https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

Probabilistic deep learning models for streamflow forecasting: A case study in the Rhine basin

Patrick Bentheimer[1], Bastian Klein[1], and Dennis Meißner[1]

[1]Federal Institute of Hydrology, Koblenz, Germany

Correspondence: Patrick Bentheimer (buddy.bent@web.de)

Abstract. Along the Rhine River, low-flow conditions are among the most frequent and economically significant constraints on inland waterway transport. Reliable and accurate forecasts are critical for informed decision-making. Although deep learning has shown great promise in hydrological modelling, process-based models remain the operational standard at many institutions, including the German Federal Institute of Hydrology.

In this study, we show that a regionally trained probabilistic deep learning model improves operational streamflow forecasts for the Rhine basin. We further investigate a hybrid configuration that combines data-driven and process-based modelling by incorporating operational forecasts as inputs to the deep learning model. Both approaches provide more accurate and reliable forecasts than the current operational system. Improvements are most pronounced for low flows, where the models reach skill levels that the operational system attains only at substantially shorter lead times, corresponding to an effective gain of approximately 2 forecast days. The hybrid approach achieves the highest overall performance by combining the short-term strengths of the operational system with the long-term generalization capabilities of deep learning.

These findings highlight that deep learning can complement established operational systems without replacing them, offering a practical pathway to integrate data-driven methods into operational practice. This, in turn, enables safer and more efficient inland waterway transport and facilitates planning under low-flow conditions.

1 Introduction

Accurate and reliable rainfall-runoff modelling remains a fundamental challenge in hydrology and plays a key role in water management (Dobriyal et al., 2017; McMillan et al., 2017), ecosystem conservation (Bunn and Arthington, 2002), and prediction of extreme events, particularly floods and low flows (Smakhtin, 2001; Alfieri et al., 2013; Nearing et al., 2024). Traditional conceptual and process-based hydrological models are widely used for these tasks, but they face well-documented limitations (Beven, 1987, 2001; Blöschl et al., 2019; Nearing et al., 2021).

One particularly challenging aspect is regionalizing rainfall-runoff models across multiple catchments (Kratzert et al., 2019b). Traditional hydrological models perform best when calibrated for individual catchments, but their performance deteriorates when applied regionally or globally (Mizukami et al., 2017; Kratzert et al., 2019b, 2024). This problem arises from the difficulty of extrapolating hydrological information from one catchment to another (Blöschl and Sivapalan, 1995) and is closely related to the problem of prediction in ungauged basins (PUB, Hrachowitz et al., 2013; Kratzert et al., 2024). Numerous model-dependent regionalization strategies have been developed (Razavi and Coulibaly, 2013), ranging from multi-scale parameter regionalization schemes (Samaniego et al., 2010) to donor-catchment approaches (Beck et al., 2016). Because these approaches rely on predefined hydrological models and prior knowledge of the hydrological system, their ability to generalize across diverse catchments remains limited (Prieto et al., 2019; Kratzert et al., 2019b, 2024). Thus, achieving spatially consistent and reliable hydrological simulations across heterogeneous regions continues to be one of the most persistent challenges in hydrological modelling (Guo et al., 2021).

Another long-standing problem is the effective minimization and quantification of predictive uncertainties, which has been a major focus of research over the past decades (Pappenberger and Beven, 2006; Nearing et al., 2021; Klotz et al., 2022). Various approaches have been developed to estimate uncertainty, including meteorological ensemble forecasts (e.g. Cloke and Pappenberger, 2009; Thielen et al., 2009), statistical time-series models (e.g. Bogner and Pappenberger, 2011; Broersen and Weerts, 2005), data assimilation techniques (e.g. Reichle et al., 2002; Alvarado-Montero et al., 2017), post-processing methods (e.g. Montanari and Koutsoyiannis, 2012; Hemri and Klein, 2017), and broader uncertainty quantification frameworks (e.g. Pappenberger et al., 2005; Beven and Binley, 1992). Despite substantial methodological progress, there is currently no single, prevailing approach for obtaining distributional rainfall-runoff predictions (Klotz et al., 2022). The majority of existing methods build upon deterministic models that are augmented with some uncertainty estimation strategy (Klotz et al., 2022). In this context, distributions are added on top of model inputs, structures, or predictions to represent a lack of complete information (Klotz et al., 2022; Nearing et al., 2021). Consequently, most models do not produce probabilistic predictions intrinsically but are instead used within larger frameworks, relying on external methods such as ensemble sampling or Bayesian inference to approximate uncertainty (Klotz et al., 2022).

Concurrently, the rise of machine learning and deep learning has transformed many fields of the natural sciences (He et al., 2019; AlQuraishi, 2019; Lam et al., 2023; Ravuri et al., 2021; Zhu et al., 2017). In hydrology, the application of deep learning to rainfall-runoff modelling gained momentum with the work of Kratzert et al. (2018), who demonstrated that Long Short-Term Memory (LSTM) networks can predict streamflow from meteorological observations with accuracies comparable to well-established conceptual models.

LSTMs are a type of recurrent neural network (RNN) first proposed by Hochreiter and Schmidhuber (1997). Their architecture includes memory cells that are conceptually analogous to the state variables of traditional dynamical system models, making them particularly well suited for simulating natural systems such as watersheds. Unlike conventional RNNs, LSTMs mitigate the problem of exploding and vanishing gradients, enabling stable learning over long sequences. This ability to capture long-term dependencies is especially valuable for modelling catchment processes that evolve over extended timescales, such as snow accumulation and seasonal vegetation dynamics (Kratzert et al., 2019b).

On this basis, Kratzert et al. (2019a, b) demonstrated that a regionally trained LSTM, leveraging both time series data and static catchment attributes, can achieve state-of-the-art performance in rainfall-runoff modelling, outperforming multiple locally and regionally calibrated benchmark models. Their findings marked a paradigm shift by showing that a single LSTM can learn hydrological behaviors at both local and regional scales (Kratzert et al., 2019b). This capability to capture similarities and differences among catchments directly advances predictions in ungauged and data-scarce basins (Kratzert et al., 2019a).

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

By learning temporal and spatial dynamics simultaneously within a single predictive framework, deep learning effectively circumvents many limitations of traditional regionalization methods, such as the need to transfer parameters or define catchment similarities in advance (Kratzert et al., 2019b). In contrast to conceptual and process-based hydrological models, deep learning models perform better when trained on multiple catchments, improving generalization across heterogeneous regions (Gauch et al., 2021; Nearing et al., 2021; Kratzert et al., 2024).

While these advances represent a major step forward in regional hydrological modelling, they do not yet account for uncertainty, an aspect inherent to all parts of hydrological science (Pappenberger and Beven, 2006; Beven, 2016; Sivapalan et al., 2003). Klotz et al. (2022) addressed this limitation by extending the LSTM rainfall-runoff models of Kratzert et al. (2019a, b) with mixture density networks (Bishop, 1994). In these models, rather than predicting deterministic streamflow values, the network learns the parameters of a probability distribution. This allows the network to produce probabilistic predictions that reliably express uncertainty directly from the input data – without relying on ensemble sampling or a priori probabilities (Klotz et al., 2022; Kratzert et al., 2024).

Nearing et al. (2024) expanded the LSTM-based modelling framework developed by Kratzert et al. (2018, 2019a, b, 2021), and Klotz et al. (2022) into a global operational system, illustrating that deep learning models can generalize to ungauged basins and produce reliable, real-time predictions across diverse hydrological and climatic regions. Their architecture employs an encoder-decoder LSTM setup, in which one LSTM extracts hydrologically relevant information from historical forcings (encoder), while a second LSTM uses meteorological forecasts to generate multi-day streamflow predictions (decoder). This design allows the model to leverage both past and future forcing information, resulting in forecasts that exceed the performance of current state-of-the-art global flood forecasting systems (Nearing et al., 2024). These developments highlight the transformative potential of deep learning in hydrology – not only for scientific research but also for operational practice.

Despite these advances, major hydrological institutions such as the German Federal Institute of Hydrology (BfG) continue to rely primarily on established process-based systems. Barriers include a lack of institutional trust in data-driven methods, limited expertise in artificial intelligence, and challenges in integrating deep learning approaches with established process-based modelling frameworks. Overcoming these obstacles is essential to bridge the gap between research innovations and their practical implementation.

This study aims to address these challenges by demonstrating that deep learning can improve operational forecasts at the BfG. The institute is responsible for the development and maintenance of navigation-related forecasting systems for the German inland waterways. Since low flows represent the most significant hydroclimatic threat to the efficiency and reliability of inland waterway transport, the forecasting systems primarily focus on low- to medium-flow conditions (Meißner et al., 2022). Using the Rhine basin – an important region for navigation-related forecasting – as a case study, the objectives of this study are:

(i) to benchmark the performance of the deep learning model against the operational process-based forecasting system, and

(ii) to explore the added value of a hybrid approach that integrates process-based forecasts as inputs to the deep learning model.

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

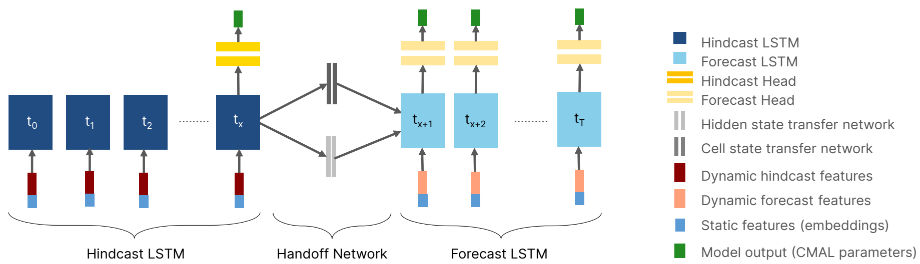

Figure 1. Overview of the LSTM-based state-handoff forecast model, showing the hindcast and forecast networks and their connection via internal state handoff. Adapted from Nearing et al. (2024).

2 Data and Methods

2.1 Deep learning model

In this study, we employ the LSTM architecture from Nearing et al. (2024) that uses a state-handoff to transition from a hindcast sequence model to a forecast sequence model and has demonstrated operational success in Google's Flood Hub system.

The hindcast LSTM processes a 365-day sequence of observed meteorological and hydrological data up to the forecast issue date. The resulting internal states, both the cell state and the hidden state, are then passed through the handoff network to initialize the forecast LSTM, which subsequently predicts streamflow over a 10-day lead time using meteorological forecasts as inputs. Conceptually, the hindcast LSTM should capture the catchment's recent hydrometeorological conditions by processing the preceding 365 days of observed data. The forecast LSTM then uses these learned states as initial conditions.

