



Testing discharge assimilation strategies to enhance short-range AI-based operational rainfall–runoff forecasts

Bob E. Saint-Fleur¹, Eric Gaume¹, Florian Surmont¹, Nicolas Akil², and Dominique Theriez²

¹GERS-EE, Université Gustave Eiffel, Allée des Ponts et Chaussées, 44344 Bouguenais, France

²Aquasys Entreprise, 2 rue de Nantes, 44710 Port-Saint-Père, France

Correspondence: Bob E. Saint-Fleur (bob.saint-fleur@univ-eiffel.fr)

Received: 29 August 2025 – Discussion started: 30 September 2025

Revised: 27 March 2026 – Accepted: 18 May 2026 – Published: 11 June 2026

Abstract. Effective discharge forecasts are essential in operational hydrology. The accuracy of such forecasts, particularly in short lead times, is generally increased through the integration of recent measurements of observed discharge; commonly known as discharge assimilation (DA). Recent studies have demonstrated the effectiveness of deep learning (DL) approaches for rainfall–runoff (RR) modeling, particularly Long Short-Term Memory (LSTM) networks, outperforming traditional approaches. However, most of these studies do not include DA procedures, which may limit their operational forecast performance. This study suggests and evaluates three DA strategies that incorporate discharge from either recent discharge measurements or forecasts from a pre-trained rainfall–runoff model. The proposed strategies, based on a Multilayer Perceptron (MLP) as orchestrator, include: (1) the integration of recently observed discharges, (2) the integration of both recent discharge observations and pre-trained model forecasts, and (3) the post-processing of model forecast errors. Experiments are implemented using two large datasets, CAMELS-US and CAMELS-FR, and two established benchmark models (BM): the trained LSTM model from Kratzert et al. (2019) and the conceptual Sacramento Soil Moisture Accounting (SAC-SMA) model from Newman et al. (2017), covering both deep learning and conceptual RR simulation approaches. The considered lead times range from 1 to 7 d, covering both short- and mid-term horizons. The approaches are evaluated within two forecast frameworks: (1) perfect meteorological forecasts over the forecasting lead time and (2) ensemble meteorological forecasts. The two frameworks yield contrasting outcomes. When evaluated under the perfect forecast framework, the application of DA leads to substantial improvements in fore-

cast performance, although the magnitude of these gains depends on the initial performance of the benchmark models and the forecasting lead time. Improvements are consistently significant for the SAC-SMA cases, while for the LSTM cases, gains are observed mainly for basins where the LSTM initially underperforms. However, the ensemble forecast evaluation yields unexpected results: the performance ranking of the tested models changes markedly compared to the perfect forecast framework. The LSTM model, in particular, appears penalized by the under-dispersion of its forecast ensembles. Although this underdispersion could be partly attributable to the underdispersion of the forecast archives tested, it persists even when the model is driven by the high spread climatology-based ensemble. This finding underscores the importance of ensuring reliable ensemble dispersion for the efficient operational deployment of AI-based hydrological forecasts.

1 Introduction

Discharge forecasting models are essential in operational hydrology, whether for water resource or related-risk management. Their importance is set to increase as climate-related threats intensify (Schiermeier, 2018; Philip et al., 2020; Rentschler et al., 2023). However, providing accurate discharge forecasts remains challenging due to the complexity of rainfall–runoff (RR) processes, model imperfections, and uncertainty in input data, particularly regarding the quality of weather forecasts.

Over decades, significant efforts have been made to address the challenges of hydrological modeling, leading to

the development of various models and approaches. In the era of artificial intelligence (AI), notable advances have been achieved, with recent studies demonstrating the outstanding performance of deep learning models (DL) relative to traditional RR models (Kratzert et al., 2019; Husic et al., 2022). Commonly used DL architectures include multilayer perceptrons (MLPs) (Jeannin et al., 2021; Saint-Fleur et al., 2023), recurrent neural networks (RNNs) such as Long Short-Term Memory (LSTM) networks (Kratzert et al., 2018, 2019; Fang et al., 2021; Wunsch et al., 2021; Rahbar et al., 2022), and more recently, Transformers (Pözl et al., 2024). Nonetheless, most hydrological models in the literature are evaluated mainly under perfect weather scenarios, which may overestimate their performance in an operational forecasting framework. Although simulation models can be incorporated into forecasting systems, either as assimilable data or driven by forecasted forcings, their development frequently overlooks key components such as discharge assimilation, persistence analysis, and ensemble (probabilistic) assessment.

Persistence analysis, introduced by Kitanidis and Bras (1980), evaluates a model's performance relative to a naive baseline, which simply translates the current observation to the target lead time. This analysis, which serves as a relevant benchmark for assessing the predictive ability of models, is rarely considered in most hydrological modeling studies. Discharge assimilation (DA), on the other hand, which consists of dynamically providing real-time discharge data to a running forecast model, is essential in operational forecasting (Bourgin et al., 2014; Boucher et al., 2020; Piazzini et al., 2021). By ensuring regular updates of the model states, DA allows one to reduce the impact of uncertainties associated with meteorological forecasts and model structures, thus keeping the model aligned with evolving hydrological conditions. Several DA techniques exist, and their efficacy often depends on the reliability of the underlying model and/or the techniques used (Feng et al., 2020; Nearing et al., 2022; Yang et al., 2025). For direct DA strategies, the importance of DA is typically more pronounced at shorter lead times. However, suboptimal models may over-rely on the assimilated discharge data, which may overshadow the contribution of the forcings, leading towards naive models (Saint Fleur et al., 2020). Thus, DA methods can improve the operational application of RR forecasting models but are not straightforward to calibrate and implement efficiently.

In the following, two benchmark models are considered to evaluate the added value of DA procedures: the regional LSTM model of Kratzert et al. (2019) and the basin-specific conceptual SAC-SMA model from Newman et al. (2017). Three different discharge assimilation (DA) strategies that take into account past observed discharges to generate forecasts will be tested. For simpler implementation, including time and resource efficiency, a MultiLayer Perceptron (MLP) network (Rosenblatt, 1958) is used as the orchestrator in these DA methods. MLP networks have been largely adopted over recent decades (Werbos, 1988, 1974), and several stud-

ies have shown their effectiveness in RR modeling (Atmaja and Akagi, 2020; Oliveira et al., 2021; Jeannin et al., 2021). Although recent studies have demonstrated the superior performance of models such as LSTM (Kratzert et al., 2018) networks or transformers (Li et al., 2024), MLPs have been used in this study not only as a forecasting orchestrator but also as an alternative for RR modeling due to the relative simplicity of their implementation. Therefore, as a possible future work, the hereby developed MLP can be involved in a comparison with other classical "data assimilation" techniques, such as the Ensemble Kalman Filter (Clark et al., 2008).

As discharge assimilation procedures generally lose effectiveness at extended lead times, forecasts are evaluated at both short- and mid-term horizons. These lead times are defined with respect to the basin response times, estimated based on a rainfall-discharge cross-correlation analysis. To ensure operational relevance and reflect real-world forecasting practices, two scenarios are considered for the weather forecast data: (1) assuming weather forecasts are perfect; (2) using ensemble-based forecasts. Accordingly, forecast performance is assessed using both deterministic and probabilistic metrics.

The experiments are based on two widely used large-scale hydrometeorological datasets, CAMELS-US (Addor et al., 2017) and CAMELS-FR (Delaigue et al., 2025). Ensemble-based forecasts are obtained using historical meteorological observations, hindcast products, and forecast archives from the ECMWF platform.

This paper is structured as follows: Sect. 2 introduces the data and methods, Sect. 2.1 presents the datasets, and Sect. 2.2 to 2.5 present the DA strategies, forecasting approaches, evaluation metrics, and experimental design. The results for the deterministic and ensemble forecasts are successively presented and discussed in Sect. 3, followed by Sect. 4 with an extension of the analysis to the French basins and using more recent forecast products. Section 5 presents the main conclusions.

2 Materials and methods

2.1 Dataset

The CAMELS-US dataset (Addor et al., 2017) consists of basin-averaged hydrometeorological time series, catchment attributes, and daily streamflow observations from the United States Geological Survey (USGS) for 671 catchments across the Contiguous United States (CONUS). The meteorological forcings are available from Daymet, NLDAS, and Maurer sources. CAMELS-FR provides the same types of data for French catchments, of which a subset of 338 basins is considered in this study. As this study builds upon the benchmark works of Kratzert et al. (2019) and Newman et al. (2017), hereafter referred to as LSTM and SAC-SMA, it is limited to the same subset of 531 basins, the Maurer forcings, and

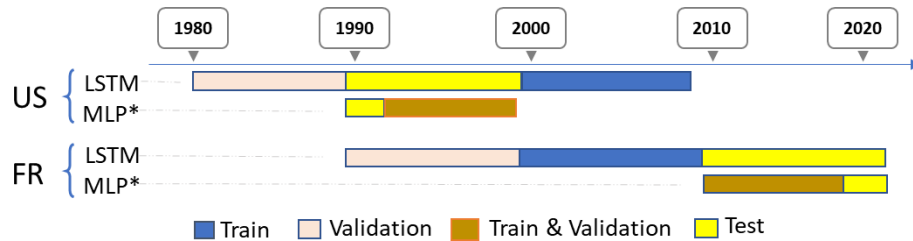


Figure 1. Train-test-validation split for both CAMELS-US and CAMELS-FR datasets. The test set (yellow), training set (blue), validation set (salmon), and combined training&validation set (marron) are indicated; the latter corresponds to training performed using cross-validation. MLP* denotes the orchestrator used for discharge assimilation. Note that the entire modeling process of the DA strategies is carried out exclusively on the test period of the initial LSTM (or the benchmark) models.

