Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2021-154-RC2, 2021

Comment on hess-2021-154

Anonymous Referee #2

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-RC2, 2021

Summary
This paper focuses on the use of several −mostly new for hydrology− concepts and methods from the machine and deep learning fields for uncertainty quantification in rainfall-runoff modelling. Specifically, it presents a large-scale application of these concepts and methods under a new framework. This large-scale application can be used as a guide for future works wishing to apply these (or similar) concepts and methods.

General comments
Overall, I believe that the paper is meaningful, very interesting, and well-prepared in general terms with room for improvements.
I recommend major revisions. To my view, these revisions should (mainly but not exclusively) be made in the following key directions for the paper to reach its best possible shape:
a) Key direction #1 (for details, see specific comments #1,2): To my view, the work's background should be better covered. In fact, to my knowledge there are two very relevant published studies, additionally to the studies already included in the "Introduction" section, that use LSTMs for uncertainty assessment in hydrological modelling. Also, there are research works presenting machine learning concepts and algorithms for uncertainty assessment in hydrological modelling (e.g., for probabilistic hydrological post-processing), including some few ones that conduct large-scale benchmark experiments using data from hundreds of catchments and several machine learning models (thereby also introducing benchmark procedures, which I agree that are very rare in the field and very important). Further, I would say that the connection with the machine and deep learning fields needs to be better highlighted as well.
b) Key direction #2 (for details, see specific comment #3): I agree with the main point raised by the other reviewer (Dr John Quilty). To my view, only through a comparison of the four deep learning methods of the paper to other statistical and machine learning methods providing probabilistic predictions (with the latter methods playing the role of benchmarks) the paper will fully achieve its aims in terms of benchmarking. I believe that this is absolutely necessary, as (i) the paper devotes a lot of space discussing its benchmarking contribution (but easier-to-apply methods are currently missing from its contents, while they have already been exploited for probabilistic hydrological modelling) and, (ii) the paper indeed offers interesting results which would mean more to the reader if compared to the results provided by easier-to-apply statistical and machine learning

methods.
c) Key direction #3 (for details, see specific comment #4): To my view, proper scores (see e.g., Gneiting and Raftery 2007) should necessarily be computed for assessing the issued probabilistic predictions. Currently, there is an important −from a practical point of view− aspect of this work's large-scale results that is not assessed. In fact, the selected scores cannot directly and objectively inform the forecaster-practitioner which method to prefer (and when), while proper scores can.

Specific comments
1) To my view, the biggest contribution of this work is that it guides the reader on how to use and combine (mostly) new deep learning concepts and methods for uncertainty assessment in hydrological modelling (type-a contribution), while the introduction of a general benchmarking framework for uncertainty assessment in hydrological modelling is (as also mentioned in the "Introduction" section) a secondary (but still important) contribution (type-b contribution). For both these types of contribution and mainly for the former one, a better coverage of the study's background is required. For instance, in lines 15 and 16 it is written that "the majority of machine learning (ML) and Deep Learning (DL) rainfall–runoff studies do not provide uncertainty estimates (e.g., Hsu et al., 1995; Kratzert et al., 2019b, 2020; Liu et al., 2020; Feng et al., 2020)". This is inarguably true; however, there are machine and deep learning rainfall-runoff studies (mostly machine learning rainfall-runoff studies) that do provide uncertainty estimates, while some of them also involve large-scale benchmarking across hundreds of catchments and also use proper scoring rules (together with more interpretable scores) to allow practical comparisons. In fact, this study is not the first one proposing and/or extensively testing machine learning algorithms for probabilistic rainfall-runoff modelling and, to my view, this should be somehow recognized in the "Introduction" section during revisions. In this latter section, information on uncertainty quantification in hydrological modelling using machine and deep learning algorithms is currently scarce, although other topics (even less relevant ones) are well-covered. Especially as regards LSTM-based methods for uncertainty quantification, to my knowledge there are two published works proposing such methods in hydrological modelling and forecasting (Zhu et al. 2020; Althoff et al. 2021). To my view, these studies should necessarily be viewed as part of this work's background.
2) Moreover, I would say that the connection with the machine and deep learning fields needs to be further highlighted for the paper to become more balanced with respect to its nature. Perhaps, this could be established by referring the reader in more places in the manuscript to the original sources of the concepts and algorithms, and by adding a few examples of research works adopting (some of) the same concepts and methods for non-hydrological applications (and possibly by highlighting features that are especially meaningful for rainfall-runoff modelling applications).
3) I should also note that I agree with the main point raised by the other reviewer (Dr John Quilty). As the paper aims to establish benchmarks and benchmark procedures for future works (and as it emphasizes its practical contribution in terms of benchmarking), it would be essential to also provide a comparison with respect to easier-to-apply methods from the statistical and machine learning fields. Such methods have already been applied in the field (mainly for probabilistic hydrological post-processing), and include (but are not limited to) the following ones: linear-in-parameters quantile regression, quantile regression forests, quantile regression neural networks and gradient boosting machine.
4) Furthermore, in lines 94−99 it is written that "the best form metrics for comparing distributional predictions would be to use proper scoring rules, such as likelihoods (see, e.g., Gneiting and Raftery, 2007). Likelihoods, however do not exist on an absolute scale (it is generally only possible to compare likelihoods between models), which makes these difficult to interpret (although, see: Weijs et al., 2010). Additionally, these can be difficult to compute with certain types of uncertainty estimation approaches, and so are not completely general for future benchmarking studies. We therefore based the assessment