We train a single regional model across subbasins in the Rhine catchment. Consequently, both networks additionally receive static attributes describing basin-specific climatic, hydrological, topographic, land cover, and soil characteristics, allowing the model to learn shared patterns across catchments. An overview of the model architecture is illustrated in Fig. 1.

The model design further enables uncertainty estimation through mixture density networks (Klotz et al., 2022). A mixture density network is a neural network that combines multiple simple distributions, called components, to generate a mixture probability density that represents uncertainty in the predictions (Bishop, 1994; Klotz et al., 2022). From input data such as precipitation and temperature, the network estimates the parameters of each component (e.g., mean and standard deviation for Gaussian component) as well as their corresponding mixing weights. The mixed density is then formed as a weighted sum of these components. The process is illustrated in Fig. 2. Prediction quantiles, confidence intervals, and other measures of uncertainty can be derived by sampling from the probabilistic distribution, providing a comprehensive representation of forecast uncertainty.

We implemented the model using a hidden size of 256 cell states for both the hindcast and forecast LSTM, a fully connected hidden-state transfer network, and a mixture density network with asymmetric Laplacian distributions as components. At each time step, the model predicts the parameters of a single asymmetric Laplacian distribution over area-normalized streamflow, and

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

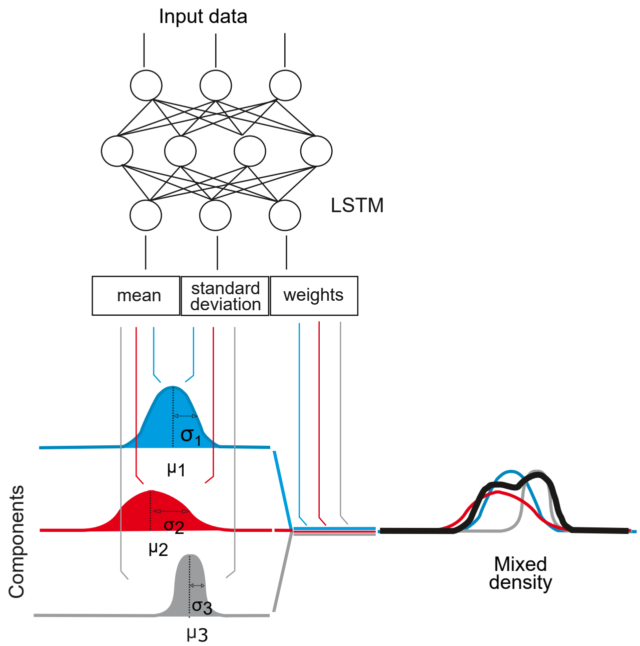

Figure 2. Schematic illustration of a mixture density network for probabilistic forecasting. For each input, the network predicts the parameters and weights of multiple components, which are combined to form a mixture density representing predictive uncertainty. Adapted from Klotz et al. (2022).

the loss function is the joint negative log-likelihood of this heteroscedastic density. The model produces a separate asymmetric Laplacian distribution for each time step and forecast lead time rather than a single deterministic prediction.

The data were split into training, validation, and test sets. The training and validation sets were used to estimate model parameters and tune hyperparameters, while the test set was reserved exclusively for performance evaluation. All input data were standardized using the mean and standard deviation of the training period.

Daily sliding windows without padding were used, with splits defined by forecast issue date. Each sample consists of a 375-day input sequence (365 days of hindcast observations followed by a 10-day forecast horizon). Due to the long hindcast window and daily sequencing, training inputs may include hindcast observations from the respective test year, and forecast dates may overlap across training and test sets near year boundaries. However, this does not imply information leakage, as only observations available at forecast initialization are used and overlapping forecast dates occur exclusively for different forecast issue dates and lead times, ensuring strict separation of forecast targets.

Because only eight full years of data were available (2017–2024), the model was trained and evaluated using leave-one-year-out cross-validation based on hydrological years (November–October). The eight yearly cross-validation runs were then concatenated to form a continuous discharge forecast for each basin and lead time. Based on these time series, metrics were computed separately for each lead time and basin, then averaged across lead times to evaluate overall forecast performance.

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

Deterministic and probabilistic forecasts were derived from an ensemble of five independently trained LSTM models with identical architectures and hyperparameters that differ only in their random initialization. Probabilistic results were obtained by pooling 100 samples per time step and lead time from each model. Deterministic forecasts were derived by taking the median of the pooled 500 samples at each time step and forecast lead time. For reference, we also tested the approach of Nearing et al. (2024), in which the median hydrograph from each individual LSTM is computed first and then averaged across the ensemble. Both approaches yield nearly identical deterministic results. However, pooling preserves the full predictive distribution required for probabilistic evaluation, which is why it is used in this study.

2.2 Data

The Rhine River basin, including its major tributaries (Fig. 3), served as the case study for this work. The Rhine is Europe's most important inland waterway. It has a total length of approximately 1,320 km and drains a basin of about 185,000 km^2 into the North Sea (Belz et al., 2007). The elevation range of the basin extends from sea level at the delta in the Netherlands to more than 4,000 m in the Alpine headwaters. The catchments considered in this study represent a variety of runoff regimes. Discharge in the Upper Rhine is strongly influenced by the Alpine part of the basin and is characterized by a pronounced unimodal glacial-nival regime, with peak flows in summer and low flows in late autumn and winter. In contrast, the other tributaries contributing to the Rhine are predominantly rainfall-dominated (pluvial). The superposition of pluvial and snowmelt-driven runoff components results in complex flow regimes along the river continuum (Belz et al., 2007). Gauging stations located in the Alpine region were excluded from the analysis because strong anthropogenic influences, such as reservoirs, regulated lakes, bypass channels, and inter-basin water transfers, negatively affect the training of regionalized deep learning models.

Thus, our dataset comprises model inputs and streamflow targets for 52 gauges and corresponding catchments, where catchments are defined by the total upstream area of each gauge and may be nested along the same river. All data are provided as catchment-aggregated (lumped) time series at daily resolution and streamflow was normalized by the upstream catchment area and expressed as specific discharge.

Observed meteorological data, consisting of precipitation and temperature, were derived from interpolated station measurements provided by the German Weather Service (DWD), Météo-France, and MeteoSwiss. Both meteorological and streamflow measurements from catchment gauges were used as inputs for the hindcast LSTM model.

Meteorological forecasts (temperature and precipitation) were obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF), including both the high-resolution deterministic forecast (HRES) and the ensemble prediction system (ENS), with 51 ensemble members. Although ENS provides forecasts up to 15 days ahead, the lead time was restricted to 10 days, matching the maximum forecast horizon of HRES. Due to major changes in the ECMWF ENS system (substantial increase in spatial resolution) introduced in March 2016 (Haiden et al., 2016), the analysis was restricted to the period from 2017–2024 to ensure a homogeneous dataset. Nevertheless, it should be noted that the considered numerical weather prediction (NWP) models were continuously updated during this period, resulting in changes in the statistical properties of the forecast variables. Rather than using all individual ENS members, ensemble summary statistics (e.g., the mean, median, minimum, and maximum across members) were computed and used as inputs. In addition, the second experiment utilized autocorrected

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

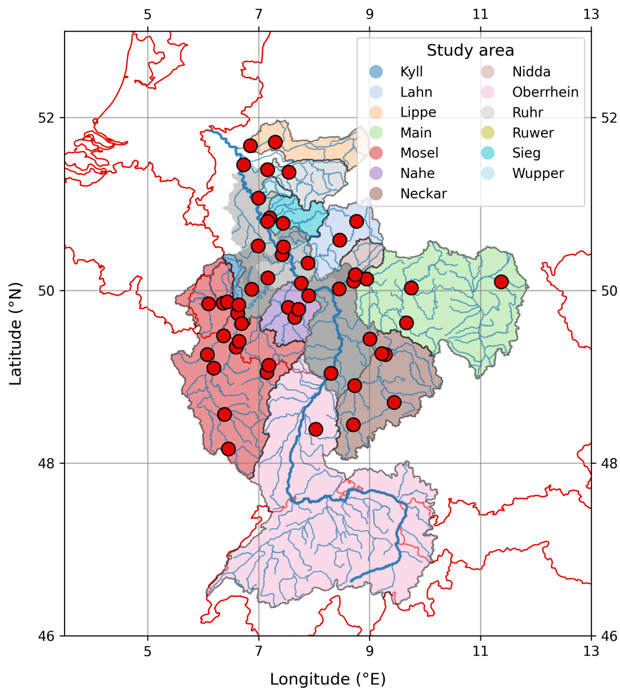

Figure 3. Overview of the study area, the Rhine River basin, showing its major tributaries and the 52 gauging stations included in this study. Gauging stations located downstream within the Netherlands and those situated in the Alpine region were excluded from the analysis.

forecasts from the operational forecast system. These forecast datasets served as inputs to the forecast LSTM model, which generates streamflow predictions over a lead time of 10 days.

Hydrologic, climatic, topographic, land use, soil, and anthropogenic basin attributes were incorporated to provide static catchment characteristics (Table 1). These attributes were created specifically for the study gauges following an approach similar to the CAMELS (Addor et al., 2017) and HydroATLAS (Linke et al., 2019) datasets.

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

Table 1. Static catchment attributes.

