We thank the reviewer for spending their time reviewing our manuscript and for providing constructive comments. Please, find below our point-to-point response to the reviewer's comments (text in black) and proposed modifications (text in blue) of the original manuscript.

This paper presents an analysis of uncertainty in transit time distributions estimated using SAS functions, including that arising from the interpolation of input tracer data, and from the SAS function parameterization. Uncertainty of each configuration of model and input data is assessed from the range of predictions made by the top 5% of monte-carlo sampled parameter sets ranked by goodness-of-fit (KGE). The fraction of young water F_{yw} obtained from the method proposed by Kirchner (2016) is used to further constrain the behavioral set. This paper aims to address an important gap in the literature. There is a need to better understand the uncertainty associated with SAS models, and how data can be best used to constrain them.

We thank the reviewer for acknowledging the important gap in the literature (uncertainty induced by SAS parameterization and input data) that we want to address in this manuscript.

However, I think there are two major problems with the approach used here, and I think the resulting conclusions are unsupported as a result.

- I don't think it makes sense to use F_{yw} to constrain the SAS model parameterizations.

We appreciate this comment, and we understand the reviewer's concern that it seems it might not make more meaningful sense to use Fyw, derived using a relatively simple approach (the sine-wave fitting approach), to constrain the results already run with the best available data obtained from a more "elaborated" model (the StorAge Selection – SAS approach). This is especially true as we cannot know if Fyw from the sine-wave approach is better than that from the SAS approach or not. We note, however, that the main goal of our study is to highlight the uncertainty in SAS-based modeled results arising from model inputs, as well as underlying model structure and parameters – that have been not thoroughly evaluated yet in previous studies. The use of Fyw from the sine-wave fitting approach as an additional and minor part of the presented work, as an attempt to suggest a further metric that might be helpful in constraining the model simulations of an already calibrated SAS model.

The reviewer's comments have stressed the strong assumptions we have made in the use of Fyw as an additional model constraint. We agree that it may need a more elaborate procedure that considers the uncertainty in sine-wave fitted Fyw and corresponding age thresholds for young water (see below for further explanations) to relax some of these assumptions. Adding this to the revised manuscript would, however, put the focus too much on the use of Fyw and distract from the first and major part of the manuscript i.e. demonstrating the appreciable uncertainty in SAS modeling. Hence, we have decided to discard the part about Fyw from this manuscript and instead plan to develop and illustrate this approach more thoroughly in a different study.

- I think the use of top 5% KGE to define behavioural parameter sets makes it impossible to meaningfully compare the uncertainty of each configuration

Thank you for this remark. As we understood, the reviewer suggest to use a fixed KGE value for defining the behavioral simulations, rather than fixing the sample size based on best 5% KGE, as we proposed.

Firstly, by doing this, we will run into the same problem raised by the reviewer – these behavioral simulations do not have the same range in goodness-of-fit i.e. KGE. In fact, if we define the behavioral simulations as those with KGE $\geq$ 0.5, the range of KGE with setup 1 is KGE = [0.5, 0.64], while for setup 4 IT is KGE = [0.5, 0.72] as it is possible to see in the range of behavioral KGE values in Fig. 2 of the original manuscript. Secondly, fixing the KGE threshold will lead to a different sample size per each model setup. For example, if we choose a fixed threshold limit of KGE $\geq$ 0.5, the behavioral solutions range between 1,300 and 2,700 across the 12 model setups. When looking at the uncertainty in the simulated outputs, the 90% confidence interval is wider for model setups that have a larger number of behavioral solutions than for those that have a smaller number. Therefore, a varying sample sizes would affect the uncertainty analysis. With a fixed sample size based on the 5% best KGE, we can ensure a meaningful comparison in uncertainty across the model scenarios. Also, we are still able to meet the requirement of a minimum acceptable KGE value (minimum KGE in the behavioral solutions across all tested setups is 0.57).

Despite this, we acknowledge that fixing the sample size is not necessarily better than imposing a threshold limit as there will be always a tradeoffs and pros/cons of each of the chosen approaches. However, given (i) the arguments provided above, (ii) the objective of our study (showing the uncertainty in the modeled outputs arising from model inputs, structure and parameters, not identifying the best simulations) and (iii) the large number of model setups explored i.e. 12, we find it more appropriate to use the top 5% simulations. Therefore, we would like to keep the definition of behavioral solution in the way we proposed in the revised manuscript. Nevertheless, during the revision, we will make it clear the reasons regarding the chosen criterion by providing the supporting motivation described above.

