https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

Unveiling the Limits of Deep Learning Models in Hydrological Extrapolation Tasks

Sanika Baste[1], Daniel Klotz[2, 3], Eduardo Acuña Espinoza[1], Andras Bardossy[4], and Ralf Loritz[1]

[1]Institute of Water and Environment, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[2]Interdisciplinary Transformation University Austria, Linz, Austria
[3]Google Research, Vienna, Austria
[4]Institut für Wasser- und Umweltsystemmodellierung, Universität Stuttgart, Stuttgart, Germany

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

Abstract. Long Short-Term Memory (LSTM) networks have shown strong performance in rainfall–runoff modelling, often surpassing conventional hydrological models in benchmark studies. However, recent studies raise questions about their ability to extrapolate, particularly under extreme conditions that exceed the range of their training data. This study examines the performance of a stand-alone LSTM trained on 196 catchments in Switzerland when subjected to synthetic design precipitation events of increasing intensity and varying duration. The model's response is compared to that of a hybrid model and evaluated against hydrological process understanding. Our study reiterates that the stand-alone LSTM is not capable of predicting discharge values above a theoretical limit, and we show that this limit (73 mm d^{-1}) is below the range of the data the model was trained on (183 mm d^{-1} when trained on CAMELS-CH). Furthermore, the LSTM exhibits a concave runoff response under extreme precipitation, indicating that event runoff coefficients decrease with increasing design precipitation—a phenomenon not observed in the hybrid model used as a benchmark. We show that saturation of the LSTM cell states alone does not fully account for this characteristic behavior, as the LSTM does not reach full saturation, particularly for the 1-day events. Instead, its gating structures prevent new information about the current extreme precipitation from being incorporated into the cell states. Adjusting the LSTM architecture, for instance, by increasing the number of hidden states, and/or using a larger, more diverse training dataset can help mitigate the problem. However, these adjustments do not guarantee improved extrapolation performance, and the LSTM continues to predict values below the range of the training data or show unfeasible runoff responses during the 1-day design experiments. Despite these shortcomings, our findings highlight the inherent potential of stand-alone LSTMs to capture complex hydro-meteorological relationships. We argue that more robust training strategies and model configurations could address the observed limitations, preserving the promise of stand-alone LSTMs for rainfall–runoff modelling.

1 Introduction

Deep learning models, particularly Long Short-Term Memory (LSTM; Hochreiter and Schmidhuber, 1997) networks, have become important tools in rainfall–runoff modelling. The current prototypical setup was introduced by Kratzert et al. (2019a), who trained a single LSTM model for 531 basins across the United States (and achieved superior performance compared to

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

several traditional process-based models). Similar results were confirmed in follow-up work, such as the study by Lees et al. (2021) in Great Britain or Loritz et al. (2024) in Germany. However, as with any model, certain best practices for setting up LSTM-based models are essential to achieve good predictive performance. Among the most important, is training the LSTMs on large, comprehensive, and diverse datasets (Kratzert et al., 2024)—such as Catchment Attributes and Meteorology for Large-sample Studies (CAMELS-US; Addor et al., 2017; Newman et al., 2015).

A behavior that LSTMs exhibit, is that their states can saturate when they ingest new inputs. The mechanism that leads to this behavior is the use of hyperbolic tangent ($\tanh$) and sigmoid activation functions inside LSTM cell. These saturate when the output approaches their asymptotic extremes (Chen and Chang, 1996; Rakitianskaia and Engelbrecht, 2015). Kratzert et al. (2024) identified the saturation of the $\tanh$ function in the computation of the hidden states ($h_t = o_t \odot \tanh(c_t)$, where c_t are the cell states and o_t is the output gate; Appendix D) as a key factor that limits the ability of the LSTMs to predict extreme discharge values. As c_t grows $\tanh$ caps them, restricting the transmission of meaningful information, such as meteorological forcing signals. The severity of this saturation effect depends on the learned weights and biases, and hence on the range and diversity of the training data. In hydrological modelling, the circumstance that model predictions are restricted to the empirical support of the data is unsatisfactory—particularly for the prediction of extremes. This is particularly true in hydrology, where predicting extremes beyond the existing observations is a key modelling aspect. Considering the rapid rise in the application of LSTMs and other deep learning models in rainfall–runoff modelling, we believe that a deeper understanding of their current limitations is essential. This study therefore aims to examine the extrapolation behavior of LSTMs to extreme rainfall–runoff events that lie outside the range of the training data. Albeit the term "extrapolation" is difficult to pinpoint technically—especially in the context of high-dimensional datasets and deep learning models (Balestriero et al., 2021)—the events that we consider in our study are by construction either at the edge of, or outside the range of the observed data (with regard to precipitation).

Previous studies (e.g., Frame et al., 2022; Acuña Espinoza et al., 2024a; Song et al., 2024) have explored the predictive accuracy of LSTMs in extreme runoff scenarios by adopting training/test splits that deliberately exclude certain high-flow values during training. In a stress test setting, Frame et al. (2022) found that, when compared with two conceptual hydrological models, a stand-alone LSTM outperformed one of the former for the most extreme rainfall–runoff events in the CAMELS-US, and was only slightly worse than the second. Acuna Espinoza et al. (2024b) used the same setting to demonstrate that a hybrid model, combining a conceptual hydrological model with an LSTM, was slightly better than a stand-alone LSTM at predicting the most extreme events in the CAMELS-US dataset. In the study, the stand-alone LSTM performed particularly well for the overall evaluation, but for the most extreme events, the LSTM's response showed major deviations from the hybrid model and a conceptual model—exhibiting a distribution of simulated extreme values with no tail (see Figure 5(a) in Acuna Espinoza et al. (2024b)). On the other hand, Song et al. (2024) (in a slightly different setting) found that a hybrid model, similar to the one used in Acuna Espinoza et al. (2024b) outperformed the stand-alone LSTM. The stand-alone LSTM, the mass-conserving LSTM (MC-LSTM in Frame et al., 2022), and hybrid models performed similarly when evaluated using standard metrics;

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

however, the studies provided notably different interpretations regarding whether, and to what extent, LSTMs can successfully extrapolate to extreme events.

Although the stress tests in Frame et al. (2022); Acuna Espinoza et al. (2024b) systematically test the model's ability to handle increasingly extreme events, it is not realistic from a practical perspective. In real-world applications, modellers would not intentionally exclude known extremes from their training datasets, particularly when using data-driven models. In this study, we propose a complementary approach for investigation: Rather than withholding extreme events during training, we force the LSTM with design precipitation values (as commonly used in infrastructure planning and engineering; Global Water Partnership (GWP) and World Meteorological Organization (WMO), 2013). These precipitation values, which are derived using statistical models, can exceed historical observations, but are considered physically possible (World Meteorological Organization (WMO), 1973, 2009). This allows us to probe the model's extrapolation capabilities without imposing artificial constraints on the training data. An intrinsic limitation of our approach is that our augmentation destroys the covariate-structure of the inputs. Hence, in theory, we cannot directly disentangle the effect of the general LSTM out-of-distribution behavior and the one introduced by an actual extreme event of the same kind. This restricts us to a certain coarseness of the analytical depth of our study. However, we argue that the pattern that emerges from our experiments is so clear that it is indicative for the extrapolation behavior of LSTMs in hydrology. Specifically, we compare the LSTM's output with that of a mass-conserving hybrid model (Feng et al., 2022) and assess how both models respond under unprecedented forcing conditions to evaluate the physical realism of the LSTM's predictions.

This study addresses the following research questions:

1. Can LSTMs extrapolate to discharge values beyond the training distribution when forced with statistically derived design precipitation events?
2. Is the saturation of LSTM memory states the primary reason, which limits their ability to extrapolate to extreme and unprecedented hydrological conditions?
3. How do the inherent assumptions and structural characteristics (inductive biases) of LSTMs influence their ability to simulate realistic hydrological responses under conditions that exceed observed training ranges?

The paper is structured as follows: we give a description of the datasets and the models in section 2. This section also details out the set-up for the design precipitation experiments and the methodology for calculating saturation in the LSTM network. This is followed by section 3, where we present the overall model performance and a comparison of model simulations from our design experiments. We discuss the findings and their implications with regard to the three research questions in section 4 and give our conclusion in section 5.

