(1) Reviewer Comment:

The study fits the scope of HESS and makes a valuable contribution to the field of transit time modelling and tracer hydrology. Illustrating the capacity of stable water isotopes to quantify older water will open up new opportunities for TT modelling in catchments that are assumed to show comparably large MTTs. Hence, I support the general motivation and objectives of the study.

Reply:

We highly appreciate this positive overall assessment of our work and we thank the reviewer for her interest in our work as well as for the thoughtful and detailed comments that helped to strengthen our analysis. Below, we provide clarifications and our perspectives to respond in detail to the individual reviewer comments.

(2) Reviewer Comment:

First, I am not sure whether a catchment (river basin) of 13,000 km² with at the same time limited availability of tracer data is the best choice for the study objectives. While individual controls on TTs remains largely elusive, it has been shown that TTs (or their metrics) vary widely depending on catchment characteristics such as elevation, topography or climate (e.g., Jasecko et al., 2016, Kumar et al., 2020). Modelling TTs in a river basin that shows a gradient of more than 800 mm yr-1 in annual precipitation, an elevation gradient of around 900 meters and varying land use types adds a lot of complexity that could have been avoided when using a much smaller and more homogeneous catchment. At the same time, the study relies on only one precipitation station for both stable water isotopes and tritium (within the basin) providing monthly composite samples. Hence, the tracer data are rather sparse both temporally and spatially, which adds another layer of uncertainty to the modelling. An alternative might be to compile data from previous TT modelling approaches that have been conducted in smaller catchments with more highly-resolved (space and/or time) stable water isotope and tritium data (e.g., Rodriguez et al., 2021 – reference already in manuscript).

Reply:

Choice of study region

We agree that it remains a defining challenge in hydrology to fully account for heterogeneities in larger systems. Unfortunately, there is no "silver bullet" to solve that problem. This is also explicitly discussed in the Discussion section of our manuscript (p.21, l.658ff). While we share the reviewer's view that studies at smaller scales are very important, these types of studies typically suffer from other limitations. Specifically for the case of stable isotope and tritium comparisons and apart from the fact that there are hardly any catchments world-wide in which data for both tracers are available, the study cited by the Reviewer (Rodriguez et al., 2021) is indeed conducted in a smaller catchment with higher tracer sampling frequency. *However*, and as explicitly mentioned in the manuscript (p.4, l.132-150), it relies on much shorter time series, i.e. 2 years, and only a handful of tritium samples, i.e. 24. In addition, conclusions from that study on the ability of stable isotopes to see older water may be hampered by the potential *absence* of older water. In other words, if there is no older water present in a catchment, stable isotopes

can also not see it, as recently pointed out by Stewart et al. (2021). We therefore believe, that in spite of potential uncertainties arising from the size of the system, our study allows us to explore aspects of the research question that could not (or not fully) be addressed by Rodriguez et al. (2021).

Role of heterogeneity for older water ages – catchment as low-pass filter

It is also important to note that in our study we are mostly interested in older water ages. As catchments act as low-pass filters, they already smooth out much of short time-scale and small spatial-scale hydro-climatic variability. The remaining higher-frequency components in the response, e.g. responses to individual rain events, then mostly affect water ages at the younger side of the spectrum. These can indeed be sensitive to spatial-temporal heterogeneities. In contrast, older water ages are mostly controlled by low frequency components of the system and thus variabilities at much larger spatial and longer temporal scales, e.g. seasonal or inter-annual changes in groundwater tables, and are thus much less sensitive to small-scale heterogeneities. This can for example be seen in the significant differences between the power spectra of stream tracer concentrations of fast responding parts of the system (i.e. short time-scales, high-frequency components and thus younger water ages) and groundwater tracer concentrations (i.e. much longer time-scales, low-frequency components of the system and older water ages), as for example demonstrated by Hrachowitz et al. (2015; Figure 8 therein) and which define the recurrently described, very characteristic 1/f scaling of stream tracer responses across many system in contrasting environmental settings across the world (e.g. Kirchner et al., 2001; Godsey et al., 2009; Hrachowitz et al., 2009; Aubert et al., 2014; Kirchner and Neal, 2013). Another piece of evidence for the lower sensitivity of older water to heterogeneity is the higher sensitivity of high-frequency components and younger water ages to hydro-climatic variability (e.g. Figure 9 in our original manuscript) as compared to the almost complete lack sensitivity to hydro-climatic in low-frequency components and thus older water (e.g. Figure 10), which has also been reported in many other studies (e.g. Hrachowitz et al., 2013, 2015; Soulsby et al., 2015). Overall, this means that while the pattern and dynamics of young water ages may indeed to some degree be affected by heterogeneities within our study basin, it is plausible to assume that they have only minor impact on the estimation of older water ages. *We will add a more detailed discussion on this in the revised version of the manuscript.*

Spatial representation of hydro-climatic and tracer input heterogeneity in the study

Notwithstanding the above and to limit adverse effects of a coarser data resolution, we here invested considerable effort into spatial adjustments of hydro-meteorological input data as well as tracer data, according to the best available information in our distributed model implementation. While the major spatial differences in precipitation are accounted for by the identification and use of four individual precipitation zones, major spatial differences in temperature (and thus also in EP) are accounted for by the additional stratification into 100m elevation zones as described in Sections 3.2.1 and 4.2.2. Similarly and more importantly, the tracer input signals were spatially adjusted, as described in Sections 3.2.2 and 3.2.3 as well as in the Supplement, following the method recently developed by Allen et al. (2018, 2019). This method identified strong relationships between multiple catchment characteristics and seasonal stable isotope signals in precipitation. These relationships thus allow a robust estimation of the spatial differences in stable isotope input, both globally (Allen et al., 2019) and perhaps more importantly, also

regionally, as demonstrated in Allen et al. (2018) who quantified spatial stable isotope input for Switzerland, which is just across the border from our study basin in Southern Germany. A comparable approach was applied for precipitation tritium concentrations, which in any case do not exhibit major spatial differences (e.g. Schmidt et al., 2020). The same applies also to water stable isotopes in precipitation for monthly sampling resolution as indicated by the similarity to isotopes for stations close by, i.e. Karlsruhe (Stumpp et al. 2014).

