and roughly 26 days (for b_l=20) in the lower box. These time scales are factors b_u and b_l smaller than the steady-state mean transit times, which are determined by the ratios between the volumes and water fluxes,

$$\frac{S_{u,ref}}{\overline{P}} \quad \text{(steady-state mean transit time, upper box)} \quad (12)$$

$$\frac{S_{l,ref}}{(1-\eta) \overline{P}} \quad \text{(steady-state mean transit time, lower box)} \quad (13)$$

From Eqs. (12)-(13) one can also directly estimate the steady-state mean travel time in the combined discharge as the weighted average of streamflow derived directly from the upper box, and water that flows through the upper and lower boxes in series,

$$\eta \frac{S_{u,ref}}{\overline{P}} + (1-\eta)\left(\frac{S_{u,ref}}{\overline{P}} + \frac{S_{l,ref}}{(1-\eta)\overline{P}} \right) = \frac{S_{u,ref} + S_{l,ref}}{\overline{P}} \quad (14)$$

which is the expected result for any system at steady state: regardless of its internal configuration, the mean transit time in any steady-state system will equal the ratio between its storage volume and its throughput rate. For the reference parameter set and Smith River precipitation, Eq. (14) becomes (100 mm + 2000 mm)/7.6 mm d^{-1}, or roughly 0.76 years, in good agreement with the whole-catchment mean transit time **of 0.74 years determined** from age tracking **(see** Fig. 5d**)**. Note, however, that the *distribution* of these transit times will be markedly different from the exponential distribution that would be expected in steady state. This makes estimating mean transit times from tracer fluctuations difficult, as shown below in Sect. 3**.3**.

Equations (12) and (13) imply that the mean transit times in the upper and lower boxes should be roughly 13 days (or 0.036 yr) and 529 days (or 1.45 yr), respectively, in good agreement with the mean transit times **of 0.03 and 1.44 years determined** from age tracking (Fig. 5d).

Deleted: 20
Deleted: 21
Deleted: 20
Deleted: 21
Deleted: 22
Deleted: 23
Deleted: (22)-(23
Deleted: 24
Deleted: 24
Deleted: quoted in
Deleted: 4
Deleted: 22
Deleted: 23
Deleted: estimated

However, Eqs. (10)-(11) imply that these transit times will differ by factors of 10 and 20 (the values of b_u and b_l, respectively) from the hydraulic response timescales that regulate catchment runoff response. The disconnect between hydraulic response times and mean transit times is the counterpart, in lumped conceptual models, to the disconnect between the velocity of water transport and the celerity of hydraulic head propagation in more realistic, physically extended systems (Beven, 1982; Kirchner et al., 2000; McDonnell and Beven, 2014). This contrast between hydraulic response times and mean transit times (or dynamic and total storage, or celerity and velocity) is a simple explanation for the apparent paradox of prompt discharge of old water during storm events (Kirchner, 2003).

3.3 Inferring MTT and F_{yw} from seasonal tracer cycles in nonstationary catchments

The analysis above shows that the simple two-box model gives hydrograph and tracer behavior that is complex and nonstationary (Figs. 3-6). Furthermore, even this simple five-parameter model exhibits strong equifinality (Appendix B). Much of this equifinality can be alleviated (compare Figs. B1 and B2) through parameter transformations based on the perturbation analysis outlined above. However, because the timescales of catchment storage and hydraulic response are controlled by different combinations of parameters, parameter calibration to the hydrograph cannot constrain the storage volumes or streamwater age (Figs. B2-B3). These model results demonstrate general principles that have been recognized for years: a) the hydrograph responds to, and thus can help to constrain, dynamic storage but not passive storage, and b) because passive storage is often large, timescales of hydrologic response and catchment water storage are decoupled from one another, such that water ages cannot be inferred from hydrograph dynamics. Thus for understanding how catchments store and mix water, tracer data are essential.

But how should these tracer data be used? One approach is to explicitly include tracers in a catchment model, and calibrate that model against both the hydrograph and the tracer chemograph (e.g., Birkel et al., 2011; Benettin et al., 2013; Hrachowitz et al., 2013). The usefulness of that approach depends on whether the model parameters can be constrained and, more importantly, whether the model structure adequately characterizes the system under study (which is usually unknown, and possibly unknowable). Except in multi-model studies, it will be unclear how much the conclusions depend on the particular model that was used,

Deleted: (20)-(21

Deleted: the simplest

Deleted: <#>Equifinality in hydraulic behavior and divergence in travel times¶
The analysis outlined in Sect. 3.2 further implies that approximate equifinality is inevitable, even in such a simple model, because variations in the exponents b_u and b_l and the reference storage levels $S_{u,ref}$ and $S_{l,ref}$ will have nearly offsetting effects on the model's runoff response. Equations (20) and (21) show that, for a given average precipitation forcing, any parameter values for which the partitioning coefficient η and the ratios $S_{u,ref}/b_u$ and $S_{l,ref}/[(1-\eta)b_l]$ are invariant would give nearly equivalent hydrograph predictions, because the hydraulic response timescales of the upper and lower boxes, and their relative contributions to discharge, would be invariant. These conditions can be achieved for widely varying values of the individual parameters b_u, b_l, $S_{u,ref}$, and $S_{l,ref}$. This equifinality problem can be readily visualized by plots like Fig. 7. To generate Fig. 7, I ran the model with Smith River precipitation forcing and the reference parameter set (shown by the red squares in Fig. 7), and used the resulting daily hydrograph (after the spin-up period) as virtual "ground truth" for model calibration. I then ran the model with 1000 random parameter sets, and used the Nash-Sutcliffe efficiency (NSE) of the logarithms of discharge to measure how well their hydrographs matched the reference hydrograph (thus the reference hydrograph has NSE=1 by definition). The 50 best-fitting parameter sets, all with NSE≥0.98, are shown as dark blue points in Fig. 7. The bottom row of scatterplots shows the conventional "dotty plots". Their flat tops are the hallmark of equifinality, i.e., wide ranges of parameter values give equally good hydrograph predictions (Beven, 2006). Only the partition coefficient η, which performs well across half its range, can be even modestly constrained by calibration. (The other precipitation drivers yield res ... [4]

Formatted: Bullets and Numbering

Formatted: Font: Italic

Formatted: Font: Italic, Subscript

Deleted: That analysis further shows that

Deleted: in the model, and thus that

Deleted: is largely decoupled from hydrograph characteristics

Deleted: 7-9

Deleted: vividly

Deleted: (as distinct from how they transmit hydraulic potentials, which are more clearly reflected in the hydrograph)

and the particular way that it was fitted to the data. Furthermore, adequate tracer data for calibrating such models are rare, particularly because dynamic models require input data with no gaps. The mismatch between model complexity and data availability means that in some cases, all the data are used for calibration and validation must be skipped, leaving the reproducibility of the model results unclear (e.g., Benettin et al., 2015).

For all of these reasons, there will be an ongoing need for methods of inferring water ages that have modest data requirements and that are not dependent on specific model structures and parameters. Sine-wave fitting of seasonal tracer cycles, for example, is not based on a particular mechanistic model, but instead is based on a broader conceptual framework in which stream output is some convolution of previous precipitation inputs. That premise is of course open to question, but nevertheless seasonal tracer cycles (of, e.g., ^{18}O, ^{2}H, and Cl^{-}) have been widely used to estimate mean catchment transit times (see McGuire and McDonnell, 2006, and references therein), largely because this particular method has modest data requirements. In particular, it does not need unbroken records of either precipitation inputs or streamflow outputs.

As detailed more fully in Paper 1, the seasonal tracer cycle method is based on the principle that when one convolves a sinusoidal tracer input with a transit time distribution (TTD), one obtains a sinusoidal output that is damped and phase-lagged by an amount that depends on the shape of the TTD and also on its scale, as expressed, for example, by its mean transit time (MTT). Conventionally one assumes an exponential TTD, which is the steady-state solution for a well-mixed reservoir. More generally, one might assume that transit times are gamma-distributed, recognizing that the exponential distribution is a special case of the gamma distribution (with the shape factor α equal to 1). A sinusoidal tracer cycle that has been convolved with a gamma TTD will be damped and phase-lagged as described in Eqs. (8) and (9) of Paper 1. These equations can then be inverted to infer the shape and scale of the TTD from the seasonal tracer cycles in precipitation and streamflow.

The procedure is as follows. One first measures the amplitudes and phases of the seasonal tracer cycles in precipitation and streamflow using Eqs. (4)-(6) of Paper 1. If one assumes an exponential TTD, one can estimate the MTT directly from the amplitude ratio A_S/A_P in streamflow and precipitation using Eq. (10) of Paper 1 with α=1. Where I plot results from this procedure (i.e., Fig. 7) the corresponding axis will say "MTT inferred from A_S/A_P". This

Deleted: in review

Deleted: ; in

Deleted: are plotted

Deleted: 10

is the approach that is conventionally used in the literature. Alternatively, as I showed in **Sect. 4.4 of** Paper 1, one can use the tracer cycle amplitude ratio A_S/A_P and phase shift ϕ_S-ϕ_P to jointly estimate the shape factor α and the MTT (assuming the TTD is gamma-distributed, which is less restrictive than assuming that it is exponential). To do this one estimates the shape factor α **from A_S/A_P and ϕ_S-ϕ_P** using Eq. 11 from Paper 1, and then estimates the scale factor β using Eq. 10 from Paper 1; the MTT is α times β. MTT's estimated by this procedure are shown in Figs. **10-12** as "MTT inferred from A_S/A_P and ϕ_S-ϕ_P".

Paper 1 shows that both of these MTT measures are extremely vulnerable to aggregation bias in spatially heterogeneous catchments. Therefore Paper 1 proposes an alternative measure of travel times, the young water fraction F_{yw}, which is designed to be much less sensitive than MTT to aggregation artifacts. F_{yw} is the fraction of streamflow that is younger than a specified threshold age. For a seasonal cycle (i.e., with a period of 1 year) and reasonable range of TTD shapes, the threshold age varies between about 0.15 and 0.25 years, or equivalently ~2-3 months (see Eq. 14 and Fig. 10 in Paper 1). As described in Sect. 2, in the model simulations the "true" F_{yw} is defined by a threshold age of 0.189 years (69 days), which equals the threshold age for seasonal cycles convolved with an exponential TTD.

One can use seasonal tracer cycles to infer the young water fraction following either of two strategies. As shown in Sect. 4.1 of Paper 1, in many situations F_{yw} is approximately equal to the amplitude ratio A_S/A_P itself (indeed, it was designed to have this property). In figures where the amplitude ratio A_S/A_P is used **as an** estimate **of** F_{yw} (e.g., Fig. **7**), the axis says simply "F_{yw} inferred from A_S/A_P". Alternatively, one can use **both** the amplitude ratio A_S/A_P and phase shift ϕ_S-ϕ_P to estimate **F_{yw}, as explained in Sect. 4.4 of Paper 1. First, one estimates** the shape factor α **from A_S/A_P and ϕ_S-ϕ_P** using **Paper 1's** Eq. (11**).** One **then determines** the threshold age **τ_{yw} from α using** Paper **1's Eq. (14), and** the scale factor β **from α** and **A_S/A_P using** Paper **1's Eq. (10). Lastly, one estimates F_{yw} as** lower incomplete gamma function **$\Gamma(\tau_{yw}, \alpha, \beta)$ (Eq. 13 of Paper 1).** Where I have followed this more complex procedure (e.g., Figs. **9-**12), the figure axes say "F_{yw} inferred from A_S/A_P and ϕ_S-ϕ_P". **All of these F_{yw}'s and MTT's are intended as temporal averages, reflecting whatever conditions (e.g., precipitation climatologies or flow regimes) have shaped the seasonal cycles that are used to estimate them.**

Deleted: 13-16

Deleted: .1 above

Deleted: to directly

Deleted: 10

Deleted:

Deleted: jointly

Deleted:) of Paper 1.

Deleted: can

Deleted: use Eq. (14) of Paper 1 to determine

Deleted: , Eq. (10) of

Deleted: 1 to determine

Deleted: ,

Deleted: Eq. (13) of

Deleted: 1 to estimate F_{yw} from the

Deleted: for

Formatted: Font: Not Italic

Formatted: German (Germany)

Deleted: , and the threshold age.

Deleted: -16

These methods for inferring the young water fraction F_{yw} are derived from the properties of gamma TTD's. However, as I showed in Sects. 4.2-4.3 of Paper 1, these methods reliably estimate F_{yw} for very wide ranges of catchment TTD's (beyond the already broad family of gamma distributions), at least in catchments that are **spatially** heterogeneous but time-invariant. Here I explore whether these methods are also reliable in nonstationary catchments (and, in Sect. 3.**5** below, in catchments that are both nonstationary and **spatially** heterogeneous).

Figure **7** shows the **true** young water fractions F_{yw} and mean transit times (MTT's) in discharge from the two-box model, compared to estimates of F_{yw} and MTT inferred from the model's seasonal tracer cycles. As Figs. **7a**-c show, the amplitude ratios A_S/A_P of seasonal tracer cycles reliably estimate the true young water fractions in the model streamflow, across 1000 random parameter sets encompassing a very wide range of nonstationary catchment behavior. The slight underestimation bias in Figs. **7a**-c is reduced when both amplitude and phase information are used to estimate F_{yw} (Figs. **7d**-f). Under strongly seasonal precipitation forcing (Smith River; right panels in Fig. **7**), the seasonal tracer cycles underestimate F_{yw} by roughly 0.1 to 0.2, although the predicted and observed values of F_{yw} remain strongly correlated. For the other two precipitation drivers (Broad River and Plynlimon), the predicted and observed values of F_{yw} correspond almost exactly. Thus Fig. **7** shows that the young water fraction is relatively insensitive to aggregation error under nonstationarity, mirroring its robustness against spatial heterogeneity (as shown in Paper 1). By contrast, estimates of MTT are strongly biased and widely scattered, even on logarithmic axes (lower panels, Fig. **7**).

One additional complication in nonstationary situations, compared to the time-invariant examples explored in Paper 1, is that the young water fraction F_{yw} and mean transit time MTT can be expressed either as simple averages over time (representing the F_{yw} or MTT of an average *day* of streamflow), or as flow weighted averages (representing the F_{yw} or MTT of an average *liter* of streamflow). These quantities will not be equivalent, since higher flows will typically have higher F_{yw}'s and shorter MTT's (Figs. 3 and 4). Likewise one can expect that amplitudes of flow-weighted and un-weighted fits to the seasonal tracer cycles will be different. As the light blue points in Fig. **7** show, amplitude ratios of flow-weighted fits to the seasonal tracer cycles accurately predict the flow-weighted F_{yw} in streamflow; likewise, as the dark blue points show, the amplitude ratios of un-weighted fits accurately predict the un-weighted F_{yw} in streamflow. **The flow-weighted fits to the seasonal tracer cycles were**

Deleted: 6
Deleted: 10
Deleted: 10a
Deleted: 10a
Deleted: 10d
Deleted: 10
Deleted: 10
Deleted: 10
Deleted:

calculated by weighted least squares, with weights proportional to streamflow or precipitation volume. (In real-world applications, a robust fitting technique like Iteratively Reweighted Least Squares (IRLS) can be used to limit the influence of outliers. An R script for performing volume-weighted IRLS is available from the author.)

The underestimation bias in F_{yw} observed under the Smith River precipitation forcing may arise because the assumed tracer cycle is correlated with the strong seasonality in precipitation, such that tracer concentrations peak during the summer, when almost no rain falls. Thus the effective variability of tracer inputs to the catchment is less than one would infer from a sinusoidal fit to the precipitation tracer concentrations (and volume-weighting the fit does not help, because in these synthetic precipitation data the fit is exact so there are no residuals on which the weighting can have any effect). Because the tracer concentration amplitude overestimates the effective variability in tracer concentrations reaching the catchment, the tracer damping in the catchment is overestimated and thus the F_{yw} is underestimated. This underestimation bias disappears if one shifts the phase of the assumed precipitation tracer concentrations so that they peak in the spring or fall, and thus are uncorrelated with the seasonality in precipitation volumes. I have not done so here, however, because stable isotope ratios in precipitation typically peak in the mid-summer at latitudes poleward of ~35° (Feng et al., 2009), where most catchment studies have been conducted. Thus Fig. 7 suggests the potential for bias in F_{yw} estimates at sites where isotope cycles are correlated with very strong precipitation seasonality. **However, even under the strongly seasonal Smith River precipitation forcing, the bias in inferred F_{yw} values is small compared to the a priori uncertainty in F_{yw} (which is of order 1), and small compared to the bias in inferred MTT's (which is large even on logarithmic axes).**

Deleted: 10

Figures 7g-i compare the MTT in streamflow with estimates of MTT as they are conventionally calculated, that is, from the seasonal tracer cycle amplitude assuming an exponential TTD. These figures show that these conventional estimates are subject to a strong underestimation bias, which can exceed an order of magnitude. Some of the MTT estimates do fall close to the 1:1 line, but these are mostly cases in which the partition coefficient η is very small, such that nearly all drainage from the upper box is routed through the lower box, thus transforming the two-box, nonstationary model into a nearly one-box, nearly stationary model. The strong aggregation bias in MTT under catchment

Deleted: 10g

nonstationarity shown in Figs. 7g-i mirrors the similarly strong bias under **spatial** heterogeneity that was demonstrated in Paper 1.

The implication of Figs. 7g-i (and of Paper 1) is that many of the MTT values in the literature are likely to **be** underestimated by large factors, and thus that real-world catchment MTT's are likely to be much longer than we thought. This observation raises the question: where is all that water being stored? In steady state, the storage volume must equal the discharge multiplied by the MTT (see Sect. 3.2). Thus if we have been underestimating MTT's by large factors, then we have also been underestimating catchment storage volumes by similar multiples. Where is the storage volume that can accommodate all this water?

One possible answer is that in a non-steady-state system, the MTT decreases with increasing discharge (e.g., Fig. 4b), and the storage volume equals the discharge multiplied by the *volume-weighted* MTT rather than the *time-averaged* MTT. Because the volume-weighted MTT is less (potentially much less) than the time-averaged MTT (see also Peters et al., 2014), the implied storage volume is correspondingly smaller. Furthermore, many MTT studies in the literature have been based on tracer sampling that excludes high flows, such that they infer the mean age of baseflow rather than of the average discharge (McGuire and McDonnell, 2006). To the extent that mean baseflow discharges are lower than mean total discharges, the stored volume of baseflow water will be less than what one might overestimate by multiplying the mean *total* discharge by the mean *baseflow* age. Beyond these general considerations, however, it makes little sense to draw precise inferences based on MTT estimates that are likely to be strongly biased and widely scattered **(as shown here, and also in Paper 1)**.

