REVIEWER 2

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.

We will respond to these points in specific comments below, where the reviewer gives details about which papers and methods they are referring to.

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) will the paper 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.

The reviewer has fundamentally misunderstood our aims for benchmarking. What the reviewer suggests is antithetical to what we are trying to do in terms of advancing benchmarking. We explained this more fully in response to reviewer #1's comment, however to summarize here, the reason that we feel strongly about not doing ad hoc, one-off benchmarking is that it is generally impossible for authors to correctly implement methods that they are not experts in (and are motivated to beat). We see this consistently in papers where people benchmark against methods that we are experts in — they almost never implement these methods correctly. This is why modern scientific disciplines use community benchmarking (which has some of its own challenges). The choice to *NOT* benchmark against an ad hoc selection of ad hoc implemented methods was conscious and deliberate — doing so goes against what we are trying to do with setting up a community benchmark on an open, public community dataset. Doing what the reviewer suggests would perpetuate the exact problem that we are trying to address.

Also, we are not aware of any UE methods that are easier to apply. It is possible to design very simple UE methods around existing hydrological models, but in these cases you still have to calibrate the hydrology model, and then afterward apply some statistical procedure. Here, we do all this in one go (and we do it with the current best-performing hydrological model that has been published). All we have to do here is add a probabilistic head to the model (less than 10 lines of code), and change the loss function (a few lines of code), and then use the same training procedure that any LSTM rainfall-runoff model uses. How could there possibly be any method that is simpler than this (either conceptually, computationally, or in terms of the effort it takes to apply)?

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.

We will provide a more detailed answer to this comment in our answer to specific comment 4 of reviewer 2.

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.

The reviewer mentioned that there are UE benchmarking studies in hydrology using large-scale, public datasets. Alas, no examples are given. But, if any such studies exist that we did not cover, we would really like to know about them — and we would include them as reference, as we did with the other UE studies we cite.

We will include Althoff et al. (2021) in the revision — this paper was published after we completed writing this paper. It is a single-basin study (which is likely not appropriate for deep learning), but it is directly relevant. We already cited the work by Zhu et al. (2020), but we cited their primary methodological paper, not the derivative study that the reviewer mentions here.

In view of specific comment 4 by reviewer 2 it is perhaps also worth mentioning that neither of these publication reports CPRS; and Zhu et al. (2020) specifically do not make use of any strictly proper scoring.

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).

All original sources for all methods were cited. If the reviewer sees that we missed one, we will definitely fix it.

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.

This specific comment is largely covered by our responses to reviewer #1. In short: As the title suggests this paper tries to establish baselines for benchmarking. As such the suggestion is not in line with our intention for the framework and the baselines.

Further, in our eyes there are no simpler UE methods available that are the ones that we propose. We cover post-processing in the paper already: We cited several seminal post-processing papers and discussed the advantages of generative probabilistic methods vs. post-processing beginning in line 20 of the manuscript. Be it as it may, we view it as more complicated than building probabilistic models directly. Quantile-regression is a form of regression and does not constitute an approach as such. Random forests and gradient boosted models are not necessarily simpler. Further, we already know that XGboost does not constitute rainfall-runoff models that are as good as LSTMs (see: Gauch, Mai, Lin; 2021). Even if we were to adopt the approach of creating our own ad hoc benchmarks (which, again, goes directly against the point we are making about community benchmarking), none of these suggestions are good examples.

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.

We admit that we did not emphasize the general deficiencies of the (single) metrics/statistics/distances enough in this paper. Taken alone every metric has deficiencies and assumptions underlying it. And, in general we did not want to imply that any of the proposed metrics are fixed, since it is very difficult to define a meaningfully complete set of metrics for hydrological (probabilistic) predictions — and every application will have its own unique purpose. Thus, we never wanted to attempt to define a universal set of benchmarking metrics here. As a matter of fact, we hope that the proposed metrics will be adapted, refined, exchanged and complemented as benchmarking efforts will be adopted by the community. In this way, maybe at some point, canonical metrics for UE benchmarking will emerge. But, we definitely would not dare to claim to do this. We discussed this briefly in the conclusion of the current version of the manuscript. However given the reading by the reviewer we now came to believe that we should also add a broad disclaimer when introducing the method and deepen the discussion. We will do so in the revision.