Also, as we understood from the reviewer, we cannot use the GLUE methodology if we consider the top 5% simulations as behavioral because "*each behavioral set would have a different total likelihood associated with it (if a formal likelihood were estimated)*". Therefore, in the revised manuscript, we will use the informal likelihood (the Sequential Uncertainty Fitting Procedure – SUFI-2; Abbaspour et al., 2004), an approach that has been widely used for estimating parameter uncertainty of eco-hydrological models (e.g., the Soil and Water Assessment Tools – SWAT, Arnold et al., 2012). In this way, we will estimate the uncertainty for the top 5% best parameters which is described by a uniform distribution, and not by a formal likelihood such as done in GLUE. We have already checked differences in results with the SUFI-2 approach versus those with the GLUE approach, and we have found insubstantial differences in the model prediction uncertainty (see below for more details).

Reference:

Abbaspour, K. C., Johnson, C. A., & van Genuchten, M. T. (2004). Estimating uncertain flow and transport parameters using a sequential uncertainty fitting procedure. Vadose Zone Journal , 3 , 1340–1352. https://doi.org/10.2136/vzj2004.1340.

Arnold, J. G., Moriasi, D. N., Gassman, P. W., Abbaspour, K. C., White, M. J., Srinivasan, R., ... & Jha, M. K. (2012). SWAT: Model use, calibration, and validation. *Transactions of the ASABE*, *55*(4), 1491-1508.

Major issues

Use of F_{yw} to constrain SAS models

- I do not think it makes sense to use the young water fraction obtained from the sine-wave ratio to constrain a SAS model. Kirchner's method for this is useful for obtaining rough estimates of the fraction of water that is roughly a quarter of a year old from tracer time series. The method might be robust (in some sense) but it isn't precise. SAS models are a more complex and sophisticated tool that have the *potential* to provide a much more precise estimate of water age distribution from the same data. It doesn't make sense to me to use the outputs of a rough-and-ready model to constrain the parameters of a more precise one.

Thank you for this comment. Please, refer to our response above for the proposed modifications in the revised manuscript.

Here, we want to add that additional complexity to constrain models does not necessarily lead to a better result than the use of simple models. This, for example, has been demonstrated and supported in the hydrological community through different studies (Michaud and Sorooshian, 1994, Orth et al., 2015, Merz et al., 2022). Also, the reviewer argues that the sine-wave fitting approach is not precise. Although we cannot generally falsify this statement, it is also difficult to prove that it is fully correct given that the level of “preciseness” is difficult to assess for both approaches (sine-wave fitting and SAS functions). To our knowledge, there are no studies proving that Fyw from the sine-wave fitting approach is not “precise”. Conversely, the sine-wave fitting approach has increasingly been acknowledged in the past years for estimating Fyw (Jasecko et al., 2016, Lutz et al., 2018; von Freyberg et al., 2018 , Stockinger et al., 2019; Gallart et al., 2020). However, we agree that there is a need for a more rigorous testing to better understand, which approach provides a better estimate of Fyw based on the available data (same as done for the transit times in a recent paper by Benettin et al., 2022). Since this topic is out of the scope of current work, we will revise our work - excluding the part on Fyw discussion - and focus on the uncertainty in the SAS models.

References:

- Michaud, J. and Sorooshian, S. (1994) Comparison of Simple versus Complex Distributed Runoff Models in a Midsized Semi-Arid Basin. Water Resources Research, 30, 593–605, https://doi.org/10.1029/93WR03218.

- Orth, R. Staudinger, S.I. Seneviratne, J. Seibert, & M. Zappa (2015) Does model performance improve with complexity? A case study with three hydrological models. J. Hydrol., 523, 147–159, https://doi.org/10.1016/j.jhydrol.2015.01.044.

Merz, R., Miniussi, A., Basso, S., Petersen, K. J. & Tarasova, L. (2022) More Complex is Not Necessarily Better in Large-Scale Hydrological Modeling: A Model Complexity Experiment across the Contiguous United States. BAMS, E1947–E1967, https://doi.org/10.1175/BAMS-D-21-0284.1.