Attribute	Description	Unit	Source
frac_snow	Fraction of precipitation falling as snow, i.e. while mean air temperature is $< 0°$C.	–	HYRAS (Rauthe et al., 2013)
high_prec_dur	Mean duration of high-precipitation events (number of consecutive days with precipitation $\geq 5\times$ mean daily precipitation).	d	
high_prec_freq	Frequency of high-precipitation days ($\geq 5\times$ mean daily precipitation).	$\mathrm{d\,yr^{-1}}$	
low_prec_dur	Mean duration of dry periods (number of consecutive days with precipitation $< 1\ \mathrm{mm\,d^{-1}}$).	d	
low_prec_freq	Frequency of dry days ($< 1\ \mathrm{mm\,d^{-1}}$).	$\mathrm{d\,yr^{-1}}$	
p_mean	Long-term mean daily precipitation from 1951 to 2020.	$\mathrm{mm\,d^{-1}}$	
p_seasonality	Seasonality and timing of precipitation estimated using sine curves representing annual temperature and precipitation cycles. Positive (negative) values indicate summer (winter) precipitation peaks, while values close to zero indicate uniform precipitation throughout the year.	–	
runoff_ratio	Runoff ratio, defined as the ratio of mean daily discharge to mean daily precipitation.	–	HYRAS (Rauthe et al., 2013); Federal state agencies; Federal Waterways and Shipping Administration
q_mean	Mean daily specific discharge.	$\mathrm{mm\,d^{-1}}$	Federal state agencies; Federal Waterways and Shipping Administration
area_metadata	Catchment area as provided by the agencies.	$\mathrm{km^2}$	
slope_fdc	Slope of the flow duration curve between the log-transformed 33rd and 66th streamflow percentiles (see Coxon et al. (2020)).	–	
snow_ice_perc	Areal coverage of snow and ice.	%	IHME1500 (Günther and Duscher, 2019)
agricultural_areas_perc	Areal coverage of agricultural areas.	%	CORINE Land Cover (European Environment Agency, 2020)
artificial_surfaces_perc	Areal coverage of artificial surfaces.	%	
forests_and_seminatural_areas_perc	Areal coverage of forests and semi-natural areas.	%	
water_bodies_perc	Areal coverage of water bodies.	%	
wetlands_perc	Areal coverage of wetlands.	%	

Continued on next page

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

Static catchment attributes (continued)

Attribute	Description	Unit	Source
bulk_density_0_30cm_mean	Bulk density of the fine earth fraction at depth 0–30 cm.	kg dm^{-3}	SoilGrids250m (Poggio et al., 2021)
clay_0_30cm_mean	Proportion of clay particles ($<$ 0.002 mm) in the fine earth fraction at depth 0–30 cm.	wt.%	
coarse_fragments_0_30cm_mean	Volumetric fraction of coarse fragments ($>$ 2 mm) at depth 0–30 cm.	wt.%	
sand_0_30cm_mean	Proportion of sand particles ($>$ 0.05/0.063 mm) at depth 0–30 cm.	wt.%	
silt_0_30cm_mean	Proportion of silt particles ($\geq$ 0.002 mm and $\leq$ 0.05/0.063 mm) in the fine earth fraction at depth 0–30 cm.	wt.%	
soil_organic_carbon_0_30cm_mean	Soil organic carbon content in the fine earth fraction at depth 0–30 cm.	wt.%	
elev_mean	Mean catchment elevation based on the MERIT Hydro geometry.	m a.s.l.	Copernicus GLO-30 DEM (European Space Agency, 2022)
elev_sd	Standard deviation of elevation values within the catchment.	m	

2.3 Operational model

In the current operational hydrological forecasting system of the BfG, hydrological processes within the catchment are simulated using the conceptual, semi-distributed rainfall-runoff model HydPy-H-HBV96-PREVAH (HydPy-H). This model integrates the PREVAH concept (Viviroli et al., 2009) for soil water balance simulations with the HBV96 concept for remaining hydrological processes (Bergström, 1995; Lindström et al., 1997). The Rhine river basin is divided into 375 subbasins which are further subdivided into hydrological response units (HRU) based on land use and elevation classes. Runoff generation processes are calculated at the HRU level. Model simulations are conducted at an hourly temporal resolution. For benchmarking, the hourly forecasts were aggregated to daily resolution.

The hydrological model uses observed meteorological data up to the forecast initialization time to determine the initial states (e.g., soil moisture and snow water equivalent). Beyond the initialization time, meteorological forecasts of precipitation and temperature are used to predict streamflow. To guarantee fair comparison, both models use the same meteorological input data. When ensemble forecasts (ENS) are used as input, the model generates a separate forecast for each ensemble member. To create a deterministic forecast for benchmarking, a hydrograph is constructed from the median (50th percentile) of all ensemble members.

In addition to raw simulations, the operational model provides autocorrected forecasts by applying an autoregressive error-correction model (Broersen and Weerts, 2005), which corrects forecast values based on residuals between past observed and simulated values using observed meteorology over a 192-hour window before forecast initialization. Each catchment is individ-

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

ually calibrated using observed meteorological input data for the period 1 October 2011 to 30 September 2019. The calibration objective was the mean Nash-Sutcliffe Efficiency (NSE) between simulated and observed flows. As a result, comparisons are inherently favorable to the operational model, representing a conservative benchmark for evaluating deep learning performance.

2.4 Experiments

To address the two objectives outlined in the introduction, we designed two experiments to evaluate (i) the performance of the deep learning model relative to the operational process-based forecasting system and (ii) the potential of combining both modelling paradigms in a hybrid forecasting approach.

2.4.1 Exp. DL – Benchmarking data-driven against process-based forecasts

The first experiment benchmarks the performance of the deep learning approach against the operational forecasting system. The data-driven model uses the following inputs:

- **Hindcast:** catchment-aggregated precipitation and temperature as well as observed streamflow,
- **Forecast:** forecasted precipitation and temperature from ECMWF HRES and ENS,
- **Static attributes:** approximately 30 catchment descriptors, including hydrologic, climatic, land use, topographic, and geological variables.

2.4.2 Exp. Hybrid – Combining data-driven and process-based forecasts

In the second experiment, we explore a hybrid forecasting approach that combines data-driven and process-based modelling by adding the operational forecasts as input to the deep learning model. Alongside the inputs used in Exp. DL, the forecast LSTM additionally receives:

- **Autocorrected streamflow forecasts** from the operational HydPy-H model.

In principle, if the LSTM were to adopt the operational forecast directly, it would match the benchmark, while still having the potential to learn improvements from the additional inputs. Hereafter, the purely data-driven setup is referred to as DL, while the hybrid configuration incorporating operational process-based forecasts is denoted Hybrid.

2.5 Metrics

The performances of the models are evaluated using a comprehensive set of metrics that capture different aspects of predictive skill. As primary efficiency measures, we use the Nash-Sutcliffe Efficiency (Nash and Sutcliffe, 1970), the NSE computed on logarithmically transformed discharge (log-NSE), and the Kling-Gupta Efficiency (KGE, Gupta et al., 2009). To assess bias components explicitly, we further include the α-NSE and β-NSE decompositions proposed by Gupta et al. (2009), together with the β-KGE term.

While these metrics assess performance over the full time series, hydrological applications often require an evaluation of specific discharge regimes. Therefore, we additionally compute three signatures derived from the flow duration curve (FDC) following Yilmaz et al. (2008):

- the bias in high flows, represented by the 2 % exceedance flow or peak-flow bias (FHV),
- the bias in the slope of the middle section of the FDC (FMS),
- the bias in low flows, quantified using the bottom 30 % of the FDC (FLV).

Probabilistic forecast performance is assessed using the Continuous Ranked Probability Score (CRPS; Hersbach, 2000) and its associated skill score (CRPSS), which quantifies improvement relative to the operational forecasting system. Moreover, forecast calibration and reliability are examined using probability integral transform (PIT) histograms (Dawid, 1984; Diebold et al., 1997; Gneiting et al., 2007).

Together, these metrics offer a comprehensive evaluation framework, capturing overall model skill, regime-specific performance, and the reliability and accuracy of probabilistic forecasts. All metrics are reported solely for the test period.

2.6 Software

All analyses in this study were conducted using open-source software. The experiments were implemented in Python 3.11 (Van Rossum and Drake Jr, 1995). Data preprocessing and management were carried out using the scientific Python libraries NumPy (Harris et al., 2020), Pandas (Wes McKinney, 2010), and Xarray (Hoyer and Hamman, 2017), which provide efficient tools for numerical computation, tabular data handling, and labeled multi-dimensional arrays, respectively. Metrics were calculated using NeuralHydrology (Kratzert et al., 2022), while CRPS was computed using properscoring (Barrett et al., 2015).

The LSTM architecture used in this study is based on the model introduced by Nearing et al. (2024), which has been incorporated into the NeuralHydrology code base (Kratzert et al., 2022). While the underlying architecture remains consistent with their implementation, we developed a customized data processing and sequencing pipeline that accepts actual forecasts as inputs. This modification allows the model to make multi-day forecasts instead of only nowcasts. Using the setup described herein, the deep learning model takes approximately 24 hours to train and evaluate on a single NVIDIA-A100 graphics processing unit. All figures were generated using Matplotlib (Hunter, 2007) and seaborn (Waskom, 2021). To improve accessibility for readers with colour vision deficiencies, we employed scientifically designed colour maps (Crameri et al., 2020).

3 Results

We first illustrate the forecasting behavior and uncertainty characteristics of the models using a representative basin and hydrological year. Building on this example, we quantify forecast performance and reliability per lead time before extending the analysis to all basins and evaluation years to assess the model's robustness and ability to generalize across basins. Lead time is expressed as the number of days after the prediction date, with day 1 corresponding to the first 24-hour forecast and day 10 to

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

forecasts 10 days ahead. While the main figures display NSE and log-NSE, KGE was evaluated analogously and is discussed where relevant. An overview of model performance across all basins, evaluation years, and lead times is summarized in Table 2.

Table 2. Overview of key performance metrics for all models, averaged across all basins and lead times. Median values are reported for each metric.

Model	NSE[a]	log-NSE[b]	KGE[c]	α-NSE[d]	β-NSE[e]	β-KGE[f]	FHV[g]	FMS[h]	FLV[i]
HydPy-H	0.765	0.851	0.725	0.777	-0.063	0.926	-25.328	-6.392	-156.980
LSTM DL	0.765	0.893	0.760	0.821	-0.063	0.923	-22.482	-2.264	22.078
LSTM Hybrid	0.781	0.899	0.779	0.834	-0.062	0.929	-20.157	-2.947	22.004

[a] Nash-Sutcliffe Efficiency $(-\infty, 1]$, values closer to 1 are desirable. [b] NSE computed on logarithmically transformed discharge $(-\infty, 1]$, values closer to 1 are desirable. [c] Kling-Gupta Efficiency $(-\infty, 1]$, values closer to 1 are desirable. [d] α-NSE decomposition $(0, \infty)$, values close to 1 are desirable. [e] β-NSE decomposition $(-\infty, \infty)$, values close to 0 are desirable. [f] β-KGE decomposition $(0, \infty)$, values close to 1 are desirable. [g] Top 2% peak flow bias $(-\infty, \infty)$, values close to 0 are desirable. [h] 30% low flow bias $(-\infty, \infty)$, values close to 0 are desirable. [i] Bias of FDC mid-segment slope, $(-\infty, \infty)$, values close to 0 are desirable.

3.1 Illustrative discharge forecasts for a representative basin

In this section, we analyze discharge forecasts for gauge Kalkofen in the Lahn basin (5,304 km^2) for the hydrological year 1 November 2019–31 October 2020.

Figure 4 presents probabilistic forecasts from the deep learning model and the operational forecasting system. For both systems, median forecasts and prediction intervals are derived from the respective ensembles and shown for lead times of 1, 5, and 10 days. As expected, forecast skill decreases and predictive uncertainty increases with lead time.

Deterministic forecast performance is evaluated using NSE, log-NSE, and KGE at each lead time, computed from the median forecast. The results (Fig. 5) indicate that the deep learning model achieves higher skill for most lead times, whereas the operational system performs better at short lead times (1–2 days).