Table 1. Used features from the two datasets.

Type	Variables	Description	Unit	CAMELS-US	CAMELS-FR
Forcings	PET	Potential Evapotranspiration	mm d ⁻¹	x	x
	PRCP	Rainfall	mm/day	x	x
	SRAD	Incident Solar radiation	W m ⁻²	x	x
	Tmax	Daily maximum temperature	°C	x	x
	Tmin	Daily minimum temperature	°C	x	x
	Vp	Vapor Pressure	Pa	x	x
Target variable	Q.OBS	Observed discharge	mm d ⁻¹	USGS	x
Model outputs	Q.SAC	SAC-SMA simulated discharge	mm d ⁻¹	Newman et al. (2017)	–
	Q.LSTM	LSTM simulated discharge	mm d ⁻¹	Kratzert et al. (2019)	Current study

the 1989–2008 period used in these previous works using the CAMELS-US dataset. For the CAMELS-FR dataset, an LSTM has been developed from scratch under the same approach as in Kratzert et al. (2019), then considered an equivalent benchmark. The usage of these variables is summarized in Table 1.

The added value of the proposed DA strategies is evaluated for two types of RR models: (a) the LSTM proposed in Kratzert et al. (2019), which was trained regionally and incorporates basin-specific static attributes, and (b) the conceptual global model SAC-SMA from Newman et al. (2017). As in Kratzert et al. (2019), the SAC-SMA model has been chosen as a reference to illustrate the performance of conceptual RR models, which remain widely used for operational discharge forecasting.

The train-test-validation split is illustrated in Fig. 1. It depicts how the data is divided for training, validation, and evaluation of the models. While the splitting of the initial models is mainly shown for reporting purposes, it provides a clear view of how the data are positioned for the tested DA strategies.

2.2 Discharge assimilation procedures

As outlined in Fig. 2 and described in Eqs. (1) to (5), three discharge assimilation procedures are tested, integrating ei-

ther recent discharge measurements or simulations from the two RR models and using MLP as the orchestrator.

1. DA-1: direct forecast of discharges \hat{Q}_{t+hp} over the forecast horizon hp with an MLP, fed with the past observed discharges Q^o , observed meteorological variables X^o , as well as meteorological forecasts \hat{X} (see Eq. 1).

$$\hat{Q}_{t+hp} = f\left(Q_{t-p:t}^o, \hat{X}_{t-n:t+hp}, X_{t-n:t}^o\right) \quad (1)$$

2. DA-2: the same approach as in DA-1 but with the forecasts of the RR model Q^s (either SAC-SMA or LSTM) as additional input variables (see Eq. 2).

$$\hat{Q}_{t+hp} = f\left(Q_{t-p:t+hp}^s, Q_{t-p:t}^o, \hat{X}_{t-n:t+hp}, X_{t-n:t}^o\right) \quad (2)$$

3. DA-3: post-processing of the prediction errors of the RR model ε_t (again SAC-SMA or LSTM). In this strategy, the orchestrator is used to forecast the errors ($\hat{\varepsilon}_{t+hp}$) of the RR model over the horizon hp and the prediction errors are then added to the forecasts of the RR model. The assimilation procedure then proceeds in three steps (see Eqs. 3, 4, and 5).

$$\varepsilon_t = Q_t^o - Q_t^s \quad (3)$$

$$\hat{\varepsilon}_{t+hp} = f\left(\varepsilon_{t-p:t}, Q_{t-p:t}^o, \hat{X}_{t-n:t+hp}, X_{t-n:t}^o\right) \quad (4)$$

$$\hat{Q}_{t+hp} = Q_{t+hp}^s + \hat{\varepsilon}_{t+hp} \quad (5)$$

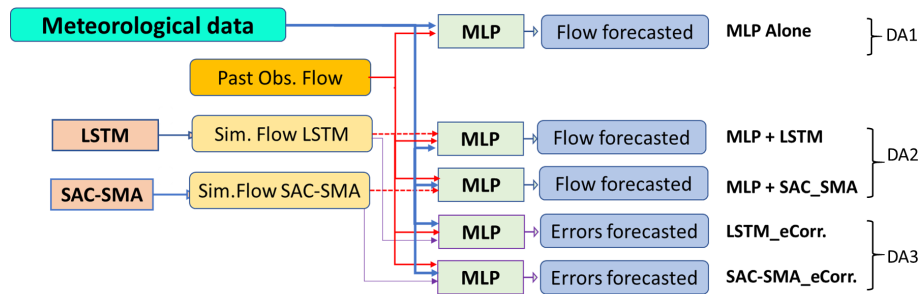


Figure 2. Discharge assimilation set-up: DA1, MLP Alone; DA2, MLP fed with RR model forecasts (MLP + LSTM or MLP + SAC – SMA); DA3, Post-treatment of RR forecasting errors noted as LSTM_eCorr and SAC-SMA_eCorr.

In the previous equations, n and p are the sequence lengths for the forcing and the assimilated discharge. These values will be fixed based on the mean response time of the basins using a RR cross-correlation analysis; see Fig. 5. As suggested in Saint Fleur et al. (2020), to prevent the models from relying disproportionately on assimilated discharge rather than forcing, we imposed $n \geq p$.

In summary, seven (7) different model configurations are compared: the five (5) DA procedures (unfolded from DA1, DA2, DA3) presented in this section, plus the two (2) direct forecasts from both pre-trained models, SAC-SMA and LSTM, which serve as benchmarks to evaluate the efficiency of the tested DA strategies. The direct forecasts from the benchmark models were assumed to be unchanged for the tested lead time; therefore, no further training was necessary.

In all the considered DA strategies and for each basin, the MLPs were trained (i.e., calibrated) 20 times, accounting for the random initialization (seeds) of their parameter values, leading to 20 different possible trained models. Likewise, 8 seeds have been considered for the LSTM and 10 for the SAC-SMA model. This aims to account for the uncertainties and variability induced by model initialization during training. The DA strategies are trained based on the series of mean simulated values of both benchmark models (SAC-SMA and LSTM). The predictions thus consist of an ensemble of 20 runs for the DA strategies and 8 and 10 runs for the LSTM and SAC-SMA benchmark forecasts without assimilation, respectively. The performances of the ensemble simulations (dispersed by random initialization) are analyzed based on their mean values in the first part of this paper (Sect. 3.1).

2.3 Forecasting setup and forecast products

2.3.1 Forecasting setup

In this study, the explored lead times range from 1 to 7 d. As illustrated in Eqs. (1), (2), and (5), the input feature selection for forecasting models incorporating discharge assimilation may be affected by the lead time (hp). In most feedforward architectures, a separate model is calibrated for

each lead time. The alternative which consists of calibrating a single model across the entire range of lead times, either jointly or recursively, is generally inefficient, as it substantially amplifies the forecast uncertainty (Chevillon, 2007; Teräsvirta et al., 2010; Liu and Wang, 2024). This behavior has also been observed in the present study (results not presented herein). It is also worth noting that single-step models may not guaranty continuity of the outputs through successive lead times.

The forecasting framework is summarized in Fig. 3, which illustrates how past observations, assimilated discharge, and forecasted forcing data are integrated. The implementation in DA1 and DA2 procedures is straightforward for both the *perfect* and *ensemble* forecast strategies. However, for DA3 under the ensemble scenario, the corrected quantity corresponds to the forecast member \hat{Q}_{t+hp}^i for which the forecasted error $\hat{\epsilon}_{t+hp}^i$ (in Eq. 4) is issued, where i indicates the forecast member.

All the proposed DA strategies are trained using the *perfect weather forecast* configuration and then evaluated under both the *perfect* and the *ensemble-based* forecast conditions. The ensemble forecast evaluation is conducted using three sources of meteorological forcing: (1) a no-skill ensemble generated from past observations using a date-to-date sampling strategy, referred to as “Climatology”; (2) hindcast (re-forecast) products; and (3) real-time forecast archives provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). Hindcasts correspond to retrospective forecasts produced for past dates to establish a stable statistical reference for ensemble analysis, whereas real-time forecasts are operational predictions issued daily for current and future conditions. Their preparation for this paper is described in Sect. 2.3.2.

2.3.2 Hindcasts, forecast archives and the climatology approach

The operational evaluation is implemented using the sub-seasonal to seasonal (S2S) dataset (Vitart et al., 2017), developed through a joint initiative project of the *World Weather Research Programme* (WWRP) and the *World*

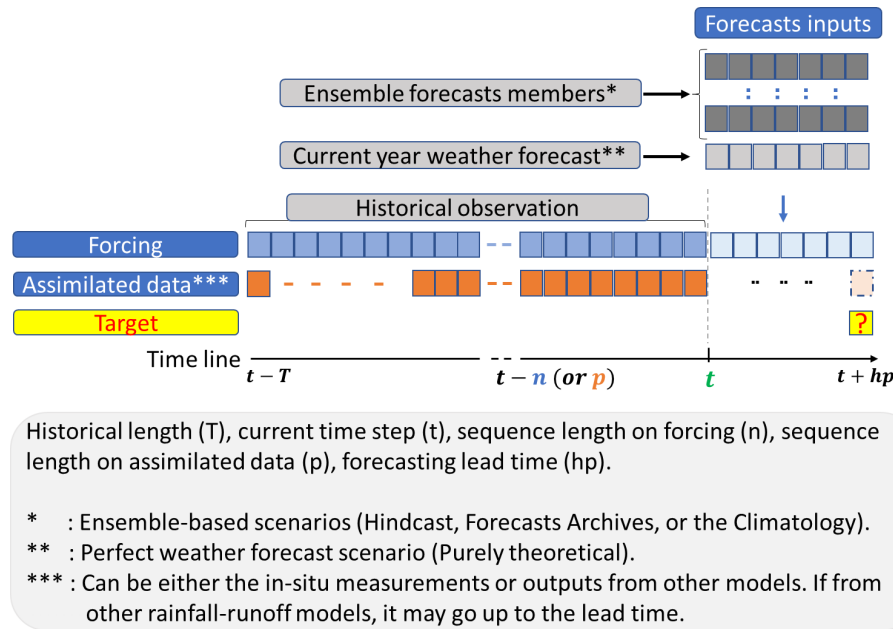


Figure 3. Forecasting assumptions setup.