Regarding the proper scoring rules (*sensu* Gneiting and Raftey, 2007), we have to say that we purposefully did not report them. This choice was made consciously and not out of laziness or oversight. The reasoning for doing so is provided in the portion of text that the reviewer cited. And, regarding CRPS specifically, one generally distinguishes between the continuous ranked probability score (CRPS) and the continuous ranked probability skill score (often abbreviated as CRPSS or CRPS score). The former integrates over the different quantiles by construction (which might be what the reviewer tries to indicate in his statement?). The latter is the usual choice in literature for providing a more interpretable score, by using the CRPS in the same style that the NSE uses variances of point estimates. This use does however require sensible baselines, such as the ones proposed in our contribution. Our approaches conceptually allow us to evaluate the performance in terms of CRPs (and also in terms of likelihood). However, computing meaningful scores is not as simple

as the reviewer thinks: Whenever a probability density function is discretized, information is lost (see for example: Gupta et al., 2021). Intuitively, (a) if the bins are very wide, a bias is introduced because the difference between the mass of the bin and of the actual continuous distribution becomes very large; (b) if the bins are very thin, almost no data can be used to estimate its properties, which induces a large variance in the estimation. In the hydrological context specifically, small bin-widths should also be distrusted because of the inherent uncertainty of the variables. Apart from a careful choice of bin-sizes, it is therefore in practice often more appropriate to evaluate the properties of the UE approaches with regard to different metrics to derive at a nuanced, paretian evaluation (see for example: Kumar, Lia and Ma, 2020). This is what we choose to do in this paper.

Finally, we want to point out that CRPS are not common in hydrology. To substantiate this claim, we did a literature review to provide a list of publications that assess UE approaches in a hydrological context. The list is not exhaustive, but we include a larger set of topics and settings than considered in our manuscript. Notably, it includes all referenced sources by the reviewer himself, which did not report proper scoring rules. In summary, from 38 references only 5 report proper proper scores; and these also examine the performance in terms of probability plots (or metrics derived thereof) and resolution.

1. Althoff et al. (2021) primarily use point-prediction metrics for evaluation, but also percentage of coverage (related to the probability plot), average width of the uncertainty intervals (i.e., a single statistic for the resolution), and average interval score (a proper scoring rule, but nut CRPRS). They also use a set of ad-hoc tests to get an intuition about the uncertainty estimation capacity.
2. Abbasour et al. (2015) report the performance in terms of point-metrics, and show the 95% quantiles visually within hydrographs.
3. Ajami et al. (2007) count the number of observations within the 95% prediction interval (i.e. a single point on the probability plot) and show visual evidence of their approach in the form of hydrographs.
4. Berthet et al (2020) report CRPS scores, but since they found them so uninformative they also evaluate in terms of the probability-plot, the sharpness and point-metrics (as we do).

5. Beven (1993) shows bounds for specific events.
6. Beven and Binley (2014) show the 90% quantiles for individual events.
7. Beven and Smith (2015) report the event-based coverage of the 95% prediction intervals (related to the probability plot).
8. Bogner and Pappenberger (2011) report metrics for point-predictions, CRPS scores (which they just interpret in terms of accuracy), and probability plots (together with derived summary statistics).
9. Coxon et al (2015) evaluate the uncertainty in terms of bound-distances (related to precision statistics) and report the performance of the point-estimates.
10. Dogulu et al. (2015) report UE estimation performance in terms of prediction interval coverage probability (related to the probability plot), mean prediction interval (related to the precision statistics), and average relative interval length related to the precision statistics).
11. Gopalan et al. (2019) report point-prediction metrics, p-factors (related to coverage, thus to the probability plot) and simulation uncertainty indices (an improper score, related to the width of the quantiles).
12. Huart and Mailhot (2008) show hydrographs and dotty-plots.
13. Kavetski et al. (2006) report RMSE and standard deviation and show the prediction bounds visually using hydrographs.
14. Liu et al (2005) show dotty plots, and provide a visual inspection for specific events.
15. Kim et al. (2020) report CRPS scores, together with probability plots and rank histograms.
16. Mantovan and Todini (2006) report percentiles of MSE values for their derived posteriors.
17. McMillan et al. (2010) use a rank histogram (similar to our deviation plot) as primary diagnostic tool. They also show exemplary hydrographs with uncertainty bounds.
18. Montanari and Koutsoyiannis (2012) report performance in terms of the probability-plot and show some exemplary hydrographs.

