General

R3: "The manuscript highlights an important aspect of inverse modeling based on stream tracer test – the uncertainty and identifiability of the estimated model parameters and the connected metrics used to assess solute transport processes in streams. Although I think that the text is generally well written, the structure of the paper (e.g. order of sections) could be streamlined and improved for clarity. As it is right now, I found myself going back and forth between the different sections in order to understand the iterative and rather complex structure. Overall, I appreciate the approach and the general topic, but I believe the manuscript would benefit from addressing a number of issues highlighted below."

Authors: We thank Reviewer 3 (R3) for the time spent in reviewing our manuscript and for the constructive and supportive feedback. In the revised manuscript version, we will re-arrange the structure of the method, results, and discussion section to improve the overall clarity. Consistent with comments and suggestions from R2, we will modify the method section as follows:

- a.1) ADE parameters
- a.2) TSM parameters
  - a.2.1 Identifiability of TSM parameters when v = vvariable
    - a.2.1.1 TSM first iterations
    - a.2.1.2 TSM last iterations
  - a.2.2 Identifiability of TSM parameters when v = vpeak
    - a.2.2.1 TSM first iterations
    - a.2.2.2 TSM last iterations
- a.3) DYNIA
  - a.3.1 DYNIA (first + last iteration - v = vvariable)
  - a.3.2 DYNIA (first + last iteration - v = vpeak)
- a.4) Comparison
  - a.4.1 Comparison with inverse modeling results (OTIS-P)
  - a.4.2 Comparison with random sampling approaches (OTIS-MCAT)
- a.5) Metrics and hydrologic interpretations of model results

The revised results and discussion sections will mirror the same order that will be used in the revised method section. This modification will be in line with the suggestion of R2 and with the requirements of R1.

Major comments:

R3: "I think that the terms "identifiability" and "sensitivity" needs to be defined early on in the introduction so that there is no doubt what is meant by these key terms in the context of the study. As it is now the words are used already from the start of the paper, but the definitions are "hidden" in Appendix A. I suggest moving the definitions (lines 902-905) and placing them in the main paper where they are first introduced."

Authors: We will implement these changes.

R3: "There are several threshold values used to select behavioral parameter sets and subsequently to define identifiability. However, these thresholds are not sufficiently motivated and discussed, which makes it difficult for the reader to assess the results. This holds for the definitions of both the global identifiability (e.g. the top 0.1-10% of the models when assessing the CDF deviation from the 1:1 line having, the grouping based on the K-S results) and the dynamic identifiability (e.g. information content > 0.66). This leads to questions about the subjectivity of these thresholds when assessing the identifiability and how sensitive the overall assessment of the model parameters are to the chosen thresholds (e.g. in Figure 8)."

Authors: Selecting a certain behavioral threshold introduces a certain degree of subjectivity to the results. Thus, it is important to remain consistent with previous literature and the previously selected behavioral thresholds. We assured consistency with previous work by selecting the same behavioral thresholds used in a wide range of transient storage studies that used a 10% threshold (Wagener et al., 2002; Wlostowski, 2013; Ward 2013; Ward 2019; Kelleher 2019) and 0.1% threshold (Ward et al., 2017). We will clarify this choice in the revised manuscript.
For K-S test we also oriented our selection of thresholds on current literature (Ouyang et al. 2014). However, K-S test was never used in TSM. We decided to adopt K-S test and the reported thresholds to assure a more robust assessment of identifiability of TSM parameters, since parameter identifiability is usually assessed via visual interpretation of global identifiability analysis results (Wagener et al., 2002; Wlostowski et al., 2013; Ward et al., 2013; Ward et al. 2018; Ward et al., 2019; Kelleher et al., 2019). We will clarify this choice in the revised manuscript.
For the dynamic identifiability analysis, we were interested if a certain parameter was poorly identifiable (information content < 0.33) or strongly identifiable (information content > 0.66) in a certain section of the BTC. While the reported thresholds are arbitrary, they assure consistency in terms of parameter interpretation compared to previous work, where the role of a certain parameter on the BTC was controlled by a certain degree of subjectivity since dynamic identifiability analysis was visually interpreted (Wagener et al., 2002; Wlostowski et al., 2013).
We will extend the discussion of the revised manuscript by accounting for the role of the thresholds in the adopted modelling approach. It is clearly of interest to explore the impact of the selection of the threshold on the physical realism of the results in future work. However, our results are physically realistic, and the selected thresholds proved to be effective in the interpretation of model outcomes and in the achievement of parameter identifiability.

