Reply to RC3: 'Comment on egusphere-2026-469', Jonathan Frame

Review: The need for uncertainty: why probabilistic LSTMs are key to improving flood predictions and enabling learned warning rules

This paper tests three different methods of predicting streamflow probabilities, and compares those against a deterministic LSTM. Interestingly, this paper demonstrates that high probability predictions usually underestimate peak flows, but the high end of the low probability predictions are more capable of matching peaks than deterministic LSTM. This is further demonstrated with a hit and miss comparison, showing LSTM misses floods more often than not, yet 99th percentile captures floods, except for "extreme floods", which are still outside the 99th percentile. This paper also includes a reinforcement learning-based method for translating streamflow probability into actionable flood warnings.

Paper organization: This paper is generally easy to read and easy to understand. There are a few sentences that are hard to follow, I'll point those out in the line comments.

Novelty: This paper tests three methods for probabilistic streamflow predictions, and in the end, none of them turn out to sufficiently capture the most extreme events. CMAL had been used before, but I wonder if the authors simply chose the other two methods almost randomly or was there consideration of the applicability to the problem. I am left wondering if probabilistic streamflow in general can't capture those extremes, or, is it just these methods. DRN was developed for weather and BQN for wind speed. I'm not suggesting more methods be added, but I think it is worth reflecting on if there is a better method specifically for streamflow? Or, would it make sense to develop a custom method? If not, is it a limitation of LSTM itself? I think these would need to be addressed in order to answer research question 1 completely.

First, we would like to elaborate on the remark, that none of our models sufficiently capture the most extreme events.

We thank the reviewer for this comment, which touches on a central challenge in probabilistic flood prediction. We agree that some extreme events are not captured within the upper predictive quantiles. However, this behavior is consistent with the probabilistic interpretation of calibration. By definition, a predicted 90^{th} (or 99^{th}) percentile event is expected to be exceeded in 10% (or 1%) of cases, respectively. Consequently, particularly rare or unprecedented events, such as those beyond the empirical range of the training data, are not expected to be reliably enclosed by lower quantiles. In other words, some extreme floods must lie beyond the predicted high quantiles.

At the same time, we fully acknowledge that, from an operational perspective, the adequate representation of extreme events is critical. This highlights a fundamental distinction between overall probabilistic calibration and the ability to accurately represent tail behavior relevant for flood risk assessment. To better address this aspect, we extended our analysis by including the FRiCA model, which enables a transition from probabilistic predictions to deterministic decision-making. This allows us to explicitly evaluate how probabilistic forecasts can be translated into actionable flood predictions.

More generally, we emphasize that reliable prediction of rare events (of any dataset) with data-driven models fundamentally requires probabilistic formulations. Deterministic approaches, or probabilistic models implicitly forced toward deterministic behavior through tailored loss functions, tend to concentrate on specific parts of the distribution (e.g., upper quantiles). While this can improve performance scores targeted at extremes (such as flood hit rates), it typically comes at the cost of degraded calibration and an increased rate of false alarms. This trade-off underscores the importance of probabilistic frameworks, which allow us to explicitly represent uncertainty and flexibly balance detection of extremes against false alarm rates, rather than implicitly encoding such trade-offs in the training objective. To address this point more comprehensively, we will extend our analysis beyond aggregate calibration metrics and include additional diagnostics, such as marginal calibration, to better characterize model behavior in the distribution tails. We will also add a section to the discussion where we explain in detail why we believe that probabilistic LSTMs are the key for flood modelling particularly when we work with regionally trained models.

… CMAL had been used before, but I wonder if the authors simply chose the other two methods almost randomly or was there consideration of the applicability to the problem.

We agree that the rationale behind the selection of probabilistic models was not sufficiently articulated in the current manuscript, and we will clarify this in the revision.

The choice of CMAL, DRN, and BQN was not arbitrary, but guided by the objective to compare fundamentally different approaches to probabilistic prediction within a consistent LSTM framework. In particular, the three models represent complementary strategies for modeling predictive uncertainty but work differently:

(i) CMAL serves as a flexible mixture-based approach that can represent complex, potentially multi-modal predictive distributions and has been successfully applied in hydrological settings. It is hence the quasi benchmark but it has not been tested particularly for its ability to predict extremes.

(ii) DRN represents an explicit parametric approach, where the conditional distribution is modeled through a predefined distribution. The choice of the logistic distribution is deliberate: it belongs to the class of exponential-tailed distributions, it exhibits heavier tails than the Gaussian distribution and therefore provides a more suitable representation for skewed and moderately heavy-tailed streamflow data. At the same time, it retains analytical tractability and allows for stable parameter estimation within the neural network framework. We note that if the true data-generating process exhibits strongly sub-exponential (i.e., heavier-than-exponential) tail behavior, our assumptions may become limiting, an aspect that is also reflected in our results.

(iii) BQN follows a fundamentally different, non-parametric strategy by directly estimating the quantile function without assuming a specific distributional form. This makes it a natural complement to the parametric approaches above, particularly in settings where the shape of the conditional distribution is difficult to specify a priori. Previous work (e.g., Bremnes, 2020) has demonstrated the flexibility and strong calibration properties of BQN-type models, which motivated its inclusion in this study.

