Reply to RC2: 'Comment on egusphere-2026-469', Sandeep Poudel & Scott Steinschneider (co-review team)

Summary:

This study addresses two objectives: (1) evaluating probabilistic LSTM networks on extreme rainfall-runoff events, and (2) introducing reinforcement learning (RL) methods on top of probabilistic forecasts to issue flood warnings. Overall, the authors find that probabilistic LSTM performs better for extreme events, as it assigns some probability mass to those events, whereas deterministic LSTM mostly underestimates them. They also demonstrate the value of RL-type methods in issuing reliable flood warnings compared to simple fixed heuristics. The paper covers important aspects and has practical applications, and therefore has merit. However, there are several major concerns that need to be addressed to improve the clarity of the work and better support the results of their analysis. See comments below.

Thank you for the positive assessment. We agree that the manuscript is probably quite technical to follow and can be improved for better readability. Your comments have brought to our attention essential aspects that need better articulation. We realize that providing more background and reasoning for certain parts of our work is essential. We would also take this opportunity to justify certain aspects that have concerned you. Please find our point-by-point responses below. Again, thank you for your very detailed and helpful comments.

Major Comments:

1. The DNR and CMAL models are not really that different. Both estimate the parameters of a distribution for flow, and it's just the distribution that differs (countable mixture of asymmetric Laplace distributions vs. logistic distribution). However, the authors use the negative log-likelihood as the loss for CMAL and CRPS as the loss for DNR. Therefore, it's unclear how the analysis separates the impact of the loss function from the structure of distribution on model performance. The authors should therefore try using the negative log-likelihood as a loss for DRN and CRPS as a loss for CMAL, and then comment on whether it's the loss or the distribution type that leads to the bigger separation in performance.

We thank the reviewer for this comment and would like to take this opportunity to elaborate on how the DRN and CMAL are different. We agree that the rationale behind the selection of probabilistic models was not sufficiently communicated (as you rightly pointed out in your minor comments), and we will clarify this in the revision.

While both CMAL and DRN parameterize a predictive distribution, they differ fundamentally in their representational assumptions. The DRN models the conditional distribution using a single

parametric distribution (here, logistic), resulting in a comparatively constrained and smooth representation with fixed tail behavior. In contrast, the CMAL represents the distribution as a mixture of asymmetric Laplace distributions, which allows for substantially greater flexibility, including the ability to approximate skewed, heavy-tailed, or even multi-modal distributions. Importantly, this distinction is not merely a matter of choosing different distributions, but reflects a structural difference in how the conditional distribution is represented. While CMAL can, in principle, reduce to a single-component model, its formulation as a mixture explicitly enables increased expressiveness when supported by the data. In contrast, the DRN is inherently restricted to a single parametric distribution, and extending it to a comparable flexibility would require a fundamentally different formulation (e.g., mixtures or alternative distributions). As such, the comparison reflects a broader trade-off between structural simplicity and strong inductive bias (DRN) versus high flexibility and data-adaptive distributional shape (CMAL). This difference is particularly relevant in hydrological applications, where the shape of the conditional distribution can vary strongly across flow regimes, especially under extreme conditions.

It would be interesting to analyze the effects of the loss function on the learnt probabilistic behavior, but we do not expect substantial differences (Gebetsberger et al., 2018). We will however train a DRN with the NLL and the CMAL with a sample-based CRPS and report our findings suitably in the revised manuscript.

2. The FRiCA reinforcement learning model feels underdeveloped and tacked on at the end. Additionally, based on the results, it does not come off as a convincing improvement over simply using a static quantile of the probabilistic predictive distribution. I still think it's interesting and a useful contribution, but the framing in the introduction needs to be adjusted to communicate to readers that this part of the paper is a bit exploratory and designed as proof of demonstration of how additional models can be used to explore how probabilistic DL hydrologic predictions could be incorporated into peak flow estimation and decision-support for flood warnings.

We thank the reviewer for this comment and agree that the role of the FRiCA model was not sufficiently clarified in the current manuscript. In the revised version, we will adjust the framing in the introduction to make explicit that this component is exploratory and intended as a proof of concept.

