Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2018-429-AC3, 2018

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Quantifying new water fractions and transit time distributions using ensemble hydrograph separation: theory and benchmark tests" *by* J. Kirchner

J.W. Kirchner

kirchner@ethz.ch

Received and published: 21 October 2018

I thank Nicolas Rodriguez (hereafter NR) for his comments. I have reproduced those comments below (in normal type), with my responses (in bold).

This paper has been really pleasant to read. The quality of the writing, of the figures, and of the mathematics is really high. The amount of meaningful explanations is impressive. I think the paper will have a strong impact on the isotope hydrology community. I think we need more approaches such as this one in tracer hydrology.

Thanks.

That being said, I would like to mention two things that could be discussed.

The first one is more context about travel time modeling. While the introduction and the discussion compare well this ensemble approach to the traditional hydrograph separation, a large part of the paper also deals with determining travel time distributions (TTDs). Yet, only little is said about travel time modeling, especially in the introduction.

This is because the paper does not deal with simulation modeling (nor is it intended to).

I think it should be mentioned that the ensemble approach deals with a current need in isotope hydrology to have more data-driven approaches and non-parametric TTDs. I think this is exactly what the proposed approach brings compared to already existing approaches, but not more. Unlike reviewer n?1, I believe we should not try to formally compare methods which have different purposes. This approach calculates only the streamflow average TTD suggested by the tracer data, without assuming its shape. This is novel and important. Yet the proposed method can only be used for the period covered by training data (i.e. "backwards"), and for streamflow only.

As with any technique for analyzing the behavior of a natural system, yes, this method does use data, and thus "can only be used for the period covered by the training data".

But the statement that the method is useful "for streamflow only" is puzzling, given that Section 5.4 explicitly mentions that "these methods could potentially be applied to infer transit times in other catchment fluxes, such as groundwater seepage or evapotranspiration." Of course, one would first need tracer data from those fluxes (as would also be required for any other method with the same objectives).

Using StorAge Selection (SAS) functions with assumed shapes allows one to obtain the time-varying TTDs at every moment and in every flux (backwards), and the Residence

Time Distributions (RTDs). But more importantly, SAS function allow one to simulate other time-varying solute fluxes (e.g. Benettin et al., 2015) with the calibrated model in a forward way (even outside the period covered by training data). Note that a model based on SAS functions can consist of just a handful of parameters (e.g. Benettin et al. 2017) which makes it really competitive. Yet, I also agree that there are clear limitations in approaches based on SAS functions.

This is a misleading comparison. Of course with a simulation model one can *simulate* all kinds of things (without any real-world constraint), but *analysis* of real-world data is a fundamentally different task. With a simulation model, *of course* one can *simulate* age distributions "at every moment and in every flux", including times and fluxes for which one has no data (and thus for which one has no idea whether the simulations are realistic or not). But what does one learn from doing so? One primarily learns about the consequences of one's modeling assumptions, but not about whether they accurately describe the real-world system.

My second comment relates to the potential limitations of the proposed approach. I think that all the choices made to derive the mathematical solutions were presented as if they are the best choices for any tracer data set, or the only choices possible.

I presented the choices that were actually made, along with the rationales for them. I neither stated nor implied that they are the best choices for any tracer data set, or the only choices possible, as NR claims.

This may not be true in all cases. The discussion would benefit from an objective assessment of the problems that could occur when trying to apply the approach to real tracer data.

I don't know what NR means by an "objective" assessment. The whole point of the paper is to quantitatively test the approach using a benchmark model (so that we know what the right answer would be), using data that are a reasonable

approximation to real tracer data. The results are presented in 17 figures and extensive discussion. What more "objective assessment" would one want?

In my opinion, this ensemble approach will be accurate only for the left tail of the TTDs, while it truncates (cf. equation 30) older ages.

It is not correct to say that this approach "truncates" the TTD; instead it (correctly) *makes no assumption* about the TTD beyond the specified range of lag times.