Ability of the model to represent the response and spatial heterogeneity therein

To reduce the potential of misrepresentations of the system and its heterogeneities by the model we have deliberately chosen to expose the model to a rigorous calibration and post-calibration evaluation procedure that goes far beyond what is done in the vast majority of studies in scientific hydrology. The use of *eight different performance indicators*, that describe the models' ability to simultaneously reproduce distinct signatures and thus distinct aspects of the system response, allowed to identify and discard solutions that in traditional model calibration/evaluation procedures, based on one or two performance metrics, would have been falsely accepted as feasible. This leads to a robust representation of the system, as can be seen by the models' ability to relatively well and simultaneously reproduce these multiple signatures – both, in the calibration as well as and more importantly in the post-calibration evaluation ("validation") periods as illustrated by Figures 3-5 and Table 4 in the original manuscript and also illustrated here below in Figure FR1, for the example of stream flow Q in Scenario 12.

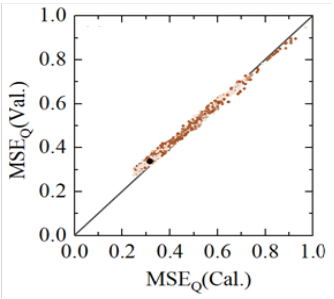

Figure FR1. Model performance of all pareto-optimal solutions accepted as feasible against to reproduce stream flow Q in model calibration vs. model evaluation periods based on the mean squared error (MSE_Q). The dark dot indicates the most balanced pareto-optimal solution. The fact that all solutions plot very close to the 1:1 line suggests that the model does reproduce Q in the model post-calibration evaluation period ("validation") almost as good as in the calibration period. This is a strong indicator of the model being a plausible representation of the system response.

However and in addition to the strict model evaluation procedure in our original manuscript, we have taken this concern of the reviewer very serious and decided to *confront the model with additional observations* to further test its ability to meaningfully represent spatial differences in the response. To do so, we have now also evaluated the model outputs against streamflow observations in three sub-catchments (C1: Kirchentellinsfurt, C2: Calw, and C3: Kocherstetten) within the Neckar basin, whereby each one of them largely represents the response from one of the precipitation zones (Figure FR2 here below).

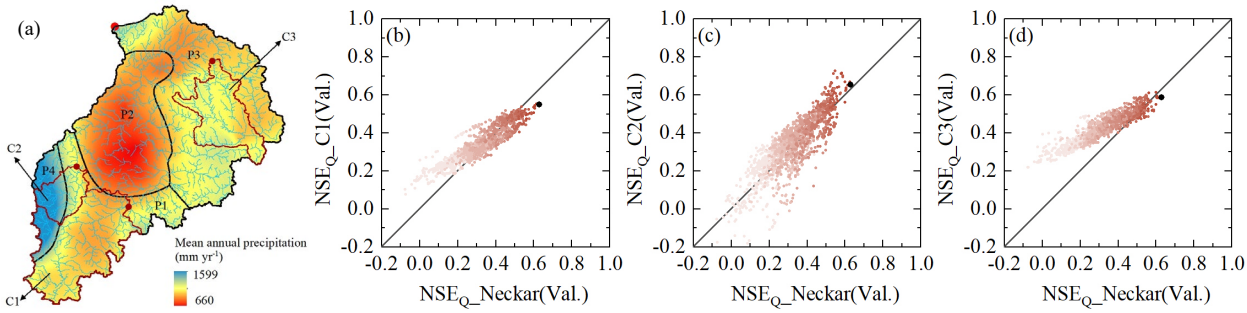

Figure FR2: (a) Sub-catchments C1 – C3 within the Neckar basin used to evaluate the model performance, (b) model performance in the Neckar basin vs. sub-catchment C1, (c) Neckar vs. C2 and (d) Neckar vs. C3, based on Scenario 10. The dots indicate all Pareto-optimal solutions in the multi-objective model performance space. The shades from dark to light indicate the overall model performance based on the Euclidean Distance D_E, with the darker solutions representing the overall better solutions (i.e. smaller D_E)

It can be seen, that the model calibrated on stream flow of the entire Neckar basin can reproduce stream flow in the 3 sub-catchments similarly well, with C2 and C3 even better reproduced with many of the solutions than the calibrated Neckar stream flow. These results suggest that the model does indeed pick up the major differences in response types due to hydro-climatic heterogeneities throughout the Neckar basin. Together with the spatial adjustments of the tracer inputs as described above, this is further evidence that the model provides an adequate representation of the major features of the hydrological response even at the larger scale of the Neckar basin and therefore also a meaningful spatial representation of the tracer circulation. We will add these additional model tests to the manuscript to better demonstrate the suitability of the model for our study.

Overall, we can and do not claim that our models generate the best possible TTD estimates. Rather, our intention in this analysis is to show the consistency between TTD estimates derived from stable isotopes and tritium, i.e. that both contain enough and comparable information which can be exploited to estimate water ages. In other words, even if TTD estimates of both tracers are subject to uncertainties, the fact that they provide similar TTD estimates when used in the same model type is evidence for a similar information content, supporting the notion that stable isotopes have indeed the potential to see older water, if used in conjunction with suitable modelling approaches. This is explicitly discussed in the text (p.19, l.600ff in the original manuscript).

(3) Reviewer Comment:

Secondly, there is a remarkably great difference in model complexities between the individual TT modelling approaches. On the one hand, simple CO models with only one compartment, no temporal/seasonal variation and two pre-defined shape parameters for the TTs have been used, while on the other hand, the SAS model consists of three hydrological response units with multiple storage volumes each, has 11 calibration parameters and is also tested in a spatially distributed implementation. As the authors are

clearly aware of, time-variant concepts of CO models (see Hrachowitz et al., 2010; and references cited therein) as well as multi-compartment models representing fast and slow flow routes have been used; using especially the latter is a common approach in CO modelling. Moreover, the SAS model with its comparably large number of parameters is calibrated simultaneously to discharge and at least one of the two tracers, while the CO models are calibrated to only one tracer. I am thus wondering to what extent results from these TT models can be compared at all. I understand that the objective of this paper is not to dismiss a specific model type, but rather to analyse the flexibility of stable water isotopes as TT model tracers. However, this requires to use model setups and data similar to those used in the papers that have demonstrated the truncation of TT distributions by calibration to stable water isotopes. To address this concern, one could think of (i) focussing on a smaller (or even headwater) catchment with preferably daily tracer data, (ii) using established CO models such as the more complex ones in Stewart et al. (2010), and (iii) using measured and modelled P, ET, storage and Q data as input for SAS modelling (potentially with non-random sampling) with one or a maximum of two SAS function compartments, as commonly done in more recent SAS modelling studies (e.g., Benettin et al., 2017; Harman, 2015; Nguyen et al., 2021).

