



Targeted data augmentation to improve the simulation of extreme floods with regional deep learning models

Alexander Dolich¹, Paul Reis^{1,2}, Sergiy Vorogushyn³, Viet Dung Nguyen³, Antara Dasgupta², and Ralf Loritz¹

¹Karlsruhe Institute of Technology (KIT), Institute for Water and Environment, Karlsruhe, Germany

²RWTH Aachen University, Data Driven Computing in Civil Engineering, Aachen, Germany

³GFZ Helmholtz Centre for Geosciences, Section Hydrology, Potsdam, Germany

Correspondence: Alexander Dolich (alexander.dolich@kit.edu)

Abstract. Deep learning models, particularly Long Short-Term Memory (LSTM) networks, achieve state-of-the-art overall performance in regional rainfall–runoff modeling. However, under unprecedented extreme meteorological forcings, purely data-driven LSTMs can suffer from artificial output saturation—a critical structural limitation that remains largely hidden during standard historical evaluations. In this study, we utilize a regional, non-stationary weather generator to produce synthetic extreme precipitation events, stress-testing a standard LSTM alongside mass-conserving conceptual HBV and hybrid models across a subset of catchments from the CAMELS-DE dataset. To mitigate the LSTM’s saturation problem, we introduce a targeted data augmentation strategy. By driving the physics-constrained models with synthetic extreme forcings from the weather generator, we create physically plausible discharge targets to enrich the LSTM’s training data.

Our results demonstrate that adding even a small fraction of synthetic extreme events successfully mitigates artificial output saturation by shifting the LSTM’s predictive limit upward, even exceeding that of an LSTM trained on the entire CAMELS-DE dataset. The augmentation improves peak-flow simulation performance without compromising overall simulation accuracy. We validate this approach using the catastrophic out-of-sample July 2021 flood event across multiple affected gauges, demonstrating that regional data augmentation can significantly reduce peak flow underestimation. However, while augmentation mitigates the practical impact of the LSTM’s architectural ceiling, it does not eliminate it. Notably, the evaluated hybrid model natively avoids saturation, yielding the best overall extreme-flow performance. Ultimately, for operational flood forecasting and design flood estimation, we emphasize the necessity of quantifying the LSTM’s predictive limit and monitoring model outputs for artificial saturation. To prevent the potentially catastrophic underestimation of unprecedented extremes, we strongly recommend deploying mass-conserving or hybrid architectures in parallel with standard LSTMs as structural safeguards.

1 Introduction

In recent years, data-driven models, particularly Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), have demonstrated strong performance in rainfall–runoff modeling, often achieving performance comparable to or better than that of conceptual hydrological models (Lees et al., 2021; Kratzert et al., 2019; Loritz et al., 2024). While LSTMs achieve superior accuracy across most of the hydrograph, they exhibit a critical, yet often overlooked, limitation: they can



only simulate peak flows up to a strict mathematical ceiling. As first described by Kratzert et al. (2024), this theoretical prediction limit arises because the LSTM's hidden-state representation is bounded by saturating activation functions such as the sigmoid and hyperbolic tangent. Consequently, once the bounded hidden states are passed through the final linear output layer, the fixed weights and biases of the trained model impose a finite upper bound on the predicted discharge. Critically, Baste et al. (2025) showed that this theoretical limit can be substantially lower than the maximum discharge observed in the training data, rendering the model mathematically incapable of reproducing some of the highest historical extremes encountered during training. The authors further demonstrated that model outputs saturate well below this theoretical maximum. By forcing LSTMs with extreme 1-, 3-, and 5-day design precipitation events, they identified an empirical ceiling, termed the "design limit", which the models could not exceed, regardless of precipitation intensity. As extreme precipitation volumes increase, the simulated discharge stagnates, leading to decreasing runoff coefficients with increasing rainfall intensity—a physically implausible behavior. This limitation is especially concerning when LSTMs are applied to estimate design discharges or for operational flood forecasting, where accurately simulating extreme events is critical for infrastructure resilience, early warning, and disaster preparedness.

Although this upper bound is architectural, its magnitude depends on the weights and biases learned from the training data. Its practical consequences are therefore closely connected to the limited exposure of LSTMs to extreme events during training. Such events are inherently rare in the historical record and remain scarce even in large-sample hydro-meteorological datasets (e.g. Addor et al., 2017; Kratzert et al., 2023; Loritz et al., 2024). In some catchments, they may be entirely absent from the observational record, although their absence does not make their future occurrence less likely (Montanari et al., 2024). It is not a new finding that neural network models struggle to generalize beyond the range of their training data (Barnard and Wessels, 1992; Hettiarachchi et al., 2005). While several strategies exist to improve this limitation, data augmentation represents a particularly promising pathway. In machine learning, data augmentation is a standard data-space approach used to improve generalization and robustness by expanding the training set with synthetic samples that preserve valid input-target relationships (Dao et al., 2019; Iwana and Uchida, 2021). Rather than altering the underlying model architecture, it enriches the training data so that the network can adapt to a wider input space (Wen et al., 2021). In hydrology, general data augmentation techniques (e.g., noise addition, scaling, or mixup) have been shown to improve median model performance in low-data regimes, but their benefits diminish when applied to large-sample training datasets (Kenne et al., 2026). This suggests that once a training dataset is sufficiently large and diverse, the primary remaining bottleneck is no longer general variability, but the severe and asymmetric deficit of rare high-flow events caused by the pronounced heavy-tail behavior of precipitation and flood discharge distributions (Merz et al., 2022). Previous studies have shown that physically based models and stochastic rainfall can be used to enrich the availability of such events and improve peak-flow simulations (Yang et al., 2020; López-Chacón et al., 2023). However, these studies were limited to simpler machine learning architectures in localized, data-scarce settings, leaving the potential of targeted augmentation for regional, state-of-the-art deep learning models unexplored—particularly as a mechanism to mitigate the inherent structural extrapolation ceiling of the LSTM.

Building on this idea, we propose a targeted augmentation strategy focused specifically on the right tail of the discharge distribution. We use a non-stationary regional weather generator (nsRWG), introduced by Nguyen et al. (2024), to generate a



large ensemble of spatially consistent synthetic extreme rainfall events. We compare three hydrological modeling architectures: a stand-alone LSTM, the conceptual HBV model (Bergström and Forsman, 1973; Seibert, 2005), and a hybrid architecture in which an LSTM dynamically parametrizes an HBV model (Feng et al., 2022). While the stand-alone LSTM is subject to a finite output limit, the HBV and hybrid models retain a mass-conserving rainfall–runoff response and do not impose a fixed architectural upper limit on simulated discharge. We therefore drive these models with extreme nsRWG rainfall events to generate synthetic peak-discharge targets and inject the resulting events into the regional LSTM training dataset. This targeted augmentation directly addresses the deficit of extreme events while maintaining spatially and temporally consistent meteorological forcing and physically plausible discharge responses. Ultimately, this allows us to investigate whether augmenting the training data with synthetic extreme scenarios can raise the prediction limit of the LSTM and improve its ability to simulate unprecedented extreme discharges.

In this study we address the following research questions:

1. How do the different hydrological model architectures respond to unprecedented extreme meteorological forcings from the weather generator?
2. Can targeted data augmentation mitigate the inherent limitations of the LSTM?
3. What are the implications of these findings for large-sample, regional LSTM training strategies?

2 Data and Methods

This section introduces the observational data from the CAMELS-DE dataset (Sect. 2.1) and the synthetically generated weather data (Sect. 2.2). We then detail the three modeling architectures evaluated: the stand-alone LSTM (Sect. 2.3), the conceptual HBV model (Sect. 2.4), and the hybrid model (Sect. 2.5). The section closes with an outline of the data augmentation methodology designed to target the LSTM’s structural limits (Sect. 2.6).

2.1 Observed historical data: CAMELS-DE

The historical data basis for model training and calibration is the CAMELS-DE dataset (Loritz et al., 2024; Dolich et al., 2025). CAMELS-DE provides daily-resolution meteorological forcing data derived from the HYRAS gridded dataset (Rauthe et al., 2013), alongside observed discharge records spanning the period from 1951 to 2020. Potential evapotranspiration (PET, mm d^{-1}), which is required by the HBV and hybrid models, is calculated from daily extreme temperatures using the modified Hargreaves formula (Droogers and Allen, 2002; Clerc-Schwarzenbach et al., 2024). Previous benchmark studies (Loritz et al., 2024; Dolich et al., 2026c) have demonstrated that both LSTM and conceptual HBV models can achieve high performance when trained or calibrated on this dataset (median NSE: 0.85 for the LSTM, 0.72 for the HBV). For this study, we filter the CAMELS-DE dataset to align with the spatial domain of the weather generator. A catchment is included if at least 80 % of its area is covered by the gridded weather generator data. This spatial filtering results in an initial selection of 180 catchments located within and slightly beyond the borders of North Rhine-Westphalia (Fig. 1). While the LSTM and the hybrid model

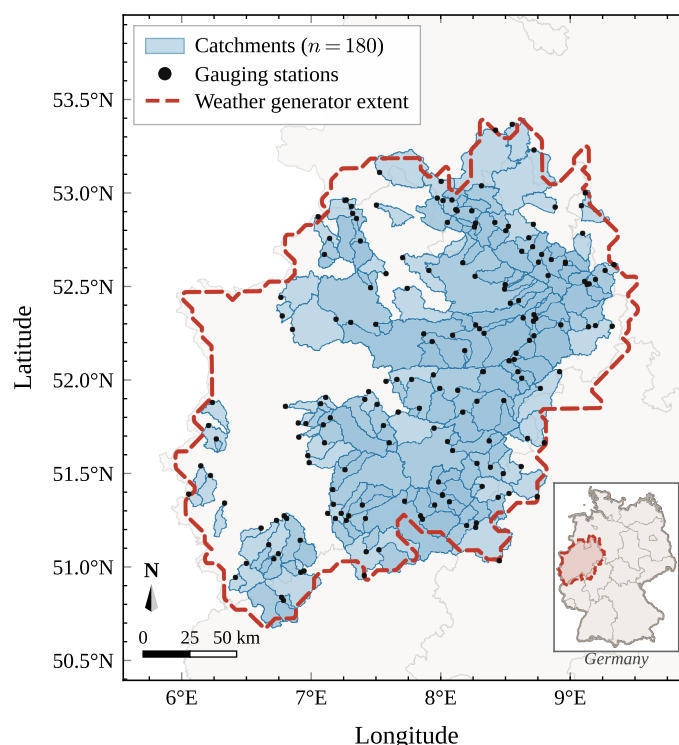


Figure 1. Overview of the study region, comprising North Rhine-Westphalia and adjacent headwater catchments. The main map illustrates the spatial distribution of the $n = 180$ CAMELS-DE catchments (shaded blue) alongside their respective gauging stations (black dots). The geographical extent of the weather generator domain is indicated by the dashed red line. The bottom-right inset shows the location of the study area within Germany. Borders of Germany: © GeoBasis-DE / BKG (2023).

are regionally trained on this full set of 180 catchments, 13 locations lack sufficient observational discharge data records for reliable HBV model calibration. Consequently, the HBV model is only calibrated for 167 catchments, restricting comparative evaluations across the three models to this subset. In addition to the meteorological time series, we utilize the static catchment attributes provided in CAMELS-DE as input features for the regional training of the LSTM and hybrid models.