This is a critical problem in travel time modeling in general (Stewart et al., 2012; Stewart & Morgenstern, 2016). It is already mentioned in the reply to reviewer n?1, but I think it should be clearly written in the discussion as well.

As I have already replied to reviewer #1, the problem with estimating long tails is intrinsic to the use of conservative tracers. It is a problem of the (low) information content in the tracer data on those time scales, and that problem cannot be solved by clever analytical tricks. This can be mentioned in the discussion. Whether it is "a critical problem" depends on whether one is interested in the long tail and the mean transit time (which are difficult or impossible to constrain with conservative tracers), or the shorter-term behavior (which is the focus of my approach, and which we can actually learn a lot about from tracer data).

1. To give more context in the introduction you could mention and describe briefly the common methods to estimate TTDs, namely the Lumped Parameter Models (e.g McGuire & McDonnell, 2006, and references therein), flux tracking in conceptual models (e.g. Hrachowitz et al., 2013), SAS functions applied to a single control volume (e.g. Benettin et al., 2017), and particle tracking in distributed models (e.g. Davies et al., 2013; Danesh-Yazdi et al., 2018). Doesn't the ensemble approach answer the need to have alternatives to these methods, which all need to assume an underlying model for water transport?

Yes, exactly. These methods not only require a *physically correct* underlying model for water transport (which can be highly problematic in practice), they also require *continuous input data* (because in any time-integrating model, errors accumulate). These methods can be briefly mentioned in the introduction or discussion, but the paper is not (and should not become) a review and comment on the broad topic of transit time modeling.

2. P7, L11-13: Least squares regression means that any real data set with " outliers" (which may just be tracer values one did not expect) is likely to adversely affect the results from the ensemble approach, as it is suggested here. Same for the least squares solution in equation 38. This is in my opinion one of the limitations of the proposed method.

A limitation compared to what? *Every other method* for separating hydrographs, estimating TTD's, or calibrating SAS functions (etc.) is *also* potentially vulnerable to outliers. And of course parameter calibrations in simulation models are *also* affected by outliers, in ways that are often poorly understood.

This should perhaps be mentioned in the discussion. Can iteratively reweighted least squares or another robust regression technique be used instead?

This is something I am investigating. Briefly, although iteratively reweighted least squares (IRLS) can be applied straightforwardly to conventional multiple regressions (where you have complete data), the same is not true when you have missing data (because then you have missing residuals too, and thus no obvious way to identify outliers). In principle IRLS can be used term-by-term to estimate each of the covariances in a multiple regression (in place of Eqs. 40 and 41), but the effects of doing so are not easy to determine a priori. As always, the choice between robust methods (like IRLS) and non-robust methods (like least squares) represents a tradeoff: with robust methods, you get reliable results even if your

data are messy, but you lose precision and sensitivity if your data are not messy.

I agree that this approach assumes no model for the transport of tracers, yet it does assume a model for the errors between the regression and the measurements (i.e. the residuals).

So does *any* data analysis method (although technically we are not minimizing a sum of squares, because we will always have missing data during rainless periods).

This is similar to the choice of an objective function in traditional model calibration, and deserves attention. For example, commonly used assumptions about streamflow residuals were shown to be often violated, because of autocorrelation, non-normality, and heteroscedasticity (Schoups & Vrugt, 2010). Are the tracer residuals in this work likely to show non-normality, autocorrelation, and heteroscedasticity as well?

This rhetorical question chooses to ignore the fact that I have *explicitly provided a framework* for quantifying the effects of autocorrelation.

In any case, having taught environmental data analysis for many years, I can reassure NR that I am quite familiar with the statistical assumptions underlying regression. I am also familiar with the rather large literature on how badly one needs to violate those assumptions in order to substantially affect the results (which is really the question in practice, not whether some theoretically ideal assumptions exactly hold or not - they almost never do). The general message from that literature is that standard regression is surprisingly robust unless its assumptions are very badly violated, or unless you are trying to make inferences about extremely improbable events (very small p-values), or unless you don't estimate standard errors (and keep them in mind in your interpretation), or unless you are overfitting (in which case you have lots of problems to worry about).

