https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

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

Sanika Baste[1], Sebastian Lerch[2], Daniel Klotz[3], and Ralf Loritz[1]

[1]Institute of Water and Environment, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[2]Department of Mathematics and Computer Science, Marburg University, Marburg, Germany
[3]Machine Learning in Earth Science, Interdisciplinary Transformation University Austria, Linz, Austria

Correspondence: Sanika Baste (sanika.baste@kit.edu)

Abstract. Deterministic model predictions can struggle to adequately capture extreme events such as floods and droughts, which are of particular relevance in hydrology. This limitation arises because deterministic models collapse the conditional runoff distribution to a single point estimate. Probabilistic modeling provides a way to address this issue by explicitly representing uncertainty and assigning non-zero probabilities to a range of possible outcomes, including rare and extreme events, thereby capturing the full range of plausible hydrological responses. Motivated by this perspective, we examine whether probabilistic Long Short-Term Memory (LSTM) models improve the representation of extreme events in rainfall—runoff simulations across Switzerland. Overall, the probabilistic models show good calibration, although some miscalibration remains for the extremes. Differences between models mainly manifest in how uncertainty is distributed: some approaches produce narrower and lighter-tailed distributions, while others yield broader distributions with heavier tails. These trade-offs highlight that probabilistic models differ not only in sharpness but also in how their calibration for rare events. We observe this tradeoff also in models' accuracy metrics. When evaluating the mean of the probabilistic predictions using the Nash–Sutcliffe efficiency (NSE), none of the probabilistic approaches outperform the deterministic LSTM in terms of average predictive accuracy. However, a clear advantage over the determinsitc models emerges when focusing on the tail of the discharge distribution. For the most extreme events (top 0.1% of the discharge distribution), the deterministic LSTM underestimates more than 90% of observed values (since it provides estimates of an expectation), whereas probabilistic predictions can capture a substantially larger fraction of these extremes within their upper predictive bounds. Building on the additional information provided by probabilistic runoff predictions, we further show how they can be translated into actionable flood warnings using reinforcement learning. To this end, we introduce a Flood Risk Communication Agent (*FRiCA*) that operates on probabilistic runoff predictions and learns decision rules for issuing warnings of varying intensity. The *FRiCA* is implemented as an LSTM-based policy network and is trained by rewarding correct warning levels while penalizing the underestimation of flood severity. Results indicate that the *FRiCA* outperforms simple fixed heuristics, such as issuing warnings based on the predictive mean or a fixed high quantile (e.g., the 99th percentile). While this behavior already demonstrates the potential of reinforcement learning for improved flood risk communication, it also motivates further exploration of better reward design and policy network definition for context-dependent decision policies that adapt to varying hydrological and societal contexts.

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

1 Introduction

Deep learning (DL) models, particularly Long Short-Term Memory networks (LSTMs; Hochreiter and Schmidhuber, 1997), have established themselves as a powerful alternative to traditional hydrological modelling approaches and are now the benchmark for rainfall–runoff modeling across varying hydrological settings (Kratzert et al., 2019a). Ever since, much research has focused on applying LSTMs in different regions of the world (Lees et al., 2021; Loritz et al., 2024; Frank et al., 2023; Clark et al., 2024) and on further improving predictive performance by implementing various strategies (Feng et al., 2022; Kapoor et al., 2023; Patra et al., 2024; Li et al., 2023; Khoshkalam et al., 2023; Ma et al., 2024). However, as their application becomes increasingly widespread, a growing body of research is now examining the conditions under which LSTMs reach their limits (Baste et al., 2025; Heudorfer et al., 2025). In particular, an important open question concerns their ability to reliably predict hydrological extremes. Although LSTMs often outperform conceptual and process-based models for the majority of the flow regime — performance being assessed as an aggregation over long continuous hydroclimactic time series — their advantage can weaken substantially when confronted with rare, high-magnitude events. Frame et al. (2022) employed a stress-testing approach in which models were trained on an intentionally truncated dataset, excluding all years containing floods exceeding a 5-year recurrence interval. They found that the LSTM performance gains over conceptual models diminish with increasing event magnitude and, for the most extreme floods, the conceptual model even outperformed the LSTM (see corrigendum from Frame et al., 2022). Using a similar experimental setup, Acuna Espinoza et al. (2024b) compared an LSTM with hybrid model variant (from Feng et al., 2022) and a conceptual model (the Hydrologiska Byråns Vattenbalansavdelning (HBV); Bergström, 1992). Once again, their results showed that the performance advantage of the LSTM decreases as flood magnitude increases, with all three models showing similar underestimation for the events with the largest recurrence period. These findings suggest that the strong performance of LSTMs relative to traditional models does not automatically extend to the tail of the flood distribution. This poses a significant limitation given the importance of extremes in hydrological forecasting, risk assessment, and operational decision-making.

To further investigate the behavior of LSTMs when predicting extremes, we analyzed their extrapolation performance in Swiss catchments (Baste et al., 2025) and found that, although the training data contained peak flows of up to ~180 mm d^{-1}, the LSTM failed to produce discharge beyond a theoretical ceiling of ~75 mm d^{-1}. Building on this finding, we designed a new stress test approach using artificially constructed precipitation time series with statistically derived extreme precipitation events. In this experiment, the most extreme precipitation events in the test period were replaced with design storms with annual return intervals of 50, 100, and 300 years. Such an approach directly relevant for infrastructure design and the evaluation of rare, high-impact floods. These experiments showed that the LSTM not only underestimates the flood response for these unseen extremes but also exhibits an empirical upper bound of ~60 mm d^{-1}. These findings are consistent with the gating-related limitations discussed in Kratzert et al. (2024). This systematic underestimation persists across the tail of the discharge distribution, even when flood-peak magnitudes expected for the design storms are present in the training data. Together with the results of Frame et al. (2022) and Acuna Espinoza et al. (2024b), these findings provide converging evidence that LSTM architectures, and the way in which they are currently applied, exhibit constraints when used for extreme events prediction, particularly in situations

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

that require extrapolating beyond historically observed precipitation magnitudes. However, this does not represent a dead end; several promising modelling strategies offer potential pathways to overcome these limitations.

Baste et al. (2025) explored some of these strategies to improve the extrapolation behavior of LSTMs, including increasing the amount of training data following Kratzert et al. (2024), adapting the loss function, and modifying the activation functions within the LSTM. While each of these adjustments resulted in small improvements, none of them resolved the core issue, and the LSTMs continued to exhibit restricted simulations for the highest floods and a persistent bias for extreme values. One important factor contributing to these biases in hydrological modeling is that identical or similar sequences of meteorological forcing can produce multiple, non-unique hydrological responses. For example, a 50 mm d^{-1} rainfall event may trigger an extreme flood response depending on antecedent moisture, snow conditions, seasonality, and the catchment state. From a learning perspective, the model is asked to infer a deterministic mapping from inherently ambiguous inputs conditions and systematic underestimation of rare events becomes likely. Hydrological systems are, in principle, deterministic dynamical systems whose future evolution is uniquely defined, given complete knowledge of all relevant states. However, this theoretical determinism breaks down in practice. Owing to errors in the meteorological and hydrological data, uncertainty in the internal storage states, and limitations in how data-driven models represent or retain these states, models fail to accurately map the one forcing-to-many responses relations. This ambiguity in response generation further highlights why probabilistic models are conceptually better suited for runoff prediction, particularly for the extreme floods. To exemplify, if out of ten superficially similar precipitation events, perhaps only one produces an extreme flood, a deterministic model is forced to collapse this ensemble of possible outcomes into a single expected value. Given this uncertainty, it seems logical to work with probabilistic models; a fact that has long been accecpted in traditional hydrology (Beven and Freer, 2001). Cloke and Pappenberger (2009) reviewed the use of ensemble prediction systems (EPS) for flood forecasting and compiled the advantage of ensemble hydrological predictions for key case studies across Europe. Bálint et al. (2006) reported that sets of hydrological predictions increased the capapbility to issue flood warnings for the floods in August 2002 at the Devin and Budapest gauging stations. Roulin (2007) looked at long term hindcasts for two Belgian catcments and report the superiosity of probabilistic forecasts in terms of greater Brier Skill scores and economic value based on a cost-loss decision model, than deterministic ones. Several studies which investigated the use of EPS for hydrological forecasting also highlight the advantage of additional information from ensemble forecasts in providing "clear indication of the possibile occurence of the event" (Roulin and Vannitsem, 2005) or "the indication of the possibility of an extreme event" (Gouweleeuw et al., 2005).

A first step in addressing ambiguity also in hydrological deep learning was taken by Klotz et al. (2022), who introduced three methods for uncertainty-estimation in a regional rainfall–runoff model trained on the CAMELS-US (Catchment Attributes and Meteorology for Large-sample Studies - US Addor et al., 2017) dataset. Their results showed that mixture density networks (MDNs), particularly the Countable Mixture of Asymmetric Laplace (CMAL) model, produced the most reliable predictive distributions. This is highly relevant for flood forecasting, as the key benefit of probabilistic models is not necessarily improved median performance, but that the upper quantiles from the predictive distributions are better able to represent the uncertainty associated with flood-generating events. In other words, probabilistic models can allocate meaningful probability mass to extreme runoff outcomes, while deterministic models collapse the probable outcomes into a single, biased point estimate. Chaves

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

et al. (2025) introduced the variational LSTM which makes non-parametric probabilistic predictions and learns uncertainty representations directly from the data. Their approach have no a priori assumptions about the modeled distribution and their results are consistent with Klotz et al. (2022). Klotz et al. (2022) was primarily focused on benchmarking different approaches for capturing overall predictive uncertainty. In this study, we take a step further and investigate whether alternative probabilistic modeling approaches may better capture hydrological extremes. We draw inspiration from the Distributional Regression Network (DRN; Rasp and Lerch, 2018) and the Bernstein Quantile Network (BQN; Bremnes, 2020), which was introduced in the literature on post-processing ensemble weather forecasts and adapted for wind gust forecasting by Schulz and Lerch (2022). These approaches broaden the space of probabilistic models, both in terms of the objective functions used for training and the predictive distributions they can learn.

Although probabilistic forecasts are, in theory, capable of capturing the likelihood of extreme events by allocating probability mass to rare but high-impact outcomes, operational flood management still requires binary or categorical decisions. Consequently, the full predictive distribution must ultimately be collapsed into a single actionable value to determine, for example, whether a warning should be issued and at what severity. For a well-calibrated model, this choice reflects the level of risk decision-makers are willing to accept. Conventionally, the mean, median, or a fixed quantile is selected as a single-value summary of the probabilistic predictions. However, this approach may be suboptimal under varying hydrological conditions. Deterministic choices based on fixed distribution summaries, such as the mean or median, can leave decision-makers vulnerable to high-risk events, while selecting a very high quantile (e.g., the 99th) may trigger unnecessary warnings (false alarms). To address this trade-off, we complement the three probabilistic modelling approaches by additionally testing a Flood Risk Communication Agent (*FRiCA*), a reinforcement-learning-based decision module.

In summary, while deep learning has substantially advanced hydrological modelling, its capability to predict extremes remains constrained. To address this limitation, we investigate whether probabilistic modeling and reinforcement learning together can enhance the identification, quantification, and communication of floods. To address this, we postulate the following research questions:

1. Which of the investigated probabilistic models provides the most reliable rainfall–runoff simulation?
2. Why are probabilistic models more suitable to capture hydrological extremes and the tail regions of the discharge distribution?
3. Can reinforcement learning aid operational flood-warning communication by learning a risk-aware mapping from predictive distributions to actionable warnings based on flood-intensity categories?

The paper is structured as follows: we give a description of the dataset and the models in section 2. This section also describes the Flood Risk Communication Agent (*FRiCA*). Section 3 first presents the global model performance results followed by the probabilisitc predictions for representative floods in Switzerland. Lastly, we present the results from the *FRiCA*. We discuss our research questions in light of our results in section 4 and give our conclusion in section 5.

