

Letter to the Editor: egusphere-2026-469

Dear Editor and Reviewers,

We thank the editor for enabling a rigorous review of our work and we thank the reviewers for their time and for providing thorough reviews. The reviewers' comments motivated us to re-examine the details of our work and we are happy to present a revised manuscript. As a result of this review, the manuscript has benefitted greatly in terms of articulation and better readability. We believe that the revised introduction and discussion convey our core message more clearly. We also made some methodological revisions that address the concerns pointed out by the review comments.

The major changes in the revised manuscript are as follows:

- 1. General readability:** A common theme from the three reviews addressed the readability of our work. We have now split the text into shorter, more concise paragraphs that should help our audience grasp the concepts and our message with ease. We hope the reviewers find the revised manuscript more coherent and easy to follow.
- 2. Change in CMAL implementation:** Following the review comments, we investigated the poor performance of the CMAL. We found out that the parameterisation of the Asymmetric Laplace distributions (ALDs) within the CMAL resulted in the poor calibration for the CMAL. Originally, the ALDs were parameterized by an asymmetry parameter $\tau \in (0, 1)$, which is modelled by a sigmoid activation function (adopted from NeuralHydrology, (Kratzert et al., (2022))). In the revised version, the ALDs are parameterized by a skewness parameter $\kappa \in (0, \infty)$, which is modelled by a softplus activation function. We have modified the methods section accordingly and discuss the effects of the two parameters (τ and κ) on the learnt behavior of the CMAL in Appendix B. Additionally, the change in CMAL implementation has affected the results sections 3.1 and 3.2, Figures 1, 2 and 3, and Tables 3 and 4.
- 3. Change in FRiCA implementation:** Although our previous implementation of the FRiCA was technically sound, it was complex and difficult to follow. Following the reviewers' suggestions, we have now simplified the approach and implemented the FRiCA using Gymnasium (Towers et al., 2025) and Stable-Baselines3 (Raffin et al., 2021). Simplifications include modified inputs to the FRiCA and reduced action space. We mention the relevant details in section 2.4. We also conducted a sensitivity study for the reward function (as suggested in RC2) and present these results in Appendix E. As suggested by the reviewers, we mention the exploratory nature of FRiCA clearly in the

introduction and the discussion. However, we believe the revised implementation portrays its potential more strongly.

- 4. Discussion for Research Question 2:** RC3 pointed out that our research question 2 seemed “either trivial or poorly worded”. We have now modified the introduction and the discussion to address this concern. Apart from changing the wording for research question 2, the introduction clearly establishes the need for probabilistic models and the discussion is revised for clarity and coherence.

Apart from the major changes listed above, we have incorporated other reviewer suggestions as well. We give a point-by-point account of all major and minor changes (in response to [reviewer comments](#)) in the Attachments to this letter.

Once again, we are grateful for the detailed feedback we received, and we made exhaustive efforts to address your questions and comments. Although the revision doesn’t affect our original core message, the reviews compelled us to put our work through a fine comb. As a result, we believe the revised manuscript to be technically stronger and an easier reading experience for our audience.

We hope to have addressed your concerns satisfactorily and look forward to the second round of discussion.

On behalf of all authors,

Sanika Baste

Attachment 1:

RC1: 'Comment on egusphere-2026-469', Anonymous Referee #1

[B] Conceptual and Methodological Questions

1. How can deterministic objective functions be adapted to better represent extreme flow regimes? What are the key bottlenecks in doing so?

Response: We address your question in **L477-485** in the discussion. We mention how strategies such as streamflow adaptations, weighting schemes within the squared loss formulations, quantile and expectile losses create specialist models which are not robust of median and low flows. We believe that, this is the key motivation which necessitates the use of probabilistic objective functions for LSTM based hydrological modeling.

2. Would a schematic flowchart illustrating the blueprint of the proposed framework improve clarity and accessibility?

Response: We simplified the FRiCA approach and a schematic representation is now redundant.

[C] Readability Suggestion

The information presented is professional and technically solid. However, I suggest breaking up some of the longer paragraphs into smaller sub-paragraphs—particularly in sections such as the Results—to improve readability.

This is only my second time reviewing an EGU-style preprint, so this suggestion may partly reflect personal preference.

Response: We have revised the MS for better readability considering the above remarks.