We subsequently assess probabilistic calibration and reliability using PIT histograms. Figure 6 reveals pronounced differences between the two systems. The operational model exhibits a strongly U-shaped PIT distribution across lead times, indicating that the model is overconfident and consistently underestimates the true uncertainty of the forecasts. In contrast, the deep learning model produces PIT distributions that are considerably closer to uniform, suggesting improved probabilistic calibration and more reliable forecasts.

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

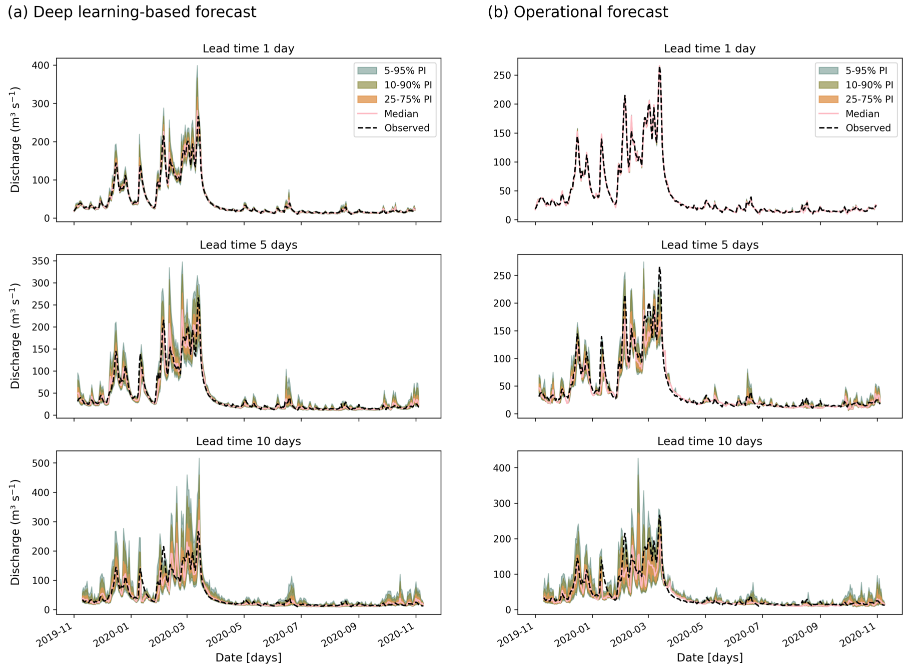

Figure 4. Comparison of probabilistic discharge forecasts from (a) the deep learning model (Exp. DL) and (b) the operational forecasting system at gauge Kalkofen in the Lahn basin for one hydrological year (November 2019–October 2020). Forecasts are shown for lead times of 1, 5, and 10 days. Shaded bands indicate prediction intervals (PI, 5–95%, 10–90%, and 25–75%), illustrating forecast uncertainty. For the deep learning model, they are derived from pooling five independent model runs (N=500), and for the operational forecast, from aggregating ensemble members (N=51). Forecast performance decreases and predictive uncertainty increases with increasing lead time.

3.2 Exp. DL – Benchmarking data-driven against process-based forecasts

3.2.1 Overview

Figure 7 illustrates lead-time-aggregated performance differences between the deep learning model and the operational forecasting system across 52 gauges in the Rhine basin over all evaluation years. Positive values indicate higher skill of the deep learning model.

For overall hydrological performance, NSE indicates that the deep learning model generally performs worse, with only 18 gauges (35 %) showing equal or improved skill. By contrast, KGE – another metric for overall hydrological performance, which accounts for correlation, bias, and variability – shows improvement at 32 gauges (62 %). For low flows, evaluated using log-NSE, the deep learning model consistently outperforms the benchmark at all gauges.

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

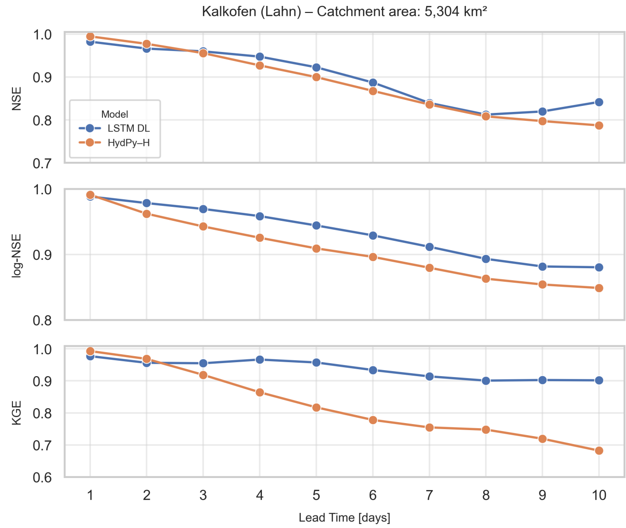

Figure 5. Performance comparison of the deep learning model (Exp. DL) and the operational forecasting model at gauge Kalkofen in the Lahn basin for the hydrological year November 2019–October 2020. Forecast skill of discharge predictions is compared per lead time using NSE, log-NSE, and KGE. The deep learning model shows higher skill than the operational model for most lead times across all metrics, while the operational model performs better at lead times of 1 to 2 days.

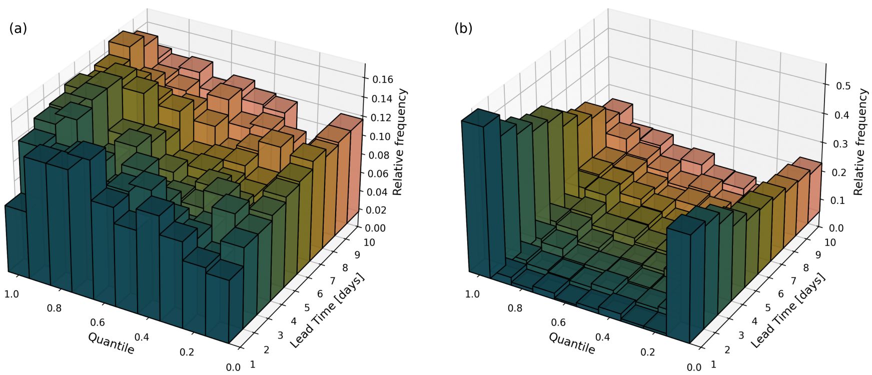

Figure 6. Three dimensional probability integral transform (PIT) histograms for (a) the deep learning model (Exp. DL) and (b) the operational forecasting model at gauge Kalkofen in the Lahn basin for the hydrological year November 2019–October 2020. The PIT distribution highlights differences in probabilistic calibration and forecast reliability across lead times. For a perfectly calibrated forecast, the PIT values are uniformly distributed, corresponding to a relative frequency of 0.1 in each bin.

3.2.2 Key metrics

In Fig. 8 we present cumulative density functions (CDFs) for various evaluation metrics. Metrics are aggregated over all lead times, and CDFs are smoothed using kernel density estimation solely to aid visual interpretation.

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

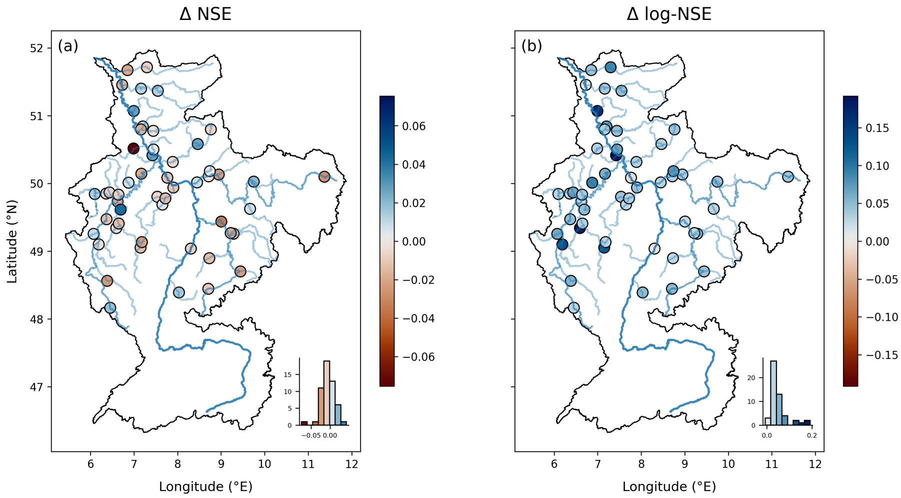

Figure 7. Overview of performance differences between the deep learning model (Exp. DL) and the operational forecasting system across 52 gauges in the Rhine basin. Differences are shown for (a) NSE and (b) log-NSE and are computed from discharge forecasts over 2017–2024, aggregated across all lead times. Positive values indicate higher skill of the deep learning model. Improvements are consistently observed for log-NSE, whereas differences in NSE vary, with both positive and negative values.

Both the deep learning model and the operational system show generally high skill across all gauges. For NSE and KGE, both models perform comparably, with median NSE of 0.765 and median KGE of 0.760 (deep learning) and 0.725 (operational). Differences are more pronounced for log-NSE. The deep learning model achieves higher median log-NSE (0.893 vs. 0.851), and no gauge falls below 0.8, whereas nine gauges (17 %) do so for the operational system.

Looking at the decomposition of NSE and KGE, the deep learning model better captures flow variability, though the model still underestimates variability (α-NSE). Both models tend to underestimate discharge as captured by β-NSE and β-KGE.

Metrics derived from the flow duration curve indicate that the deep learning model better represents low- and normal-flow conditions (FLV and FMS). In contrast, improvements for high-flow conditions, quantified by FHV, are limited, with both models showing similar performance in representing peak flows.

3.2.3 Forecast skill across lead times

Analyzing forecast performance by lead time provides a more detailed picture. Figure 9 summarizes basin-wise skill differences between the deep learning model and the operational forecast system. The ridge plot illustrates the distribution of model skill differences. Positive values indicate higher skill of the deep learning model.

Although the overall NSE indicates that the deep learning model outperforms the operational system in only 35 % of basins, a lead-time-resolved perspective reveals important deviations. At a 1-day lead time, nearly all basins (except three, 6 %) show reduced skill, whereas from lead time 2 onward, between 50 % and 75 % of basins show improved performance.