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

2 Data and Methods

In this section, we describe the CAMELS-CH dataset (section 2.1) and the CAMELS-US dataset (section 2.2) used for model training and testing. The subsequent subsections (section 2.3 and section 2.4) briefly describe the LSTM networks, the hybrid model, and their respective model configurations employed in this study. Following these, the section 2.5 details out the selection of catchments and experimental setup for the design precipitation events. Finally, section 2.6 explains how we estimate network saturation in the LSTM.

2.1 The CAMELS-CH Dataset

The CAMELS-CH dataset (Höge et al., 2023) provides daily hydro-meteorological time series data for 331 basins within Switzerland and neighboring countries, along with static catchment attributes which include topographic, climate, hydrology, soil, land cover, geology, glacier, hydrogeology, and human influence attributes. Due to its diverse topography and climate, Switzerland is often referred to as the 'water tower of Europe' (Höge et al., 2023) and despite its small size, it exhibits significant hydrological variability across different regions. CAMELS-CH includes data for 298 river catchments and 33 lakes. The available data spans from 1 January 1981 to 31 December 2020. In this study, we exclude the lakes and 102 river catchments belonging to France, Germany, Austria, and Italy and focus only on the 196 catchments in Switzerland. From this subset, we exclude another four catchments where preliminary model simulations had negative Nash-Sutcliffe efficiency (NSE). We trained an ensemble of 5 LSTMs (see section 2.3) and 5 hybrid models (see section 2.4) for the period from 01.10.1995 to 30.09.2005 (training period; see Table 1). The input for the models consists of 5 dynamic forcing variables and 22 static catchment attributes (see appendix A), and we trained both models to target specific discharge. For the CAMELS-CH dataset, the maximum precipitation during the training period is 234 mm d^{-1} and was recorded for the Krummbach stream located in southern Switzerland. The maximum observed specific discharge is 183 mm d^{-1} which occurred during a flood in the Chli Schliere stream in the Alpnach village in central Switzerland triggered by torrential rains in August 2005 (Federal Department for the Environment and DETEC, 2005).

2.2 The CAMELS-US Dataset

We use a subset of 531 catchments from the CAMELS-US dataset, which was originally identified by Newman et al. (2015). This provides daily meteorological forcing from three data sets, Daymet, Maurer, and NLDAS, and daily stream flow measurements from the United States Geological Survey (USGS) spanning from 1980 to 2015. Catchment topographical characteristics, climate and hydrological indices, and soil, land-cover and geological characteristics are also provided. We use the dataset in combination with the CAMELS-CH dataset to train an ensemble of 5 LSTMs. We use 3 dynamic forcing variables from the Daymet meteorological forcing and 12 static catchment characteristics (see appendix A) as inputs and the daily stream flow data as the target. We use the same training period from 01.10.1995 to 30.09.2005. The maximum observed specific discharge for this training dataset is 299 mm d^{-1}, which is recorded for the Medina river in Texas. The precipitation observed in Krummbach stream (234 mm d^{-1}) in Switzerland is also the maximum precipitation for this combined training dataset.

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

Table 1. Hyperparameters for LSTM network and hybrid model ensemble

Hyperparameter	Value	
	LSTM	Hybrid Model
Number of layers	1	
Number of nodes	64	
Dropout rate	0.4	
Initial forget gate bias	3	
Initial learning rate	0.001	
Sequence length	365	730
Batch size	256	
No. of epochs	20	
Training period	1 October 1995 to 30 September 2005	
Test period	1 October 2010 to 30 September 2015	

2.3 LSTM model

The hyperparameters of our LSTM network (see Table 1) are guided by the work of Lees et al. (2021) and Acuña Espinoza et al. (2024a) and the model implementation is done using PyTorch (Paszke et al., 2019). We train an ensemble of 5 LSTMs, all with a single layer of 64 nodes, to account for random initialization and stochasticity in the network optimization algorithm. The head-layer for our LSTMs is a fully connected linear layer with a dropout rate of 0.4. We use a batch size of 256 and a sequence length of 365 days for training our LSTMs for a total of 20 epochs. We use a learning rate of 1×10^{-3} for the first ten epochs, and 5×10^{-3} for the remaining ten epochs. The basin averaged Nash-Sutcliffe efficiency (NSE*) proposed by Kratzert et al. (2019a) is used as a loss function and the algorithm for optimization is ADAM (Kingma and Ba, 2017). We refer the reader to Kratzert et al. (2019a) for a detailed description of the LSTM architecture and about specific details as to how it is typically applied in hydrology. For easy reference, we present the equations describing the forward pass of the LSTM in appendix D.

2.4 The Hybrid Model

We use a type of hybrid model introduced by Feng et al. (2022). The hybrid model uses a modified version of the Hydrologiska Byråns Vattenbalansavdelning (HBV) model (Aghakouchak and Habib, 2010; Beck et al., 2020; Bergström, 1976, 1992; Seibert and Vis, 2012) as a backbone conceptual model. Differentiable parameter learning (dPL) using a single LSTM is used to parameterize a number of modified HBVs. The discharge signal produced by the modified HBVs is averaged and routed through a unit hydrograph, which produces the final simulated discharge. We implement the $\delta_n(\beta^t, \gamma^t)$ model with a collection of 16 modified HBV models with dynamic parameterization. A detailed description of this model can be found in Feng et al. (2022). While the stand-alone LSTM produces specific discharge as the output, in the hybrid model, the LSTM produces as

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

many outputs as is the number of parameters required by 16 HBVs and the unit hydrograph routing. In our hybrid model, the LSTM produces 210 outputs (13 HBV parameters*16 HBV models+2 routing parameters). The hyperparameters of the LSTM component in the hybrid model are described in Table 1. The hybrid model receives a sequence length of 730 days, the first 365 values from which are used to initialize the internal states of the HBV models (warm-up period) and do not contribute to loss calculation. As mentioned in Table 1, the data split implemented for training and testing is the same for both the hybrid and the LSTM model. The optimizer and learning rate schedule is also the same. The main difference between the stand-alone LSTM network and the hybrid model, besides the sequence length, is that the hybrid model gets potential evapotranspiration (mm d^{-1}) as an additional dynamic input, along with the 5 dynamic and 22 static inputs used while training the LSTM. The daily time series for potential evapotranspiration is obtained from the simulation based hydrometeorological time series of the CAMELS-CH dataset.

2.5 Design Precipitation Events: Selection and Experimental Set-up

In this study, we use design precipitation values from an extreme value analysis published by the Federal Office of Meteorology and Climatology (MeteoSwiss; MeteoSwiss, 2022). This includes 1- to 5-day precipitation analyses with annual return interval (ARI) from 1 to 300 years at more than 300 meteorological observation stations. Given that the design precipitation values are only valid on the exact location of the stations (Frei and Fukutome, 2022), we identified a smaller subset of 25 CAMELS-CH catchments that have a meteorological observation station within or at a distance of 2.5 km from the catchment boundary. We acknowledge that, given the diversity in terrain and elevation in Switzerland, and its small-scale spatial climate patterns, access to sophisticated tools enabling better interpolation of the extreme values would be ideal (Bárdossy and Pegram, 2013). However, due to the lack of such methods and the explicit admission of added uncertainty in the related documentation (Frei and Fukutome, 2022), we proceed with the chosen subset of catchments. This is reasonable since this study is focused on better understanding the limitations of LSTM-based hydrological simulations, rather than addressing actual infrastructure design issues in Switzerland.

To systematically analyze the simulations of our models in extreme scenarios, we force our models with precipitation events of varying ARI during the test period. For each of the above-mentioned 25 catchments, we identified dates, where the observed precipitation value (mm d^{-1}) belonged to the top 99.5th percentile of the distribution of precipitation values during the test period in the respective catchment. The minimum replaced precipitation is 34 mm d^{-1} and the maximum is 139 mm d^{-1}. We replaced these by the 1-, 3-, and 5-day design precipitation values with ARI of 50, 100, and 300 years. In the case of 3- and 5-day values, the precipitation volume was distributed uniformly over three and five days, respectively, centered around the identified dates. The LSTM and hybrid model then received this synthetic input for discharge simulations. This approach allows us to test the impact of extreme, but physically plausible, magnitudes of precipitation input for the LSTM-based discharge simulations, under different initial conditions. Our experimental set-up is constrained by the fact that we only manipulate precipitation. Given that other meteorological variables, such as temperature or radiation, are not fully independent of precipitation, our approach does not account for the complex correlation among climate inputs. However, by only replacing

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

precipitation values at times when observed extremes had already occurred, we try to minimize inconsistencies in other meteorological inputs. While this approach has its limitations, it provides a controlled setting to examine how the LSTM and hybrid models respond to unprecedented precipitation magnitudes and reflects to a certain degree a classical hydrological use case, which is the design of infrastructure.