R3: "The objective function used to assess the model performance, i.e. the RMSE, is based on the difference between observed and modelled BTC over a given time scale/number of observations. What are the effects on the RMSE for concentration values that may differ more than an order of magnitude over assessed time window, i.e. for high values (peak of the BTC corresponding to 50 mg/l) compared to the low values (tail of the BTC corresponding to <1 mg/l)? How does this affect (i) the global identifiability analysis and (ii) the dynamic identifiability? Previously alternative objective functions have been suggested, including RMSE with a logarithmic (e.g. Ward et al., 2018) or mixed scale (e.g. Bottacin-Busolin et al., 2011; Riml et al., 2013), to account for the different magnitudes of the concentrations across the BTC. I suggest that the authors assess this and discuss the implications.
Authors: We used RMSE because it is an equivalent form of RSS that is used in the calibration of OTIS-P and of mean absolute error used in the dynamic identifiability analysis. Thus, RMSE allowed us a consistent comparison of our TSM results with OTIS-P and DYNIA outcomes. For the same reason, RMSE is the preferred objective function in studies where random-sampling TSM simulation are compared to OTIS-P results (Ward et al., 2017) or coupled with DYNIA approach (Wlostowski et al., 2013). We will clarify our choice in the revised manuscript.
However, our model tested several objective functions for every TSM iteration (Nash-Sutcliff efficiency, Pearson's R2, Kling–Gupta efficiency, normalized RMSE, log-transformed RMSE, log-transformed R2, log-transformed NSE). As expected, there are some consistencies and some inconsistencies between the different objective functions. This has been shown for other modelling applications (Gupta et al., 2009; Ouyang et al. 2014; Wagener et al., 2002). It is surely a valuable follow-up study to evaluate the consistency of identifiability of TSM parameters as function of the selected objective function, as this problem has never been directly tackled in TSM. However, this analysis is far beyond the scope of the paper. We will highlight this important point in the revised discussion section.

R3: "Moreover, I miss a visual comparison of the observed and simulated BTC as a complement to the presented RMSE values, preferable using log-transformed concentrations to highlight how well the

model captures the tail of BTC that is argued to be of importance for transient storage processes (e.g. lines 55-57)."

Authors: We will add in the appendix the following figure (Figure A). Despite the achieved parameter identifiability and the relatively low RMSE values obtained at the end of the proposed iterative approach, it is evident that the best-fitting BTCs are unable to reproduce the last section of the tail of the observed BTCs. This might be driven by the selected exponential RTD in the formulation of the TSM, and/or by the selected objective function (RMSE). As an example, the best-fitting BTC obtained at the end of the second TSM iteration shows a visually better fit on the BTC tail (Figure B) despite the large RMSE (1.5197 mg/l). We will implement the manuscript discussing the role of the used objective function and of the exponential RTD in the BTC fitting. We will extend the discussion accordingly.