Jasechko, S., Kirchner, J. W., Welker, J. M., & McDonnell, J. J. (2016). Substantial proportion of global streamflow less than three months old. Nat Geosci, 9, 126–129, https://doi.org/10.1038/ngeo2636.

Lutz, S. R., Krieg, R., M¨uller, C., Zink, M., Kn¨oller, K., Samaniego, L., & Merz, R. (2018). Spatial patterns of water age: using young water fractions to improve the characterization of transit times in contrasting catchments. Water Resour. Res., 54 , 4767–4784, https://doi.org/10.1029/2017WR022216.

von Freyberg, J., Allen, S. T., Seeger, S., Weiler, M., & Kirchner, J. W. (2018). Sensitivity of young water fractions to hydro-climatic forcing and landscapeproperties across 22 Swiss catchments. Hydrol. Earth Syst. Sci., 22 , 3841–3861, https://doi.org/10.5194/hess-22-3841-2018.

Stockinger, M. P., Bogena, H. R., L¨ucke, A., Stumpp, C., & Vereecken, H. (2019). Time variability and uncertainty in the young water fraction in a small headwater catchment. Hydrol. Earth Syst. Sci., 23, 4333–4347, https://doi.org/10.5194/hess-23-4333-2019. 2018.

Gallart, F., Valiente, M., Llorens, P., Cayuela, C., Sprenger, M., & Latron, J. (2020). Investigating young water fractions in a small mediterranean mountain catchment: Both precipitation forcing and sampling frequency matter. Hydrol. Process., 34 , 3618–3634, https://doi.org/10.1002/hyp.13806.

Benettin, P., Rodriguez, N. B., Sprenger, M., Kim, M., Klaus, J., Harman, C. J., et al. (2022). Transit time estimation in catchments: Recent developments and future directions. Water Resour. Res., 58, e2022WR033096, https://doi.org/10.1029/2022WR033096.

- I believe the fact that the authors do find that F_{yw} has power to constrain the SAS parameters is largely because the uncertainty in the associated age threshold τ_{yw} is not accounted for. The method that F_{yw} relies on is based on a variety of assumptions, including that the inputs are sinusoidal and that the transit time distribution is approximately a gamma distribution. Two important *and distinct* sources of uncertainty here are:

- The threshold age of the young water fraction τ_{yw} is not 75 days, as suggested by the authors. Rather it depends on the shape parameter of the assumed gamma distribution. As Figure 10 of Kirchner (2016) shows, for a shape parameter of 0.2 it is around 40 days, while for a shape parameter of 2 it is more like 100 days. This considerable uncertainty is not accounted for in the present paper.

- The estimates of amplitudes A_q and A_p obtained from fitting sinusoids to the observed tracer timeseries are uncertain, and that uncertainty ought to be estimated and propagated into uncertainty in F_{yw}. The authors may have accounted for this (if I understand the brief statement on line 165) but they claim that in doing so they have also accounted for the uncertainty in τ_{yw}, which is not the case. These errors are independent of each other. The errors obtained for F_{yw} were only 0.07-0.08 (line 325), which I suspect contributes far less uncertainty than the 60-day window bracketing τ_{yw} paper.

Thank you for the above observations on the uncertainty in Fyw and the young water threshold (τ_{yw}). Although we will remove the Fyw part from the manuscript, we acknowledge that the uncertainty in τ_{yw} was not properly addressed in our original manuscript. We agree with the reviewer that not only the uncertainty in Fyw (which we accounted for in the original manuscript) should be considered, but also in τ_{yw} (which we did not) when fitting the sine function to the tracer data in inflow and outflow.

- Furthermore, the theory behind F_{yw} and τ_{yw} rests on the assumption that flows through the system are steady, the transit time distribution is invariant, and that the input signal is a perfect sinusoid. These are not the case in general in real watersheds, which results in additional epistemic uncertainty into the estimates of F_{yw} and τ_{yw}. These particular sources of uncertainty do not necessarily apply to the SAS models, since they can allow for variable flows, variable transit time distributions, and make use of the observed input signal.