Because the standard CAMELS-DE dataset ends in 2020, it does not include the catastrophic July 2021 flood, which we utilize as an independent, high-impact real-world test case. We simulate this event for multiple heavily affected gauges in North Rhine-Westphalia and northern Rhineland-Palatinate. These include the Opladen (Wupper) and Kreuzberg (Sahrbach) gauges, alongside the Altenahr gauge (Ahr), which recorded the devastating Ahrthal flood (Vorogushyn et al., 2022). To enable these simulations, we aggregate hourly streamflow and temperature data from the CAMELS-DE-1h dataset (Dolich et al., 2026b) to a daily resolution (summing specific discharge and averaging temperature) and process the corresponding daily HYRAS precipitation data (Rauthe et al., 2013) using spatial catchment means. This allows us to run the pre-trained LSTM models on these out-of-sample unprecedented extreme events.



2.2 Generated weather data: nsRWG

To probe model behavior under extreme forcings, we employ a non-stationary Regional Weather Generator (nsRWG) developed by Nguyen et al. (2024), specifically calibrated for our study area (Fig. 1). The nsRWG is a stochastic weather model capable of generating multiple synthetic realizations of gridded precipitation and air temperature. By fitting heavy-tailed marginal distributions to the observed historical data, the weather generator can produce statistically plausible and spatio-temporally coherent extreme meteorological events that significantly exceed the magnitude of those in the historical record (Nguyen et al., 2021). While such generators are typically used for flood frequency and risk assessments, we leverage the nsRWG both to systematically stress-test the structural limits of our models and to generate extreme forcing sequences for data augmentation.

Unlike traditional weather generators that assume a fixed climate, the nsRWG incorporates non-stationarity, achieved by conditioning its generations on large-scale atmospheric circulation patterns and temperature trends derived from ERA5 reanalysis data (Hersbach et al., 2020), while being calibrated locally against the HYRAS dataset. Because HYRAS is the meteorological data source used in CAMELS-DE, this calibration ensures statistical consistency between the generated data and the observed historical data from CAMELS-DE. In total, the nsRWG provides 100 alternative 71-year realizations (1951–2021) of 5×5 km gridded daily precipitation and air temperature. By simulating continuous weather fields rather than isolated storms, the generator produces dynamic antecedent moisture states prior to extreme events and provides the meteorological drivers for diverse flood-generating mechanisms, including snow-melt, convective storms, and prolonged heavy rainfall. To work with the lumped hydrological models, the gridded outputs are aggregated to spatial catchment means. Consistent with the CAMELS-DE data, potential evapotranspiration (PET, mm d^{-1}) is calculated using the modified Hargreaves formula to force the HBV and hybrid models.

2.3 LSTM model

The data-driven baseline is a regionally trained LSTM model, largely following the CAMELS-DE benchmark setup by Loritz et al. (2024) with specific modifications for our experimental design. The model is trained regionally on the 180 catchments in the study area (Fig. 1). To ensure compatibility between the historical training data and the nsRWG data during inference, the dynamic input variables are reduced to solely mean catchment precipitation (mm d^{-1}) and mean air temperature ($^{\circ}\text{C}$). The set of static attributes is unchanged from the CAMELS-DE benchmark, and the target variable is the specific discharge (mm d^{-1}).

The training and evaluation splits remain consistent with the CAMELS-DE benchmarks: the training period is set from 1 October 1970 to 31 December 1999, the validation period from 1 October 1965 to 30 September 1970, and the testing period from 1 January 2000 to 31 December 2020. The network architecture consists of a single hidden layer with a hidden size of 128. The model is trained for 30 epochs using the Adam optimization algorithm with a sequence length of 365 days, a batch size of 256, a learning rate of 0.001, and a dropout rate of 0.4. The loss function is the basin-averaged Nash-Sutcliffe efficiency (NSE*) proposed by Kratzert et al. (2019). To mitigate uncertainties associated with random weight initialization, we train an ensemble of five independent LSTMs using different random seeds. Throughout this study, all reported LSTM simulations and computed metrics refer to the ensemble median at each time step. After initial evaluation on the historical testing period,



135 the trained ensemble is forced with the nsRWG data to investigate the LSTM's ability to simulate unprecedented extreme events and to identify its simulation limits. Finally, to test whether the architectural constraints of the LSTM can be mitigated, the LSTM setup is retrained in a targeted data augmentation experiment, where the historical training data is enriched with synthetic extreme events (detailed in Sect. 2.6).

2.4 HBV model

140 To provide a physically constrained baseline, we utilize the conceptual HBV model, specifically the implementation by Acuña Espinoza et al. (2024), which is based on Seibert (2005). In addition to precipitation and mean air temperature, the HBV model receives potential evapotranspiration (PET, mm d^{-1}) as an input forcing. Targeting improved high-flow performance, we apply two structural adjustments to the standard CAMELS-DE HBV setup: the inclusion of a calibrated direct runoff component (which accelerates the runoff reaction) and a saturation-excess runoff mechanism, which routes precipitation
145 directly into the stream when the soil storage is filled.

The calibration and testing periods for the HBV model are identical to those used for the LSTM. The model is calibrated individually for each catchment using the Differential Evolution Adaptive Metropolis (DREAM) algorithm (Vrugt, 2016), implemented via the SPOTPY library (Houska et al., 2015), with the Gaussian Likelihood serving as the objective function. To account for parameter uncertainty and the issue of equifinality in hydrological modeling (Beven, 2006), where multiple distinct
150 parameter sets can yield similarly good performance, we generate an ensemble of five calibrated parameter sets per catchment. As with the LSTM setup, all reported HBV simulations and metrics refer to the ensemble median at each time step.

2.5 Hybrid model

In addition to the stand-alone LSTM and HBV models, we evaluate a hybrid architecture that combines both approaches. Specifically, we utilize a variant of the $\delta_n(\gamma_t, \beta_t)$ model with dynamic parameterization as proposed by Feng et al. (2022) and
155 implemented by Acuña Espinoza et al. (2026a). In this setup, a neural network (LSTM) processes the meteorological inputs and static catchment attributes to dynamically set the parameters of conceptual backbone models at each time step. Rather than a single conceptual model, the LSTM simultaneously parametrizes an internal ensemble of 16 parallel HBV models. This multi-model configuration aims at representing spatial heterogeneity through virtual sub-catchments, and to directly account for parameter equifinality. The mean discharge produced by these 16 members is subsequently routed through a two-parameter
160 unit hydrograph to yield the final simulated discharge. The HBV backbone in this hybrid setup also includes the structural modifications (i.e., the direct and saturation-excess runoff components) introduced in our stand-alone HBV model (Sect. 2.4). The LSTM part learns and dynamically sets the following parameters: runoff shape coefficient ($BETA$), evapotranspiration shape coefficient ($BETAET$), recession coefficients (K_0, K_1, K_2), and fraction of direct runoff (F_{dir}). Consequently, the LSTM predicts 98 parameters at each time step (16 ensemble members $\times$ 6 HBV parameters, plus 2 routing parameters).

165 The hybrid model uses the same training, validation and testing periods and receives the same dynamic and static inputs as the stand-alone LSTM, with the addition of potential evapotranspiration (PET), which is required for the HBV component's internal water balance (Sect. 2.4). The hyperparameters of the hybrid model are similar to those of the LSTM; the hybrid model



has a hidden size of 256, is trained for 30 epochs, uses a batch size of 256, a dropout rate of 0.4, and applies a scheduled learning rate of 0.001 initially, reduced to 0.0005 at epoch 10 and to 0.0001 at epoch 20. The model is trained using the Adam optimizer and the basin-averaged Nash-Sutcliffe efficiency (NSE*) as the loss function. We also train an ensemble of five hybrid models using different random seed initializations for the neural network component to mitigate initialization uncertainty. All reported hybrid model results correspond to the ensemble median simulation at each time step.

While this architecture possesses significantly more degrees of freedom than a standard conceptual model, the neural network's outputs are strictly forced to operate within a physically consistent framework dictated by minimum and maximum parameter bounds and the mass conservation laws of the HBV structure. After initial training and evaluation on the historical CAMELS-DE dataset, the trained hybrid ensemble is forced with the 100 realizations of the nsRWG data. By comparing its response to extreme precipitation to the responses of the stand-alone LSTM and the conceptual HBV model, we specifically evaluate the extrapolation capabilities of the hybrid architecture. Ultimately, we aim to test whether this hybrid approach offers the "best of both worlds": bypassing the artificial saturation limits of the stand-alone LSTM while maintaining higher overall predictive accuracy than a purely conceptual model.