It is important to recognize that the predicted F_{yw} values are really predictions, unlike many "predictions" from calibrated models. The horizontal axes in Fig. 7 are calculated solely from the age-tracking within the model, with no information about the tracer concentrations. Likewise the vertical axes in Fig. 7 are calculated from the modeled tracer cycles alone, without any information about the model that generated them, and in particular without any information about the modeled age of streamflow. Thus Fig. 7 gives some basis for confidence that estimates of F_{yw} will also be reliable in real-world catchments, where the true "model" can never be known.

Deleted: 10g

Deleted: 10g

Deleted: (as shown here, and in Paper 1)

Deleted: 10

Deleted: 10

Deleted: 10

3.4 Young water fractions in discrete flow regimes

Figures 3 and 4 show that high-flow periods are characterized by shorter mean transit times and higher young water fractions, reflecting the increased dominance of drainage from the upper box with its younger water ages. Although instantaneous transit time distributions (TTD's) can be highly variable, and thus instantaneous mean transit times and young water fractions can exhibit scattered relationships with discharge (Fig. 4), the marginal (time-averaged) TTD's in Fig. 6 clearly show systematically stronger skew toward younger water ages in higher ranges of streamflow. Thus, as Fig. 6 shows, the TTD varies in shape, not just in scale, between different flow regimes.

This observation leads naturally to the question of whether these variations in TTD's are also reflected in streamflow tracer concentrations, and whether those tracer signatures can be used to draw inferences about the TTD's that characterize individual flow regimes. Figure 3 shows that high-flow periods typically exhibit wider variations in tracer concentrations, reflecting greater contributions from the upper box, which has shorter residence times and thus more labile tracer concentrations than the lower box does. To test how systematic these variations in concentrations are, I ran the model with the reference parameter set and Plynlimon **(temperate maritime)** precipitation forcing, and separated the resulting time series into six discharge ranges. Figure **8** shows these six discharge ranges and the corresponding tracer concentrations in dark blue, superimposed on the entire discharge and concentration time series in light gray. As Fig. **8** shows, seasonal tracer cycles at higher flows are systematically less damped and phase-shifted (relative to the tracer cycle in precipitation, shown by the dotted gray line), implying shorter MTT's and larger young water fractions.

To test whether these changes in the seasonal tracer cycles are quantitatively consistent with the shifts in water age across the six flow regimes, I **fitted sinusoids separately to the tracer concentrations in each individual discharge range (Fig. 8). I compared these with a single sinusoid fitted to the entire precipitation tracer time series (because it is not possible to assign discrete precipitation events to individual discharge ranges). From the resulting amplitude ratios and phase shifts for each discharge range, I then estimated F_{yw} and MTT using** the methods outlined in Sect. 3**.3.** Figure **9** presents the results of this thought experiment, showing that the **time-averaged (but flow-specific)** young water

Formatted: Bullets and Numbering

Deleted: 11

Deleted: 11

Deleted: applied

Deleted: 4 to each discharge range.

Deleted: 12

fraction F_{yw} in each discharge range is accurately predicted by the damping and phase shift of the corresponding seasonal tracer cycle.

To test whether this result is general, I repeated this thought experiment for 200 random parameter sets and all three precipitation drivers. The results are shown in Fig. 10, with each discharge range plotted in a different color. The colors overlap because the discharge ranges, F_{yw}'s and MTT's all vary substantially from one parameter set to the next. The amplitudes and phase shifts of the seasonal tracer cycles predict the **time-averaged** young water fractions F_{yw} in each discharge range with reasonable accuracy (upper panels, Fig. 10). Somewhat surprisingly, the F_{yw} underestimation bias seen in Figs. 7c and 7f under the highly seasonal Smith River precipitation forcing does not arise in the predicted F_{yw} values for the separate discharge ranges (Fig. 10c). In contrast to the generally close correspondence between the predicted and observed F_{yw} values, predicted MTT's are very widely scattered for all discharge ranges and all precipitation forcings (lower panels, Fig. 10).

Deleted: 13

Deleted: 13

Deleted: 10c

Deleted: 10f

Deleted: 13c

Deleted: 13

Formatted: Bullets and Numbering

3.5 Combined effects of nonstationarity and spatial heterogeneity

Paper 1 explored whether mean travel times and young water fractions can be reliably inferred from tracer dynamics in spatially heterogeneous (but stationary) catchments, **composed of diverse subcatchments with different (but time-invariant) TTD's. The** sections above have presented a similar analysis for nonstationary (but spatially homogeneous) catchments. However, real-world catchments are not *either* heterogeneous *or* nonstationary; instead they are *both* heterogeneous *and* nonstationary. **That is, their subcatchments each exhibit nonstationary dynamics that may vary greatly from one to the next.** To explore the combined effects of nonstationarity and **spatial** heterogeneity, I merged the approach developed in Paper 1 with the model developed in Sect. 2 above.

As **illustrated** in Fig. 11, I ran eight copies of the nonstationary model developed in Sect. 2, representing eight different tributaries, each with a different, randomly chosen parameter set. I chose the number eight to provide a reasonable degree of complexity and heterogeneity while preserving a reasonable degree of computational efficiency. I supplied the same precipitation forcing (Fig. 11a) to all eight models (Fig. 11b) to simulate the behavior of the eight hypothetical tributary streams (Fig. 11c). I then simulated the merging of these streams by averaging their discharges, and taking volume-weighted averages of their tracer concentrations, young water fractions, and water ages (Fig. 11d). Because the instantaneous

Deleted: and the

Deleted: ilustrated

Deleted: 14

Deleted: 14a

Deleted: 14b

Deleted: 14c

Deleted: 14d

flows from the eight tributaries vary differently through time, their mixing ratios also fluctuate. The individual random parameter sets create a wide range of model structures at the whole-catchment level, since the eight parallel subcatchments in Fig. 11 jointly comprise a 16-box, 40-parameter model incorporating wide ranges of large and small reservoirs with varying degrees of nonlinearity.

In any spatially heterogeneous catchment (which is to say, any real-world catchment), one will typically only have observations from the merged whole-catchment streamflow (i.e., the blue time series in Fig. 11d). One will typically have no information about the behavior of the individual tributaries (i.e., the colored time series in Fig. 11c), and if one did, then those tributaries would themselves have their own spatially heterogeneous tributary streams or flowpaths, and so on. Thus the heterogeneity of any real-world catchment will remain poorly quantified (and possibly even unrecognized), and rigorously reductionist attempts to fully characterize such complex multiscale heterogeneity would be impractical.

Thus we face the problem: how much can we infer from the behavior of the merged whole-catchment streamflow, given that it originates from processes that are heterogeneous and nonstationary (to a degree that is unknown and unknowable)? Figure 12 explores this general question in the specific context of young water fractions and mean travel times, presenting results from 200 iterations of the heterogeneous nonstationary model shown in Figure 11 with all three precipitation drivers. In Fig. 12 the merged streamflow is separated into discrete flow regimes, following the approach outlined in Sect. 3.4. As Fig. 12 shows, F_{yw} values inferred from the tracer cycles in each discharge range accurately predict the true fraction of young water in that discharge range, as determined from age tracking.

Figure 12 is analogous to Fig. xxf10, with the difference that Fig. 10 shows model runs for individual random parameter sets, whereas Fig. 12 shows results from eight runs merged together. Merging the model outputs will tend to average out the idiosyncrasies of the individual parameter sets, which is why the clusters of points in Fig. 12 are more compact than the corresponding point clouds in Fig. 10. As a result, the individual discharge ranges overlap less in Fig. 12 than in Fig. 10. The compact scatterplots shown in Fig. 12 show only small deviations from the 1:1 line for estimates of the young water fraction F_{yw}. **By contrast, estimates of mean transit times in Fig. 12 exhibit substantial bias and scatter (note the logarithmic axes in Fig. 12d-f).**

Deleted: 14

Deleted: 14d

Deleted: 14c

Deleted: 15

Deleted: 1000

Deleted: 14,

Deleted: ¶
As

Deleted: 15 shows, the amplitude ratios and phase shifts of the seasonal tracer cycles accurately predict the young water fractions F_{yw} in

Deleted: (Fig. 15a-c), but yield strongly biased and highly scattered estimates of mean transit times (Fig. 15d-f; note logarithmic scale). When the model is driven by the strongly seasonal Smith River precipitation forcing (Fig. 15c), inferred values of F_{yw} show a persistent underestimation bias of roughly 5-20 percent. As was also seen in Fig. 10, this bias arises from the correlation between the seasonality in the assumed tracer cycle and the strong seasonality in precipitation, and it would largely disappear if the tracer concentrations peaked in the spring or the fall, instead of the summer. In any event, even under the strongly seasonal Smith River precipitation forcing, the bias in inferred F_{yw} values

Deleted: small compared to the a priori uncertainty in F_{yw} (which is of order 1), and small compared to the bias in inferred MTT (which is large even on logarithmic axes).¶
One can also separate the merged streamflow

Deleted: 5.

Deleted: 16

Deleted: Figures 15 and 16 are

Deleted: s. 10 and 13

Deleted: s.

Deleted: and 13 show

Deleted: s. 15-16 show

Deleted: s. 15-16

Deleted: s

Deleted: and 13

Deleted: 16

Deleted: 13.

Deleted: s. 15-16

3.6 Hydrological and hydrochemical implications of young water fractions

The results reported above, together with the results reported in Paper 1, show that unlike mean transit times, young water fractions can be estimated reliably from seasonal tracer cycles in catchments that are **spatially** heterogeneous, nonstationary, or both. These findings then raise the obvious question: we can measure young water fractions reliably, but what are they good for? One answer is that young water fractions can be considered as a catchment characteristic, analogous (but far from equivalent) to MTT. **In theory** MTT **should** be particularly useful as a catchment descriptor, because the MTT times the mean annual discharge yields the total catchment storage. But **because** estimates of MTT **will** often **be substantially** in **error**, estimates of catchment storage derived from MTT are likely to be equally unreliable. If the shape of the transit time distribution (TTD) were known, of course, there would be a clear functional relationship between MTT and F_{yw}, and one could be calculated from the other. But if the shapes of the TTD were known, estimating the MTT itself would also be easy; the problem in estimating the MTT is the fact that the TTD's shape – particularly the length of its tail – is poorly constrained by tracer data. This is why F_{yw} can be estimated much more reliably than MTT. F_{yw}, like **the** amplitude of the seasonal tracer cycle, depends on the relative proportions of younger and older water, but is insensitive to how old the "older" water is. MTT depends critically on the age of the older water, which cannot be reliably determined because it has almost no effect on the seasonal tracer cycle (or on more elaborate convolution analyses: see Seeger and Weiler, 2014).

Because the young water fraction is indifferent to the age of the older water, it cannot be used to estimate residual storage. What F_{yw} estimates, instead, is the fraction of water reaching the stream by relatively fast **(less than ~2-3 month)** flowpaths. In the context of the present model, this is reflected in the correlation between F_{yw} and the partitioning parameter η (Fig. **B2**). This correlation is not exact, because F_{yw} will depend not only on how much streamflow comes from the upper box, but also on how much of the upper box is young water. That, in turn, will depend on precipitation climatology and the size of the upper box.

One can use F_{yw} not only to make comparisons across catchments, but also, in an individual catchment, to compare how the proportions of flow traveling by fast flowpaths change across different flow regimes, as shown in Figs. **8-10** and **12.** In turn it may be possible to draw inferences about how catchment processes change with flow regime. In this model, variations

Formatted: Bullets and Numbering

Deleted: would

Deleted: if

Deleted: are

Deleted: wrong by large factors (as the analyses above and

Deleted: Paper 1 suggest)

Deleted: 8

Deleted: 11-13

Deleted: 16.

in F_{yw} across different flow regimes are strongly correlated with the fractional contributions of the upper box to streamflow (Fig. 13). The slopes and intercepts of the relationships vary among parameter sets, principally reflecting variations in the partitioning parameter η and the sizes of the upper and lower boxes. The strong correlations shown in Fig. 13 are typical. Repeating the analysis shown in Fig. 13 for 200 random model "catchments" (i.e., different random parameter sets) yields an average correlation of over 0.99 (again, with different linear relationships for different parameter values). Of course these results – and, more generally, the interpretation of F_{yw} in terms of upper-box flow – are model-dependent. They are meant to demonstrate only that process inferences can be drawn from F_{yw}, not that these particular inferences should be applied literally to real-world catchments. Indeed one must remember that in the real world there is no "upper box"; it, like all model abstractions, should not be confused with reality.

The young water fraction F_{yw} may also be helpful in inferring chemical processes from streamflow concentrations of reactive chemical species. Many reactive species exhibit clear concentration-discharge relationships. Because one can determine how F_{yw} varies, on average, across different ranges of discharge (as demonstrated in Figs. 8-10 and 12), one can potentially construct mixing relationships between F_{yw} and the concentrations of reactive species. If the measurable range of F_{yw} is wide enough, one may even be able to estimate the end-member concentrations corresponding to idealized "young water" (F_{yw}=1) and "old water" (F_{yw}=0).

Figure 14 illustrates a preliminary proof of concept for this approach, based on 20-28 years of weekly precipitation and streamflow samples from three catchments at Plynlimon, Wales (Neal et al., 2011) with contrasting geochemical behavior. I separated the streamflow samples into five discharge ranges (lowest 20 percent, next 20 percent, and so on), then fitted the seasonal chloride concentration cycles in each discharge range and calculated the corresponding young water fractions using the approach outlined in Sect. 3.4 above. I then examined the relationships between these young water fractions and the mean streamwater concentrations of reactive chemical species in each discharge range. Figure 14 shows three different views of how reactive tracer chemistry varies with discharge across the three catchments. The left-hand panels show the average concentrations in each discharge range, as functions of the logarithm of discharge. The middle panels show the same concentrations as functions of the inferred F_{yw}, with the vertical axis at F_{yw}=0 indicating the

Deleted: 17

Deleted: 17

Deleted: 17

Deleted: 11-13

Deleted: 16

Deleted: 18

Deleted: . Using

Deleted:), I used seasonal cycles of chloride concentrations to calculate young water fractions for five discharge ranges (lowest 20 percent, next 20 percent, and so on) at three catchments with contrasting (bio)geochemical behavior.

Deleted: 18

hypothetical old water end-member. The right hand panels show the concentrations plotted against the reciprocal of F_{yw}; here, the vertical axis at $1/F_{yw}$=1 indicates the hypothetical young water end-member. The gray lines are fitted by hand to indicate general trends, and to suggest potential end-member concentrations.

The three catchments are characterized by contrasts in soil hydrology, with the abundance of impermeable gley soils and boulder clay tills increasing in the rank order Hafren < Hore < Tanllwyth. The same rank order is observed in the calculated young water fractions at high flows, reflecting the greater high-flow variability in chloride concentrations at sites with more impermeable soils. The three sites also exhibit contrasting concentration-discharge relationships for nitrate and aluminum (Fig. 14a and d), two solutes that are relatively abundant in near-surface soil solutions. When plotted against the young water fraction, however, these catchment-specific concentration-discharge relationships collapse to single concentration-F_{yw} relationships (Fig. 14b and e) in which the three sites are generally indistinguishable within error. These relationships can be extrapolated to reasonably well-constrained old water end-member concentrations of ~0.1 mg L^{-1} NO_3-N and ~50 µg L^{-1} Al, and to comparably well-constrained young water end-member concentrations of ~0.45 mg L^{-1} NO_3-N and ~600 µg L^{-1} Al (Fig. 14c and f). In the case of calcium, the three catchments have markedly different concentration-discharge relationships (Fig. 14g), reflecting differences in the abundance of calcite in their bedrock. As a result, the three catchments have different old water end-member calcium concentrations, ranging from ~1 to ~4 mg L^{-1} (Fig. 14h). However, all three streams converge to similar concentrations of ~0.5 mg L^{-1} Ca in the young water end-member (Fig. 14i).

It is tempting to interpret the concentration differences between the young and old end-members as reflecting chemical kinetics, but this should be approached with caution. A kinetic interpretation makes sense if the young and old end-members differ only in age (albeit by an unspecified amount since we cannot know how old the "old" end-member is), but not if they differ in other respects as well. At Plynlimon, for example, porewaters in the acidic soil layers have relatively high concentrations of aluminum and transition metals, and relatively low concentrations of base cations and silica, whereas waters infiltrating deep into the fractured bedrock react with calcite and layer lattice silicates and thus become enriched in base cations and silica, and depleted in aluminum and transition metals (Neal et al., 1997). Thus one must also consider the alternative hypothesis that the young end-member represents

Deleted: 18a,
Deleted: 18b,
Deleted: /
Deleted: ug/
Deleted: /
Deleted: ug/
Deleted: 18c,
Deleted: 18g
Deleted: /
Deleted: 18h
Deleted: /
Deleted: 18i

mostly soil water, and the old end-member represents mostly deeper groundwater, and that the two end-members exhibit different chemistry because of their sources rather than their ages. In this case, the end-member compositions identified through plots like Fig. 14 may help in characterizing the chemistries, and thus localizing the physical sources, of the young and old waters. In this proof-of-concept example, all three catchments appear to have geochemically similar young water end-members, with a composition suggesting a shallow soil source, but each has a different old water end-member, suggesting deeper groundwater sources with differing amounts of carbonate minerals. This is consistent with independent geochemical evidence at Plynlimon (Neal et al., 1997).

It is also important to note that if the ideal end-member mixing assumptions hold (i.e., the young and old end-members are invariant, and the mixture undergoes no further chemical reactions), then the mixing relationships in the middle plots of Fig. 14 should be straight lines, and they should extrapolate to physically realistic (non-negative) concentrations at both F_{yw}=0 and F_{yw}=1. To the extent that the mixing relationships are not straight, or imply unrealistic end-members, they indicate that these assumptions are not met.

3.7 General observations and caveats

It is important to recognize that the inferred young water fractions F_{yw} plotted in Figs. 7-12 are not in any way calibrated to the true values determined by age tracking. Nor do they make use of any information about the models that transform precipitation into streamflow (neither their structure, nor their parameter values). Thus there is nothing artifactual about the close correspondence between predicted and observed values of F_{yw} in Figs. 7-12. Instead, these thought experiments provide strong evidence that seasonal tracer cycles can be used to reliably partition streamflow into young and old fractions (F_{yw} and 1-F_{yw}, respectively), even in catchments that are both nonstationary and spatially heterogeneous, and whose real-world "models" (i.e., whose underlying processes) are poorly understood.