Thank you for this comment. Although we will remove the Fyw part from the revised manuscript, we would like to comment on the aspect of the steady state assumption - correctly highlighted by the reviewer. In the original manuscript we first estimated the transient daily transit time distribution (TTD) and then derived the marginal TTD, from which we calculated the Fyw values. By estimating the marginal TTD, we assume to reflect the steady state behavior, though admittedly not perfect, but this could be a reasonable approach. To the aspects of the input signal and the transit time distribution, we agree with the reviewer that the isotope signal in inflow and outflow does not perfectly follow the sinusoidal as the marginal TTD might not perfectly follow a gamma distribution. Therefore, we acknowledge that the approach presented in the original manuscript has some limitations and there is uncertainty in Fyw (which we accounted for in the original manuscript) and in τ_{yw} (which we did not) when fitting the sine function to the tracer data in inflow and outflow.

However, it could also be argued the other way around: SAS functions have uncertainties (e.g. lack of agreement on which model parameterization to use, equifinality of parameters, assumptions regarding age distributions of evapotranspiration) that, in contrast, do not apply to Fyw obtained with the sine-wave fitting approach. Indeed, we explored and highlighted some of these uncertainties in the current study (i.e. tracer data interpolation and choice for SAS parameterization), which have not been emphasized in detail in previous studies.

- In fact, it is possible to reproduce the model used to justify Kirchner's method as a SAS model. This can be done by approximating the flows as constant, replacing the inputs concentrations with

sinusoids, and choosing a SAS function whose corresponding steady-state TTD is a gamma. From this perspective F_{yw} and τ_{yw} can be viewed as outputs of a particular SAS model parameterization run with degraded data. Why should the results of that parameterization be used to constrain other parameterizations run with the best available data?

Thanks for this obsveration. We agree with the reviewer that, being the SAS parameters already calibrated and being the model already run with the best available data, there may be no reason for further constraining the model with any additional metrics e.g. Fyw. For this reason, as we have already argumented in the first response, we will remove the Fyw part from the manuscript.

Use of top 5% KGE as the 'behavioural' parameter set

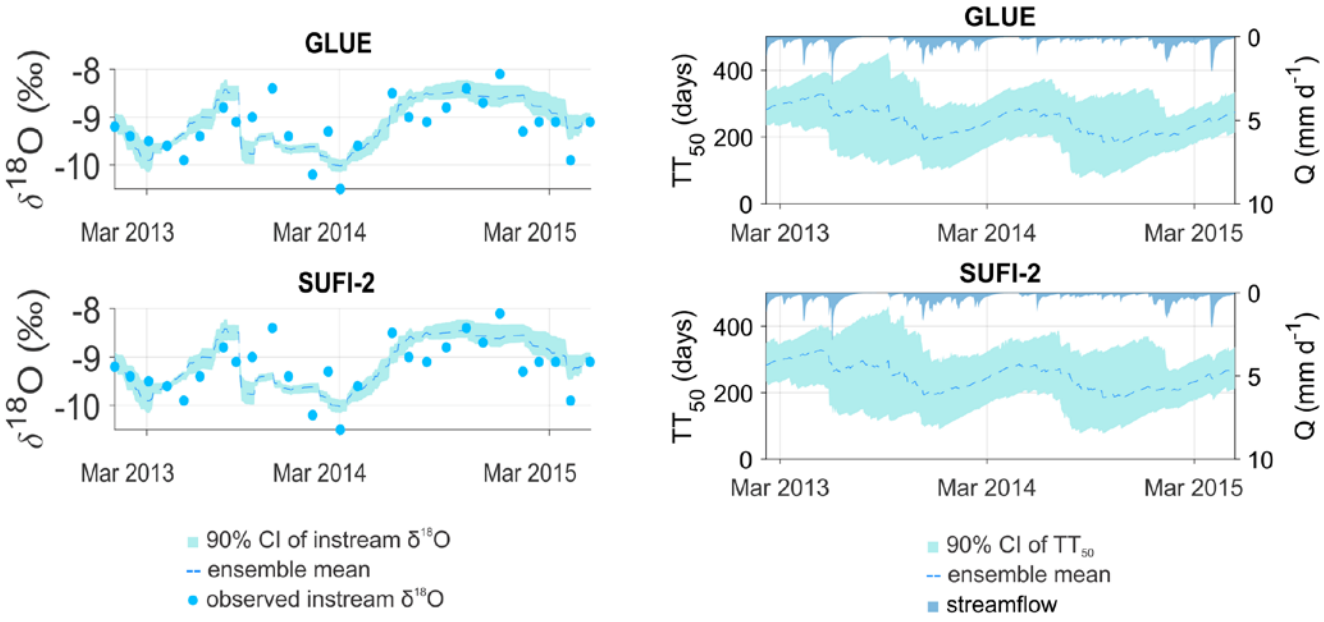
- The use of the top 5% KGE as the 'behavioural' parameter set makes it impossible to make meaningful comparisons between the different parameterizations (i.e. PLTI, PLTV, BETA). This is because the range of goodness-of-fit (i.e. the KGE) of each model's behavioral set depends on the size of the pool from which it was taken, in addition to how well it actually fits the data. The range of KGE in the top 5% depends on the assumed prior distribution of the parameter set, since that determines what the 5% is a percentage of. Since each parameterization has fundamentally incommensurate parameters, there isn't an obvious way to normalize for this dependence across different parameter spaces. As a result each behavioral set would have a different total likelihood associated with it (if a formal likelihood were estimated). Comparing these different behavioral parameterizations therefore makes no sense, since they have been held to different standards.