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

2 Data and Methods

We begin this section by describing the CAMELS-CH dataset (section 2.1). We then describe the models used in this study, namely the Long Short-Term Memory network (LSTM_{det}; see section 2.2.1), and the three different probabilistic frameworks: the Countable Mixture Asymmetric Laplacians (CMAL in section 2.2.2), the Distributional Regression Network (DRN in section 2.2.3) and the Bernstein Quantile Network (BQN in section 2.2.4). In the subsequent subsections (sections 2.3 and 2.4), we explain the processes for combining the predictive distributions from an ensemble and their evaluation. Lastly, we describe the reinforcement learning-based decision module in section 2.5.

2.1 The CAMELS-CH Dataset

The CAMELS-CH dataset (Höge et al., 2023) is a large sample dataset of daily hydro-meteorological time series and static catchment attributes for 331 catchments and lakes from Switzerland and neighboring countries. The static attributes include climate, hydrology, topographic, soil, geology, glacier, hydrogeology, land cover, and human influence attributes. This time-series data is available for the period from 1 January 1981 to 31 December 2020. For this study, we exclude 33 lakes and 102 catchments belonging to France, Germany, Austria, and Italy and proceed with the remaining 196 for training and testing our models. For all the models, we use four meteorological forcings, namely, precipitation (mm d^{-1}), minimum temperature (°C), maximum temperature (°C), and relative sunshine duration (%) along with 22 catchment attributes (see Appendix A1) as the input and specific discharge (mm d^{-1}) as the target.

Table 1. Description of models used in this study. All models are variants of the state-of-the-art LSTM_{det}, differing only in the head layer.

Model	Description
LSTM_{det}	State-of-the-art deterministic LSTM. The head layer is a fully connected linear layer mapping the hidden states to a single output. The output is the specific discharge (mm d^{-1}).
CMAL	Probabilistic LSTM producing a predictive distribution of specific discharge (mm d^{-1}), which is a mixture of three asymmetric Laplace distributions. The head layer is a fully connected linear layer mapping the LSTM hidden states to weights and (location, scale, and asymmetry) parameters for the three component distributions.
DRN	Probabilistic LSTM producing a single logistic distribution as the predictive distribution of specific discharge (mm d^{-1}). The head layer is a fully connected linear layer mapping the LSTM hidden states to location and scale parameters of a logistic distribution.
BQN	Probabilistic LSTM which produces a non-parametric quantile function in the form of a linear combination of Bernstein polynomials, for specific discharge (mm d^{-1}). The head layer (a fully connected linear layer) outputs raw coefficients, which are converted into Bernstein basis coefficients to obtain the linear combination of Bernstein basis polynomials.

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

Table 2. Hyperparameters for the models used: LSTM_{det}, CMAL, DRN and BQN

Hyperparameter	LSTM_{det}	CMAL	DRN	BQN
Number of layers		1		
Number of nodes	64	250	256	256
Dropout rate	0.4	0.5	0.4	0.4
Initial forget gate bias		3		
Initial learning rate	0.001	0.0005	0.0001	0.00005
Sequence length		365		
Batch size		256		
No. of epochs	20	10	10	10
Training period		1 October 1980 to 30 September 1995		
Validation period		1 October 1995 to 30 September 2000		
Test period		1 October 2000 to 30 September 2020		
Number of outputs	1	12	2	13
Training loss	NSE*	Negative log-likelihood	Continuous ranked probability Score	Quantile loss

2.2 Models

The study compares four models, which are all, based on the state-of-the-art LSTM. The models have differing head layers, each of which are described in Table 1. Sections 2.2.1 to 2.2.4 explain the models in further depth, along with the predictive distributions modeled by the probabilistic models for the target variable (specific discharge mm d^{-1}. For every model in the study, we train a five-member ensemble to account for stochasticity and random initialization. All model implementation is done in PyTorch (Paszke et al., 2019) and optimized using the ADAM (Kingma and Ba, 2017). The training period for all the models extends from 1 October 1980 to 30 September 1995, and the models are tested for the period from 1 October 2000 to 30 September 2020. The sequence length and the batch size are 365 and 256 respectively. All models are initialized with a forget gate bias of 3. Table 2 gives all the hyperparameters for the models, and we explain the hyperparameter search in Appendix A2. All models instances are *regional*; implying that the models are trained collectively on all 196 catchments mentioned in section 2.1.

2.2.1 LSTM_{det}

The LSTM_{det} is the state-of-the-art LSTM, as proposed by Kratzert et al. (2019a) and makes deterministic predictions for specific discharge (mm d^{-1}) at each time step. The head layer in the LSTM_{det} thus maps the hidden states to a single value at every time step. For a detailed description of the LSTM_{det}, we refer the reader to Kratzert et al. (2019a). The catchment averaged Nash-Sutcliffe efficiency (NSE*) (Kratzert et al., 2019a) is the loss function during training.

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

2.2.2 Countable Mixture of Asymmetric Laplace Distributions

Klotz et al. (2022) implemented the CMAL for the CAMELS-US (Addor et al., 2017) dataset, which predicts the probabilistic distribution for the target (specific discharge mm d^{-1}) at every time step. Among the four probabilistic models implemented by Klotz et al. (2022), the CMAL is the best performing model in terms of the calibration of the predictive distributions. The predictive distribution from the CMAL is a weighted sum of three asymmetric Laplace distributions. An asymmetric Laplace distribution is a parametric distribution characterized by asymmetry, location and scale parameters, and for three such compoment distributions, we have a total of 9 parameters. To obtain the weighted sum, we also require one weight associated with every component distribution. Thus, for the CMAL, the model produces 12 outputs. The CMAL is trained by minimizing the negative log-likelihood (NLL) of the observed data with respect to the predictive distribution. In Appendix B, we describe the asymmetric Laplace distribution (see Eq. B1), the mixture distribution (see Eq. B2), and the negative log-likelihood (NLL; see Eq. B3) for the mixture distribution. The hyperparameters of the CMAL are given in Table 2, and we explain the hyperparameter search for the CMAL in Appendix A2.

2.2.3 Distributional Regression Network

Rasp and Lerch (2018) proposed a neural network-based distributional regression network (DRN) for post-processing ensemble weather forecasts. The forecast distribution is chosen as a suitable parametric distribution, the parameters of which are obtained as the output of a neural network. The DRN approach has been extended to various target variables, including wind speed (Schulz and Lerch, 2022), precipitation (Ghazvinian et al., 2021), and solar energy (Horat et al., 2024). In this study, we adapt the DRN for rainfall-runoff modeling using an LSTM as the neural network. The LSTM predicts the parameters of a logistic distribution, namely the location and the scale parameters. The DRN is trained by optimizing the Continuous Ranked Probability Score (CRPS; see Eq. B4) and Appendix C1) of the training data with respect to the estimated predictive distribution. The DRN has the hyperparameters mentioned in Table 2 and the hyperparameter search is described in Appendix A2. We used the scorinrules library (Zanetta and Allen, 2024), as it has a readily available implementation of the CRPS for logistic distribution, compatible with PyTorch.

2.2.4 Bernstein Quantile Network

The BQN is a nonparametric approach that predicts the quantile function of the target. This approach was originally proposed by Bremnes (2020) for wind speed forecasting, and has been extended to other target variables, including solar energy forecasts (Schulz et al., 2021). Its flexibility renders it applicable for runoff predictions as well. The said quantile function is a linear combination of Bernstein basis polynomials (Bremnes, 2020; Schulz and Lerch, 2022) and is given by

$$Q(q|\mathbf{x}) := \sum_{l=0}^{d} \alpha_l(\mathbf{x}) B_{l,d}(q), \quad q \in [0,1], \tag{1}$$

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

where $\alpha_l(\mathbf{x})$ are basis coefficients and are a function of the inputs $\mathbf{x}$, and

$$B_{l,d}(q) = \binom{d}{l} q^l (1-q)^{d-l}, \quad l = 0, 1, \ldots, d, \tag{2}$$

are the Bernstein basis polynomials of the degree $d \in N$ for individual quantiles $q \in [0, 1]$.

For a given degree d, the quantile function is fully defined by $d+1$ basis coefficients. Schulz and Lerch (2022) suggest that the choice of d is critical, with larger d providing for more flexibility but at the cost of larger estimation variance. We adopt a 12^{th} degree Bernstein polynomial following Schulz and Lerch (2022). For the BQN, the head layer maps the hidden states of the base LSTM to 13 (12+1) raw coefficients $\tilde{\alpha}_l$. One important deviation from Schulz and Lerch (2022) that our implementation assumes, is the non-enforcement of strictly positive support for the predicted distribution. Thus, we allow for the first basis coefficient (α_0) to be negative. We maintain the requirement that the rest of the basis coefficients are strictly positive and increasing by implementing a softplus activation and recursively estimating them as follows:

$$\alpha_0 = \tilde{\alpha}_0, \quad \alpha_l = \alpha_{l-1} + \tilde{\alpha}_l, \quad l = 0, 1, \ldots, d, \tag{3}$$

The BQN is trained by optimizing the average quantile loss (QL) over 99 equidistant quantiles. The QL is a consistent scoring function for the corresponding quantile, and the integral of the QL over the unit interval is proportional to the CRPS (Gneiting and Ranjan, 2011). The QL for a quantile q is given by

$$\mathrm{QL}_q(y_{true}, y_{q|x}) = \max(q(y_{true} - y_{q|x}), (1-q)(y_{q|x} - y_{true})) \tag{4}$$

where y_{true} is the true target value, and y_q is the predicted target quantile q obtained from Eq. (1). Table 2 gives the hyperparameters of the BQN and the hyperparameter search is explained in Appendix A2.

2.3 Combining predictive distributions

All results presented in section 3 and subsequent analyses are based on ensemble averages. For the LSTM_{det}, the ensemble average is the average of the predicted discrete discharge of individual member models. For the probabilistic models, we adopt the aggregation process implemented by Schulz and Lerch (2022), which varies depending on the model in question, and refer the reader to Schulz et al. (2024) for a detailed discussion on combining predictive distributions from ensembles of neural network-based models.

For the CMAL, the weights of the three component distributions as well as the location, scale, and asymmetry parameters are averaged over the ensemble members. These are then used to calculate the resulting ensemble-averaged discharge distribution. Similarly, for the DRN, the location and the scale parameters from the ensemble members are averaged to produce the ensmble averaged discharge distribution. For the BQN, the raw coefficients are first converted into Bernstein basis coefficients (according to Eq. 3) for every ensemble member. The basis coefficients are then averaged over the ensemble members and the quantile function is then obtained according to equation 1.

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

2.4 Evaluation of predictive distributions

Gneiting et al. (2007) suggest that the goal of probabilistic forecasting is to maximize the *sharpness*, subject to *calibration*. Although they use different terminology, Klotz et al. (2022) evaluate their models similarly based on *reliability* (*calibration*) and *resolution* (*sharpness*). *Calibration* is a joint property of the predicted distributions and the observed events, and refers to the statistical consistency between the two, whereas *sharpness* refers to the concentration of the predicted distribution, and is a property of the predicted distribution alone. Quantitatively, calibration and sharpness can be simultaneously evaluated with proper scoring rules, where lower values indicate better predictive performance (Gneiting and Raftery, 2007). The CRPS (see Appendix C1 for a detailed description) is a strictly proper scoring rule and is used in this study to evaluate the probabilistic predictions. For the DRN, the CRPS is computed analytically. For the CMAL, the CRPS is estimated using 5,000 Monte Carlo samples. For the BQN, the CRPS is approximated from predicted quantiles using the quantile loss. Torch-compatible implementations of these functions are provided by the Python library scoringrules (Zanetta and Allen, 2024). The CRPS is calculated for every every time step and averaged over the test period for every catchment. We report the mean and the median catchment CRPS scores as a measure of overall model performance. For a visual assessment of the calibration alone, we compute the the Probability integral transform (PIT) values by evaluating the predictive cumulative distributions at the observed values and plot PIT histograms for the entire test period and all catchments combined in Section 3.1. For the BQN, we resort to the unified PIT (uPIT) as only a set of quantiles or the quantile function is available (Vogel et al., 2018). For more details about the PIT histograms, we refer the reader to Appendix C2 where we show the of PIT histograms for calibrated and (typical instances of) miscalibrated probabilistic predictions. We also adopt some metrics from Klotz et al. (2022) to evaluate sharpness of the probabilistic predictions. For every time step in the test period, we calculate the mean absolute deviation, standard deviation and variance from 5000 Monte Carlo samples drawn from the predictive distribution. From the quantiles of the distributions, we calculate the interquartile range and the interdecile range. The metrics are presented as a mean aggregation over all catchments for the entire test period. Finally, evaluate the distributions based on single-point metrics such as the Mean Absolute Error (MAE) evaluated on the median of the distributions and the Root Mean Squared Error (RMSE) evaluated on the mean of the distributions. To align with standard hydrological practice, where overall model performance is primarily assessed on discrete predictions, we also calculate the Nash-Sutcliffe efficiency (NSE) and the Kling-Gupta efficiency (KGE) (following Klotz et al., 2022). For every time step, the mean of the probabilistic distributions is considered as a point estimator to calculate these metrics and we report the aggregated metrics for the entire test period. We acknowledge that the NSE and KGE are not consistent scoring functions (Vrugt, 2024) and as such, should not primarily be used to evaluate probabilistic predictions.