Climate Research Programme (WCRP). At the time of writing this paper, the S2S database is hosted at ECMWF as an extension of the TIGGE archive. Overall, two forecast products are used: hindcast and real-time forecast archives. Since we evaluated the benchmark models (LSTM and SACSMA) over the 1989–1991 period, only hindcast-based evaluation is implemented on the CAMELS-US dataset because real-time archives are not provided for that period. Consequently, the hindcast product used is from the Bureau of Meteorology (BoM) database (Hudson et al., 2020). Nevertheless, to complement this analysis, the ensemble evaluation has been extended to the french basins using the CAMELS-FR dataset (Delaigue et al., 2025). This extension was specifically implemented on the two main DA approaches tested (LSTM and DA1), using both hindcast and forecast archives for the recent period of 2018–2021. On the ECMWF data portal (<https://apps.ecmwf.int/datasets/data/s2s-realtime-instantaneous-accum-ecmf/levtype=sfc/type=cf/>, last access: 10 February 2026), BoM and ECMWF forecasts are provided as separate products, allowing the use of both hindcast products and real-time forecast archives.

For the present analysis, the perturbed forecast from the BoM dataset was retrieved for up to 7 d of lead time, with all its 32 members. The same method was applied to gather the ECMWF forecast archives (50 members) and hindcast (10 members) products. It is worth noting that these open data are available mostly for 6 to 8 d a month.

We also implement the “Climatology” approach as a baseline, which represents the simplest alternative to archived weather forecasts. It is constructed by resampling historical meteorological observations. Although more sophisticated

sampling strategies could be implemented, for example, by selecting periods of similar hydrological conditions (Hidalgo and Jougla, 2018), the present study adopts a simple date-to-date sampling strategy. For a current date (t_0) within the evaluation period (1989–1991 for CAMELS-US or 2018–2021 for CAMELS-FR), the sequence spanning the lead time ($t_0 : t_0 + h_p$) is defined. The same calendar sequence (day and month) is then extracted for each complete year in the remaining period (1991–2008 or 1989–2017), generating 18 ensemble members for the CAMELS-US cases and 29 members for the CAMELS-FR. This approach constitutes a typical *no-skill* or *poor-man’s* ensemble, as its construction does not explicitly account for the predictability of non-periodic variables such as rainfall data. Nevertheless, it is conceptually similar to the Ensemble Prediction (ESP) framework introduced by Day (1985) and widely used in previous studies (Hidalgo and Jougla, 2018; Crochemore et al., 2017).

At the other end of the evaluation spectrum, the “Perfect forecast” configuration is also implemented. In this case, the forecasted meteorological variables are assumed to be equal to the actual observed values at the corresponding future lead time in the evaluation year. This configuration is particularly useful for estimating the theoretical upper bound of the performance of the models. Overall, four forecast configurations are considered in this study: *Perfect Mode*, *Climatology Mode*, *Hindcast Mode*, and *Real-time Forecast Mode* based on meteorological forecast archives.

2.4 Evaluation metrics

Numerous metrics are proposed in the literature to evaluate the skills of hydrometeorological forecasting models (Mur-

phy, 1993; Seillier-Moiseiwitsch and Dawid, 1993; Bradley and Schwartz, 2011; Lai et al., 2011; Harold et al., 2015; Petropoulos et al., 2022): evaluating the efficiency for deterministic and ensemble predictions, as well as reliability and resolution for ensemble predictions (Bradley and Schwartz, 2011; Slater et al., 2019). The selected evaluation metrics are presented below.

2.4.1 Forecasting efficiency

The *efficiency* is a measure of the proximity between the observed values Q_t and the predicted values \hat{Q}_t . The commonly used metrics for deterministic forecasts are based on the sum of squared errors: Nash–Sutcliffe Efficiency (NSE), Eq. (6) (Nash and Sutcliffe, 1970), the Kling–Gupta Efficiency (KGE) (Gupta et al., 2009), and the Persistence Criterion (PERS), Eq. (7) (Kitanidis and Bras, 1980; Corradini et al., 1986; Anctil et al., 2004).

$$\text{NSE} = 1 - \frac{\sum_{t=1}^T (Q_t - \hat{Q}_t)^2}{\sum_{t=1}^T (Q_t - \bar{Q})^2} \quad (6)$$

$$\text{PERS} = 1 - \frac{\sum_{t=hp}^T (Q_t - \hat{Q}_t)^2}{\sum_{t=hp}^T (Q_t - Q_{t-hp})^2} \quad (7)$$

NSE and PERS are scores that measure the proportion of the sum of square errors of an unskilled model explained by the calibrated (or trained) forecasting model. The unskilled benchmark model for NSE is the trivial mean model ($\hat{Q}_t = \bar{Q}$), and for PERS the persistent model ($\hat{Q}_{t+hp} = Q_t$). Both criteria range from 1 (perfect model) to $-\infty$. A negative value indicates that the model produces higher errors and, consequently, performs worse than the unskilled benchmark models. It should be noted that it is more difficult to achieve a positive PERS than a positive NSE, particularly at short lead times.

For ensemble forecasts, the Continuous Ranked Probability Score (CRPS), Eq. (8) (Hersbach, 2000; Matheson and Winkler, 1976), is commonly used.

$$\begin{aligned} \text{CRPS} &= \frac{1}{T} \sum_{t=1}^T \text{CRPS}_t \text{ with } \text{CRPS}_t \\ &= \int_{-\infty}^{\infty} [F_t(y) - \mathbf{1}_{\{y \geq Q_t\}}]^2 dy \end{aligned} \quad (8)$$

where, for time step t , F_t is the cumulative distribution of the ensemble forecasts, Q_t is the observed value, \hat{Q}_t is the predicted value, \bar{Q} is the time average of the observed values, and $\mathbf{1}_{\{y \geq Q_t\}}$ is the Heaviside-step function for a binary

0|1 outcome. The CRPS ranges from 0 (perfect models) to $+\infty$ (low-quality models). Note that the CRPS is the mean absolute error of the model in the case of a deterministic forecast (i.e. ensemble constituted of a unique member).

2.4.2 Forecasting reliability

An ensemble forecast is considered reliable (or statistically consistent) when the ensemble spread adequately reflects forecast uncertainty, such that the observations are statistically indistinguishable from the ensemble members (Talagrand et al., 1997; Whitaker and Loughé, 1998; Hamill, 2001; Buizza et al., 2005). The resulting distribution of the ranks of a sufficient number of observations, as proposed in Hamill (2001) and Talagrand et al. (1997), provides a visual verification of the reliability of the ensemble forecasts. The lack of reliability may take different forms: (i) a tendency to overestimate (resp. underestimate), leading to an overrepresentation of the lower (resp. higher) ranks in the rank diagram; (ii) under- or over-dispersions of the ensembles, resulting in a *U-shape* or *dome* shape of the rank diagrams. Figure 4 shows the rank diagrams of the evaluation period (1989–1991) throughout the remaining period (1991–2008), for the daily rainfall and PET data.

The rank diagram of the climatological ensemble does not reveal any major deviations from the expected uniform distribution (Fig. 4), suggesting the absence of obvious biases in this ensemble. However, the uniformity is not observed in the hindcast product, which exhibits noticeable underdispersion that may be reflected in the forecasted discharges. This underdispersion, which varies within the lead times (see Appendix A8), remains an open question in the present study. Nonetheless, as mentioned by Hamill (2001), rank diagrams may mask certain defaults in ensemble forecasts; therefore, they will be complemented here by spread-skill ratio (SSR) analysis.

$$\begin{aligned} \text{SSR} &= \frac{\sqrt{\frac{1}{T} \sum_{t=1}^T \sigma_t^2}}{\sqrt{\frac{1}{T} \sum_{t=1}^T (\bar{x}_t - y_t)^2}} \text{ with} \\ \sigma_t^2 &= \frac{1}{N-1} \sum_{i=1}^N (x_{t,i} - \bar{x}_t)^2 \end{aligned} \quad (9)$$

where x and y denote forecast and observed values; N and i , the full-set and individual forecast members; T and t , the evaluation period length and time step. The spread-skill ratio (Eq. 9) is a widely used metric to evaluate the reliability of ensemble forecasts. It compares the ensemble spread (the forecast uncertainty) with the actual forecast error (skill) of the ensemble mean. As formalized by Whitaker and Loughé (1998), it is typically calculated as the ratio of the square root of the mean of the ensemble variance (spread) to the

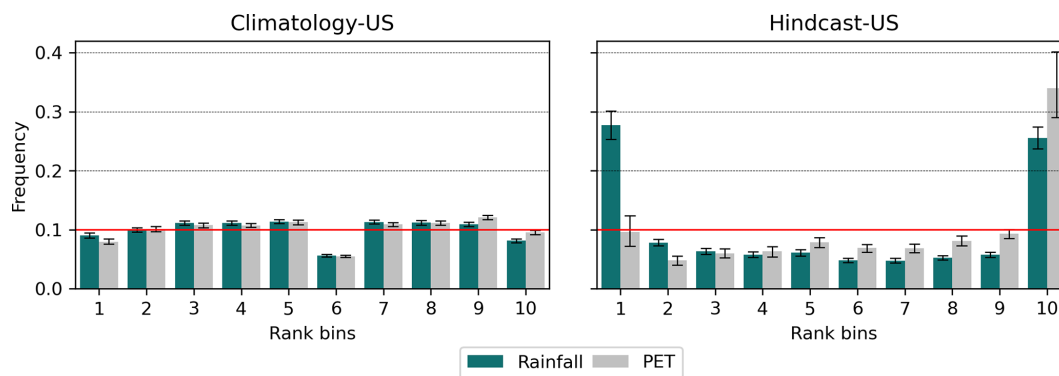


Figure 4. Rank diagrams for daily precipitation and PET for the Climatology-based ensemble (left panel) and Hindcast products (right panel) over the CAMELS-US basins for 3 d lead time. The plots correspond to the evaluation of the test-period (1989–1991) within the remaining 1991–2008 period. The error bars represent variability across the 56 basins considered, and the red line denotes the expected uniform distribution. For ease comparison, the ensembles have been condensed into 10 classes from 17 and 32 members, respectively.

root mean squared error (RMSE) of the ensemble mean. Values close to one indicate a well-calibrated ensemble, while values below (above) one reveal under- (over-) dispersion.