19. Murphy and Winkler (1984) discuss the utility and usage of the probability plot for weather forecasting. They are one of the earliest sources that we are aware of.
20. Mustafa et al. (2019) do not evaluate the predictive uncertainty as such.
21. Papacharalampous, Tyralis, and Langousis et al. (2019) use proper scores, but only report their relative performance with respect to an arbitrary benchmark model (for the sake of clarity). They also report reliability scores (related to the probability plot), the average width of the prediction interval (related to our precision statistics).
22. Schoups and Vrugt (2010) report the UE performance in terms of probability plots.
23. Shrestha and Solomatine (2008) report interval coverage probability (related to the probability plot) and mean prediction interval (related to our precision statistics). They use a derivative of the probability plot to relate model error with probability of occurrence, as well as the model residuals in dependence of the input variables.
24. Shrestha and Solomatine (2009) use a probability plot and the mean size of the prediction intervals. They also show cumulative densities for specific events.
25. Shrestha, Kayastha, and Solomatine (2009) report interval coverage probability (related to the probability plot) and mean prediction interval (related to our precision statistics). They also show the 90% prediction intervals for specific events.
26. Shortridge, Guikema, and Zaitchik (2016) do not use explicit diagnostics to assess the UE estimation capacity, but use a scenario based approach.
27. Srivastav, Sudheer, and Chaubey (2007) show the quantiles within hydrographs for specific events.
28. Teweldebrhan, Burkhart, and Schuler (2018) report point-metrics, critical success index (similar to coverage ratio, thus related to the probability plot) and show some exemplary hydrographs.
29. Tian et al. (2018) report point-estimation metrics, average size of the uncertainty interval (related to our precision statistics) and the coverage ratio differentUE quantiles (related to the probability plot).

30. Thyer et al. (2009) report the uncertainty estimation performance in terms of the probability plots.
31. Tolson and Shoemaker (2008) show prediction bounds explicitly.
32. Vrugt et al. (2005) report point-estimation metrics together with error distribution plots.
33. Vrugt et al. (2008) report point-estimation metrics, show the obtained bounds over the training and validation periods, and the standard deviation of the estimated parameters.
34. Woldemeskel et al. (2018) report CRPSS (a scoring rule), a probability plot derivative, and the 99% interquartile range (related to our precision statistics). The authors also show the 98% quantiles for specific events.
35. Westerberg and McMillan (2015) show individual runs, and quantile deviations (similar in kind to our deviation plot).
36. Zink et al. (2017) report the coefficient of variation and normalized 5%-95% quantile ranges (both related to our precision statistics).
37. Zhu et al. (2020) primarily use point-prediction metrics and visually inspect the uncertainty estimation capacities of the model.
38. Vaysse and Lagacherie (2017) report point-prediction metrics in conjunction with probability plots.

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.

Since the reviewer gave no reason or explanation as to why they hold this opinion it is difficult for us to decide how (or whether) to act on this comment. What content is missing from the conclusions and outlook section? Just asking to “add more” is not helpful.

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:

1. … "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).
2. "We struggled with finding suitable benchmarks for the DL uncertainty estimation approaches explored here" (lines 51 and 52).
3. "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).
4. "However, as of now, there is no way to assess different uncertainty estimation strategies for general or particular setups" (lines 332 and 333).

To our knowledge, all of the above quoted statements from the paper are correct. The reviewer provides no references that do any of these things — the reviewer only asserts that such references exist. We looked extensively for such references and found none.