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

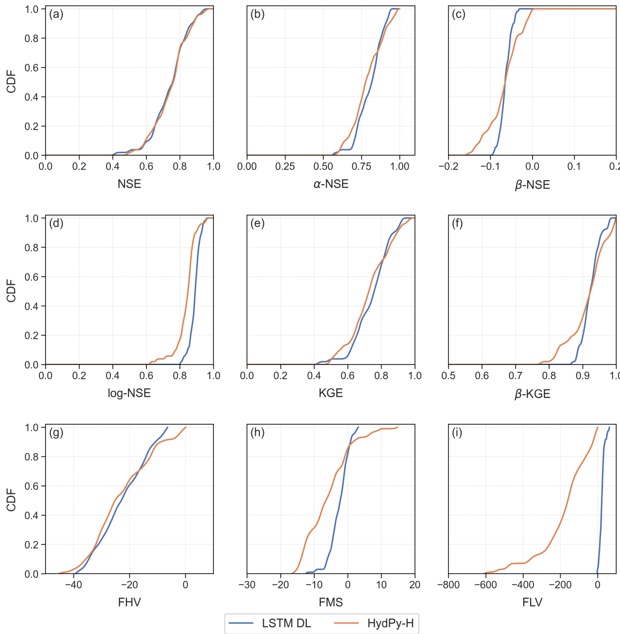

Figure 8. Cumulative density functions of key metrics for the deep learning model (Exp. DL) and the operational forecasting system across all 52 gauges. For each basin metrics are computed from discharge forecasts over 2017–2024, aggregated across all lead times. Curves are smoothed using a kernel density estimate. The deep learning model performs comparably to or better than the operational forecast system.

In contrast, the lead-time-aggregated log-NSE suggests improved performance at all gauges, while a lead-time-specific analysis highlights more mixed behavior at a 1-day lead time, where improvements are observed at only 29 gauges (56 %). From lead time 2 onward, the results again align with the overall assessment, with almost all basins exhibiting higher skill. Importantly, beyond a lead time of 4 days, the deep learning model reaches skill levels that the operational system attains only at substantially shorter lead times. In practice, this corresponds to a shift of approximately 2 forecast days – for example, the 9-day log-NSE matches or exceeds the operational model's 7-day performance.

Like NSE, KGE exhibits reduced forecast skill at 1-day lead time for almost all basins. From lead time 4 onward, most basins show pronounced improvements, effectively shifting the forecast skill by 1–2 days. This behavior mirrors that of log-NSE, confirming consistent gains of the deep learning model at extended horizons.

3.3 Exp. Hybrid – Combining data-driven and process-based forecasts

Next, we assess the impact of combining data-driven and process-based forecast paradigms by adding the operational forecasts as input to the deep learning model. Performance is compared to both the operational benchmark and Exp. DL.

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

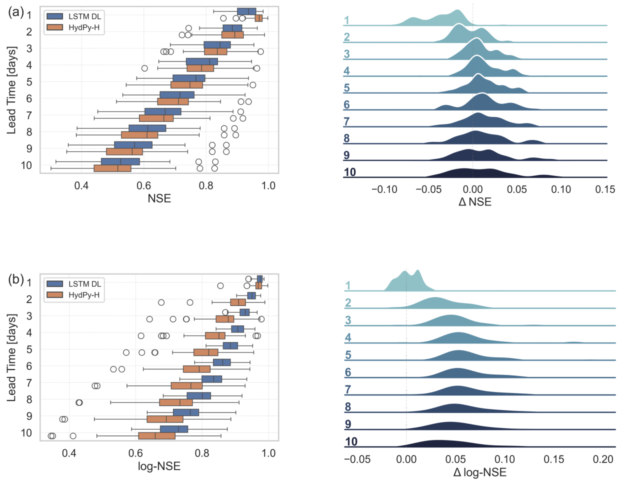

Figure 9. Lead-time-dependent skill comparison between Exp. DL and the benchmark model for (a) NSE and (b) log-NSE. Metrics are computed per basin from discharge forecasts over all evaluation years. Boxplots summarize the distribution of skill, while ridge plots show smoothed distributions of model skill differences using kernel density estimates. Positive values indicate higher skill of the deep learning model.

3.3.1 Overview

Using the same diagnostics as in the first experiment, we assess the model's ability to generalizes across the Rhine Basin.

Figure 10 shows that the hybrid model outperforms the operational forecasting system at the majority of gauges, with higher NSE at 45 stations (87 %). This represents a clear improvement over Exp. DL, where only 18 gauges (35 %) exhibited improved or equivalent NSE. As in Exp. DL, log-NSE increases at all stations. Improvements in KGE are less pronounced but remain consistent with this overall trend. In total, 37 gauges (71 %) show higher KGE values than the operational forecast, compared to 32 gauges (62 %) in Exp. DL.

3.3.2 Key metrics

In Fig. 11 we again evaluate CDFs for various evaluation metrics. For comparison results from Exp. DL are additionally shown.

Relative to the operational forecasting system, the hybrid model shows improved skill across most metrics. Median NSE and KGE increase to 0.781 and 0.779, respectively, compared to 0.765 and 0.725 for the operational system. Log-NSE remains high, with all gauges being above 0.8 and a median value of 0.899, representing a clear improvement over the operational system (0.851). Compared to Exp. DL, these results indicate consistent gains in NSE and KGE, with a modest improvement in log-NSE.

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

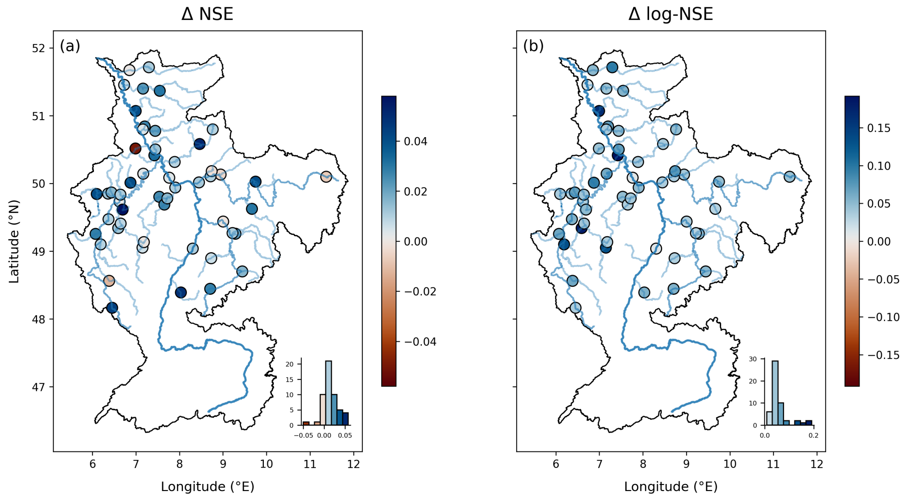

Figure 10. Overview of performance differences in the Rhine basin between the deep learning model (Exp. Hybrid) and the operational forecasting model across all basins for (a) NSE and (b) log-NSE. Metric differences are computed from discharge forecasts over all evaluation years and aggregated across all lead times. Positive values indicate higher skill of the deep learning model. Improvements are consistently observed for log-NSE across all basins. Similarly so for NSE, however a few have decreased skill.

Decomposition of NSE and KGE indicates that bias and flow variability remain largely unchanged by incorporating the operational forecasts as inputs. Similar to Exp. DL, both models tend to underestimate discharge. The hybrid model appears to capture flow variability slightly better relative to Exp. DL.

Regime-specific performance shows improved low- and normal-flow representation, consistent with Exp. DL, and now additionally exhibits a small improvement in peak flows.

3.3.3 Forecast skill across lead times

Skill differences across lead times for the hybrid model are summarized in Fig. 12.

Similar to the first experiment, NSE exhibits a drop in skill at a 1-day lead time, with only 28 gauges (54 %) showing improved performance, compared to 45 gauges (87 %) overall. From lead time 2 onward, however, the hybrid model demonstrates clear gains, with over 80 % of basins exhibiting higher skill.

For log-NSE, improvements are consistently observed across all lead times. In contrast to Exp. DL, this includes the 1-day lead time, where almost all basins show higher skill. Consistent with Exp. DL, starting from a 4-day lead time, the hybrid model attains skill levels that the operational system reaches only at substantially shorter lead times, corresponding to an effective gain of about 2 forecast days.

Compared to Exp. DL, for both metrics the hybrid model approach leads to improved performance across all lead times, with the largest gains at the 1-day lead time. The improvement gradually decreases with increasing lead time, suggesting that

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

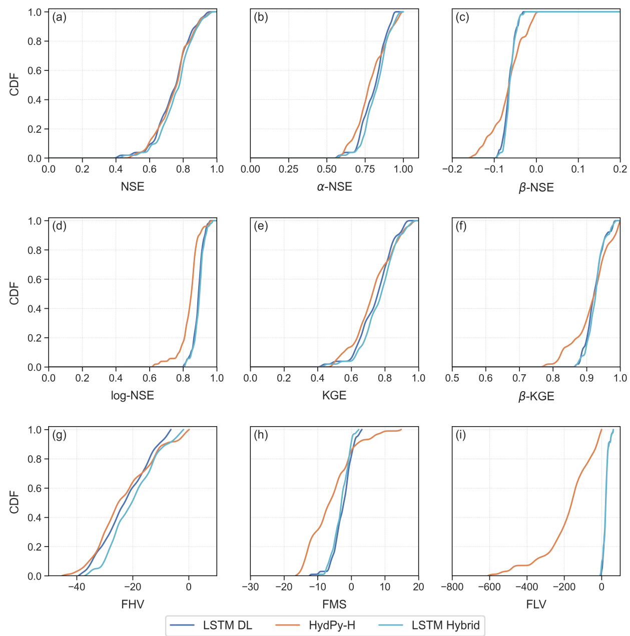

Figure 11. Cumulative distribution functions of key metrics comparing Exp. Hybrid to the benchmark model. Metrics are computed per basin from discharge forecasts across all evaluation years and aggregated over all lead times. Curves are smoothed using a kernel density estimate. Results from Exp. DL are shown for comparison. The hybrid LSTM model generally outperforms the benchmark model.

the additional information from operational forecasts is most valuable at shorter horizons. Improvements in log-NSE are more modest but consistent across all lead times.

KGE shows a pattern consistent with Exp. DL, with reduced skill at 1-day lead time and clear improvements at longer horizons. As before, these gains correspond to an effective shift of roughly 1 to 2 forecast days.

3.3.4 Spatial distribution

No obvious spatial pattern emerges in Fig. 7 and 10. However, plotting performance differences against basin area highlights that the relative improvements decline with increasing basin size, with smaller basins showing the most pronounced improvements. In other words, the greatest gains occur in small basins, even though absolute performance remains highest in the largest basins.

3.4 Probabilistic evaluation

So far, we have focused on the deterministic performance of the models and have not yet considered their probabilistic skill. To address this, we evaluate the Continuous Ranked Probability Skill Score (CRPSS) of the deep learning models relative to the operational forecast system. Positive CRPSS values indicate higher accuracy of the deep learning model.