Reply:

We agree with the reviewer that the model approaches are different and we also agree that comparisons need to be consistent and systematic to be meaningful.

However, we also want to point out here – as correctly mentioned by the reviewer – that the objective of our analysis is to analyse the potential of stable isotopes to see older water and not a full-fledged comparison of different model approaches. This is explicitly stated in the research hypothesis "[…] that ^{18}O as tracer generally and systematically cannot detect tails in water age distributions and that this truncation leads to systematically younger water age estimates than the use of ^{3}H" (p.5, l.151-152)

Please note that therefore what is actually compared here are models of the *same type* (and same complexity) run with stable isotopes and subsequently with tritium. The comparison is not made between models of different types and/or complexities. In other words, we compare water age estimates obtained from e.g. a CO model with exponential TTD run with ^{18}O with those obtained from the same model but run with ^{3}H. In contrast, we do not compare water ages from that CO model with ages estimated from another model, e.g. IM-SAS. This is also emphasized by the last four columns of table 5.

To further clarify, we have estimated water ages based on CO models to check if we would find differences in water ages between ^{18}O- and ^{3}H-based model runs in the study basin, using the *same* types of *lumped*, *time-invariant* models that Stewart et al. (2010) based their argument on. The fact that we found significant differences between these estimates, would, without further analysis, further support the observation of Stewart et al. (2010) that ^{18}O *generally* truncates water ages.

Our intention is *not* to show that CO models are generally not capable to estimate older ages. Perhaps, time-variant implementations can do that very well, but exploring this was not the objective of our study. Also the combined use of ^{18}O and ^{3}H in CO models has previously been shown to be useful to estimate older ages. But this is outside the scope of our study. Instead, as clearly stated in the research hypothesis, we test if ^{18}O can generally be considered to be useless for the determination of ages older than ~4 years. Our results then further suggest, that, if used in combination with IM-SAS models, the hypothesis needs to be rejected, as these models produce similar water ages with ^{18}O and ^{3}H that are much older than 4 years. Given that the results of Stewart et al. (2010) as well as our own CO scenarios are based on lumped,

time-invariant CO model implementations, our results eventually also allow the observation that the perceived failure of ^{18}O to see older ages is not a general limitation of that tracer, but rather a consequence of its use in ***lumped***, ***time-invariant*** CO models.

However, we agree with the reviewer that we have not tested the more complex CO model implementations from Stewart et al. (2010) in our original manuscript. We therefore took up this advice of the reviewer and did additional model runs, with full calibrations (and evaluations) of a wider range of common time-invariant implementations of CO models, also including more complex ones. Our analysis now includes in addition to exponential (EM) and gamma (GM) models also two parallel reservoir (2EM; scenarios X1-2), three parallel reservoir (3EM; scenarios X3-4) and exponential piston flow (EPM, scenarios X5-6) implementations. The TTD estimates from these additional model implementations are consistent with those in the original analysis: for all tested lumped, time-invariant CO models, the TTDs derived from ^{18}O indicated with MTTs ~ 1-2 yrs significantly younger water than those derived from ^{3}H, which suggest MTTs ~10 yrs throughout (see Table TR1 and Figure FR3 below). This further strengthens our previous results, suggesting that ^{18}O when used in lumped, time-invariant CO models underestimates water ages, as suggested by Stewart et al. (2010).

Table TR1. Metrics of stream flow TTDs derived from the 10 model scenarios with the different associated calibration strategies based on different CO models, where $C_{\delta^{18}O}$ indicates calibration to $\delta^{18}O$, $C^{3}H$ calibration to ^{3}H. The TTD metrics represent the best fits of the respective time-invariant TTD. The water fractions are shown as the fractions of below a specific age T. The columns with absolute difference Δ illustrate the differences in TTDs from the same models calibrated to $\delta^{18}O$ and ^{3}H, respectively. The subscripts indicate the scenarios that are compared (e.g., $\Delta_{3,4}$ compares scenarios 3 and 4).

Scenario		3	4	5	6	X1	X2	X3	X4	X5	X6	$\Delta_{3,4}$	$\Delta_{5,6}$	$\Delta_{X1, X2}$	$\Delta_{X3, X4}$	$\Delta_{X5, X6}$
Model		CO-EM		CO-GM		CO-2EM		CO-3EM		CO-EPM		Absolute difference				
Calibration strategy → TTD metrics ↓		$C_{\delta^{18}O}$	$C^{3}H$	$C_{\delta^{18}O}$	$C^{3}H$	$C_{\delta^{18}O}$	$C^{3}H$	$C_{\delta^{18}O}$	$C^{3}H$	$C_{\delta^{18}O}$	$C^{3}H$	$\Delta TT_{\delta^{18}O\text{-}^{3}H}$ $\Delta F(T<x)_{\delta^{18}O\text{-}^{3}H}$				
Percentiles (yr)	Mean (yr)	1.4	10.4	2.4	9.7	1.9	9.5	2.1	9.4	1.8	10	-9.0	-7.3	-7.6	-7.3	-8.2
	10^{th}	0.1	1.1	<0.1	0.3	<0.1	<0.1	<0.1	0.9	1.0	1.1	-1.0	-0.2	0.0	-0.8	-0.1
	25^{th}	0.4	3.0	0.2	1.3	0.2	0.3	0.2	2.8	1.1	2.9	-2.6	-1.1	-0.1	-2.6	-1.8
	50^{th} (median)	1.0	7.2	1.0	5.0	1.1	3.6	1.3	7.3	1.5	7	-6.2	-4.0	-2.5	-6.0	-5.5
	75^{th}	1.9	14.4	3.2	13.1	2.7	13.8	3.1	15.0	2.2	13.9	-12.5	-9.9	-11.1	-11.9	-11.7
	90^{th}	3.2	26.3	6.8	25.4	4.8	27.3	5.6	25.6	3.0	23.1	-23.1	-18.6	-22.5	-20.0	-20.1
Water fractions (%)	F(T<3 m)*	16	2	28	10	26	25	25	3	0	2	14	18	1	22	-2
	F(T<6 m)	30	5	38	14	34	34	32	6	0	5	25	24	0	26	-5
	F(T<1 yr)	51	9	50	21	47	40	44	10	13	9	42	29	7	34	4
	F(T<3 yr)	88	25	74	39	78	48	74	26	90	26	63	35	30	48	64
	F(T<5 yr)	97	38	85	50	91	55	88	38	99	39	59	35	36	50	60
	F(T<10 yr)	100	62	95	68	99	68	98	60	100	63	38	27	31	38	37
	F(T<20 yr)	100	85	100	85	100	84	100	84	100	86	15	15	16	16	14