2.4.3 Forecasting resolution

In ensemble forecast verification, resolution refers to the ability of a model to discriminate between events and non-events: i.e., the exceedance or non-exceedance of a given threshold discharge for hydrological predictions. Commonly used metrics for such evaluation include the Brier score (Brier, 1950) and the AUC score (Area Under the Curve) estimated based on a ROC (Receiver Operating Characteristic) curve.

– *Brier score (BS)*

$$BS = \frac{1}{N} \sum_{i=1}^N (f_i - o_i)^2 \quad (10)$$

N is the number of time steps, f_i is the forecast probability of the event according to the ensemble, and o_i is the observed boolean outcome (1 if the event occurs and 0 otherwise).

The Brier score values range from 0 (perfect) to 1 and are equal to 0.25 for a random detection model (i.e., the no-skill model).

– *ROC curves and AUC*

To elaborate on the ROC curve; given a selected target discharge threshold, each rank of the ensemble is selected in turn as the forecast value for event detection. The True positive rate (TPR: proportion of observed events predicted as events) and the False positive rate (FPR: proportion of non-events predicted as events) are computed for each ensemble rank over the evaluation period. The ROC curve relates TPR and FPR. The AUC is the estimated area under the ROC curve. It takes its value between 1 (perfect model, TPR = 1 and FPR = 0 for all ranks) and 0. The ROC

curve of a random detection model corresponds to the diagonal (i.e., TPR = FPR = proportion of predicted events). The AUC value of this random detection model is equal to 0.5.

The forecast resolution may depend on the chosen discharge threshold. To evaluate the ensemble forecasts, several threshold values are tested based on the quantile of the observed discharge series. The considered quantile probabilities are x of 0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, and 0.99. For a given discharge threshold Q_x , an event is recorded whenever discharge values cross this threshold. For thresholds below the median ($x \leq 0.5$), events correspond to low-flow conditions, whereas high-flow (flood) conditions correspond to thresholds above the median ($x > 0.5$). Exceedance is defined based on crossings from above (recession curve) or below (rising curve) the threshold, respectively.

2.5 Experimental settings

2.5.1 Input sequence size and lead time selection strategy

The sizes of the input sequences of the MLPs have been set based on cross-correlation diagrams; see Fig. 5 for the CAMELS-US dataset and Appendix A1 for the CAMELS-FR dataset. The median cross-correlation coefficients were considered in the 531 basins. Following Mangin (1984), a limit value has been chosen for the autocorrelation coefficient for discharges of 0.2 to fix the length p of the input sequence for the assimilated discharges. The sequence size (n) of the forcing has been set to 30 d as an arbitrary value along the flattened portion of the RR cross-correlogram.

The correlation coefficients between observed discharges and daily rainfall amounts are highest for lag times between 1 and 3 d, suggesting that the basins of the CAMELS-US sample have, on average, short response times, typically of less than 3 d. This ensures that the evaluated 1 to 7 d lead times cover both short- and mid-range forecast horizons. Ac-

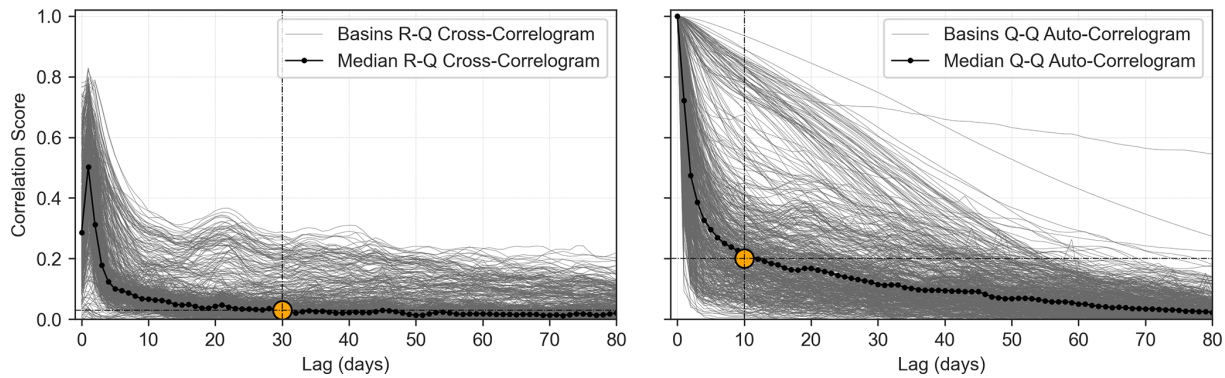


Figure 5. Rainfall–Discharge cross- and auto-correlation on the CAMELS-US dataset; see Appendix A1 for the CAMELS-FR case. The chosen sizes (n and p) of the input sequences are marked with the dashed lines and an orange-colored dot.

According to the response times of the basins, it is anticipated that short-term predictions 1 d ahead will be partly controlled by past observed rainfalls, whereas mid-term 3 to 7 d forecasts will be mostly determined by predicted rainfalls.

2.5.2 Basin sub-sampling for the climatological ensemble runs

The evaluation of the ensemble-based forecast may be numerically demanding: 7 lead times, 7 model configurations, 20 randomly initialized models, 10 to 50 forecast members, and numerous trials for model hyperparameter searching and training. To keep reasonable computation times, the ensemble-based evaluations were conducted on a subset of 56 basins from the initial set of 531 basins. This subset of basins was selected uniformly according to their NSE rank from Kratzert et al. (2019), covering the same range of basins as the initial sample of 531 basins (Fig. 6). For the CAMELS-FR basins, the selection was based on the completeness of the discharge time series, with total missing data not exceeding 90 d, while ensuring that all regions (basins coded from A to Y) are represented. The lists of selected basins are provided within the code availability.

2.5.3 Softwares and hyperparameter settings

For the orchestrator (MLP) configurations, the hyperparameters listed in Table 2 were optimized using exhaustive grid search and cross-validation with respect to the used datasets. The hyperparameter subset was derived from a larger space using 20 randomly selected basins, retaining the most frequent configurations. The hidden sizes ranged from a single layer of 30 neurons to four layers with multiples of 30 neurons. Five levels of learning rates (10^{-1} to 10^{-5}) were primarily tested, and two have been retained based on their occurrences as the best values.

The experiments developed in this study are essentially based on open-source software and the Python 3.9 programming language (van Rossum, 1995). Our modeling frame-

Table 2. Model hyper-parameter setup.

Parameters	Parameter space
Hidden layers [size,]	[120, 90] [120, 90, 60]
Activation	[relu, tanh]
Learning rate	[0.01, 0.001]
Solver	ADAM
Early-stopping	True
No_iter_no_change	15
Validation fraction	[0.2]
Random number of random initialisation	20
Sequence depth	30 on forcings, 10 on assimilated data

work is based on the Scikit-Learn library (Pedregosa et al., 2012). Data analysis, processing, and visualization are performed mainly using Pandas (McKinney, 2010), Numpy (van der Walt et al., 2011), Seaborn (Waskom, 2021), Matplotlib (Hunter, 2007), and Xskillscore (Bell et al., 2021). The model development was carried out using Jupyter Notebook (Kluyver et al., 2016), Anaconda (Anon, 2020), and PyCharm (JetBrains, 2024).

3 Results on the CAMELS-US dataset

The performance of the three discharge assimilation (DA) approaches is evaluated against the benchmark models across the considered forecast scenarios. This comparison emphasizes the differences in model behavior between an idealized setting (perfect forecast scenario) and several ensemble-based forecasts. The variability of the scores across the basins is illustrated using boxplots and error bars. To intro-

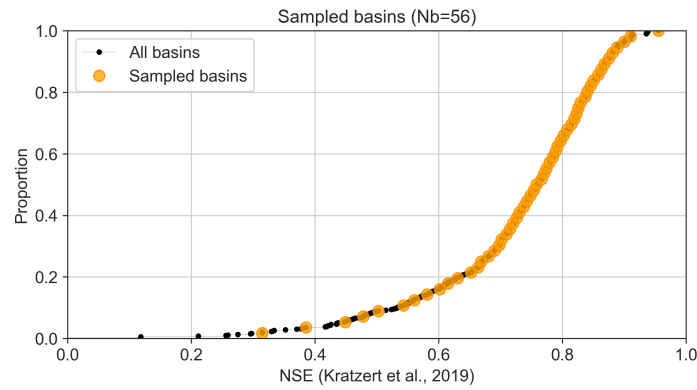


Figure 6. Cumulative distribution of the NSE scores of the 531 US-basins from the regional LSTM of Kratzert et al. (2019) and the selected subset of 56 basins.