Regarding non-normality and heteroskedasticity, of course these could be

looked at (at the cost of greatly expanding the length and complexity of the paper), but the results would be highly assumption-dependent. What should you assume about your sampling and laboratory errors? What should you assume about the nature of the mis-match between the assumptions that underlie the method, and the behavior of the real-world system? One could rapidly get lost in a high-dimensional assumption space.

And, again, I will point out that if these issues are a problem for this approach, they are likely to be an equally bad (or even worse) problem for many other methods of estimating transit times.

Look, let's keep this in perspective. What *other* hydrograph separation or transit time methods have been tested as comprehensively as those that are presented in this paper? One can always ask to investigate an endless list of statistical conjectures. But why should other approaches get a free pass when they've hardly been tested at all?

Although CQ(j) and Cnew(j) are both "normalized" by subtracting CQ(j-1), there could be autocorrelation of higher order than just 1.

Of course, but the autocorrelation in the *variables* isn't relevant; it's the autocorrelation in the *residuals* that matters.

How does the variance of errors change with larger flashy events.

It's hard to give a general answer. But remember, these are concentrations, not water fluxes, so a lot will depend on whether big events have higher or lower tracer variance in precipitation.

3. P12, L19-20 All the benchmarking is done for a catchment without evapotranspiration.

That is false. See Section 3.6.

This points to a more general concern with the ensemble approach. No assumption is made explicitly about what happens to the tracer masses between precipitation and streamflow. This means that the method may try to find direct "connections" (in a loose statistical sense here) between tracer inputs from the past and current tracer fluctuations in the stream. Intermediary (unconsidered) processes may still be important to explain the transformation from one to the other. I especially think of processes affecting the lumped catchment tracer mass balance, which is an expression that was not considered in the approach. In that regard, how are the results expected to change if ET is actually used in the benchmark tests? Is the approach robust for real catchments where ET can be a major part of the water balance? Here I am not considering the effects of fractionation which were already dealt with, but the selective removal of certain tracer masses (associated with particular ages, i.e. different soil/groundwater mixtures) by ET, which will hence not be available for streamflow.

As the manuscript points out, this approach will determine the lagged fractions of whatever input is sampled (presumably precipitation, but could be generalized to multiple inputs) that appear in whatever output is sampled (presumably streamflow, but could be something else). There is no requirement that these are the only inputs and outputs. Specifically, Eq. 29 says that the discharge Q_j is the sum of contributions to discharge q_jk across each time lag k, but it does not say that this is the only way that water can leave the system (that is, the q_jk do not necessarily add up to P).

The method is not based on an input-output mass balance for the catchment. This implies that there should be no problem if (as often happens in nature) ET is a significant fraction of the water balance, and if the age distribution of the ET flux is different from the age distribution of the discharge flux. In such a case, of course, the age distribution of discharge will be different from what it would be with ET=0, but that will be reflected in tracer concentrations that correlate differently with precipitation. Thus the proposed method will estimate the age

distribution of the sampled output (in this case Q); it will not estimate the age distribution of the un-sampled output (ET).

4. P20, L4-6: These estimates seem to differ as much as 50% from the known values for the damped catchment and weekly data on figure 4.

Yes, by cherry-picking one comparison (from among 8 in Fig. 4), one can find a roughly 50% discrepancy (which is actually a small discrepancy between two small numbers). This concerns the average "forward" new water fraction (the blue lines in Fig. 4g). The true value from age tracking in the model is 0.11, and the estimate from ensemble hydrograph separation is 0.07. One way to look at this is that ensemble hydrograph separation underestimates the "forward" new water fraction by about 40% (of the true value), but the other way to look at it is that the discrepancy is 0.04 in absolute terms.

Putting the matter differently, what would be the likely error of an *a priori* "guesstimate" of the *forward* new water fraction in this system, without any formal analysis? How does a discrepancy of 0.04 look compared to *that*?