When these results are applied in practice, however, one must keep in mind that in contrast to typical field studies, these thought experiments are based on synthetic data sets that are dense (daily measurements for 10 years) and error-free. Furthermore, these thought experiments use a sinusoidal precipitation tracer signal that varies only seasonally, with no confounding variation on shorter or longer time scales. Further benchmark testing will be needed to test the accuracy of F_{yw} estimates derived from shorter, sparser, and messier data sets.

Deleted: 18

Deleted: 18

Formatted: Bullets and Numbering

Deleted: 10-16

Deleted: 10-16.

Deleted: where the

Deleted: the

One can of course also question the realism of the particular model that I have used for these thought experiments. This model can be calibrated to reproduce the stream discharge with a Nash-Sutcliffe efficiency of better than 0.85 at two of the three sites, but there is no guarantee that it is getting the right answer for the right reasons. All models – whether lumped conceptual models or "physically based" spatially explicit models – necessarily involve approximations and simplifications. In plain language: any model, including this one, incorporates assumptions that are false and are known to be false. One obvious idealization (a less euphemistic word would be *fiction*) is the well-mixed boxes that form the core of most lumped conceptual models, including the model presented here. Assuming that everything in each box is completely mixed or, equivalently, that it is randomly sampled in the outflow – regardless of where it is physically located in the landscape – clearly strains credibility, but this is what typical conceptual models must assume for mathematical convenience. The model presented here is no different.

What is different, however, is that here the model is used for purposes that make its literal realism unnecessary. Typical modeling studies draw conclusions about real-world systems from model behavior; thus those conclusions depend critically on the realism of the model. But here, the primary goal is not to test how catchments work, but instead to test specific methods for inferring water ages from complex, nonstationary time series **of tracer concentrations**. All the model must do is generate outputs with reasonable degrees of complexity and nonstationarity; it is not essential that the model generates these time series by the same mechanisms that real-world catchments do. The only inductive leap is the inference that if a method correctly infers water ages from tracer patterns in these complex, nonstationary time series, it will also correctly infer water ages in complex, nonstationary time series generated by real-world catchments.

Deleted: tracer concentrations in

It is important to highlight an essential difference between the approach developed here and typical studies that infer water ages or transit time distributions from calibrated models (e.g., Birkel et al., 2011; van der Velde et al., 2012; Heidbüchel et al., 2012; Hrachowitz et al., 2013; Benettin et al., **2015**; Benettin et al., 2013). When one draws inferences from a model, their validity depends on whether **that model is** structurally adequate and whether **its** parameter values are realistic, both of which are usually in doubt. Here, by contrast, I have developed an inferential method (for estimating the young water fraction F_{yw} from seasonal tracer cycles) that is not drawn from – and thus does not depend on – the model's structure or

Deleted: in review

Deleted: those models are

Deleted: their

its parameter values. The model is used only to create synthetic data to test the inferential method.

The results reported here, together with those in Paper 1, show that mean transit times (MTT's) cannot be estimated reliably by fitting sine waves to seasonal tracer cycles from nonstationary or **spatially** heterogeneous catchments. These results do not imply that other methods for estimating MTT's are any better; instead, they imply only that sine wave fitting has been subjected to rigorous benchmark testing and has failed. The other methods have not yet been similarly tested, and it is unclear whether they too will fail. Efforts to fill this knowledge gap are underway. But in the meantime, ignorance is not bliss; one should not simply assume that these other methods work as intended, just because they have not yet been rigorously tested. In that regard, the most general contribution of this analysis is not that it reveals specific problems with MTT estimation from seasonal tracer cycles, or that it demonstrates the reliability of F_{yw} as an alternative metric of catchment transit times, but rather that it illustrates the clarifying power of well-designed benchmark tests.

4 Summary and conclusions

The age of streamflow – i.e., the time that has elapsed since it fell as precipitation – is an essential descriptor of catchment functioning with broad implications for runoff generation, contaminant transport, and biogeochemical cycling (Kirchner et al., 2000; McGuire and McDonnell, 2006). The age of streamflow is commonly measured by its mean transit time (MTT), which in turn has often been estimated from the damping of seasonal cycles of chemical and isotopic tracers (such as Cl^-, $\delta^{18}O$, or **δ^2H**). In a companion paper **("Paper 1": Kirchner, 2015)**, I demonstrated that MTT cannot be reliably estimated from seasonal tracer cycles in spatially heterogeneous catchments, and I proposed an alternative water age metric, the young water fraction F_{yw}, which is relatively immune to the errors and biases that afflict **the** MTT.

Deleted: $\delta^{18}H$

Here I have explored how catchment nonstationarity affects estimates of MTT and F_{yw}, using simple thought experiments based on a simple two-box conceptual model (Fig. 1), driven by three precipitation time series representing a range of precipitation climatologies (Fig. 2). The model exhibits complex nonstationary behavior (Fig. 3), with striking volatility in tracer concentrations, young water fractions, and mean transit times as the mixing ratio between the

upper and lower boxes shifts in response to precipitation events. This mixing ratio is both hysteretic and nonstationary, varying in response both to precipitation forcing and to the antecedent moisture status of the two boxes (Fig. 4).

Marginal (time-averaged) age distributions in drainage are skewed toward younger ages than the storage distributions they come from, because storage is flushed more quickly (and thus is younger) during periods of higher discharge (Fig. 5). The age distributions in whole-catchment storage and discharge are approximate power laws, with markedly different slopes (Fig. 5). **The** age distribution in streamflow becomes increasingly skewed at higher discharges, with a marked increase in the young water fraction and decrease in the mean water age (Fig. 6), reflecting the increased dominance of the upper box at higher flows. Flow-weighted average MTT's are typically close to the steady-state MTT, estimated as the ratio of the total storage to the throughput rate. However, the **marginal age** distributions are markedly different from the distributions that would be expected in steady state**, demonstrating that steady-state approximations are misleading guides to the non-steady-state behavior of the system, *even on average***.

Even this simple two-box model exhibits strong equifinality (Fig. **B1**), with four of its five parameters having virtually no identifiability through hydrograph calibration. However, scaling arguments based on simple perturbation analyses (Sect. 3.2) reveal ratios of parameters that can be constrained through hydrograph calibration (Fig. **B2), greatly reducing the** equifinality in the parameter space. Unfortunately, water age is primarily controlled by residual storage, which cannot be constrained through hydrograph calibration (Fig. **B2**). Thus, parameter sets that yield virtually identical hydrographs imply widely differing young water fractions and mean water ages (Fig. **B3**).

The simple two-box model was used to simulate discharge, **water ages, and the propagation of seasonal tracer cycles through the catchment,** across wide ranges of random parameter sets. **MTT's inferred from** the damping and phase shift of the seasonal tracer cycles exhibited strong underestimation bias and large scatter (Fig. **7**). This result implies that many literature MTT values **(and thus also residual storage volumes)** may have been underestimated by large factors. **By contrast, the seasonal tracer cycles accurately predicted the actual F_{yw} in streamflow, as determined by age tracking within the model (Fig. 7).**

Deleted: not a simple function of discharge; instead it is

Deleted: , and thus to the prior history of precipitation

Deleted: In steady state, non-age-selective drainage (i.e., the well-mixed assumption) would yield an exponential distribution of ages in the upper box, and thus in the short-time

Formatted: English (U.K.)

Deleted: . However, when the system is not in steady state and we aggregate its behavior over time, we are combining different age distributions from different time intervals with different precipitation forcing. Thus these distributions illustrate aggregation errors in the time domain, in the sense that the steady-state approximation will be a misleading guide to the non-steady-state behavior of the system, *even on average*.¶
The transit time distribution in streamflow

Deleted: Residual storage is typically much larger than dynamic storage, so residence times are much longer than timescales of hydraulic response (Kirchner, 2009; Birkel et al., 2011). This contrast between hydraulic response times and mean transit times (or dynamic and total storage, or celerity and velocity) is the simplest explanation for the apparent paradox of prompt discharge of old water during storm events (Kirchner, 2003).

Deleted: of these transit times

Deleted: 7

Deleted: 8). When the parameter space is re-projected onto this different set of coordinate axes, three of the five new parameters are identifiable by calibration. Thus much of the

Deleted: can be eliminated by a simple transformation of variables

Deleted: 8

Deleted: 9

Deleted: tracer concentrations, and water ages

Deleted: , assuming that tracer concentrations in precipitation follow a sinusoidal seasonal cycle. From

Deleted: , I estimated the F_{yw} and MTT in streamflow. The seasonal tracer cycles accurately predicted the actual F_{yw} in streamflow (as determined by age tracking within the model). MTT's inferred from seasonal tracer cycles, by contrast,

Deleted: 10

Deleted: (together with Paper 1)

Deleted: , and thus that catchment residual storage volumes may have also been underestimated by similarly large factor

Flow-weighted fits to the seasonal tracer cycles accurately predicted the flow-weighted average F_{yw} in streamflow, while unweighted fits to the seasonal tracer cycles accurately predicted the unweighted average F_{yw}. The streamflow time series can be separated into distinct flow regimes with their own seasonal tracer cycles (Fig. **8),** which **accurately reflect** the F_{yw} in each flow regime **(Figs. 9** and **10**). Seasonal tracer cycles **also** accurately predicted the F_{yw} in the merged streamflow from spatially heterogeneous **assemblages of** nonstationary model catchments (Fig. **12**). Importantly, all of these F_{yw} predictions were really predictions; they were not calibrated in any way.

The relationship between F_{yw} and the flow regime reflects how the fluxes from short-term storages vary with hydrologic forcing (Fig. 13). In a preliminary proof of concept (Fig. 14), I showed that one can construct mixing relationships between solute concentrations and F_{yw}'s for discrete flow regimes. From these mixing relationships one can estimate the chemical composition of idealized "young water" and "old water" end-members (Fig. 14).

These findings extend the results of Paper 1 by showing that estimates of MTT from seasonal tracer cycles are unreliable under nonstationarity as well as spatial heterogeneity. These findings also extend the results of Paper 1 by showing that F_{yw} can be reliably estimated in nonstationary catchments as well as spatially heterogeneous ones, and can also be reliably estimated for discrete flow regimes. These results further demonstrate that F_{yw} can **be reliably estimated for discrete flow regimes, and can** provide helpful insights into the hydrological and hydrochemical functioning of catchments. Most generally, these results, along with those of Paper 1, illustrate how well-posed benchmark tests can be essential in clarifying what is knowable – and, conversely, unknowable – in environmental research.

Appendix A: Solution scheme

For simplicity and efficiency, the hydrological model is solved on a fixed daily time step. This requires some care with the numerics, given the clear (though often overlooked) dangers in naive forward-stepping simulations of nonlinear equations (Clark and Kavetski, 2010; Kavetski and Clark, 2010, 2011). Here I use a weighted combination of the trapezoidal method (which is partly implicit, for enhanced accuracy) and the backward Euler method (which is fully implicit, for guaranteed stability). The

Deleted: 11), from

Deleted: can be estimated (Fig. 12). These regime-specific seasonal tracer cycles accurately predicted the actual young water fractions in each flow regime, as determined by age tracking (Fig. 13). The variation in F_{yw} with flow regime can give insight into how the fluxes from short-term storages vary with hydrologic forcing (Fig. 17). In a preliminary proof of concept (Fig. 18), I showed that one can construct mixing relationships between solute concentrations

Deleted: F_{yw}'s for discrete flow regimes. From these mixing relationships one can estimate the chemical composition of idealized "young water" and "old water" end-members (Fig. 18). ¶
To study the behavior of synthetic catchments that are both nonstationary and spatially heterogeneous, I combined the streamflow generated by multiple copies of the simple two-box model, each representing a different tributary subcatchment with its own random parameter set (Fig. 14

Deleted: these

Deleted: ,

Deleted: , both for individual flow regimes

Deleted: 16) and for all flows combined (Fig. 15

hydrological solution scheme is illustrated here for the upper box; the lower box is handled analogously. The storage in the upper box is updated using the following equation:

$$S_u(t_{i+1}) - S_u(t_i) = \Delta t \left(P - \rho k_u S_u(t_{i+1})^{b_u} - (1-\rho) k_u S_u(t_i)^{b_u} \right) \quad , \qquad \text{(A1)}$$

where $S_u(t_i)$ is the storage in the upper box at the beginning of the i^{th} time interval (with length Δt), $S_u(t_{i+1})$ is the storage at the end of that interval (and thus the beginning of the next), and P is the average precipitation rate over the interval. Equation (A1) is implicit and nonlinear; there is no closed-form solution for the future storage $S_u(t_{i+1})$, which instead is found using Newton's method. The relative dominance of the trapezoidal and backward Euler solutions is determined by the weighting factor ρ, which takes on values between ρ=0.5 (trapezoidal method) and ρ=1 (backward Euler method). The value of ρ in Eq. (A1) is determined for each time step using the simple stability criterion,

$$\rho = \min\left(0.5 + 0.5 \frac{\left(P - k_u S_u(t_i)^{b_u}\right)\Delta t}{\left(P/k_u\right)^{1/b_u} - S_u(t_i)} \ , \ 1 \right) \quad , \qquad \text{(A2)}$$

where the numerator represents the amount that S_u would change during one time step if the instantaneous drainage rate L in Eq. (1) were projected forward in time, and the denominator represents the difference between S_u's current value and its equilibrium value at the precipitation rate P. Equation (A2) says that if the trapezoidal method would move S_u by only a small fraction of the distance to its equilibrium value (at the precipitation rate P), then the stability advantages of the backward Euler method are unnecessary and the more accurate trapezoidal method should dominate the solution instead ($\rho \approx 0.5$). On the other hand, if the trapezoidal method would overshoot the equilibrium value, then ρ=1 and the fully implicit backward Euler method is used to solve Eq. (A1). The closer the trapezoidal method would come to overshooting the equilibrium, the larger the value of ρ and the greater the weight that is given to the backward Euler solution. The guaranteed stability of the backward Euler method is important when b_u or b_l is large, because the underlying equations can become quite stiff. After the final value of S_u is determined by Eq. (A1), the drainage from S_u between t_i and t_{i+1} is determined by mass balance:

$$L = P + \left(S_u(t_i) - S_u(t_{i+1})\right)/\Delta t \quad , \qquad \text{(A3)}$$

where L is the average drainage rate over the interval Δt between t_i and t_{i+1}.

The tracer concentrations are determined under the assumption that each box is well mixed, implying that individual water parcels within each box do not need to be tracked, and also that the concentration draining from each box equals the average concentration within the box. I make the simplifying assumption that each box's inflow and outflow rates (and also inflow concentrations) are constant over each day. Again taking the upper box as an example, these assumptions imply that starting from $t=t_i$ the tracer concentration will evolve as

$$\frac{dC_u}{dt} = \frac{P(C_P - C_u)}{S_u(t_i) + (P - L)(t - t_i)} \quad , \qquad \text{(A4)}$$

where C_P and C_u are the concentrations in precipitation and the upper box, respectively, and the denominator expresses how the volume in the box changes with time from its initial value of $S_u(t_i)$. Integrating Eq. (A4) over an interval Δt yields the concentration updating formula:

$$C_u(t_{i+1}) = C_P + \left(C_u(t_i) - C_P\right)\left(\frac{S_u(t_i)}{S_u(t_{i+1})}\right)^{\left(P/(P-L)\right)} \quad , \qquad \text{(A5)}$$

where any quantities that are not shown as functions of time are constant at their average values over the interval. Equation (A5) could potentially become difficult to compute when P and L are nearly equal (differing by, say, less than 1 part in 1000), and the power function approaches its exponential limit. In such cases the change in volume in Eq. (A4) becomes trivially small, and one can replace Eq. (A5) with the more familiar exponential formula for a well-mixed box of constant volume:

$$C_u(t_{i+1}) = C_P + \left(C_u(t_i) - C_P\right)\exp\left(-P\,\Delta t / S_u\right) \quad . \qquad \text{(A6)}$$

After the tracer concentrations are updated, the average concentrations in drainage are calculated by mass balance, as follows:

$$C_L = \left[C_P(t_i)\,P + C_u(t_i)\,S_u(t_i) - C_u(t_{i+1})\,S_u(t_{i+1})\right]/L \quad , \qquad \text{(A7)}$$

where C_L is the average concentration in drainage over the time interval between t_i and t_{i+1}.

The mean age within each box is modeled analogously to the tracer concentrations, following the "age mass" concept widely used in groundwater hydrology. Here I will illustrate the approach using the example of the lower box, since it is the more complex case (for the upper box, the input age in precipitation is zero, but this is not true for the upper-box drainage that recharges the lower box). Assuming that the inflow and outflow rates $L(1-\eta)$ and Q_l are constant over a day, as is the average age $\bar{\tau}_L$ of the inflow from the upper box, the mean age in the lower box should evolve according to

$$\frac{d\bar{\tau}_l}{dt} = \frac{L(1-\eta)(\bar{\tau}_L - \bar{\tau}_l)}{S_l(t_i) + (L(1-\eta) - Q_l)(t - t_i)} + 1 \quad , \qquad \text{(A8)}$$

which is directly analogous to Eq. (A4), except for additional term of +1, which accounts for the continual aging of the water in the box. The solution to Eq. (A8) is

$$\bar{\tau}_l(t_{i+1}) = \bar{\tau}_L + \frac{S_l(t_{i+1})}{2L(1-\eta) - Q_l} + \left(\bar{\tau}_l(t_i) - \bar{\tau}_L - \frac{S_l(t_i)}{2L(1-\eta) - Q_l} \right) \left(\frac{S_l(t_i)}{S_l(t_{i+1})} \right)^{\left(\frac{L(1-\eta)}{L(1-\eta) - Q_l} \right)} \quad , \qquad \text{(A9)}$$

where $\bar{\tau}_l(t_i)$ and $\bar{\tau}_l(t_{i+1})$ are the mean age of the water in the lower box at the beginning and end of the time interval. Analogously to tracer concentrations, one can calculate the mean age of the drainage from the box based on the inputs and the change in mean age inside the box, using conservation of "age mass":

$$\bar{\tau}_{Q_l} = \left[\bar{\tau}_L(t_i)\,(1-\eta) + \bar{\tau}_l(t_i)\,S_l(t_i) - \left(\bar{\tau}_l(t_{i+1}) - \Delta t \right) S_l(t_{i+1}) \right] / Q_l \quad , \qquad \text{(A10)}$$

where the factor of $-\Delta t$ accounts for the aging of the contents of the box.