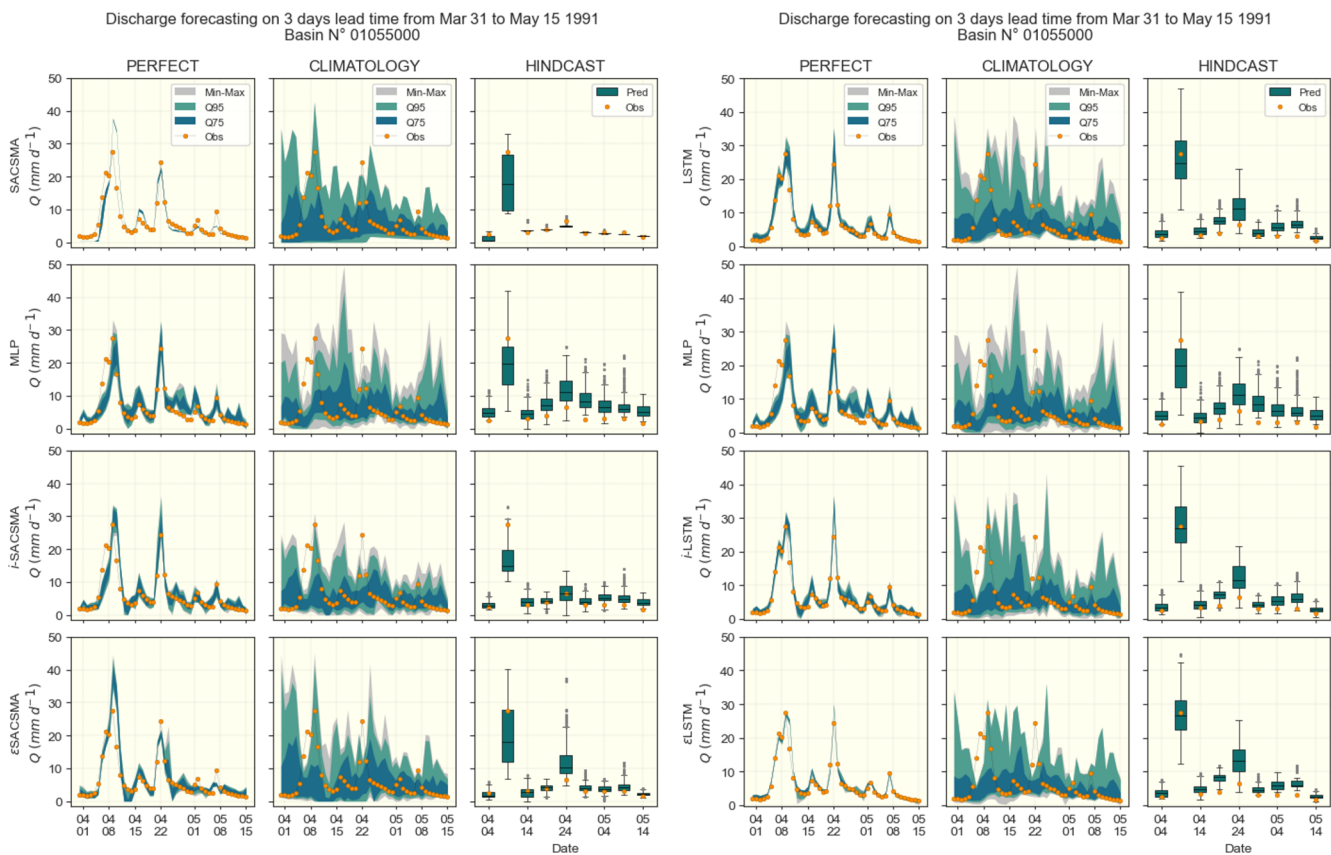


Figure 7. Examples of hydrographs on basin 01055000 of the CAMELS-US dataset for a 3 d lead time rainfall–runoff forecasting. SAC-SMA (left panels) and LSTM (right panels) cases are shown separately; rows indicate (benchmark models, DA1, DA2, DA3), and columns points to the meteorological forecasting approaches (Perfect, Climatology and Hindcast). Since hindcast products are available only 6 times a month, their outputs are discontinued and shown through box-plots for easier display. Color ranges are used to highlight ensemble quantile ranges $Q_{75} = [0.125, 0.875]$, $Q_{95} = [0.025, 0.975]$ and [Min, Max] values, while the observed discharge values are marked with orange-dots.

duce the results, an example of hydrographs (observed and forecasted) is presented in Fig. 7. This example corresponds to basin No. 01055000 from the CAMELS-US dataset over the period from 31 March to 15 May 1991, and includes the three tested forecasting approaches (Perfect, Climatology,

and Hindcast) along with both benchmark models (LSTM and SAC-SMA). All presented results concern the test set.

Figure 7 shows what each of the forecast results looks like. For illustration purposes, we selected a case in which both the benchmark models and the meteorological hindcast pro-

vide accurate results. The performance metrics and scores of the different approaches for the various lead times, evaluated in the full set of 531 CAMELS-US basins, are presented and discussed in the following sections.

3.1 Performances of the DA approaches based on perfect meteorological forecasts

3.1.1 Forecasting efficiency

As an introduction, Table 3 displays an overview of the NSE values and gains of the discharge assimilation methods tested in this study for the 1 d lead time forecast, compared to the results published in Nearing et al. (2022), which tested discharge assimilation using an LSTM on the same CAMELS-US dataset. Note that the test period differs between the study of Nearing et al. (2022) (1989–1999) and the present study (1989–1991). Table 3 also includes the results obtained on the CAMELS-FR dataset, which are presented in more detail in Sect. 4. It shows that NSE scores are highly dependent on the datasets and that the relative gains from discharge assimilation methods tend to be greater when the benchmark models have lower NSE values.

Overall, the improvements achieved by the DA strategies developed here are globally consistent with those reported in Nearing et al. (2022). NSE gains range from 8 % to 12 % for the LSTM model and reach up to 22 % for the conceptual SAC-SMA model. As explained in Sect. 2, the remaining analysis is based on the persistence analysis (Fig. 8), which provides more contrasted results than the NSE score.

As expected, the PERS scores (Fig. 8) are lower at short lead times. This is a common outcome in persistence analysis, as models generally struggle to outperform the persistent model at very short horizons: the smaller the discharge variations, the more difficult they are to predict. Furthermore, in agreement with previous studies, performances are significantly higher for the LSTM-cases than for the SAC-SMA and this trend persists even when DA procedures are implemented.

The DA1 method appears to be more effective than the SAC-SMA benchmark in all the lead times tested. However, it only clearly outperforms the benchmark LSTM model in the 1 d lead time when the initial PERS scores of the LSTM are modest: the median PERS values lower than 0.5 and the PERS values lower than 0 for 20 % of the basins.

For further clarity, Fig. 9 summarizes the gain in PERS scores achieved by the different DA procedures relative to their corresponding benchmark models. Almost without surprise, these gains are highly dependent on the initial PERS value of the benchmark model. Three classes of initial benchmark PERS values are considered in the figure to illustrate this dependency: $(-\infty, 0]$, $[0, 0.5]$, and $[0.5, 1]$. The two other DA strategies, DA2 and DA3, both based on the benchmark models (i.e., MLP-informed and error post-processing), prove to be effective, as they consistently im-

prove the performance of the benchmark forecasting models on which they are based. DA2 outperforms DA1, while the DA3 approach generally enhances performance or, at least, preserves performance when it is already high.

The two Figs. 8 and 9 show that the gains are larger for shorter lead times, more pronounced for SAC-SMA than for LSTM due to the overall lower scores of the SAC-SMA model, and lower for basins where the initial model already performs well. In general, the ranking of the approaches tested is strongly dependent on the initial skills of the benchmark models. DA2 appears to be the most effective approach overall, followed by DA3 (Fig. 8).

These results demonstrate the effectiveness of the proposed DA strategies in improving forecast performance under a perfect meteorological forecast scenario. Gains are particularly significant for the 1 d lead times. However, the added value of the proposed DA decreases rapidly with increasing lead times (Fig. 9). This decline can be explained by both the increase in the PERS benchmark models with lead time and the overall short response times of the basins in the CAMELS-US dataset, which typically range from one to a few days. This is depicted by the cross-correlation analysis (Fig. 5), which explains the reduced influence of past discharges on future trajectories for horizons exceeding these response times.

The next step consists of assessing whether these conclusions remain valid when taking into account uncertainties in meteorological forecasts, a situation that corresponds to the operational implementation of rainfall–runoff forecasting models. To streamline the discussion while considering the superiority of the LSTM model compared to Sac-SMA and other possible conceptual rainfall–runoff models, only the LSTM cases are considered in the remainder of the present manuscript.

3.2 Performances of the DA approaches under ensemble-based forecast scenarios

As a reminder, for the examples based on the CAMELS-US dataset, the ensemble forecast scenario is implemented using the historical meteorological records (i.e., Climatology) and the BoM hindcast data. According to the adopted sampling strategy, the Climatology-ensemble consists of 18 members (1991–2008), whereas the hindcast ensemble consists of 32 members, as provided by the data source. The hindcast product is discontinuous, as only 6 predictions are issued within a month. To account for model uncertainties, these ensembles are further expanded through random model initialization: 8 realizations for the LSTM and 20 for the DA approaches. To limit computational costs, the evaluation is conducted in a representative subset of 56 basins, selected to cover the range of LSTM NSE (test) values observed in the 531 basins of the CAMELS-US dataset, as shown in Fig. 6. Three key properties of the forecast ensembles are evaluated here: (1) their efficiency based on the CRPS score,

Table 3. NSE and improvements at a 1 d lead time across tested discharge assimilation approaches and benchmark models from several studies.