2.5 Flood Risk Communication Agent (*FRiCA*)

The *FRiCA* learns to map the DRN predictive distributions to single quantiles and is based on the REINFORCE (Williams, 1992) policy-gradient algorithm. We chose the DRN in this experiment because it showed the best calibration for the test period among the three probabilistic models. The goal of the agent is to trigger flood warnings by extracting a single actionable quantile from the predictive distribution. The decision-making policy is parameterized by a single-layer LSTM (LSTM_{π_θ})with

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

an input size of 26 (same as the other models in this study), a hidden size of 256, no dropout, and an output dimension equal to the number of available actions. To maintain hydrological context, the inputs to the *FRiCA* are the same as the inputs to the probabilistic models, with a sequence length of 365. The learning rate is set at 0.0005. The action space indicates the quantiles available for the agent to select from, and is a non-uniform set of 32 quantile levels ranging from 0.8 to 0.9999. In other words, at every time step, the agent can choose from 32 quantiles between 0.8 and 0.9999. The reinforcement-learning formulation follows a single-step decision process. At each time step, the LSTM outputs one real-valued score for each possible action. These scores do not yet represent probabilities for the action space, and are converted into a probability distribution by constructing a categorical distribution. For every time step, an action is sampled from this distribution, and the log-probability of the sampled action is used to compute the policy-gradient loss. A pretrained DRN provides the full predictive distribution of discharge for that time step, and the selected ation is the selected quantile from this distribution. This yields a single deterministic discharge estimate for the corresponding time step. This estimate is then compared to the observed discharge for its annual return interval (ARI). The reward function assigns a value of +100 for a *hit*, defined as the case where the ARI category of the predicted discharge matches that of the observation. A *miss*, corresponding to an underestimation of the ARI, is penalized with -5000, while a *false alarm*, corresponding to an overestimation, receives a reward of 0. Following the standard policy-gradient formulation, the loss for each training batch is computed as the product of the reward and the negative log-probability of the selected actions, and training proceeds by minimizing the average batch loss. Reward computation is restricted to time steps for which the 85th quantile of the DRN predictive distribution exceeds the HQ2 threshold (the discharge associated with an ARI of 2 years) for the respective catchment. This masking ensures that the *FRiCA* focuses exclusively on learning quantile-selection behavior for rare events. For all other time steps, the agent is not activated and the mean of the predictive distribution is used instead. During validation and testing, actions are selected by choosing the quantile with the highest policy probability. Model selection is based on validation performance, and the model achieving the highest hit rate is retained. Test performance is reported using contingency tables of flood warnings and is compared against results obtained using deterministic LSTM predictions as well as an arbitrary fixed choice of the 99th quantile of the DRN predictive distributions. For the purpose of evaluation, extreme events in the test period are categorized into *floods* and *extreme floods*. Events with ARIs between 2 and 10 years (HQ2–HQ10) are classified as *floods*, while events with ARIs between 10 and 100 years (HQ10–HQ100) are classified as *extreme floods*. The ARI thresholds (HQ2, HQ10, HQ30, HQ100, and HQ300) used for this categorization are derived from catchment-specific extreme value analyses reports published by the Bundesamt für Umwelt (BAFU, 2024) of the Swiss confederation.

3 Results

In this section, we present the overall evaluation for the probabilistic models based on the CRPS, PIT histograms, resolution metrics and single-point accuracy metrics. Section 3.2 follows with a comparison of the model performance for representative flood events recorded in Switzerland. Finally in section 3.3 we present the results from the *FRiCA*.

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

3.1 Global Model Performance

As mentioned in section 2.4 we calculate the mean CRPS for the entire test period per catchment and begin this section by examining the mean and median CRPS across catchments. The CMAL has the lowest mean CRPS of 0.50 as compared to the DRN (0.53) and the BQN (0.56) and outperforms the two models in terms of overall prediction calibration and sharpness. The difference in performance is further pronounced if we consider the median CRPS across all catchments. For the CMAL the median CRPS is 0.38, whereas the DRN and the BQN predictions have a median CRPS of 0.41 and 0.44 respectively. The superior performance of the CMAL can be attributed to its ability to model asymmetry and multimodality in probabilistic runoff predictions. For all the three models, the mean CRPS across all catchments is bigger than the median CRPS, indicating a right-skewed distribution of per-catchment CRPS for all the models. For a plot of the cumulative density function (CDF) of the catchment-wise CRPS refer Appendix D. Fig. 1 presents the PIT histograms for the (a) CMAL, the (b) DRN and the (c) BQN aggregated over all catchments. Visually, the PIT histogram of the DRN (see panel (b) in Fig. 1), resembles a uniform distribution most closely. We can say that the DRN shows the best over all calibration among the three models, however, the differences are not too large in comparison to the CMAL. The mean absolute percentage deviation from a uniform distribution (red dashed line) for the three models is 25% (CMAL), 18% (DRN) and 31% (BQN). For the CMAL and the DRN, the PIT interval (0.95, 0.99] shows the maximum deviation of 143% and 107% respectively, indicating that the models respetively assign probabilities belonging to (0.95, 0.99] 143% and 107% more frequently than truly observed. This shows that events are being underestimated throughout the range of observed data, and especially at higher flow levels. Extreme values occur less often than expected, pointing to under-dispersion or an insufficient probability mass in the upper tail of the forecast distributions. For the BQN, however, the maximum deviation of 102% is for the PIT interval (0, 0.05], indicating that the BQN assigns 102% more probabilities from this range, than actually observed. The unified PIT (uPIT) is obtained from the quantiles explicitly and is hence influenced by the range of quantiles. About 9% of the test data is lower than the 1st quantile of the BQN predictions, which results in uPIT values of 0. This also represents an overestimation of these discharges, especially low flows, by the BQN. The CMAL and BQN are seen to predict more events within the probability intervals (0.4, 0.6] than actually observed, whereas such a behavior is not observed for the DRN. For the uPIT interval (0.05, 0.3], the BQN predicts these probabilities, on average, 50% less frequently, than actually observed — thereby indicating a significant mis-calibration for the interval. The PIT histograms answer the question: *"how trustworthy is our calibration?"*, and deviations from a uniform distribution indicate that the probability assigned to a certain event is not reliable. In the case of the DRN, the calibration can be said to be mostly biased. In case of the CMAL and the BQN, the calibration is also slightly over-defensive while being biased. For comparison, we also the plot probability plots following Klotz et al. (2022) and present them in Appendix C.

To complement the calibration evaluation of the probabilistic predictions based on the PIT histograms, Table 3 reports the resolution metrics for the probabilistic models. The DRN produces the sharpest predictive distributions, reflected in its consistently lower mean absolute deviation, standard deviation, and variance relative to the other models. While the CMAL exhibits larger overall standard deviation and variance than the DRN, its interquartile and interdecile ranges are smaller. This indicates that, although CMAL generates narrower central distributions than both, the DRN and BQN, it also produces substantially

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

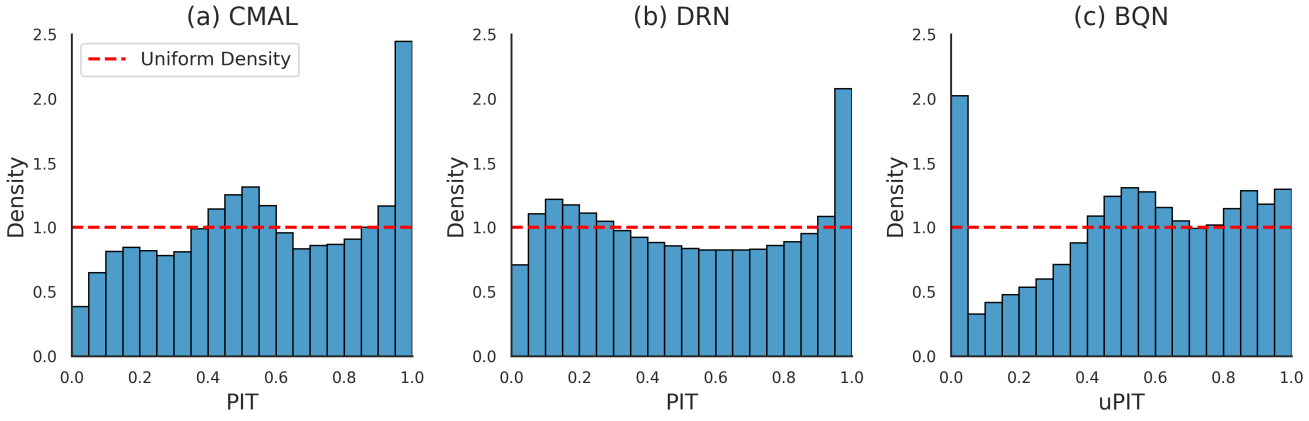

Figure 1. PIT histograms for the (a)CMAL, the (b)DRN and the (c)BQN. The x-axes indicate the PIT values from the entire testing data (a total of 1340737 values) binned into 20 equal-sized bins. The y-axes give the density of these bins. The red dashed line indicates the uniform distribution — a case of perfect calibration. Bins with density less than 1 imply that the model predicts probabilities of occurence within this bin *lesser frequently* than actually observed. Similarly, bins with density more than 1 indicate that the model assigns these probability values *more frequently* than actually observed.

Table 3. Metrics indicating sharpness of the predictive distributions. The mean absolute deviation is the average deviation of 5000 Monte Carlo samples (drawn from the distribution) from the mean of the distribution; the standard deviation and the variance are also calculated for these 5000 samples. The inter quantile and inter decile ranges are calculated from the distribution quantiles directly without sampling. All metrics are reported as averages over the test period and all catchments. All metrics can take values in the interval $(0, \infty)$ and lower values are better for all metrics. The lowest metric per row is indicated in **bold**.

Resolution Metric	DRN	CMAL	BQN
Mean absolute deviation	**0.502**	0.582	0.537
Standard deviation	**0.657**	0.766	0.690
Variance	**1.236**	1.655	1.384
Inter quantile range (distance between 0.25 and 0.75 quantiles)	0.796	**0.642**	0.860
Inter decile range (distance between 0.10 and 0.90 quantiles)	1.592	**1.510**	1.769

heavier and longer tails. As a general interpretation (see Appendix C2), PIT histograms with a central *hump* are over-dispersed or unsharp, while being poorly calibrated. In this regard, the over-dispersed (less sharp) CMAL and BQN distributions show high resolution metrics, especially the mean absolute deviation, the standard deviation and the variance. The fact that the CMAL predictions have the smallest CRPS is further corroborated by the resolution metrics. The CRPS jointly accounts for the calibration and the sharpness (resolution) of the predictive distributions. While the CMAL is seen to calibrated slightly poorly as compared to the DRN (see PIT histograms in (a) and (b) in Fig. 1), the lower interquantile and interdecile ranges indicate sharper distributions, which overcome the poor calibration to produce generally lower CRPS as compared to the DRN.

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

Table 4. Evaluation of different single-point accuracy metrics. The LSTM$_{det}$ predictions are determninistic and are as such single-point. For the probabilistic models, the median of the predictive distributions is used to calculate the MAE, and the mean is considered for calculating the RMSE, the NSE, and the KGE. The metrics are calculated for every time step and averaged over the test period for every catchment. The median catchment metrics are reported here.