The approach used here for concentrations and water ages requires the assumption that input fluxes to each box are constant within each time interval (but constant at their average values, not their initial values). This is a reasonable approximation, particularly when we have no sub-daily precipitation data. And in exchange for this simplifying assumption, equations (A5), (A6), and (A9) provide something important, namely, the exact analytical solution for the evolution of concentration and age during each time interval. Thus these equations directly solve for the correct result even if, for example, an individual day's rainfall is much greater than the total volume of the upper box. The equations above will correctly calculate the consequences of the (potentially many-fold) flushing that occurs in such cases. The approach outlined above also

guarantees exact consistency between stocks and fluxes (but note, not in the usual way by updating stocks with fluxes, but rather by calculating output fluxes from inputs and changes in stocks). Readers should keep in mind that all stocks and properties of stocks (i.e., storage volumes, concentrations, and ages) are expressed as the instantaneous values at the beginning of each time interval, and that fluxes and properties of fluxes (i.e., water fluxes and their concentrations and ages) are expressed as averages over each time interval. Otherwise it could be difficult to make sense of the equations above.

Appendix B: Equifinality in hydraulic behavior and divergence in travel times

The analysis outlined in Sect. 3.2 implies that approximate equifinality is inevitable, even in such a simple model, because variations in the exponents b_u and b_l and the reference storage levels $S_{u,ref}$ and $S_{l,ref}$ will have nearly offsetting effects on the model's runoff response. Equations (10) and (11) show that, for a given average precipitation forcing, any parameter values for which the partitioning coefficient η and the ratios $S_{u,ref}/b_u$ and $S_{l,ref}/[(1-\eta)b_l]$ are invariant would give nearly equivalent hydrograph predictions, because the hydraulic response timescales of the upper and lower boxes, and their relative contributions to discharge, would be invariant. These conditions can be achieved for widely varying values of the individual parameters b_u, b_l, $S_{u,ref}$, and $S_{l,ref}$.

This equifinality problem can be readily visualized by plots like Fig. B1. To generate Fig. B1, I ran the model with Smith River precipitation forcing and the reference parameter set (shown by the red squares in Fig. B1), and used the resulting daily hydrograph (after the spin-up period) as virtual "ground truth" for model calibration. I then ran the model with 1000 random parameter sets, and used the Nash-Sutcliffe efficiency (NSE) of the logarithms of discharge to measure how well their hydrographs matched the reference hydrograph (thus the reference hydrograph has NSE=1 by definition). The 50 best-fitting parameter sets, all with NSE≥0.98, are shown as dark blue points in Fig. B1. The bottom row of scatterplots shows the conventional "dotty plots". Their flat tops are the hallmark of equifinality, i.e., wide ranges of parameter values give equally good hydrograph predictions (Beven, 2006). Only the partition coefficient η, which performs well across half its range, can be even modestly

constrained by calibration. (The other precipitation drivers yield results similar to those shown in Fig. B1.)

The other panels of the scatterplot matrix also give important clues to the origins of the observed equifinality. In particular, the best-fitting parameter sets show strong correlations between $S_{u,ref}$ and b_u, and between $S_{l,ref}$ and b_l, as expected from the perturbation analysis presented in Sect. 3.2. Thus good model performance can be obtained across almost the entire range of these parameters, but only for specific parameter combinations. These parameter combinations correspond to "valleys" in the model's response surface, a longstanding problem in model calibration (e.g., Ibbitt and O'Donnell, 1974). The interdependence of the parameters is visually obvious in the scatterplot matrix, but is invisible in the conventional "dotty plots".

This information can be exploited to design parameter spaces that are more identifiable through calibration (e.g., Ibbitt and O'Donnell, 1974). An ideal parameter space would be one in which 1) all parameters are highly identifiable, meaning the goodness-of-fit surface is strongly curved along each parameter axis, and 2) in the best-fitting parameter sets, no parameters are strongly correlated with one another. The second of these criteria is necessary (although not sufficient) for the first, as Fig. B1 illustrates. A third criterion is that all parameters that are needed for simulating any quantities of interest must be determined somehow within the parameter space, either individually or through combinations of other parameters. Thus, for example, although the volumes of the boxes ($S_{u,ref}$ and $S_{l,ref}$) are strongly correlated with their exponents (b_u and b_l), the parameter space must allow them to be individually determined, because as Eqs. (12-14) suggest, the mean transit times will be controlled primarily by the volumes alone (not in combination with the exponents), whereas the runoff response will be controlled primarily by the ratios of volumes to exponents (Eqs. 10-11). These criteria, plus some trial and error, lead to a more identifiable parameter space, whose five axes are $S_{u,ref}$, $S_{l,ref}$, $S_{u,ref}/(\eta \cdot b_u)$, $S_{l,ref}/b_l$, and η.

Figure B2 shows that this parameter space exhibits much less equifinality than the parameter space shown in Fig. B1, although the underlying parameter sets and model simulations are exactly the same. All that has been done is to re-project the parameter space onto a different set of coordinate axes in which the curvature of the goodness-of-fit

surface is more clearly visible. Thus, much of the apparent equifinality in the parameter space has been eliminated by simple transformations of variables. These transformations can be designed by eye in this case, because the dimensionality of the original parameter space is low. In higher-dimension parameter spaces, multivariate techniques such as factor analysis may be helpful. Nonetheless, given the obvious utility of this simple correlation analysis and the perturbation analysis of Sect. 3.2, it is surprising that they are not more widely used in hydrological modeling.

Despite the improved identifiability of the parameter space, however, it is still not possible to constrain the mean transit time by calibration to the hydrograph. As the bottom row of scatterplots in Figure B2 shows, the mean transit time (MTT) is almost entirely determined by the lower box's reference volume $S_{l,ref}$, as one would expect from Eq. 14. However, as predicted by the perturbation analysis in Sect. 3.2, and as shown by Fig. B2, the runoff response of the model system is essentially independent of $S_{l,ref}$ and therefore cannot be used to constrain it. The runoff response does depend on the ratio of $S_{l,ref}$ to b_l, and thus can be used to constrain that ratio, but it cannot constrain $S_{l,ref}$ by itself, and thus it cannot constrain the MTT. For the young water fraction F_{yw} the outlook is not quite as bleak, because F_{yw} is correlated with the partition coefficient η, which can be constrained somewhat by calibration. As a result, it appears that F_{yw} could potentially be constrained within roughly 1/3 of its full range by parameter calibration to the hydrograph.

Figure B3 provides a different visualization of the same equifinality problem. Figure B3 shows a two-year excerpt from the simulated time series of streamflows, tracer concentrations, young water fractions, and mean transit times for the reference parameter set (the blue curves), along with the 50 parameter sets that gave the best fit to the reference hydrograph (the gray curves). Because these 50 parameter sets were those that matched the reference hydrograph best, it is unsurprising that the 50 gray hydrographs generally follow the blue reference hydrograph in Fig. B3a. The 50 gray tracer concentration time series also follow the blue reference time series (Fig. B3b), but with somewhat greater variability than the hydrographs, indicating that the parameter values affect the chemographs and the hydrographs in somewhat different ways. But the most striking feature of Fig. B3 is the much greater variability among the young water fractions F_{yw} and (especially) the mean transit times MTT for these same

parameter sets (Fig. B3c-d). Although all the parameter sets fit the reference hydrograph nearly perfectly, they vary over a range of 0.3 in F_{yw} (out of a total possible range of 1.0), and over a factor of 9.5 in MTT, on average for the whole time period. Thus these time series demonstrate, consistent with Fig. B2, that there are wide ranges of variability in F_{yw} and especially MTT that cannot be constrained by calibration to the hydrograph.

Acknowledgements

I thank Scott Jasechko and Jeff McDonnell for the intensive discussions that motivated this analysis, and **Markus Weiler and an anonymous reviewer for their comments.** I thank the Centre for Ecology and Hydrology for making the Plynlimon data available.

References

Benettin, P., van der Velde, Y., van der Zee, S., Rinaldo, A., and Botter, G.: Chloride circulation in a lowland catchment and the formulation of transport by travel time distributions, Water Resour. Res., 49, 4619-4632, doi: 10.1002/wrcr.20309, 2013.

Benettin, P., Kirchner, J., Rinaldo, A., and Botter, G.: Modeling chloride transport using travel-time distributions at Plynlimon, Wales, Water Resour. Res., **51, 3259-3276, doi: 10.1002/2014WR016600, 2015**.

Bethke, C. M., and Johnson, T. M.: Groundwater age and groundwater age dating, Annual Review of Earth and Planetary Sciences, 36, 121-152, doi: 10.1146/annurev.earth.36.031207.124210, 2008.

Beven, K.: On subsurface stormflow: predictions with simple kinematic theory for saturated and unsaturated flows, Water Resour. Res., 18, 1627-1633, 1982.

Beven, K.: A manifesto for the equifinality thesis, J. Hydrol., 320, 18-36, doi: 10.1016/j.jhydrol.2005.07.007, 2006.

Birkel, C., Soulsby, C., and Tetzlaff, D.: Modelling catchment-scale water storage dynamics: reconciling dynamic storage with tracer-inferred passive storage, Hydrological Processes, 25, 3924-3936, 2011.

Birkel, C., Soulsby, C., Tetzlaff, D., Dunn, S., and Spezia, L.: High-frequency storm event isotope sampling reveals time-variant transit time distributions and influence of diurnal cycles, Hydrological Processes, 26, 308-316, doi: 10.1002/hyp.8210, 2012.

Botter, G., Bertuzzo, E., and Rinaldo, A.: Transport in the hydrological response: Travel time distributions, soil moisture dynamics, and the old water paradox, Water Resour. Res., 46, W03514, doi: 10.1029/2009WR008371, 2010.

Botter, G., Bertuzzo, E., and Rinaldo, A.: Catchment residence and travel time distributions: The master equation, Geophys. Res. Lett., 38, L11403, doi: 10.1029/2011GL047666, 2011.

Botter, G.: Catchment mixing processes and travel time distributions, Water Resour. Res., 48, 15, W05545, doi: 10.1029/2011wr011160, 2012.

Deleted: in review

Clark, M. P., and Kavetski, D.: Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes, Water Resour. Res., 46, W10510, doi: 10.1029/2009wr008894, 2010.

Feng, X. H., Faiia, A. M., and Posmentier, E. S.: Seasonality of isotopes in precipitation: A global perspective, Journal of Geophysical Research-Atmospheres, 114, D08116, doi: 10.1029/2008jd011279, 2009.

Harman, C. J.: Time-variable transit time distributions and transport: Theory and application to storage-dependent transport of chloride in a watershed, Water Resour. Res., 51, 1-30, doi: 10.1002/2014WR015707, 2015.

Heidbüchel, I., Troch, P. A., Lyon, S. W., and Weiler, M.: The master transit time distribution of variable flow systems, Water Resour. Res., 48, W06520, doi: 10.1029/2011WR011293, 2012.

Hrachowitz, M., Soulsby, C., Tetzlaff, D., Malcolm, I. A., and Schoups, G.: Gamma distribution models for transit time estimation in catchments: Physical interpretation of parameters and implications for time-variant transit time assessment, Water Resour. Res., 46, W10536, doi: 10.1029/2010wr009148, 2010.

Hrachowitz, M., Savenije, H., Bogaard, T. A., Tetzlaff, D., and Soulsby, C.: What can flux tracking teach us about water age distribution patterns and their temporal dynamics?, Hydrol. Earth Syst. Sci., 17, 533-564, doi: 10.5194/hess-17-533-2013, 2013.

Ibbitt, R. P., and O'Donnell, T.: Designing conceptual catchment models for automatic fitting methods, in: Mathematical Models in Hydrology, International Association of Hydrological Sciences Publication, 101, Wallingford, U.K., 461-475, 1974.

Kavetski, D., and Clark, M. P.: Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction, Water Resour. Res., 46, W10511, doi: 10.1029/2009wr008896, 2010.

Kavetski, D., and Clark, M. P.: Numerical troubles in conceptual hydrology: Approximations, absurdities and impact on hypothesis testing, Hydrological Processes, 25, 661-670, doi: 10.1002/hyp.7899, 2011.

Kirchner, J. W., Feng, X., and Neal, C.: Fractal stream chemistry and its implications for contaminant transport in catchments, Nature, 403, 524-527, 2000.

Kirchner, J. W., Feng, X., and Neal, C.: Catchment-scale advection and dispersion as a mechanism for fractal scaling in stream tracer concentrations, J. Hydrol., 254, 81-100, 2001.

Kirchner, J. W.: A double paradox in catchment hydrology and geochemistry, Hydrological Processes, 17, 871-874, 2003.

Kirchner, J. W.: Catchments as simple dynamical systems: catchment characterization, rainfall-runoff modeling, and doing hydrology backward, Water Resour. Res., 45, W02429, doi:02410.01029/02008WR006912, 2009.

Kirchner, J. W.: Aggregation in environmental systems: Seasonal tracer cycles quantify young water fractions, but not mean transit times, in spatially heterogeneous catchments, Hydrol. Earth Syst. Sci., submitted manuscript, 2015.

Kreft, A., and Zuber, A.: On the physical meaning of the dispersion equation and its solutions for different initial and boundary conditions, Chem. Eng. Sci., 33, 1471-1480, 1978.

McDonnell, J. J., and Beven, K.: Debates-The future of hydrological sciences: A (common) path forward? A call to action aimed at understanding velocities, celerities and residence time distributions of the headwater hydrograph, Water Resour. Res., 50, 5342-5350, doi: 10.1002/2013wr015141, 2014.

McGuire, K. J., and McDonnell, J. J.: A review and evaluation of catchment transit time modeling, J. Hydrol., 330, 543-563, 2006.

Neal, C., Wilkinson, J., Neal, M., Harrow, M., Wickham, H., Hill, L., and Morfitt, C.: The hydrochemistry of the River Severn, Plynlimon, Hydrol. Earth Syst. Sci., 1, 583-617, 1997.

Neal, C., Reynolds, B., Norris, D., Kirchner, J. W., Neal, M., Rowland, P., Wickham, H., Harman, S., Armstrong, L., Sleep, D., Lawlor, A., Woods, C., Williams, B., Fry, M., Newton, G., and Wright, D.: Three decades of water quality measurements from the Upper Severn experimental catchments at Plynlimon, Wales: an openly accessible data resource for research, modelling, environmental management and education, Hydrological Processes, 25, 3818-3830, doi: DOI: 10.1002/hyp.8191, 2011.

Peters, N. E., Burns, D. A., and Aulenbach, B. T.: Evaluation of high-frequency mean streamwater transit-time estimates using groundwater age and dissolved silica concentrations in a small forested watershed, Aquatic Geochemistry, 20, 183-202, 2014.

Deleted: 1

Deleted: the

Seeger, S., and Weiler, M.: Reevaluation of transit time distributions, mean transit times and their relation to catchment topography, Hydrol. Earth Syst. Sci., 18, 4751-4771, doi: 10.5194/hess-18-4751-2014, 2014.

Tetzlaff, D., Malcolm, I. A., and Soulsby, C.: Influence of forestry, environmental change and climatic variability on the hydrology, hydrochemistry and residence times of upland catchments, J. Hydrol., 346, 93-111, 2007.

Van der Velde, Y., De Rooij, G. H., Rozemeijer, J. C., van Geer, F. C., and Broers, H. P.: The nitrate response of a lowland catchment: on the relation between stream concentration and travel time distribution dynamics, Water Resour. Res., 46, W11534, doi: doi:10.1029/2010WR009105, 2010.

van der Velde, Y., Torfs, P. J. J. F., van der Zee, S. E. A. T. M., and Uijlenhoet, R.: Quantifying catchment-scale mixing and its effect on time-varying travel time distributions, Water Resour. Res., 48, W06536, doi: doi:10.1029/2011WR011310, 2012.

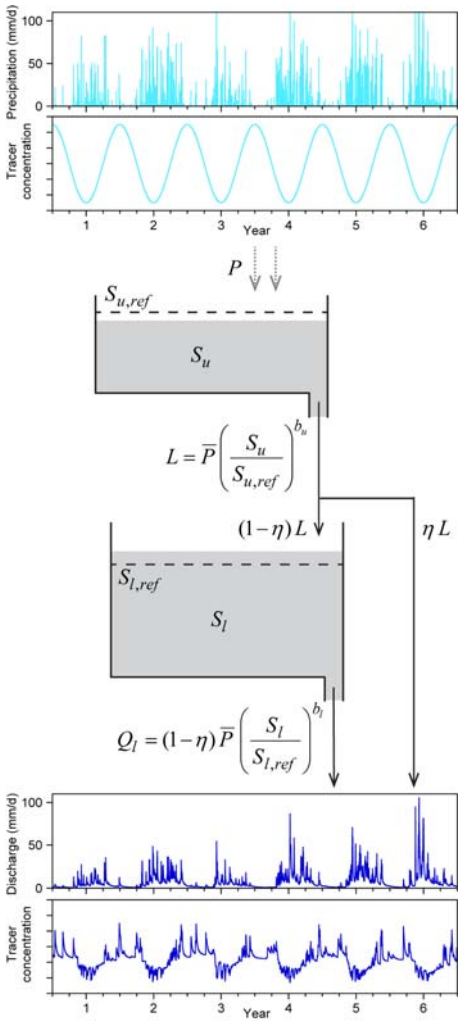

Figure 1. Schematic diagram of conceptual model. Drainage from the upper and lower boxes is determined by power functions of the storage volumes S_u and S_l (depicted by gray shaded regions) as ratios of the reference storage levels $S_{u,ref}$ and $S_{l,\,ref}$ (depicted by dashed lines). The partition coefficient splits the upper box drainage L into direct discharge and infiltration to the lower box.

