REVIEWER 3: Anna Sikorska-Senoner

GENERAL COMMENTS

This paper proposes a novel method for benchmarking uncertainty in river flow simulations via using novel deep learning (DL) methods and an extensive sample of 531 catchments. The manuscript is generally well written and structured and it is of a value for hydrological community and HESS readers. The great value of this work is a combination of a large sample study with novel deep learning methods for benchmarking uncertainty in rainfall-runoff models. Nevertheless, some issues as described below should be addressed before possible publication. Thus, my recommendation is a moderate to major revision.

SPECIFIC COMMENTS

1. The authors based their analysis on a large sample of CAMELS catchments (subset of 531 catchments), which gives a great potential for the analysis they are conducting. Thus, I found it a bit disappointing to see the results of the analysis reported only as averaged values (i.e. averaged over all catchments). I think the usage of such a large sample together with novel DL methods here applied creates a great potential to present their results in a bit more detailed way. For instance, evaluation metrics or probability plots could be presented not only for the averaged values but also giving some sample details. One way could be to present ensemble of probability plots or some ranges to give a reader a better feeling about the individual catchments' results. In a similar way, tabular values could be presented for some ranges and not only for averaged values.

Adding variance to the evaluation table is a great idea and we will include it in the revised manuscript.

Regarding the suggestion to include an ensemble of probability plots, Figure 9 in the original manuscript shows almost the same information: it shows the distributions of the results for the different quantiles over the different basins. This plot actually provides a *more* detailed look at the different quintiles and basins than an ensemble QQ plot would. We did actually try an ensemble of probability plots (before the original submission) and packing ranges or all solutions directly into the probability plot was more confusing than helpful, which is why we decided to show the results over the different quantiles and as deviations from the 1:1 line in the form of densities (which makes it much easier and comparable than point clouds or line-plots). Anyway, it is very difficult to present results for 531 basins in a constructive way, but the paper does already include almost exactly the information that the reviewer requested.

2. It is not quite clear, which period the reported values of results for four tested models referred to. Ideally, values and plots could be presented for all three periods, i.e., for training, validation and test periods with sufficient details (see comment #1).

Thank you for pointing this out. As is convention for DL/ML approaches, we only report the test period. Training and validation periods can exhibit arbitrarily good performance, thus it is generally discouraged to report the model performance there. We will add a statement about all statistics (except hypertuning) being from the test period to all figures, tables and the textual description to make this clear for readers.

3. The method section is very well written. However it provides mostly details from a single catchment perspective. Some additional details for a large sample study, as used here, would be very useful, specifically for readers without sufficient background in the methods applied here.

Thank you for this compliment. Alas, it might hint at a deficiency in our description as no part of the method section was written with a single-catchment perspective in mind. That is, the distributional predictions are certainly made for each basin and time-steps, but the DL based model as such is and should not be trained on the basis of individual basins. As is stated in line 191, all (training) data from all 531 catchments were used to train each model. Training a model per-basin would yield bad solutions.

4. Finally, I agree with both previous reviewers that a comparison to other simpler data-driven model(s) would be very useful for assessing the methods presented here. At the current stage, one can only see which method among four tested performs best. However it is difficult to judge their overall value as a comparison to simpler methods is missing. Such analysis would also add a value to the "Conclusions and Outlook" section.

Please see our answer to John Quilty's general comments and his specific comment 4.

MINOR COMMENTS

Figure 1: make clear whether the figure presents all CAMELS catchments or the subset you used in this study.

Thank you. The figure shows the entire CAMELS dataset. We will make sure that this becomes clear in the revised manuscript version.

Figure 2: add a & b in the figure caption for a higher readability.

Good idea, we will do this.

Table 1: remove the index a with its notation as it duplicates information from the figure caption.

It does not, but we are happy to move the information to the table caption nevertheless.

Figure 7: what is 'clipping' here? It is also not quite clear what m and n refer to. Maybe it would be easier to present figure as a scheme, when example is given for a basin 1, 2, … and then n=531. Also it should be: "In total we have 531 basins….". Add t to "For each time step t we…"

Clipping here means that samples that are below zero are set to zero. We will mention this in the figure description and weave in your suggestions.

Line 201: why do you take 7500 samples and not any other number?

The number itself is however not crucial here. We tested different cutoff-points for the sampling during the preparation of the manuscript, both by sampling different amounts of points and by using a Gaussian simulation (so that we can control the

actual underlying uncertainty. This way we found that at around 5000 points the evaluation was relatively stable. To this we added 2500 points as a margin of safety and thus obtained the 7500. The number might thus be seen as a compromise between a relatively small number of samples provided and a relatively stable statistical estimation that can be derived from the samples.

Table 3: Text ”a) All metrics are computed for the samples of each timestep and then averaged over time and basins.” could be removed as it is already mentioned in the table caption.

We will remove it. Our understanding is that table captions should not present new information (except about how to read the table).

Table 4: for which period are these values presented?

We would only ever report values for the test period. No paper should ever, under any circumstances, report values for training periods, unless there is a particular and clearly stated reason. We will mention this in the table caption, but it is redundant with strict rules of practice.

Figure 10: the figure presents an example of an event of some catchment. Maybe it could be useful to pick up one catchment as an example and provide detailed results for this catchment from probability plots to events.

Albeit interesting, that would be a post-hoc model examination with a different goal. We do not see how it contributes at this point.

Conclusions and Outlook: as there is no discussion section, this part could be extended. Particularly, the discussion of obtained (averaged) results is quite vague.

This part would also benefit from comparing the tested methods to a simpler data-driven model.

We will extend the conclusions and outlook with regard to the limits of diagnostics. That said, we are not aware of simpler data-driven models that could be used in this context or would be beneficial here. The proposed approaches are quite simple (either a direct estimation of the likelihood or a sampling based approach that can be used for models that estimate the maximum likelihood) and can be used in context with all models that are differentiable and able to provide the necessary estimates. Finally, ad hoc benchmarking is antithetical to what we view as critical scientific ethics, as discussed in our responses to reviewer #1.

Line 417: remove the word ‘single’ which is used twice.

Will be removed.

Line 430: the expression ‘the training data’ is used twice.

Will be removed.

Line 438: the word ‘intermediate’ is used twice.

Will be removed.