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

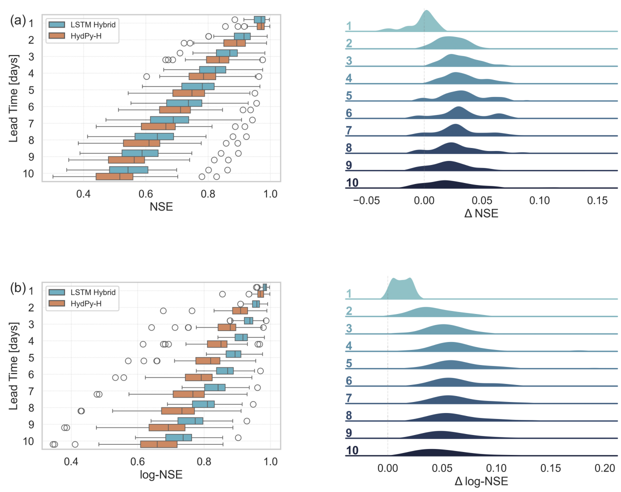

Figure 12. Lead-time-dependent skill comparison between Exp. Hybrid and the benchmark model for (a) NSE and (b) log-NSE. Metrics are computed per basin from discharge forecasts over all evaluation years. Boxplots summarize the distribution of skill, while ridge plots show smoothed distributions of model skill differences using kernel density estimates. Positive values indicate higher skill of the deep learning model.

Figure 13 shows a clear improvement in probabilistic forecast accuracy across most basins. The median CRPSS improvement amounts to approximately 11 % for Exp. DL and 16 % for Exp. Hybrid.

A more differentiated picture across lead times is provided in Fig. 14. For Exp. Hybrid, improvements in CRPSS are largest at short lead times, with a median increase in accuracy of around 30 %. The magnitude of these improvements steadily decreases as lead time increases. Exp. DL shows a comparable decline in CRPSS with increasing lead time, starting from a lead time of 2 days. At a lead time of 1 day, however, Exp. DL deviates from this general pattern, exhibiting markedly weaker performance, with positive skill gains observed for only 27 gauges (52 %). Nevertheless, even at the longest lead times, probabilistic accuracy is improved at 63 % and 88 % of the basins for Exp. DL and Exp. Hybrid, respectively.

4 Discussion

Despite the recent success of deep learning in hydrological modelling (Kratzert et al., 2018; Nearing et al., 2024), major hydrological institutions continue to rely predominantly on established process-based systems. This study demonstrates that deep learning-based forecasting can effectively complement these systems in a regional, real-world context, providing more accurate and reliable forecasts. Importantly, the purely data-driven LSTM model achieves forecast skill that is comparable to, and in many cases exceeds, that of the operational system. These results are consistent with previous studies (Kratzert et al., 2019b; Nearing et al., 2024; Feng et al., 2020; Loritz et al., 2024; Lees et al., 2021). The hybrid model, which incorporates operational

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

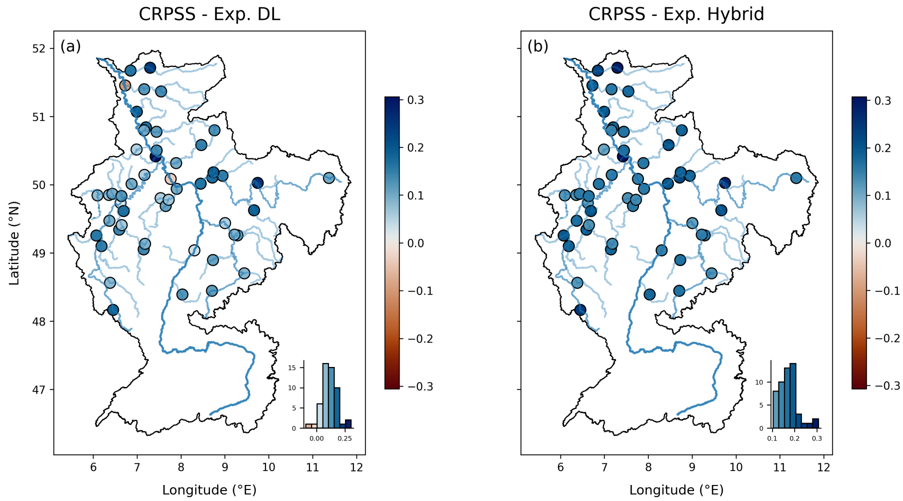

Figure 13. Overview of the probabilistic forecast performance of the deep learning models, Exp. DL (a) and Exp. Hybrid (b), relative to the operational forecasting system across all 52 gagues. The Continuous Ranked Probability Skill Score quantifies improvements relative to the operational benchmark. Scores are computed from discharge forecasts over all evaluation years and aggregated across all lead times. Both experiments show improved CRPSS at the majority of gauges.

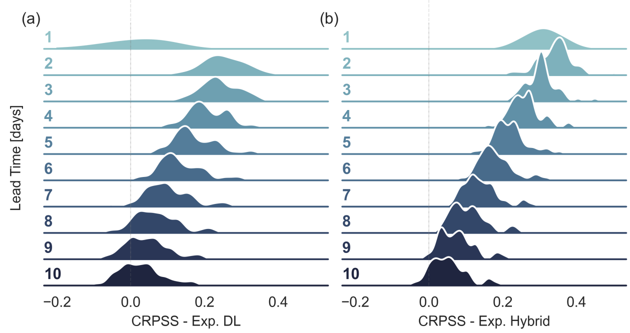

Figure 14. Lead-time-dependent probabilistic forecast performance of the deep learning models, Exp. DL (a) and Exp. Hybrid (b), relative to the operational forecasting system. The Continuous Ranked Probability Skill Score quantifies improvements relative to the operational benchmark. Scores are computed from discharge forecasts over all evaluation years. The ridge plots show smoothed distributions using kernel density estimates.

process-based forecasts as inputs, further improves the overall forecast skill, consistently outperforming the operational model across the majority of catchments and lead times.

A consistent pattern emerges across both experimental setups. At short lead times, particularly at 1 day ahead, the operational system proves difficult to outperform in terms of deterministic skill. The purely data-driven deep learning model generally performs worse, and even the hybrid approach provides only modest improvements. This behavior likely reflects the

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

autoregressive structure of the operational model, which tightly anchors forecasts to the most recent discharge observations. As a consequence, short-term forecasts are sharp but underdispersed, indicating that reliability is sacrificed for accuracy. In contrast, the deep learning-based forecasts provide a more reliable representation of uncertainty, with probabilistic performance improving most noticeably at short lead times. In contrast, for longer lead times the deep learning models consistently outperform the operational system. Improvements are most pronounced for low flows, where the models reach skill levels that the operational system attains only at substantially shorter lead times, corresponding to an effective gain of approximately 2 forecast days. Overall model performance, as measured by KGE, shows a similar extension of the effective forecast horizon by roughly 1–2 days. This indicates that the deep learning approach reduces the rate at which predictive skill deteriorates with lead time. Improvements in NSE are more moderate, which may, on one hand, reflect the fact that the operational model is explicitly calibrated to maximize mean NSE. On the other hand, the comparatively larger gains in KGE suggest that the deep learning model provides a more balanced representation of correlation, variability, and bias. Beyond deterministic skill, the deep learning models produce markedly more reliable and accurate probabilistic forecasts. The PIT analysis indicates substantially improved calibration across forecast horizons, demonstrating that uncertainty is represented more consistently than in the operational system.

Importantly, the deep learning approach not only improves forecast skill but also addresses key structural limitations of traditional hydrological models, namely regionalization and intrinsic uncertainty estimation. By learning distributional parameters directly from input data, the model eliminates the need for ensemble-based strategies to estimate uncertainty (Klotz et al., 2022). Despite being trained as a single regional model across all basins, it achieves strong performance at individual catchments. In many cases, it even outperforms operational predictions, which are calibrated separately for each basin. This supports the growing body of evidence that deep learning models fundamentally benefit from regional training and can mitigate long-standing challenges related to parameter transfer and defining catchment similarities (Kratzert et al., 2024; Guo et al., 2021).

The hybrid approach achieves the best overall performance. Compared to the purely data-driven model, this integration particularly enhances predictions at short and intermediate lead times. This highlights the complementary strengths of the two approaches: the operational system captures short-term autoregressive dynamics, while the deep learning model improves generalization over longer horizons. In essence, our hybrid approach can be interpreted as a post-processing strategy. Our findings are partly in contrast with Frame et al. (2021), where LSTM post-processors did neither consistently nor significantly outperform purely data-driven approaches. Differences in results may be attributed to model architecture, research area, basin coverage, forecast horizon, and evaluation strategy. More sophisticated methods exist to combine process-based and data-driven approaches, such as adding physical constraints into the deep learning model structure (Hoedt et al., 2021). However, all attempts to add physics to (properly trained) streamflow LSTM models in hydrology have produced lower-performing models (Kratzert et al., 2024). Moreover, their implementation can be technically complex. Given the persistent barriers highlighted in the introduction, a simple approach that allows both models to run in parallel while retaining access to the process-based forecast may offer practical advantages. For instance, physics-based models provide a consistent fallback and serve as a plausi-

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

bility check in cases where deep learning models fail due to unforeseen events not represented in the training data. In addition, parallel operation helps forecasters build trust in data-driven predictions.

Several limitations of our study should be acknowledged. First, the modelling framework is lumped, aggregating meteorological inputs and catchment characteristics over entire basins. This implicitly assumes spatial homogeneity and may limit performance in larger catchments, where spatial variability in forcing and hydrological response becomes more pronounced. The observed decrease in performance gains with increasing basin size supports this interpretation. Second, forecasts are provided at daily resolution. In operational practice, sub-daily predictions, particularly hourly forecasts, are highly relevant for both flood and low-flow management. Extending LSTM models to hourly resolution substantially increases computational demand (Acuña Espinoza et al., 2025), and recurrent architectures can struggle with very long sequences (Zhang and You, 2020; Zhao et al., 2020). Multi-frequency approaches (Acuña Espinoza et al., 2025), which process older inputs at coarser temporal resolution, may provide a promising pathway to overcome these constraints. Third, the framework is inherently restricted to gauged basins. Because observed discharge and operational forecasts are used as inputs, the approach cannot be directly transferred to ungauged catchments. While a purely data-driven configuration could exclude discharge observations, our hybrid setup fundamentally relies on operational forecasts. Consequently, this constraint substantially limits its broader applicability, given that only a small fraction of the world's river basins are equipped with discharge gauges (Nearing et al., 2024)

5 Conclusion

Overall, the results demonstrate that deep learning can enhance even well-established, state-of-the-art operational hydrological forecasts. The model extends the effective forecast horizon and improves the reliability of probabilistic predictions without compromising short-term forecast performance. At the same time, it addresses two longstanding limitations of traditional hydrological models: regionalization and uncertainty estimation. These findings highlight the relevance of deep learning not only for developing new forecasting frameworks, but also for enhancing existing operational systems.