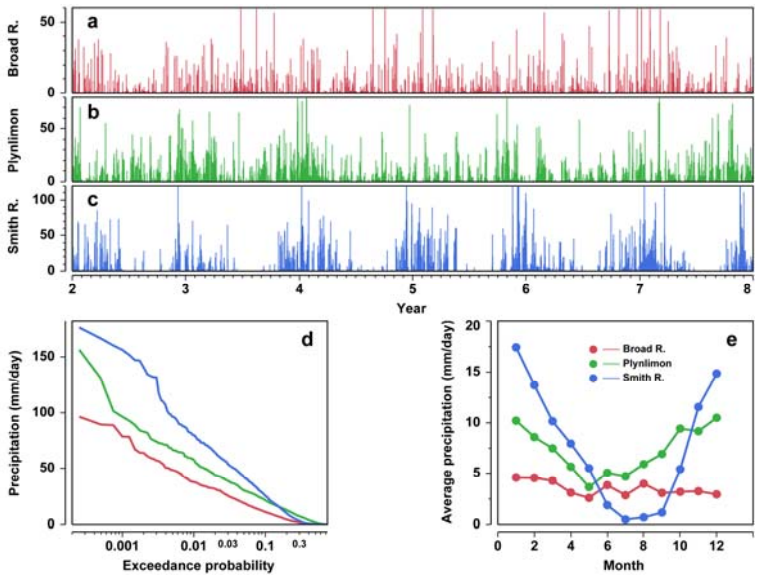

Figure 2. Excerpts of daily precipitation records used to drive the model: (a) Broad River, Georgia, USA **(humid temperate climate; Köppen climate zone Cfa)** in red, (b) Plynlimon, Wales **(humid maritime climate; Köppen climate zone Cfb)** in green, and (c) Smith River, California, USA **(Mediterranean climate; Köppen climate zone Csb)** in blue. Axes are expanded to make typical storms visible; thus the largest storms, some of which extend to roughly twice the axis limits, are cut off. Exceedance probability plot (d) shows a steeper magnitude-frequency relationship for Smith River than for the other two records. Monthly precipitation averages (e) show clear differences in seasonality among the three sites.

Deleted: (

Deleted:),

Deleted: (

Deleted:),

Deleted: (

Deleted:),

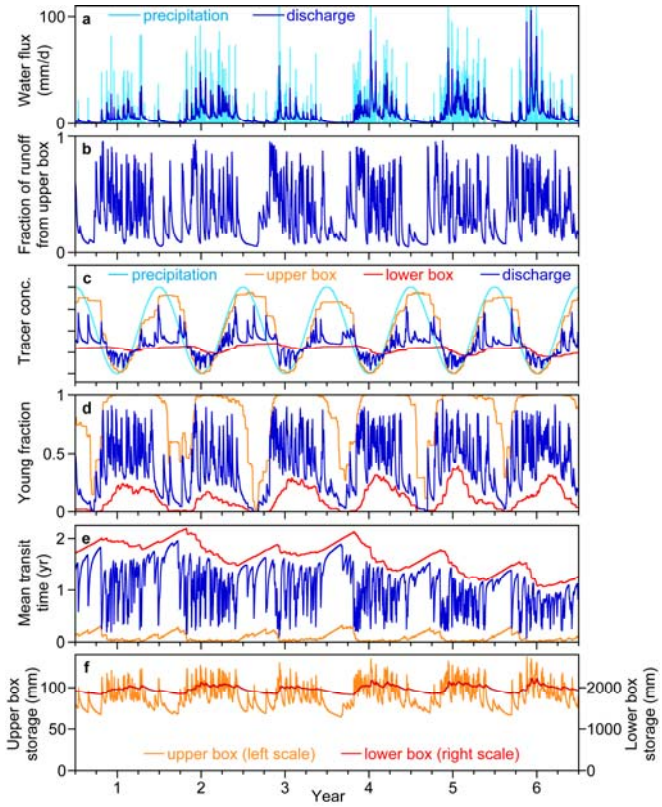

Figure 3. Illustrative time series from the two-box model, using the reference parameter set and the Smith River **(Mediterranean climate)** precipitation time series. Responses to precipitation events (a) entail rapid shifts in the proportions of discharge coming from the upper and lower boxes (b). The smaller, upper box, shown in orange, has a larger young water fraction (d) and a younger mean age (e) than the larger, lower box, shown in red, and thus its tracer concentration (c) is less lagged and damped relative to the hypothetical precipitation concentration, shown by the cosine wave in (c). Mean ages increase (e) and young water fractions decrease (d), in both boxes, throughout the dry summer periods. The proportions of streamflow originating from the upper and lower boxes shift dramatically in response to transient precipitation inputs; thus the tracer concentrations, young water fractions, and mean ages in discharge (dark blue, panels c-e) vary widely between the time-varying end members represented by the upper and lower boxes. Storage volumes fluctuate in a relatively narrow range (f) while discharge varies by orders of magnitude, because the

drainage rates from both boxes are strongly nonlinear functions of storage. Thus both boxes have sizeable residual storage, which is not drained even under extreme low-flow conditions.

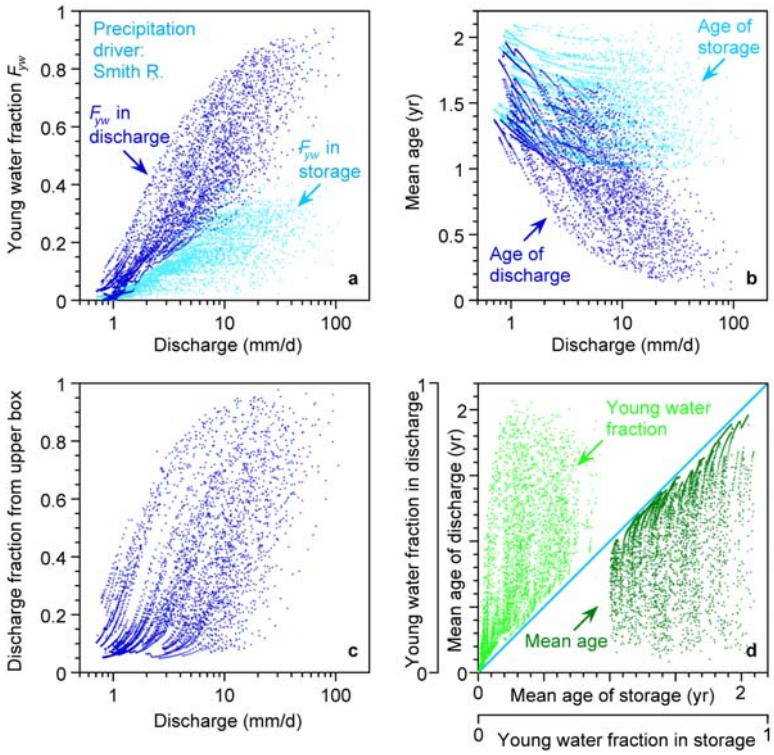

Figure 4. Daily values of young water fractions F_{yw} (a) and mean water ages (b) in storage (light blue) and discharge (dark blue) in the two-box model with reference parameter values and Smith River **(Mediterranean climate)** precipitation. The young water fraction and mean age are both highly scattered functions of discharge (a, b), as is the fractional contribution **from** the upper box to streamflow (c), reflecting the effects of variations in antecedent rainfall. The average age and F_{yw} of water in discharge are strongly biased, and highly scattered, measures of the same quantities in storage (d).

Deleted: a

Deleted: "

Deleted: "

Deleted: They are biased because most of the storage is in the lower box (which derives its contents from the upper box and thus is systematically older), and they are highly scattered because the fractional contributions to streamflow from the (younger, smaller) upper box and the (older, larger) lower box are highly variable.

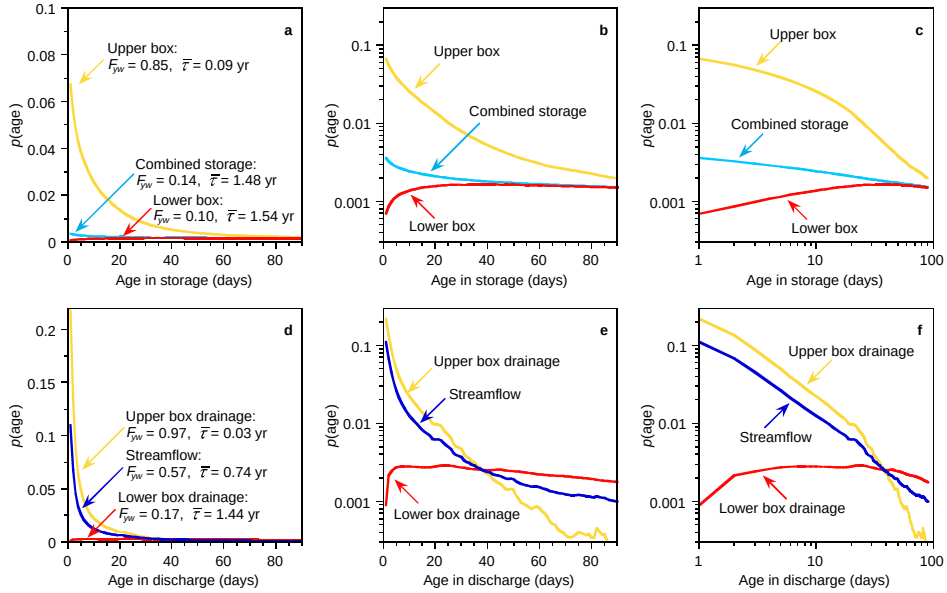

Figure 5. Marginal (time-averaged) age distributions in storage (a-c) and drainage (d-f) in the reference case simulation (Fig.3), shown on linear (a, d), log-linear (b, e), and double-log (c, f) axes. Distributions in drainage (lower plots) are skewed toward younger ages than the storage distributions that they come from (upper plots). This arises, even though drainage is not age-selective, because storage is flushed more quickly (and thus is younger) during periods of higher discharge. Age distributions in the upper box, combined storage, and streamflow are more skewed than exponentials (i.e., they are upward-curving in the middle plots). The age distributions in the combined storage and streamflow (blue lines) are approximate power laws; i.e., they are nearly straight in the right-hand plots, with markedly different power-law slopes. The light blue line in the upper plots shows the age distribution of the combined upper and lower boxes, which resembles the age distribution of the lower box because the reference parameter values **imply that** the lower box comprises about 95 percent of total storage. However, direct drainage from the upper box comprises 50 percent of streamflow; thus the streamflow age distribution (shown by the dark blue line in lower plots) reflects the strong skew of the upper box age distribution. Although both boxes are well mixed and have nearly constant volumes, the age distribution of discharge clearly differs from the distribution that would be expected in steady state, which would be exponential in short time.

Deleted: shown in

Deleted: (assuming

Deleted:)

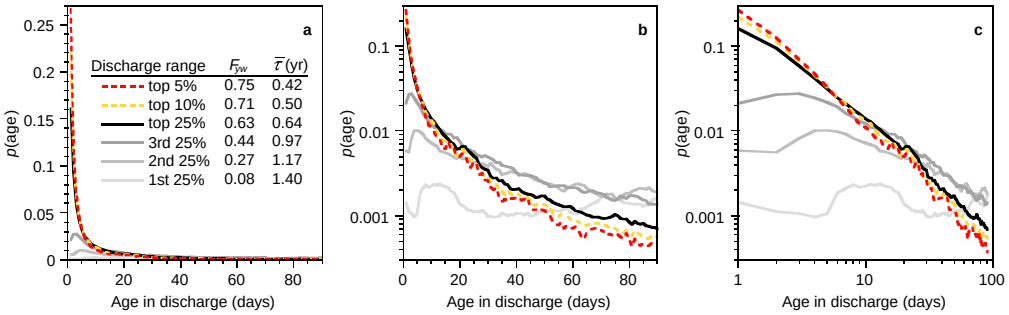

Figure 6: Marginal (time-averaged) transit-time distributions (TTD's) for selected ranges of daily discharges in the two-box model, with the reference parameter set and Smith River **(Mediterranean climate)** precipitation forcing, on linear (a), log-linear (b), and double-log (c) axes. The TTD becomes increasingly skewed at higher discharges (a), with a marked increase in the young water fraction F_{yw} and decrease in the mean water age $\bar{\tau}$. For the upper half of all discharges, the age distribution is upward-curving on log-linear axes (b), implying that it is more skewed than exponential. Discharges in the top 25% and above have approximately power-law age distributions, plotting as nearly straight lines on double-log axes (c).

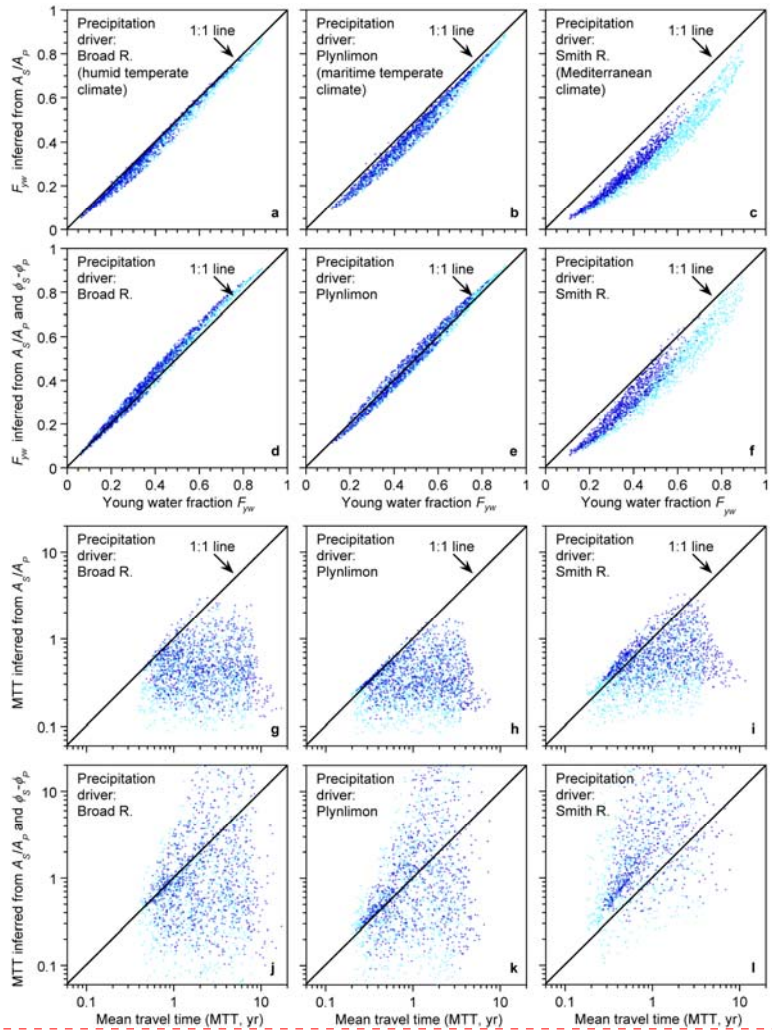

Figure 7. Young water fractions (F_{yw}, **left** panels) and mean transit times (MTT, **right** panels – note log scale) in streamflow from the two-box model. Upper panels compare **the average** F_{yw} in discharge, determined by age tracking within the model (on the horizontal axes) with the seasonal tracer cycle amplitude ratio A_S/A_P (panels a-c), and with F_{yw} inferred from the tracer cycle amplitude ratio A_S/A_P and phase shift ϕ_S-ϕ_P (panels d-f). Lower panels compare **the** average MTT in discharge (again from age tracking) with MTT inferred from the tracer amplitude ratio (panels g-i) and from amplitude ratio and phase shift (panels j-l). Light blue points show flow-weighted average F_{yw}'s and MTT's for each simulation, compared to estimates from flow-weighted fits to seasonal tracer cycles. Dark blue points show un-

Formatted: English (U.S.)

Deleted: Equifinality in discharge predictions. The scatterplot matrix shows relationships among 1000 random parameter sets and the Nash-Sutcliffe efficiency (NSE) of discharge time series driven by Smith River precipitation forcing. The red square indicates the "reference" parameter set that was used to generate the discharge time series that the other parameter sets were tested against; these reference parameters thus correspond to NSE=1.00 by definition. The dark blue dots show the best-fitting 50 (or 5%) of the parameter sets, all with NSE≥0.98. Excellent discharge predictions can be obtained across almost the full range of all five model parameters, except the partition coefficient η, which performs well across only about half its range. The dark blue dots show clear correlations between the reference storage levels in each box ($S_{u,ref}$, $S_{l,ref}$) and the corresponding drainage function exponents (b_u, b_l); these correlations delimit regions with nearly constant hydraulic response time scales, as defined by Eqs. (20)-(21). ¶

… [5]

Deleted: top

Deleted: bottom

weighted average F_{yw}'s and MTT's, compared to estimates from un-weighted fits to seasonal tracer cycles. Panels show results from 1000 random parameter sets and three contrasting precipitation drivers: Broad River (humid, temperate, with very little seasonality), Plynlimon (wet maritime climate with slight seasonality), and Smith River (Mediterranean climate with pronounced winter-wet, summer-dry seasonality). Seasonal tracer cycle amplitudes generally predict the **average** young water fraction, although they exhibit some systematic bias under strongly seasonal precipitation regimes like Smith River, where seasonal cycles in precipitation volume are correlated with seasonal cycles in tracer concentration. By contrast, mean transit time estimates from seasonal tracer cycles are highly unreliable in all precipitation regimes.

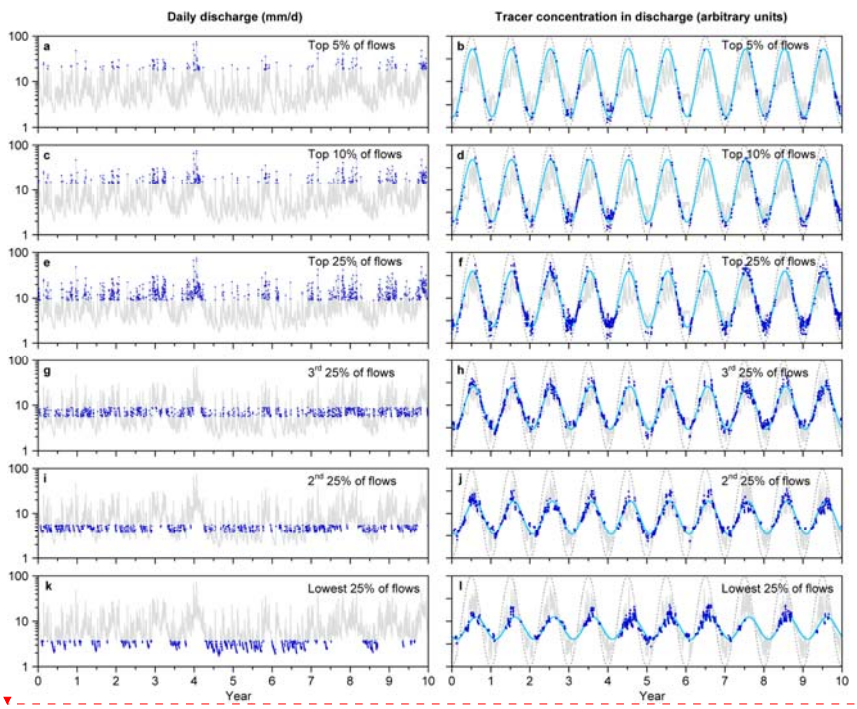

Figure 8. Daily discharges (left panels) and tracer concentrations (right panels) in streamflow from two-box model with reference parameter values and Plynlimon precipitation forcing. Individual discharge ranges and corresponding tracer concentrations are highlighted in dark blue. In right-hand panels, precipitation tracer concentrations are shown by dashed gray lines and best-fit sinusoidal fits to streamflow tracer concentrations are shown in light blue. At

Deleted: ¶

Deleted: 11

1 higher discharges, tracer cycles are less damped and less phase-shifted, indicating greater
2 fractions of young water in streamflow.