The motivation for including FRiCA is rooted in a fundamental property of probabilistic forecasts: rare and extreme events are, by definition, expected to lie outside high predictive quantiles (e.g., the 90^{th} or 99^{th} percentile) with a frequency consistent with their observed probability level. While this behavior is statistically correct and reflects probabilistic calibration, it also implies that probabilistic predictions alone do not directly translate into actionable decisions in an

operational flood forecasting context. In practice, decisions such as issuing flood warnings require an explicit mapping from predictive distributions to actions, which involves balancing competing objectives such as detection of extreme events and avoidance of false alarms. Static quantile thresholds represent one simple strategy to perform this mapping, but they implicitly fix this trade-off and do not adapt to changing conditions.

The FRiCA framework is therefore introduced to explore a more flexible alternative, where decision rules are learned directly from the probabilistic forecasts in a data-driven manner. We agree that, in its current form, this component should not be interpreted as a fully developed or superior operational solution. Rather, its purpose is to demonstrate how probabilistic hydrological predictions can be coupled with decision-making frameworks, and to highlight the potential for moving beyond static threshold-based approaches.

We will revise the manuscript accordingly to better position FRiCA as an exploratory extension and to clarify its role within the overall study.

3. Furthermore, I found the methodological description of FRiCA is little difficult to follow. This section (Section 2.5) would benefit from clearer conceptual framing. In particular, the state–action–reward structure is not clearly defined, making it hard to understand what the agent is actually learning. In addition, the reward design (e.g., +100 vs. −5000 vs. 0) and the masking based on the 85th quantile could be better justified, as it's not clear whether those choices influence the learned behavior. Overall, I recommend adding a concise conceptual overview of the decision problem, defining the RL components better, and providing clearer justification (or sensitivity analysis) for the reward structure and activation threshold.

We will add a flow chart/schematic representation of the framework in Section 2.5 (see also conceptual/methodological question 5 in RC1). On top of that, we will revise this section for clarity following your suggestions.

4. The introduction contains a comprehensive discussion of the saturation limit problem of LSTM networks, which is also one of the major motivations of this study. The authors also point out a few strategies they have previously implemented to partially overcome these limitations. However, in the analysis of this study, no focus has been made on the saturation limit problem, as seen from the result that although probabilistic LSTMs are somewhat better than deterministic LSTMs, the saturation problem still persists with them. Therefore, either some supporting analysis, perhaps incorporating ideas the authors have previously hinted at, or at the very least a detailed discussion of this limitation as it applies to the probabilistic LSTM, is necessary.

We agree that we missed the opportunity to link the discussion to the saturation problem raised in the introduction. As a matter of fact, we expect saturation to occur in the probabilistic models

as well since they use the LSTM as their backbone. Saturation in the probabilistic LSTMs, should be relatively less of an issue, because the model to accounts for the whole distribution and not just the expectation. Saturation will occur in the predicted parameters of the CMAL and the DRN and the raw coefficients of the BQN, and ultimately manifest as miscalibration. We will add this discussion in the revised manuscript.

Minor comments:

In general, paragraphs are far too long, making it hard to separate out key ideas and natural breaks in the narrative. I would introduce more paragraph breaks throughout to improve readability.

The manuscript will be modified for better readability.

Lines 14-16: Provide quantifiable estimate rather than only stating that probabilistic predictions capture extreme events within their bounds.

We will rephrase this in the revised manuscript.

Lines 97-98: The motivation here is not clear. We already have a probabilistic LSTM model that predicts the parameters of the CMAL distribution, which is flexible enough and should in principle be capable of capturing extremes. You introduce two alternative approaches, but why are these necessary? What is the limitation of CMAL, does it have a property that causes it to struggle with extremes, and do the two methods you introduce (DRN and BQN) have property that could potentially better represent extremes? In short, out of the many distributions and methods available, the rationale for why these two in particular are used here is missing.