Many tracer data sets are at weekly resolution and come from "damped" catchments (e.g. Tetzlaff et al., 2009; Pfister et al., 2017). Data-driven approaches are by nature highly sensitive to the quality of data (e.g. variability, resolution, and measurement uncertainty). The proposed approach could thus show some limitations due to its strict data needs in some cases. This could be mentioned in the discussion.

I don't know what NR means by "strict data needs". Strict compared to what? I've demonstrated that the method yields quantitatively realistic estimates across a wide range of catchment behaviors, with both weekly and daily data (that contain both errors and gaps). Thus the data needs here are considerably *less* strict than those of many other approaches. I will also note that among the few other benchmark tests that have been published, some make the remarkably unreal-

istic assumption that the input data (and sometimes also the output data) are completely error-free.

Thus it's not clear how to respond to NR's comment, beyond noting that of course any empirical approach requires data, and thus will depend on the quality of the data. (Conversely, any approach that does not depend on data, and thus is free of data quality concerns, probably doesn't teach us much about the real world!).

5. P20, L19-20: A weekly sampling routine is likely to contain more "baseflow" samples which reflect older water contributions.

That is not correct. In regular weekly sampling, baseflow will be sampled proportionally to its frequency of occurrence, because regular sampling does not preferentially include or exclude baseflow samples.

This results in an underestimation of QFnew as shown here.

The premise of that statement is incorrect. The underestimation in the figure has nothing to do with preferential oversampling of base flow, because that oversampling did not occur.

Yet how can the fraction of new water with respect to discharge be under- estimated while the fraction of new water with respect to precipitation is overestimated? These quantities refer to the same mass of water in streamflow.

Yes, they are the same mass of water, but expressed as a ratio to two different things (discharge and precipitation). One can be overestimated while the other is underestimated because uncertainties propagate differently in Eq. 9 vs. Eq. 28.

6. P28, L2: Here it is assumed that values of Cp and CQ at all times corresponding

to indexes j or j-k are known (except a few, which require the solutions proposed in 4.2 and 4.3).

That is not correct. There is no requirement that Cp or CQ is sampled at all times, or that the number of missing values is just "a few". The benchmark tests here assume 5% missing values from rainfall and streamflow, in addition to the much larger number of rainfall samples that are missing because there was not enough rain to make an isotope measurement. One can change the percentage of missing values from 5% to 10%, 20%, or more, of course. The uncertainties will grow accordingly (depending in part on how large m is).

In practice it is very likely that the sampling interval is irregular, such that there is not a perfect correspondence between measurement times, and required times indexed by j or j-k. Any recommendation on how to best adjust the measurement time series so that these terms are defined properly would be welcome. Similarly, the method requires the same number of measurements in precipitation and in streamflow.

That is not correct. Nothing in the text or the math states this or implies it. The calculations presented here assume for convenience that the time bases of the measurements are the same (that is, if precipitation is sampled weekly then streamflow is also sampled weekly). But there is no requirement that even if (for example) streamflow is sampled every week, precipitation must be sampled every week.

How could we deal with this in various research catchments as this is often not the case?

The only general answer is to say that one can use benchmark tests to look at the effects of various scenarios of missing data. It is difficult to generalize about the wide range of possible scenarios.

7. P29, L12: Here it is assumed that the most recent precipitation events have more

weight in the current tracer fluctuations in the stream than older inputs.

That is not correct. The math does not assume this, the text does not say this, and counter-examples are presented (See, for example, Figure 12b-c).

This is implicitly reflected by the truncation of the sum in equation 30.

That is not correct. The truncation of the sum does not assume that the remaining terms are necessarily small.

This is also reflected by the estimated travel times that mostly stay below a few months. Although this assumption about tracer contributions is likely to be valid, catchment travel times are known to be generally in the order of magnitude of a few years and even decades (McGuire & McDonnell, 2006). This is all the more true when age estimates are based on tritium measurements (Stewart et al., 2010). Would the ensemble approach be robust in catchments where streamflow is volumetrically dominated by water older than a few months?