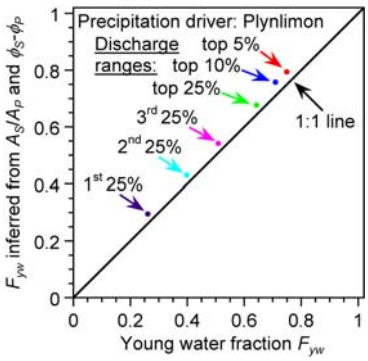

Figure 9. Time-averaged, flow-specific young water fractions F_{yw} for the six discharge ranges shown in Fig. 8, measured by age tracking in the model (with Plynlimon precipitation forcing and the reference parameter set), compared to F_{yw} values estimated from the amplitude ratios A_S/A_P and phase shifts ϕ_S-ϕ_P of the tracer cycles shown in Fig. 8.

Deleted: 12. Young

Deleted: 11

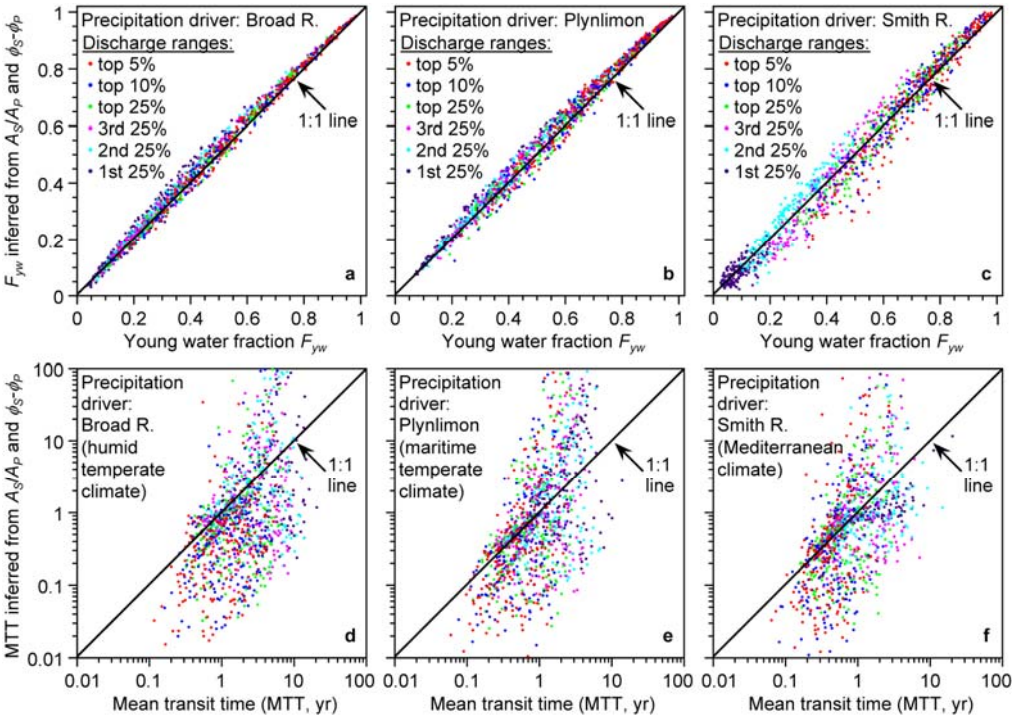

Figure 10. Young water fractions (F_{yw}) and mean transit times (MTT) in separate discharge ranges in streamflow from two-box model. Upper panels compare the time-averaged, flow-specific F_{yw} for each discharge range (measured by age tracking in the model) with F_{yw} values estimated from the amplitude ratios A_S/A_P and phase shifts ϕ_S-ϕ_P of the best-fit tracer cycle sinusoids in those discharge ranges (analogously to Fig. 8) using Eqs. (10)-(11) and (13)-(14) of Paper 1. Similar results (not shown) are also obtained for flow-weighted F_{yw} and flow-weighted tracer cycle sinusoids. Results obtained from tracer cycle amplitude alone (without phase information) are also similar, except in some cases where the amplitude ratio is small (particularly with Smith River precipitation forcing). Lower panels compare the MTT, determined by age tracking, with the MTT inferred from tracer amplitude ratios and phase shifts using Eqs. (10)-(11) from Paper 1. Each panel shows results from 200 random parameter sets and three contrasting precipitation drivers: Broad River (humid, temperate, with very little seasonality), Plynlimon (wet maritime climate with slight seasonality), and Smith River (Mediterranean climate with pronounced winter-wet, summer-dry seasonality). Tracer cycle amplitudes and phases generally predict the young water fractions in each

Deleted: 13.

Deleted: in

Deleted: 11

discharge range, although with some modest scatter. Mean transit time estimates, by contrast, are highly unreliable, exhibiting large scatter (note log scales).

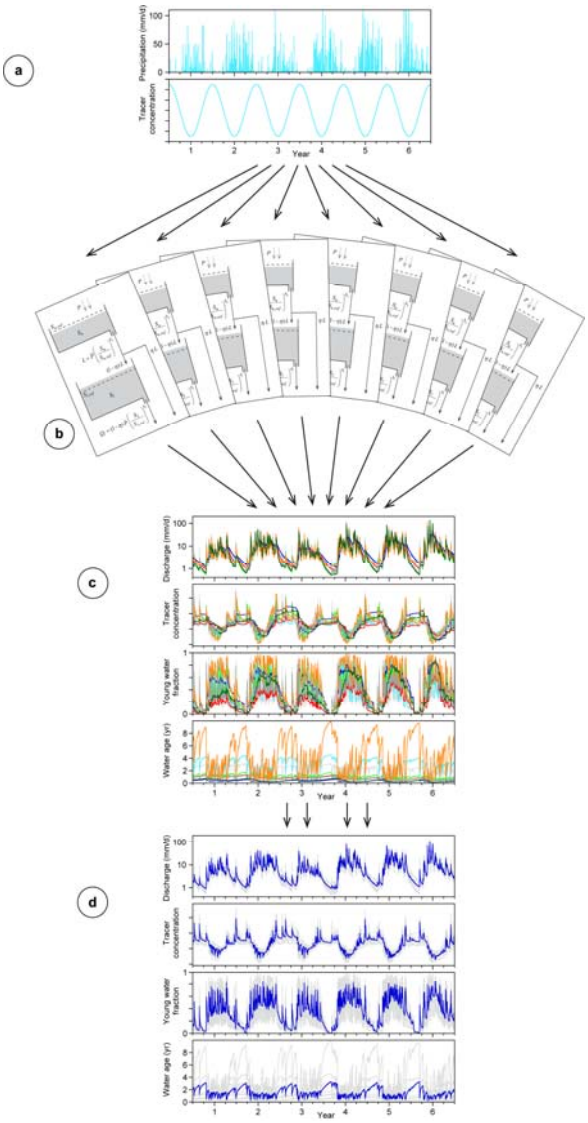

Figure 11. Scheme for simulating spatially heterogeneous catchments with nonstationary tributary subcatchments. A single precipitation time series (a) is used to drive eight copies of the model representing eight tributary streams (b), each with a different set of random parameter values. Streamflows, tracer concentrations, young water fractions, and water ages from these eight nonstationary tributaries (c, with each color representing a separate tributary

Deleted: 14.
Deleted: catchment
Deleted: 8

stream) are mass-averaged to determine the time series that would be observed in the merged streamflow (d, with blue lines showing the merged streamflow and gray lines showing the tributaries).

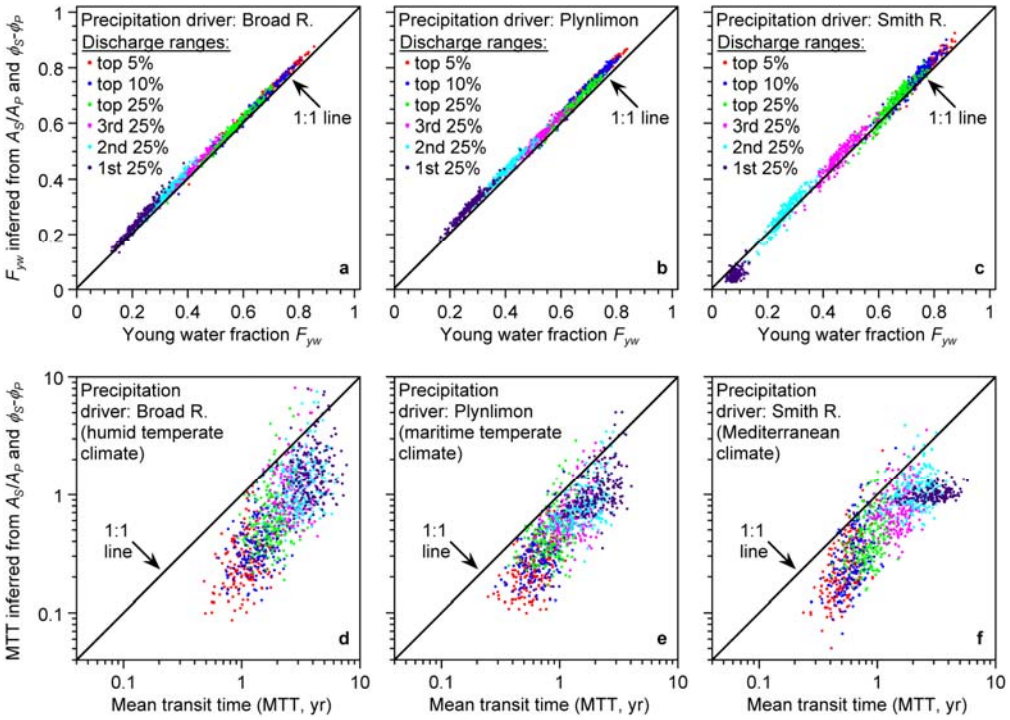

Figure 12. Actual and inferred young water fractions (F_{yw}, top panels) and mean transit times (MTT, bottom panels) in separate discharge ranges, under combined effects of nonstationarity and spatial heterogeneity. Panels show results for 200 synthetic catchments, each consisting of 8 copies of the two-box model with independent random parameter sets (Fig. 11). Upper panels compare average F_{yw}'s with F_{yw}'s predicted from amplitudes and phases of best-fit tracer cycle sinusoids **for each discharge range** (e.g., Fig. 8) using Eqs. (10)-(11) and (13)-(14) of Paper 1. Similar results (not shown) are also obtained for flow-weighted F_{yw}'s and flow-weighted tracer cycle sinusoids. Results obtained from tracer cycle amplitude alone (without phase information) are also similar, but exhibit slightly greater bias. Lower panels compare MTT with MTT predicted from tracer amplitude ratios and phase shifts using Eqs. (10)-(11) from Paper 1. Seasonal tracer cycle amplitudes and phases accurately predict young water fractions in separate flow regimes; the corresponding estimates of mean transit times exhibit **substantial** bias and scatter.

Deleted: 15. Combined effects of nonstationarity and spatial heterogeneity on actual and inferred young water fractions (F_{yw}, top panels) and mean transit times (MTT, bottom panels – note log scale). Panels show results for 1000 synthetic catchments, each consisting of 8 copies of the two-box model with independent random parameter sets (Fig. 14). Upper panels compare the average F_{yw} in discharge, determined by age-tracking in the model (on the horizontal axes) with F_{yw} values inferred from the amplitude ratios A_S/A_P and phase shifts ϕ_S-ϕ_P of the best-fit tracer cycle sinusoids, using Eqs. (10)-(11) and (13)-(14) of Paper 1. Results obtained from amplitude ratios alone (without phase information) are broadly similar, but exhibit somewhat greater bias. Lower panels compare MTT with MTT inferred from tracer amplitude ratios using Eqs. (10)-(11) from Paper 1. As in Fig. 10, light blue points show flow-weighted average F_{yw}'s and MTT's for each simulation, compared to estimates from flow-weighted fits to seasonal tracer cycles. Dark blue points show un-weighted average F_{yw}'s and MTT's, compared to estimates from un-weighted fits to seasonal tracer cycles. Seasonal tracer cycle amplitudes and phases generally predict young water fractions accurately, although systematic bias is evident in strongly seasonal precipitation regimes like Smith River, where seasonal cycles in precipitation volume are correlated with seasonal cycles in tracer concentration. Mean transit time estimates exhibit strong bias and large scatter across all precipitation regi ... [6]

Deleted: 14

Deleted: 11

Deleted: strong

Deleted: large

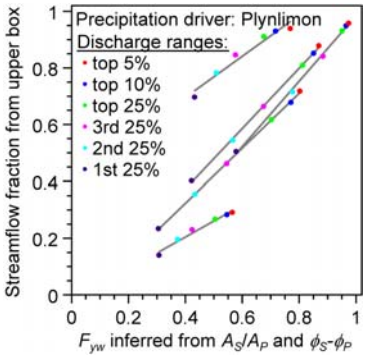

Figure 13. Correlations between flow-weighted young water fractions F_{yw} and fractional contributions of the upper box to streamflow across different discharge ranges, for five parameter sets illustrating the diversity of relationships that can arise in the model. The upper box contribution is strongly correlated with F_{yw} in all cases, although the slopes and the intercepts vary among parameter sets.

Deleted: 17.

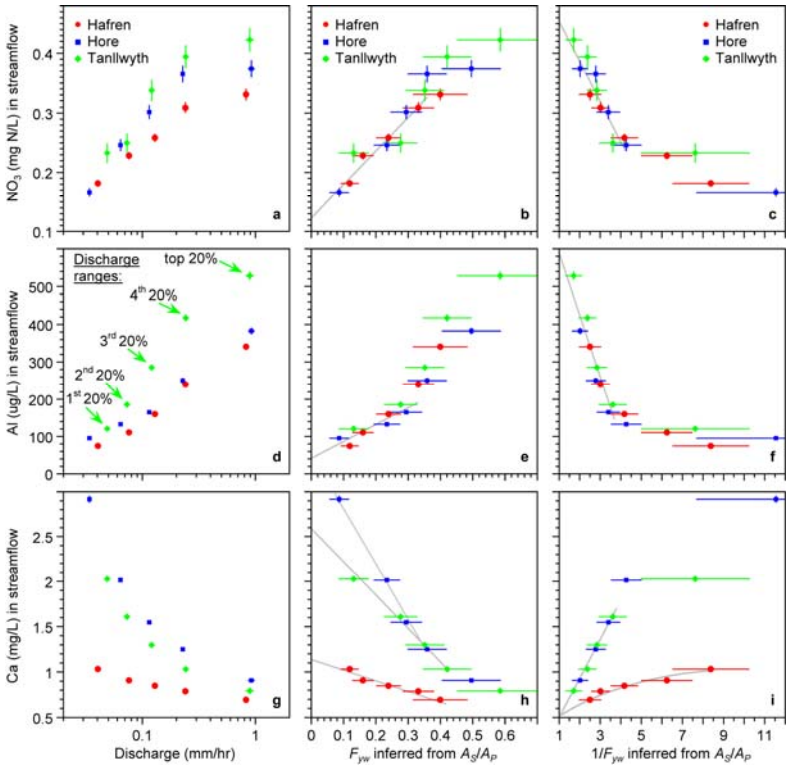

Figure 14. Concentrations of reactive chemical species as functions of discharge (left panels), young water fractions (middle panels), and reciprocal young water fractions (right panels) for streams draining three contrasting catchments at Plynlimon, Wales. Symbols show means for 20% intervals of each catchment's discharge distribution, and error bars indicate ±1 standard error. Gray lines are drawn by hand to indicate general trends. Concentration-discharge relationships in nitrate and aluminum differ among the three catchments (a and d), but collapse to single concentration-F_{yw} relationships (b-c and e-f). These concentration-F_{yw} relationships extrapolate to broadly consistent old water end-members (F_{yw}=0, panels b and e) and young water end-members (F_{yw}=1, panels d and f). Calcium follows different concentration-F_{yw} relationships in the three streams, which extrapolate to three different old water end-members (h) but roughly the same young water end-member (i).

Deleted: 18.

Deleted: placed

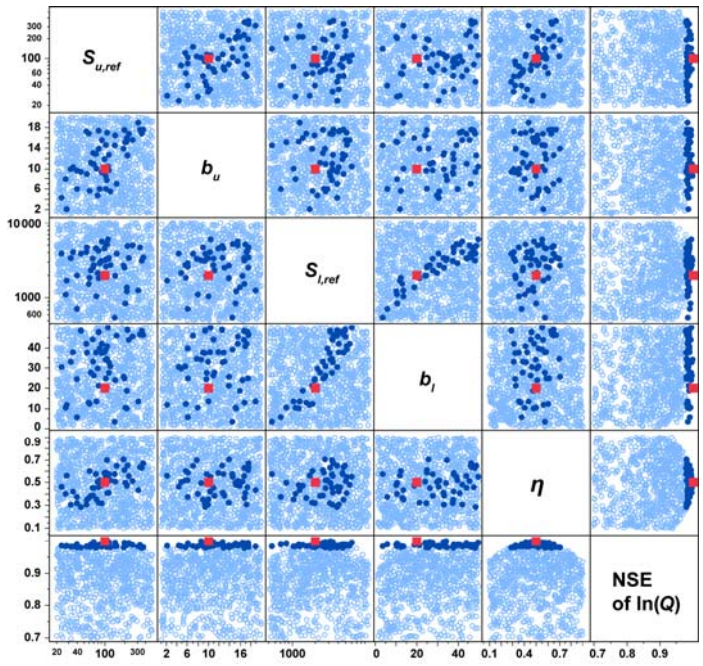

Figure B1. Equifinality in discharge predictions. The scatterplot matrix shows relationships among 1000 random parameter sets and the Nash-Sutcliffe efficiency (NSE) of discharge time series driven by Smith River (Mediterranean climate) precipitation forcing. The red square indicates the "reference" parameter set that was used to generate the discharge time series that the other parameter sets were tested against; these reference parameters thus correspond to NSE=1.00 by definition. The dark blue dots show the best-fitting 50 (or 5%) of the parameter sets, all with NSE≥0.98. Excellent discharge predictions can be obtained across almost the full range of all five model parameters, except the partition coefficient η, which performs well across only about half its range. The dark blue dots show clear correlations between the reference storage levels in each box ($S_{u,ref}$, $S_{l,ref}$) and the corresponding drainage function exponents (b_u, b_l); these correlations delimit regions with nearly constant hydraulic response time scales, as defined by Eqs. (10)-(11).