The hybrid model demonstrates that the integration of deep learning does not necessarily require a choice between data-driven and process-based modelling. Instead, the combination of both approaches makes use of their complementary strengths. Given the persistent challenges associated with the operational adoption of data-driven models, a parallel implementation of traditional and deep learning methods may offer important practical advantages. Physics-based models can provide a reliable fallback and serve as a plausibility check when deep learning models encounter unforeseen events or conditions that are not sufficiently represented in the training data. At the same time, operating both models in parallel allows forecasters to gradually build trust in data-driven predictions. Rather than replacing established systems, hybrid approaches therefore provide a pragmatic pathway for integrating deep learning into operational practice, bridging the gap between research advances and real-world decision support.

Furthermore, the results indicate that the model proposed by Nearing et al. (2024), originally developed and evaluated for global floods in ungauged basins, can be successfully applied to gauged basins at a regional scale. Its applicability also extends beyond flood forecasting, showing substantial potential for improving the prediction of low-flow conditions.

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

From an operational perspective, the improvements in low-flow forecasting are particularly relevant for navigation on the Rhine, where low-flow conditions represent the most frequent and economically significant constraint on inland waterway transport (Klein and Meißner, 2016). More accurate and better-calibrated probabilistic forecasts support improved planning for navigation, water management, and recreational use, thereby contributing to more efficient inland waterway transport and reduced emissions through optimized operations.

Code availability. The code developed for this study is stored in a Gitlab repositroy hosted by OpenCode. Access is currently restricted to the authors. A permanent DOI will be assigned upon completion of the institutional approval process.

Data availability. All underlying data used in this research study are openly available. The sources are listed in Sect. 2.2. We plan to publish the pre-processed model input data and the model results on Zenodo. However, we lack the distribution license for data from certain gauges. These data will therefore be omitted from the published datasets and represented by NaN values where applicable. The missing gauge data can be obtained directly from the responsible operators. A permanent DOI for the published datasets will be assigned upon completion of the institutional approval process.

Author contributions. PB developed the model code and performed the simulations. PB, BK and DM designed all the experiments. PB conducted all the experiments and analyzed the results, together with BK and DM. PB prepared the manuscript with contributions from all co-authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors acknowledge the Federal Ministry of Transport for providing funding for the presented research in the context of the Research & Development Project “MALPROG”. An AI-based language assistance tool was used for spelling and grammar checking and for suggesting alternative sentence formulations. The authors take full responsibility for the content of the manuscript.

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

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, Hydrology and Earth System Sciences, 29, 1749–1758, https://doi.org/10.5194/hess-29-1749-2025, 2025.

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.

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., and Pappenberger, F.: GloFAS–global ensemble streamflow forecasting and flood early warning, Hydrology and Earth System Sciences, 17, 1161–1175, 2013.

AlQuraishi, M.: End-to-End Differentiable Learning of Protein Structure, Cell Systems, 8, 292–301.e3, https://doi.org/https://doi.org/10.1016/j.cels.2019.03.006, 2019.

Alvarado-Montero, R., Schwanenberg, D., Krahe, P., Helmke, P., and Klein, B.: Multi-parametric variational data assimilation for hydrological forecasting, Advances in Water Resources, 110, 182–192, https://doi.org/https://doi.org/10.1016/j.advwatres.2017.09.026, 2017.

Barrett, L., Hoyer, S., Kleeman, A., and O'Kane, D.: properscoring: Proper Scoring Rules in Python, https://github.com/properscoring/properscoring, 2015.

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., and Bruijnzeel, L. A.: Global-scale regionalization of hydrologic model parameters, Water Resources Research, 52, 3599–3622, https://doi.org/https://doi.org/10.1002/2015WR018247, 2016.

Belz, J. U., Brahmer, G., Buiteveld, H., Engel, H., Grabher, R., Hodel, H., Krahe, P., Lammersen, R., Larina, M., Mendel, H., et al.: Das Abflussregime des Rheins und seiner Nebenflüsse im 20. Jahrhundert: Analyse, Veränderungen, Trends, CHR/KHR, 2007.

Bergström, S.: The HBV model., 1995.

Beven, K.: Towards a new paradigm in hydrology, IN: Water for the Future: Hydrology in Perspective. IAHS Publication, 1987.

Beven, K.: How far can we go in distributed hydrological modelling?, Hydrology and Earth System Sciences, 5, 1–12, https://doi.org/10.5194/hess-5-1-2001, 2001.

Beven, K.: Facets of uncertainty: epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication, Hydrological Sciences Journal, 61, 1652–1665, 2016.

Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction, Hydrological Processes, 6, 279–298, https://doi.org/https://doi.org/10.1002/hyp.3360060305, 1992.

Bishop, C. M.: Mixture density networks, 1994.

Blöschl, G. and Sivapalan, M.: Scale issues in hydrological modelling: A review, Hydrological Processes, 9, 251–290, https://doi.org/https://doi.org/10.1002/hyp.3360090305, 1995.

Blöschl, G., Bierkens, M. F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J. W., McDonnell, J. J., Savenije, H. H., Sivapalan, M., Stumpp, C., Toth, E., Volpi, E., Carr, G., Lupton, C., Salinas, J., Széles, B., Viglione, A., Aksoy, H., Allen, S. T., Amin, A., Andréassian, V., Arheimer, B., Aryal, S. K., Baker, V., Bardsley, E., Barendrecht, M. H., Bartosova, A., Batelaan, O., Berghuijs, W. R., Beven, K., Blume, T., Bogaard, T., de Amorim, P. B., Böttcher, M. E., Boulet, G., Breinl, K., Brilly, M., Brocca, L., Buytaert, W., Castellarin, A., Castelletti, A., Chen, X., Chen, Y., Chen, Y., Chifflard, P., Claps, P., Clark, M. P., Collins, A. L., Croke, B., Dathe, A., David, P. C., de Barros, F. P. J., de Rooij, G., Baldassarre, G. D., Driscoll, J. M., Duethmann, D., Dwivedi, R., Eris, E., Farmer, W. H., Feiccabrino, J., Ferguson, G., Ferrari, E., Ferraris, S., Fersch, B., Finger, D., Foglia, L., Fowler, K., Gartsman, B., Gascoin, S., Gaume, E., Gelfan, A., Geris, J., Gharari, S., Gleeson, T., Glendell, M., Bevacqua, A. G., González-Dugo, M. P., Grimaldi, S., Gupta, A. B., Guse, B., Han, D., Hannah, D., Harpold, A., Haun, S., Heal, K., Helfricht, K., Herrnegger, M., Hipsey, M., Hlaváčiková, H., Hohmann, C., Holko, L., Hopkinson, C., Hrachowitz, M., Illangasekare, T. H., Inam, A., Innocente, C., Istanbulluoglu, E., Jarihani, B., Kalantari, Z., Kalvans, A., Khanal, S., Khatami, S., Kiesel, J., Kirkby, M., Knoben, W., Kochanek, K., Kohnová, S., Kolechkina, A., Krause, S., Kreamer, D., Kreibich, H., Kunstmann, H., Lange, H., Liberato, M. L. R., Lindquist, E., Link, T., Liu, J., Loucks, D. P., Luce, C., Mahé, G., Makarieva, O., Malard, J., Mashtayeva, S., Maskey, S., Mas-Pla, J., Mavrova-Guirguinova, M., Mazzoleni, M., Mernild, S., Misstear, B. D., Montanari, A., Müller-Thomy, H., Nabizadeh, A., Nardi, F., Neale, C., Nesterova, N., Nurtaev, B., Odongo, V. O., Panda, S., Pande, S., Pang, Z., Papacharalampous, G., Perrin, C., Pfister, L., Pimentel, R., Polo, M. J., Post, D., Sierra, C. P., Ramos, M.-H., Renner, M., Reynolds, J. E., Ridolfi, E., Rigon, R., Riva, M., Robertson, D. E., Rosso, R., Roy, T., Sá, J. H., Salvadori, G., Sandells, M., Schaefli, B., Schumann, A., Scolobig, A., Seibert, J., Servat, E., Shafiei, M., Sharma, A., Sidibe, M., Sidle, R. C., Skaugen, T., Smith, H., Spiessl, S. M., Stein, L., Steinsland, I., Strasser, U., Su, B., Szolgay, J., Tarboton, D., Tauro, F., Thirel, G., Tian, F., Tong, R., Tussupova, K., Tyralis, H., Uijlenhoet, R., van Beek, R., van der Ent, R. J., van der Ploeg, M., Loon, A. F. V., van Meerveld, I., van Nooijen, R., van Oel, P. R., Vidal, J.-P., von Freyberg, J., Vorogushyn, S., Wachniew, P., Wade, A. J., Ward, P., Westerberg, I. K., White, C., Wood, E. F., Woods, R., Xu, Z., Yilmaz, K. K., and Zhang, Y.: Twenty-three unsolved problems in hydrology (UPH) – a community perspective, Hydrological Sciences Journal, 64, 1141–1158, https://doi.org/10.1080/02626667.2019.1620507, 2019.

Bogner, K. and Pappenberger, F.: Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system, Water Resources Research, 47, https://doi.org/https://doi.org/10.1029/2010WR009137, 2011.

Broersen, P. M. and Weerts, A. H.: Automatic error correction of rainfall-runoff models in flood forecasting systems, in: 2005 IEEE Instrumentationand Measurement Technology Conference Proceedings, vol. 2, pp. 963–968, IEEE, 2005.

Bunn, S. E. and Arthington, A. H.: Basic principles and ecological consequences of altered flow regimes for aquatic biodiversity, Environmental management, 30, 492–507, 2002.

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

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J., Lane, R., Lewis, M., Robinson, E. L., et al.: CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain, Earth System Science Data, 12, 2459–2483, 2020.

Crameri, F., Shephard, G. E., and Heron, P. J.: The misuse of colour in science communication, Nature communications, 11, 5444, 2020.

Dawid, A. P.: Present position and potential developments: Some personal views statistical theory the prequential approach, Journal of the Royal Statistical Society: Series A (General), 147, 278–290, 1984.

Diebold, F. X., Gunther, T. A., and Tay, A. S.: Evaluating Density Forecasts, Working Paper 215, National Bureau of Economic Research, https://doi.org/10.3386/t0215, 1997.