2.6 Measuring saturation in the LSTM

Although saturation can occur at any $\tanh$ or sigmoid activation within an LSTM, we focus on the saturation that arises during the computation of the hidden state (the second term in Eq. (D6) in appendix D) as discussed by Kratzert et al. (2024). Defining a precise threshold for when $\tanh$ saturates is challenging due to its continuous nature. However, previous studies have noted that the useful (non-saturated) region extends until approximately 90% of the saturation level (Chen and Chang, 1996). We hence identify saturation in the said activation when the absolute of its output equals or exceeds 0.9. We define network saturation as the total number of saturated activations (out of the 64 units in the hidden layer). In the following, we will use the term "cell state saturation" to refer specifically to the saturation of the $\tanh$ activation function when computing hidden states ($h_t = \tanh(c_t) \cdot o_t$).

3 Results

3.1 LSTM and hybrid model performance

Fig. 1 presents the test performance of the LSTM and hybrid model ensemble as a cumulative distribution function (CDF) of individual catchment performance measured by the NSE. The models' testing is spatially in-sample but temporally out-of-sample, which means that the models are tested using the same 196 catchments used during the training process, but in a different test period (gauged simulations). The average median NSE achieved by the LSTM ensemble is 0.84 while that for the hybrid model ensemble is slightly lower at 0.79. Both models perform better than the PREVAH model (Viviroli et al., 2009) (median NSE = 0.60), simulated discharge time series which are provided with the CAMELS-CH dataset. It is worth noting that the hybrid model performed similarly to the LSTM ensemble in studies by Feng et al. (2022) and Acuna Espinoza et al. (2024b) on the CAMELS-US dataset. However, in this study, we could not replicate the same performance, despite using the exact same model setup and training procedure. Our investigations did not reveal a specific cause for the slightly lower NSE observed. Interestingly, in four specific catchments where the hybrid model exhibited a pronounced drop in performance compared to the LSTM ensemble, the hybrid accurately predicted timing patterns (high correlation) but showed an increasing bias over the duration of the test period. This suggests larger mass balance errors in these catchments that could not be corrected by the hybrid model's mass-conserving structure. Given that the hybrid model primarily serves as a benchmark for the LSTM ensemble, the observed difference in NSE is considered negligible for the objectives of this study.

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

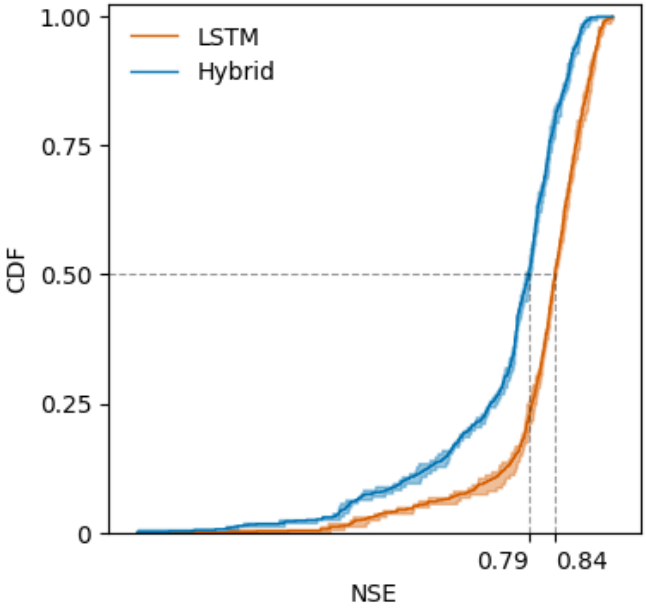

Figure 1. Cumulative Density Function (CDF) showing the NSE of the LSTM and hybrid model ensemble tested on 229 CAMELS-CH catchments during the test period from 01.10.2010 to 30.09.2015. The solid line represents the mean of the ensemble, and the shaded region depicts the variation within the ensemble. The average median NSE achieved by the LSTM network ensemble is 0.84, while that for the hybrid model ensemble is 0.79

3.2 Theoretical prediction limit and maximum simulated value of the LSTM ensemble

Kratzert et al. (2024) discuss the existence of a theoretical prediction limit for a trained LSTM network and provide a mathematical derivation (Appendix C in Kratzert et al., 2024). This theoretical prediction limit depends on the learnable parameters (weights and biases) of the linear head layer that maps the LSTM's hidden states to a single output value. For our LSTM ensemble, the mean theoretical prediction limit is 73 mm d^{-1}. This limit means that under no circumstances can the stand-alone LSTM produce a simulated discharge higher than 73 mm d^{-1}. This theoretical prediction limit is notably smaller than the maximum specific discharge observed during the training period, about 183 mm d^{-1}, which occurred during a flood in the Chli Schliere stream, located in central Switzerland. In total, there are 66 days in the training period during which discharge values exceed 73 mm d^{-1}, representing approximately 0.01% of the total training data.

Our design experiments revealed that the maximum simulated discharge value from the LSTM ensemble is not the theoretical limit of 73 mm d^{-1}, but 60 mm d^{-1}. This maximum was reached during a 1-day design precipitation event, which had a total precipitation volume of 304 mm, in the Magliaso-Ponte catchment located in southern Switzerland. To further investigate how closely the stand-alone LSTM can approach its theoretical maximum, we tested scenarios with extremely high precipitation

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

intensities up to 1000 mm d^{-1} sustained over 3- and 5-day durations. Such values exceed realistic conditions by far, especially considering the fact that the highest total annual precipitation recorded in Switzerland is 4173 mm a^{-1} (MeteoSwiss, 2024). Even under these extreme forcing conditions, the model did not produce a discharge value beyond 60 mm d^{-1}. We hence refer to this simulated maximum as the "design limit" of the LSTM. The "design limit" being smaller than the theoretical prediction limit, can be understood as a consequence of not all linear head-layer units contributing fully to the final output.

Training LSTMs with a higher number of hidden states and on a larger, more diverse dataset (as recommended in Kratzert et al., 2024) can raise the theoretical limit, but does not necessarily affect the "design limit". For instance, a single LSTM network with 256 hidden states, compared to one with 64 hidden states, trained on the CAMELS-CH dataset, demonstrates a theoretical prediction limit of 120 mm d^{-1}. The "design limit" also increased to 75 mm d^{-1}. Similarly, a single LSTM with 256 hidden states, trained on both the CAMELS-CH and CAMELS-US datasets together, achieves a theoretical prediction limit of 194 mm d^{-1} and a raised "design limit" of 110 mm d^{-1}. Despite these improvements, the "design limits" remain significantly lower than the maximum discharges encountered during training: 299 mm d^{-1} in CAMELS-US and 183 mm d^{-1} in CAMELS-CH. While the theoretical limit reflects the maximum potential output based on model parameters, the "design limit" is constrained by the interplay of network weights and activations during inference. Thus, increasing the theoretical maximum by expanding the number of hidden states does not necessarily translate to a higher "design limit".

In contrast, the hybrid model used in our experiments does not exhibit a theoretical limit. The highest simulated value observed was 144 mm d^{-1}, which is still lower than the maximum discharge seen during training. However, when forced with increased precipitation, the model's outputs scale more or less linearly with the forcing, demonstrating greater flexibility than the standalone LSTM.