We agree that the introduction currently lacks our reasoning for choosing the three methods. Our primary goal with this study was to add to the existing methods of uncertainty quantification in deep learning-based hydrology. We chose to test the DRN and the BQN based on their successful application in post processing of weather predictions. A good choice of distribution for the DRN is the one that represents the conditional distribution of the target data. If we expect the tails of the to-be-estimated distribution to be sub-exponential it is a somewhat bad choice. The logistic distribution is an exponential-tailed distribution and has heavier tails than the gaussian. Hence it is a natural alternative to a Laplace in the streamflow setting (this also somewhat shows in our results, since we just have to use one distribution to get good results). Moreover, it is advantageous as it offers analytical simplicity and stable parameter estimation. The alternative route is to not explicitly parametrize the conditional probability, and one way to do this is to estimate the quantiles of the distribution. The BQN represents a relatively simple non-naive approach (i.e., not directly estimating the quantiles), and was therefore another natural choice. As for the limitations of the CMAL for extremes, Klotz et al. (2022) do not analyze the CMAL for these situations specifically.

In the revised manuscript, we will add the above points, as well as our explanation to major comment #1 to provide a clear motivation for the models we selected.

Lines 111-112: You do not provide any background on the RL method here. Why would RL be a suitable candidate for issuing more reliable flood warnings? There should be references to support this, from hydrology or similar domains, to provide the reader with useful context. On this point and the one above, the authors are assuming a great deal of prior knowledge on the part of readers, which might not be the case. Please provide brief context and background so that the study is more accessible.

While there might exist alternative approaches, the choice of reinforcement learning was intuitive. A reward-based learning set up is advantageous as it eliminates the need of optimization based on squared errors – an approach we tried before resorting to RL methods. We realize that this explanation is currently missing and we will add our rationale for selecting a reinforcement learning based module in the revised version.

Table 2: The hidden size of the deterministic LSTM (64) is substantially smaller than that of the other probabilistic models (256 or 250). Do you have any comment on this? Was the size for the deterministic LSTM tuned specifically for this region, while the size for the probabilistic models was adopted directly from studies using CAMELS-US? For the CAMELS-US dataset, which is much larger than your data, the common hidden size choice is similar for both deterministic and probabilistic LSTMs (256 vs. 250). If 64 hidden units are sufficient for deterministic LSTM in your region, a similar size may also suffice for the probabilistic models, and the larger size may simply be overfitting. I would be interested to hear your explanation for this or see any supporting analysis.

The choice of 64 hidden states for the deterministic LSTM was guided by our previous study (Baste et. al., 2025) where we obtained state of the art performance with this hidden size. As for the CMAL, we wanted to replicate the model from Klotz et al. (2022) as closely as possible and chose the hidden size of 250. For similarity within the probabilistic models, we chose a similar hidden size (256) for the DRN and the BQN. We will conduct some hyperparameter tests for the two hidden sizes and if the results differ substantially, we will report them suitably in the revised manuscript.

Lines 174-178: DRN is not clearly explained based on your description here. Is this a family of many regression methods, or is logistic regression always used with DRN? If the choice to use logistic regression was yours, clarify why?

As we mentioned earlier, the DRN can, in principle, model any distribution of our choice. We justify the choice of the logistic distribution for the DRN in our reply to your minor comment for L. 97-98. The log-normal distribution could be another choice for the to-be-estimated distribution in the DRN. Although it is a sub-exponential distribution, it somewhat represents a boundary to

power-law distributed densities and has all moments defined. It also naturally respects the fact that streamflow cannot be below 0.

We assume, that you mean *logistic distribution* when you say *logistic regression* in your comment here. Logistic regression and DRN differ fundamentally in their scope and applicability. A standard logistic regression applies to a binary classification setting only, while the DRN framework yields a full predictive probability distribution for a real-valued target quantity (which here is a logistic *distribution*) as its output.

Lines 193-196: Streamflow is strictly positive, so why did you decide to remove the strictly positive enforcement? This seems contrary to what should be done, so the reasoning needs to be clarified.