2.6 Data Augmentation

Building on the rationale established in Sect. 1, our data augmentation strategy is specifically designed to enrich the LSTM's training dataset with synthetic extreme events that maintain spatiotemporal and statistical consistency. The augmentation samples are constructed by pairing extreme meteorological forcings from the nsRWG (input features) with corresponding extreme discharge simulations (target labels). These target labels are extracted from either the conceptual HBV model or the hybrid model simulations. These synthetic samples represent a plausible, albeit unobserved, hydrological reality; however, they inherently carry the structural assumptions and potential biases of their respective source models. To maintain a strict separation between training and evaluation, the 100 nsRWG realizations are divided: 70 realizations are utilized exclusively for potential augmentation samples, while the remaining 30 realizations are reserved as an independent test set to evaluate the augmented models and their prediction limits (Sect. 3.3).

We sample time steps where the HBV or hybrid model simulations exceed the capacity of the standard, non-augmented LSTM when driven by the nsRWG forcings. We define two extremity thresholds based on the baseline LSTM's behavior under nsRWG forcing:

1. **Simulated Maximum:** Targets exceeding the highest discharge simulated by the base LSTM.
2. **Theoretical Limit:** Targets exceeding the calculated mathematical upper bound of the base LSTM.

Applying these two thresholds across the two target-generating models (HBV and hybrid) results in four distinct experimental sets for data augmentation (Tab. 1).

For each identified time step, a sample is constructed comprising a 365-day meteorological input sequence ending on that day, paired with the corresponding simulated discharge target. These synthetic sequences are added to the full histori-



Table 1. Overview of experimental setups, training datasets, and corresponding evaluation sections.

Model	Training Forcing	Training Targets	Augmentation Threshold	Section
LSTM baseline	Observations	Observations	–	Sect. 3.1, 3.2
HBV baseline	Observations	Observations	–	Sect. 3.1, 3.2
Hybrid model baseline	Observations	Observations	–	Sect. 3.1, 3.2
LSTM _{HBV > Theo. Limit}	Obs. + Synthetic	Obs. + HBV sim.	Theoretical Limit	Sect. 3.3
LSTM _{HBV > Sim. Max.}	Obs. + Synthetic	Obs. + HBV sim.	Simulated Maximum	Sect. 3.3
LSTM _{Hybrid > Theo. Limit}	Obs. + Synthetic	Obs. + Hybrid sim.	Theoretical Limit	Sect. 3.3
LSTM _{Hybrid > Sim. Max.}	Obs. + Synthetic	Obs. + Hybrid sim.	Simulated Maximum	Sect. 3.3

cal CAMELS-DE training dataset. During model training, the fraction of augmented samples is randomly shuffled into the training batches, rendering them indistinguishable from the historical observations to the neural network.

3 Results

This section presents the evaluation of the LSTM, the HBV model, and the hybrid model. Table 1 provides an overview of all experimental setups and their respective training and data augmentation configurations. We first assess baseline model performances using observed data from the CAMELS-DE dataset (Sect. 3.1), followed by an analysis of how the models behave under the extreme meteorological forcing conditions provided by the weather generator (Sect. 3.2). Building on these initial findings, we apply the data augmentation strategies described in Sect. 2.6, incorporating additional synthetic extreme events into the LSTM’s training data. The performance of these augmented LSTMs is then evaluated on both the observed CAMELS-DE data and an independent set of weather generator realizations. Finally, we demonstrate the practical implications of the data augmentation strategies by simulating the catastrophic July 2021 flood event, including the devastating Ahrthal flood (Sect. 3.3).

As described in Sect. 2.1, discharge data availability limits the number of catchments for model comparison from 180 to 167. In a further filtering step, we restrict our analysis to catchments where all three base models achieve a Nash-Sutcliffe Efficiency (NSE) ≥ 0.6 , yielding a final subset of 152 catchments. Applying this performance threshold ensures that the samples subsequently used for data augmentation originate from catchments where the baseline models provide reliable simulations.

3.1 Models tested on observed data

To establish a baseline model performance, the three models are evaluated over a 21-year test period using observed data from the CAMELS-DE dataset (2000 – 2020). Based on the NSE, the data-driven approaches generally outperform the conceptual HBV model (Fig. 2a). The LSTM achieves the highest overall performance with a median NSE of 0.90, closely followed by

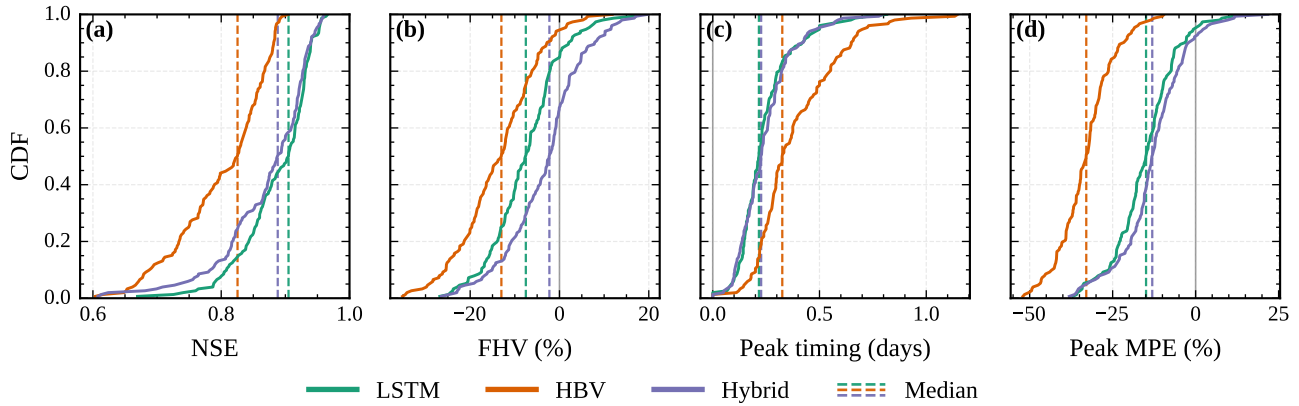


Figure 2. Empirical cumulative distribution functions of NSE and peak-flow performance metrics across CAMELS-DE catchments for the HBV, LSTM, and hybrid models in the observed test period (2000-2020). Panels show (a) Nash-Sutcliffe efficiency (NSE), (b) peak flow error (FHV), (c) Peak Timing error, and (d) Peak mean percentage error (MPE); all panels share the same y-axis (cumulative probability). Solid colored curves denote model-specific distributions, and dashed vertical lines indicate the corresponding model medians. The vertical black dashed line marks the reference value at 0 (ideal value for FHV, Peak Timing, and Peak MPE).

the hybrid model at 0.89. In contrast, the HBV model achieves a median NSE of 0.83 for the 152 selected catchments. While the LSTM and hybrid models perform similarly well in the highest-scoring catchments, the hybrid model exhibits a wider spread towards lower NSE values, indicating that the hybrid model performs worse than the pure LSTM in a subset of the more challenging catchments. We further evaluate the simulations using high- and peak-flow specific metrics (Fig. 2b-d). For the high-flow volume bias (FHV, Yilmaz et al., 2008), which represents peak flow bias for the highest 2 % of discharges, the hybrid model performs best with a median value of -2.26 %. This notably outperforms both the LSTM (-7.54 %) and the HBV model (-13.01 %). Regarding the timing of peak flows, the LSTM and hybrid models show similar capabilities, missing the observed peaks by a median lag of 0.22 and 0.23 days, respectively. The HBV model shows a larger lag, missing peaks by a median of 0.33 days. For the peak mean percentage error (Peak MPE), the hybrid model again demonstrates a slight advantage over the LSTM, yielding a median error of -13.06 % compared to -14.93 % for the LSTM, while the HBV again performs worst with a median error of -32.95 %.

Based on this standard metrics-based comparison on the observed record, we conclude that while the LSTM yields the best overall simulation quality (as measured by the NSE over 21 years), the hybrid model demonstrates superior performance for high- and peak-flows. The conceptual HBV model consistently achieves the lowest performance across all evaluated metrics. These baseline findings align with recent literature (e.g., Acuña Espinoza et al., 2025; Feng et al., 2022), which similarly highlights the strong overall performance of pure LSTMs alongside the specific advantages of hybrid architectures for simulating extreme flows.



3.2 Model limits under observed and synthetic extreme forcings

To evaluate the physical plausibility and extrapolation limits of the models, we compare their high-flow responses under historical conditions to those under unprecedented synthetic extremes. For the latter, we utilize the nsRWG dataset, whose 100 realizations provide 100 to 250 times more extreme precipitation events of varying durations (≥ 100 -year return period) than the observed CAMELS-DE record (Table A1). Because this synthetic meteorology lacks corresponding observed discharge targets, we evaluate the models based on physical plausibility, specifically examining whether simulated discharge scales appropriately with increasing precipitation. Figure 3 illustrates this by plotting the 250 highest-magnitude discharge simulations for each catchment against their Antecedent Precipitation Index (API, Kohler and Linsley, 1951), alongside the calculated theoretical upper limit of the trained LSTM (Kratzert et al., 2024). The API is used as a proxy for total multi-day preceding rainfall and catchment wetness, computed over a 14-day window with a decay constant of 0.9.

When forced with the observed CAMELS-DE meteorological record in the testing period (Fig. 3a-c), all three models exhibit similar, plausible relationships between API and peak discharge. In this observational space, the LSTM functions without issue; its maximum simulated discharge remains well below its calculated theoretical limit and the expected general trend of increasing simulated discharge with higher precipitation intensities is evident. However, the model responses diverge significantly when forced with the extreme precipitation events generated by the nsRWG, revealing clear patterns of saturation in case of the LSTM (Fig. 3d). In line with previous studies (Baste et al., 2025; Acuña Espinoza et al., 2025), the HBV and hybrid models continue to exhibit physically plausible behavior under extreme forcing, with a trend of simulated discharge increasing continuously as API increases, showing no apparent upper boundary (Fig. 3e-f). While this continuous scaling confirms that the models do not artificially saturate, the accuracy of the simulated peak magnitudes cannot be verified due to the lack of corresponding observations. In contrast, the standard LSTM hits a hard simulation ceiling that it cannot breach. Despite the presence of much higher precipitation forcings, the LSTM's maximum simulated discharge caps at 49.1 mm d^{-1} . Notably, this practical ceiling is reached even below the theoretical limit of 53.7 mm d^{-1} , stressing the extrapolation constraint of the purely data-driven model under unprecedented extreme events.