Panels (a)-(c) in Fig. 2 show the evolution in the simulated specific discharge for three catchments for a particular, catchment-specific, 1-day design precipitation event with varying ARI from 50 to 300 years. We chose these three catchments, as they have the highest flows among the 25 catchments. Notably, the maximum simulated discharge by the stand-alone LSTM ensemble increase only marginally from ARI 50-year to ARI 300-year in all three catchments. For these catchments the simulations increase on average by 6% in contrast to the precipitation, with different ARIs, that increase by 39%. The maximum simulated values of these three catchments, which are 48 mm d^{-1}, 43 mm d^{-1}, and 60 mm d^{-1} respectively, are well below the theoretical limit of the LSTM ensemble, but close to the "design limit". From a hydrological viewpoint, this entails that, although rainfall increases significantly, the LSTM simulations have decreasing runoff coefficients. In contrast, we typically observe an increase in runoff coefficients with increasing intensity of extreme events, as increasing area of a catchment becomes saturated (Beven et al., 2021). The hybrid model ensemble on the other hand responds considerably more to the increasing precipitation input, and there is an increase of 51% from ARI 50-year to ARI 300-year. The identified patterns in the three most runoff reactive test catchments shown in Fig. 2 are on average also true for most of the 25 test catchments. While the precipitation increases by 43% from ARI 50 to ARI 300, the LSTM simulations show an average increase of 25%. Whereas, the hybrid simulations increase by 48%. In some catchments with particularly low runoff values, the LSTM ensemble occasionally produces even

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

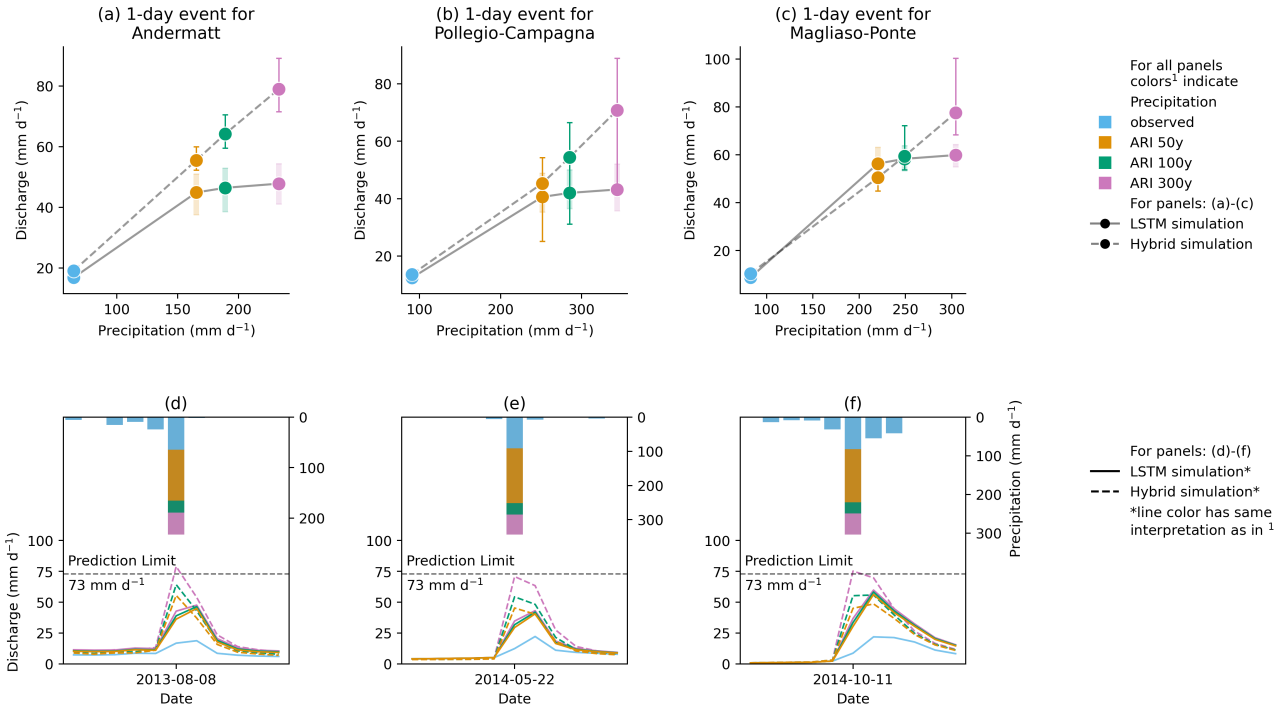

Figure 2. Evolution of LSTM and hybrid model ensemble simulation for three, catchment specific, 1-day events with increasing ARI for gauges located at (a)Andermatt, (b)Pollegio-Campagna and (c)Magliaso-Ponte and their respective hydrographs (d)-(f). The LSTM ensemble doesn't simulate discharge higher than its theoretical prediction limit (panels (d)-(f)). The increase in the hybrid model simulation is more consistent with hydrological expectation than the LSTM (panels (a)-(c)).

higher runoff estimates than the hybrid model. The closer the estimates approach the theoretical prediction limit, the greater the difference between the hybrid model and the LSTM becomes.

Fig. 3 shows the results of a 3-day (panels (a), (c)) and a 5-day (panels (b), (d)) event at the Magliaso-Ponte gauge, one of the test catchments exhibiting the most pronounced runoff responses. Consistent with observations from the 1-day events, the LSTM network simulations reveal certain characteristic limitations. Nonetheless, for both the 3-day and 5-day events, the hybrid model's peak discharge simulations increase with higher ARIs (see panels (a) for the 3-day event and (b) for the 5-day event in Fig. 3), a pattern also evident—though somewhat weaker—in the standalone LSTM results. The discrepancy between the hybrid and the LSTM simulations is much smaller for the 3-day events compared to the 1-day events, and even further reduced for the 5-day events.

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

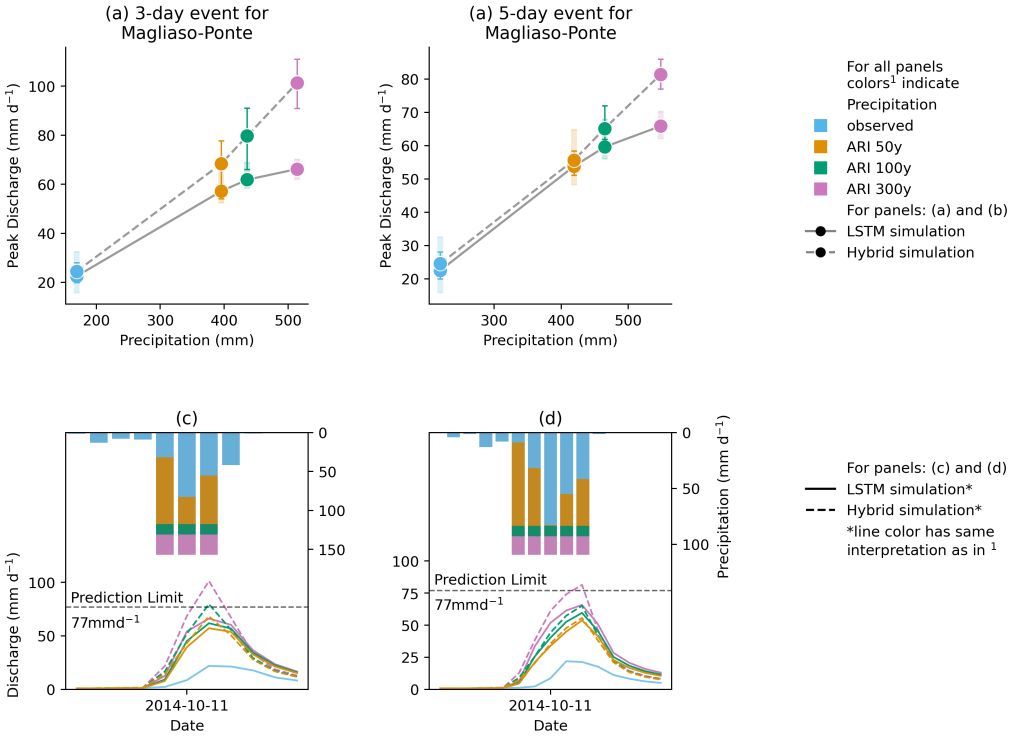

Figure 3. Evolution of LSTM and hybrid model ensemble simulation for gauge located at Magliaso-Ponte for a (a)3-day event and a (b)5-day event with their respective hydrographs (c) and (d).

Table 2. Number of nodes (out of 64) of the LSTM network such that output of the $|\tanh(c_n)| \geq 0.90$. Ensemble maximum (ensemble minimum) values are reported for single events in each catchment. Due to poor reliability of 5-day extreme precipitation analyses for Andermatt (MeteoSwiss, 2022), the corresponding results are not reported here.

ID	Gauge Name	Event Date	Number of Saturated Nodes								
			Design Experiment ARI								
			50y			100y			300y		
			1d	3d	5d	1d	3d	5d	1d	3d	5d
2087	Andermatt	08.08.2013	37(28)	45(42)	-	35(27)	46(43)	-	34(26)	45(43)	-
2494	Pollegio-Campagna	22.05.2014	32(26)	51(42)	50(44)	32(26)	52(39)	50(45)	32(26)	50(40)	51(45)
2461	Magliaso-Ponte	11.10.2014	48(40)	50(41)	47(41)	48(40)	51(42)	49(42)	48(37)	51(44)	51(43)

3.3 Evolution of saturation in the LSTM ensemble

Table 2 shows the maximum (and minimum) number of saturated LSTM cells (out of 64) for three test catchments across various design events. Notably, in none of the cases do the LSTM's cell states fully saturate. For the 1-day events, on average,

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

the maximum saturation across the ensemble ranged from about 50% to 75%, while the minimum ranged from approximately 41% to 63%. Interestingly, this degree of saturation remained nearly unchanged even as the ARI increased, and the associated precipitation became more intense. Even pushing the model with a very high 1-day precipitation of 1000 mm d^{-1} did not cause the cell states to approach complete saturation.

