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

[A] Overall Assessment

This is an interesting and well-executed study. I personally learned a great deal while reading it. Strictly speaking, I don't see any major concerns that would prevent publication. However, I encourage the authors to reflect on and further clarify the following points to help strengthen and polish the manuscript.

Thank you, for your positive assessment! We find your conceptual questions in [B] thought provoking, as they help us approach the core of our work from different perspectives that are lacking in the current manuscript. We will incorporate them in the relevant sections of the manuscript.

[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?

Over the past several decades, deterministic objective functions in hydrological modeling have been adapted to better represent extreme flow regimes. Strategies include weighting schemes within least squares loss functions that implicitly or explicitly prioritize high flows, transformation of target variables (Hunter et al. (2021), Thirel et al., (2024)), and the development of alternative performance metrics (Hunter et al. (2021)). Quantile (Tyralis and Papacharalampous (2021)) and expectile-based (Tyralis et al. (2023)) losses have been proposed recently to target different regions of the flow regimes. An important aspect of these methods is that they achieve improved performance for specific flow regimes, which might be accompanied by reduced global performance. Thank you for mentioning this and we will address this in the revised manuscript.

2. What are the advantages and limitations of uncertainty quantification (UQ) compared to multi-criterion calibration?

UQ and multi-criterion calibration address complementary aspects of model evaluation. UQ provides explicit estimates of predictive uncertainty and the approaches we use could be optimized by using multiple error metrics. In contrast, multi-criterion calibration constrains model performance across multiple hydrological signatures, yet it does not necessarily yield probabilistic predictions. We therefore view both approaches as complementary rather than interchangeable, with UQ targeting predictive reliability and multi-criterion calibration targeting structural and process consistency.

One interesting formulation could however be UQ as a form of multi-criterion decision in that we do not only care about the precision of the point estimate, but also higher order properties of the predictions.

3. What insights can deterministic modeling–based ensemble approaches provide for likelihood-based training criteria?

As a general practice, as well as in our study, we implement a 5-member LSTM ensemble to account for randomness in initialization. The ensemble members do not differ significantly in global model performance. However, from previous studies (Baste et. al., 2025), we know that there can be significant variance within the ensemble for several events in the discharge time series. We are currently running some experiments with much bigger ensembles (100 members), and the results should tell us more about the insights of ensemble approaches for uncertainty quantification.

Note that, while dealing with deterministic model ensembles, one must deal with the notions of epistemic and aleatoric uncertainties. If such ensembles are not specifically trained to capture aleatoric uncertainty, the difference in the ensemble members is usually interpreted as indicators for epistemic uncertainty.

4. Would it be useful to expand the introduction to clarify how "events" are defined within the hydrologic science literature?

Generally speaking, in hydrology, "(rare/extreme) events" mean rare realizations in the tails of the discharge distribution and the exact definition of "how rare" varies with context. In the context of the study, we specifically focus on the upper tail. While the latter sections of the manuscript define events based on annual recurrence interval (such as in Coles S. (2001) and Wilks D S. (2011)), we keep the term rather general in the introduction. We will modify the introduction to include a brief definition of "events" in the context of the study and their general definition in hydrologic literature.

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

As per your specific comment on Section 2.5, we assume you speak about the FRiCA here. We agree that a schematic figure will make the FRiCA framework easier to understand and we will add one in the revised manuscript.

[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.

We will revise the manuscript for better readability.

[D] Specific Comments

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

“Rare events” are events with low probability of occurrence/exceedance. This probability threshold can be flexible and is context-dependent. In our manuscript, we take a closer look at discharge values belonging to the top 0.1% of the distribution. We will revise the abstract to clearly define “rare events” in the context of this study.

Line 14: “The deterministic LSTM underestimates more than 90% of observed values.” Is this conclusion derived from results obtained across all single-objective functions? What is the basis for this statement?

This is a result from a state-of-the-art implementation of the LSTM which makes deterministic predictions for discharge and is trained by minimizing the mean squared error. We will rephrase this line to include these details.