3.3 Impact of data augmentation on extreme flow simulations

Table 2 summarizes the characteristics of the four data augmentation sets, generated based on the strategies detailed in Sect. 2.6. Depending on the threshold and the source model used, these sets provide between 62 and 632 additional extreme samples for training, corresponding to the points above the thresholds in Figure 3e-f. Because the hybrid model tends to simulate higher peak flows than the conceptual HBV model under extreme nsRWG forcing, its respective augmentation sets yield a larger number of high-magnitude extreme samples.

We first evaluate the augmented LSTMs on the observed CAMELS-DE test period to calculate standard performance metrics and allow for a direct comparison against the baseline LSTM. As shown in Table 2, the addition of the augmentation samples slightly improve all evaluation metrics, including NSE and peak-flow specific measures, across all four augmentation sets.

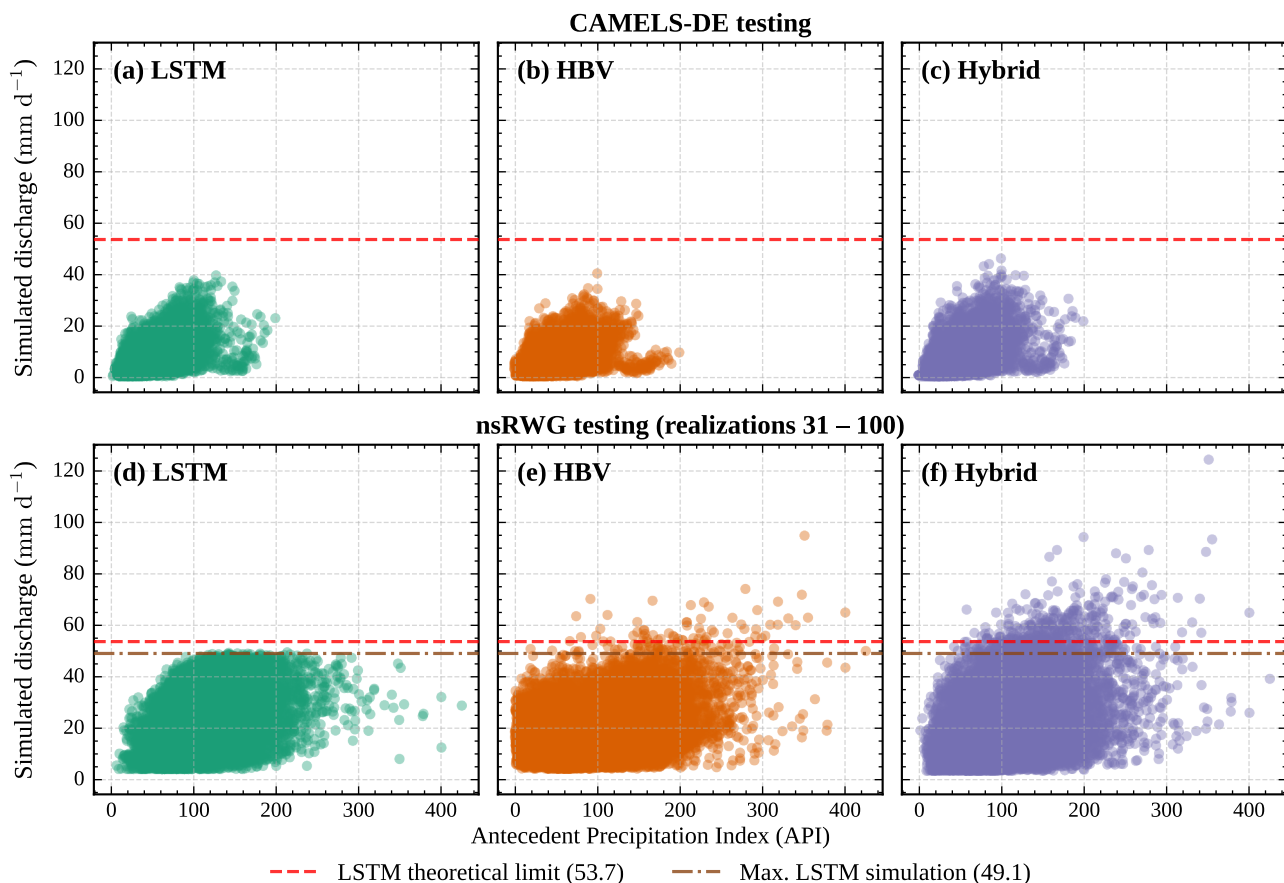


Figure 3. Relationship between the Antecedent Precipitation Index (API) and peak simulated discharge (mm d^{-1}) for the 250 largest simulated values per catchment. The top row (a-c) shows model responses driven by the observational CAMELS-DE dataset, while the bottom row (d-f) shows responses driven by the nsRWG weather generator data. Columns correspond to the standalone LSTM, conceptual HBV, and hybrid models, respectively. The horizontal red dashed line denotes the theoretical limit of the LSTM. For the weather generator forcing (d-f), the horizontal brown dash-dotted line additionally indicates the maximum absolute discharge simulated by the LSTM across the dataset. These horizontal lines also depict the sampling thresholds for the data augmentation experiments.

While data augmentation generally increases baseline performance, no single sampling strategy stands out as significantly superior within this observed data space.

We then test the augmented models on the remaining 30 nsRWG realizations that are not used for data augmentation sampling (Table 2 and Fig. 4). Compared to the baseline LSTM, the theoretical limits for all augmented models are raised substantially, with the sets derived from the hybrid model inducing the largest increases. Most notably, training with hybrid model targets exceeding the theoretical limit increases the theoretical limit to 92.65 mm d^{-1} (+72.6 %) and raises the maximum simulated discharge to 72.37 mm d^{-1} (+47.5 %). Figure 4f-j demonstrates that while the baseline LSTM hits a strict simulation ceiling

Table 2. Model limits and performance metrics comparing the baseline LSTM with four targeted data augmentation strategies. The augmentation sets consist of additional training samples where the HBV or hybrid models simulated discharges exceeding the baseline LSTM’s theoretical limit or maximum simulated value. The theoretical limit and maximum LSTM simulation are derived by forcing the models with the synthetic nsRWG sequences, whereas predictive performance metrics (NSE, FHV, Peak Timing, Peak MPE) are calculated against observed targets over the historical CAMELS-DE testing period. Best performing values in each column are highlighted in bold.

Model	Augmentation Samples	nsRWG		CAMELS-DE			
		Theoretical Limit (mm d ⁻¹)	Max. Simulated (mm d ⁻¹)	NSE	FHV (%)	Peak Timing (days)	Peak MPE (%)
LSTM baseline	0 (1,517,235)	53.67	49.06	0.905	-7.54	0.22	-14.93
LSTM _{HBV} > Theo. Limit	62	78.60	65.10	0.907	-6.41	0.21	-13.97
LSTM _{HBV} > Sim. Max.	130	83.99	62.17	0.915	-7.28	0.20	-14.04
LSTM _{Hybrid} > Theo. Limit	312	92.65	72.37	0.915	-7.46	0.20	-14.46
LSTM _{Hybrid} > Sim. Max.	632	89.05	70.40	0.912	-6.54	0.20	-14.00

under extreme nsRWG forcing, the augmented LSTMs react much stronger to high API values. Although an upper theoretical ceiling must mathematically still exist, it has shifted so far upward that it is no longer reached by the extremes resulting from the nsRWG-driven LSTM simulations.

Finally, to test the value of this approach on a real-world, high-impact event, we simulate the catastrophic July 2021 flood across three heavily affected catchments: Altenahr, Kreuzberg, and Opladen (Fig. 5). Across all three gauges, the baseline LSTM substantially underestimates the observed daily discharge peaks. The data augmentation strategies consistently improve the simulations, generating higher peak discharges than the baseline models. Although not shown in the plot, augmentation variants using the theoretical limit as a threshold performed very similarly to those based on the simulated maximums. At the Kreuzberg and Opladen gauges, while the augmented LSTMs simulate slightly higher peak flows, they still largely under-
 285 derestimate the absolute magnitude of the extreme event. In contrast, the improvement at the Altenahr gauge is much more pronounced: here, the LSTM augmented with hybrid-model targets captures the flood peak with only minor underestimation and a reduced ensemble spread. Notably, this reduction in peak error at Altenahr occurred even though the peak simulated by the baseline LSTM is well below its theoretical limit, and the data augmentation sets contained no data from the Altenahr catchment itself.

290 4 Discussion

The discussion section addresses the research questions formulated in Sect. 1.