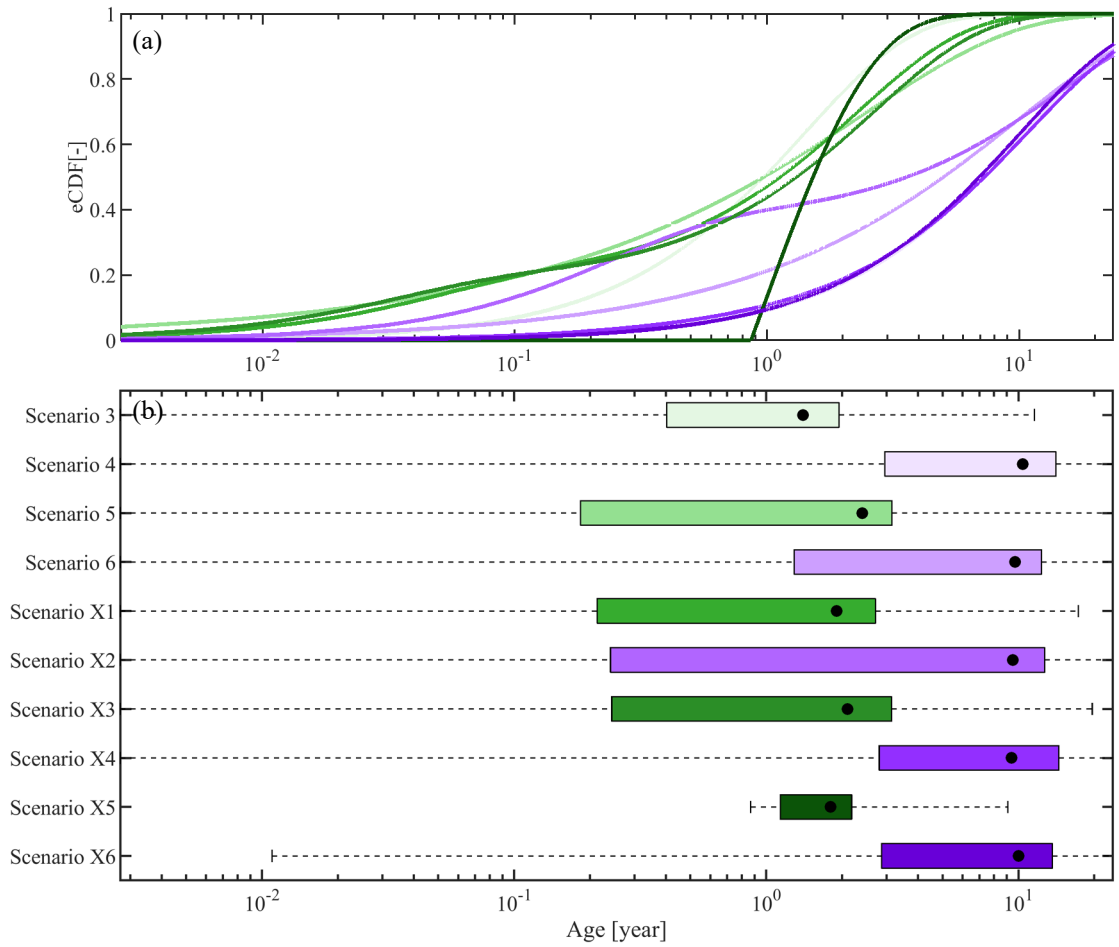

Figure FR3. Stream flow TTDs derived from the 10 model scenarios with the different associated calibration strategies based on different CO models. The TTDs represent the best fits of the respective time-invariant TTD. Green shades represent the TTDs inferred from $\delta^{18}O$ based on different CO models (from lighter to darker for scenarios 3, 5, X1, X3 and X5) in (a) and (b); the purple shades represent TTDs inferred from ^{3}H based on different CO models (from lighter to darker for scenario 4, 6, X2, X4 and X6); the black dots in (b) indicate the mean transit time for each model scenario.

In addition, and as requested by the reviewer, we have also included a "pure" SAS implementation (scenarios X7-9) with one compartment as described in Benettin et al. (2017), using observed Q to account for storage variations (as opposed to modelled Q in the IM-SAS implementations in scenarios 7-12) and one power-law shaped SAS function to route tracers through the system. Also, the results from this model implementation supports our original interpretation: the SAS model, similar to all other IM-SAS implementations (scenarios 7-12), provides similar TTDs for ^{18}O and ^{3}H. Both estimates are with MTT ~ 11 yrs also broadly consistent with the higher MTTs obtained from the other IM-SAS implementations (see Figure FR4 and Table TR2 here below).

Overall, all results and TTD estimates from additional model implementations are highly consistent with our previous results and considerably strengthen our conclusions to reject the hypothesis that stable isotopes underestimate water ages. We will add all additional model scenarios in the revised manuscript.