Metric	LSTM$_{det}$	DRN	CMAL	BQN
MAE	0.52	0.57	0.55	0.61
RMSE	0.82	0.99	1.01	1.01
NSE	0.87	0.83	0.77	0.82
KGE	0.84	0.78	0.67	0.76

Table 4 gives the single-point metrics for the four models. None of the probabilistic models supersede the LSTM$_{det}$ in any of the accuracy metrics. While the RMSE from the three probabilistic models are very similar, a bigger difference in performance is seen in the MAE. The CMAL, which has the lowest overall CRPS, also has the lowest MAE of 0.55. However, the DRN and the BQN have better performance than the CMAL predictions in terms of the NSE and the KGE. This contrasts with the results of Klotz et al. (2022) on CAMELS-US where the CMAL model reached a similar NSE as the deterministic LSTM.

One of objectives of this study is to determine the advantage of probabilistic predictions for tail regions of the discharge distribution. To explicitly examine model performance in the extremes, Fig. 2 shows the top 0.1% of the sorted discharge values from the test period (1233 discharge values). The maximum observed discharge during the testing period is about 180 mm d^{-1} (the largest flood in CAMELS-CH). The LSTM$_{det}$ simulations underestimate about 94% of these values, with a mean underestimation of about 21 mm d^{-1}. The figure also shows the DRN predictive distributions for every observation, and about 67% of the observed values are within the 99th quantile of their respctive predictive distributions. This underscores the potential advantage of probabilistic approaches for flood modelling. They can represent a broader range of possible outcomes and assign non-zero probability to extreme floods that deterministic predictions tend to underestimate.

3.2 Probabilistic predictions for representative floods in Switzerland

Figure 3 shows three representative flood events together with their corresponding predictive distributions and the LSTM$_{det}$ predictions. The selected events span different seasons and were recorded at diverse locations across Switzerland.

Panels (a.1) to (a.3) in Fig. 3 show a summer flood recorded at the Seedorf gauge on the river Reuss. The associated catchment is a high altitude alpine catchment located close to Luzern. The average catchment elevation is about 2000 m above sea level and its area is about 800 km^2, out of which 6.4% is glaciated. The event shown here recorded a total rainfall of about 132 mm during the period from 09$^{\text{th}}$ to 12$^{\text{th}}$ June 2019. On the 10$^{\text{th}}$ of June, a total rainfall of 102 mm was recorded, which represents a precipitation event with an annual return interval of 30 years (MeteoSwiss, 2022). The resulting flood had a peak specific discharge of 32 mm d^{-}1 recorded on the 11$^{\text{th}}$ of June. The LSTM$_{\text{d}}$et prediction follows the rising limb of the observed hydrograph quite well, until it underestimate this peak flow by 6 mm, and this bias is carried forward throughout the recession. The CMAL prediction shows large uncertainty for the discharge values on the 10th and the 11th of June, and

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

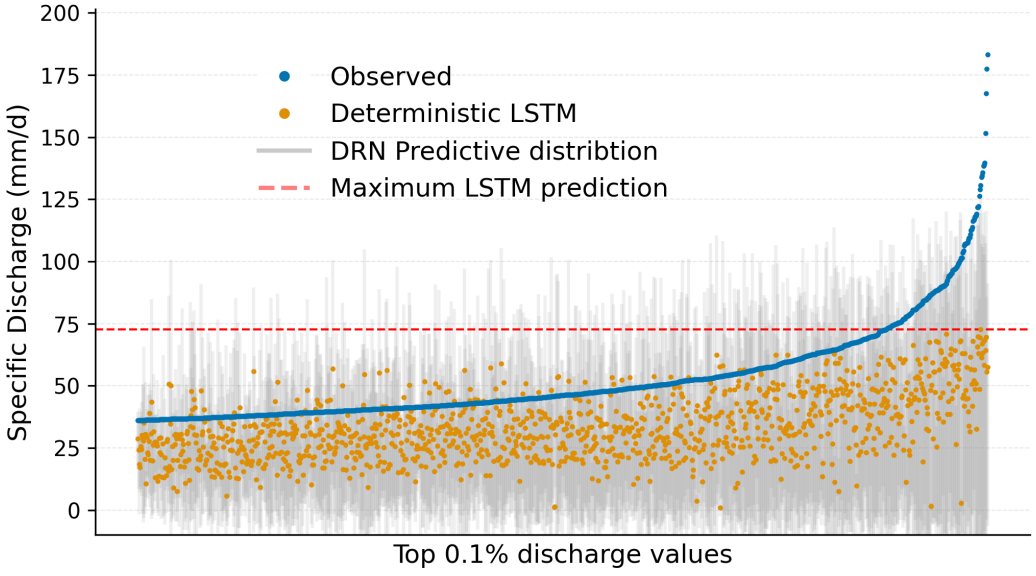

Figure 2. Observed discharge values from top 0.1% tail (1233 discharge values) of the testing distribution, sorted in an increasing order (blue dots). The corresponding LSTM$_{det}$ predictions are shown by orange dots. The maximum LSTM$_{det}$ prediction of 72 mm d^{-1} is given by the red dashed line. The DRN predictive distribution between the 1st and the 99th quantile is a vertical gray line for every observed value.

the distribution is narrower for the recession. The most likely values (dark green region of the predictive distribution) show a greater underestimation than the discrete LSTM$_{\mathrm{det}}$. The DRN predictions for the flood period are narrower than those from the CMAL. The median values from these predictions are comparable to the LSTM$_{\mathrm{det}}$ predictions. The 96th quantile predicts the peak flow with the least error and overcomes the underestimation of the discrete LSTM$_{\mathrm{det}}$. The panel (a.2) also shows the prediction from the *FRiCA*. The *FRiCA* is only activated for the flood events, and hence for the rising limb and the recession the *FRiCA* hydrograph is the mean of the predictive distribution. The *FRiCA* selects the 95th quantile for 11$^{\mathrm{th}}$ June, which predicts the peak flow with an underestimation of less than 1 mm. The BQN prediction does not capture the flood hydrograph, and the 99th quantile underestimates the peak flow by 2 mm.

The second flood shown in panels (b.1) to (b.3) in Fig. 3 is a winter flood recorded at the Dardagny, Les Granges gauge on the river Allondon. The catchment is a small, low altitude catchment with an area of 119 km^2 and mean elevation of about 760 m above sea level. It is located in south-west Switzerland near Geneve. From the 10$^{\mathrm{th}}$ to the 14$^{\mathrm{th}}$ of January 2004, a total of 141 mm rainfall was recorded at the gauge. Rainfall accumulation on the 12$^{\mathrm{th}}$ (49 mm) and the 13th (56 mm) of January was close to the 2-year return precipitation in the area (MeteoSwiss, 2022). The peak discharge for the resulting flood was about 51 mm d^{-1}. This is almost comparable to a flood estimated to occur once in 10 years (HQ10) at this gauge. The LSTM$_{\mathrm{det}}$ prediction for the peak discharge is 37 mm d^{-1}. The CMAL prediction fails to capture the observed hydrograph within the 99th quantile, and the median values show a large bias for the entire flood duration. The DRN and the BQN predictions, assign

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

a finite probability of occurrence to the flood. On 13$^{\text{th}}$ January, the flood peak is best predicted by the 96th quantile from the DRN predictions and the 93rd quantile of the BQN predictions, with almost no bias. While the *FRiCA* hydrograph predicts the recession (~3 mm overestimation on 14$^{\text{th}}$ January) and the peak (1.53 mm overestimation) quite well, it overestimates the total volume on the 12$^{\text{th}}$ of January by about 20 mm.

The third flood shown in Fig. 3 (panels (c.1) to (c.3)) was recorded on the Brig gauge on the Saltina river. The associated catchment is a small alpine catchment and is located in southern Switzerland. During the period from 12$^{\text{th}}$ to 15$^{\text{th}}$ October, a total rainfall of 435 mm d^{-1} was recorded at the gauge Brig. The rainfall accumulation on the 13$^{\text{th}}$ and the 14$^{\text{th}}$ was more than that estimated to occur once in 100 years (MeteoSwiss, 2022). It rained about 160 mm on the two days, which indicates an extreme precipitation for two consecutive days. The resulting flood was a HQ30 flood, meaning that the flood peak value is estimated to occur once in 30 years and is about 87 mm d^{-1}. The LSTM$_{det}$ peak prediction of 50 mm largely underestimates the observed flood. Only the DRN predictions capture the observed hydrograph, with the 99$^{\text{th}}$ quantile overestimating the flood peak by 2 mm. The BQN predictions show a consistent underestimation for the 14$^{\text{th}}$ and 15$^{\text{th}}$ leading to an underestimation of total flood volume and the 99th quantile misses the flood peak by 7 mm. The *FRiCA* is unable to select the closest DRN quantile and predicts the best quantile as the 95$^{\text{th}}$ quantile for the flood peak, which is still an improvement over the deterministic model by 24 mm. The *FRiCA* is specifically trained to predict the flood category, which justifies this behavoir since selecting the 95$^{\text{th}}$ quantile by the *FRiCA* already triggers a warning for *extreme flood* for the catchment. On the 14$^{\text{th}}$ of October, as well, the *FriCA* selects the 95$^{\text{th}}$ quantile to produce a better fitted rising limb of the flood hydrograph. On the recession, however, the *FRiCA* overestimates the discharge on the 16$^{\text{th}}$ of October by about 12 mm. We see an improved performance from the *FRiCA* for the flood peak estimation, but it fails to capture the flood volume accurately.

In summary, Fig. 3 highlights the performance gain from probabilistic predictions for high-stake rare events. The panels (a.1), (b.1) and (c.1) exemplify the resolution metrics for the CMAL predictions reported in Table 4 — though, the CMAL predictions have smaller inter quartile and inter decile ranges, presence of long and heavy tails account for the large standard deviation and variance values. On the days of and leading up to the flood peak, the CMAL predictions are wide, with probability masses being attributed to negative discharge values. For instance, for the flood in Allondon and Saltina, the 1st quantile of the CMAL distributions is -19 mm and -36 mm respectively. Overall, for the entire test period, the lowest 1st quantile across the entire test data from the CMAL, the DRN and the BQN predictions is -95 mm d^{-1}, -20 mm d^{-1}, and -0.6 mm d^{-1} respectively. Negative discharge values are non-physical, and we do not induce this bias by delibrately choosing distributions that are restricted to positive values (including zero). Owing to symetric distirbutions from the DRN, some distributions extending into the negative space could be valid, especially when the upper tails extend to higher values. Given the ability of the CMAL to model asymmetry and multimodality. long and heavy tails extending into the negative space were contrary to our expectation, but this could possibly be a matter of better calibration. Despite allowing for it, the BQN predictions extend minimally into negative discharge values, which might indicate that the model has learned the inherent non-negativity of specific discharge values.