Approach	Nearing et al. (2022)			This study CAMELS-US				This study CAMELS-US				CAMELS-FR	
	LSTM	AR	DA	LSTM	DA1	DA2	DA3	SAC-	DA1	DA2	DA3	LSTM	DA1
NSE	0.80	0.88	0.86	0.74	0.80	0.83	0.82	0.67	0.80	0.82	0.80	0.91	0.95
Gain*		10 %	8 %		8 %	12 %	11 %		19 %	22 %	19 %		4 %

* Gains are estimated relative to the benchmark model NSE. In Nearing et al. (2022), AR and DA refer to the methods called auto-regressive and discharge assimilation respectively. SAC- refers to SAC-SMA model.

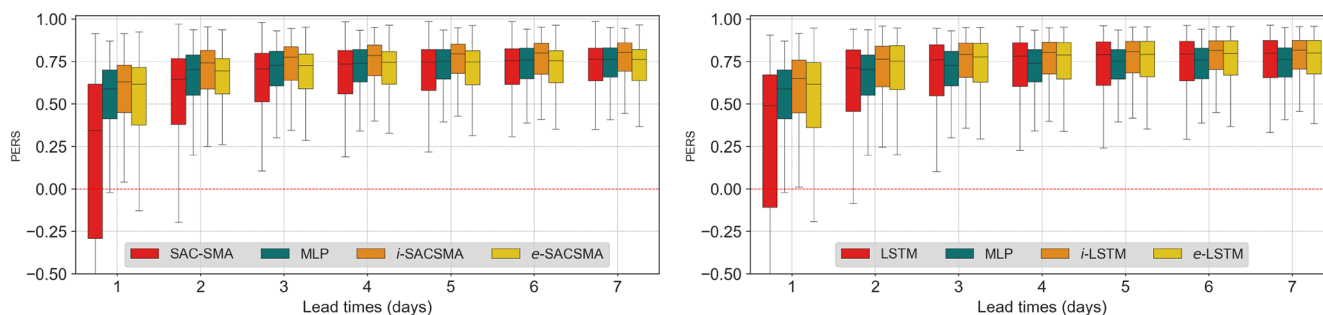


Figure 8. Box-plots of the persistence (PERS) scores. The figures are ordered with SACSMA-cases first, followed by LSTM-cases. Lead times (1–7 d) are shown on *x* axis, while scores are displayed on the *y* axis. Color-codes distinguish the approaches: benchmark models (red), DA1 or MLP (green), DA2 (orange) and DA3 (gold). DA1 is replicated in both benchmark cases. In the legend, *i*-LSTM stands for DA2 or MLP-informed by LSTM, *e*-LSTM stands for DA3 or error post-processing approach on the LSTM case.

(2) their reliability based on rank diagrams complemented with spread/skill ratios (SSR), and (3) their resolution using Brier and AUC scores.

3.2.1 Forecast efficiency

Figure 10 presents the CRPS values for both forecast scenarios based on climatological ensembles and forecasts in all approaches and lead times tested for DA. Two baseline models are included in the analysis: the persistent model (forecast equal to the last observed discharge) and the past-observed (P.O) discharge model, which consists of discharge observations from previous years on the considered date. Since the Persistent-Model produces a single deterministic prediction, its CRPS is reduced to the mean absolute error (MAE). In contrast, the P.O model comprises 18 members and is therefore treated like all other ensemble forecasts.

According to Fig. 10, all tested models appear more efficient than the persistent baseline model for all lead times, even when accounting for uncertainties in the ensemble forecasts. The performance gap between these models and the persistent model becomes larger as the forecast lead time increases.

In the climatology-based scenario (left-most), the models consistently outperform the baseline observed in the past (P.O.), signifying that all tested models and approaches remain informative even at the larger lead times. However, this pattern is not consistently observed in the hindcast-based

scenario for lead times exceeding 3 d. The observed biases in the BoM hindcast products for the period 1989–1991, and for the estimated basin average daily PET and rainfall (Fig. 4), clearly limit the efficiency of ensemble forecasts based on these hindcasts for the CAMELS-US basins. A detailed analysis of the structure of these biases, along with the development of an appropriate bias correction method (Zalachori et al., 2012; Yang et al., 2020) would be essential to fully exploit the potential of these hindcasts. However, this likely complex task was beyond the scope of the present study.

Finally, the observed trends in Fig. 10 (climatology) are consistent with those observed using the PERS criterion under the perfect meteorological forecast scenario (Fig. 8) with some nuances. The DA approaches, including DA1 (simple MLP), remain effective, as they globally improve the performance of the LSTM model or, at least, do not degrade it for any of the tested lead times. Their added values are also more pronounced at shorter lead times.

3.2.2 Forecast reliability

Figure 11 shows the rank diagrams for both the *climatology*- and *hindcast-based* scenarios with the CAMELS-US dataset. The ensemble members have been grouped into 10 classes for all models to facilitate comparisons. The figure is organized vertically, and the results of the LSTM benchmark model are shown in the first row, followed by the three tested DA approaches.

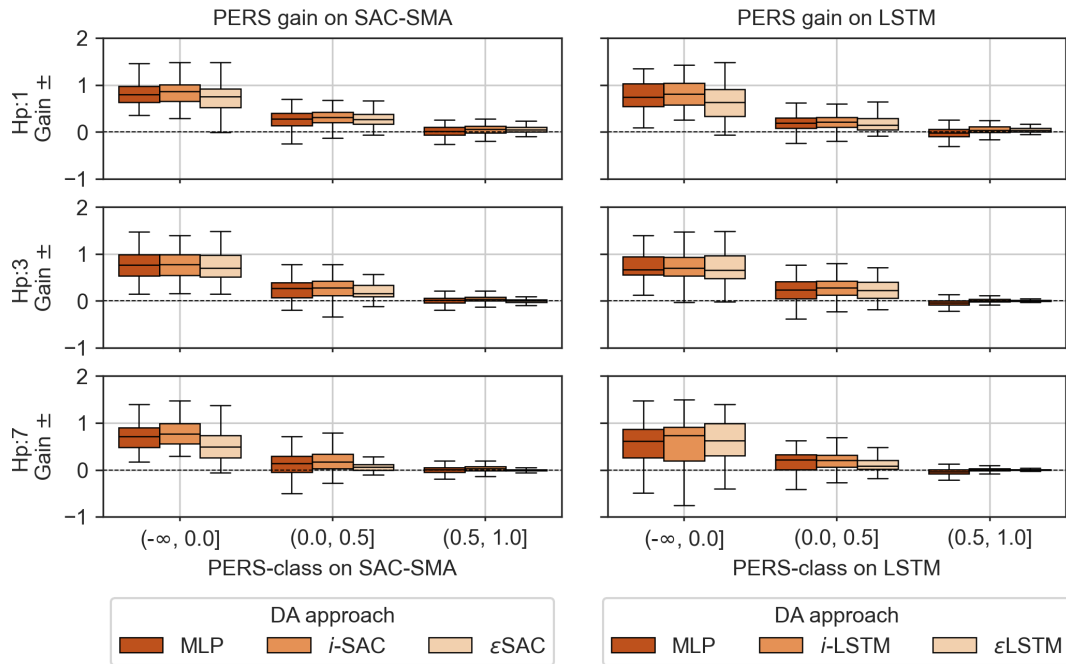


Figure 9. Gain on Persistence. Where, Gain = DA – Benchmark. For lighter nomenclature, the following names have been respectively used: MLP simple (MLP), MLP Informed by BM (*i*-SAC or *i*-LSTM), benchmarks error post-processed (ϵ SAC or ϵ LSTM).

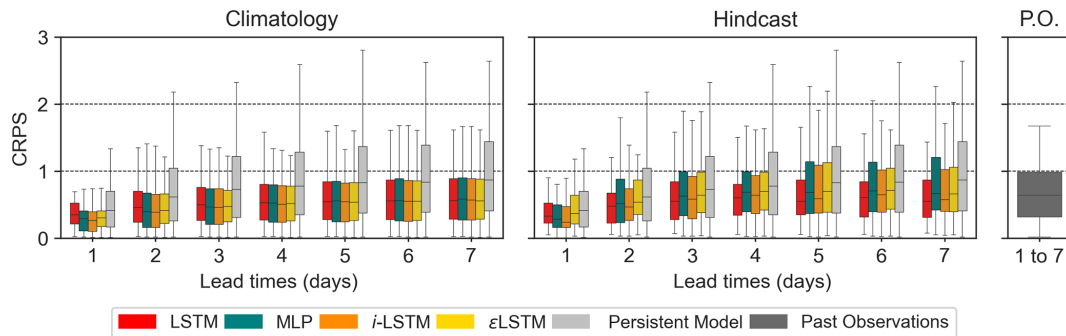


Figure 10. Box-plots of the CRPS scores for the 56 tested basins for the 1989–1991 test period. Lead times of 1 to 7 d are shown in *x* axis, while scores are displayed in *y* axis where 0 denotes perfect forecasts. Colors indicated the LSTM benchmark (red), DA1 (green), DA2 (dark-orange), DA3 (gold), Persistent Model (gray) and Past-Observed discharge (dark gray).