A different pattern emerged, however, when we examined longer-duration events. For the 3-day events, we observed a substantial increase in cell state saturation. This indicates that some cells require more than a single day to accumulate sufficient input signals to reach higher saturation levels. This is thereby controlled by the input and forget gates in an LSTM (Eqs. (D1) and (D2) in appendix D). The input gate controls how much new information enters the cell state, while the forget gate determines how much past information is retained or discarded. Over multiple days, the continued influx of rainfall data (regulated by the input gate) and the retention of previously encoded information (controlled by the forget gate) allow the cell states to build up more gradually. With this prolonged input, more cell states move closer to saturation. For the 5-day events, saturation did not increase further, which at first seems contradictory. However, the total precipitation of the 5-day events does not greatly exceed that of the 3-day events. Since the rainfall is spread uniformly over a longer period, it results in a lower daily precipitation intensity. Without sufficiently large daily inputs, the cell states do not accumulate to higher saturation levels, even over multiple days. Thus, while longer durations can facilitate higher saturation when daily precipitation is intense, simply extending the time frame without maintaining high-intensity input does not necessarily lead to further saturation. The number of saturated cell states, hence, provides useful insights. However, the saturation of the cell states is not the only kind of saturation that limits the LSTM.

4 Discussion

We structure our discussion around the three research questions posed at the end of our introduction.

1. *Can LSTMs extrapolate to discharge values beyond the training distribution when forced with statistically derived design precipitation events?*

Our study highlights limitations in current training strategies. While LSTMs are undeniably powerful tools for modelling complex relationships in hydrological systems (Kratzert et al., 2018, 2019a; Loritz et al., 2024; Nearing et al., 2024), their response to inputs outside the training range exposes critical challenges (Acuna Espinoza et al., 2024b; Song et al., 2024). In order to use ML models responsibly, users should be aware of how the training data limit the model applicability (see also: Meyer and Pebesma, 2021).

Although we train the LSTM ensemble using state-of-the-art methods following the current benchmarks (Kratzert et al., 2019a; Lees et al., 2021; Acuna Espinoza et al., 2024b), it still underestimates discharge values with low exceedance probabilities (high floods), even when these are present in the training data. For instance, although the model saw the largest flood in the training period of 183 mm d^{-1} and 66 other events higher than the theoretical limit 20 times during training (once every epoch

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

of training), the maximum value it could simulate is much lower (73 mm d^{-1}). Extreme hydrological events often coincide with distinct regime shifts, which may necessitate the model to adopt a completely different set of network weights and a unique mapping of inputs to outputs to accurately capture these phenomena. However, reallocating network capacity in this way could compromise the model's ability to simulate more common flow conditions. Thus, the model is potentially disincentivized from fitting to these rare but critical extremes effectively. Another contributing factor may be the inherent bias of minimizing the mean squared error (MSE), which disproportionately penalizes rare outliers and can lead to systematic underestimation of their magnitude. Furthermore, both the inputs and targets are frequently noisy, adding another layer of complexity to accurately capturing extreme events. While our experiments cannot definitively determine which of these factors—or their combination—is primarily responsible for the observed underestimation of extreme floods, the inherent flexibility of LSTMs suggests that this limitation is not intrinsic to the model itself. Instead, it highlights the need for an improved training strategy that better balances the representation of rare extremes and common flow conditions.

Scaling the LSTM by increasing the number of hidden states, and/or providing more training data from a broader range of hydrologic conditions, seems to be an avenue to mitigate this problem. For instance, our LSTM with 256 hidden states, trained on a combined CAMELS-US and CAMELS-CH dataset, results in improved simulations of the extreme events in our test catchments. This corroborates the intuition given by Kratzert et al. (2019a) and studied in Kratzert et al. (2024). However, the theoretical limit of the ensemble, in this case, was still well below the maximum observed training data in Switzerland and far below that of CAMELS-US. Once again, it is imprudent to state with certainty, the underlying reason or combinations thereof—whether it is the rarity of the extreme events or the training strategy which minimizes a squared error. Our study provides some indications on how we can overcome these limits: For one, our results show that stronger structural priors—as for example implemented by the hybrid-approach—can lead to more behavior that is more plausible. However, we do not yet know how strong or weak the structural choices need to be (the study by Frame et al. (2022) indicates that mass conservation alone is not enough). Another potential avenue could come from the training itself: During the training process, there are no technical limits to a prediction made by the LSTM. Hence, the issue could most likely be reduced by a well-chosen training strategy. This could, for example, involve changing the loss function (for instance by weighting high flow events more; Tanrikulu et al., 2024). Alternatively, one can also think to directly train for the warranted behavior. We leave the exploration of these potential solutions to future work. Our results show that there is, indeed, a need for improvement in how we train and setup LSTMs in hydrology.

2. *Is the saturation of LSTM cell states the primary reason, which limits their ability to extrapolate to extreme and unprecedented hydrological conditions?*

Our multi-day design precipitation experiments highlight that, saturation of the cell states can be an important reason for the threshold behavior as increasing inputs led to large values of c_t (Eq. (D5)) for certain cells—which are then asymptotically limited to $-1, 1$ by the $tanh$ function. However, the theoretical limit of the LSTM derived in Kratzert et al. (2024) can only partly explain why the model does not respond to increasing inputs. The reason for this is that the gating mechanisms can in practice

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

saturate much earlier. Hence, one has to consider the model response as a whole and empirically, the design limit lies below the theoretical maximum from Kratzert et al. (2024). As a matter of fact, a deeper examination of the internal mechanisms—particularly the behavior of the gating functions (see appendix D)—showed that, most 1-day design precipitation events never reach the cell state because the input gate (Eq. (D1)) in the LSTM filters them out, or the forget gate (Eq. (D2)) discards most of the historical information. This suggests that the LSTM's inherent assumptions and structural characteristics can prevent it from effectively processing extreme inputs, leading to an underestimation of extreme high-flow events, as additional mass is effectively "deleted" (in contrast, we posit that, for low-flow events this property should not be antagonistic to the hydrological intuition, since saturation behavior naturally occurs there). In principle, an LSTM could also be built with its gating functions employing non-saturating activation functions, but this would typically introduce significant new challenges (e.g., due to vanishing gradients; Hochreiter and Schmidhuber, 1997). Non-saturating functions (e.g., Rectified Linear Units) do not naturally bound the values that flow through the network, making it harder to control the internal state dynamics. Without the built-in constraints provided by sigmoid or $tanh$ activations, the cell states could grow without bound, potentially leading to exploding gradients and destabilized training. In this regard, it is of interest to compare the mechanism of the original LSTM with its latest iteration, the xLSTM (Beck et al., 2024). More specifically, the sLSTM variant. It incorporates a non-saturated exponential function for the input gate. However, it also relies on additional stabilizing mechanisms that also leads to a form of saturation, ensuring that values remain within manageable ranges. In this way, while alternative architectures and activation functions might circumvent certain limitations, they often introduce new challenges related to stability and training dynamics. Ultimately, these findings again highlight that, when it comes to purely data-driven models, there is no simple, one-size-fits-all solution; rather, careful architectural choices, tailored activation functions, and potentially new inductive biases are needed to effectively capture and represent extreme events within LSTM-based models.