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

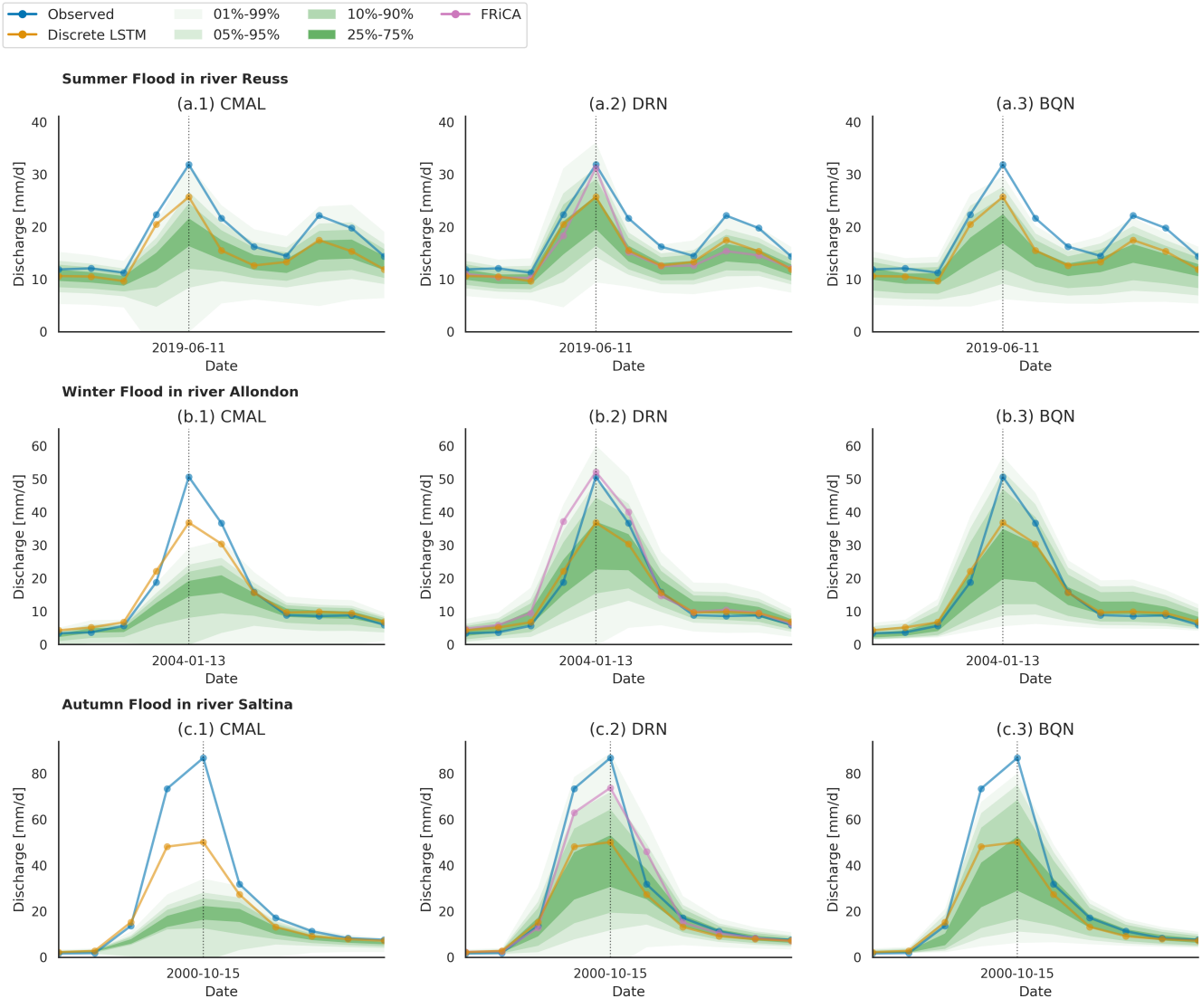

Figure 3. Probabilistic predictions for selected floods recorded in rivers (a.1 to a.3) Reuss in summer, (b.1 to b.3) Allondon in winter, and (c.1 to c.3) Saltina in Autumn. The subplots are in 3X3 grid, with columns representing probabilistic models and rows representing flood events. The x-axis is for time and the discharge is plot on the y-axis. The blue solid line shows the observed hydrograph, and the orange represents the deterministic LSTM_{det} prediction. The predictive distribution is bounded by the 1st and the 99th quantile. For panels (a.2), (b.2) and (c.2) pink solid line indicates the hydrograph simulated from the *FRiCA*.

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

Table 5. Comparison of correct (hits), misses and false warnings for floods and extreme floods given by *FRiCA*. The warnings are compared to those in the case of discrete LSTM_{det} predictions and an arbitrary choice of the 99th quantile of the DRN predictive distributions.

Event Category	Hits	Misses	False Alarms	Total	Method
Floods	605	566	132	1303	FRiCA
	337	941	25	1303	LSTM_{det}
	704	338	261	1303	99th Percentile
Extreme Floods	22	83	34	139	FRiCA
	21	107	11	139	LSTM_{det}
	35	54	50	139	99th Percentile

3.3 *FRiCA* based flood warnings

Table 5 summarizes the contingency statistics (hits, misses, and false alarms) for the *FRiCA*, the LSTM_{det}, and the 99^{th} percentile of the DRN predictions. During the test period, the discharge time series was classified using two ARI-based thresholds, resulting in 1303 identified *flood* and 139 *extreme floods* across all catchments. Because the flood categories are defined using aggregated ARI ranges (BAFU, 2024) over 158 catchments, the corresponding discharge thresholds exhibit substantial variability. For the *flood* category, the minimum lower threshold is 3.4 mm d^{-1} at the Oberkirch gauge on the Suhre, with a corresponding upper threshold of 4.6 mm d^{-1}. The maximum lower and upper thresholds are 288 mm d^{-1} and 448 mm d^{-1}, respectively, observed at the Roveredo–Bacino di Compenso gauge on the Riale di Roggiasca, where no floods occurred during the test period. Overall, approximately 50% of the catchments experienced no *floods* in the testing period, while *extreme floods* were observed in only about 8% of the catchments. For the flood category, the LSTM_{det} achieves a hit rate of only 26%, while producing almost no false alarms (false-alarm rate of 2%). This behavior is consistent with earlier findings indicating that the LSTM_{det} underestimates approximately 94% of events in the upper 0.1% tail of the observed discharge distribution. Using the 99th percentile of the DRN predictive distribution increases the hit rate substantially to 54%, but at the expense of a markedly higher false-alarm rate of 20%. In comparison, the *FRiCA* agent attains a slightly lower hit rate of 46% while simultaneously reducing the false-alarm rate to 10%. Although this reflects a more risk-aware behavior with fewer false alarms, it does not lead to a higher number of correctly identified floods. The *FRiCA* misses 17% more flood events than a strategy that consistently uses the DRN's 99^{th} percentile. The *FRiCA* resorts to the 95^{th} quantile from the DRN distributions in most of the cases. In other words, the policy assigns highest probability to the 95^{th} quantile from the possible action space. This also explains how the number of hits, misses and false-alarms from the *FRiCA* is exactly intermediate to those from the LSTM_{det} and the choice of the 99^{th} quantile. This might be a direct result of the reward design and the fact that the *FRiCA* is not specifically trained to accurately predict the flood magnitudes, as long it predicts the category correctly.

As event severity increases, the performance of deterministic predictions degrades further. For *extreme floods*, the LSTM_{det} misses approximately 77% of events, highlighting its strong tendency to underestimate rare extremes. The *FRiCA* exhibits a

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

low hit rate of 16%, but achieves a 17% reduction in miss rate relative to the LSTM_{det}. The DRN's 99^{th} quantile yields the highest hit rate for *extreme floods* at 25%, albeit with a higher false-alarm rate. In contrast, *FRiCA* produces 12% fewer false alarms than the fixed 99^{th}-quantile strategy, indicating a more conservative warning behavior for the most extreme events.

4 Discussion

We frame the discussion around the three research questions stated in the introduction to this paper (see Section 1).

1. Which of the investigated probabilistic models provides the most reliable rainfall–runoff simulation?

None of the three probabilistic models shows perfect calibration over the test period as can be seen from the PIT histograms in Fig. 1). Overall, the DRN exhibits the best calibration among the three probabilistic models — although the differences from the other two models are modest. The slight advantage of the DRN may be attributed to the use of a single logistic distribution, the CRPS loss function, or the interaction of both components. In contrast, the overconfidence of CMAL is consistent with well-known limitations of likelihood-based training objectives (NLL), which tend to produce overly sharp predictive distributions (Klotz et al., 2022). Whether the DRN's improvemed calibration arises solely from the CRPS loss cannot be conclusively determined in our experimental setup. However, our findings are in line with Schulz and Lerch (2022), who systematically compared probabilistic machine learning approaches for wind gust forecasts in Germany and showed that the DRN had the best performance for 55% stations, followed by the BQN (37% stations). The BQN offers flexibility in terms of the predictive distribution being non-parametric, but the very fact can be disadvantageous as certain properties of a distribution such as bimodality can only be visible implicitly. It is noteworthy that CMAL's original performance advantage in Klotz et al. (2022) relied partly on the ability of the asymmetric Laplace distribution mixture to capture hydrologically meaningful distributional skewness. The fact that our experiments are conducted on CAMELS-CH rather than CAMELS-US may therefore contribute to the comparatively weaker performance of CMAL, given the pronounced differences in hydrological regimes, runoff-generation mechanisms and the quality of the data between the two datasets. Another important deviation from Klotz et al. (2022) is that we assess the calibration for the entire distribution from the CMAL, whereas Klotz et al. (2022) sample 7500 samples from the distribution per time step and truncate the sample by setting negative values to zero. In our experiments, evaluating CMAL calibration using such a truncating sampling strategy led to substantial changes in the PIT histograms: the mean deviation from a uniform distribution decreased to 18.45%, and the deviation in the highest PIT bin (0..95, 0.99] reduced from 143% to 39%. This improvement in calibration would likely also translate into better performance of CMAL during the analyzed flood events. This sampling-based post-processing is computationally expensive, and the resulting performance gains would likely render CMAL broadly comparable to, rather than clearly superior to, the DRN.

All models perform worse than the deterministic LSTM_{det} when compared in terms of the NSE. This finding contrasts with Klotz et al. (2022), who reported comparable performance between CMAL and the state-of-the-art deterministic LSTM for CAMELS-US. We acknowledge that the NSE is not a consistent scoring function (Vrugt, 2024) and should not be used to evaluate the performance of probabilistic models, whose main objective is to represent predictive uncertainty. To assess

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

whether these differences could be attributed to model configuration, we conducted a targeted hyperparameter search (see Appendix A2), but did not observe improvements in performance on CAMELS-CH. Given the large and mostly unexplored hyperparameter space, further tuning may nevertheless improve model performance in terms of both calibration and accuracy. We found that particularly the regularization by adding noise to the training data (inputs and target) influences the overall performance of CMAL for CAMELS-CH.

To our best knowledge, this study is the first to use CRPS for training a DL model in a hydrology context. In meteorological modeling, however, CRPS is common for model training (Gneiting et al., 2005; Vannitsem et al., 2021). During training DRN, we observed that the use of CRPS as a loss function can often lead to prioritizing sharpness of the predictions while compromising calibration. The preference of CRPS for sharp distributions is documented empirically and theoretically by Buchweitz et al. (2025), but in a correcly specified model, the CRPS-based parameter estimation will be consistent (as will NLL-based estimation), and the resulting distribution will be calibrated. While the BQN performed slightly worse than the DRN, it nevertheless shows considerable potential due to its flexibility and and its lack of reliance on a predefined distribution. We hypothesize that further performance gains can be achieved through a more exhaustive hyperparameter search, but we view that as outside the scope of this study. While using the DRN with a single parametric distribution, our assumption of the underlying distribution is crucial. It would be valuable to explore how other heavy-tailed distributions perform for this task. For example, generalized extreme value distributions have been successfully applied in the weather forecasting literature (Lerch and Thorarinsdottir, 2013; Scheuerer and Hamill, 2015). The use of proper scoring rules as objective functions also presents a promising avenue for future research. And, in theory, it is possible to train a single network with multiple heads and losses so that its internal representation considers multiple considerations of the aleatoric uncertainty (Buchweitz et al., 2025). We leave it for future work to examine whether such an approach would lead to synergistic effects or has to be treated as a multi-objective problem.

2. Why are probabilistic models more suitable to capture hydrological extremes and the tail regions of the discharge distribution?

In principle, the rainfall–runoff transformation is a deterministic process. If all relevant system states were fully observed, the governing dynamics perfectly known, and both inputs and targets free of measurement errors, a probabilistic formulation would indeed be unnecessary. In practice, however, none of these conditions are met. Hydrological modeling relies on spatially aggregated forcing data and discharge observations that are affected by measurement uncertainty, representativeness errors, and unresolved sub-grid-scale processes. As a result, the conditional relationship between the meteorological forcing and runoff is inherently uncertain and often highly skewed. This can be illustrated by a simple example from CAMELS-CH: during the training period, precipitation events of approximately 100 mm d^{-1} occurred 27 times. However, floods exceeding 40 mm d^{-1} were observed in only 6 of these cases. On the other hand, there are floods that are triggered without extreme precipitation. This is a highly simplified view of rainfall–runoff modeling, as other metreorological forcing as well as catchment conditions matter crucially. Nevertheless, it illustrates that runoff generation at the catchment scale in large-sample datasets is not unique. The resulting conditional discharge distribution during flood conditions is therefore strongly right-skewed, with a few extreme responses forming the tail. As a consequence, deterministic runoff simulations will underestimate these events if they try to

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

capture only a functional of the distribution such as the mean or the median. In other words, by collapsing the conditional target distribution to a functional, deterministic models effectively disregard rare but plausible responses for statistical reasons alone. In the example above, the deterministic LSTM accurately represents the bulk of observed responses but systematically underestimates the few high-discharge realizations that constitute the tail of the distribution. In contrast, the probabilistic DRN explicitly represents the full conditional distribution and assigns a finite probability of occurrence to these rare events, placing them within its upper predictive quantiles (up to the 99^{th} percentile).