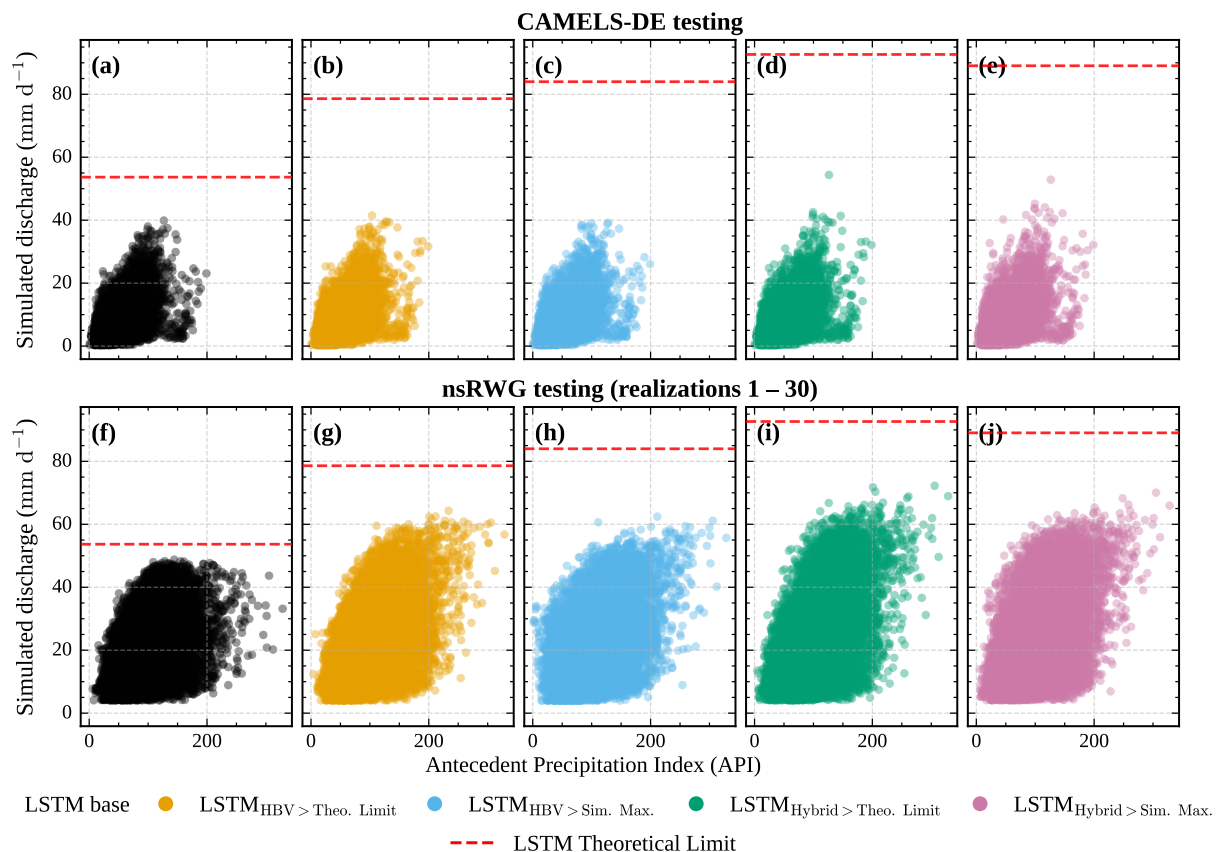


Figure 4. Relationship between antecedent precipitation index (API) and simulated peak discharge for the baseline LSTM and four data-augmentation variants (HBV > Theo. Limit, HBV > Sim. Max., Hybrid > Theo. Limit, Hybrid > Sim. Max.). The top row (a-e) shows CAMELS-DE testing and the bottom row (f-j) shows nsRWG testing using realizations 1-30. In each panel, points represent the top 250 events per catchment. Dashed red horizontal lines indicate the ensemble-median theoretical LSTM output limit.

1. How do the different hydrological model architectures respond to unprecedented extreme meteorological forcings from the weather generator?

Standard evaluations on historically observed data consistently show that purely data-driven LSTMs outperform conceptual models like HBV in overall metrics (Lees et al., 2021; Frame et al., 2022), with hybrid models sometimes excelling at high-flow simulations (Acuña Espinoza et al., 2025; Song et al., 2026)—a pattern replicated in our study (Section 3.1, Fig. 2). Furthermore, when evaluating the highest discharge simulations under these historical forcings (Fig. 3a-c), no fundamental simulation limitations are apparent for the LSTM. If we were to base model selection solely on this historical data space, the HBV model would not be chosen; instead, the LSTM would be selected for general simulation quality, and the hybrid model could potentially serve as an expert for high-flow simulations. Based on this historical testing perspective, one might

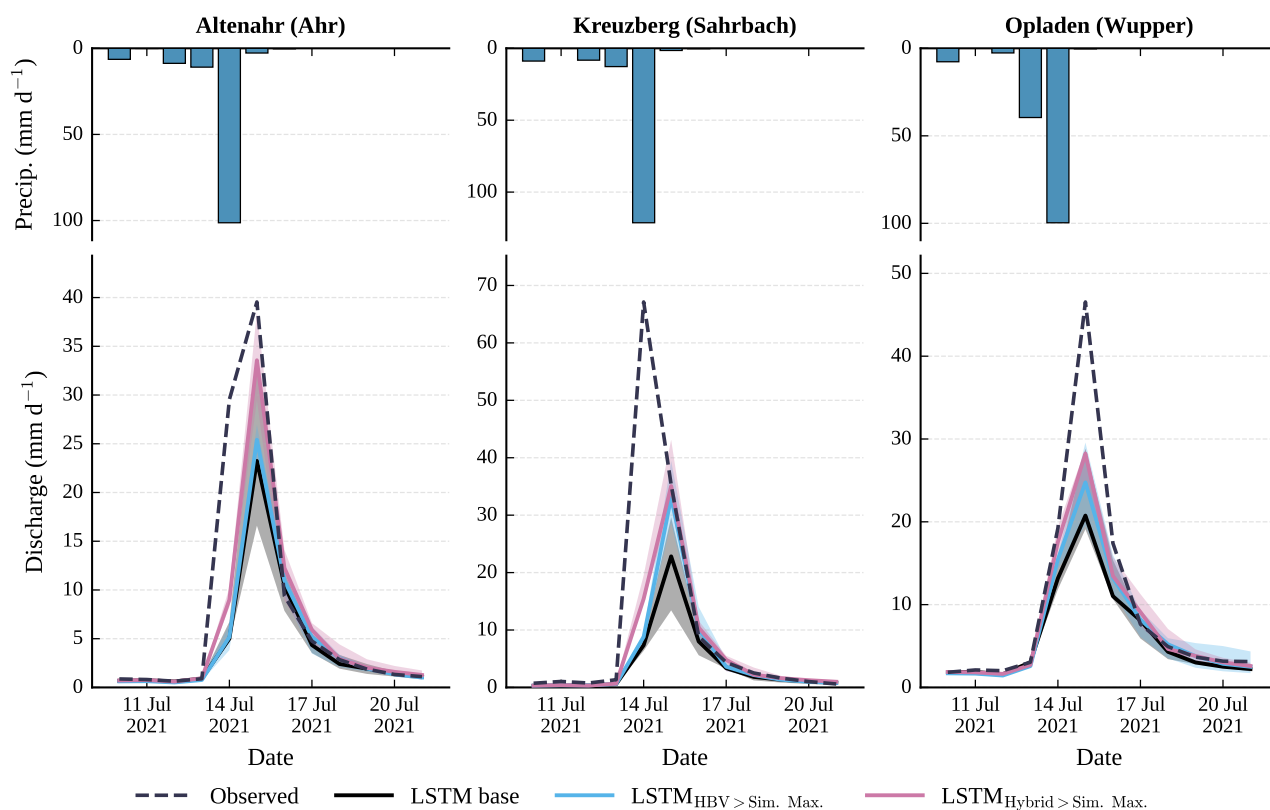


Figure 5. Precipitation and discharge during the July 2021 flood event (10-21 July 2021) for the catchments Altenahr, Kreuzberg, and Opladen. The upper panels show daily precipitation, and the lower panels compare observed discharge with simulations from the baseline LSTM and two augmented LSTM variants based on HBV-derived and hybrid-model-derived samples. Lines show ensemble median discharge and shaded envelopes show the minimum-maximum ensemble spread across seed members.

reasonably assume that the theoretical prediction limits of the LSTM no longer pose a practical constraint, especially since regional training can raise this upper bound (Kratzert et al., 2024).

However, historical records, which in our case are limited to the testing period from 2000 to 2020, contain only a fraction of possible extreme events across a limited number of catchments. To test the true boundaries of the models, we utilize a weather generator to produce diverse, unprecedented, yet statistically plausible precipitation extremes, encompassing single-day and multi-day events under various antecedent conditions (Section 3.2, Table A1). When forcing the models with these synthetically generated extremes, the fundamental limitations of the LSTM become apparent. Under these extreme forcings, the LSTM hits a hard ceiling that it cannot overcome (Fig. 3d). In our simulations, the maximum discharge generated by the LSTM stalled at an absolute value of 49.1 mm d^{-1} , despite the theoretical capacity, calculated based on the trained model weights, being 53.7 mm d^{-1} . This reiterates that the LSTM output saturates even below its absolute mathematical limit. While Baste et al. (2025) demonstrated a similar limitation using artificial design precipitation events, our study proves that this



behavior also occurs under the temporally consistent time series generated by the weather generator. In contrast, the HBV and hybrid models exhibit physically plausible behavior under the same extreme forcings, showing continuous increases in simulated discharge in response to increasing precipitation intensities (Fig. 3e-f). In these models, additional precipitation input must strictly be accounted for (e.g., routed to storage or translated to runoff), whereas the standard LSTM architecture lacks a structural constraint to prevent it from dissipating mass. In the context of unprecedented extremes, our results suggest that this physical constraint is a critical advantage over purely data-driven approaches.

Our results align with recent findings in the broader meteorological domain, where Zhang et al. (2026) showed that physics-based models consistently outperform AI-based models in forecasting record-breaking extreme weather events. In hydrology, evaluations of physical constraints on historical high flows yield mixed results: while some studies report improved high-flow performance for hybrid models and mass-conserving architectures (Acuña Espinoza et al., 2025; Hoedt et al., 2021), others find that standard LSTMs still perform better in this regard (Frame et al., 2022). However, standard metrics such as FHV evaluate the broader upper percentiles of the flow record, revealing little about model behavior at the absolute highest, unprecedented peaks where extrapolation limits are actually reached. Our results suggest that the advantage of physical constraints becomes pronounced under unprecedented conditions. Although rare, such events are statistically inevitable and projected to intensify under climate change (Blöschl et al., 2019; Brunner et al., 2021; Astagneau et al., 2026). Accurately simulating these magnitudes is critical for operational flood forecasting and design discharge estimation, where we must increasingly account for seemingly "impossible" floods (Montanari et al., 2024). Our study shows that relying solely on unconstrained data-driven models for such tasks is risky, unless their inherent extrapolation limits are explicitly addressed to ensure that extreme precipitation translates into proportionate runoff. Synthetically generated weather forcings provide a framework to reliably stress-test hydrological models, particularly for exposing the structural boundaries of standard LSTMs.