To answer your question briefly, the choice of not truncating the predicted distributions allowed for uniformity within models, as Klotz et al. (2022) do not implement a truncated CMAL.

The opinion on whether or not the predictive distribution should be truncated at zero is split. Such truncation introduces an inductive bias in the model, and in doing so the model doesn't need to learn that the data has a lower predictive bound. The model might not learn as much about the underlying processes or the data, as it would otherwise do. While such a physically informed bias is preferred from a statistic's point of view, it might hinder the model from generalizing better. We will include this clarification here in the revised manuscript.

Line 206-215: This is a bit confusing. The authors first direct us to Schulz and Lerch (2022) and Shulz et al. (2024) to read about how predictive distributions are combined, but then in the next paragraph, they describe the process.

This section will be modified in the revised manuscript.

Lines 245-276: This entire section is somewhat confusing, and the working mechanism of FRICA is not clearly communicated. I suggest breaking it into smaller paragraphs to more clearly explain how it works, and, if possible, adding a conceptual figure illustrating the FRICA mechanism.

We will modify this section for better readability and comprehension. We will also add a schematic diagram explaining the FRiCA framework.

Figure 1: Looking at this PIT histogram, what stands out is that for tail flood events, BQN actually performs better than both CMAL and DRN, both of which behave very similarly to each other. For the most extreme flood events, the histogram mass is much higher than 1 for both CMAL and DRN, suggesting that extreme flood observations more frequently fall in the right tail of the forecast distribution, indicating that extremes are being under forecasted. Given that the main focus of this paper is capturing extreme flood events with probabilistic models, why is DRN considered the best probabilistic method based on average behavior rather than extreme

performance? Based on this figure and the CRPS values provided, I am not entirely convinced that DRN is the best method.

Thank you for this remark! We would like to elaborate on why we consider the DRN as the best method in terms of *overall calibration,* as we mention in L292. Note, that this assessment of the DRN is distinct from the assessment of the three models in section 3.2, where we consider only a few representative floods to exhibit the suitability of the probabilistic models for extreme floods.

Figure 1 and section 3.1 are intended to address the global model performance for the entire test period, before we look at some representative extreme floods in Section 3.2. To answer our first research question, it is essential that we rank the models based on their global performance. For probabilistic models, this performance addresses the calibration as well as sharpness of distributions, and sharp predictive distributions as the cost of poor calibration are undesirable. Thus, we assess the models' global calibration separately (in Figure 1) apart from considering the CRPS scores. We rank the DRN as the best, based on the deviation from a standard uniform distribution. In our opinion, a global assessment is important, because though we would like to focus on the extremes, we want a model which is not necessarily poor for the other flows. A "specialist" model which captures the extreme flows perfectly can also be achieved with several other strategies such as weighting the squared loss functions, discharge transformations (Hunter et. al. (2021), Thirel et al., (2024)) and the quantile- (Tyralis and Papacharalampous (2021)) or expectile-based (Tyralis et al. (2023)) training. Such models however might perform poorly for the other flow regimes or have deteriorated global performance. Though the studies were conducted for conceptual/process-based models, we also had similar results when we tried some of these strategies for the LSTM.

The fact that the CMAL has the lowest CRPS amongst the three models but slightly worse PIT histogram than the DRN can be justified by its low-resolution metrics. Overall, the CMAL produces sharper distributions which are slightly less reliable than the DRN. The preference for sharper predictive distributions is conditional on the fact that they are calibrated (reliable). Thus, despite low CRPS scores, the DRN is still considered a better model than the CMAL in terms of global calibration.

We hope our response sufficiently resolves your concerns here. We will add these details in the revised manuscript as well. We will make the distinction in the global performance assessment and local performance for extremes more distinct.

Lines 327 – 335: The authors argue that the DRN predictive distributions provide an advantage for extremes, but they cite that only 67% of the extreme flow observations are within the 99th percentile bounds of the predicted distribution. While better than a deterministic model that strongly underestimates the peaks, the performance of the probabilistic bounds of the DRN

model aren't great either. Overall, I would recommend the authors not overstate the performance of the probabilistic models for the most extreme events. This also extends to the title of the paper; it's not clear that probabilistic models 'are key' to improving flood predictions based on these results.

