Review of “Towards a community-wide effort for benchmarking in subsurface hydrological inversion: benchmarking cases, high-fidelity reference solutions, procedure and first comparison”

General comments

The paper presents a very interesting topic and can be useful for the community. I think it needs some clarification in the writing as it is difficult to follow the full discussion. There is the need to provide some clarification for when the authors reference to the truth case, the reference and the benchmark scenarios. Consistency on how the different conditions are referred to will help the reader.

One question is about the added information that would come from the two synthetic truths: they only have a different standard deviation and I am not sure there is much to gain from comparing the two of them. I wonder if it would be more useful to compare different hydrogeological conditions. Something like bookend scenarios that look at very different assumptions would be very beneficial for the reader to understand the power of the methods presented.

Looking at the scenarios in table 3, the regular well distribution scenario does not seem to add more information and S0, S2 and S3 do not really represent a variety of conditions.

My comments below are mostly about Figures: the paper would really benefit from a more consistent representation and discussion of the results and more consistency in the figures with results.

I would also suggest to provide some more informative conclusions, maybe summarized in a bullet point lists.

Specific comments

Line: suggest to change to “across multiple scales”

Line 37: I would not say that data scarcity and subsurface heterogeneity are the sole responsible for uncertainty. Depending on the model, we need to add uncertainty on other terms of the water budget.

Line 413: why S1 is a fallback scenario? I wonder if S1 is needed, and if yes I think there is the need to have at least to scenarios with regular well distribution to be able to make a comparison

Figure 2: as mentioned earlier: is it a limiting factor having two truths that are so similar and only different for standard deviation?

Figure 3: I think that the use of a different scale for S2 is misleading, but if the authors believe that it is important I would suggest to add explanation

Figure 4: same comment on the scale

Figure 5: same comment: it would be good to add an explanation

Line 514: it is really hard to see the sharp decrease

Figure 6: the legend for S0 says “reference” and for the other scenarios “Synthetic truth”. Is this correct? Also, are there some conclusion to be derived from the mean always being above the reference/synthetic truth?

Figure 9: is the reference solution, the true solution?

Technical corrections

No technical corrections.