[D] Specific Comments

Line 10: What is the definition of “rare events”?

Response: In **L8**, we include a definition of *extreme discharges* in the context of this study.

Line 34–35: “Although LSTMs often outperform conceptual and process-based models for the majority of the flow regime.” Could references be provided to support this claim?

Response: References added in **L29**.

Line 49–54: It may be helpful to move these experimental results to the Appendix.

Response: We revised **L40 - 49** and present our previous findings more concisely.

Line 71–73: The statement regarding one-forcing-to-many-responses may require references.

Response: We added appropriate references in **L71**

Line 111–112: Are there alternative modeling perspectives to this reasoning? Could the authors further explain why a reinforcement-learning–based decision module was chosen?

Response: We provide a background about RL application from water resources management and cite its application for heat wave warning communication in the methods and the discussion sections.

Line 224–225: Why were 5,000 Monte Carlo samples selected?

Response: We revised our calculations based on 7500 samples, and this choice is guided by Klotz et al. (2022)

Line 248: For readers less familiar with reinforcement learning, it would be helpful to clarify: Why is the decision policy parameterized by an LSTM? How is the decision-making policy defined in hydrologic terms?

Response: We updated the FRiCA and have implemented it using Gymnasium (Towers et al., 2025) and Stable-Baselines3 (Raffin et al., 2021). The original description was difficult to follow, and we hope to have addressed that issue with the current implementation with standard libraries.

Line 252: Why were 32 quantiles used?

Response: The action space for the FRiCA is simplified. The agent now chooses from 6 possible actions which correspond to no flood, or flood severity ranging from HQ2 to HQ300.

Line 253–254: Could the authors elaborate on what a “single-step decision process” looks like in the hydrologic context?

Response: We have updated the description of the FRiCA in section 2.4..

Line 260–262: Why were +100 and –5000 chosen as reward values?

Response: We updated the reward structure for the FRiCA and now present the sensitivity analyses for 9 reward configurations in the Appendix E.

Section 2.5: This section contains many implementation details. An illustrative figure could significantly improve clarity.

Response: We had initially agreed to include an illustrative figure, but we believe the new simplified approach is easy to follow from the text itself and a figure would now be redundant.

[E] Performance Analysis

Line 281–288: Have the authors evaluated performance by grouping catchments according to functional traits or hydrologic regimes?

Response: We did not check the spatial distribution of the model performance, in terms of the test period CRPS across Switzerland. While such an analysis would be interesting, it expands the scope of the current work substantially, and we prefer leaving it to further studies.

Line 325–326: Do the authors have insights into why results differ from Klotz et al. (2022) on CAMELS-US?

Response: We traced the poor performance of CMAL to the parameterization of its component distributions. We found out that the parameterisation of the Asymmetric Laplace distributions (ALDs) within the CMAL resulted in the poor calibration for the CMAL. Originally, the ALDs were parameterized by an asymmetry parameter $\tau \in [0, 1]$, which is modelled by a sigmoid activation function (adopted from NeuralHydrology, (Kratzert et al., (2022))). In the revised version, the ALDs are parameterized by a skewness parameter $\kappa \in [0, \infty)$, which is modelled by a softplus activation function. The CAMELS-CH differs considerably from the CAMELS-US dataset, and this might explain why the parameterization matters. We give more details about the effect of the CMAL parameterisation in Appendix B.

Section 3.2: It may strengthen the manuscript to further analyze how the three approaches learn (or fail to learn) uncertainty bounds during events.

Response: We analysed the models based on their test period distributions and gained some insights into the learnt behavior. With the CMAL, the parameterization affects the learnt behavior, and we mention the details in Appendix B. For the BQN, we realised that the model struggles to learn the raw parameters with enough spacing which leads to the miscalibration in the lower PIT bins (mentioned in section 4.1). A larger LSTM network (bigger hidden size) or lower drop-out rate helps address this issues. As for the DRN, truncating the logistic distribution at zero provides improved calibration in lower PIT bins, but not necessarily in the last PIT bin. We mention our findings in section 4.1.

Line 336 (Figure 3): Was any attempt made to improve the poor uncertainty bound of CMAL? Is this behavior associated with the model structure? How generalizable is this finding across locations?

Response: Please refer to our response to your comment on [Line 325–326](#).

