Review hess-2021-154

TITLE

Uncertainty Estimation with Deep Learning for Rainfall–Runoff Modelling

RECOMMENDATION

Moderate Revision

REVIEWER

John Quilty

GENERAL COMMENTS

This paper introduces several variants of long short-term memory networks (LSTM) for the estimation of predictive uncertainty in rainfall-runoff modelling. The methods are explored using the CAMELS dataset (Catchment Attributes and MEteorology for Large-sample Studies) (Addor et al., 2017) as a means to provide a large-scale benchmark for the proposed methods as well as others that may be explored in the future. My takeaway is that both of these items are the main contributions of the paper, the latter, in my opinion, being the most relevant.

In general, the paper is well-written and clear, the methodology is reasonable (however, some suggestions are included below), and the results seem promising. The main point I feel important to raise is that the authors adopt LSTM as the sole model and do not compare it against a plethora of other 'simpler' data-driven models that can provide estimates of the predictive uncertainty, likely with much less computational cost and onerous hyper-parameter tuning. Since the authors are looking to create a benchmark for comparing different predictive uncertainty estimation techniques in the context of rainfall-runoff modelling, it seems reasonable that very simple approaches be explored as a baseline before considering more complex methods. This critique is not meant to take anything away from the work that is done, since the reviewer understands LSTM may be a preferred method within this group (as is the case with process-based models within other groups), but more to seek justification for using more complicated (computationally intensive) models when it may suffice to use simpler ones that end up providing similar results, which seems relevant to consider from a benchmarking perspective.

Aside form the above point, I consider most other comments below relatively minor but still important to address.

If the authors can suitably address the comments noted below, I would be happy to recommend that the paper be published in Hydrology and Earth System Sciences.

SPECIFIC COMMENTS

1. Terminology: throughout the paper the authors mention both 'prediction' and 'forecasting', which have different purposes and (potentially) different modelling setups. The authors should be consistent in their use of terminology throughout the manuscript. It appears the term 'prediction' is more appropriate considering their application.

2. Introduction: in L20-31 it would be good for the authors to acknowledge other established and recent methods for probabilistic prediction (and forecasting) in hydrology including: quantile regression-based neural networks (Cannon, 2018, 2011), copula-based approaches (Li et al., 2021; Liu et al., 2021), and (parametric/distributional) probabilistic decision tree methods (Başağaoğlu et al., 2021; Schlosser et al., 2019), the latter methods allow for the predictive distribution of the target variable to be easily estimated as part of the training procedure. Although I have yet to undertake such an analysis myself, I suspect the latter group of methods could estimate predictive uncertainty with much less computational expense than the LSTM variants adopted here.

3. Introduction: the authors may want to acknowledge the recent paper by Althoff et al. (2021) whom also used MCD with LSTMs for daily streamflow forecasting.

4. Section 2 and 2.3: while I appreciate the authors have been exploring LSTM in many former studies and are evolving a research program around this method, from my own experience, these methods are extremely computationally intensive (due to it's recurrent formulation, gradient-based learning, need for careful hyper-parameter tuning, etc.), and thus tend to be a 'turn-off' for those who do not have adequate computing resources to explore such methods. It would seem beneficial for the reader to have a 'simpler' non-LSTM baseline to compare against the proposed LSTM variants. If the results of the LSTM variants significantly outperform these simpler methods, then it may serve as a motivation for other researchers to devote more time and resources to incorporating LSTM into their research endeavours. My feeling is that a simple benchmark method could include one of the many variants of quantile regression forests (see for example, https://scikit-garden.github.io/examples/QuantileRegressionForests/). This would not seem out of scope as the authors mention in Section 2 that their work is devoted to data-driven methods, of which LSTM are only a small (but relevant) fraction.

5. L177-180: This somewhat confusing. If I understand correctly (based on the Althoff et al. (2021) paper), dropout is used during training at each iteration but it does not create a separate model at each iteration, only a 'thinned' network. However, performing dropout during testing (or model implementation, evaluation, or whatever other term you wish to use), each time you make a prediction you simply turn on/off nodes according to the pre-specified probabilities used during training and you repeat this as many times as you desire, creating a number of 'sister' predictions. Again, the model does not change, you simply 'thin' the network each time you create a 'sister' prediction. If this is how MCD was used in the experiments described in this paper, it is not apparent and would be helpful to clarify.

6. L193-4: why not use a regularized squared loss function? Is it not a standard practice to perform L2 (and potentially L1) regularization to improve LSTM performance and reduce overfitting? Was this considered? If not, why?

