We thank reviewer #2 for these comments on our manuscript. Below we respond (in bold type) to the reviewer's comments (in normal type).

The Technical Note: "Calculation scripts for ensemble hydrograph separation" by Kirchner and Knapp, presents an ensemble hydrograph separation tool, useful to estimates new water fractions and transit time distributions (TTDs). The authors developed user-friendly scripts that perform EHS calculations in two broadly used platforms (MATLAB and R).

The authors used an impressive synthetic data set, that despite the limitations they clearly stated in the manuscript, mimics reasonably the real word behavior of isotope time series.

The authors made an important contribution to the scientific community by helping to solve the common problem of lack of monitored/non-stationary end end-members while performing hydrograph separation. Moreover, they put great effort into describing the method, providing examples, and addressing uncertainties issues. I was delight by reading this technical note that is well-structured and clearly written.

Some of my main suggestions matched those of Reviewer RC1 (specifically related to IRLS and the overestimation of Fnew when the TTDs are humped) and were already clarified by the authors by including them as supplementary material.

I found this work in very good form and suggest the Editor accept this publication after a single additional clarification.

Thanks very much for these comments.

L 380-392 Could the authors please further explain the mismatch between the discharge age tracking using the benchmark model and the new ensemble hydrograph separation? As well as the potential implications for sampling size and frequency. This will be useful for users who will apply the method with real-world data.

As we indicated in our response to reviewer #1, it is difficult to generalize here. What we show in Fig. 11, and comment on in lines 380-392, are results for one specific set of benchmark model parameters, and for the particular time series of precipitation fluxes and tracer concentrations shown in Fig. 1. It would of course be interesting to undertake a more systematic exploration of how the sample size and frequency, as well as the various benchmark model parameters and the characteristics of the precipitation time series, all affect the uncertainties and errors in ensemble hydrograph separation estimates.

But that would be an entirely different (and probably much longer) paper. The point of this paper is to make the codes for the method publically available and to briefly illustrate their possible uses and limitations. We should also keep this matter in perspective: both in the earlier paper (Kirchner, 2019) and in the present manuscript, we have already tested this method more rigorously than many others have been tested.

What we can do is to expand the discussion in Section 4.5 to also include some comments on sample size, based on our recent experience applying ensemble hydrograph separation to real-world data from Plynlimon (Knapp et al., 2019). In the revised manuscript, we will therefore begin Section 4.5 with the following:

"Prospective users of ensemble hydrograph separation may naturally wonder what sample sizes and sampling frequencies are needed to estimate new water fractions and transit time distributions. The answers will depend on many different factors, including the time scales of interest to the user, the desired precision of the F_new and TTD estimates, the logistical constraints on sampling and analysis, the frequency and intermittency of precipitation events, the variability of the input tracers over different time scales, and the time scales of storage and transport in the catchment itself (that is, what the TTD is and how non-stationary it is, which of course can only be guessed before measurements are available). Ideally one should sample at a frequency that is high enough to capture the shortest time scales of interest, and sample much longer than the longest time scales of interest. One should also aim to capture many diverse transport events, spanning many different catchment conditions and precipitation characteristics.

Beyond these generalizations, it is difficult to offer concrete advice. We can, however, report our recent experience applying ensemble hydrograph separation to weekly and 7-hourly isotope time series at Plynlimon, Wales (Knapp et al., 2019). We were generally able to estimate TTDs out to lags of about three months based on four years of weekly sampling. The same four years of weekly samples yielded about 100 precipitation-discharge sample pairs (after samples corresponding to below-threshold precipitation were removed), which were sufficient to estimate weekly event new water fractions with an uncertainty of about 1% (e.g., $^{\mathbf{Qp}}\boldsymbol{F}_{\mathbf{new}}\sim$ 8±1%). When these were split into four seasons, we could estimate event new water fractions with an uncertainty of about 2-3% using 20-30 weekly precipitation-discharge pairs, and when they were split into 4-6 different ranges of precipitation and discharge, we could reasonably well constrain the profiles of new water response to catchment wetness and precipitation intensity (Fig. 10 of Knapp et al., 2019). We were able to estimate 7-hourly TTDs out to lags of 7 days based on about 17 months of 7-hourly isotope samples, including almost 1500 discharge samples and 540 above-threshold precipitation samples, and splitting these data sets in half allowed us to distinguish the TTDs for summer and winter conditions (Figs. 11 and 12 of Knapp et al., 2019). However, these numbers should not be uncritically adopted as rules of thumb for other catchments, since precipitation at Plynlimon is frequent and weakly seasonal, and the catchment is characterized by rapid hydrological response but relatively long storage timescales (Kirchner et al., 2000). All of these characteristics could potentially affect the sample sizes needed for estimating new water fractions and transit time distributions. As more experience is gained at more catchments, general rules of thumb may emerge. Until then, however, benchmark tests like those described here can potentially provide a more reliable site-specific guide to sample size requirements."