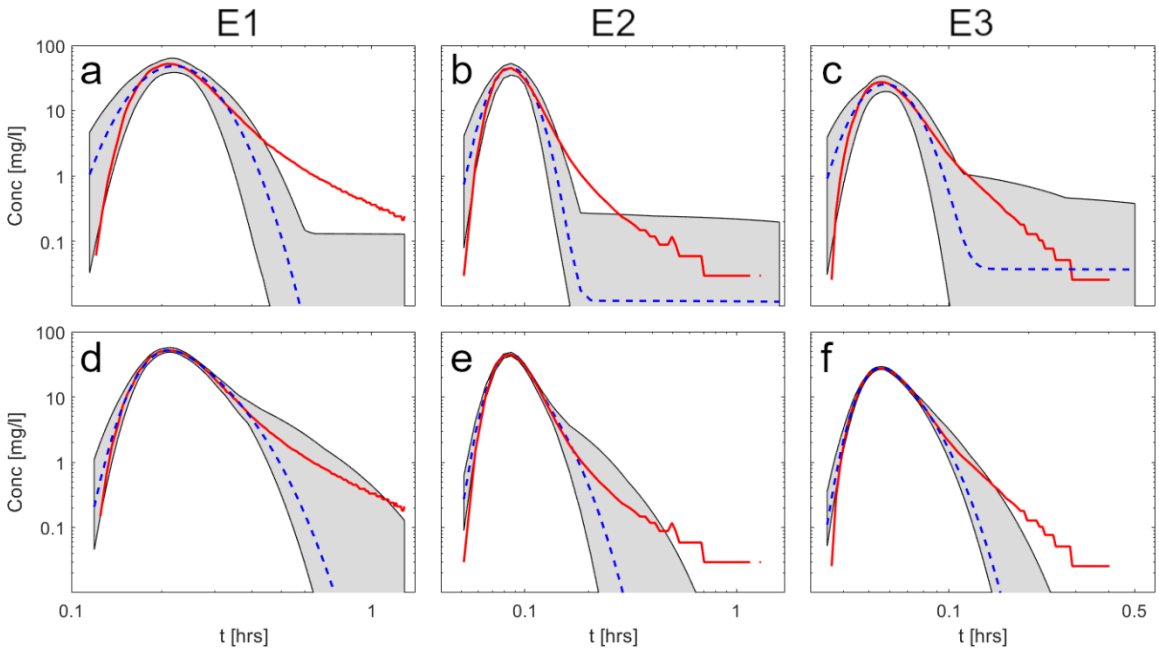

Figure A: Observed BTC (red line) together with the grey area comprised between the top 100 simulated BTCs and the best-fitting BTC (blue dashed line) for (a, d) E1, (b, e) E2, and (c, f) E3. Results reported for the first (a, b, c) and last (d, e, f) TSM iterations.

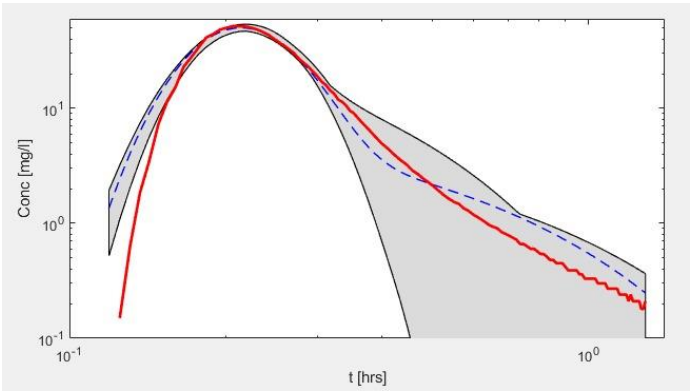

Figure B: Observed BTC (red line) together with the grey area comprised between the top 100 simulated BTCs and the best-fitting BTC (blue dashed line) for the second TSM iteration (E1).

R3: "The fact that using erroneous model parameter estimates (obtained either from the literature, from a simplified model (ADE) or from a Monte-Carlo simulation with to wide parameter ranges and/or not sufficient iterations) leads to uncertainty/errors when estimating the transport metrics (Eqn 5-8) is rather intuitive. Firstly, I find (the rather long) discussion in Section 4.1 as well as the conclusion (line

24-26) and abstract (lines 21-26) about misinterpretations/uncertainty when comparing the different models "unfair". The conditions for the OTIS-MCAT simulations seems to be equivalent to the first iteration of the proposed methodology. Thus the conditions when OTIS-MCAT was used differ substantially from the 3-4 iterations with a successive refinement of the parameter ranges of the proposed methodology. I understand that the authors would like to make a point and compare the results against an existing model framework, but I think that manuscript would benefit from significantly downplaying the role of OTIS-MCAT. I would prefer a stronger focus on how the refinement of parameter ranges using the existing model framework resulted in a reduced uncertainty/increased identifiability of the model parameters."

Authors: We fully agree with R3. One of the major points of the manuscript is the show how our approach can clearly reduce the non-identifiability of the parameters via the proposed iterative approach. We will revise the manuscript to emphasize the benefits of the refinement of the parameter range to increase parameter identifiability. To do so, we will include in the appendix the following figure (Figure C). This figure emphasizes the role of the defined parameter range over the role of the number of parameter sets. In the original version of the manuscript, we felt the need to compare our results with a currently existing framework (OTIS-MCAT). However, in the revised manuscript we will downplay the results of OTIS-MCAT, and we will rather emphasize the role of the iterative narrowing of the parameter range to achieve parameter identifiability.