Table TR2. Metrics of stream flow TTDs derived from the 9 model scenarios with the different associated calibration strategies based on different SAS models, where $C_{\delta^{18}O}$ indicates calibration to $\delta^{18}O$, C^{3}_{H} calibration to ^{3}H, while $C_{\delta^{18}O,Q}$, $C^{3}_{H,Q}$ and $C_{\delta^{18}O,^{3}H,Q}$ indicate multi-objective, i.e. simultaneous calibration to combinations of $\delta^{18}O$, ^{3}H and stream flow. The TTD metrics represent the mean and standard deviations of all daily streamflow TTDs during the modelling period 01/10/2001 – 31/12/2016 are given. The mean transit time was estimated by fitting Gamma distributions to the volume-weighted mean TTDs of each individual scenario. The water fractions are shown as the fractions of below a specific age T. The columns with absolute difference Δ illustrate the differences in TTDs from the same models calibrated to $\delta^{18}O$ and ^{3}H, respectively. The subscripts indicate the scenarios that are compared (e.g., $\Delta_{7,8}$ compares scenarios 7 and 8). *Note that the fraction of water younger than 3 months is comparable to the fraction of young water as suggested by Kirchner (2016).

Scenario		7	8	9	10	11	12	X7	X8	X9	$\Delta_{7,8}$	$\Delta_{10,11}$	$\Delta_{X7,X8}$
Model		IM-SAS-L			IM-SAS-D			P-SAS			Absolute difference		
Calibration strategy → TTD metrics ↓		$C_{\delta^{18}O,Q}$	$C^{3}_{H,Q}$	$C_{\delta^{18}O,^{3}H,Q}$	$C_{\delta^{18}O,Q}$	$C^{3}_{H,Q}$	$C_{\delta^{18}O,^{3}H,Q}$	$C_{\delta^{18}O}$	C^{3}_{H}	$C_{\delta^{18}O,^{3}H}$	$\Delta TT_{\delta^{18}O,^{3}H}$ $\Delta F(T<x)_{\delta^{18}O,^{3}H}$		
Percentiles (yr)	Mean (yr)	17.4	11.9	11.2	15.6	13.2	12.8	11.4	11.0	11.0	5.5	2.4	0.4
	10^{th}	0.5±0.7	0.5±0.8	0.4±0.6	0.3±0.5	0.3±0.5	0.3±0.4	0.04±0.03	0.02±0.02	0.02±0.02	0.0	0.0	0.02
	25^{th}	2.1±2.1	1.9±2.1	1.5±1.8	2.1±1.7	1.5±1.7	1.4±1.5	0.4±0.1	0.2±0.1	0.2±0.1	0.2	0.6	0.2
	50^{th} (median)	9.0±3.3	6.5±4.8	5.7±4.3	8.6±2.6	6.7±3.7	6.6±3.5	3.2±0.2	2.4±0.2	2.5±0.2	2.5	1.9	0.7
	75^{th}	22.2±3.3	17.6±6.5	16.3±6.2	20.8±2.8	18.8±4.6	17.8±4.2	13.7±0.3	12.5±0.4	12.5±0.3	4.6	2.0	1.2
	90^{th}	31.3±4.3	29.2±5.0	28.6±5.1	31.1±4.2	30.4±4.3	29.9±4.2	33.4±0.4	33.4±0.4	32.7±0.2	2.1	0.7	0.0
Water fractions (%)	F(T<3 m)*	18±12	23±19	21±15	16±10	22±13	23±15	22±3	26±3	26±2	-5	-6	-5
	F(T<6 m)	21±13	29±22	30±19	20±11	27±16	27±16	27±2	32±2	32±2	-8	-7	-5
	F(T<1 yr)	24±13	32±22	35±21	22±11	30±16	29±15	34±2	39±2	39±1	-8	-8	-5
	F(T<3 yr)	31±11	39±20	42±19	30±10	37±14	37±14	49±1	53±1	52±1	-8	-7	-4
	F(T<5 yr)	38±10	46±18	49±17	38±9	44±13	44±12	57±1	60±1	60±1	-8	-6	-3
	F(T<10 yr)	52±8	59±13	62±12	53±7	58±10	58±9	69±1	71±1	71±1	-7	-5	-2
	F(T<20 yr)	71±5	77±7	79±7	74±4	76±5	77±5	82±0	83±0	83±0	-6	-2	-1