Reliable forecasts are expected to yield uniformly distributed rank diagrams, indicating ensemble forecasts in which actual events are evenly distributed across all forecast member ranks. It should be noted first that the rank diagrams are similar across all lead times for a given model and meteorological ensemble product, and they differ between models, indicating that the rainfall–runoff forecasting model, including the discharge assimilation procedure, has an impact on the spread and possible biases of the forecast ensembles. The rank diagrams indicate that the hindcast biases (Appendix A6) propagate in all models and methods tested, providing an explanation for the lower observed CRPS values compared to the climatology-based scenario. The U-shape of the hindcast-based forecast rank diagram suggests that

the forecast ensembles may be, on average, under-dispersed. This pattern is not evident when looking at model outputs (Fig. 7), but it seems to be confirmed by the spread-skill ratios, which are significantly lower for hindcast-based forecasts than those of the climatology-based forecasts (Fig. 12). A slight deviation from the uniform distribution also appears in the rank diagram of the LSTM climatology-based ensemble forecasts. The spread-skill ratios of the LSTM model appear similar to those of the DA1 (MLP), DA2, and DA3 approaches, suggesting that the possible biases in the ensembles generated by the rainfall–runoff model are more complex than simple systematic under-dispersion.

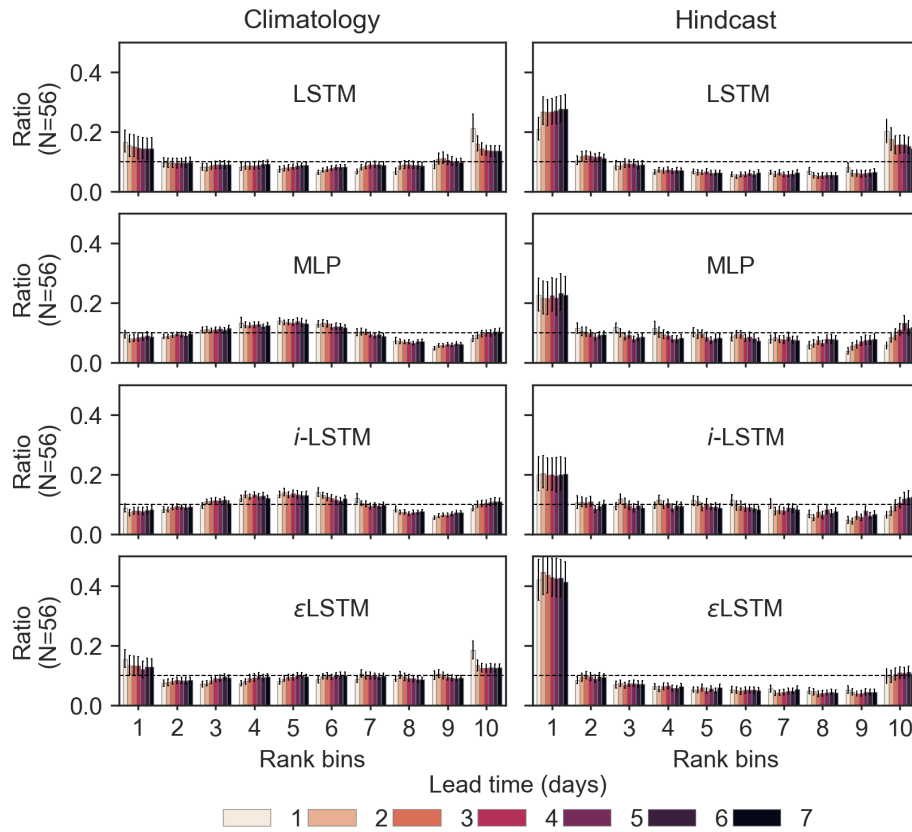


Figure 11. Rank diagrams for the LSTM-cases models and the DA strategies. x axis (10 rank classes), y axis (proportion of observed values in each class), median ratio and error-bars indicating the distributions for the 56 basins. Colors indicate the lead times.

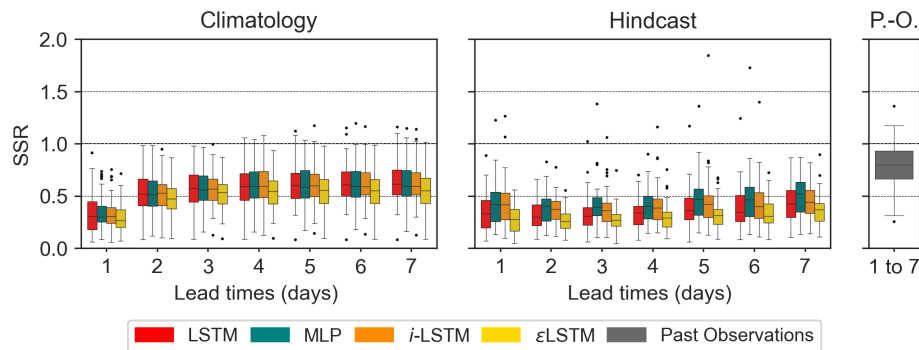


Figure 12. Box-plots of the spread-skill ratio for both climatology- (left panels) and hindcast-based (right panels) forecast scenario for the subset of 56 basins. LSTM-cases (LSTM, DA1, DA2 or i -LSTM, DA3 or ϵ -LSTM) cases are shown including the Past-observed (P.O.) discharge model.

3.2.3 Forecast resolution

The Brier’s and AUC scores evaluate the ability of an ensemble forecast to anticipate events and non-events; for instance, the exceedance or non-exceedance of a selected discharge threshold. Their values are presented in Fig. 13 and 14. As expected, the resolution of all forecasting approaches tested decreases with increasing lead times; i.e., the computed brier scores and AUC get closer to the values obtained based on

past observations only for all thresholds as the lead time increases. The resolution analysis also confirms the poor skill of the ensemble forecasts based on the hindcast as used here, which is particularly noticeable in the Brier scores (Fig. 13).

All the approaches tested outperform the random detection model (Brier = 0.25) and generally surpass the P.O model under *climatology-based* forecasts, with little to no improvement under the *hindcast-based* forecast. The earlier find-

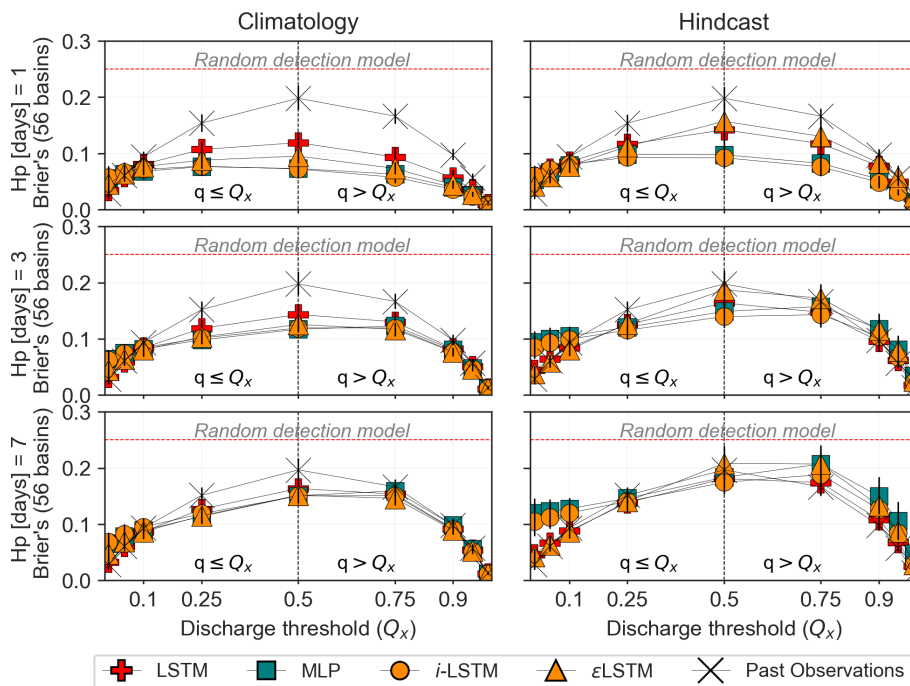


Figure 13. Brier's Scores for event detection based on thresholds using discharge quantile (Q_x) for both low flow ($q \leq Q_x$) and high flow ($q > Q_x$) values. Median scores and error-bars are shown, indicating the dispersion across the subset of 56 test basins. Past-observed discharge is also evaluated as a poor man's discharge forecast and represented by the X symbols.

ing that climatology-based ensembles tend to outperform hindcast-based forecasts in terms of forecast resolution is also observed here.

The previous conclusions also hold for the model resolution: the proposed DA strategies prove effective. They either significantly improve the skill of the LSTM benchmark or, at least, do not degrade its initial performance.

This is particularly clear in the Brier scores (Fig. 13), especially for short lead times and intermediate discharge thresholds. The AUC graph shows less pronounced contrasts (Fig. 14). For a more in-depth comparison between methods, an example of the Roc curves obtained for the threshold quantile 0.95 is presented in Appendix B2. It illustrates the complexity of the comparison: the relative ranking of the models depends on the lead times, criteria, range of considered discharge values or thresholds, and also the target probability of detection in the ROC curve.

Two additional observations can be drawn from the AUC figure (Fig. 14), despite its limited contrast. First, the gap between AUC values based on past observed discharges and those of the tested forecasting approaches is particularly pronounced for high-threshold quantiles at a 1 d lead time. This suggests that the tested approaches are particularly well-suited to predicting the exceedance of high discharge values, which is consistent with the fact that the standard root mean square error criterion, known to place greater emphasis on large discharge values (Terven et al., 2025), has been used to train all models and methods.

More surprisingly, for large discharge thresholds at 3 and 7 d lead times, the AUC scores obtained with hindcast-based approaches exceed those associated with climatology-based forecasts. This indicates that, despite their apparently lower overall skill, the hindcast products used contain valuable information compared to climatology for predicting intense rainfall-triggered events.