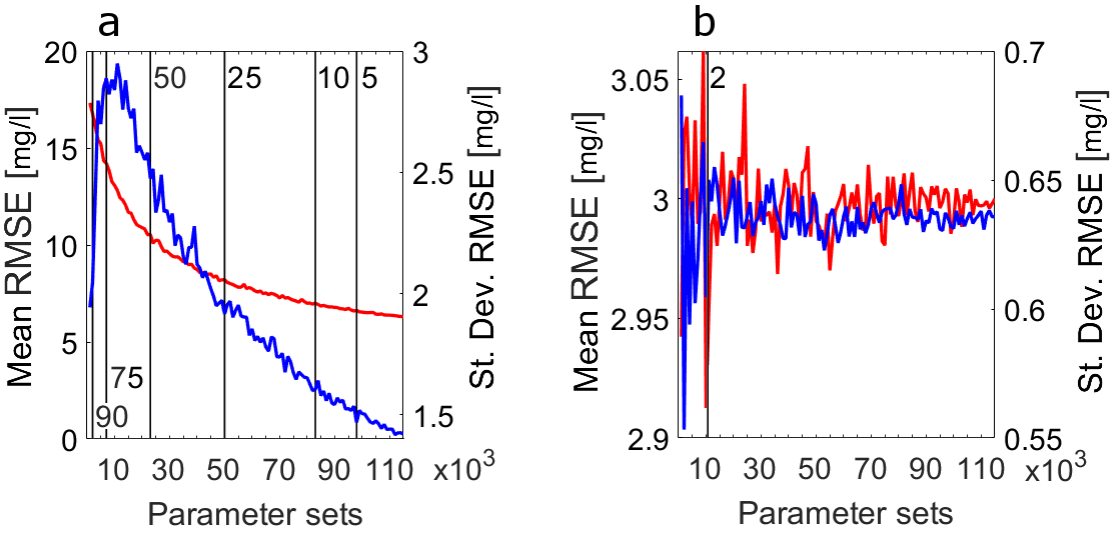

Figure C: Mean (red lines, left axes) and standard deviation (blue lines, right axes) for RMSE values relative to the top 10% of the modelling results as a function of the number of parameter sets used in the TSM. The results are reported for the (a) first TSM iteration and the (b) last TSM iteration (E1). Vertical black lines indicate the number of parameter sets needed to have the shown percentage difference between the mean RMSE value calculated at a certain number of parameter sets and at 115'000 parameter sets. Eg: In plot (a) after 50'000 parameter sets there is less than 25% difference in top 10% RMSE values compared to results using 115'000 parameter sets.

R3: "Secondly, although the authors successfully reduced the uncertainty of the model parameters by an iterative and smart sampling of the parameter space, it was surprising to see that the results from the OTIS-P outperformed (2 out of 3 experiments) the proposed methodology when using the objective function preferred by the authors (RMSE, table 2). This is something that is not sufficiently discussed in the paper and, in my view, opens for questions when the iterative sampling procedure is needed."

Authors: We thank R3 for this comment. Indeed, the similarity with OTIS-P results is interesting and we will discuss this more clearly in the revised manuscript. This finding raises the question *if* and *when*

the random sampling approach for TSM is really needed and *if* and *when* OTIS-P results can be considered reliable.

R3: "Could the DYNIA approach be combined with OTIS-P using a given confidence interval as input for the parameters ranges to assess the identifiability in parameter estimates from OTIS-P?
Authors: We thank R3 for this comment. The idea to use the DYNIA together OTIS-P is really interesting and could be a novel modelling practice and a strong follow-up of this project.

R3: "I guess that similarity in performance between parameters obtained by OTIS-P and the proposed sampling procedure might connect to the objective function used to evaluate the performance (see major comment #3), the limited number of experiments (in a single reach) and how initial values (and the possibility of finding a local minimum when optimizing) were defined in the OTIS-P model."
Authors: We will implement the discussion with these observations. However, OTIS-P was used starting from several initial values and it was iteratively run until convergence of the model results. Thus, we exclude the finding of local minimum during the optimization. We will clarify this in the revised manuscript.