[F] Event Definition and Aggregation

Line 397: Please specify which two ARIs are used.

Response: We have mentioned the definition of floods and extreme floods with respect to the ARIs in section 2.4.

Line 401–404: The event definition is unclear. It would be helpful to clearly compare this definition with previous literature and explain how flood categories are identified following BAFU (2024).

Response: We have mentioned the definition of floods and extreme floods with respect to the ARIs in section 2.4.

[G] Additional Technical Clarifications

Line 454–455: More detail on the role of noise-based regularization could strengthen the main argument.

Response: With the new CMAL, a detailed discussion on the role of noise-based regularization is not required anymore.

Line 518–519: A reference supporting the statement about subjective risk choice in reinforcement learning would be useful.

Response: We have added references from the reinforcement learning literature in the discussion for research question 3 (section 4.3).

Line 510–541: Given the technical complexity of reinforcement learning, expanding literature support in this section would improve rigor.

Response: We have revised the methods section 2.4 and the discussion paragraph on research question 3 (section 4.3) to include additional literature references for reinforcement learning.

Attachment 2:

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

Major Comments:

1. The DNR and CMAL models are not really that different. Both estimate the parameters of a distribution for flow, and its 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, its 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.

Response: We implemented the DRN using NLL and found its validation-period calibration to be similar to that of the DRN trained with CRPS. Thus, for the DRN, the choice of loss function — NLL or CRPS — does not lead to a substantial difference in performance. A similar insensitivity to loss function choice has also been reported in the weather post-processing literature (Gebetsberger et al., 2018).

For the CMAL, however, no analytical solution for the CRPS exists, and training with a sample-based CRPS proved non-trivial: we observed exploding gradients within the first few batches, and stabilising training would require careful additional tuning. While we agree that Section 3.1 would benefit from a complete loss-function comparison across both models, we ultimately decided not to invest further effort here, since the comparison would not affect our core message.

We would also like to note that, contrary to what we proposed in our initial response, we do not report these findings in the revised manuscript, as our efforts did not yield a meaningful conclusion. We do, however, acknowledge this limitation — our inability to fully separate the effects of loss function choice from distribution choice — in **L425-427** and **L450-459**.

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.

Response: We have now simplified the approach and implemented the FRiCA using Gymnasium (Towers et al., 2025) and Stable-Baselines3 (Raffin et al., 2021). The simplifications include change in the inputs to the FRiCA and change in the action space. Earlier, the FRiCA would take in long meteorological sequences, but all information from these is already extracted in the probabilistic predictions. Thus, we now input a sequence of only 7 days of CMAL predictive distributions and 1-day lagged discharge. We now choose the CMAL for the FRiCA, instead of the old choice of the DRN, as the CMAL predictions are over-dispersed and indicate an over-safe stance from the perspective of a warning-issuing system.

We also made several sensitivity tests with the reward function, ran every reward configuration with three different random initialisations. Moreover, for every reward configuration and random seed, we tested three choices of the entropy coefficient, which controls the FRiCA's exploration vs. exploitation. However, we only present the results from the best performing configuration in the main text. As the FRiCA is meant to be a proof of concept and we would like to not deviate from our main message a lot, we mention all these additional details about the FRiCA in Appendix E.

We believe the new approach is simpler to follow, and the results present a convincing argument in favour of the FRiCA over the static rules. Nonetheless, we agree that the FRiCA is still exploratory in nature and mention this suitably throughout the manuscript.

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 its 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.

Response: Please see our response to major comment 2. We have revised Section 2.4 to provide clear definitions of our states, actions, and rewards, and we hope the description is now easier to follow. In our original response, we had proposed adding a schematic diagram of the FRiCA (as suggested by RC1); however, we believe this would be redundant given the new, simplified approach.

As with the earlier approach, where events were masked based on the 85th quantile, we now filter episodes based on the exceedance probability of the HQ2 threshold. While this filtering is somewhat arbitrary, it mimics a realistic decision-making scenario: the need for a FRiCA-like agent, or for decision-making informed by human expertise, only arises when there is a substantial chance of the lowest HQ threshold being exceeded. We set this threshold at 15%. We do not expect the learned behaviour to be highly sensitive to this choice — filtering events differently would primarily affect the frequency of class 0 and 1 events, and the resulting class imbalance is already, to some extent, offset by the use of class weights during training. Since factors such as reward function design, random initialisation, and the entropy coefficient have a greater influence on the learned policy than this filtering threshold, we did not consider a separate sensitivity analysis on the threshold choice necessary.