3. *How do the inherent assumptions and structural characteristics (inductive biases) of LSTMs influence their ability to simulate realistic hydrological responses under conditions that exceed observed training ranges?*

LSTMs are not just general function approximators, but are also proven to be Turing complete (Siegelmann and Sontag, 1992; Chung and Siegelmann, 2021). However, the inherent assumptions and structural characteristics of an LSTM introduce an inductive bias that can limit its ability to simulate hydrological responses when conditions strongly deviate from those observed during training. In essence, the LSTM's model structure acts as a form of prior knowledge that guides its predictions toward states that reflect its training experience (Hochreiter and Schmidhuber, 1997). The LSTM design, however, does not focus on yielding model behavior that reflects hydrological intuitions in extrapolation regimes. In case of the LSTM and the maximum runoff reaction, this is due to its reliance on saturating activation functions (which, for large precipitation values, results in an input-concave behavior) and in case of the hybrid and its use of linear reservoirs, close to linear (if the parameters remain unchanged during the extreme event; which empirically they do, due to the saturation of the LSTM). In contrast to both models, in hydrology, we might assume a convex model behavior with increase in precipitation (ceteris paribus no changes in the other input features). This is because we typically assume that runoff coefficients increase with increasing intensity of extreme events, as increasing area of a catchment becomes saturated (Beven et al., 2021; Kirchner, 2024). In other words, if

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

we plotted runoff as a function of precipitation for increasingly intense events, we might observe a curve that bends upward (convex). This shape reflects the fact that once critical saturation thresholds are reached, each additional unit of rainfall generates disproportionately more runoff than before. If we trust our hydrological theory, this knowledge should also be reflected in the "inductive bias" of the model we are using. In reality, hydrology is much more complex, and we could observe concave hydrological responses to increasing precipitation, but the a-priori assumption of a convex reaction seems reasonable.

The hybrid model effectively avoids the unrealistic behavior observed in the stand-alone LSTM by enforcing an almost linear behavior due to its use of linear reservoirs. The LSTM component within the hybrid model does saturate (showing a similar behavior as the pure machine learning approach, so that estimated parameters of the hydrological model typically reach their predefined constraints when exposed to precipitation values beyond the training range). Crucially, the conceptual structure of the hybrid model ensures that predicted discharges increase consistently with increasing precipitation. This alignment with hydrological principles allows the hybrid model to provide predictions that remain hydrologically plausible even when the model is forced with inputs outside the observed regime. In other words, the structural choices of the hybrid-model effectively mitigate the saturation behavior observed in the stand-alone LSTM—making the hybrid approach more suitable for applications like infrastructure design where plausible extrapolation behavior is essential (whether the actual behavior reflects a real-world response of the underlying basin and whether it is actually meaningful to use models in this way, is beyond the scope of this study).

For operational flood forecasting, the situation may differ. Recent work by Nearing et al. (2024) highlights the potential advantages of LSTMs over classical hydrological models, particularly when trained on a global database. Our results support this, showing that in catchments with low runoff generation, the LSTM behaves in a hydrologically consistent manner. Additionally, the stand-alone LSTM offers numerous advantages over classical hydrological models. For instance, its flexible use of embedding layers enables the model to seamlessly transition between different temporal frequencies and switch between simulation and forecasting modes (Acuña Espinoza et al., 2024). This adaptability makes LSTMs a powerful tool in operational settings, where diverse conditions and forecasting needs must be addressed efficiently. By emphasizing on high-flow events (Tanrikulu et al., 2024) during training or employing data augmentation techniques like weather generators combined with classical hydrological models (Nguyen et al., 2021), the simulation of extreme events included in the training data could probably be improved.

5 Conclusion

This study investigates the ability of LSTMs to extrapolate under extreme rainfall–runoff conditions and compares their performance with a hybrid model. Based on our findings, we conclude the following:

- Limitations of LSTMs: State-of-the-art LSTMs struggle to predict discharge values beyond a theoretical limit, and this limit is below the range of the training data.

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

- Saturation of LSTM states: While saturation of LSTM cell states contributes to limiting the model's ability to simulate extreme hydrological events, the gating mechanisms play a significant role in filtering or discarding information, especially during 1-day design precipitation events.
- Inconsistent runoff responses: Increasing (extreme) design precipitation events lead to decreasing runoff coefficients, contrary to the hydrological expectation. This highlights structural limitations in the LSTM architecture for hydrological extreme value simulation.
- Hybrid model benchmark: The hybrid model aligns better with hydrological principles, demonstrating consistent scaling of discharge with increasing extreme precipitation. Its mass-conserving structure and use of conceptual hydrological components make it more robust under extreme forcing conditions.
- Potential for improvement: Increasing the number of LSTM hidden states and training on larger, more diverse datasets can raise the theoretical and design prediction limits. However, these adjustments do not fully address the observed limitations, particularly during the 1-day events. Incorporating stronger structural priors, or adapting training strategies which weigh extreme events more during optimization, could mitigate these issues.

Every modeling approach has inherent limitations within its scope of application. While the constraints of conceptual hydrological models are well understood, the same cannot be said for deep learning models, where such limitations remain less explored. We argue that addressing these gaps is crucial for advancing their utility in hydrological applications. The limitations outlined above are not beyond resolution; they represent opportunities for further development. Future research should focus on refining LSTM architectures to better align with hydrological principles, improving training strategies to give greater weight to extreme events during optimization, and exploring innovative hybrid approaches that combine the strengths of data-driven and process-based models. By addressing these challenges, we can move closer to unlocking the full potential of deep learning in hydrological modelling, particularly under extreme forcing conditions. All of the above stated limitations can potentially be fixed, and we believe that future research should focus on refining LSTM architectures, improving training strategies, and exploring and optimizing new hybrid approaches.

Code availability. All the codes for model training, testing, design experiments and plotting the results presented in this paper are available at https://doi.org/10.5281/zenodo.14771377. This also contains the CAMELS CH and the CAMELS US dataset for the ease of reproduction of results.

Data availability. The CAMELS US dataset is freely available at https://doi.org/10.5065/D6MW2F4D (Newman et al., 2015; Addor et al., 2017). The CAMELS CH dataset is freely available at https://doi.org/10.5281/zenodo.7784632 (Höge et al., 2023). Extreme value analyses for Switzerland is available at https://www.meteoswiss.admin.ch/services-and-publications/applications/standard-period.html (MeteoSwiss, 2022)

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

Appendix A: Static and Dynamic Inputs

Table A1 gives the description of the static and dynamic inputs to the LSTM and hybrid models. This description follows from the CAMELS CH dataset (Höge et al., 2023). Where inputs from the CAMELS US (Addor et al., 2017) are listed, they have similar and corresponding interpretation in Addor et al. (CAMELS-US 2017).

Table A1: Dynamic and static inputs used to train the [1]LSTM ensembles using the CAMELS CH dataset, [2]LSTM ensembles using CAMELS CH and CAMELS US dataset combined and [3]hybrid model ensembles.(Addor et al., 2017)

CAMELS CH	CAMELS US	Description	Unit
Dynamic Inputs			
precipitation	prcp	Observed daily summed precipitation[1,2,3]	mm d^{-1}
temperature_min	tmin	Observed daily minimum temperature[1,2,3]	°C
temperature_max	tmax	Observed daily maximum temperature[1,2,3]	°C
rel_sun_dur		Observed daily averaged relative sunshine (solar irradiance $\geq$ 200 W m-2) duration[1,3]	%
swe		Observed daily averaged snow water equivalent[1,3]	mm
pet_sim		Simulated daily averaged potential evapotranspiration (Penman–Monteith equation without interception correction)[3]	mm d^{-1}
area	area_gages2	catchment area	m^2
elev_mean	elev_mean	Mean elevation within catchment	m a.s.l.
slope_mean	slope_mean	Catchment mean slope over all grid cells	°
sand_perc	sand_frac	Percentage sand	%
silt_perc	silt_frac	Percentage silt	%
clay_perc	clay_frac	Percentage clay	%
porosity	soil_porosity	Volumetric porosity	-
conductivity	soil_conductivity	Saturated hydraulic conductivity	cm h^{-1}
glac_area		Glacier area of Swiss glaciers per catchment	km^2
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		Total storage capacity of reservoirs in megaliters	ML
p_mean	p_mean	Mean daily precipitation	mm d^{-1}

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

CAMELS CH	CAMELS US	Description*	Unit
Static Inputs			
pet_mean	pet_mean	Mean daily potential evapotranspiration (PET; Penman–Monteith equation without interception correction)	mm d^{-1}
p_seasonality	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))	-
frac_snow	frac_snow	Fraction of precipitation falling as snow, i.e., while temperature is < 0 °C	-
high_prec_freq	high_prec_freq	Frequency of high-precipitation days ($\geq$ 5 times mean daily precipitation)	d yr^{-1}
low_prec_freq	low_prec_freq	Frequency of dry days (< 1 mm d^{-1})	d yr^{-1}
high_prec_dur	high_prec_dur	Average duration of high-precipitation events (number of consecutive days $\geq$ 5 times mean daily precipitation)	d
low_prec_dur	low_prec_dur	Average duration of dry periods (number of consecutive days < 1 mm d^{-1} mean daily precipitation)	d

Appendix B: LSTM network and hybrid model ensemble results for 1-day design precipitation event for 30 catchments

Fig. B1 presents the results from the 1-day design experiment for all 25 test catchments selected in section 2.5 of this paper. Catchments with a strong rainfall–runoff generation show a concave increase in the runoff with increasing intensity of design precipitation, whereas for all the test catchments, the response of the hybrid model increases linearly.