The requirements for a suitable, standardized benchmark are (Nearing et al., 2018): (i) that the benchmark uses a community-standard data set that is publicly available, (ii) the model or method is applied in a way that conforms to community standards of practice for that data set (e.g., standard train/test splits), and (iii) that the results of the standardized benchmark runs are publicly available. To these we added a post-hoc model examination step in our framework, which aims at exposing the intrinsic properties of the model. Although this last step is important, especially for ML approaches and imperfect approximations, we do not view it as a requirement for benchmarking in general (and therefore would have included any paper that did items i-iii but not this).

We spent considerable time searching for such UE benchmarks for this paper. We do not believe that we "misrepresented the current research landscape" and we wrote

the quoted sentences in good faith. As a matter of fact, the difficulty to find such benchmarks was a reason why we decided to include a focus on establishing a community UE benchmark in the first place.

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)

This text passage does not propose to estimate uncertainty by using ensembles. It proposes to build ensembles of uncertainty estimators. We are not aware of any publication in the hydrological sector that has done this so far. If a reference had been provided we could cite it in this context, however we are unaware of such a study.

References

- Abbaspour, K. C., Rouholahnejad, E., Vaghefi, Srinivasan, R., Yang, H., & Kløve, B. (2015). A continental-scale hydrology and water quality model for Europe: Calibration and uncertainty of a high-resolution large-scale SWAT model. *Journal of Hydrology*, *524*, 733-752.
- Althoff, D., Rodrigues, L. N., & Bazame, H. C. (2021). Uncertainty quantification for hydrological models based on neural networks: the dropout ensemble. *Stochastic Environmental Research and Risk Assessment*, *35*(5), 1051-1067.
- Ajami, N. K., Duan, Q., & Sorooshian, S. (2007). An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water resources research*, *43*(1).

- Berthet, L., Bourgin, F., Perrin, C., Viatgé, J., Marty, R., & Piotte, O. (2020). A crash-testing framework for predictive uncertainty assessment when forecasting high flows in an extrapolation context. *Hydrology and Earth System Sciences*, *24*(4), 2017-2041.
- Beven, K. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling. *Advances in water resources*, *16*(1), 41-51.
- Beven, K., & Binley, A. (2014). GLUE: 20 years on. *Hydrological processes*, *28*(24), 5897-5918.
- Beven, K., & Smith, P. (2015). Concepts of information content and likelihood in parameter calibration for hydrological simulation models. *Journal of Hydrologic Engineering*, *20*(1), A4014010.
- Bogner, K., & Pappenberger, F. (2011). Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. *Water Resources Research*, *47*(7).
- Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R., & Smith, P. J. (2015). A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. *Water resources research*, *51*(7), 5531-5546.
- Dogulu, N., López López, P., Solomatine, D. P., Weerts, A. H., & Shrestha, D. L. (2015). Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC methods and their comparison on contrasting catchments. *Hydrology and Earth System Sciences*, *19*(7), 3181-3201.
- Gopalan, S. P., Kawamura, A., Amaguchi, H., Takasaki, T., & Azhikodan, G. (2019). A bootstrap approach for the parameter uncertainty of an urban-specific rainfall-runoff model. *Journal of Hydrology*, *579*, 124195.
- Huard, D., & Mailhot, A. (2008). Calibration of hydrological model GR2M using Bayesian uncertainty analysis. *Water Resources Research*, *44*(2).
- Kavetski, D., Kuczera, G., & Franks, S. W. (2006). Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water resources research*, *42*(3).
- Kim, T., Shin, J. Y., Kim, H., & Heo, J. H. (2020). Ensemble-Based Neural Network Modeling for Hydrologic Forecasts: Addressing Uncertainty in the Model Structure and Input Variable Selection. *Water Resources Research*, *56*(6), e2019WR026262.