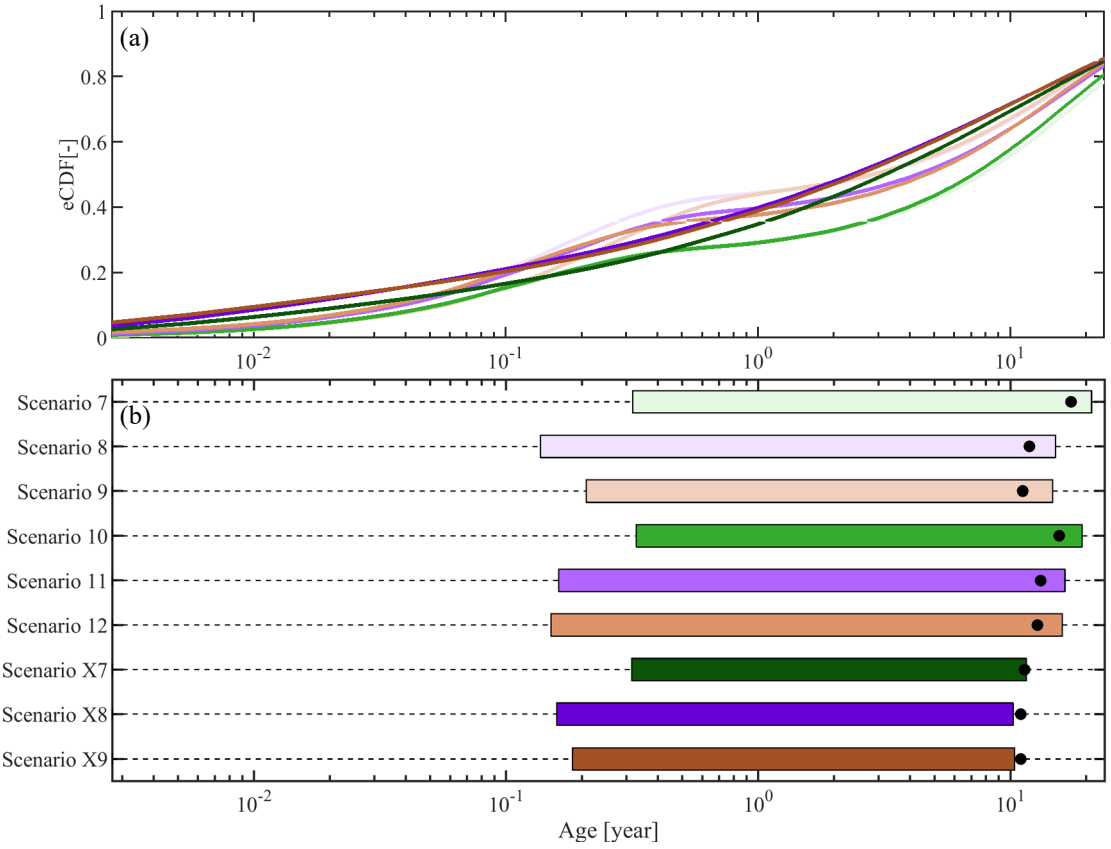

Figure FR4. Stream flow TTDs derived from the 9 model scenarios with the different associated calibration strategies based on different SAS models (i.e., scenarios7-9 based on model IM-SAS-L, scenarios 10-12 based on model IM-SAS-D, scenariosX7-X9 based on model P-SAS which is same as that described in Benettin et al. (2017)). The TTDs represent the volume weighted average daily TTDs during the modelling period 01/10/2001 – 31/12/2016 are given. Green shades represent the TTDs inferred from $\delta^{18}O$ based on different SAS models (from lighter to darker for scenario 7,10, X7) in (a) and (b); the purple shades represent TTDs inferred from ^{3}H based on different models (from lighter to darker for scenario 8, 11, X8), the brown lines represent TTDs inferred from combined $\delta^{18}O$ and ^{3}H based on different models (brown shades from lighter to darker for scenario 9, 12, X9); the black dots in (b) indicate the mean transit time for each model scenario. Note that the mean transit time was estimated by fitting Gamma distributions to the volume-weighted mean TTDs of each individual scenario.

(4) Reviewer Comment:

Thirdly, the fact that spatial aggregation introduces bias in CO model-based MTTs, as stated also by the authors, raises the question to what extent comparison of MTT estimates is meaningful. I understand that the authors would like to test the validity of stable water isotopes in TT modelling particularly of older water ages, and that MTT has been a metric commonly reported for CO models. Nonetheless, according to Kirchner (2016 – reference already in manuscript), sine-wave fitting to seasonal isotope data does give robust estimates of the young water fraction Fyw. Hence, it might be more meaningful to compare Fyw estimates by the different TT model approaches, or, even better, to add this as further TT metric in the comparison.

Reply:

We agree, that MTT estimates from stable isotopes may be less robust than previously assumed *if* they are estimated using CO-type of models and *if* there is a large contrast in MTTs from sub-parts of the system (which we do not know in reality), as demonstrated by Kirchner (2016). This, however, can at this point not (yet) be generalized as it does not imply that MTT estimates obtained from different model approaches and/or systems with little internal contrast in MTTs suffer similar uncertainties.

But we also completely agree with the reviewer that the exclusive comparison of MTT has the potential to conceal interesting pattern. In that sense there seems to be a misunderstanding: our analysis was never limited to MTTs. Instead, throughout the experiment and the reporting of the results, we always analyse the *full range of TTDs*, i.e. percentiles and fractions of water of different ages. This can be seen in Table 5, as well as Figures 7 – 10 in the original manuscript but also in Figures FR3-4 and Tables TR1-2 here above. As water ages throughout all percentiles show *similar pattern* between the individual scenarios, we used the MTT for communicative purposes in the text (note that the use of any other percentile would have resulted in equivalent descriptions) as this has traditionally been the most commonly used metric. For the purpose of our analysis we believe that the emphasis on MTT in the text instead of using multiple metrics improves the readability of the manuscript. In addition, we think that MTT is more suitable here than the fraction of young water, because the core of the analysis is older water instead of young water. In any case, the young water fractions F_{yw} are of course also part of the analysis in the original manuscript (Table 5, Figures 7 – 10) but also here above (Tables TR1-2, Figures FR3-4). Please note that we used a different symbol to represent it – F(T<3m) (see p.17, l.536) – to remain consistent with the notation of other metrics throughout the manuscript. We will clarify this in the text.

(5) Reviewer Comment:

Finally, I would highly appreciate if the authors could increase traceability of their results and provide the underlying tracer data as well as model codes. Traceability is one of the main criteria for HESS nowadays and given that the authors address such a fundamental claim in tracer hydrology and TT modelling, I find it necessary for the entire TT community to benefit from this study not only via the paper, but also in terms of data and code accessibility.

Reply:

We agree, and we will upload the model code to an open access repository. Most tracer data are available via open access databases as explicitly highlighted in text and the Data availability section. The water stable isotopes in stream samples will be available soon, together with other stream data from Germany, as those data are currently prepared for publication in a data paper. Still, the data from the Neckar can be shared upon request.

Minor Comments

(6) Reviewer Comment:

Lines 35—37: if this refers to the findings by Kirchner (2016), one could be more precise by specifying that the MTT (as commonly reported metric) derived from CO models is affected by spatial aggregation errors.

Reply:

Agreed. We will adjust that in the revised manuscript.

(7) Reviewer Comment:

Line 59: in what sense is there more coherence?

Reply:

There is more coherence in the sense that tracer circulation is explicitly linked to and described by the movement of water (i.e. storage and release), which is the actual agent of physical transport in terrestrial hydrological systems.

(8) Reviewer Comment:

Line 70: does Cl- have a clear seasonal cycle? I assume both weathering and anthropogenic effects (e.g., application of road salt) govern its concentrations. Another possible distinction would be radioactive vs. conservative tracers.

Reply:

The chloride ion has a pronounced seasonal cycle, in particular in coastal and maritime influenced climates. It has been successfully applied as age tracer in many previous studies (e.g. Kirchner et al., 2001, 2010; Page et al., 2007; Shaw et al., 2008; Hrachowitz et al., 2009; Soulsby et al., 2010; McMillan et al., 2012; Benettin et al., 2015; Harman, 2015; Wilusz et al., 2017; Cain et al., 2019; Kaandorp et al., 2021; Meira Neto et al., 2022). Anthropogenic effects, such as road gritting, can indeed influence the chloride concentrations. That is why the above studies are limited to catchments with minor human influence.