Appendix C

As mentioned in section 3.2 of this paper, increasing the number of hidden states, and/or training the LSTMs on larger datasets, increases the theoretical prediction limit as given in Table C1. LSTMs with more hidden states and/or trained on larger dataset also simulate higher runoff for the design precipitation values. Nevertheless, this response, too, is concave (Fig. C1), unlike the hybrid model response.

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

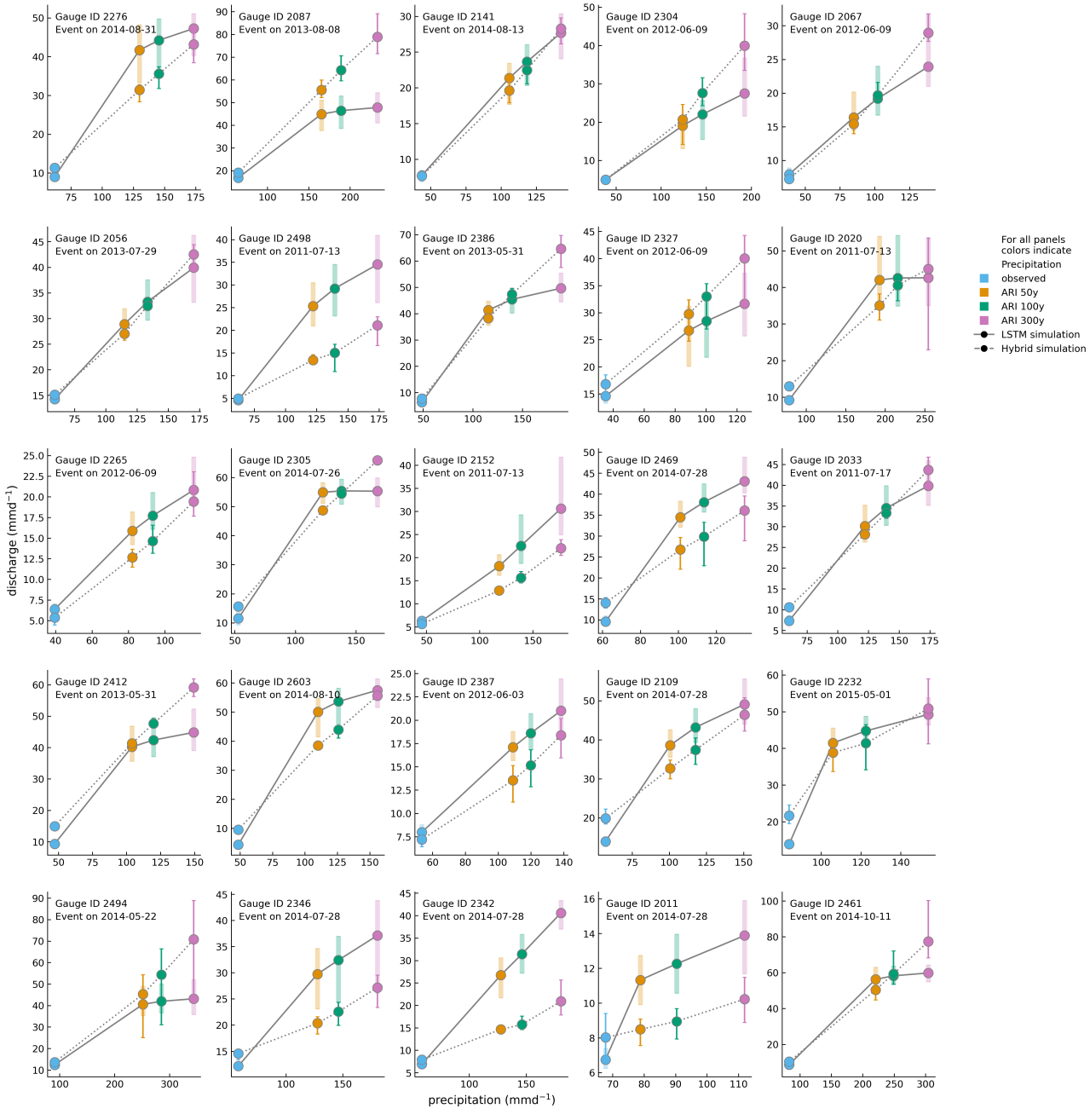

Figure B1. LSTM and hybrid model ensemble simulation for 25 catchment specific events with varying ARI.

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

C1 Theoretical Prediction Limits for LSTM networks with more nodes and trained on different datasets

Table C1. Theoretical prediction limits for different LSTM networks. max(y_{obs}) indicates the maximum observed target value during the training period from 01.10.1995 to 30.09.2005.
*used in this study [1]ensemble of 5 LSTMs. [2]single LSTM

LSTM Network	Number of Nodes	Training Dataset	Theoretical Prediction Limit mm d^{-1}	max(y_{obs}) mm d^{-1}
LSTM_CH*[1]	64	229 CAMELS-CH catchments	73	183
LSTM_CH[2]	256		120	
LSTM_US_CH[1]	64	229 CAMELS-CH	115	299
LSTM_US_CH[2]	256	and 531 CAMELS-US catchments	193	

C2 Additional LSTM networks' and hybrid model ensemble results for 1-day design precipitation event for four catchments with gauge IDs 2087, 2494 and 2461

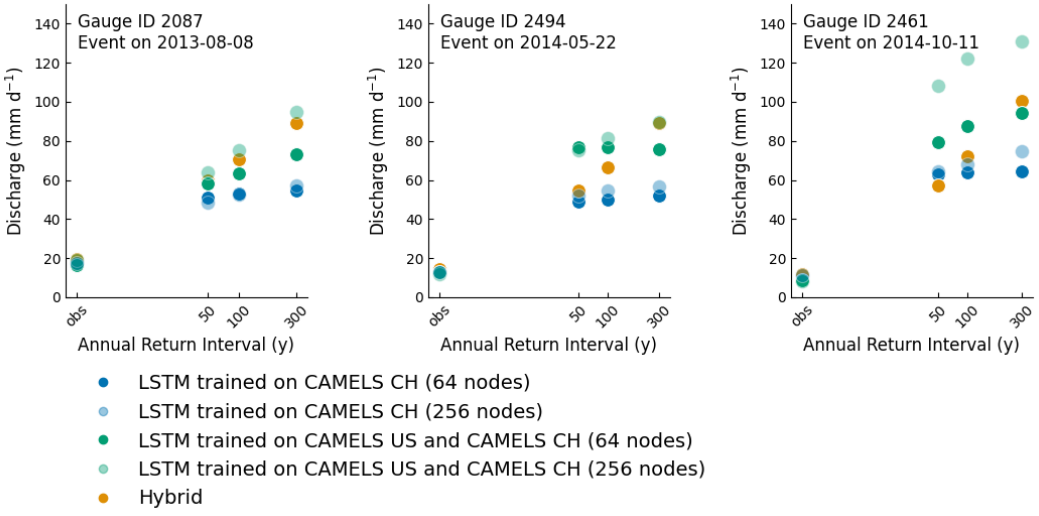

Figure C1. Additional LSTM networks' and hybrid model ensemble simulation for 3 catchment specific events.