Dobriyal, P., Badola, R., Tuboi, C., and Hussain, S. A.: A review of methods for monitoring streamflow for sustainable water resource management, Applied Water Science, 7, 2617–2628, 2017.

European Environment Agency: CORINE Land Cover 2018 (raster 100 m), Europe, 6-yearly – version 2020_20u1, May 2020, https://doi.org/10.2909/960998c1-1870-4e82-8051-6485205ebbac, 2020.

European Space Agency: Copernicus GLO-30 DEM, https://doi.org/10.5270/esa-c5d3d65, 2022.

Feng, D., Fang, K., and Shen, C.: Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks With Data Integration at Continental Scales, Water Resources Research, 56, e2019WR026 793, https://doi.org/https://doi.org/10.1029/2019WR026793, e2019WR026793 2019WR026793, 2020.

Frame, J. M., Kratzert, F., Raney II, A., Rahman, M., Salas, F. R., and Nearing, G. S.: Post-Processing the National Water Model with Long Short-Term Memory Networks for Streamflow Predictions and Model Diagnostics, JAWRA Journal of the American Water Resources Association, 57, 885–905, https://doi.org/https://doi.org/10.1111/1752-1688.12964, 2021.

Gauch, M., Mai, J., and Lin, J.: The proper care and feeding of CAMELS: How limited training data affects streamflow prediction, Environmental Modelling & Software, 135, 104 926, 2021.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic Forecasts, Calibration and Sharpness, Journal of the Royal Statistical Society Series B: Statistical Methodology, 69, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x, 2007.

Günther, A. and Duscher, K.: Extended vector data of the International Hydrogeological Map of Europe 1: 1,500,000 (Version IHME1500 v1. 2), Technical note.(BGR, 2019), 2019.

Guo, Y., Zhang, Y., Zhang, L., and Wang, Z.: Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review, WIREs Water, 8, e1487, https://doi.org/https://doi.org/10.1002/wat2.1487, 2021.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80–91, https://doi.org/https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Haiden, T., Janousek, M., Bidlot, J., Ferranti, L., Prates, F., Vitart, F., Bauer, P., and Richardson, D.: Evaluation of ECMWF forecasts, including the 2016 resolution upgrade, European Centre for Medium Range Weather Forecasts Reading, UK, 2016.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E.: Array programming with NumPy, Nature, 585, 357–362, https://doi.org/10.1038/s41586-020-2649-2, 2020.

He, S., Li, Y., Feng, Y., Ho, S., Ravanbakhsh, S., Chen, W., and Póczos, B.: Learning to predict the cosmological structure formation, Proceedings of the National Academy of Sciences, 116, 13 825–13 832, https://doi.org/10.1073/pnas.1821458116, 2019.

Hemri, S. and Klein, B.: Analog-Based Postprocessing of Navigation-Related Hydrological Ensemble Forecasts, Water Resources Research, 53, 9059–9077, https://doi.org/https://doi.org/10.1002/2017WR020684, 2017.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather and Forecasting, 15, 559 – 570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.

Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural computation, 9, 1735–1780, 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, edited by Meila, M. and Zhang, T., vol. 139 of *Proceedings of Machine Learning Research*, pp. 4275–4286, PMLR, https://proceedings.mlr.press/v139/hoedt21a.html, 2021.

Hoyer, S. and Hamman, J.: xarray: N-D labeled arrays and datasets in Python, Journal of Open Research Software, 5, https://doi.org/10.5334/jors.148, 2017.

Hrachowitz, M., Savenije, H., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J., Arheimer, B., Blume, T., Clark, M., Ehret, U., et al.: A decade of Predictions in Ungauged Basins (PUB)—a review, Hydrological sciences journal, 58, 1198–1255, 2013.

https://doi.org/10.5194/egusphere-2026-4476
Preprint. Discussion started: 10 September 2026

Hunter, J. D.: Matplotlib: A 2D Graphics Environment , Computing in Science & Engineering, 9, 90–95, https://doi.org/10.1109/MCSE.2007.55, 2007.

Klein, B. and Meißner, D.: Vulnerability of Inland Waterway Transport and Waterway Management on Hydro-meteorological Extremes, Rapport, EU-Horizon2020 IMPREX, Deliverable, 9, 2016.

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

Kratzert, F., Klotz, D., 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., 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, 11 344–11 354, https://doi.org/https://doi.org/10.1029/2019WR026065, 2019a.

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, 2019b.

Kratzert, F., Klotz, D., Hochreiter, S., and Nearing, G. S.: A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling, Hydrology and Earth System Sciences, 25, 2685–2703, https://doi.org/10.5194/hess-25-2685-2021, 2021.

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

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.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al.: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, 2023.

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.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, Journal of Hydrology, 201, 272–288, https://doi.org/https://doi.org/10.1016/S0022-1694(97)00041-3, 1997.

Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H., et al.: Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution, Scientific data, 6, 283, 2019.

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 1582 catchments in Germany, Earth System Science Data, 16, 5625–5642, https://doi.org/10.5194/essd-16-5625-2024, 2024.

McMillan, H., Seibert, J., Petersen-Overleir, A., Lang, M., White, P., Snelder, T., Rutherford, K., Krueger, T., Mason, R., and Kiang, J.: How uncertainty analysis of streamflow data can reduce costs and promote robust decisions in water management applications, Water Resources Research, 53, 5220–5228, https://doi.org/https://doi.org/10.1002/2016WR020328, 2017.

Meißner, D., Klein, B., and Frielingsdorf, B.: Implementing Hydrological Forecasting Services Supporting Waterway Management and Transportation Logistics Relating to Hydroclimatic Impacts, Atmosphere, 13, https://doi.org/10.3390/atmos13101606, 2022.

Mizukami, N., Clark, M. P., Newman, A. J., Wood, A. W., Gutmann, E. D., Nijssen, B., Rakovec, O., and Samaniego, L.: Towards seamless large-domain parameter estimation for hydrologic models, Water Resources Research, 53, 8020–8040, https://doi.org/https://doi.org/10.1002/2017WR020401, 2017.

Montanari, A. and Koutsoyiannis, D.: A blueprint for process-based modeling of uncertain hydrological systems, Water Resources Research, 48, https://doi.org/https://doi.org/10.1029/2011WR011412, 2012.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I — A discussion of principles, Journal of Hydrology, 10, 282–290, https://doi.org/https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., et al.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, 2024.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What Role Does Hydrological Science Play in the Age of Machine Learning?, Water Resources Research, 57, e2020WR028 091, https://doi.org/https://doi.org/10.1029/2020WR028091, e2020WR028091 10.1029/2020WR028091, 2021.

Pappenberger, F. and Beven, K. J.: Ignorance is bliss: Or seven reasons not to use uncertainty analysis, Water Resources Research, 42, https://doi.org/https://doi.org/10.1029/2005WR004820, 2006.

Pappenberger, F., Beven, K. J., Hunter, N. M., Bates, P. D., Gouweleeuw, B. T., Thielen, J., and de Roo, A. P. J.: Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS), Hydrology and Earth System Sciences, 9, 381–393, https://doi.org/10.5194/hess-9-381-2005, 2005.

Poggio, L., De Sousa, L. M., Batjes, N. H., Heuvelink, G., Kempen, B., Ribeiro, E., and Rossiter, D.: SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, Soil, 7, 217–240, 2021.

Prieto, C., Le Vine, N., Kavetski, D., García, E., and Medina, R.: Flow Prediction in Ungauged Catchments Using Probabilistic Random Forests Regionalization and New Statistical Adequacy Tests, Water Resources Research, 55, 4364–4392, https://doi.org/https://doi.org/10.1029/2018WR023254, 2019.

Rauthe, M., Steiner, H., Riediger, U., Mazurkiewicz, A., Gratzki, A., et al.: A Central European precipitation climatology–Part I: Generation and validation of a high-resolution gridded daily data set (HYRAS), Meteorologische Zeitschrift, 22, 235–256, 2013.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., et al.: Skilful precipitation nowcasting using deep generative models of radar, Nature, 597, 672–677, 2021.

Razavi, T. and Coulibaly, P.: Streamflow prediction in ungauged basins: review of regionalization methods, Journal of hydrologic engineering, 18, 958–975, 2013.

Reichle, R. H., McLaughlin, D. B., and Entekhabi, D.: Hydrologic data assimilation with the ensemble Kalman filter, Monthly weather review, 130, 103–114, 2002.

Samaniego, L., Kumar, R., and Attinger, S.: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, Water Resources Research, 46, https://doi.org/https://doi.org/10.1029/2008WR007327, 2010.

Sivapalan, M., Takeuchi, K., Franks, S., Gupta, V., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J., Mendiondo, E. M., O'connell, P., et al.: IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences, Hydrological sciences journal, 48, 857–880, 2003.

Smakhtin, V.: Low flow hydrology: a review, Journal of Hydrology, 240, 147–186, https://doi.org/https://doi.org/10.1016/S0022-1694(00)00340-1, 2001.

Thielen, J., Bartholmes, J., Ramos, M.-H., and de Roo, A.: The European Flood Alert System – Part 1: Concept and development, Hydrology and Earth System Sciences, 13, 125–140, https://doi.org/10.5194/hess-13-125-2009, 2009.

Van Rossum, G. and Drake Jr, F. L.: Python tutorial, vol. 620, Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

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.

Waskom, M. L.: seaborn: statistical data visualization, Journal of Open Source Software, 6, 3021, https://doi.org/10.21105/joss.03021, 2021.

Wes McKinney: Data Structures for Statistical Computing in Python, in: Proceedings of the 9th Python in Science Conference, edited by Stéfan van der Walt and Jarrod Millman, pp. 56 – 61, https://doi.org/10.25080/Majora-92bf1922-00a, 2010.

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/https://doi.org/10.1029/2007WR006716, 2008.

Zhang, X. and You, J.: A Gated Dilated Causal Convolution Based Encoder-Decoder for Network Traffic Forecasting, IEEE Access, 8, 6087–6097, https://doi.org/10.1109/ACCESS.2019.2963449, 2020.

Zhao, J., Huang, F., Lv, J., Duan, Y., Qin, Z., Li, G., and Tian, G.: Do RNN and LSTM have Long Memory?, in: Proceedings of the 37th International Conference on Machine Learning, edited by III, H. D. and Singh, A., vol. 119 of *Proceedings of Machine Learning Research*, pp. 11 365–11 375, PMLR, https://proceedings.mlr.press/v119/zhao20c.html, 2020.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., and Fraundorfer, F.: Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources, IEEE Geoscience and Remote Sensing Magazine, 5, 8–36, https://doi.org/10.1109/MGRS.2017.2762307, 2017.