(9) Reviewer Comment:

Lines 80—98: the focus on the amplitude ratio for the "traditional" TT approaches is fine for simple one-compartment gamma (and thus also exponential) models, but is this also relevant for multiple-compartment CO models and other pre-defined TT shapes such as the dispersion model? This suggests that CO models are exclusively based on the amplitude ratio and shift in seasonal isotope ratios.

Reply:

We are not entirely sure what the reviewer wants to express here. The concept of seasonal tracers as means to estimate stream water ages is rooted in the attenuation of seasonal tracer precipitation amplitudes in the stream water. This is independent of the model application. *Any* model that aims to represent the movement of such a seasonal tracer through a catchment will have to reproduce these observed attenuation between precipitation stream tracer amplitudes, i.e. the amplitude ratio.

(10) Reviewer Comment:

Lines 84—85: "practically" and "feasibly" twice?

Reply:

Indeed. We will correct that.

(11) Reviewer Comment:

Lines 97: to what extent could a spatial aggregation bias also affect spatially lumped (one-compartment) SAS models?

Reply:

This is unknown and to some extent also investigated here, as explicitly mentioned in the original manuscript (e.g. p.5, l.147ff; p.21, l.636ff; p.22, l.698ff).

(12) Reviewer Comment:

Lines 197: you used the CORINE dataset from 2018. To what extent has land use remained stable since 2001?

Reply:

There was no significant change between the here defined land use classes over the 2001-2018 period, as shown in Table TR3 below.

Table TR3: Landuse in the Neckar basin between 1990 and 2018 based on CORINE landcover data.

Landcover percentage	1990	2000	2006	2012	2018
Forest (%)	35	35	35	36	36
Grass/Crop (%)	53	53	52	50	50
Urban (%)	11	12	13	14	14
Water (%)	1	~0	~0	~0	~0

(13) Reviewer Comment:

L*ine 374: we do not necessarily see passive storage volumes in the most recent SAS model studies.*

Reply:

This seems to be a misunderstanding. Indeed, studies based on the "pure" SAS approach that do not model Q, typically define a mixing/sampling storage S_{tot}, although the symbols and terminology vary between individual papers (e.g. Benettin et al., 2017). This S_{tot} represents the total storage available for mixing/sampling in a component and is thereby fully equivalent with our $S_{S,tot}$. The difference is that we have to distinguish a hydraulically active part S_S of that storage that represents the hydraulic head above the river bed to generate Q in our model as visualized in e.g. Zuber (1986, Figure 1 – "dynamic" and "minimum" volume) or Hrachowitz et al. (2016; Figure 2), so that $S_{S,tot}=S_S+S_{S,p}$. As "pure" SAS models do not generate Q they also do not need this distinction. Besides that, two definitions of storage are completely identical.

(14) Reviewer Comment:

Lines 398—414: I am wondering to what extent we can trust the spatially distributed implementation, given that there is only one calibration gauge at the outlet of the entire catchment. This also relates to my general comment about the considerable size and few data for the study basin.

Reply:

This is indeed an important comment. To further test the IM-SAS implementations for their ability to reflect the spatial differences in the study basin, we have now evaluated the models' ability to reproduce observed stream flow in several sub-catchments within the Neckar river basin. As described in detail in reply to Comment (2) above and as can be seen in Figure FR2, the results suggest that the model provides a rather robust representation of the hydrological response and its spatial variability throughout the Neckar basin. We will add this analysis to the revised version of the manuscript.

(15) Reviewer Comment:

Line 411: could you specify what the distributed moisture accounting approach is?

Reply:

This type of model implementation, elsewhere also referred to as "semi-lumped" as in detail described by Ajami et al. (2004), runs a model with spatially distributed forcing data but using the same model parameters. For example, here, each precipitation zone receives different precipitation, but the model parameters are the same in all four precipitation zones. This approach has in past been shown to be very effective for improving the representation of spatially variable response dynamics while limiting the amount of necessary model parameters (e.g. Fenicia et al., 2008; Euser et al., 2015).

(16) Reviewer Comment:

Lines 420—421: why have the authors not applied a multi-objective calibration to the CO models?

Reply:

We are not sure what the reviewer intends to express here. The CO models in our study exclusively model the tracer circulation in the basin. They generate only one single output variable, i.e. the tracer concentration in the stream. We therefore cannot perform the same multi-objective calibration as for the IM-SAS models that besides tracer concentrations also reproduce streamflow Q. If the reviewer had a simultaneous calibration of ^{18}O and ^{3}H in mind, we would like to emphasize that the objective of this paper is to test if the *exclusive* use of ^{18}O underestimates water ages. A simultaneous calibration to both tracers in CO models will not add any additional information to answer this question. Please also note that the simultaneous calibration to ^{18}O and ^{3}H in the IM-SAS models was only done to test if/how it affects parameters that control water fluxes in the model. Major differences in model parameters between the different calibration approaches would have been an indication for differences of how the individual models route water and tracers through the system and thus a source of potential uncertainty in the interpretation.

(17) Reviewer Comment:

Line 424: this is interesting but I think, as stated in my general comments, that TTs should be obtained from a SAS model with storage, input and output fluxes defined a priori (as if they were "real" data), rather than computing TTs from simultaneous calibration against flow and tracers. I think that this would be a more straightforward methodology given the scope of TT modelling and tracers. As presented here, we do not know to what extent simulated TTs are affected by equifinality in the hydrological model parameters.

Reply:

Please see above: as replied to Comment (3) we have now added such a model implementation (scenario X7-8; Figure FR4 and Table TR2). The results lead to the same conclusions as the IM-SAS model implementations: ^{18}O and ^{3}H lead to similar TTDs, and there is no indication for ^{18}O truncating water ages. This further strengthens our original conclusions. We will add this model implementation to the revised manuscript.

(18) Reviewer Comment:

Lines 553—555: not a complete sentence

Reply:

We well correct this.

(19) Reviewer Comment:

Line 571: not only, but also…?

Reply:

We will correct this.

(20) Reviewer Comment:

Lines 577—578: I think you could easily implement the multi-objective calibration for CO models as well.

Reply:

Indeed. It would be easily to implement that, but as explained in response to Comment (16) it does not add any additional information to test the research hypothesis.