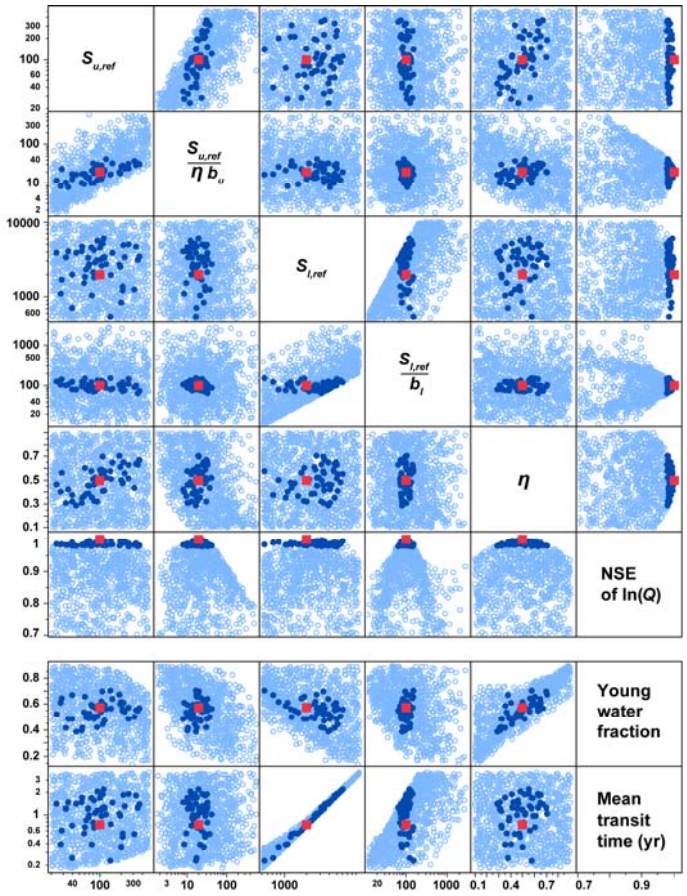

Figure B2. Equifinality partly cured by parameter transformations. The scatterplot matrix shows relationships among 1000 random parameter sets and the Nash-Sutcliffe efficiency (NSE) of discharge time series driven by Smith River (Mediterranean climate) precipitation forcing, along with two key model outputs, the young water fraction and mean transit time in discharge (bottom two rows). As in Fig. B1, the red square indicates the "reference" parameter set that was used to generate the discharge time series that the other parameter sets were tested against; these reference parameters thus correspond to NSE=1.00 by definition. The dark blue dots show the best-fitting 50 (or 5%) of the parameter sets, all with NSE≥0.98. In contrast to Fig. B1, three of the five parameters can be constrained by calibration against discharge (as shown by the clear peaks in NSE), and none of the parameters are strongly correlated with one another. However, the two reference storage volumes $S_{u,ref}$ and $S_{l,ref}$ remain poorly constrained.

The mean transit time is determined almost entirely by $S_{l,ref}$, so it cannot be constrained by parameter calibration against the streamflow hydrograph.

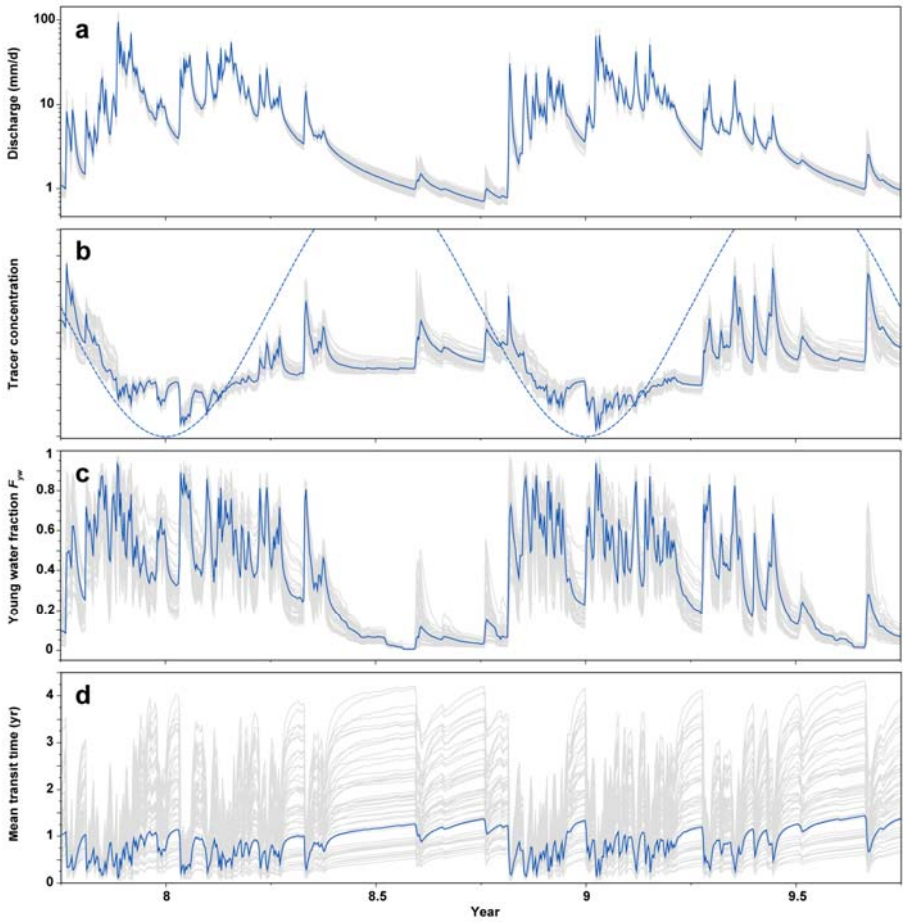

Figure B3. Excerpts from time series of discharge, tracer concentrations, young water fractions, and mean travel times in the two-box model with Smith River (Mediterranean climate) precipitation forcing and the reference parameter set (the dark lines, for the parameter values shown by the red squares in Figs. B1 and B2) and the 50 parameter sets that come closest to matching the reference discharge time series (the light gray lines, for the parameter sets shown by the solid blue dots in Figs. B1 and B2). The 50 gray hydrographs (panel a) cluster closely around the blue hydrograph (which is unsurprising because they have been selected to do so). The 50 gray tracer concentration curves (panel b) also generally follow the blue curve (the precipitation

tracer sinusoid is shown for comparison by the dashed line). By contrast, the young water fraction F_{yw} (panel c) and mean transit time (panel d) are much more variable; the gray curves vary by an average range of 0.3 in F_{yw} and a factor of 9.5 in mean transit time.

is illustrated here for the upper box; the lower box is handled analogously. The storage in the upper box is updated using the following equation:

$$S_u(t_{i+1}) - S_u(t_i) = \Delta t \left(P - \rho k_u S_u(t_{i+1})^{b_u} - (1-\rho) k_u S_u(t_i)^{b_u} \right) \quad , \qquad (4)$$

where $S_u(t_i)$ is the storage in the upper box at the beginning of the i^{th} time interval (with length Δt), $S_u(t_{i+1})$ is the storage at the end of that interval (and thus the beginning of the next), and P is the average precipitation rate over the interval. Equation (4) is implicit and nonlinear; there is no closed-form solution for the future storage $S_u(t_{i+1})$, which instead is found using Newton's method. The relative dominance of the trapezoidal and backward Euler solutions is determined by the weighting factor ρ, which takes on values between ρ=0.5 (trapezoidal method) and ρ=1 (backward Euler method). The value of ρ in Eq. (4) is determined for each time step using the simple stability criterion,

$$\rho = \min\left(0.5 + 0.5 \frac{\left(P - k_u S_u(t_i)^{b_u} \right) \Delta t}{\left(P / k_u \right)^{1/b_u} - S_u(t_i)} \ , \ 1 \right) \quad , \qquad (5)$$

where the numerator represents the amount that S_u would change during one time step if the instantaneous drainage rate L in Eq. (1) were projected forward in time, and the denominator represents the difference between S_u's current value and its equilibrium value at the precipitation rate P. Equation (5) says that if the trapezoidal method would move S_u by only a small fraction of the distance to its equilibrium value (at the precipitation rate P), then the stability advantages of the backward Euler method

unnecessary and the more accurate trapezoidal method should dominate the solution instead ($\rho \approx 0.5$). On the other hand, if the trapezoidal method would overshoot the equilibrium value, then ρ=1 and the fully implicit backward Euler method is used to solve Eq. (4). The closer the trapezoidal method would come to overshooting the equilibrium, the larger the value of ρ and the greater the weight that is given to the backward Euler solution. The guaranteed stability of the backward Euler method is important when b_u or b_l is large, because the underlying equations can

become quite stiff. After the final value of S_u is determined by Eq. (4), the drainage from S_u between t_i and t_{i+1} is determined by mass balance:

$$L = P + \left(S_u(t_i) - S_u(t_{i+1})\right)/\Delta t \quad , \tag{6}$$

where L is the average drainage rate over the interval Δt between t_i and t_{i+1}.

The tracer concentrations are determined under the assumption that each box is well mixed, implying that individual water parcels within each box do not need to be tracked, and also that the concentration draining from each box equals the average concentration within the box. I make the simplifying assumption that each box's inflow and outflow rates (and also inflow concentrations) are constant over each day. Again taking the upper box as an example, these assumptions imply that starting from $t=t_i$ the tracer concentration will evolve as

$$\frac{dC_u}{dt} = \frac{P(C_P - C_u)}{S_u(t_i) + (P-L)(t-t_i)} \quad , \tag{7}$$

where the denominator expresses how the volume in the box changes with time from its initial value of $S_u(t_i)$. Integrating Eq. (7) over an interval Δt yields the concentration updating formula:

$$C_u(t_{i+1}) = C_P + \left(C_u(t_i) - C_P\right)\left(\frac{S_u(t_i)}{S_u(t_{i+1})}\right)^{\left(P/(P-L)\right)} \quad , \tag{8}$$

where any quantities that are not shown as functions of time are constant at their average values over the interval. Equation (8) could potentially become difficult to compute when P and L are nearly equal (differing by, say, less than 1 part in 1000), and the power function approaches its exponential limit. In such cases the change in volume in Eq. (7) becomes trivially small, and one can replace Eq. (8) with the more familiar exponential formula for a well-mixed box of constant volume:

$$C_u(t_{i+1}) = C_P + \left(C_u(t_i) - C_P\right)\exp\left(-P\,\Delta t / S_u\right) \quad . \tag{9}$$

After the tracer concentrations are updated, the average concentrations in drainage are calculated by mass balance, as follows:

$$C_L = \left[C_P(t_i)\,P + C_u(t_i)\,S_u(t_i) - C_u(t_{i+1})\,S_u(t_{i+1})\right]/L \quad , \tag{10}$$

where C_L is the average concentration in drainage over time interval between t_i and t_{i+1}.

The mean age within each box is modeled analogously to the tracer concentrations, following the "age mass" concept widely used in groundwater hydrology. Here I will illustrate the approach using the example of the lower box, since it is the more complex case (for the upper box, the input age in precipitation is zero, but this is not true for the upper-box drainage that recharges the lower box). Assuming that the inflow and outflow rates $L(1\text{-}\eta)$ and Q_l are constant over a day, as is the average age $\bar{\tau}_L$ of the inflow from the upper box, the mean age in the lower box should evolve according to

$$\frac{d\bar{\tau}_l}{dt} = \frac{L(1-\eta)(\bar{\tau}_L - \bar{\tau}_l)}{S_l(t_i) + (L(1-\eta) - Q_l)(t - t_i)} + 1 \quad , \qquad (11)$$

which is directly analogous to Eq. (7), except for additional term of +1, which accounts for the continual aging of the water in the box. The solution to Eq. (11) is

$$\bar{\tau}_l(t_{i+1}) = \bar{\tau}_L + \frac{S_l(t_{i+1})}{2L(1-\eta) - Q_l} + \left(\bar{\tau}_l(t_i) - \bar{\tau}_L - \frac{S_l(t_i)}{2L(1-\eta) - Q_l} \right) \left(\frac{S_l(t_i)}{S_l(t_{i+1})} \right)^{\left(\frac{L(1-\eta)}{L(1-\eta) - Q_l} \right)} \quad , \quad (12)$$

where $\bar{\tau}_l(t_i)$ and $\bar{\tau}_l(t_{i+1})$ are the mean age of the water in the lower box at the beginning and end of the time interval. Analogously to tracer concentrations, one can calculate the mean age of the drainage from the box based on the inputs and the change in mean age inside the box, using conservation of "age mass":

$$\bar{\tau}_{Q_l} = \left[\bar{\tau}_L(t_i)\,(1-\eta) + \bar{\tau}_l(t_i)\,S_l(t_i) - \left(\bar{\tau}_l(t_{i+1}) - \Delta t \right) S_l(t_{i+1}) \right] / Q_l \quad , \qquad (13)$$

where the factor of $-\Delta t$ accounts for the aging of the contents of the box.

The approach used here for concentrations and water ages requires the assumption that input fluxes to each box are constant within each time interval (but constant at their average values, not their initial values). This is a reasonable approximation, particularly when we have no sub-daily precipitation data. And in exchange for this simplifying assumption, equations (8), (9), and (12) provide something important, namely, the exact analytical solution for the evolution of concentration and age during each time interval. Thus these equations directly solve for the correct result even if, for example, an individual day's rainfall is much greater than the total volume of the upper box. The equations above will correctly calculate the consequences of the (potentially many-fold) flushing that occurs in such cases. The approach

Page 5: [3] Deleted **Author**

above also guarantees exact consistency between stocks and fluxes (but note, not in the usual way by updating stocks with fluxes, but rather by calculating output fluxes from inputs and changes in stocks). Readers should keep in mind that all stocks and properties of stocks (i.e., storage volumes, concentrations, and ages) are expressed as the instantaneous values at the beginning of each time interval, and that fluxes and properties of fluxes (i.e., water fluxes and their concentrations and ages) are expressed as averages over each time interval. Otherwise it could be difficult to make sense of the equations above

Page 14: [4] Deleted **Author**

Equifinality in hydraulic behavior and divergence in travel times

The analysis outlined in Sect. 3.2 further implies that approximate equifinality is inevitable, even in such a simple model, because variations in the exponents b_u and b_l and the reference storage levels $S_{u,ref}$ and $S_{l,ref}$ will have nearly offsetting effects on the model's runoff response. Equations (20) and (21) show that, for a given average precipitation forcing, any parameter values for which the partitioning coefficient η and the ratios $S_{u,ref}/b_u$ and $S_{l,ref}/[(1-\eta)b_l]$ are invariant would give nearly equivalent hydrograph predictions, because the hydraulic response timescales of the upper and lower boxes, and their relative contributions to discharge, would be invariant. These conditions can be achieved for widely varying values of the individual parameters b_u, b_l, $S_{u,ref}$, and $S_{l,ref}$.

This equifinality problem can be readily visualized by plots like Fig. 7. To generate Fig. 7, I ran the model with Smith River precipitation forcing and the reference parameter set (shown by the red squares in Fig. 7), and used the resulting daily hydrograph (after the spin-up period) as virtual "ground truth" for model calibration. I then ran the model with 1000 random parameter sets, and used the Nash-Sutcliffe efficiency (NSE) of the logarithms of discharge to measure how well their hydrographs matched the reference hydrograph (thus the reference hydrograph has NSE=1 by definition). The 50 best-fitting parameter sets, all with NSE≥0.98, are shown as dark blue points in Fig. 7. The bottom row of scatterplots shows the conventional "dotty plots". Their flat tops are the hallmark of equifinality, i.e., wide ranges of parameter values give equally good hydrograph predictions (Beven, 2006). Only the partition coefficient η, which performs well

across half its range, can be even modestly constrained by calibration. (The other precipitation drivers yield results similar to those shown in Fig. 7.)

The other panels of the scatterplot matrix also give important clues to the origins of the observed equifinality. In particular, the best-fitting parameter sets show strong correlations between $S_{u,ref}$ and b_u, and between $S_{l,ref}$ and b_l, as expected from the perturbation analysis presented above. Thus good model performance can be obtained across almost the entire range of these parameters, but only for specific parameter combinations. The interdependence of the parameters is visually obvious in the scatterplot matrix, but is invisible in the conventional "dotty plots".

This information can be exploited to design parameter spaces that are more identifiable through calibration. An ideal parameter space would be one in which 1) all parameters are highly identifiable, meaning the goodness-of-fit surface is strongly curved along each parameter axis, and 2) in the best-fitting parameter sets, no parameters are strongly correlated with one another. The second of these criteria is necessary (although not sufficient) for the first, as Fig. 7 illustrates. A third criterion is that all parameters that are needed for simulating any quantities of interest must be determined somehow within the parameter space, either individually or through combinations of other parameters. Thus, for example, although the volumes of the boxes ($S_{u,ref}$ and $S_{l,ref}$) are strongly correlated with their exponents (b_u and b_l), the parameter space must allow them to be individually determined, because as Eqs. (22-24) suggest, the mean transit times will be primarily controlled by the volumes alone (not in combination with the exponents), whereas the runoff response will be primarily controlled by the ratios of volumes to exponents (Eqs. 20-21). These criteria, plus some trial and error, lead to a more identifiable parameter space, whose five axes are $S_{u,ref}$, $S_{l,ref}$, $S_{u,ref}/(\eta \cdot b_u)$, $S_{l,ref}/b_l$, and η.

Figure 8 shows that this parameter space exhibits much less equifinality than the parameter space shown in Fig. 7, although the underlying parameter sets and model simulations are exactly the same. All that has been done is to re-project the parameter space onto a different set of coordinate axes in which the curvature of the goodness-of-fit surface is more clearly visible. Thus, much of the apparent equifinality in the parameter space has been eliminated by simple transformations of variables. These transformations can be designed by eye in this case, because the dimensionality of the original parameter space is low. In higher-dimension parameter spaces, multivariate techniques such as factor analysis may be helpful. Nonetheless, given the obvious

utility of this simple correlation analysis and the perturbation analysis of Sect. 3.2, it is surprising that they are not more widely used in hydrological modeling.