- Liu, Z., Martina, M. L., & Todini, E. (2005). Flood forecasting using a fully distributed model: application of the TOPKAPI model to the Upper Xixian Catchment. *Hydrology and Earth System Sciences*, *9*(4), 347-364.
- Mantovan, P., & Todini, E. (2006). Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology. *Journal of hydrology*, *330*(1-2), 368-381.
- McMillan, H., Freer, J., Pappenberger, F., Krueger, T., & Clark, M. (2010). Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions. *Hydrological Processes: An International Journal*, *24*(10), 1270-1284.
- Montanari, A., & Koutsoyiannis, D. (2012). A blueprint for process-based modeling of uncertain hydrological systems. *Water Resources Research*, *48*(9).
- Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association*, *79*(387), 489-500.
- Mustafa, S. M. T., Hasan, M. M., Saha, A. K., Rannu, R. P., Uytven, E. V., Willems, P., & Huysmans, M. (2019). Multi-model approach to quantify groundwater-level prediction uncertainty using an ensemble of global climate models and multiple abstraction scenarios. *Hydrology and Earth System Sciences*, *23*(5), 2279-2303.
- Nearing, G. S., Ruddell, B. L., Clark, M. P., Nijssen, B., & Peters-Lidard, C. (2018). Benchmarking and process diagnostics of land models. *Journal of Hydrometeorology*, *19*(11), 1835-1852.
- Papacharalampous, G., Tyralis, H., Langousis, A., Jayawardena, A. W., Sivakumar, B., Mamassis, N., ... & Koutsoyiannis, D. (2019). Probabilistic hydrological post-processing at scale: Why and how to apply machine-learning quantile regression algorithms. *Water*, *11*(10), 2126.
- Schoups, G., & Vrugt, J. A. (2010). A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research*, *46*(10).
- Shrestha, D. L., & Solomatine, D. P. (2008). Data-driven approaches for estimating uncertainty in rainfall-runoff modelling. *International Journal of River Basin Management*, *6*(2), 109-122.
- Shrestha, D. L., Kayastha, N., & Solomatine, D. P. (2009). A novel approach to

parameter uncertainty analysis of hydrological models using neural networks. *Hydrology and Earth System Sciences*, *13*(7), 1235-1248.

- Shortridge, J. E., Guikema, S. D., & Zaitchik, B. F. (2016). Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences*, *20*(7), 2611-2628.
- Solomatine, D. P., & Shrestha, D. L. (2009). A novel method to estimate model uncertainty using machine learning techniques. *Water Resources Research*, *45*(12).
- Srivastav, R. K., Sudheer, K. P., & Chaubey, I. (2007). A simplified approach to quantifying predictive and parametric uncertainty in artificial neural network hydrologic models. *Water Resources Research*, *43*(10).
- Teweldebrhan, A. T., Burkhart, J. F., & Schuler, T. V. (2018). Parameter uncertainty analysis for an operational hydrological model using residual-based and limits of acceptability approaches. *Hydrology and Earth System Sciences*, *22*(9), 5021-5039.
- Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., & Srikanthan, S. (2009). Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. *Water Resources Research*, *45*(12).
- Tian, Y., Xu, Y. P., Yang, Z., Wang, G., & Zhu, Q. (2018). Integration of a parsimonious hydrological model with recurrent neural networks for improved streamflow forecasting. *Water*, *10*(11), 1655.
- Tolson, B. A., & Shoemaker, C. A. (2008). Efficient prediction uncertainty approximation in the calibration of environmental simulation models. *Water Resources Research*, *44*(4).
- Vaysse, K., & Lagacherie, P. (2017). Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma*, *291*, 55-64.
- Vrugt, J. A., Diks, C. G., Gupta, H. V., Bouten, W., & Verstraten, J. M. (2005). Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. *Water resources research*, *41*(1).
- Vrugt, J. A., Ter Braak, C. J., Clark, M. P., Hyman, J. M., & Robinson, B. A.

(2008). Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research*, *44*(12).

- Westerberg, I. K., & McMillan, H. K. (2015). Uncertainty in hydrological signatures. *Hydrology and Earth System Sciences*, *19*(9), 3951-3968.
- Woldemeskel, F., McInerney, D., Lerat, J., Thyer, M., Kavetski, D., Shin, D., ... & Kuczera, G. (2018). Evaluating post-processing approaches for monthly and seasonal streamflow forecasts. *Hydrology and Earth System Sciences*, *22*(12), 6257-6278.
- Zink, M., Kumar, R., Cuntz, M., & Samaniego, L. (2017). A high-resolution dataset of water fluxes and states for Germany accounting for parametric uncertainty. *Hydrology and Earth System Sciences*, *21*(3), 1769-1790.
- 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*(9), 1313-1329.