2. Can targeted data augmentation mitigate the inherent limitations of the LSTM?

After establishing that the standard LSTM hits a strict simulation ceiling under extreme forcings, a limitation the conceptual HBV and hybrid models do not suffer from, we leverage the capacity of the physics-constrained models to generate targeted augmentation samples. Specifically, we use the weather generator sequences as input features, paired with the corresponding discharge simulations from the HBV and hybrid models as targets. Although these synthetic extremes represent a small fraction of the overall dataset (adding a maximum of 632 events to the approximately 1.5 million historical CAMELS-DE training samples), their inclusion has a profound impact on the network's behavior.

As demonstrated in Table 2, all four data augmentation strategies successfully raise both the LSTM's theoretical limit and its maximum simulated values when tested on unseen weather generator realizations. Furthermore, the inclusion of the synthetic extremes in the training data improves high- and peak-flow metrics, and in most cases, overall NSE, when evaluated on the observed CAMELS-DE data. While all four strategies prove to be effective, augmentation utilizing targets from the hybrid model raise the theoretical limit more than those from the HBV model. This likely occurs because the hybrid model simulates higher magnitude discharges that were used for data augmentation, pushing the LSTM's predictive boundary further upward.



345 The practical benefit of this shifted boundary becomes evident when analyzing the 250 most extreme events per catchment in the CAMELS-DE testing data (Fig. 4a-e). The augmented LSTMs gain substantial headroom toward their newly raised theoretical limits. Even though the historical CAMELS-DE forcings rarely push the models to their mathematical boundaries, the augmented models, particularly those trained on hybrid model targets, simulate higher peak flows, with one instance surpassing the original baseline LSTM's theoretical maximum. More importantly, when testing these augmented LSTMs against the much
 350 more extreme, held-out weather generator sequences, the previously observed hard ceiling is no longer observable (Fig. 4f-j). The raised theoretical maximum provides ample headroom, allowing the augmented LSTMs to simulate significantly higher discharges without showing any signs of artificial saturation. Visual assessments of simulated hydrographs revealed that the augmented LSTM does not learn to mimic the responses of the HBV and hybrid models. Instead, it leverages its expanded headroom, while its core rainfall-runoff response still mirrors that of the non-augmented LSTM.

355 To validate these findings under severe real-world conditions, we test the models on the catastrophic July 2021 flood event across the Altenahr, Kreuzberg, and Opladen catchments (Fig. 5). Across all gauges, the baseline LSTM underestimates the peak discharge of this unprecedented event. While the augmented LSTMs still underestimated the absolute peaks, the magnitude of the error as well as the ensemble spread are reduced. This improvement is particularly pronounced at the Altenahr gauge using the hybrid-model-augmented LSTM, leading to a better simulation of the extreme hydrograph. However, the results at
 360 the Kreuzberg and Opladen gauges highlight that the simulation of unprecedented extreme events remains a challenge that data augmentation alone cannot entirely solve. A key limiting factor is the temporal resolution of our simulations. To reliably capture such rapid flood dynamics, especially the rising limb of the hydrograph, hourly data would likely be necessary (Dolich et al., 2026b).

Further exploration of these extreme responses revealed that the actual maximum discharge simulated by the standard LSTM
 365 often stalled well below its absolute mathematical limit. Using the design precipitation methodology of Baste et al. (2025), we observed varying catchment-specific predictive limits that frequently fell far below the global limits shown in Fig. 3. Our targeted data augmentation techniques successfully raised these local limits alongside the absolute theoretical boundaries of the LSTM. This increase in predictive limits occurred across the entire region, not just in the specific catchments that contributed data augmentation samples. This is also the case for the evaluated catchments Altenahr and Opladen, which did not
 370 contribute to the synthetic augmentation dataset. The improved LSTM simulations for this historic event are therefore the direct result of regional data augmentation effects. This demonstrates that the LSTM successfully leverages its expanded boundaries to generalize the relationship between extreme forcings and high discharge, rather than simply memorizing the augmented catchments.

Our results demonstrate that targeted data augmentation, pairing synthetic meteorological extremes with discharge targets
 375 from physically informed models, can effectively mitigate the practical impact of the LSTM's theoretical limit. By using mass-conserving models as target generators during extreme synthetic events, we can actively push the boundaries of the purely data-driven model. This approach simultaneously improves general high- and peak-flow performance without compromising overall model accuracy, as shown by stable or even slightly increased NSE scores. In contrast, Kenne et al. (2026) showed that untargeted data augmentation techniques (such as noise addition or resampling) offer only marginal benefits when applied to



380 large, already diverse hydrological datasets. Together, these results suggest that while the LSTM does not suffer from a general data deficit, it is specifically constrained by a lack of physically plausible samples in the extreme right tail of the distribution. Data augmentation in the field of hydrology is therefore most promising when directly targeting this deficit. A possible alternative pathway would be defaulting to a hybrid model for extreme event simulation as a more straightforward solution, since it inherently avoids saturation and excels at high flows. However, the standard LSTM remains advantageous in several aspects: 385 its architectural simplicity reduces uncertainties associated with model assumptions (Poudel and Steinschneider, 2026), its computational efficiency, its highly active community development, particularly for efficient large-scale and high-frequency data handling (e.g., Gauch et al., 2025; Acuña Espinoza et al., 2026b), and the ease with which its outputs can be made probabilistic, a crucial aspect for simulating extreme floods (Baste et al., 2026). While the use of non-saturating functions, such as in the xLSTM (Beck et al., 2024), might appear to be a straightforward remedy to the ceiling issue, it did not resolve the saturation problem and even reduced the theoretical limit, as shown in Baste et al. (2025). Therefore, the exploration of pathways 390 that mitigate the LSTM's theoretical limits, such as our proposed targeted data augmentation, remains relevant for advancing data-driven hydrology. It must be acknowledged, however, that the theoretical limit of a standard LSTM is an inherent architectural property that can be shifted upward into regions of less practical concern, but cannot be completely eliminated. An open question for future research is whether a sophisticated weather generator is necessary to achieve this shift. Simpler 395 methods, such as utilizing artificial design storms or applying mathematical multipliers to historical targets, could potentially yield similar boundary-shifting effects. Yet, we hypothesize that the spatial and temporal consistency uniquely supplied by the weather generator ensures the neural network learns realistic rainfall–runoff dynamics under extreme conditions, rather than merely mathematically hacking its upper boundary.

3. What are the implications of these findings for large-sample, regional LSTM training strategies?

400 The results from our targeted data augmentation experiments highlight a crucial consideration for the standard train-test-validate data splitting procedure in large-sample hydrology. We demonstrate that adding merely 62 to 632 synthetic extreme samples, a small fraction compared to the approximately 1.5 million total CAMELS-DE training samples, substantially shifts the theoretical limit of the LSTM, while also improving overall and high-flow specific model performance (Table 2). This impact underscores that in order to develop a model that performs well during high-flow situations and delays artificial saturation, 405 the training dataset must encompass as many extreme, high-magnitude historical flood events as possible.

This insight adds nuance to recent discussions regarding sampling strategies and dataset sizes for regional deep learning models. For instance, Kratzert et al. (2024) concluded that increasing the number of catchments in the training data generally improves model performance. While this might hold true for overall predictive quality, our findings suggest that for robust high-flow simulations, ensuring the inclusion of the highest flood events is more critical than simply maximizing the sheer 410 number of catchments. Furthermore, Heudorfer and Loritz (2026) concluded that smart sampling does not significantly outperform random sampling, suggesting that utilizing just 10 % of available training data can yield strong performance. Although the authors explicitly noted their findings were based on NSE evaluations over multiple years, our results warrant caution against utilizing reduced, randomly sampled training sets when aiming to simulate extremes. Because rare, catastrophic floods



constitute such a minute fraction of the historical record, a small random sample is likely to miss them entirely. Omitting
 415 these specific extremes from the training phase directly constrains the LSTM's theoretical limit, thereby rendering it inherently
 incapable of simulating rare, high-magnitude floods.

To further investigate if the saturation issue could be mitigated through a larger regional training dataset, we calculate the
 theoretical limit of the benchmark CAMELS-DE LSTM (Loritz et al., 2024), which is trained on the entire dataset of 1,582
 German catchments rather than the regional subset of 180 catchments covered by the weather generator domain. As expected
 420 from the findings of Kratzert et al. (2024), this expansion of the training set raises the baseline ceiling, yielding a median
 theoretical limit of 74.8 mm d^{-1} (standard deviation 0.78) across ten model seeds. However, this limit still falls below the limit
 of the LSTM trained on just the regional subset combined with our data augmentation strategies. This discrepancy highlights
 that the relative occurrence of extreme events in the training data is just as crucial as their absolute number. While training
 on all of Germany introduces more high-magnitude historical events in absolute terms, the ratio of extreme to normal events
 425 remains small. In contrast, targeted data augmentation explicitly increases the relative density of these critical training signals.
 Ultimately, expanding the historical training set scales the problem rather than eliminating it. While including all of Germany's
 catchments raises the theoretical limit, evaluating this model against a hypothetical, Germany-wide weather generator would
 inevitably push the LSTM against its newly elevated ceiling. The fundamental architectural constraint remains; the problem
 is merely shifted to a higher magnitude. At the absolute extreme tail, where historical data is exhausted, data augmentation
 430 remains the only pathway to provide the necessary high-flow training signals.

Our study emphasizes the need to evaluate models based on their intended use case (Yilmaz et al., 2008; Beven and Lane,
 2022). For extreme flood forecasting or critical infrastructure design, standard historical performance metrics (e.g., NSE or
 peak flow error) can provide a false sense of security because they fail to expose the LSTM's structural limitations. Before
 operational deployment, purely data-driven models must be explicitly stress-tested against unprecedented extremes, whether
 435 through sophisticated weather generators or accessible alternatives like synthetic design precipitation (Baste et al., 2025).