R3: "I believe that the paper lacks a thorough discussion regarding different model representations of transient storage. Eq. 3 assumes an exponential residence time distribution (RTD) in the transient storage zone as originally defined in the TSM model (e.g. Bencala and Walters 1983). Subsequently other type of exchange models and RTDs in the transient storage zone have been introduced (e.g. Wörman et al., 2002; Haggerty et al., 2002). I see a great advantage of the proposed model framework – compared to the OTIS-MCAT and OTIS-P – to explore alternative model formulations (including multiple transient storage zones) and alternative RTDs. This flexibility, when discussed properly, could levitate the readers understanding of the usefulness of the model framework."
Authors: This is a very interesting suggestion. As also suggested by R1, our discussion section lacks a paragraph exploring modelling implication for reactive solutes or for models with different RTD or with multiple storage zones. We will implement the discussion section with a paragraph exploring possible implication for other TSM formulations.

Detailed comments:

R3: "Line 26-29: This is unclear, what is meant by "clear potential"?"
Authors: By "clear potential for increasing identifiability" we referred to the used iterative modelling approach to improve identifiability of TSM parameters. We will clarify.

R3: "Line 53: "The numerous contradictory outcomes", in what context? Please clarify"
Authors: We referred to contrasting interpretation of TSM outcomes and the change of TSM parameters values with hydrologic conditions and scales. We will clarify this in the revised introduction.

R3: "Line 109-111: Unclear. "to keep constant a rather identifiable parameter". Please rephrase."
Authors: We will reformulate this sentence.

R3: "Eq. 3: I believe that CS should be replaced by CTS in the bottom equation
Authors: We will correct it in the revised manuscript.

R3: "Line 216: I miss information of the width of the window in the DYNIA.
Authors: Following Wagener et al. (2002), we used a window size of three time steps (~1 min for E1 and E2, and ~15 secs for E3). We will clarify this detail in the revised manuscript.

R3: "Line 233-234 "The best 1% of the results were used to define its parameter space in the successive TSM iteration". Not in agreement with Figure 1 that says "New parameter range defined from the top 10% of the results". Please revise.
Authors: The new parameter range is defined from the top 10% of the results. We will revise in the text.

R3: "Line 245-246: vpeak is not clearly defined. Is this the time from injection to the BTC peak divided by the stream length (i.e. 55 m)? Please clarify."
Authors: We will clarify that v_peak is the velocity obtained via v_peak=L/tpeak. Where L is the length of the investigated reach and tpeak is the time for concentration peak to travel through study reach. We used v_peak as it is commonly used in many transient storage studies (Ward et al., 2013; Kelleher et al., 2013; Wlostowski et al., 2017; Ward et al., 2017; Ward et al., 2019a; Ward et al., 2019b).

R3: "Line 249: How was the set up of the OTIS-P simulations in terms of initial parameter values? Does "multiple OTIS-P iterations" mean that several initial conditions were tested to reduce the risk of ending up with in local minimum in the optimization? Please clarify."
Authors: We will clarify that we used different initial parameter values to avoid a local minimum in the OTIS-P simulation. Following Runkel (2008), we ran OTIS-P multiple times and used, as initial values of a subsequent run, the final estimate from the previous run. We finalized the simulations when the parameter estimates changed less than 0.1% between subsequent runs. We will clarify this in the revised manuscript.

R3: "Line 255-258: What parameter ranges were used and how many iterations were performed with the OTIS-MCAT? It is difficult to compare the results if the simulation conditions are not provided."
Authors: OTIS-MCAT results are obtained using the range indicated in Table 1 with the condition of velocity considered as a fixed-parameter equal to v_peak. Thus, OTIS-MCAT results are the results of the 1st TSM iteration for the case v=v_peak (cf. lines 257-258).

R3: "Line 293-295: From Figure 3, it is unclear how ATS < 5.356 m2 has information content > 0.66, Figure 3 i,j). Previously (line 223-224) it is stated that the information content is expressed as one minus the width of the 90 confidence interval, which I assume uses the entire parameter distribution. Please clarify why there is no lower bound on the confidence interval."
Authors: In the revised results we will specify that the lower boundaries of the confidence interval for A_TS on the tail of the BTC were 0.77 m^2 for E1, 0.32 m^2 for E2, and 0.33 m^2 for E3 after the first TSM iteration.