7. L195-6: what dataset partition was used to select optimal hyper-parameters?

8. Section 2.4.3: the authors should formally define first and second order uncertainties.

9. Table 3: it's not clear how the 'obs' data is to be used to 'contextualize' the results from the different models. More detail should be provided (perhaps with an example in the relevant

section).

10. L286-289: it would seem like a good idea for the authors to investigate how best to use the CMAL and UMAL variants for improving predictive performance at low and high (tail) flows (from the point prediction point of view), as this tends to be a major impetus for creating models with uncertainty assessment capabilities. In other words, what's the point of going through the trouble of designing these more sophisticated methods if they cannot outperform the base approach (LSTMp) when assessed on highly relevant metrics. Please don't get me wrong, I am not trying to downplay the very interesting work done by the authors, I am simply trying to help them more fully explore the merits of their work and better 'sell' their approach to a 'skeptic'.

11. Section 3.2.2 is very interesting!

12. Section 3.2.3: once first and second order uncertainties are formally defined (see comment 8), this section should give a good description of what is *shown* in Figure 11 but it is unclear what message the authors expect the reader to take-away from this figure. What's the relevance of this figure and why should the reader 'care' about it?

13. Section 3.3: my understanding is that this is the time needed to make predictions with a trained model. What is the training time for the different models? Can the authors provide an example calculation for the overall run-time in Appendix A (or at the very least in their reply, it's not clear how the 365 and 174 days were calculated)?

14. Section 4: after L335 the authors may wish to very briefly summarize the adopted models and datasets used in the study before continuing with L336 onwards. This should help the interested reader with a short 'time-budget', who may only jump from the abstract to the conclusion, get a decent idea of the methods and dataset involved (the dataset being one of the key strengths of this paper).

15. Each equation in the appendices (B1, B2, etc.) should be properly cited (the rule is to cite all equations that are not developed by the authors in the paper).

TECHNICAL CORRECTIONS

- L4: 'This contribution**s**…'
- L9: Please rephrase '…and show that learn nuanced behaviors in different situations.'
- L182-3 seems to be repeated at L193-4?
- Table 2 should be placed after it is mentioned in the text.
- L207: remove 'of'.
- Spell out NSE and KGE at first use.
- L216: perhaps indicate where the median aggregator is used in the paper?
- L223: 'produced'.
- L33: 'in **the** form of'.
- Figure 10 and 11 legends: remove the 0 from '05'.
- Figure 11 x-axis: 'Dec' not 'Dez'.

- L417: remove ‘single’ (duplicated).
- L420-1: ‘They have seen…’?
- L430 should be reviewed for duplicated words (‘the training data’) and typos.
- L450: not clear what ‘i’ pertains to…
- L470: remove ‘can be’ (duplicated).

REFERENCES

Addor, N., Newman, A.J., Mizukami, N., Clark, M.P., 2017. The CAMELS data set: Catchment attributes and meteorology for large-sample studies. Hydrol. Earth Syst. Sci. 21, 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Althoff, D., Rodrigues, L.N., Bazame, H.C., 2021. Uncertainty quantification for hydrological models based on neural networks: the dropout ensemble. Stoch. Environ. Res. Risk Assess. 35, 1051–1067. https://doi.org/10.1007/s00477-021-01980-8

Başağaoğlu, H., Chakraborty, D., Winterle, J., 2021. Reliable evapotranspiration predictions with a probabilistic machine learning framework. Water (Switzerland) 13. https://doi.org/10.3390/w13040557

Cannon, A.J., 2018. Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. Stoch. Environ. Res. Risk Assess. 32, 3207–3225. https://doi.org/10.1007/s00477-018-1573-6

Cannon, A.J., 2011. Quantile regression neural networks: Implementation in R and application to precipitation downscaling. Comput. Geosci. 37, 1277–1284. https://doi.org/10.1016/j.cageo.2010.07.005

Li, H., Huang, G., Li, Y., Sun, J., Gao, P., 2021. A C-Vine Copula-Based Quantile Regression Method for Streamflow Forecasting in Xiangxi River Basin, China. Sustainability 13. https://doi.org/10.3390/su13094627

Liu, Z., Cheng, L., Lin, K., Cai, H., 2021. A hybrid bayesian vine model for water level prediction. Environ. Model. Softw. 142, 105075. https://doi.org/10.1016/j.envsoft.2021.105075

Schlosser, L., Hothorn, T., Stauffer, R., Zeileis, A., 2019. Distributional regression forests for probabilistic precipitation forecasting in complex terrain. Ann. Appl. Stat. 13, 1564–1589. https://doi.org/10.1214/19-AOAS1247