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

Appendix D: Equations describing the LSTM forward pass

The LSTM forward pass can be mathematically represented by the following:

$$\boldsymbol{i_t} = \sigma\left(\boldsymbol{W_i x_t} + \boldsymbol{U_i h_{t-1}} + \boldsymbol{b_i}\right), \tag{D1}$$

$$\boldsymbol{f_t} = \sigma\left(\boldsymbol{W_f x_t} + \boldsymbol{U_f h_{t-1}} + \boldsymbol{b_f}\right), \tag{D2}$$

$$\boldsymbol{g_t} = \tanh\left(\boldsymbol{W_g x_t} + \boldsymbol{U_g h_{t-1}} + \boldsymbol{b_g}\right), \tag{D3}$$

$$\boldsymbol{o_t} = \sigma\left(\boldsymbol{W_o x_t} + \boldsymbol{U_o h_{t-1}} + \boldsymbol{b_o}\right), \tag{D4}$$

$$\boldsymbol{c_t} = \boldsymbol{f_t} \odot \boldsymbol{c_{t-1}} + \boldsymbol{i_t} \odot \boldsymbol{g_t}, \tag{D5}$$

$$\boldsymbol{h_t} = \boldsymbol{o_t} \odot \tanh\left(\boldsymbol{c_t}\right), \tag{D6}$$

where i_t, f_t, and o_t are the input gate, forget gate, and output gate, respectively, g_t is the cell input and x_t is the network input at time step t, and h_{t-1} is the recurrent input, c_{t-1} the cell state from the previous time step. W, U, and b are learnable parameters for each gate, where subscripts indicate which gate the particular weight matrix/vector is used for, σ is the sigmoid function, $tanh$ is the hyperbolic tangent function, and $\odot$ is element-wise multiplication.

Author contributions. The idea for the paper was proposed by RL. Codes developed by EAE were used for training the models. Model training and testing, the design experiments and analysis were done by SB, and results were discussed with RL. The draft was prepared by SB and reviewed and edited by all authors. Funding was aquired by RL. 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-2025-425
Preprint. Discussion started: 6 February 2025

References

Acuña Espinoza, E., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., Loritz, R., and Ehret, U.: Technical note: An approach for handling multiple temporal frequencies with different input dimensions using a single LSTM cell, EGUsphere, 2024, 1–12, https://doi.org/10.5194/egusphere-2024-3355, 2024.

Acuña Espinoza, E., Loritz, R., Álvarez Chaves, M., Bäuerle, N., and Ehret, U.: To bucket or not to bucket? Analyzing the performance and interpretability of hybrid hydrological models with dynamic parameterization, Hydrology and Earth System Sciences, 28, 2705–2719, https://doi.org/10.5194/hess-28-2705-2024, 2024a.

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.

Aghakouchak, A. and Habib, E.: Application of a Conceptual Hydrologic Model in Teaching Hydrologic Processes, International Journal of Engineering Education, 26, 963–973, 2010.

Balestriero, R., Pesenti, J., and LeCun, Y.: Learning in High Dimension Always Amounts to Extrapolation, https://doi.org/10.48550/arXiv.2110.09485, 2021.

Beck, H. E., Pan, M., Lin, P., Seibert, J., van Dijk, A. I. J. M., and Wood, E. F.: Global Fully Distributed Parameter Regionalization Based on Observed Streamflow From 4,229 Headwater Catchments, Journal of Geophysical Research: Atmospheres, 125, e2019JD031 485, https://doi.org/https://doi.org/10.1029/2019JD031485, e2019JD031485 10.1029/2019JD031485, 2020.

Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., and Hochreiter, S.: xLSTM: Extended Long Short-Term Memory, https://arxiv.org/abs/2405.04517, 2024.

Bergström, S.: DEVELOPMENT AND APPLICATION CONCEPTUAL RUNOFF MODEL FOR SCANDINAVIAN CATCHMENTS, Ph.D. thesis, Swedish Meteorological and Hydrological Institute (SMHI), https://www.smhi.se/en/publications/the-hbv-model-its-structure-and-applications-1.83591, 1976.

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. J., Kirkby, M. J., Freer, J. E., and Lamb, R.: A history of TOPMODEL, Hydrology and Earth System Sciences, 25, 527–549, https://doi.org/10.5194/hess-25-527-2021, 2021.

Bárdossy, A. and Pegram, G.: Interpolation of precipitation under topographic influence at different time scales, Water Resources Research, 49, 4545–4565, https://doi.org/https://doi.org/10.1002/wrcr.20307, 2013.

Chen, C.-T. and Chang, W.-D.: A feedforward neural network with function shape autotuning, Neural Networks, 9, 627–641, https://doi.org/https://doi.org/10.1016/0893-6080(96)00006-8, 1996.

Chung, S. and Siegelmann, H.: Turing Completeness of Bounded-Precision Recurrent Neural Networks, https://proceedings.neurips.cc/paper_files/paper/2021/file/ef452c63f81d0105dd4486f775adec81-Paper.pdf, 2021.

Federal Department for the Environment, Transport, E. and DETEC, C.: The Floods of 2005 in Switzerland, 2005.

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

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.

Frei, C. and Fukutome, S.: Extreme Point Precipitation, https://hydromaps.ch/#en/8/46.830/8.190/bl_hds--precip_24h_2a$4/NULL, 2022.

Global Water Partnership (GWP) and World Meteorological Organization (WMO): Integrated Flood Management Tools Series No. 20 Flood Mapping, Tech. rep., World Meteorological Organization (WMO), https://library.wmo.int/idurl/4/37083, 2013.

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.

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

Kirchner, J. W.: Characterizing nonlinear, nonstationary, and heterogeneous hydrologic behavior using ensemble rainfall–runoff analysis (ERRA): proof of concept, Hydrology and Earth System Sciences, 28, 4427–4454, https://doi.org/10.5194/hess-28-4427-2024, 2024.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks, Hydrology and Earth System Sciences, 22, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018, 2018.

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., 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.

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.

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.

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

MeteoSwiss, F.: Records and extremes, https://www.meteoswiss.admin.ch/climate/the-climate-of-switzerland/records-and-extremes.html, 2024.

Meyer, H. and Pebesma, E.: Predicting into unknown space? Estimating the area of applicability of spatial prediction models, Methods in Ecology and Evolution, 12, 1620–1633, https://doi.org/https://doi.org/10.1111/2041-210X.13650, 2021.

https://doi.org/10.5194/egusphere-2025-425
Preprint. Discussion started: 6 February 2025

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T. Y., Weitzner, D., and Matias, Y.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, https://doi.org/10.1038/s41586-024-07145-1, 2024.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, Hydrology and Earth System Sciences, 19, 209–223, https://doi.org/10.5194/hess-19-209-2015, 2015.

Nguyen, V. D., Merz, B., Hundecha, Y., Haberlandt, U., and Vorogushyn, S.: Comprehensive evaluation of an improved large-scale multi-site weather generator for Germany, International Journal of Climatology, 41, 4933–4956, https://doi.org/https://doi.org/10.1002/joc.7107, 2021.

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.

Rakitianskaia, A. and Engelbrecht, A.: Measuring Saturation in Neural Networks, in: 2015 IEEE Symposium Series on Computational Intelligence, pp. 1423–1430, https://doi.org/10.1109/SSCI.2015.202, 2015.

Seibert, J. and Vis, M. J. P.: Teaching hydrological modeling with a user-friendly catchment-runoff-model software package, Hydrology and Earth System Sciences, 16, 3315–3325, https://doi.org/10.5194/hess-16-3315-2012, 2012.

Siegelmann, H. T. and Sontag, E. D.: On the computational power of neural nets, https://doi.org/10.1145/130385.130432, 1992.

Song, Y., Sawadekar, K., Frame, J. M., Pan, M., Clark, M., Knoben, W. J. M., Wood, A. W., Patel, T., and Shen, C.: Improving Physics-informed, Differentiable Hydrologic Models for Capturing Unseen Extreme Events, https://doi.org/10.22541/essoar.172304428.82707157/v1, 2024.

Tanrikulu, O. D., Ehret, U., Haag, I., Loritz, R., and Badde, U.: Untersuchungen zum Potenzial maschineller Lernverfahren für die hydrologische Simulation und Vorhersage am Beispiel von LSTM und LARSIM in Baden-Württemberg, https://doi.org/10.5675/HYWA_2024.3_1, publisher: Federal Institute of Hydrology, 2024.

Viviroli, D., Zappa, M., Gurtz, J., and Weingartner, R.: An introduction to the hydrological modelling system PREVAH and its pre- and post-processing-tools, Environmental Modelling & Software, 24, 1209–1222, https://doi.org/https://doi.org/10.1016/j.envsoft.2009.04.001, 2009.

World Meteorological Organization (WMO): Manual for Estimation of Probable Maximum Precipitation, Tech. rep., World Meteorological Organization (WMO), 1973.

World Meteorological Organization (WMO): Guide to Hydrological Practices, Volume II Management of Water Resources and Applications of Hydrological Practices, Tech. rep., World Meteorological Organization (WMO), https://library.wmo.int/idurl/4/36066, 2009.