The three flood events shown in Fig. 3 illustrate how these conceptual differences manifest in practice. For the summer flood at Seedorf (panels a.1 to a.3), the deterministic LSTM reproduces the overall hydrograph shape but underestimates the peak discharge, whereas the DRN (panel a.2) captures the event within its upper predictive quantiles. A comparable behavior is observed for the winter flood at Dardagny (panels b.1 to b.3): while deterministic and some probabilistic models fail to encompass the observed peak, the DRN and BQN assign non-negligible probability mass to the flood response. The highly extreme alpine flood at Brig (panels c.1 to c.3) further accentuates this contrast, with the deterministic LSTM severely underestimating the peak and the DRN again placing the observation within its upper tail. Across all cases, probabilistic models do not eliminate all errors, particularly for the most extreme peaks or total volumes, but they consistently broaden the range of plausible responses in a physically meaningful way. These results underpin that the principal benefit of probabilistic formulations lies not in improving median predictions, but more in providing a structured and interpretable representation of uncertainty under extreme forcing conditions.

The examples from Fig. 3 demonstrate how probabilistic models are advantageous in simulating hydrographs for extreme events. Importantly, this behavior is not limited to a few selected cases. As seen in Fig. 2, only about 67% of the highest discharge values (top 0.1% of the sorted values) are assigned a finite probability of occurence. This indicates that the challenge posed by rare events is not entirely overcome by probabilistic approaches alone. Importantly, part of this limitation is likely attributable not only to model structure and training objectives, but also to errors and uncertainties in the input and target data, especially in precipitation estimates. Improvements in forcing data quality are therefore expected to translate directly into more reliable probabilistic predictions, particularly in the upper tail.

3. Can reinforcement learning aid operational flood-warning communication by learning a risk-aware mapping from predictive distributions to actionable warnings based on flood-intensity categories?

The *FRiCA* results obtained in this study are promising, and while they demonstrate its potential as a novel decision layer within early flood warning systems, they also open directions for further improvement. Firstly, this study did not pursue an exhaustive hyperparameter optimization to identify an optimal policy network configuration for the *FRiCA*. Instead, a limited and pragmatic set of hyperparameters was adopted to demonstrate feasibility. Similarly, the reward function was defined in a largely heuristic manner. Our experiments indicate that the learning behavior of *FRiCA* is highly sensitive to the reward design, which is a well-known challenge in reinforcement learning and on purpose a subjective choice of the decision maker and the risk one wants to take. In particular, small changes in the magnitude of negative rewards substantially altered the learned policy, tipping the model toward aggressively minimizing either the miss rate or the false-alarm rate. In this work, we deliberately pri-

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

oritized the reduction of missed warnings by penalizing misses disproportionately more than rewarding correct hits, reflecting the higher societal cost typically associated with undetected extreme flood events. We further observed that the sparsity of the reward signal has a pronounced influence on training stability and convergence. The chosen action space and activation criterion appear well aligned with the operational objective of flood warning systems. However, because these high-impact events were not fully isolated from the remainder of the time series, the resulting reward signal remained sparse. As a consequence, approximately 67% of the training batches yielded zero reward and therefore did not meaningfully contribute to policy updates. This sparsity amplified the inherent stochasticity and noise associated with policy-gradient methods, complicating the learning process. Despite these challenges, the policy network learned to assign a non-uniform probability distribution over the action space. This indicates that it did extract meaningful information from the predictive distributions. Nevertheless, for the majority of events, the most probable action converged to a single choice, namely the 95th percentile. While this behavior is consistent with the asymmetric reward structure favoring the avoidance of misses, it also suggests that the current formulation lacks sufficient incentives for more adaptive, event-specific decision-making. Future work should therefore focus on a more systematic exploration of the hyperparameter space, refinements of the action space, and the development of reward functions that explicitly balance the trade-off between maximizing the hit rate and minimizing false alarms, ideally driving the latter toward zero without sacrificing detection performance. Finally, within the present *FRiCA* scheme, the primary objective is to correctly identify the intensity class of flood events rather than to accurately reproduce the exact discharge values. To address this, we incorporated an additional value-based reward component, formulated in terms of a mean squared error. This led to marginal improvements in overall performance of the *FRiCA*, but rather made the reward signal more noisy. This suggests that the reward design based on classification-oriented rewards and value-based rewards requires further investigation. Ultimately, one would expect these improvements to yield a more diverse selection of optimal quantiles and a near-perfect hit rate under operational constraints.

5 Conclusion

This study evaluated the performance of three LSTM-based probabilistic models for rainfall–runoff simulation across Switzerland, with a particular focus on their ability to represent uncertainty for extremes in the discharge distribution. Specifically, we investigated to what extent probabilistic predictions provide added value over deterministic simulations, and why such advantages are most pronounced in the tails of the distribution rather than in bulk performance metrics. In addition, we explored a reinforcement learning–based communication method that translates probabilistic runoff forecasts into actionable flood warnings for operational early warning systems. Based on our findings, we draw the following conclusions:

- Probabilistic LSTM models provide generally reliable predictive distributions, although they tend to underestimate the most extreme events. Among the three probabilistic approaches considered, no substantial differences in overall calibration were observed. The CMAL has the best overall CRPS, but the DRN exhibited slightly better calibration as compared to the other two models.

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

- When evaluated using deterministic performance metrics such, none of the probabilistic models matched the performance of the state-of-the-art deterministic LSTM. 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, which is essential for risk-informed decision-making.

- The biggest value addition from the probabilistic models emerges in the upper tail of the discharge distribution. In particular, these models assign finite probabilities to very rare events, corresponding to the top 0.1% of observed discharges. This reflects an improved capacity to capture uncertainty in hydrological processes across various hydrological regimes and climatic conditions and to represent the plausibility of extreme floods.

- The reinforcement learning–based agent demonstrates promising potential as decision-support for early warning systems. By condensing full predictive distributions into actionable warning levels, the agent offers a principled alternative to fixed heuristics. Nevertheless, substantial scope for improvement remains, including systematic hyperparameter optimization of the policy network, refinement of the action space, better reward design, and the integration of categorical rewards with value-based rewards to better balance detection performance and warning accuracy.

Code availability. We shall make all the codes for model training, testing, and analysing and plotting the results presented in this paper available upon publishing.

Data availability. The CAMELS-CH dataset is freely available at https://doi.org/10.5281/zenodo.7784632 (Höge et al., 2023)

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

Appendix A: Description of Inputs and Hyperparameter Search

A1 List of the CAMELS-CH forcing variables and catchment attributes used for training

Table A1 gives the description of the static and dynamic inputs to the LSTM_{det} and the probablistic models in the paper.

Table A1: Dynamic and static inputs used to train the LSTM_{det} and the probablistic models.

CAMELS-CH	Description
Dynamic Inputs	
precipitation (mm d^{-1})	Observed daily summed precipitation[1,2,3]
temperature_min (°C)	Observed daily minimum temperature[1,2,3]
temperature_max (°C)	Observed daily maximum temperature[1,2,3]
rel_sun_dur (%)	Observed daily averaged relative sunshine (solar irradiance $\geq$ 200 W m-2) duration[1,3]
Static Inputs	
area (m^2)	Catchment area
elev_mean (m a.s.l.)	Mean elevation within catchment
slope_mean (°)	Catchment mean slope over all grid cells
sand_perc (%)	Percentage sand
silt_perc (%)	Percentage silt
clay_perc (%)	Percentage clay
porosity (-)	Volumetric porosity
conductivity (cm h^{-1})	Saturated hydraulic conductivity
glac_area (km^2)	Glacier area of Swiss glaciers per catchment
dwood_perc (%)	Percentage of deciduous forest
ewood_perc (%)	Percentage of coniferous forest (evergreen)
crop_perc (%)	Percentage of agriculture
urban_perc (%)	Percentage of urban and settlements
reservoir_cap (ML)	Total storage capacity of reservoirs in megaliters
p_mean (mm d^{-1})	Mean daily precipitation
pet_mean (mm d^{-1})	Mean daily potential evapotranspiration (PET; Penman–Monteith equation without interception correction)
p_seasonality (-)	Seasonality and timing of precipitation (estimated using sine curves to represent the annual temperature and precipitation cycles, positive (negative) values indicate that precipitation peaks in summer (winter), and values close to zero indicate uniform precipitation throughout the year). See Eq. (14) in Woods (2009))

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

CAMELS-CH	Description
Static Inputs	
frac_snow (-)	Fraction of precipitation falling as snow, i.e., while temperature is < 0 °C
high_prec_freq (d yr^{-1})	Frequency of high-precipitation days (≥ 5 times mean daily precipitation)
low_prec_freq (d yr^{-1})	Frequency of dry days (< 1 mm d^{-1})
high_prec_dur (d)	Average duration of high-precipitation events (number of consecutive days ≥ 5 times mean daily precipitation)
low_prec_dur (d)	Average duration of dry periods (number of consecutive days < 1 mm d^{-1} mean daily precipitation)

A2 Hyperparameter Search

In the following, we explain the basis for the hyperparameters (see Table 2) of the models from this study. We did not conduct any hyperparameter search for the $LSTM_{det}$, and adopted the hyperparameters from Baste et al. (2025).

For the CMAL, in the absence of a completely reproducible code from Klotz et al. (2022), we first recreated a pipeline based on Klotz et al. (2022) and the online repository NeuralHydrology (Kratzert et al., 2022) for the CAMELS-US. We were unable to reproduce the reliability plots from Klotz et al. (2022) for the best hyperparameters reported in the study. We then checked the other combinations from their hyperparameter search regime and were able to reproduce almost similar reliability results for the learning rate of 0.0001 and noise regularization of 0.05. The slight deviation might be because we analyze the entire distributions per time step, whereas Klotz et al. (2022) analyze a sample of 7500 random draws from the distribution where negative values are set to zero. We checked the this hyperparameter combination (lr=0.0001, noise=0.05) and the best one reported by Klotz et al. (2022) (lr=0.0005, noise=0.2) for the CAMELS-CH and picked the best out of the two based on the validation period PIT histograms. The learning rate was thus set to 0.0005 and the noise was 0.2.

To maintain consistency within the probabilistic models, we decided for a hidden size of 256 for the DRN. We checked for two learning rates 0.0001 and 0.00005 and selected the best learning rate based on the validation period PIT histograms. A low CRPS can be a result of sharp but uncalibrated predictive distributions. Selecting a model only based on the training and validation period CRPS can lead to the prioritization of sharpness at the cost of poor calibration.

For the BQN, the degree of the Bernstein polynomial is critical, as a higher degree polynomial provides more flexibility, but also increases the variance in the predictions. We performed a hyperparameter search for the degree of the Bernstein polynomial and the learning rate. Due to its computational cost, a full grid search was not feasible, and we first evaluated four degrees of the Bernstein polynomial (8, 12, 16, and 20) and found that the 12^{-1} degree polynomial performed best based on the validation period PIT histograms. Next, we tested learning rates of 0.01, 0.005, 0.0001, and 0.00005, ultimately selecting 0.00005 as the optimal value based on the validation period PIT histograms.

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

Appendix B: Mathematical Basis for the CMAL and the DRN

In this section, we give the mathematical basis for the CMAL and the DRN. We begin with the probability density function for a countable mixture of asymmetric Laplace distributions and the NLL. We then give the closed-form expression for the CRPS for a single logistic distribution. The scoringrules library (Zanetta and Allen, 2024) provides an implementation of the CRPS that is compatible with PyTorch.