(21) Reviewer Comment:

Lines 619—620: so here one could at least test how time-variant/seasonal CO models perform

Reply:

This would indeed be an interesting analysis. However, it is outside the scope of this study as explained in response to Comment (3) above.

(22) Reviewer Comment:

Lines 642—644: could this not be an indication of the fact that there are too many degrees of freedom and the model succeeds to fit the tracer data, regardless of whether it is spatially lumped or semi-distributed?

Reply:

As shown in Figure FR1 above, there is little indication of model overfitting that could results from "too many degrees of freedom". One explanation of the observed similarity between the lumped and distributed models could be that much of the climatic and topographic heterogeneity within the catchment is filtered out in the response (see also reply to Comment (2) above), so that a lumped representation may be sufficient to pick up the major features of the hydrological response in the study basin.

(23) Reviewer Comment:

Lines 656—657: see, e.g., Nguyen et al. (2022) who found substantial differences in SAS-based transport models between spatially lumped and semi-distributed setup.

Reply:

We will refer to that study as an example of a setting where spatial differences seem to be more relevant.

References:

Ajami, N. K., Gupta, H., Wagener, T., & Sorooshian, S. (2004). Calibration of a semi-distributed hydrologic model for streamflow estimation along a river system. Journal of hydrology, 298(1-4), 112-135.

Aubert, A. H., Kirchner, J. W., Gascuel-Odoux, C., Faucheux, M., Gruau, G., & Mérot, P. (2014). Fractal water quality fluctuations spanning the periodic table in an intensively farmed watershed. Environmental Science & Technology, 48(2), 930-937.

Benettin, P., Kirchner, J. W., Rinaldo, A., & Botter, G. (2015). Modeling chloride transport using travel time distributions at Plynlimon, Wales. Water Resources Research, 51(5), 3259-3276.

Benettin, P., Soulsby, C., Birkel, C., Tetzlaff, D., Botter, G., & Rinaldo, A. (2017). Using SAS functions and high‐resolution isotope data to unravel travel time distributions in headwater catchments. Water Resources Research, 53(3), 1864-1878.

Cain, M. R., Ward, A. S., & Hrachowitz, M. (2019). Ecohydrologic separation alters interpreted hydrologic stores and fluxes in a headwater mountain catchment. Hydrological Processes, 33(20), 2658-2675.

Euser, T., Hrachowitz, M., Winsemius, H. C., & Savenije, H. H. (2015). The effect of forcing and landscape distribution on performance and consistency of model structures. Hydrological Processes, 29(17), 3727-3743.

Fenicia, F., Savenije, H. H., Matgen, P., & Pfister, L. (2008). Understanding catchment behavior through stepwise model concept improvement. Water Resources Research, 44(1).

Godsey, S. E., Aas, W., Clair, T. A., De Wit, H. A., Fernandez, I. J., Kahl, J. S., ... & Kirchner, J. W. (2010). Generality of fractal 1/f scaling in catchment tracer time series, and its implications for catchment travel time distributions. Hydrological Processes, 24(12), 1660-1671.

Hrachowitz, M., Soulsby, C., Tetzlaff, D., Dawson, J. J. C., & Malcolm, I. A. (2009). Regionalization of transit time estimates in montane catchments by integrating landscape controls. Water Resources Research, 45(5).

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.

Hrachowitz, M., Fovet, O., Ruiz, L., & Savenije, H. H. (2015). Transit time distributions, legacy contamination and variability in biogeochemical 1/fα scaling: how are hydrological response dynamics linked to water quality at the catchment scale?. Hydrological Processes, 29(25), 5241-5256.

Kaandorp, V. P., Broers, H. P., Van Der Velde, Y., Rozemeijer, J., & De Louw, P. G. (2021). Time lags of nitrate, chloride, and tritium in streams assessed by dynamic groundwater flow tracking in a lowland landscape. Hydrology and Earth System Sciences, 25(6), 3691-3711.

Kirchner, J. W., Feng, X., & Neal, C. (2000). Fractal stream chemistry and its implications for contaminant transport in catchments. Nature, 403(6769), 524-527.

Kirchner, J. W., Tetzlaff, D., & Soulsby, C. (2010). Comparing chloride and water isotopes as hydrological tracers in two Scottish catchments. Hydrological Processes, 24(12), 1631-1645.

Kirchner, J. W., & Neal, C. (2013). Universal fractal scaling in stream chemistry and its implications for solute transport and water quality trend detection. Proceedings of the National Academy of Sciences, 110(30), 12213-12218.

McMillan, H., Tetzlaff, D., Clark, M., & Soulsby, C. (2012). Do time‐variable tracers aid the evaluation of hydrological model structure? A multimodel approach. Water Resources Research, 48(5).

Meira Neto, A. A., Kim, M., & Troch, P. A. (2022). Physical Interpretation of Time‐Varying StorAge Selection Functions in a Bench‐Scale Hillslope Experiment via Geophysical Imaging of Ages of Water. Water Resources Research, 58(4), e2021WR030950.

Page, T., Beven, K. J., Freer, J., & Neal, C. (2007). Modelling the chloride signal at Plynlimon, Wales, using a modified dynamic TOPMODEL incorporating conservative chemical mixing (with uncertainty). Hydrological Processes: An International Journal, 21(3), 292-307.

Shaw, S. B., Harpold, A. A., Taylor, J. C., & Walter, M. T. (2008). Investigating a high resolution, stream chloride time series from the Biscuit Brook catchment, Catskills, NY. Journal of Hydrology, 348(3-4), 245-256.

Soulsby, C., Tetzlaff, D., & Hrachowitz, M. (2010). Are transit times useful process‐based tools for flow prediction and classification in ungauged basins in montane regions?. Hydrological Processes, 24(12), 1685-1696.

Soulsby, C., Birkel, C., Geris, J., Dick, J., Tunaley, C., & Tetzlaff, D. (2015). Stream water age distributions controlled by storage dynamics and nonlinear hydrologic connectivity: Modeling with high‐resolution isotope data. Water Resources Research, 51(9), 7759-7776.

Wilusz, D. C., Harman, C. J., & Ball, W. P. (2017). Sensitivity of catchment transit times to rainfall variability under present and future climates. Water Resources Research, 53(12), 10231-10256.