R3: "Moreover, although I realize that this is the first iteration, to have a transient storage area several orders of magnitude larger than the cross-sectional area of the stream makes little sense."
Authors: The rather large order of magnitude (10^1 m^2) for the A_TS upper limit in the first TSM iteration is justified by two observations: 1) groundwater monitoring in the study site let us observe the occurrence of water gradients pointing from the stream channel toward the adjacent groundwater several meters away the stream talweg (see red arrows on the right bank in figure 7f in Bonanno et al., 2021). This suggests the occurrence of a hyporheic zone adjacent the stream channel potentially extending for more than 10 m^2 despite the relatively small size of the investigates reach; 2) The chosen order of magnitude is consistent with previous literature for TSM applied in headwater stream reaches (Wagner and Harvey, 1997; Ward et al., 2013; cf. Table 1).

R3: "Figure 3. I suggest to show the y-axis of the alpha plot (Figure 3 g) using a log scale, due to the small values."
Authors: We will modify this plot as suggested by R3.

R3: "Line 312-314: How much of this result can be derived to the used parameter ranges and the number of iterations? If the MC analysis would been set up differently (smaller parameter ranges, larger number of iterations), how would the result differ?"
Authors: We outlined the role of parameter range and number of parameter sets in the answer before by introducing Figure C. This figure will be inserted in the revised Appendix. We believe Figure C highlights the pivotal role of the parameter range over the number of parameter sets. It also increases the value of the used iterative approach in constraining the parameter range in random sampling TSM.

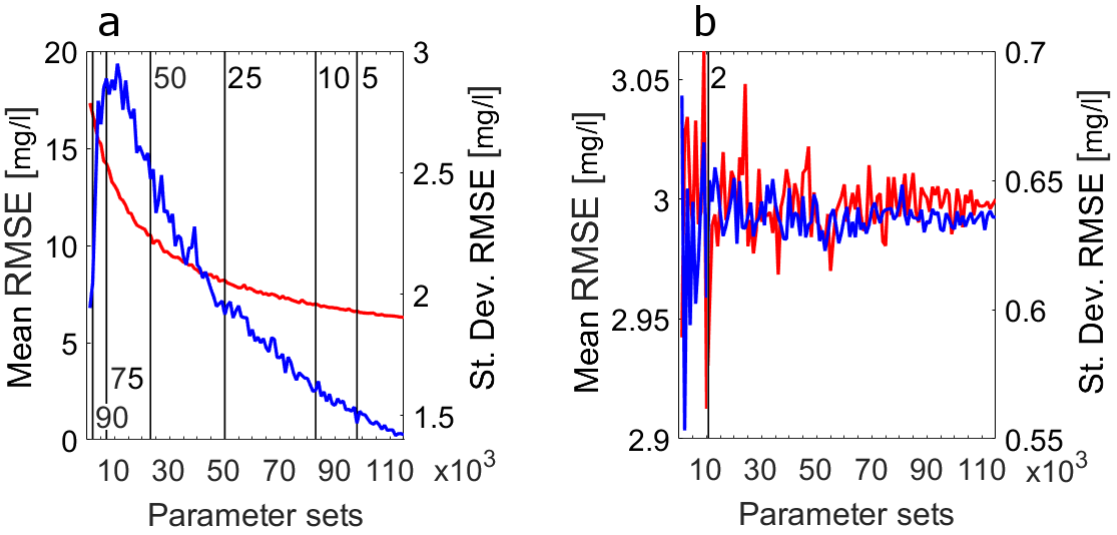

Figure C: Mean (red lines, left axes) and standard deviation (blue lines, right axes) for RMSE values relative to the top 10% of the modelling results as a function of the number of parameter sets used in the TSM. The results are reported for the (a) first TSM iteration and the (b) last TSM iteration (E1). Vertical black lines indicate the number of parameter sets needed to have the shown percentage difference between the mean RMSE value calculated at a certain number of parameter sets and at 115'000 parameter sets. Eg: In plot (a) after 50'000 parameter sets there is less than 25% difference in top 10% RMSE values compared to results using 115'000 parameter sets.

R3: "Line 338: "orange boxplots" is labeled "red boxplots" in Figure 7. Please revise."
Authors: We will revise it.