The probability density function for a single asymmetric Laplace distribution is

$$\mathcal{A}_{LD}(y|\mu,s,\tau) = \frac{\tau\cdot(1-\tau)}{s}\cdot\begin{cases}\exp[-(y-\mu)\cdot(\tau-1)/s], & \text{if } y<\mu,\\ \exp[-(y-\mu)\cdot\tau/s], & \text{if } y\geq\mu,\end{cases} \tag{B1}$$

For the CMAL the weighted sum of three such distributions is given by

$$p(y|\mathbf{x}) = \left(\sum_{k=1}^{3}\alpha_k(\mathbf{x})\cdot A_{LD}(y|\mu_k(\mathbf{x}),s_k(\mathbf{x}),\tau_k(\mathbf{x}))\right) \tag{B2}$$

where, the mixture weights (a_k), mixture location parameters (μ_k), mixture scale parameters (s_k), and mixture asymmetry parameters (τ_k) are outputs from the LSTM, and are thus a function of the inputs $\mathbf{x}$, at every time step. The CMAL is trained by maximizing the negative log-likelihood of the training data (y_{true}) from an estimated mixture density conditional on the input $\mathcal{A}_{LD}(\mu(\mathbf{x}),s(\mathbf{x}),\tau(\mathbf{x}))$; for three components this yields

$$NLL_{\mathcal{A}_{LD}}(\mu(\mathbf{x}),s(\mathbf{x}),\tau(\mathbf{x}),y_{true}) = -\log\left(\sum_{k=1}^{3}\alpha_k(\mathbf{x})\cdot A_{LD}(y_{true}|\mu_k(\mathbf{x}),s_k(\mathbf{x}),\tau_k(\mathbf{x}))\right) \tag{B3}$$

The DRN is trained by optimizing the CRPS of the training data (y_{true}) for an estimated logistic distribution $\mathcal{L}(\mu(\mathbf{x}),s(\mathbf{x}))$ conditional on the input $\mathbf{x}$. For a single logistic distribution, a closed-form equation for the CRPS exists:

$$\text{CRPS}(Logistic(\mu(\mathbf{x}),s(\mathbf{x})),y_{true}) = s(\mathbf{x})\left(\frac{y_{true}-\mu(\mathbf{x})}{s(\mathbf{x})} - 2\log F\left(\frac{y_{true}-\mu(\mathbf{x})}{s(\mathbf{x})}\right) - 1\right), \tag{B4}$$

where

$$F\left(\frac{y-\mu(\mathbf{x})}{s(\mathbf{x})}\right) = \left(1+\exp\left(-\frac{y-\mu(\mathbf{x})}{s(\mathbf{x})}\right)\right)^{-1} \tag{B5}$$

is the CDF of the logistic distribution. A variety of closed-form analytical expressions of the CRPS for various parametric families of distributions is available in Jordan et al. (2019).

Appendix C: CRPS and PIT Histograms

C1 Proper Scoring Rules and the Continuous Ranked Probability Score (CRPS)

Consider y that can take any value from the true distribution $\mathcal{Y}$. The model's *best judgment* for the distribution of y is G, but the model reports a different distribution F. When y is realized, the (negatively oriented) function $S(F,y)$ imposes a penalty

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

for on the model for reporting the distribution F. The expected penalty for reporting a distribution F, under their best guess G, is given by $\mathbb{E}_{y\sim G}S(F,y)$. The penalty assigning function S is proper, if for every (mis)reported F, the expected penalty is more than the expected penalty if the model were to report its best guess G. In other words, a *proper scoring rule* incentivizes the model to report its best guess or, its true belief. Thus, for a *proper scoring rule*, the inequality in Eq. C1 holds true for all F and G. For a *strictly proper scoring rule*, the equality is only possible if and only is $F = G$.

$$\mathbb{E}_{y\sim G}S(G,y) \leq \mathbb{E}_{y\sim G}S(F,y) \tag{C1}$$

The negative log likelihood and the CRPS are strictly proper scoring rules. The latter is a widely used in atmospheric sciences. The CRPS for a probabilistic forecast given through a cumulative distribution function (CDF) F and an observation y is given by

$$\text{CRPS}(F,y) = \int_{-\infty}^{\infty} [F(z) - \mathbf{1}\{z \geq y\}]^2 \, dz \tag{C2}$$

The CRPS has the same unit as the observation and generalizes to the absolute error in case of a deterministic forecast. The integral can be calculated analytically for multiple distributions, and the closed form solution for the logistic distribution is given in Eq. (B4). The NLL is also a strictly proper scoring rule. Given the non-availability of an analytical formula for CRPS of a CMAL, the NLL (see Eq. (B3)) is a convenient alternative.

C2 Probability Integral Transform (PIT) Histogram

The PIT histogram (Diebold et al., 1998) and the probability plots in Klotz et al. (2022), both give an assessment of the calibration of the predictive distribution. Both the methods involve the calculation of the probability integral transform ($PIT = F(y)$), which is the distribution CDF F evaluated at the observed data y. Diebold et al. (1998) proposed a simple, more revealing graphical representation of the PIT values, in the form of a PIT histogram. For a calibrated prediction, the PIT is uniformly distribution, and a histogram presents a straightforward density estimate. While Laio and Tamea (2007) agree that the PIT histograms are more intuitive, they introduced the probability plot representation, which eliminates subjective binning. Nonetheless, Klotz et al. (2022) preferred discrete steps for the probability plots, instead of continuous ones, in order to prevent falsely reporting overly precise results. Owing to their simplicity and ease of interpretation, we resort to the PIT histograms for visual assessment of the calibration of the probabilistic forecast.

Fig. C1 shows three representative PIT histograms which represent the (a)*overconfident*, (b)*overdefensive* and (c)*biased* predictions respectively. In Fig. C1 panel (a), the predicted distribution assigns more high and low probabilities to observed data, than actually observed. The predicted distribution is sharp, but a false representation of reality, as more observed data falls in the tail regions of such a forecast, than truly observed. The users of such a prediction expose themselves to unwanted surprises and strongly underestimated risks. Such a prediction, albeit sharp, is *overconfident* and not worthy of trust from the decision makers. The prediction in panel (b), on the other hand, is *overdefensive* and plays it safe by assigning the majority of the observed data probabilities closer to 50%. From a hydrological perspective, this implies that, floods which are not to be

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

650 expected as often, are predicted with higher expectation. This calls for more than required disaster preparedness and investment in mitigation efforts. A predicted distribution can be also be biased. Panel (c) in Fig. C1 shows a histogram skewed to the left, which indicates that the predicted distribution assigns lower probabilities to more observed data than the truly observed. The predicted probability densities for such events are shifted to the right, indicating an overestimation of observed data. When the histogram is right-skewed, it implies an underestimation of observed data.

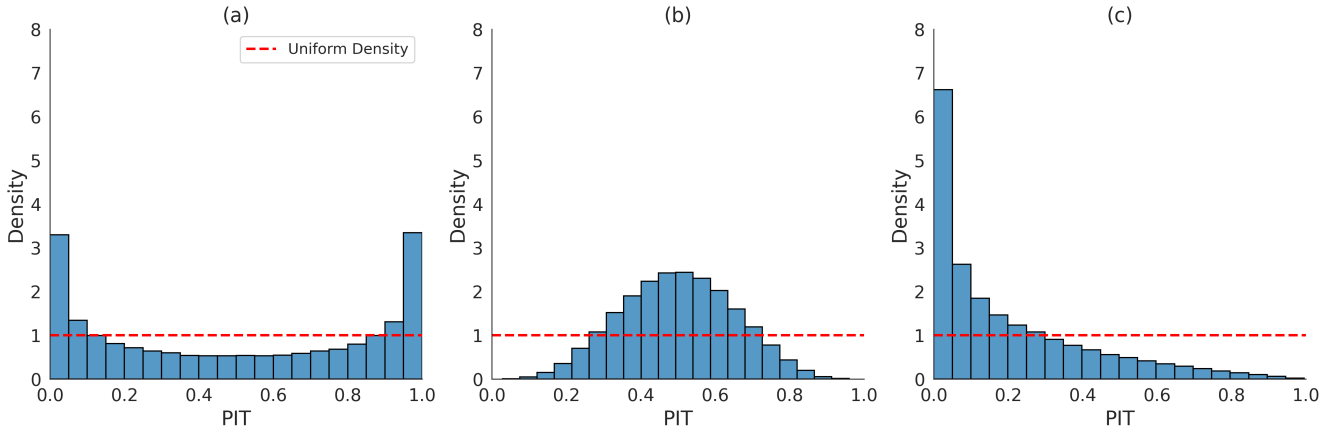

Figure C1. Interpreting probability integral transform (PIT) histograms. A (a) U-shaped PIT histogram is indicative of an overconfident forecast, whereas a (b) bell-shaped histogram indicates an overdefensive forecast. Bias in the forecast is indicated by a histogram with highly populated bins on the either side of the histogram (c).

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

Appendix D: CDF for CRPS and PIT histograms with probability plots

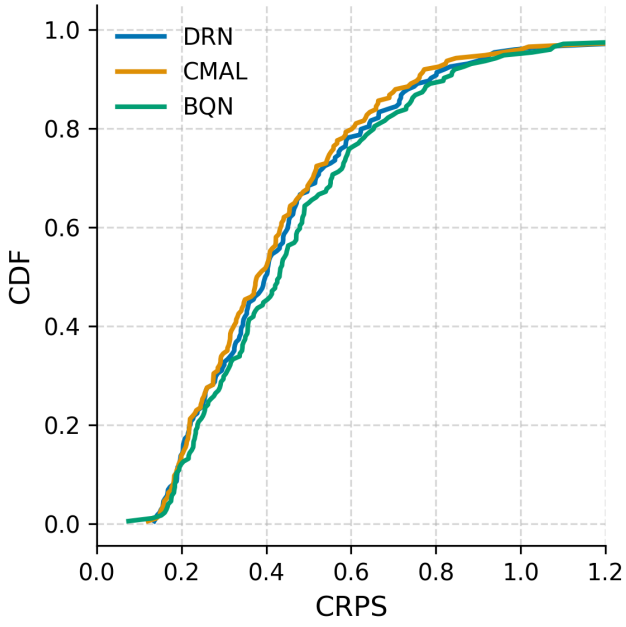

Figure D1. Cumulative density function (CDF) of Continuous ranked probability score (CRPS) for CMAL, DRN and BQN predictions. The CRPS is calculated for every catchment and averaged over the test period. The median CRPS from the three models is 0.38, 0.41 and 0.44 respectively.

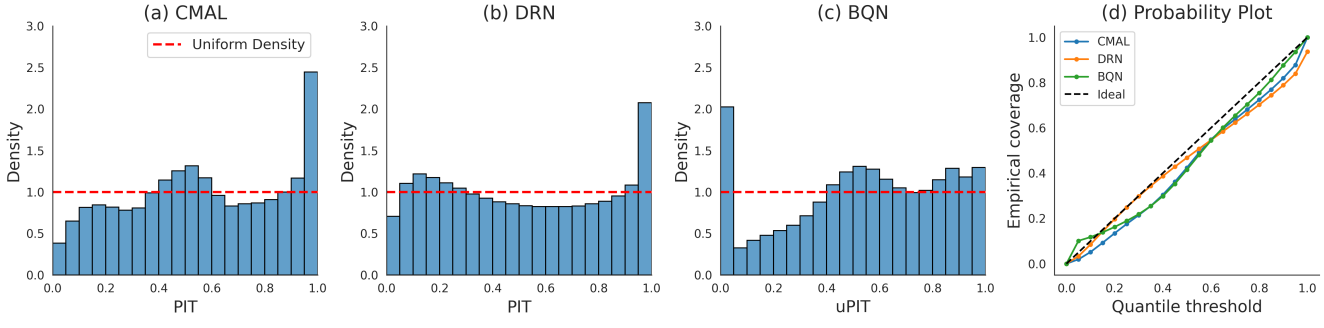

Figure D2. PIT histograms for the (a)CMAL, the (b)DRN and the (c)BQN. The x-axes indicate the PIT values from the entire testing data (a total of 1340737 values) binned into 20 equal-sized bins. The y-axes give the density of these bins. The red dashed line indicates the ideal uniform distribution. Bins with density less than 1 imply that the model predicts a lesser number of these probability values than actually observed. Similarly, bins with density more than 1 indicate that the model assigns these probability values more frequently than actually observed.