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?

The said improved performance of the LSTM as compared to conceptual models, is a common finding in the studies mentioned in L28-31. We will, however, cite them appropriately again in the above lines.

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

We will rephrase this section to highlight the contributions of our previous study without affecting the readability and the essence of the introduction.

Line 69–70: “Hydrological systems are, in principle, deterministic dynamical systems...” Could the authors clarify what theoretical framework or principle this statement refers to?

Following classical mechanics and dynamical system theory a hydrological system is a deterministic dynamical system. We will add these additional theoretical framework details here.

Line 71–73: The statement regarding one-forcing-to-many-responses may require references. We will support this claim with appropriate reference, if any.

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?

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. Model-based reinforcement learning essentially studies how agents can learn to make decisions if they have a model of the environment. In our case the task is to choose a suitable quantile based on a probabilistic prediction and environmental information in the form of the meteorology and then get a reward signal based on the decisions that chose well. We agree that this explanation is currently missing and we will add our rationale for selecting a reinforcement learning based module in the revised version.

Section 2.2: Is it necessary to clarify whether the network is operating in an extrapolation setting (e.g., PUB)?

The models are not operating in an extrapolation or PUB setting. All our inferences are drawn from spatially in-sample and (only) temporally out-of-sample testing. This means that the same set of catchments are used for training, validation and testing of the models, with the entire data being split (temporally) into training, validation and testing periods.

Line 146: Does “five-member ensemble” refer only to different random seeds?

Yes.

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

This was a choice dictated by computational limitations and was made after discussion with experts. We will mention that in the revised manuscript.

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?

We will include a more comprehensive description of the LSTM based decision-making policy network in the given hydrological context.

Line 252: Why were 32 quantiles used?

The use of a continuous or a larger discrete action space did not provide improved results, but required more computational resources. While we did not conduct a systematic tuning of the

action space, the choice of 32 unevenly spaced quantiles gave the best results without excessive computational demand. We did not check if a similar performance can be achieved for lesser quantiles (and thus, even lesser computation requirement). But discretizing the action space further (more than 32 quantiles) or having a continuous action space (quantile selection modeled by beta distribution) provided negligible improvement in results at the cost of a larger computational demand.

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

In simple terms, a 'single-step decision process' can be understood as follows: the agent's task is complete after choosing an appropriate quantile (decision) for a single day. In a hydrological context, this would mean the agent has relevant input information about the meteorological forcing and it should decide the appropriate quantile from the predictive distribution of a single day. We will describe this in more detail in the revised manuscript.

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

The choice of these values was guided by some preliminary tests. The reward calculation is limited to certain events for which the 85^{th} percentile of the predictive distribution is beyond a certain threshold. Despite this restriction the reward signal was found to be noisy and the choice of penalizing missed alarms more than rewarding hits was deliberate, which led to the choice of the said rewards.

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

We will add a schematic representation of the *FRiCA* in the revised manuscript.

[E] Performance Analysis

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

We made limited efforts to check the spatial distribution of the model performance, in terms of the test period CRPS across Switzerland. We will conduct such an analysis, and if meaningful conclusions can be drawn, we will include them in the revised manuscript.

Line 310–313: The manuscript mentions that 9% of test data fall below the 1st quantile of BQN predictions. Is this behavior visible in any figure?

In the present version of the manuscript, we do not have a figure that shows these results. We could add a figure that is similar to Figure 2 for the BQN predictive distributions for the sorted test discharge data. However, readers would have a hard time spotting said 9%. Isolating these events in a figure would not be very meaningful either. Hence, we prefer not supporting these results by any figure.

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

The difference in performance is most likely because the CAMELS-CH differs considerably from the CAMELS-US dataset. The CAMELS-CH has less climatological variance across basins and snow processes play a central role in almost all basins.

Section 3.2: It may strengthen the manuscript to further analyze how the three approaches learn (or fail to learn) uncertainty bounds during events. For example: How does LSTM parameterization affect head-layer behavior? What are we learning mechanistically from this comparison?

