**Report on the manuscript "Identification of Parameter Importance for Benzene Transport in the Unsaturated Zone Using Global Sensitivity Analysis" submitted to *Hydrology and Earth System Sciences* (all authors answers are in blue font).**

It is my pleasure to review the manuscript entitled "Identification of Parameter Importance for Benzene Transport in the Unsaturated Zone Using Global Sensitivity Analysis". How to accurately quantify the parameter importance in the complex contaminant transport models has always been an important topic in groundwater research. In this study, GSA methods of Morris and Sobol were implemented to investigate the important parameter for benzene transport model in the unsaturated zone.

I believe this paper is well written with high quality and good logic. I would be in favor of publication after the author addressed the comments given below.

**Thank you for the positive evaluation.**

**Major Comments:**

1. In the section 2.3.2, I believe authors should provide more details about the calculation process and algorithm implemented for the Sobol indices. And I believe it is too obvious that the sample sizes were dramatically different for different problems since they were calculating different variances based on different models, it is unnecessary to list these different size numbers.

More details about the Sobol indices calculation process were in the Appendix (A3). Following the comment above, those details were moved to the body of the text (line 270 in the revised manuscript), which now reads as:

> "Parameter spaces were sampled using the Sobol quasi-random, cross-sampling strategy (Sobol, 2001). Rather than generating random numbers, this technique generates a uniform distribution in the probability space. The distribution appears qualitatively random, but sampling only takes place in regions of the probability function that were not previously sampled.
>
> In order to assess the accuracy of the Sobol indices, confidence intervals of the indices should be constructed. The analytical procedure for confidence interval calculation involves repeating the model runs several times, which is too time consuming and computationally demanding in this case. Therefore the bootstrapping approach was used instead (Efron & Tibshirani, 1986). Archer et al. (1997) suggested using bootstrap confidence intervals to produce confidence intervals of complicated data structures. The bootstrapping approach is based on resampling the parameter space of the already available data many times with replacement (randomly selecting values and allowing for duplicates), and constructing a distribution of the output (Efron & Tibshirani, 1986). Here, resamples were taken from the existing dataset with replacement, and the indices' values were recalculated. That gives an estimate of the mean and variance of each of the indices and allows calculation of the confidence interval. The method thus relies on computational cost rather than on an analytical cost (running the model again). Here, the samples used for the model evaluation were resampled 1000 times with replacement, and 95% confidence intervals were constructed (Archer et al., 1997).

Still, confidence intervals for the first-order indices (S1), with the Sobol sampling method gave values of more than 100%. This was also observed by Brunetti et al. (2016) and Hartmann et al. (2018) who also studied transport in unsaturated media. This may be a result of insufficient sample size, since Sobol's convergence requires a very large sample size (Saltelli et al., 2004). Therefore, here the S1 values were extracted using the delta method of Plischke et al. (2013), calculating S1 values from a given data through emulators and bootstrapping rather than running the model itself multiple times"

The details of the Sobol sample size were moved to the appendix.

2. I am not sure why the authors focused on the crashed simulations; it seems to be the problem of numerical model instead of sensitivity analysis problem. And the differences of GSA results using different methods to fulfill these "bad" input samples are more like pure numerical fluctuations to me.

The problem of crashed simulations is indeed a problem of numerical instability rather than a sensitivity analysis problem. However, as was demonstrated previously by Sheikholeslami et al. (2019), and by Clark and Kavetski, (2010) and as we also show in this study, these numerical artefacts are very difficult to tackle when running large number of simulations on large parameter space and they can affect the assessment of parameter sensitivities. This is because GSA sampling order and size is important for the GSA results. Therefore, it is important to devise solutions that minimize the effect of crashes on GSA (Razavi et al., 2021). Yet, very few strategies for handling simulation crashes have been proposed in the literature and identified for their shortcomings. This is why this data should be handled carefully and the way to treat this data in GSA is of importance.

3. In the section 3.2, have the authors tried the geostatistics tools to build different samples of heterogeneous media structure? I don't understand how the different samples representing the vadose zone media were generated. For the sentence in line 464, I believe Dai et al., (2017) has done some similar work and please check if it is helpful to improve this research for the heterogeneous media structure generation through geostatistics implementation.

We thank the reviewer for pointing our attention to Dai et al. (2017). Dai et al. (2017), used geostatistical approach to estimate the spatial distributions of three main parameters from point measurements. It was used to estimate the elevations of the contact between aquifer and aquitard from geological logs, the hourly head boundary conditions from monitored water elevations, and the hydraulic conductivity field from permeability data.