From a probabilistic calibration point of view, we expect that the predicted probabilities match the observed probabilities. Thus, we should expect some events to be predicted beyond the 99th percentile, as not all observed probabilities are within the 99th percentile. To fully address your concern, we require additional notions of calibration, which we will include in the revised manuscript. We do not wish to oversell our results in any way and communicate the deficiencies of our models clearly in the results and discussion. At the same time, we believe that a more detailed discussion including other notions of calibration is essential. We will add additional analyses in the revised manuscript which will address the representation of the most extreme events.

As for the title: we thank you for your comment, as it brings to our attention that the manuscript lacks a convincing discussion as to why we believe that probabilistic LSTMs are key to improving flood predictions. Allow us to briefly justify our title here. In the past, we have tried several strategies in terms of different (deep-learning) model architectures, data augmentation, loss function adaptations and found none to significantly improve our knowledge about the extreme events, while maintaining performance (or not losing it too much) for the other flows. Calibrated probabilistic LSTMs address this issue by giving us reliable information across all flow regimes. Thus, from a knowledge discovery point of view, we believe probabilistic LSTMs are key for improving flood predictions.

We will include a better discussion and analysis of our results, and hope that it convincingly addresses your remark here.

Figure 2: This figure shows the limitations of the deterministic LSTM model and some advantage of the probabilistic model, but it is still not very informative. It is expected that by using a probabilistic LSTM which can assign some probability mass to extreme events, performance would improve to some degree compared to deterministic model. However, given that a large portion of the introduction focused on the saturation limit of LSTM, I was expecting more experimentation or analysis on that topic. This figure simply shows that probabilistic LSTM can be somewhat better, but the saturation problem is still there. The extremes are still clearly underestimated, and most extreme events remain beyond the reach of even the probabilistic LSTM. Also, the figure shows 99th percentile interval for probabilistic LSTM but it is unclear whether the model is assigning meaningful probability mass to those extremes. At a minimum, showing the bounds for the 25th to 75th percentile interval would be important here. And, given that saturation limit was a major focus of the introduction and motivation for this study, I was

expecting some analysis of this issue in the context of the probabilistic LSTM as well, but it is entirely absent.

We will add more analyses here. Please see our response to your major comment 4. We will include an evaluation of our models based on the notion of marginal calibration to address the issue of whether meaningful probability mass is being assigned to the events. We will also include a discussion about the saturation phenomenon of the LSTM and how it will appear as miscalibration in the models.

Typo correct improvemed to improvement

Typo will be corrected.

Line 343 and elsewhere throughout the manuscript: $LSTM_{det}$ is written $LSTM_{d}et$

Typo will be corrected.

Line 396: It might be useful to have a 1-line reminder about how hits, misses, and false alarms are defined

Noted and will be added in the revised manuscript.

Line 398-403: I'm not sure this text is necessary

Line 405-412: This gets at one of my main comments above. It's not clear based on this result that the FRiCA model actually is providing meaningful benefit over simply using the 99^{th} percentile of the DNR predictive distribution. Also, it seems to collapse to just using a different static percentile most of the time (95^{th} percentile). So, it's not clear what the benefit is over just calibrating a static quantile threshold to use from the probabilistic predictions.

We agree that the FRiCA performance can be improved further. Several other algorithms can be tried out with better reward design. We tested a baseline model which calibrates a static quantile threshold, and found that such a model was worse than the FRiCA.

As mentioned in our reply to major comment 2, we will rephrase the description of the FRiCA and mention its exploratory nature in the revised manuscript.

References:

Baste S, Klotz D, Espinoza EA, Bardossy A, Loritz R. Unveiling the Limits of Deep Learning Models in Hydrological Extrapolation Tasks. *Catchment hydrology/Modelling approaches*. Preprint posted online February 6, 2025. doi:10.5194/egusphere-2025-425

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