I had the pleasure to read the paper “Incorporating experimentally derived streamflow contributions into model parameterization to improve discharge prediction” published in HESSD under https://doi.org/10.5194/hess-2021-179 by Hartmann et al. The paper is generally well-written and the quest for reducing model uncertainty and increasing model realism is of interest to the HESS readership. I found the combination of a relatively simple conceptual rainfall-runoff model constrained by observation-based water source contribution estimates worthwhile and adequate to respond the research questions asked.

Having said that, I have some comments and suggestions that I put forward for the authors to consider:

- While I appreciate the tracer sampling effort and new data from a lesser known study site, the paper structure does not reflect this at all. The mixing model and some mentioning of the tracer data used appears in the study site description even with the results of the water source contributions. I strongly suggest to separately put these into the methods and later into results assuming this data and analysis was not previously published (no reference suggests this!). Furthermore, it would be instructive to actually see some of this data to get a notion of the space-time variability, e.g. in form of a bi-variate plot since the water source estimates are crucial for the analysis. I also wonder why no throughfall end-member was included in the mixing model and if the 20% margin for F_{HZ} and F_{GW} used to accept/reject models is based on an uncertainty estimate of the mixing model results (none are presented in Table 1)?

- The fact that throughfall was sampled made me wonder about the importance of interception at this forested catchment and the effect it might have on the modelling since this is not included in the model structure. I also have a bit of an issue with the model structure itself and how the storage outflows are used to represent water sources: The model is essentially lumped with a vertical two-storage cascade fed by a soil reservoir that re-distributes the water for runoff generation. Now, I was thinking conceptually that all the hillslope outflow must feed into the riparian zone and from there runoff is generated that together with the groundwater flow constitutes streamflow (two end-members only). However, the latter would require a minimal semi-distributed model structure with two-storages in parallel (hillslopes draining into the riparian zone) and a third groundwater reservoir. In contrast, your HZ and GW is coming from the same source (storage V1). I would definitely appreciate some more explanations here.

- My main concern however, is that the paper falls a little short in terms of the analysis related to many assumptions that are currently not sufficiently justified. For example, the choice of the KGE statistic that clearly influences the validation of mostly low flows in 2014, which is almost unfair as there is visibly no information content in these measurements. Without tracer measurements for 2014 it almost bags the question of why 2014 was included in the first place. The threshold of KGE>0.8 to accept models seems arbitrary. All three recession constants have the same initial parameter limits, but you would certainly accept a slower response of the groundwater reservoir outflow. Or you could think of fixing the groundwater recession constant based on a Master recession curve such as suggested in work by Hrachowitz et al. On the importance of the modeler’s choices. As a matter of fact there is more literature on previous work (you could potentially cite) that attempted to reduce parameter uncertainty through constraining parameters with additional information such as tracers that did not necessarily included the need for more model complexity in terms of number of parameters. I would therefore suggest to try and test different statistics to see how they perform and apply the different criteria for model parameter selection also to the full 2million parameter sets for a more comprehensive assessment of information content. Furthermore, throughout the paper you suggest quantitative assessments of information content, uncertainty in the context of a likelihood-weighted uncertainty estimate (GLUE), parameter identifiability and sensitivity, but this was not really done. Here, I would suggest to consistently use terminology and maybe provide some quantitative analysis such as e.g. a Shannon criterion for information content and/or a sensitivity metric such as Sobol and/or a measure of the width of the likelihood-weighted uncertainty bound used for prediction. With that you more comprehensively support your interpretations and allow the reader to really assess your statements in the discussion and conclusion.

- Figure 5 is quite hard to interpret and I suggest to use a log-scale for streamflow visualization.

- There are some occasions in the paper where you wrote “be”, but I think it should be “by”.

For the above reasons, I would recommend major revisions before potential publication of this paper.

Sincerely,
Christian Birkel