Dai et al. (2017), relied on the results of comprehensive field characterization measurements, and numerous field experiments and modeling studies performed at their site which provided them with a strong technical basis for their test case. They used that data to find the sources of parameters uncertainty through time and space in a 3-D model representing their site and solute transport in a ~200-hours tracers test.

Unlike the study of Dai et al. (2017), who examined the uncertainty of 3 parameters over a 3-D space in a specific site and limited time frame, in this study we deal with a more general problem examining the risk for contamination of the entire Israel's coastal plain vadoze zone over the course of 50 years and the importance of 17 parameters. While it is possible that geostatistcal tools could have provided a more realistic representation of the layers in each site, interpolation and representation of the rest of the space between the sites would require more data that we lacked and would have made the model and GSA much more computationally demanding than it already was. The more general objective of the study, has led us to adopt a more general 1-D representation of our space, where the ranges of the number of clay layers and their thickness were taken from field data.

For that purpose, we used a dataset of soil data of contaminated fuel sites along Israel coastal plain. From the dataset we extracted the average number and thickness of clay layers interbeds along the vadoze zone at each site, and calculated the total average for all sites. We then examined how much the number and thickness of these interbeds can affect benzene transport using the GSA.

Regarding line 464, we rephrased it to clarify we are dealing with clay interbeds and added the spatial distribution uncertainty tested by Dai et al., (2017) to the text (line 491 in the revised manuscript) which now reads as follow: "

We are not aware of other studies that tested the distribution of clay layers interbeds in a SA for contaminants transport. Yet, Dai et al., (2017) tested the spatial distribution uncertainty of other parameters in a GSA like the elevations of the contact between aquifer and aquitard, the hourly head boundary conditions and the hydraulic conductivity field.

**Minor Comments:**

1. Line 84-85: I don't think these reference papers are all focus on the sensitivity analysis for the unsaturated zone.

   True, thank you very much for this remark - Ciriello et al., 2013 considered only saturated flow and was therefore removed. All other SA were done for unsaturated soils (line 85).

2. Line 90: I don't understand the term "properties of benzene itself", what properties?

   Properties related to benzene and not only to the porous media like - the degradation rate or adsorption coefficient. Rephrased in the text (line 91 in the corrected manuscript).

3. Line 99-100: I believe authors should provide more details about these two methods, especially about their algorithms, in the introduction.

   We appreciate the comment. However, it seems to us that the introduction section is long enough as is. All the details regarding the two methods appear in section 2.3 in the Materials and methods section. We added to the introduction a statement that more details regarding the two methods can be found in section 2.3 (line 102 in the revised manuscript) and it now reads as follow:

“Two GSA methods were tested for the homogenous media simulations, to analyze the parameter importance: the Morris method (Morris, 1991), a reliable, computationally-cheap alternative to variance-based GSA; and the Sobol method (Sobol, 2001), a computationally heavy, variance-based GSA (see Material and Methods section 2.3 for a description of these methods).”

4. Line 103-104: please provide some references, I don’t know this is a common problem.

   (Razavi et al., 2021; Sheikholeslami et al., 2019) – were added as references (line 107 in the corrected manuscript). These authors suggested this is a common, though usually overlooked problem.

5. Some formats of titles of subsections are incorrect or inconsistent, please check all of them.

   Thank you very much, we went over it and fixed it.

**Reference**


Clark, M. P. and Kavetski, D.: Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes, Water Resour. Res., 46, 2010.

Dai, H., Chen, X., Ye, M., Song, X., and Zachara, J. M.: A geostatistics-informed hierarchical sensitivity analysis method for complex groundwater flow and transport modeling, Water Resour. Res., 53, 4327– 4343, https://doi.org/10.1111/j.1752-1688.1969.tb04897.x, 2017.

Razavi, S., Jakeman, A., Saltelli, A., Prieur, C., Iooss, B., Borgonovo, E., Plischke, E., Lo Piano, S., Iwanaga, T., Becker, W., Tarantola, S., Guillaume, J. H. A., Jakeman, J., Gupta, H., Melillo, N., Rabitti, G., Chabridon, V., Duan, Q., Sun, X., Smith, S., Sheikholeslami, R., Hosseini, N., Asadzadeh, M., Puy, A., Kucherenko, S., and Maier, H. R.: The Future of Sensitivity Analysis: An essential discipline for systems modeling and policy support, Environ. Model. Softw., 137, https://doi.org/10.1016/j.envsoft.2020.104954, 2021.

Sheikholeslami, R., Razavi, S., and Haghnegahdar, A.: What should we do when a model crashes? Recommendations for global sensitivity analysis of Earth and environmental systems models, Geosci. Model Dev., 12, 4275–4296, https://doi.org/10.5194/gmd-12-4275-2019, 2019.