The mean transit times reported by McGuire and McDonnell (2006) are mostly *baseflow* mean transit times, whereas Fig. 3 reports mean transit times *only for rainy days* (so that these can be directly compared to the new water fraction). The distinction is important, because baseflow mean transit times will always be longer (sometimes much longer) than mean transit times that are averaged over all flows, which in turn will always be longer than mean transit times averaged over rainy days.

For example, the mean transit time as shown in Fig. 3 for the benchmark model with the "flashy" parameter set is 189 days (or about six months), but the *baseflow* mean transit time (defined for these purposes as the mean transit time when precipitation has been less than 1 mm/day for the previous three days) is 563 days, almost *three times longer*. Among the Monte Carlo parameter sets underlying Figures 2, 3, 5, and 10, baseflow mean transit times range as high as

three years or more.

8. P29, L12: Regarding the linear algebra, how large can the truncation index m be in practice, given that computationally intensive large matrix operations are carried out? This is especially of interest since the matrices grow with the number of measurements in both dimensions, while m needs to be as large as possible for the ensemble method to work well.

The premise is false. Nowhere does the paper say that m "needs to be as large as possible for the ensemble method to work well", for the simple reason that it isn't true. If m is too large, the TTD estimates will become too uncertain (and the standard error estimates will show this). Conversely, the ensemble method works very well for m=1 (that is, calculations of Fnew), thus rather clearly demonstrating that there is no need for m to be large.

Solving matrix problems is computationally intensive, but not at a scale that matters for this problem (geophysicists routinely solve matrices that are orders of magnitude larger than those that will be relevant here). Even an excel spreadsheet can do these calculations for matrices with dimensions of m$>$100 (I know because I've done it), and scripting languages like R or python invoke fast low-level solvers that can efficiently handle much, much larger problems. The practical limitation on m will be data, not computer power.

In my opinion, the discussion should encourage the reader to consider if this approach shows limitations for his/her considered travel times, which may be up to a decade. Can this approach go beyond the left-hand tail of the TTDs or is it limited to the left tail?

By definition the approach handles the left tail of the distribution, because it estimates TTDs to some maximum lag m. The question is how much of the interesting behavior of the TTD is within the range of m for which the TTD can be reliably

estimated. That will depend on many factors, including (a) what behaviors one considers to be interesting, (b) how many tracer samples are available, and at what frequency, and (c) perhaps most importantly, the timescales of variability in the input tracer time series (for any fluctuation-based analysis, the input must be variable on the relevant time scales). These limitations are inherent in the use of stable isotope tracers. There is no intrinsic time limit in the method itself.

9. P33, L8: Over several years of data, doesn't neglecting 1 mm of precipitation per day sum up to a large value? It could be useful to include some discussion on the effect of that threshold on the results. Are the results highly sensitive to that choice or not?

Remember, the method does not rely on mass balances (this is essential, because any method based on mass balance - and there are many - will be inherently vulnerable to biases from ET and from the un-representativeness of precipitation measurements). Neglecting small precipitation inputs has almost no effect on the results, except for "forward" new water fractions calculated by Eqs. 21 and 22 under certain circumstances (where a precipitation threshold is needed to avoid giving huge weight to tiny precipitation inputs). Remember, too, that the real-world catchment will also "neglect" small precipitation inputs because they will typically evaporate from the canopy or ground surface, and thus contribute next-to-nothing to streamflow and stream tracer concentrations.

10. P39, Figure 11: Deviations between the benchmark TTDs and the estimated ones are visible here. How could these deviations be described more quantitatively to be more objective?

Of course one could quantify the deviations in terms of a root-mean-square error or median absolute deviation, if one wanted to. This seems obvious enough

that it doesn't need to be explicitly stated. The deviations are generally within the reported standard errors, so there is not much to be learned from them (because any measurements of the deviations would themselves have uncertainties of roughly 100%).