- One consequence of effectively holding each parameterization to a different standard is that the error associated with the more flexible parameterizations (PLTV, BETA) is larger than that associated with the less flexible one (PLTI), when we would expect the opposite to hold. This is particularly true given that PLTI represents a special case of both PLTV and BETA (when $k_{Q1}=k_{Q2}=k$ and when $\alpha=k, \beta=1$ respectively). However, as seen in Figure 2 the behavioral sets of BETA (and to a lesser extent PLTV) seem to include models that are considerably worse fits to the data than the worst models in the behavioral set of PLTI.

- To make meaningful comparisons between different parameterizations the analysis would need to be redone with a standard for 'behavioral' that is consistent across the different parameterizations. This might be as simple as choosing a cutoff value of KGE to define the behavioral set, but it would likely change the resulting conclusions about the merits of each parameterization.

Thank you for this observation. Please, refer to our response above for the reasons why we want to keep the definition of behavioral solution based on the 5% best simulations in terms of KGE, and the proposed modifications in the revised manuscript.

Here, we just want to show that there is no significant differences in the results with the SUFI-2 approach compared to those with the GLUE approach for quantifying the model prediction uncertainty (e.g. as shown below for the simulated instream isotope and median transit times for one of the 12 tested model setups).



Minor issues

- Line 57: The gamma distribution has also seen some use

We will add the gamma distribution to the list of commonly used parameterizations employed to approximate the SAS functions.

- Line 64: I don't think that the statement that F_{yw} is useful for short-term data is quite right, since the method does require data covering multiple cycles of sinusoidal variation to fit to reliably

As we will remove the Fyw part in the revised manuscript, this phrase will not be part of the revised manuscript.

- Line 111: S_{T_0} is a function of age: $S_{T_0}(T)$

We will change S_{T_0} to $S_{T_0}(T)$.

- Line 130: k_{Q1} and k_{Q2}

Here we do not use the subscript *Q* referring to streamflow, because we describe the parameters of SAS functions in general, without referring to a specific flux. Therefore, we prefer to leave *k* rather than *kQ* in lines 127-131. However, in the rest of the text, we specify which parameterization (e.g. PLTI and PLTV) we apply to each flux (i.e. streamflow and evapotranspiration), so we write *kQ*, *kQ1*, *kQ2* and *kET*.

- Table 2: Why are k_{Q1} and α grouped together? Same with k_{Q2} and β

There is no specific reason: we simply decided to group *kQ1* with *alpha* and *kQ2* with *beta* because the two correspond to the shape parameters of PLTV and BETA, respectively. Upon the reviewer's suggestion, we will disaggregate them into separate rows (in Table 1) in the revised manuscript.

- Line 188: Is TT_{50} is the median of the *backward* transit time distribution $p_Q(T,t)$ as defined in equation (5)? In that case this statement is incorrect, and should be "the maximum time elapsed *since* the youngest 50% of the water in outflow first entered the catchment", or perhaps "the age that half the outflow is older than, and half younger than, as measured from the time it fell as precipitation".

Here we consider the backward formulation of the transit time distribution. We will clarify this in line 188 and modify the definition of median transit time accordingly.

- Figure 3: A legend explaining the colors and a reference to Table 1 would aid interpretation here

We will add a legend and a reference to Table 1 in Fig. 3.

- Line 221: Parameters for the *SAS function* of Q...

We will add "SAS functions" in the revised manuscript.