5 Conclusions

In this study, we demonstrate that synthetically generated meteorological extremes provide an effective framework for stress-
 testing hydrological models. Using a weather generator, we reiterate that standard LSTMs suffer from artificial output saturation
 under unprecedented extremes, a structural limitation hidden during standard historical evaluations. Targeted data augmentation
 440 using synthetic extreme meteorological forcings together with targets from mass-conserving models shifts this theoretical limit
 upward, mitigating but not entirely eliminating the architectural ceiling. Training on larger datasets (e.g., the entire CAMELS-
 DE dataset) also raises the theoretical limit by introducing more historical extremes, but not as effectively as our data augmen-
 tation strategies, which actively increase the relative density of extreme events. Augmenting a Germany-wide LSTM with a
 nationwide weather generator represents a promising pathway to push the predictive boundaries even further. Furthermore, our
 445 results highlight the potential of hybrid modeling when simulating extreme events. The hybrid model evaluated here showed
 no signs of output saturation under unprecedented extremes and yields the best overall high- and peak-flow performance. As



the community continues to develop technical improvements for the standard LSTM, such as probabilistic models or data assimilation strategies, these advancements should be adapted for hybrid architectures to combine physics-based robustness with data-driven efficiency. Ultimately, consistent with numerous prior studies, the standard LSTM demonstrated strong over-
all simulation performance. However, practitioners must be aware of its architectural limitations. Because the LSTM's global theoretical limit can be easily calculated from its trained weights, this value should be computed and monitored continuously, while remaining aware that the limit in some catchments can be much lower. Where practically feasible, actively expanding these boundaries through targeted data augmentation can significantly mitigate this vulnerability. If an LSTM is deployed operationally for flood forecasting or design flood estimation, we recommend running a mass-conserving or hybrid model in parallel as a structural safeguard. This is particularly critical when forecasted forcings approach unprecedented magnitudes or when the LSTM nears its theoretical limit. Being aware of the LSTM's limit problem is essential; without this awareness, purely data-driven modeling can lead to the potentially catastrophic underestimation of extreme events.

Code and data availability. The historical meteorological and hydrological data used in this study are provided by the CAMELS-DE dataset (Dolich et al., 2025) and its hourly counterpart, CAMELS-DE-1h (Dolich et al., 2026a). The catchment-aggregated synthetic time series generated by the non-stationary regional weather generator (nsRWG) are available from the corresponding author upon request. The source code used to train the LSTM and the hybrid model is available in the *Hy2DL* package (Acuña Espinoza et al., 2026a) at <https://github.com/eduardoAcunaEspinoza/Hy2DL> (last access: 9 September 2026). The code for calibrating the conceptual HBV model is available in the *HyMC* package (Acuña Espinoza et al., 2024) at <https://github.com/eduardoAcunaEspinoza/HyMC> (last access: 9 September 2026).



Table A1. Comparison of exceedance counts for observed CAMELS-DE and synthetic weather-generator (nsRWG) precipitation events across return-period classes of 25–50, 50–100, and >100 years. Exceedance counts were derived from KOSTRA-DWD-2020 return-period thresholds applied to rolling precipitation sums. Ratios indicate the multiplicative increase in rare-event frequency in the weather generator relative to the historical record.

Duration (days)	Return period (years)	CAMELS-DE		nsRWG		Ratio (nsRWG/CAMELS)
		Exceedances	Catchments	Exceedances	Catchments	
1	25–50	76	61	10062	180	132.4
1	50–100	43	41	6151	178	143.0
1	>100	31	31	3184	166	102.7
2	25–50	95	72	18786	180	197.7
2	50–100	53	49	10858	180	204.9
2	>100	38	37	5312	180	139.8
4	25–50	73	60	28376	180	388.7
4	50–100	48	44	15662	180	326.3
4	>100	29	29	7112	180	245.2
7	25–50	140	90	35338	180	252.4
7	50–100	64	56	18277	180	285.6
7	>100	34	34	7617	180	224.0

Appendix A: Comparison of CAMELS-DE precipitation to nsRWG precipitation

465 We compare rainfall exceedances in the historical CAMELS-DE record to the nsRWG realizations, using catchment-specific
 return-period thresholds derived from the gridded KOSTRA-DWD-2020 dataset (Junghänel et al., 2023). Driven primarily by
 the dataset’s volume of 100 parallel realizations and the unbounded nature of the weather generator’s marginal distributions,
 the nsRWG produces substantially more events in the 25–50, 50–100, and >100 year return-period classes. Depending on event
 duration, absolute event counts exceed the single CAMELS-DE record by a factor of 100 to nearly 400 (Table A1). This shows
 470 that the synthetic data include a much wider and more extreme precipitation tail, suitable for hydrological model stress testing
 and targeted augmentation of rare-event samples.



Author contributions. RL and SV proposed the initial research idea for the study. AD and PR conducted the model training and testing as well as the formal analysis. VDN produced the weather generator realizations. AD prepared the original draft of the manuscript. All authors reviewed and edited the draft, discussed the results, and approved the final version of the manuscript.

475 *Competing interests.* At least one of the (co-)authors is a member of the editorial board of Hydrology and Earth System Sciences. The peer-review process will be guided by an independent editor, and the authors also have no other competing interests to declare.

Acknowledgements. The authors acknowledge the support of the state of Baden-Württemberg through bwHPC. Preliminary modeling and conceptual exploration for this research were conducted as part of the Master's thesis of PR.

480 *Financial support.* The results were funded by the German Federal Ministry of Research, Technology and Space (BMFTR) as part of the project “AI-based flood prediction in small river basins in Germany” (KI-HopE-De, 16IS24088A). We are grateful to the Volkswagen Foundation for funding the project “Invigorating Hydrological Science and Teaching: Merging Key Legacies with New Concepts and Paradigms” (ViTamins, 0069685). Viet Dung Nguyen has been funded by the German Federal Ministry of Research, Technology and Space (BMFTR) in the framework of the project FLOOD (project no. 01LP2324E) as a part of the ClimXtreme Research Network on Climate Change and Extreme Events within the framework programme Research for Sustainable Development (FONA3).



485 References

- 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>, 2024.
- Acuña Espinoza, E., Loritz, R., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., and Ehret, U.: Analyzing the generalization capabilities of a hybrid hydrological model for extrapolation to extreme events, *Hydrology and Earth System Sciences*, 29, 1277–1294, <https://doi.org/10.5194/hess-29-1277-2025>, 2025.
- Acuña Espinoza, E., Álvarez Chaves, M., Dolich, A., Heudorfer, B., J, A. M., Wiktorja, and Mälicke, M.: eduardoAcunaEspinoza/Hy2DL: v2.0.1, <https://doi.org/10.5281/ZENODO.22083944>, 2026a.
- Acuña Espinoza, E., Álvarez Chaves, M., Kratzert, F., Klotz, D., Gauch, M., Lang, R., Elfgang, D., Dolich, A., Loritz, R., and Ehret, U.: Everything everywhere all at once: A single-cell LSTM network unifying multi-frequency, missing data, and discharge assimilation for robust operational flood forecasting, <https://doi.org/10.5194/egusphere-2026-4885>, 2026b.
- 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.
- Astagneau, P. C., Wood, R. R., and Brunner, M. I.: Rethinking future flood hazard: Hourly data challenge daily flood projections in Alpine catchments, *Science Advances*, 12, eaed6012, <https://doi.org/10.1126/sciadv.aed6012>, 2026.
- Barnard, E. and Wessels, L. F. A.: Extrapolation and interpolation in neural network classifiers, *IEEE Control Systems*, 12, 50–53, <https://api.semanticscholar.org/CorpusID:20074078>, 1992.
- Baste, S., Klotz, D., Acuña Espinoza, E., Bardossy, A., and Loritz, R.: Unveiling the limits of deep learning models in hydrological extrapolation tasks, *Hydrology and Earth System Sciences*, 29, 5871–5891, <https://doi.org/10.5194/hess-29-5871-2025>, 2025.
- Baste, S., Lerch, S., Klotz, D., and Loritz, R.: The need for uncertainty: why probabilistic LSTMs are key to improving flood predictions and enabling learned warning rules, <https://doi.org/10.5194/egusphere-2026-469>, 2026.
- 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://doi.org/10.48550/ARXIV.2405.04517>, version Number: 2, 2024.
- Bergström, S. and Forsman, A.: Development Of A Conceptual Deterministic Rainfall-Runoff Model, *Hydrology Research*, 4, 147–170, <https://doi.org/10.2166/nh.1973.0012>, 1973.
- Beven, K.: A manifesto for the equifinality thesis, *Journal of Hydrology*, 320, 18–36, <https://doi.org/10.1016/j.jhydrol.2005.07.007>, 2006.
- Beven, K. and Lane, S.: On (in)validating environmental models. 1. Principles for formulating a Turing-like Test for determining when a model is fit-for-purpose, *Hydrological Processes*, 36, e14 704, <https://doi.org/10.1002/hyp.14704>, 2022.
- BKG: Verwaltungsgebiete 1:250 000 - Stand 01.01., Federal Agency for Cartography and Geodesy (BKG), <https://gdk.gdi-de.org/geonetwork/srv/api/records/93a98c5c-cf03-4a95-bf0a-54001fbf3949> (last access: 12 February 2026), https://opendata.dwd.de/climate_environment/CDC/grids_germany/hourly/hostrada/, 2023.
- Blöschl, G., Hall, J., Viglione, A., Perdighão, R. A. P., Parajka, J., Merz, B., Lun, D., Arheimer, B., Aronica, G. T., Bilibashi, A., Boháč, M., Bonacci, O., Borga, M., Čanjevack, I., Castellarin, A., Chirico, G. B., Claps, P., Frolova, N., Ganora, D., Gorbachova, L., Gül, A., Hannaford, J., Harrigan, S., Kireeva, M., Kiss, A., Kjeldsen, T. R., Kohnová, S., Koskela, J. J., Ledvinka, O., Macdonald, N., Mavrova-Guirguinova, M., Mediero, L., Merz, R., Molnar, P., Montanari, A., Murphy, C., Osuch, M., Ovcharuk, V., Radevski, I., Salinas, J. L.,