Taken together, these three models span a spectrum from single parametric (DRN), to mixture-based parametric (CMAL), to distribution-free quantile estimation (BQN). This design allows us to systematically assess how different assumptions about the predictive distribution affect performance in the context of flood prediction. In the revised manuscript, we will make this rationale explicit in the introduction and better motivate the connection between model choice and the characteristics of hydrological extremes.

Finally, we acknowledge that architectural aspects of the LSTM, such as potential saturation effects, may influence the representation of extremes through their impact on distribution parameters (e.g. mixture parameters or quantile function coefficients). We did not address this aspect sufficiently in the discussion, and was also brought to our attention in RC2. In the revised manuscript we will add a discussion on how the saturation phenomenon in the LSTM can manifest as miscalibration.

Research question 2 seems either 1) trivial or 2) poorly worded. This discussion piece on question 2 is also kind of strange. Particularly line 481-483. I'm not sure what is meant by this: "it illustrates that runoff generation at the catchment scale in large-sample datasets is not unique".

We would like to differ with the reviewer's view about the second research question being trivial. Having said that, perhaps "... probabilistic **LSTMs** more suitable to capture hydrological extremes…?" would be a better formulation for the question.

Deterministic deep-learning models, in particular LSTMs, are known to underestimate in the tail regions of the streamflow distribution. In the realm of non-deep-learning models, studies

investigate the use of weighted squared loss formulations, streamflow transformations (Hunter et. al. (2021), Thirel et al., (2024)) or quantile- (Tyralis and Papacharalampous (2021)) or expectile-based (Tyralis et al. (2023)) training for streamflow predictions. We implemented some of these strategies for the LSTM and found that, improved predictive performance for the extreme events is generally accompanied by a decrease in global performance. As such, we are left with a *specialist* model which predicts extremes very well, but necessitates the need of another better model elsewhere. Probabilistic modeling gives us more information about the possibility of extremes, without losing performance for other flows, enabling a single model to give us reliable information for the entire discharge distribution.

We hope this is a convincing argument as to why the question is non-trivial in our opinion. We will refine the discussion for this question, such that it presents our thoughts more clearly and comprehensively.

Line 38-40, LSTM outperforms standard calibration of static parameters for conceptual models. Actually, there is a slight correction: The conceptual model in Frame 2022 (including corrigendum) never outperforms LSTM. The NWM does, but we were not able to re-calibrate that without the extreme events, so it isn't really a fair comparison.

We will revise this for correctness.

Line 554-556: "This highlights that the primary benefit of probabilistic modeling does not lie in improving point-prediction accuracy, but rather in providing a structured and interpretable representation of predictive uncertainty" I'm not sure you can you claim this as a general conclusion. This reads to me more of a specific interpretation of this study. For instance, I'd love to see 68% of peak streamflow observations within the standard deviation of your probability distribution. Just because this result doesn't achieve that doesn't mean it isn't a benefit of probabilistic modeling in general, if the objective is met. …

We would like to contend that, in general, all probabilistic modeling gives a structured and interpretable representation of uncertainty. At the same time, we agree that its inability to improve point-prediction accuracy is a conclusion specific to this study. For instance, the UMAL model from Klotz et. al. (2022) reports an NSE better than the LSTM. We will rephrase this sentence.

As for your latter thoughts, the PIT histograms are indeed meant to check if 68% of the observed data is within the standard deviation. Although they are visualized on a quantile basis, the interpretation is similar – predicted quantiles resemble observed quantiles of the data.

… And again, on Line 557: “biggest value addition from the probabilistic models emerges in the upper tail of the discharge distribution”, this seems like a specific result to this study, not a general conclusion, but it is written as something general. I guess it should be “biggest value addition from THESE probabilistic models”.

We will rephrase this. Indeed, it cannot be said for all probabilistic models, that they will have a superior performance for the tail regions.

I trust that the code will be released by the authors, but I am surprised not to see it linked here already.

We made our Zenodo repository accessible exclusively to the reviewers in the review process. We will also link it in the final manuscript version.

References:

Bremnes JB. Ensemble Postprocessing Using Quantile Function Regression Based on Neural Networks and Bernstein Polynomials. Monthly Weather Review. 2020;148(1):403-414. doi:10.1175/mwr-d-19-0227.1

Hunter J, Thyer M, McInerney D, Kavetski D.: Achieving high-quality probabilistic predictions from hydrological models calibrated with a wide range of objective functions. Journal of Hydrology. 2021; 603:126578. doi:10.1016/j.jhydrol.2021.126578

Klotz D, Kratzert F, Gauch M, et al. Uncertainty estimation with deep learning for rainfall–runoff modeling. *Hydrol Earth Syst Sci*. 2022;26(6):1673-1693. doi:10.5194/hess-26-1673-2022

Thirel, G., Santos, L., Delaigue, O., and Perrin, C.: On the use of streamflow transformations for hydrological model calibration, Hydrol. Earth Syst. Sci., 28, 4837–4860, https://doi.org/10.5194/hess-28-4837-2024, 2024.

Tyralis H, Papacharalampous G. Quantile-Based Hydrological Modelling. Water. 2021;13(23):3420. doi:10.3390/w13233420

Tyralis H, Papacharalampous G, Khatami S. Expectile-based hydrological modelling for uncertainty estimation: Life after mean. Journal of Hydrology. 2023; 617:128986. doi:10.1016/j.jhydrol.2022.128986