of reliability on probability plots, and evaluated resolution with a set of summary statistics". However, to my view proper scores (Gneiting and Raftery 2007) should necessarily be computed in this paper, as at the moment its large-scale results cannot be directly useful to forecasters-practitioners (despite the fact that the currently computed scores provide information that could be also of interest to the reader). For example, the continuous ranked probability score—CRPS score could be computed across multiple quantiles. As these scores are indeed difficult to interpret when stated in absolute terms, in the literature they are mostly presented in relative terms by computing relative improvements offered by an algorithm with respect to another (benchmark). Therefore, one of the compared methods could serve as a benchmark for the others, and the mean (or median) relative improvements could be computed. These computations will reveal the method that forecasters would choose among the four compared ones.

5) Also, my general feeling is that the type-b contribution of the paper (see specific comment #1) is emphasized somewhat more than its type-a contribution (see again specific comment #1) throughout the paper. To my view, the opposite would be more befitting to the contents of the paper. In any case, the type-a contribution could at least be further discussed in the "Conclusions and Outlook" section.

6) Moreover, the following lines (and other similar statements) do not describe the literature accurately (as some existing works on uncertainty assessment in hydrological modelling and forecasting offer benchmarks and benchmarking procedures; see also specific comment #1) and could be rephrased a bit (or removed) to recognize the relevant work made so far in the field:

o … "while standardized community benchmarks are becoming an increasingly important part of hydrological model development and research, similar tools for benchmarking uncertainty estimation are lacking" (lines 3 and 4).

o "We struggled with finding suitable benchmarks for the DL uncertainty estimation approaches explored here" (lines 51 and 52).

o "Note that from the references above only Berthet et al. (2020) focused on benchmarking uncertainty estimation strategies, and then only for assessing postprocessing approaches" (lines 55−57).

o "However, as of now, there is no way to assess different uncertainty estimation strategies for general or particular setups" (lines 332 and 333).

7) Lastly, to my view the same holds for the following lines, as there are research works using machine learning ensembles for uncertainty quantification in hydrological modelling: "A perhaps self-evident example for the potential of improvements are ensembles: Kratzert et al. (2019b) showed the benefit of LSTM ensembles for single-point predictions, and we believe that similar approaches could be developed for uncertainty estimation" (lines 367−369).

References

Althoff D, Rodrigues LN, Bazame HC (2021) Uncertainty quantification for hydrological models based on neural networks: The dropout ensemble. Stochastic Environmental Research and Risk Assessment. https://doi.org/10.1007/s00477-021-01980-8

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102(477):359–378. https://doi.org/10.1198/016214506000001437

Zhu S, Luo X, Yuan X, Xu Z (2020) An improved long short-term memory network for streamflow forecasting in the upper Yangtze River. Stochastic Environmental Research and Risk Assessment 34:1313–1329. https://doi.org/10.1007/s00477-020-01766-4