- Sauquet, E., Šraj, M., Szolgay, J., Volpi, E., Wilson, D., Zaimi, K., and Živković, N.: Changing climate both increases and decreases European river floods, *Nature*, 573, 108–111, <https://doi.org/10.1038/s41586-019-1495-6>, 2019.
- Brunner, M. I., Swain, D. L., Wood, R. R., Willkofer, F., Done, J. M., Gilleland, E., and Ludwig, R.: An extremeness threshold determines the regional response of floods to changes in rainfall extremes, *Communications Earth & Environment*, 2, 173, <https://doi.org/10.1038/s43247-021-00248-x>, 2021.
- Clerc-Schwarzenbach, F., Selleri, G., Neri, M., Toth, E., van Meerveld, I., and Seibert, J.: Large-sample hydrology – a few camels or a whole caravan?, *Hydrology and Earth System Sciences*, 28, 4219–4237, <https://doi.org/10.5194/hess-28-4219-2024>, 2024.
- Dao, T., Gu, A., Ratner, A., Smith, V., De Sa, C., and Re, C.: A kernel theory of modern data augmentation, in: Proceedings of the 36th international conference on machine learning, edited by Chaudhuri, K. and Salakhutdinov, R., vol. 97 of *Proceedings of machine learning research*, pp. 1528–1537, PMLR, <https://proceedings.mlr.press/v97/dao19b.html>, 2019.
- Dolich, A., Espinoza, E. A., Ebeling, P., Guse, B., Götze, J., Hassler, S., Hauffe, C., Kiesel, J., Heidbüchel, I., Mälicke, M., Müller-Thomy, H., Stölzle, M., Tarasova, L., and Loritz, R.: CAMELS-DE: hydrometeorological time series and attributes for 1582 catchments in Germany, <https://doi.org/10.5281/ZENODO.16755906>, 2025.
- Dolich, A., Acuña Espinoza, E., Ehret, U., Kraft, M., Bondy, J., Meuer, J., and Loritz, R.: CAMELS-DE-1h: hourly hydro-meteorological time series, weather forecasts, and attributes for 1611 catchments in Germany, <https://doi.org/10.5880/FIDGEO.2026.045>, 2026a.
- Dolich, A., Acuña Espinoza, E., Ehret, U., Kraft, M., Bondy, J., Meuer, J., and Loritz, R.: CAMELS-DE-1h: hourly hydro-meteorological time series, weather forecasts, and attributes for 1611 catchments in Germany, <https://doi.org/10.5194/essd-2026-289>, 2026b.
- Dolich, A., Ehret, U., and Loritz, R.: To Learn or to calibrate? Potentials and limitations of data-driven and conceptual hydrological models in a Germany-wide comparison, *Hydrologie und Wasserbewirtschaftung / BfG*. - 70.2026, pp. 4 MB, 17 Seiten, https://doi.org/10.5675/HYWA_2026.4_2, artwork Size: 4 MB, 17 Seiten Medium: application/pdf, 2026c.
- Droogers, P. and Allen, R. G.: Estimating Reference Evapotranspiration Under Inaccurate Data Conditions, *Irrigation and Drainage Systems*, 16, 33–45, <https://doi.org/10.1023/A:1015508322413>, 2002.
- 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, e2022WR032404, <https://doi.org/10.1029/2022WR032404>, 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.
- Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Cohen, D., and Gilon, O.: How to deal with missing input data, *EGUsphere*, pp. 1–21, <https://doi.org/10.5194/egusphere-2025-1224>, 2025.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., De Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.: The ERA5 global reanalysis, *Quarterly Journal of the Royal Meteorological Society*, 146, 1999–2049, <https://doi.org/10.1002/qj.3803>, 2020.
- Hettiarachchi, P., Hall, M. J., and Minns, A. W.: The extrapolation of artificial neural networks for the modelling of rainfall–runoff relationships, *Journal of Hydroinformatics*, 7, 291–296, <https://doi.org/10.2166/hydro.2005.0025>, 2005.



- Heudorfer, B. and Loritz, R.: Is smart sampling worth it? Impact of training data selection on the performance of LSTMs in streamflow prediction, *Hydrology Research*, 57, 501–517, <https://doi.org/10.2166/nh.2026.119>, 2026.
- 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.
- Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G. S., Hochreiter, S., and Klambauer, G.: MC-LSTM: Mass-Conserving LSTM, in: *Proceedings of the 38th International Conference on Machine Learning*, pp. 4275–4286, PMLR, <https://proceedings.mlr.press/v139/hoedt21a.html>, 2021.
- Houska, T., Kraft, P., Chamorro-Chavez, A., and Breuer, L.: SPOTting Model Parameters Using a Ready-Made Python Package, *PLOS ONE*, 10, e0145180, <https://doi.org/10.1371/journal.pone.0145180>, 2015.
- Iwana, B. K. and Uchida, S.: An empirical survey of data augmentation for time series classification with neural networks, *PLOS ONE*, 16, e0254841, <https://doi.org/10.1371/journal.pone.0254841>, 2021.
- Junghänel, T., Ostermüller, J., and Deutschländer, T.: Grids of return periods of heavy precipitation (design precipitation) over Germany (KOSTRA-DWD) Raster der Wiederkehrintervalle für Starkregen (Bemessungsniederschläge) in Deutschland (KOSTRA-DWD), <https://doi.org/10.5676/DWD/KOSTRA-DWD-2020>, 2023.
- Kenne, A. D., Hochreiter, S., and Klotz, D.: Exploring inductive transfer approaches for machine learning based rainfall-runoff modelling, *International Journal of River Basin Management*, pp. 1–23, <https://doi.org/10.1080/15715124.2026.2729475>, 2026.
- Kohler, M. A. and Linsley, R. K.: Predicting the runoff from storm rainfall, Research Paper 34, U.S. Department of Commerce, Weather Bureau, Washington, D.C., 1951.
- Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, *Water Resources Research*, 55, 11344–11354, <https://doi.org/10.1029/2019WR026065>, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2019WR026065>, 2019.
- Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., and Matias, Y.: Caravan - A global community dataset for large-sample hydrology, *Scientific Data*, 10, 61, <https://doi.org/10.1038/s41597-023-01975-w>, 2023.
- 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., Hauße, C., Heidbüchel, I., Kiesel, J., Mälicke, M., Müller-Thomy, H., Stölzle, M., and Tarasova, L.: CAMELS-DE: hydro-meteorological time series and attributes for 1582 catchments in Germany, *Earth System Science Data*, 16, 5625–5642, <https://doi.org/10.5194/essd-16-5625-2024>, 2024.
- López-Chacón, S. R., Salazar, F., and Bladé, E.: Combining Synthetic and Observed Data to Enhance Machine Learning Model Performance for Streamflow Prediction, *Water*, 15, 2020, <https://doi.org/10.3390/w15112020>, 2023.
- Merz, B., Basso, S., Fischer, S., Lun, D., Blöschl, G., Merz, R., Guse, B., Viglione, A., Vorogushyn, S., Macdonald, E., Wietzke, L., and Schumann, A.: Understanding Heavy Tails of Flood Peak Distributions, *Water Resources Research*, 58, e2021WR030506, <https://doi.org/10.1029/2021WR030506>, 2022.



- Montanari, A., Merz, B., and Blöschl, G.: HESS Opinions: The sword of Damocles of the impossible flood, *Hydrology and Earth System Sciences*, 28, 2603–2615, <https://doi.org/10.5194/hess-28-2603-2024>, 2024.
- 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/10.1002/joc.7107>, 2021.
- 600 Nguyen, V. D., Vorogushyn, S., Nissen, K., Brunner, L., and Merz, B.: A non-stationary climate-informed weather generator for assessing future flood risks, *Advances in Statistical Climatology, Meteorology and Oceanography*, 10, 195–216, <https://doi.org/10.5194/ascmo-10-195-2024>, 2024.
- Poudel, S. and Steinschneider, S.: An argument for parsimony in differentiable hydrologic models, *Hydrology and Earth System Sciences*, 30, 5173–5193, <https://doi.org/10.5194/hess-30-5173-2026>, 2026.
- 605 Rauthe, M., Steiner, H., Riediger, U., Mazurkiewicz, A., and Gratzki, A.: A Central European precipitation climatology Part I: Generation and validation of a high-resolution gridded daily data set (HYRAS), *Meteorologische Zeitschrift*, 22, 235–256, <https://doi.org/10.1127/0941-2948/2013/0436>, 2013.
- Seibert, J.: HBV Light Version 2, User's Manual, Tech. rep., Department of Physical Geography and Quaternary Geology, Stockholm University, Stockholm, https://www.geo.uzh.ch/dam/jcr:c8afa73c-ac90-478e-a8c7-929eed7b1b62/HBV_manual_2005.pdf, 2005.
- 610 Song, Y., Sawadekar, K., Frame, J. M., Pan, M., Clark, M. P., Knoben, W. J. M., Wood, A. W., Lawson, K. E., Patel, T., and Shen, C.: Physics-Informed, Differentiable Hydrologic Models for Capturing Unseen Extreme Events, *Water Resources Research*, 62, e2025WR040414, <https://doi.org/10.1029/2025WR040414>, 2026.
- Vorogushyn, S., Apel, H., Kemter, M., and Thielen, A. H.: Analysis of flood hazard in the Ahr Valley considering historical floods, *Hydrologie und Wasserbewirtschaftung / BfG*, 66, 2022, p. 5. Seite 244, https://doi.org/10.5675/HYWA_2022.5_2, 2022.
- 615 Vrugt, J. A.: Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation, *Environmental Modelling & Software*, 75, 273–316, <https://doi.org/10.1016/j.envsoft.2015.08.013>, 2016.
- Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., and Xu, H.: Time Series Data Augmentation for Deep Learning: A Survey, in: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 4653–4660, International Joint Conferences on Artificial Intelligence Organization, Montreal, Canada, <https://doi.org/10.24963/ijcai.2021/631>, 2021.
- 620 Yang, S., Yang, D., Chen, J., Santisirisomboon, J., Lu, W., and Zhao, B.: A physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data, *Journal of Hydrology*, 590, 125 206, <https://doi.org/10.1016/j.jhydrol.2020.125206>, 2020.
- Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, *Water Resources Research*, 44, <https://doi.org/10.1029/2007wr006716>, 2008.
- 625 Zhang, Z., Fischer, E., Zscheischler, J., and Engelke, S.: Physics-based models outperform AI weather forecasts of record-breaking extremes, *Science Advances*, 12, eaec1433, <https://doi.org/10.1126/sciadv.aec1433>, 2026.