Despite the improved identifiability of the parameter space, however, it is still not possible to constrain the mean transit time by calibration to the hydrograph. As the bottom row of scatterplots in Figure 8 shows, the mean transit time (MTT) is almost entirely determined by the lower box's reference volume $S_{l,ref}$, as one would expect from Eq. 24. However, as predicted by the perturbation analysis in Sect. 3.2, and as shown by Fig. 8, the runoff response of the model system is essentially independent of $S_{l,ref}$ and therefore cannot be used to constrain it. The runoff response does depend on the ratio of $S_{l,ref}$ to b_l, and thus can be used to constrain that ratio, but it cannot constrain $S_{l,ref}$ by itself, and thus it cannot constrain the MTT. For the young water fraction F_{yw} the outlook is not quite as bleak, because F_{yw} is correlated with the partition coefficient η, which can be constrained somewhat by calibration. As a result, it appears that Fyw could potentially be constrained within roughly 1/3 of its full range by parameter calibration to the hydrograph.

Figure 9 provides a different visualization of the same equifinality problem. Figure 9 shows a two-year excerpt from the simulated time series of streamflows, tracer concentrations, young water fractions, and mean transit times for the reference parameter set (the blue curves), along with the 50 parameter sets that gave the best fit to the reference hydrograph (the gray curves). Because these 50 parameter sets were those that matched the reference hydrograph best, it is unsurprising that the 50 gray hydrographs generally follow the blue reference hydrograph in Fig. 9a. The 50 gray tracer concentration time series also follow the blue reference time series (Fig. 9b), but with somewhat greater variability than the hydrographs, indicating that the parameter values affect the chemographs and the hydrographs in somewhat different ways. But the most striking feature of Fig. 9 is the much greater variability among the young water fractions F_{yw} and (especially) the mean transit times MTT for these same parameter sets (Fig. 9c-d). Although all the parameter sets fit the reference hydrograph nearly perfectly, they vary over a range of 0.3 in F_{yw} (out of a total possible range of 1.0), and over a factor of 9.5 in MTT, on average for the whole time period. Thus these time series demonstrate, consistent with Fig. 8, that there are wide ranges of variability in F_{yw} and especially MTT that cannot be constrained by calibration to the hydrograph. This observation naturally leads one to ask whether F_{yw} and MTT can be estimated at all, and if so how. I will now turn to these questions in the next sections.

Equifinality in discharge predictions. The scatterplot matrix shows relationships among 1000 random parameter sets and the Nash-Sutcliffe efficiency (NSE) of discharge time series driven by Smith River precipitation forcing. The red square indicates the "reference" parameter set that was used to generate the discharge time series that the other parameter sets were tested against; these reference parameters thus correspond to NSE=1.00 by definition. The dark blue dots show the best-fitting 50 (or 5%) of the parameter sets, all with NSE≥0.98. Excellent discharge predictions can be obtained across almost the full range of all five model parameters, except the partition coefficient η, which performs well across only about half its range. The dark blue dots show clear correlations between the reference storage levels in each box ($S_{u,ref}$, $S_{l,ref}$) and the corresponding drainage function exponents (b_u, b_l); these correlations delimit regions with nearly constant hydraulic response time scales, as defined by Eqs. (20)-(21).

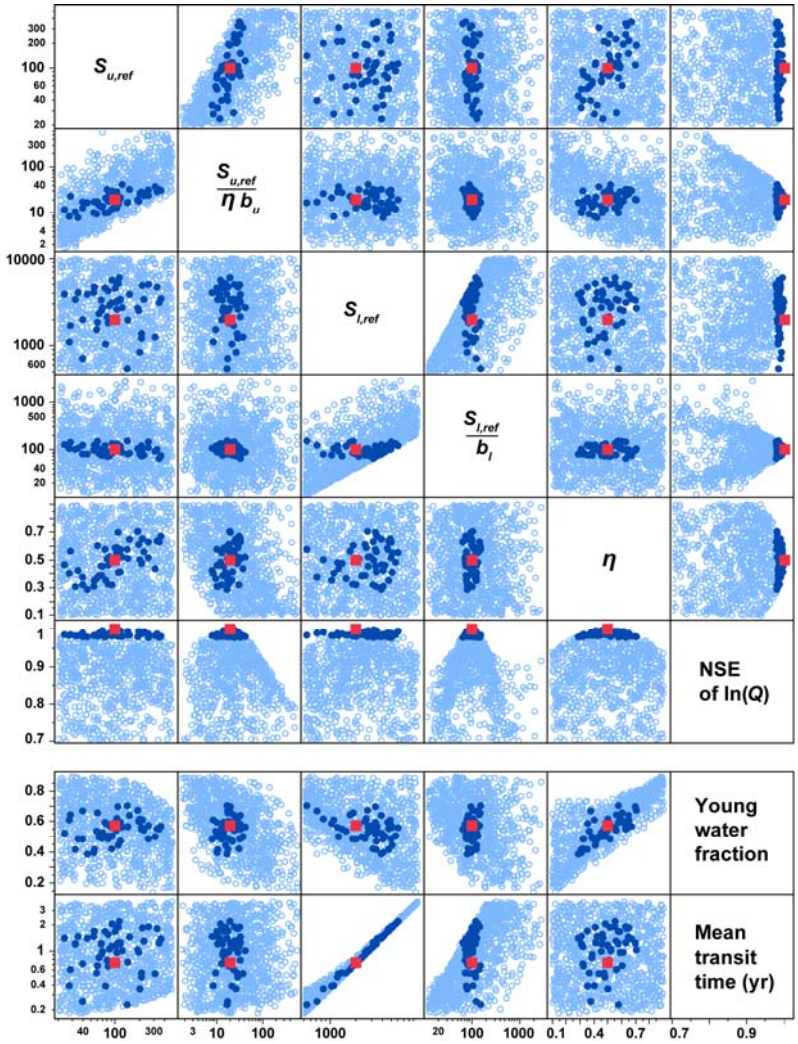

Figure 8. Equifinality partly cured by parameter transformations. The scatterplot matrix shows relationships among 1000 random parameter sets and the Nash-Sutcliffe efficiency (NSE) of discharge time series driven by Smith River precipitation forcing, along with two key model outputs, the young water fraction and mean transit time in discharge (bottom two rows). As in Fig. 7, the red square indicates the "reference" parameter set that was used to generate the discharge time series that the other parameter sets were tested against; these reference parameters thus correspond to NSE=1.00 by definition. The dark blue dots show the best-fitting 50 (or 5%) of the parameter sets, all with NSE≥0.98. In contrast to Fig. 7, three of the five parameters can be constrained by calibration against discharge (as shown by the clear peaks in NSE), and none of the parameters are strongly correlated with one another. However, the two reference storage volumes $S_{u,ref}$ and $S_{l,ref}$ remain poorly constrained. The mean transit time is determined almost

entirely by $S_{l,ref}$, so it cannot be constrained by parameter calibration against the streamflow hydrograph.

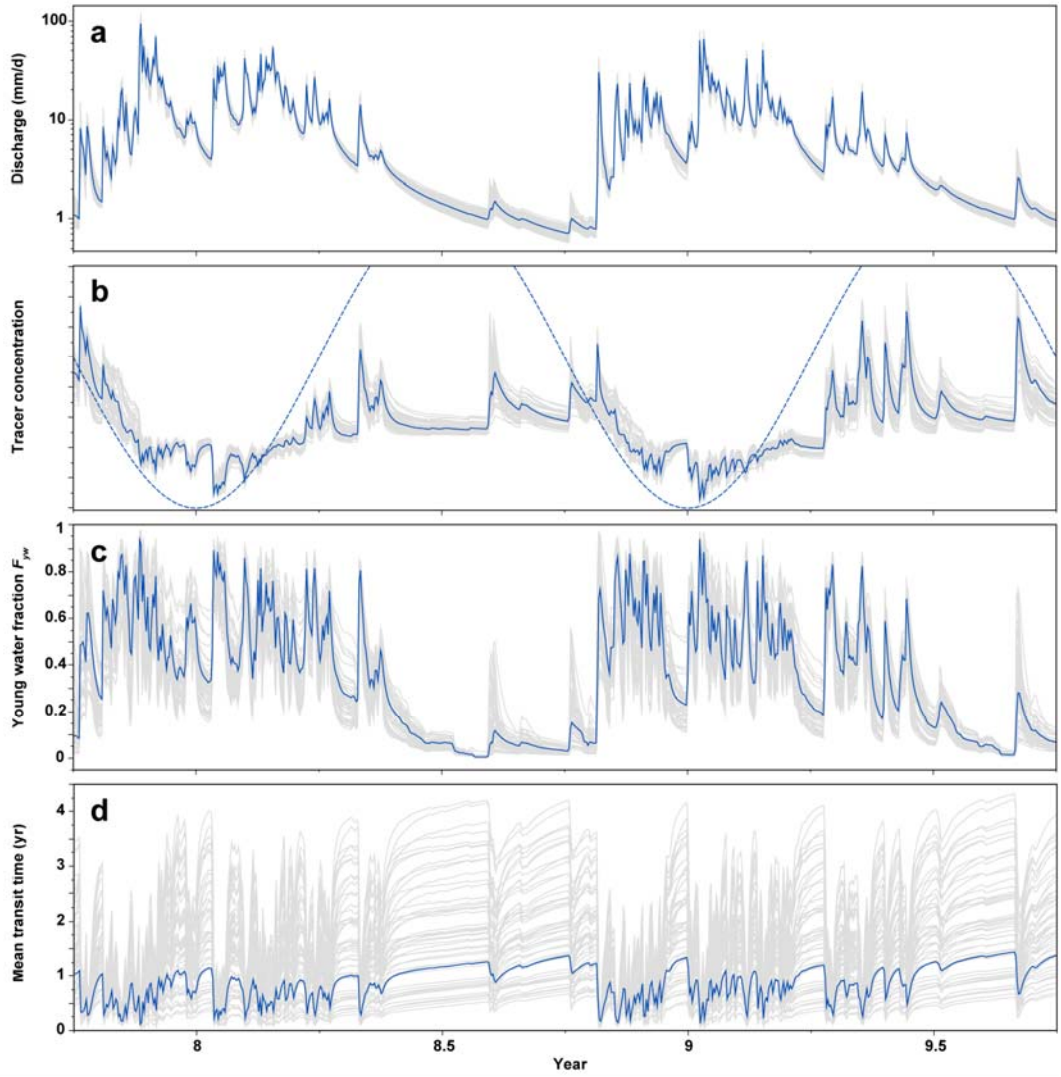

Figure 9. Excerpts from time series of discharge, tracer concentrations, young water fractions, and mean travel times in the two-box model with Smith River precipitation forcing and the reference parameter set (the dark lines, for the parameter values shown by the red squares in Figs. 7 and 8) and the 50 parameter sets that come closest to matching the reference discharge time series (the light gray lines, for the parameter sets shown by the solid blue dots in Figs. 7 and 8). The 50 gray hydrographs (panel a) cluster closely around the blue hydrograph (which is unsurprising because they have been selected to do so). The 50 gray tracer concentration curves (panel b) also generally follow the blue curve (the precipitation tracer sinusoid is shown for comparison by the dashed line). By contrast, the young water fraction F_{yw} (panel c) and mean

transit time (panel d) are much more variable; the gray curves vary by an average range of 0.3 in F_{yw} and a factor of 9.5 in mean transit time.

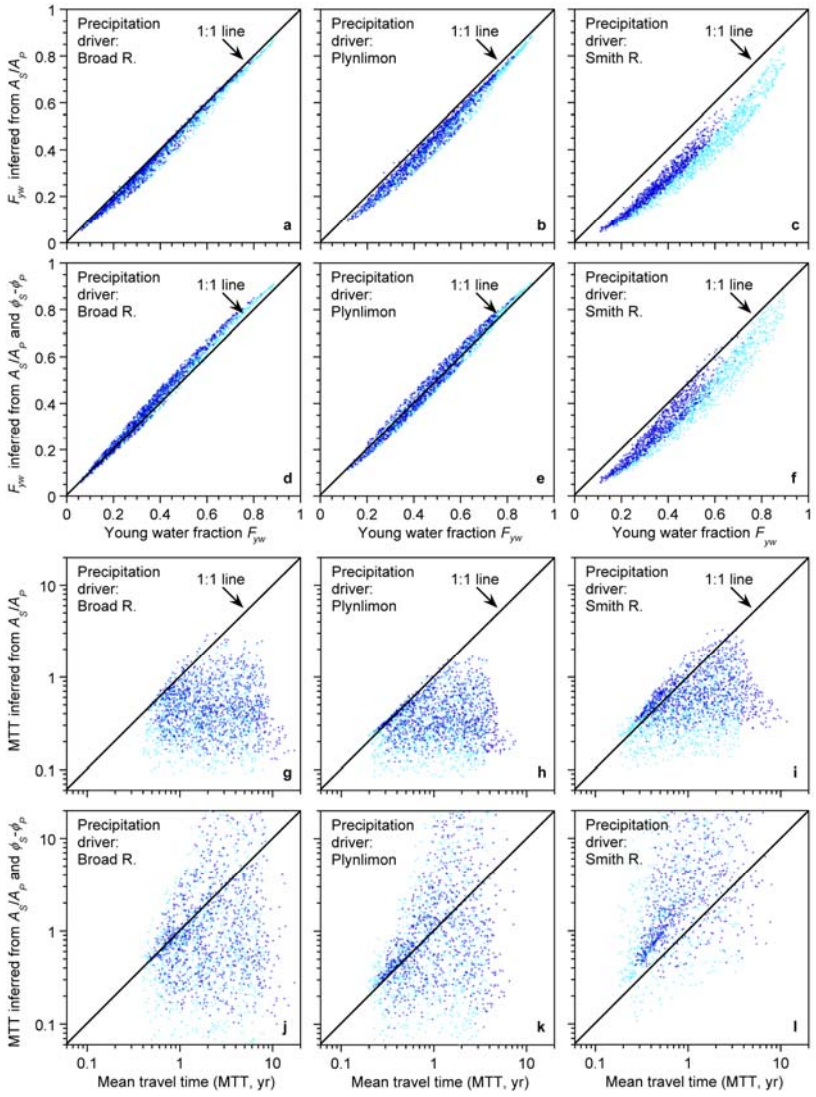

Figure 10.

Page 54: [6] Deleted **Author**

15. Combined effects of nonstationarity and spatial heterogeneity on actual and inferred young water fractions (F_{yw}, top panels) and mean transit times (MTT, bottom panels – note log scale). Panels show results for 1000 synthetic catchments, each consisting of 8 copies of the two-box model with independent random parameter sets (Fig. 14). Upper panels compare the average F_{yw} in discharge, determined by age-tracking in the model (on the horizontal axes) with F_{yw} values inferred from the amplitude ratios A_S/A_P and phase shifts ϕ_S-ϕ_P of the best-fit tracer cycle

sinusoids, using Eqs. (10)-(11) and (13)-(14) of Paper 1. Results obtained from amplitude ratios alone (without phase information) are broadly similar, but exhibit somewhat greater bias. Lower panels compare MTT with MTT inferred from tracer amplitude ratios using Eqs. (10)-(11) from Paper 1. As in Fig. 10, light blue points show flow-weighted average F_{yw}'s and MTT's for each simulation, compared to estimates from flow-weighted fits to seasonal tracer cycles. Dark blue points show un-weighted average F_{yw}'s and MTT's, compared to estimates from un-weighted fits to seasonal tracer cycles. Seasonal tracer cycle amplitudes and phases generally predict young water fractions accurately, although systematic bias is evident in strongly seasonal precipitation regimes like Smith River, where seasonal cycles in precipitation volume are correlated with seasonal cycles in tracer concentration. Mean transit time estimates exhibit strong bias and large scatter across all precipitation regimes.

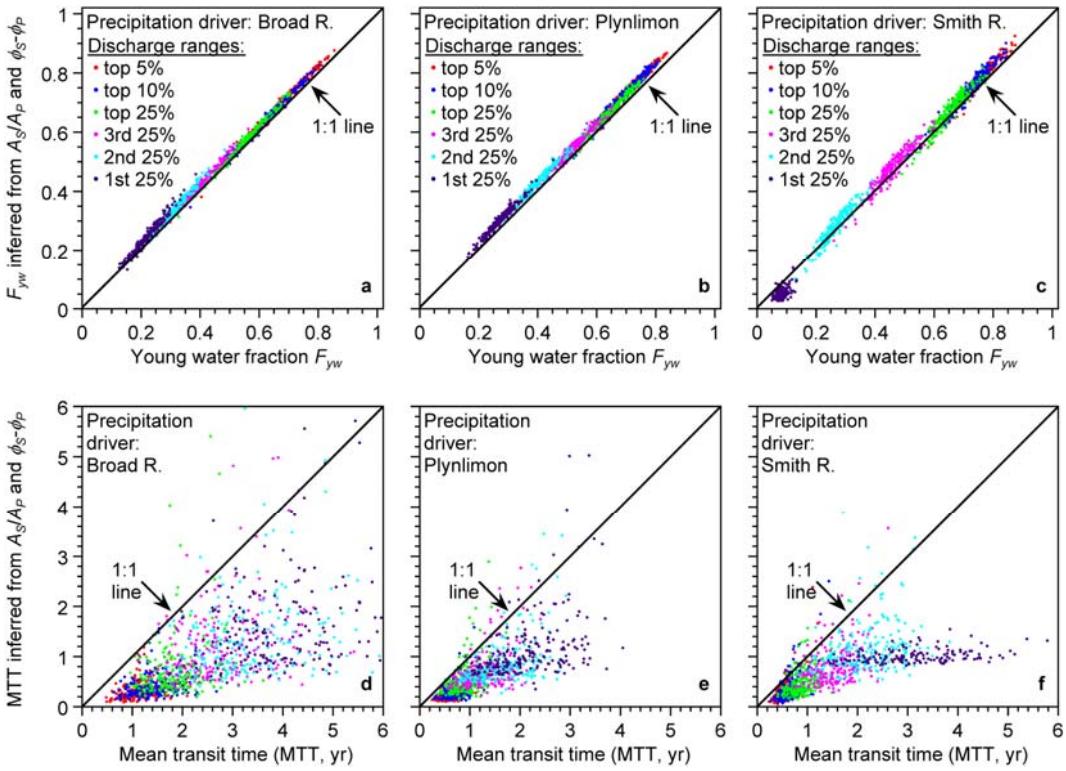

Figure 16.