The effect of the LSTM parameterization, such as the number of hidden states, can be seen as saturation effects in the head layer which will affect the parameters of the distributions (for CMAL and DRN) and the basis coefficients (for the BQN). We will try to conduct an analysis as per your suggestion and if we can draw substantial conclusions, we will report them in the revised manuscript.

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?

We only conducted a limited set of sanity checks. That is, we tested the hyperparameter ranges from Klotz et al. (2022) (especially the noise regularization), but did not consider a wider search beyond these ranges. From a scientific perspective it was important for us that each approach is allocated roughly the same number of resources. It might be possible to improve the CMAL predictions with a lot of resources and time, but then the question arises whether this would not also be possible for the other approaches. We leave it to future work to focus on this question. That said, the finding is generalizable across Switzerland and we find that for events with very low probability of occurrence, the predictive distributions are very wide with more probability mass associated with lower or negative values.

Line 343: Were rising and falling limbs quantitatively separated during analysis?

No, we did not separate the hydrograph during analysis. The reference to rising and falling limbs in this context is only inferred visually from the hydrographs shown in Figure 3.

Line 346: How does CMAL learn heteroscedasticity? Is this assumption embedded in the objective function?

CMAL estimates a mixture of densities for each input provided to the model. Hence, if the variance or other properties of the conditional distribution change with some input (e.g., more precipitation) the model can adjust for them. Since mixtures of densities are (in theory) universal density approximators they are able to account for many different forms of heteroscedasticity and modalities.

Line 338–379: Are there insights regarding spatial and temporal performance differences? It may also be helpful to report deterministic LSTM performance per location rather than only median values.

We will mention the deterministic LSTM performance in the text. We will also check for any spatial patterns in the probabilistic model performance across Switzerland. If substantial inferences can be drawn from such an analysis, we will include the results in the revised manuscript.

[F] Event Definition and Aggregation

Line 397: Please specify which two ARIs are used.

In our experiments, *floods* are events with annual recurrence intervals of 2 to 10 years and *extreme floods* are events with annual recurrence intervals of 10 to 30 years. This classification is based on BAFU (2024). We will describe the event definition more clearly in the revised manuscript.

Line 399: Does “over 158 catchments” imply that 38 basins were excluded? What is the impact of including/excluding them?

Yes, 38 catchments could not be included in the FRiCA scheme, as extreme flood statics (BAFU, 2024) was not available for these catchments. We speculate that this exclusion has not affected the current FRiCA performance significantly. As mentioned in L400-404, the current selection exhibits significant diversity, which is less likely to change much, were we to include the 38 catchments. Maybe the percentage of catchments where *floods* and *extreme floods* were observed might change (in L403 and L404), but this might not affect the FRiCA performance too much. Thus, we believe the current selection of 158 catchments for the FRiCA is representative

enough. Your question merits a discussion and we will include it in section 3.3 of the revised manuscript.

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).

We will describe the event definition with respect to BAFU (2024) in the revised manuscript.

[G] Additional Technical Clarifications

Line 435–437: Was there any attempt to train likelihood-based models using higher-order moments?

No, we did not consider any methods which use higher-order moments. That said, studying how a regularization for higher-order moments would influence the predictive quality would indeed be interesting. We will mention it in future work.

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

We agree with your suggestion. We will add more details about noise-based regularization to further support our findings here.

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

We will include appropriate reference(s) here.

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

We will include more supporting literature.

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

BAFU: Hochwasserwahrscheinlichkeiten (Jahreshochwasser), BAFU, 2024.

Chaves MA, Espinoza EA, Klotz D, Gupta HV, Ehret U, Guthke A. A variational approach at uncertainty estimation in data-driven rainfall-runoff modeling.

Coles S. *An Introduction to Statistical Modeling of Extreme Values*. Springer; 2001.

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

Wilks D S. Statistical Methods in the Atmospheric Sciences. 3rd ed. Academic Press; 2011.