Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC3
https://doi.org/10.5194/hess-2021-154-RC3, 2021

Comment on hess-2021-154

Anna E. Sikorska-Senoner (Referee)

Referee comment on "Uncertainty estimation with deep learning for rainfall–runoff modeling" by Daniel Klotz et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-154-RC3, 2021

Referee's 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

- 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.
- 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).
- 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.

- 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.

Minor comments

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

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

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

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…"

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

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.

Table 4: for which period are these values presented?

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.

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.

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

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

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

Best regards,
Anna Sikorska-Senoner