11. P40, Figure 12: It looks like the uncertainties are larger for the TTDs which shape is not a classical "L" anymore. The explanation given here is that the effective sample size neff (equation 13) is small because of tracer autocorrelation. Can we not say that an autocorrelation in the tracer time series is universal, as well as a shape of the TTD far from a simple "L"? This seems related to the issue described in comment (2).

We certainly *cannot* say that TTD's are universally far from L-shaped. In cases where TTD shapes have been evaluated with methods that are sensitive to the shape of the TTD (rather than with methods that just assume a given shape a priori), L-shaped distributions appear to be quite common in real-world watersheds.

The actual uncertainties (as reflected in the scatter clouds surrounding each data point) are not much larger for the humped distributions; they just look larger because the axis scales are different. However, the standard errors (the error bars) are indeed overestimated for these humped distributions. As the manuscript says, I think this has to do with the estimation of n_eff.

12. There are not many data-driven methods that can yield non-parametric TTDs, which explains why this new approach is really beneficial to estimate TTDs. Yet, I believe that the problem solved here is somewhat similar to what Turner et al. (1987) solved as well, using Kalman filtering approaches (see the parallel between equations 30 and 31 here, and their equations 1 and 2) (see also Turner

and McPherson, 1990). They unfortunately did not detail the math behind their approach. Nevertheless, their work present time-varying average transit times, including uncertainties, also derived without assumptions on the shape of the TTDs. This is worth mentioning and comparing to the presented approach in the discussion.

NR is correct that Turner et al.'s equations 1 and 2 and my equations 30 and 31 both express convolutions, but convolutions also underlie essentially every method for using tracer data to infer transit times. As NR notes, it is not very clear what Turner et al. actually did. Nonetheless, it's clear that their approach is not really nonparametric; instead they have apparently fitted a parametric model whose parameters are allowed to vary smoothly over time, as estimated with the Kalman filter. Nonetheless, the fact that their fits to the stream isotope data had R^2>0.999 suggest that the underlying model was massively overfitted.

Furthermore, Klaus et al., (2015) also presented a data-driven approach that could be worth mentioning and briefly comparing to the presented one.

The Klaus et al. analysis seems to me to be a calibrated model rather than a data-driven approach. As far as I can tell it appears to assume mass conservation between input and output (no ET), which, as NR notes, would be problematic in real-world applications. The "uniform mixing distribution" assumption also seems nonphysical, since it "gives *higher* weight to *less* frequent tracer concentrations" (emphasis added).

But most notably, Klaus et al.'s "proof of concept" simply runs their model in forward mode to generate a test time series, and then runs *exactly the same model* in inverse mode, on exactly the same time series (without even introducing any measurement errors). This only demonstrates the mathematical consistency of the forward and inverse models. It does essentially nothing to demonstrate that the inverse model will give realistic results when applied to real-world

data (which will *not* come from the same model, and which *will* have errors). Nonetheless, Klaus et al. conclude that, "The virtual and modeled ^{18}O time series matched exceptionally well...", apparently not recognizing that this result is mathematically inevitable because they have tested their model against itself.

Finally Kim et al. (2016) could estimate not only TTDs but also SAS functions from artificial tracer data (of course under well controlled lab conditions). Their work is worth mentioning because they are able to distinguish the " external" variability of travel times from the " internal" one, unlike the ensemble approach presented here.

The method used by Kim et al. cannot be applied to real-world field conditions, and thus is not relevant to the topic of the present paper.

13. P54, L30-32: Doesn't this mean that m should be set bigger?

Setting m bigger does not make this issue go away. One really should just take the last few lags with a grain of salt, as the manuscript says. Such edge effects often arise in various inversion problems.

Small technical comments:

(14) Figure 1 & 4: the colors are not consistent between the legend and the lines in the lower subplots.

This is intentional. The problem is that if one uses a yellow color in the legend that is as light as the yellow in the figure, the line in the legend almost disappears, whereas if one uses a dark enough yellow in the figure that it can also be seen in the legend, then the figure becomes muddy.