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.

Response: We added a discussion about the saturation phenomenon in probabilistic LSTMs and its effects on extreme flood predictions in **L509-519**

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.

Response: We have modified the text for better readability.

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

Response: **L9** now includes exact percentage of extreme events that are captured within the 99th percentile of the predictive distributions.

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.

Response: We now state our motivation to include the three probabilistic LSTMs clearly in the introduction in **L111-113**

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.

Response: We have added relevant resources from water resources management and other fields where RL has been implemented to issue warnings. We refrain from making the introduction too long as we risk distracting our readers from the essence of our work. We have however incorporated your suggestion and provide background on RL methods in section 2.4 and the discussion for research question 3 (section 4.3).

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.

Response: 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.

Following your suggestion, we tested the DRN with a hidden size of 64 and found its validation performance worse than a DRN with hidden size of 256. We mention these additional details in Appendix A. At least in the context of this study, a DRN with hidden size 256 does shows no overfitting. Guided by your comment, we conducted some more hyperparameter tests for the BQN as well. In general we found that increasing the hidden size improves the validation period

calibration for the BQN. Similar results could be obtained for a hidden size of 256 by reducing the drop-out rate. Thus, the drop-out rate for the BQN is now 0.3 instead of 0.5, and we mention this change in Table 2. Thus, it seems that, in general, larger networks might be conducive to learning, instead of overfitting to the uncertainty in our data. We mention all these details in Appendix A.

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?

Response: We have revised section 2.2.3 to include a better description of the DRN and a note about the to-be-estimated distributions. We also justify our choice of the logistic distribution for the DRN.

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.

Response: 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 generalising better.

Additionally, following your question, we tried modeling the DRN with a truncated logistic distribution. While such a truncation improves calibration of the lower PIT bins (roughly $\text{PIT} \in (0, 0.1]$), it does not improve the calibration for the last bin substantially ($\text{PIT} \in (0.95, 1.0)$). We thus maintain our original stance of modelling non-truncated distributions.

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

Response: To keep the main text concise, we have moved the original section 2.4 to Appendix D. We explain how we combine the predictive distributions, but also refer to Schulz et al. (2026) as it might be interesting for some of our audience to know more about aggregating forecast distributions.

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.

Response: Please see our response to major comments 2 and 3.

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 underforecasted. 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.

Response: As in our original response, we would like to mention that section 3.1 and Fig. 1 are meant to assess the global model performance and this assessment is separate from our analyses in section 3.2 and Figs. 2 and 3. We agree that the model with best global calibration might not be the best when it comes to predicting extreme floods and issuing warnings. In fact, the new (over-dispersed) CMAL, although poorly calibrated, is more suitable for extreme flood prediction. This is also the reason why we now use the CMAL for the FRiCA. We mention all these details suitably in the results and the discussion sections.

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.

Response: As noted in our original response, we would like to reiterate an important point regarding probabilistic calibration: it implies that predicted probabilities match observed probabilities in the long run, not that all observations will fall within a given percentile bound. We should therefore expect some observed events to fall beyond the 99th percentile — and it may simply be the case that these are precisely the events in the top 0.1% of the marginal distribution, which is consistent with our original results.

From an operational perspective, however, this is unsatisfying, since it leaves us with little confidence in our ability to anticipate these high-stakes events. This prompted us to consider whether over-dispersed — though technically miscalibrated — distributions might be preferable for operational use. Our new CMAL implementation produces such over-dispersed distributions, and its effect is visible in Fig. 3. Following your suggestion, we also revised the original Fig. 2 (now Fig. 3) to include additional detail; for legibility, we zoomed in on a smaller fraction of the marginal distribution (0.05% instead of 0.1%). We find that a substantially larger fraction of these events (~83%) now falls within the 99th percentile, a direct consequence of the distribution's over-dispersion.

