We highly appreciate the time and effort that the Reviewer has dedicated to providing feedback on our manuscript and are grateful for the insightful comments on our manuscript. Please find below our detailed replies to the individual comments.

<u>(1) Reviewer Comment:</u>

I suggest to provide more context / justification / details about the calibration procedure – for example, how do you make sure your calibrated best-fits were not local best-fits but globe ones. The best-fit results of different implementations (such as IM-SAS-L and IM-SAS-D) were similar, but that does not mean the modeled results such as MTT was true. This generally requires an analysis of the potential uncertainty. While I understand a full uncertainty analysis may be unfeasible, the impact of operational choices done in the calibration exercise need to be better discussed.

<u>Reply:</u>

We completely agree with this point. We have therefore done an uncertainty analysis to quantify the effects of parameter uncertainty on the modelled TTDs by randomly sampling from the posterior parameter distributions for both, IM-SAS-L and IM-SAS-D models. While parameter uncertainty can cause some variability in TTDs and thus in the actual magnitudes of water ages, this variability is consistently within similar age ranges for ^{18}O and ^{3}H, respectively. It does therefore not affect the overall interpretation of the results and the rejection of the hypothesis that ^{18}O underestimates water ages, as shown for scenarios 10-12 in Figure FR1 here below. We will add these results in the revised manuscript.

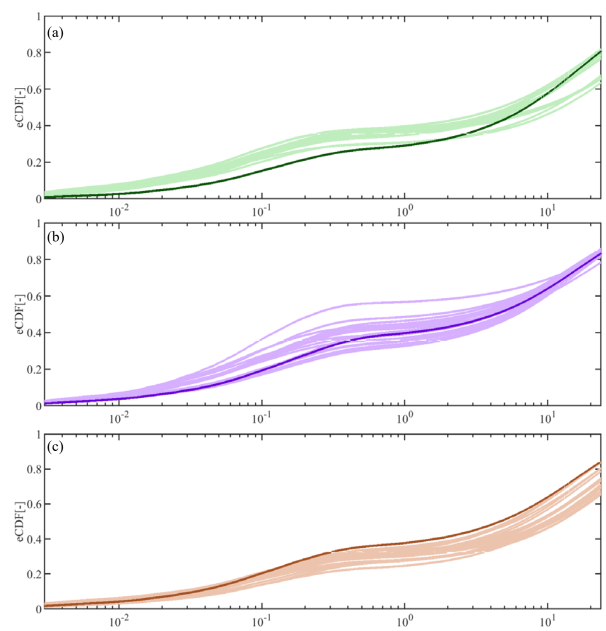

Figure FR1. Stream flow TTDs derived from the 6 model scenarios based on IM-SAS models with the different associated calibration strategies (scenarios 10-12). Each line represents the volume weighted average daily TTDs during the modelling period 01/10/2001 – 31/12/2016, generated from parameters randomly sampled from the posterior distribution (light shades) and the most balanced solution of each scenario (dark shades). (a) TTDs inferred from $\delta^{18}O$ in scenario 10; (b) TTDs inferred from ^{3}H in scenario 11; (c) The TTDs inferred from combined $\delta^{18}O$ and ^{3}H in scenario 12.

(2) Reviewer Comment:

I do agree with the authors that the ^3H and $\delta^{18}O$ tracers both are informative for the flow systems, what is needed is just a model good enough to resolve such information in a meaningful way. Especially for the catchments with strong seasonality. However, I am not sure if the model has to use combined date sets of hydrological and tracer as the author argued that "only the combined information using hydrological and tracer data and the consideration of transient flow conditions gives similar MTT, independent of the used tracer". I think the important thing is that the flow model can represent the reality in a good way, such that the tracer transport can be well reproduced. Using hydrological data in calibration may not a key control for that.

Reply:

We agree with this point. We will therefore reformulate that sentence on P.20, l.620, "only the combined information using hydrological and tracer data and the consideration of transient flow conditions gives similar MTT, independent of the used tracer" in the revised manuscript so that it better reflects that point.

(3) Reviewer Comment:

Line 160: What are Ep and P?

Reply:

Thank you for pointing this out. While Ep represents potential evaporation, P represents precipitation. We will add the definitions in the revised manuscript.

(4) Reviewer Comment:

Line 368: perhaps say that the storage component is just locally full-mixed and those local full mixtures do not lead to an overall fully mixed system

Reply:

We completely agree with this suggestion. It was mentioned on P.12, L.368ff, but we will make it clearer in the revised manuscript.

(5) Reviewer Comment:

I don't think that to reduce computational time and computer memory requirements is good reason for using uniform SAS functions rather than other shapes of SAS function. I think the right way should be describing the model of reduced complexity (parameters) was already enough for your modelling targets.

Reply:

We agree with the argument that reduced complexity here already allows to draw robust conclusions. We will reformulate the statement and add this aspect. However, we would also like to explicitly re-iterate here that computational capacity imposes major practical obstacles to testing other SAS function shapes: in contrast to uniform distributions, the sampling process then requires an explicit generation of RTDs and TTDs for each time step and to "carry" all RTDs and TTDs of all model components through the entire model period, including the warm-up period (here: 46 years). This entails for a daily modelling time-step the simultaneous handling of multiple matrices > 16.800x16.800 elements in floating number format (i.e. 8B each), which corresponds to >2 GB/matrix. With a working memory of common but good computers (i.e. 16-32 GB) this means that the generation of RTDs and TTDs alone will use (if not exceed) the memory of these computers, not to speak of other processes required.

(6) Reviewer Comment:

Line 378: could you explain in more detail how was the tracer sampled from the passive and active volumes? Also random sampling from Ss,tot ?

Reply:

The tracer and age composition of that outflow is indeed randomly sampled from the total groundwater storage volume $S_{S,tot}$. We will clarify this in the revised manuscript.

(7) Reviewer Comment:

Line 393-395: maybe simply say the lumped implementation used a single HUR to represent the entire basin. Is that what you mean? In this case the precipitation zones were not used any more, right? Maybe clarify this.

Reply:

Indeed, the lumped implementation used a single HRU (equivalent to the forest HRU described in distributed model, Fig.2) to represent the entire catchment and the precipitation zones were not used any more in this lumped case. We have will clarify this in the revised manuscript.

(8) Reviewer Comment:

Equation 14: what are $E_{mse,Q,n}$ and $E_{mse,tracer,m}$?

Reply:

Thank you for pointing this out. We will add the missing definitions in the revised manuscript.

(9) Reviewer Comment:

Line 473: it looks like that when using all the data, the lumped model (scenario 9) was even better than the distributed model (scenario 12) that has more parameters, does that mean the high model complexity is not essential for a better model performance in your case, could you clarify that.

Reply:

This is an interesting aspect. However, while the distributed implementation IM-SAS-D can indeed not be considered to outperform the lumped IM-SAS-L implementation, the opposite cannot be concluded either: as can be seen in Table 4, considering the most balanced solution, some signatures were indeed captured better by IM-SAS-L than by IM-SAS-D. Yet, others were much better reproduced by IM-SAS-D. In addition, it can be seen that the full set of pareto front solutions of IM-SAS-L includes a considerable number with poorer performance metrics (i.e. upper limit of performance ranges shown in Table S5 in the Supplementary Material).

(10) Reviewer Comment:

Line 508: Table 3?

Reply:

Indeed. We will correct that.