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

Disclaimer. We acknowledge the use of ChatGPT (version 5.0) for generating cleaner codes and for help in writing style refinement. The methods, results and analyses are based on the authors' works alone.

Author contributions. The idea for the paper was proposed by RL. Model training and testing, and analyses were done by SB, and results were discussed with RL, SL and DK. The draft was prepared by SB and reviewed and edited by all authors. All authors have read and agreed to the current version of the manuscript.

Competing interests. Some authors are members of the editorial board of HESS.

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

References

Acuna Espinoza, E., Loritz, R., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., Bäuerle, N., and Ehret, U.: Analyzing the generalization capabilities of hybrid hydrological models for extrapolation to extreme events, EGUsphere, 2024, 1–17, https://doi.org/10.5194/egusphere-2024-2147, 2024b.

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

BAFU: Hochwasserwahrscheinlichkeiten (Jahreshochwasser), https://www.bafu.admin.ch/de/hochwasserstatistik#Inhalt-der-Resultatbl%C3%A4tter, 2024.

Baste, S., Klotz, D., Espinoza, E. A., Bardossy, A., and Loritz, R.: Unveiling the Limits of Deep Learning Models in Hydrological Extrapolation Tasks, EGUsphere, 2025, 1–24, https://doi.org/10.5194/egusphere-2025-425, 2025.

Bergström, S.: THE HBV MODEL - its structure and applications, https://www.smhi.se/en/publications/the-hbv-model-its-structure-and-applications-1.83591, 1992.

Beven, K. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, Journal of Hydrology, 2001.

Bremnes, J. B.: Ensemble Postprocessing Using Quantile Function Regression Based on Neural Networks and Bernstein Polynomials, Monthly Weather Review, 148, 403–414, https://doi.org/10.1175/MWR-D-19-0227.1, 2020.

Buchweitz, E., Romano, J. V., and Tibshirani, R. J.: Asymmetric penalties underlie proper loss functions in probabilistic forecasting, arXiv preprint arXiv:2505.00937, 2025.

Bálint, G., Csík, A., Bartha, P., Gauzer, B., and Bonta, I.: APPLICATION OF METEOROLOGICAL ENSEMBLES FOR DANUBE FLOOD FORECASTING AND WARNING, in: Transboundary Floods: Reducing Risks Through Flood Management, edited by Marsalek, J., Stancalie, G., and Balint, G., pp. 57–68, Springer Netherlands, ISBN 978-1-4020-4902-6, 2006.

Chaves, M. A., Espinoza, E. A., Klotz, D., Gupta, H. V., Ehret, U., and Guthke, A.: A variational approach at uncertainty estimation in data-driven rainfall-runoff modeling, EarthArXiv, 2025.

Clark, S. R. et al.: Deep learning for monthly rainfall–runoff modelling, Hydrology and Earth System Sciences, 28, 1191–1210, https://doi.org/10.5194/hess-28-1191-2024, 2024.

Cloke, H. and Pappenberger, F.: Ensemble flood forecasting: A review, 375, 613–626, https://doi.org/10.1016/j.jhydrol.2009.06.005, 2009.

Diebold, F. X., Gunther, T. A., and Tay, A. S.: Evaluating Density Forecasts with Applications to Financial Risk Management, 39, 863, https://doi.org/10.2307/2527342, 1998.

Feng, D., Liu, J., Lawson, K., and Shen, C.: Differentiable, Learnable, Regionalized Process-Based Models With Multiphysical Outputs can Approach State-Of-The-Art Hydrologic Prediction Accuracy, Water Resources Research, 58, e2022WR032 404, https://doi.org/https://doi.org/10.1029/2022WR032404, e2022WR032404 2022WR032404, 2022.

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S.: Deep learning rainfall–runoff predictions of extreme events, Hydrology and Earth System Sciences, 26, 3377–3392, https://doi.org/10.5194/hess-26-3377-2022, 2022.

Frank, C. et al.: Short-term runoff forecasting in an alpine catchment with a Long Short-Term Memory (LSTM) neural network, Frontiers in Water, 5, 1126 310, https://doi.org/10.3389/frwa.2023.1126310, 2023.

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

Ghazvinian, M., Zhang, Y., Seo, D.-J., He, M., and Fernando, N.: A Novel Hybrid Artificial Neural Network - Parametric Scheme for Postprocessing Medium-Range Precipitation Forecasts, Advances in Water Resources, 151, 103 907, https://doi.org/10.1016/j.advwatres.2021.103907, 2021.

Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, 102, 359–378, https://doi.org/10.1198/016214506000001437, 2007.

Gneiting, T. and Ranjan, R.: Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules, Journal of Business & Economic Statistics, 29, 411–422, https://doi.org/10.1198/jbes.2010.08110, 2011.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T.: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation, Monthly Weather Review, 133, 1098–1118, https://doi.org/10.1175/MWR2904.1, 2005.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic Forecasts, Calibration and Sharpness, 69, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x, 2007.

Gouweleeuw, B. T., Thielen, J., Franchello, G., De Roo, A. P. J., and Buizza, R.: Flood forecasting using medium-range probabilistic weather prediction, 9, 365–380, https://doi.org/10.5194/hess-9-365-2005, 2005.

Heudorfer, B., Gupta, H. V., and Loritz, R.: Are Deep Learning Models in Hydrology Entity Aware?, 52, e2024GL113 036, https://doi.org/10.1029/2024GL113036, 2025.

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Computation, 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Höge, M., Kauzlaric, M., Siber, R., Schönenberger, U., Horton, P., Schwanbeck, J., Floriancic, M. G., Viviroli, D., Wilhelm, S., Sikorska-Senoner, A. E., Addor, N., Brunner, M., Pool, S., Zappa, M., and Fenicia, F.: CAMELS-CH: hydro-meteorological time series and landscape attributes for 331 catchments in hydrologic Switzerland, Earth System Science Data, 15, 5755–5784, https://doi.org/10.5194/essd-15-5755-2023, 2023.

Horat, N., Klerings, S., and Lerch, S.: Improving Model Chain Approaches for Probabilistic Solar Energy Forecasting through Post-processing and Machine Learning, Advances in Atmospheric Sciences, https://doi.org/10.1007/s00376-024-4219-2, 2024.

Jordan, A., Krüger, F., and Lerch, S.: Evaluating Probabilistic Forecasts with scoringRules, Journal of Statistical Software, 90, 1–37, https://doi.org/10.18637/jss.v090.i12, 2019.

Kapoor, N., Perrin, C., and Andréassian, V.: DeepGR4J: A hybrid deep learning and conceptual hydrological model, Journal of Hydrology, 624, 130 090, 2023.

Khoshkalam, K., Hsu, K., et al.: Transfer learning for regional rainfall–runoff modeling using LSTMs, Hydrology and Earth System Sciences, 27, 1901–1920, 2023.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, https://arxiv.org/abs/1412.6980, 2017.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G.: Uncertainty estimation with deep learning for rainfall–runoff modeling, Hydrology and Earth System Sciences, 26, 1673–1693, https://doi.org/10.5194/hess-26-1673-2022, 2022.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrology and Earth System Sciences, 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019a.

Kratzert, F., Gauch, M., Nearing, G., and Klotz, D.: NeuralHydrology — A Python library for Deep Learning research in hydrology, Journal of Open Source Software, 7, 4050, https://doi.org/10.21105/joss.04050, 2022.

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin, Hydrology and Earth System Sciences, 28, 4187–4201, https://doi.org/10.5194/hess-28-4187-2024, 2024.

Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, 2007.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., and Dadson, S. J.: Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models, Hydrology and Earth System Sciences, 25, 5517–5534, https://doi.org/10.5194/hess-25-5517-2021, 2021.

Lerch, S. and Thorarinsdottir, T. L.: Comparison of Non-Homogeneous Regression Models for Probabilistic Wind Speed Forecasting, Tellus A: Dynamic Meteorology and Oceanography, 65, 21 206, https://doi.org/10.3402/tellusa.v65i0.21206, 2013.

Li, X., Liu, Y., et al.: Improving LSTM hydrological modeling with multi-task and spatiotemporal deep learning, Journal of Hydrology, 617, 128 836, 2023.

Loritz, R., Dolich, A., Acuña Espinoza, E., Ebeling, P., Guse, B., Götte, J., Hassler, S. K., Hauffe, C., Heidbüchel, I., Kiesel, J., Mälicke, M., Müller-Thomy, H., Stölzle, M., and Tarasova, L.: CAMELS-DE: hydro-meteorological time series and attributes for 1555 catchments in Germany, Earth System Science Data Discussions, 2024, 1–30, https://doi.org/10.5194/essd-2024-318, 2024.

Ma, X. et al.: Improving predictions in data-poor basins using LSTM-based transfer learning, Journal of Hydrology, 632, 130 600, 2024.

MeteoSwiss: Extreme Value Analyses, version 2022, https://www.meteoswiss.admin.ch/services-and-publications/applications/standard-period.html, 2022.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems 32, pp. 8024–8035, Curran Associates, Inc., http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf, 2019.

Patra, S. et al.: Spatiotemporal deep learning for rainfall–runoff prediction using a ConvLSTM architecture, Journal of Hydrology, 630, 130 496, 2024.

Rasp, S. and Lerch, S.: Neural Networks for Postprocessing Ensemble Weather Forecasts, Monthly Weather Review, 146, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1, 2018.

Roulin, E.: Skill and relative economic value of medium-range hydrological ensemble predictions, 2007.

Roulin, E. and Vannitsem, S.: Skill of Medium-Range Hydrological Ensemble Predictions, 6, 729–744, https://doi.org/10.1175/JHM436.1, 2005.

Scheuerer, M. and Hamill, T. M.: Statistical Postprocessing of Ensemble Precipitation Forecasts by Fitting Censored, Shifted Gamma Distributions*, Monthly Weather Review, 143, 4578–4596, https://doi.org/10.1175/mwr-d-15-0061.1, 2015.

Schulz, B. and Lerch, S.: Machine Learning Methods for Postprocessing Ensemble Forecasts of Wind Gusts: A Systematic Comparison, Monthly Weather Review, 150, 235–257, https://doi.org/10.1175/MWR-D-21-0150.1, 2022.

Schulz, B., El Ayari, M., Lerch, S., and Baran, S.: Post-Processing Numerical Weather Prediction Ensembles for Probabilistic Solar Irradiance Forecasting, Solar Energy, 220, 1016–1031, https://doi.org/10.1016/j.solener.2021.03.023, 2021.

Schulz, B., Köhler, L., and Lerch, S.: Aggregating Distribution Forecasts from Deep Ensembles, https://doi.org/10.48550/arXiv.2204.02291, 2024.

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Bouallègue, Z. B., Bhend, J., Dabernig, M., Cruz, L. D., Hieta, L., Mestre, O., Moret, L., Plenković, I. O., Schmeits, M., Taillardat, M., den Bergh,

https://doi.org/10.5194/egusphere-2026-469
Preprint. Discussion started: 18 February 2026

J. V., Schaeybroeck, B. V., Whan, K., and Ylhaisi, J.: Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World, Bulletin of the American Meteorological Society, https://doi.org/10.1175/BAMS-D-19-0308.1, 2021.

Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A., and Gneiting, T.: Skill of Global Raw and Postprocessed Ensemble Predictions of Rainfall over Northern Tropical Africa, Weather and Forecasting, 33, 369–388, https://doi.org/10.1175/WAF-D-17-0127.1, 2018.

Vrugt, J. A.: Distribution-Based Model Evaluation and Diagnostics: Elicitability, Propriety, and Scoring Rules for Hydrograph Functionals, 60, e2023WR036 710, https://doi.org/10.1029/2023WR036710, 2024.

Williams, R. J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning, Machine Learning, 8, 229–256, https://doi.org/10.1007/BF00992696, 1992.

Zanetta, F. and Allen, S.: scoringrules: a python library for probabilistic forecast evaluation, https://github.com/frazane/scoringrules, 2024.