In our original response, we also proposed examining additional notions of calibration, and we accordingly assessed marginal calibration. The DRN predictions show near-perfect marginal calibration; however, marginal calibration is a weaker criterion than probabilistic calibration and is computationally expensive to compute. While we think that a more comprehensive assessment of calibration might be interesting, this would require additional discussions on which notions of calibration would be appropriate to evaluate our models from an applied point of view. In our opinion, this is beyond the scope of the current paper. We therefore chose not to include these additional analyses. Instead, we discuss the implications of probabilistic calibration for upper-tail events specifically, and why over-dispersion may be preferable to calibration, especially when greater confidence in extreme events is operationally more valuable than statistical calibration itself.

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.

Response: We have modified Fig. 2 to include the inter-quartile range and the range between the 1st and the 99th percentile. Please note, that for legibility, we had to change the range of

events slightly, and the new Fig. 3 now includes the top 0.05% of the marginal distribution. The interpretation is still the same, in fact more pronounced, since we now look at an even smaller fraction of the marginal distribution.

We added a discussion about the LSTM saturation phenomenon in **L509-519**.

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

Response: Please refer to our response to major comment 2 and 3.

Attachment 3:

RC3: 'Comment on egosphere-2026-469', Jonathan Frame

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.

Response: We made several changes to address this comment, please find them listed as follows:

"None of them turn out to sufficiently capture the most extreme events":

Our original results showed that only 67% of the top 0.1% of the marginal distribution fell within the DRN's 99th percentile bounds, which could be attributed to miscalibration in the DRN predictions. As we noted in our original response, we would like to reiterate an important distinction here: probabilistic calibration implies matching the true (observed) probabilities with the predicted ones. By definition, a perfectly calibrated model will still have ~1% of all observations fall outside its 99th percentile interval. Events from the upper tail of the marginal distribution are, by their nature, disproportionately likely to be among those predicted beyond the higher quantiles, precisely because they are rare. In other words, probabilistic calibration and tail-event coverage are not the same objective. A model can be well calibrated overall and still perform poorly at capturing the specific extreme events that an operational flood-warning system cares about.

This distinction, we believe, directly addresses your observation that you were "left wondering if probabilistic streamflow in general can't capture those extremes.". We propose to frame your thoughts slightly differently. It may not be that probabilistic streamflow models are incapable of capturing extremes, but rather that pursuing calibration and pursuing tail-event coverage can be, to some extent, competing goals. Our new CMAL implementation, for instance, produces over-dispersed predictions that are miscalibrated globally, yet assign greater probability mass to extreme outcomes, and consequently perform better at capturing them. We now discuss this trade-off explicitly in the manuscript.

"...but I wonder if the authors simply chose the other two methods almost randomly":

Response: We recognise, that the manuscript lacked a clear motivation for our choice of the three models — a point also raised by RC2. We have now addressed this gap in the introduction.

"...If not, is it a limitation of LSTM itself?":

Response: We have added a discussion of saturation effects in probabilistic LSTMs in **L509–519**.

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 have rephrased and revised the discussion for research question 2.

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.

Response: We revised this in **L31-32**.

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 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. ...

Response: We revised the conclusion and now report findings specific to our study, while avoiding to make blanket statements.

... 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".

Response: We revised the conclusion and now report findings specific to our study, while avoiding to make blanket statements.

References:

Kratzert et al., (2022). NeuralHydrology --- A Python library for Deep Learning research in hydrology. *Journal of Open Source Software*, 7(71), 4050, <https://doi.org/10.21105/joss.04050>

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N.: Stable-Baselines3: Reliable Reinforcement Learning Implementations, *Journal of Machine Learning Research*, 22, 1–8, <https://jmlr.org/papers/v22/20-1364.html>, 2021

Schulz, B., Köhler, L., and Lerch, S.: Aggregating Distribution Forecasts from Deep Ensembles, *Machine Learning*, 115, <https://doi.org/10.1007/s10994-025-06946-3>, 2026

Towers, M., Kwiatkowski, A., Terry, J., Balis, J. U., Cola, G. D., Deleu, T., Goulão, M., Kallinteris, A., Krimmel, M., KG, A., Perez-Vicente, R., Pierré, A., Schulhoff, S., Tai, J. J., Tan, H., and Younis, O. G.: Gymnasium: A Standard Interface for Reinforcement Learning Environments, <https://doi.org/10.48550/arXiv.2407.17032>, arXiv:2407.17032 [cs.LG], 2025.