These observations further illustrate how conclusions drawn from model comparisons depend on the target variable used to train the model, the range of values considered, and the evaluation metric used. At this stage of the analysis, the following partial conclusions can be drawn:

- The proposed discharge assimilation procedures, particularly DA2 and DA3, prove to be effective, as they either significantly improve or at least do not degrade the performance of the LSTM benchmark model across all considered lead times and evaluation criteria;
- Evaluating rainfall–runoff forecasts based on meteorological ensembles is a necessary complement to analyzes that are often conducted under the implicit assumption of perfect meteorological forecasts. In the present case, this approach reveals that the superiority of the LSTM and LSTM-based discharge assimilation methods over the proposed simpler MLP model, observed for lead times greater than two days, disappears once meteorological uncertainties are taken into account.

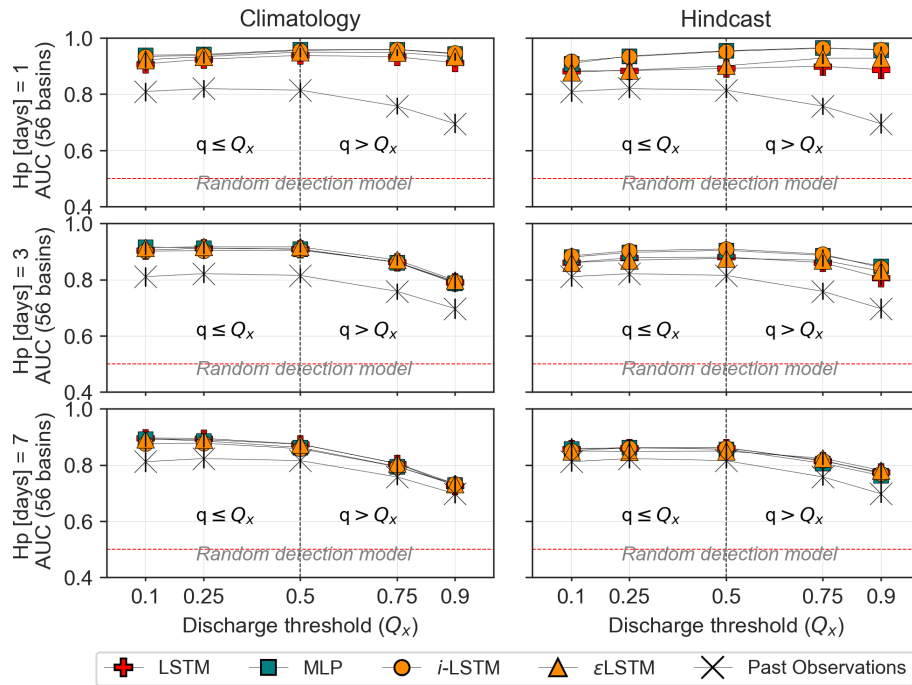


Figure 14. AUC scores for events based on flow quantile [0.1, 0.25, 0.5, 0.75, 0.9], with drought/flood detection shifting at quantile 0.5. These shown values correspond to the median AUC values across 56 basins. Climatology and Hindcast are shown in rows, lead times 1–3–7 d are presented in columns. Past-observed discharge is also evaluated as a poor man’s discharge forecast and represented by the X symbols.

Nevertheless, the analysis is limited by the low skill of the available hindcast products for the selected test period (1989–1991) in the CAMELS-US dataset. It is therefore proposed in Sect. 4 to implement some of the tested approaches on a more recent dataset (CAMELS-FR), for which additional ensemble meteorological forecast products are available. The objective of this extension is twofold: (1) to assess the robustness and generality of the conclusions drawn from the CAMELS-US case study, and (2) to evaluate ensemble forecasting skill using more recent and probably higher-quality meteorological ensemble forecasts produced by the European Center for Medium-Range Weather Forecasts (ECMWF).

In line with the conclusions of this section, and for the sake of simplicity, the analysis in Sect. 4 is restricted to the benchmark LSTM model and the DA1 (MLP) strategy, evaluated under the same framework as previously. The analysis relies on hindcast products as well as forecast archives, providing an evaluation of the predictive skill of these two ensemble rainfall–runoff forecasting models, had they been implemented in the past.

4 Extension to the CAMELS-FR dataset

To ensure consistency with previous studies, such as Kratzert et al. (2019) for the CAMELS-US dataset and Hashemi et al. (2022) for French basins, Fig. 15 illustrates the position of

the NSE values for the LSTM and DA1 approaches implemented in this extended analysis using the CAMELS-FR dataset (Delaique et al., 2025). The results indicate that the trained LSTM achieves a high level of performance on the CAMELS-FR dataset, with median NSE values reaching 0.9.

Furthermore, consistent with the comparison presented in Sect. 3, the DA1 outperforms the LSTM at the 1 d lead time and exhibits NSE values comparable to those of the LSTM model at longer lead times. The NSE values obtained on the same datasets with the conceptual GR4J model (Perrin et al., 2003), a reference model in France, are also shown. These results confirm that AI-based rainfall–runoff forecasts outperform traditional conceptual rainfall–runoff models on the French dataset, although the performance gap is less pronounced than that reported in Kratzert et al. (2019) for US basins.

It can also be observed in Fig. 15 that the NSE values increase from left to right. Since the LSTM architectures and implementation strategies are similar across the considered studies ¹, this increase may be partly explained by the improvement over time of the model training algorithms but is probably mainly attributable to the datasets; the recently published CAMELS-FR dataset consists of records from basins

¹Regionally trained LSTM models with static attributes of basins, input sequence lengths of 270 d, a loss function of mean square error, and a hidden size of 256.

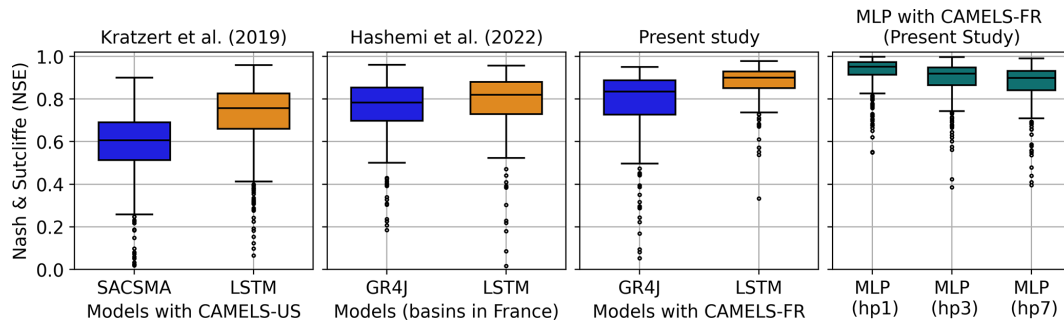


Figure 15. NSE scores comparison between LSTM and SACSMSA for 531 US-basins with Kratzert et al. (2019), LSTM and GR4J (Perrin et al., 2003) on 365 French basins with Hashemi et al. (2022), and the ongoing LSTM vs GR4J and MLP (DA1) for 338 basins from the CAMELS-FR dataset.

with limited anthropogenic influence and has undergone extensive quality control (Delaique et al., 2025).

Figure 16 illustrates, using an example of hydrographs from the CAMELS-FR experiment, what the outcomes of the various approaches tested look like. The 3 d lead time forecast is presented here, while the corresponding 1 and 7 d lead times are provided in Appendix B5. Nevertheless, no general conclusion can be drawn from this isolated example regarding the relative performance of the various methods. Furthermore, direct pairwise comparisons between *hindcast* and *forecast* archives are not possible, as the dates for which the hindcast and forecast archives are available are not strictly aligned. The aggregated evaluation metrics are presented hereafter.

4.1 Model efficiency analysis

The PERS scores obtained by the LSTM and DA1 approaches for the CAMELS-FR (Appendix B6) exhibit trends similar to those observed in the CAMELS-US analysis; however, the median PERS of the MLP (DA1) method remains higher than that of LSTM up to the 5 d lead time. This can be partly explained by the difference in hydrological inertia of the basins between the two datasets, as shown in Appendix A1 and previously discussed by Pelletier and Andréassian (2024). In the same line of thought, the spread of PERS scores for the DA1 remains more limited than that of the LSTM model across all the tested lead times. While these differences may also partly originate from variations in dataset quality and initial model performance, they also reflect the contribution of the assimilated discharges, which certainly plays a key role.

4.2 Ensemble forecast analysis (efficiency, reliability and resolution)

In this subsection, the complete ensemble analysis is provided, using CRPS scores (Fig. 17) for the efficiency analysis, the Rank diagram (Fig. 18) for reliability, and Brier's scores (Fig. 19) for the resolution of the ensemble forecasts.

The Spread-Skill ratio and the AUC scores are provided in Appendices B7 and B9, respectively.

As shown in Fig. 17, CRPS values are generally lower here than those reported previously, with most values falling below 0.5 across all forecasting approaches, including the Climatology-based method. All tested methods (LSTM and DA1) successfully outperform both the persistent model and the no-skill past observed (P.O) discharge ensembles.

Unlike in the CAMELS-US case, meteorological ensemble forecast products (hindcast and forecast archives) demonstrate better performance than the climatology-based ensemble, particularly for lead times exceeding 2 d. This is further supported by the CRPS scores (Appendix B8), estimated using the climatology-based forecast as a reference, which indicate that both forecast products outperform this baseline.

Note that this result, counterbalancing the pessimistic conclusion drawn in Sect. 3 regarding meteorological hindcasts, is obtained despite the significant biases observed in both the hindcast and forecast products used in this French experiment (see Figs. A7 and A8).

Consistent with the persistence criterion, the CRPS values obtained with the DA1 (MLP) approach are, on average, lower (better) than those of the LSTM model across all meteorological ensembles and lead times, except for the hindcast at lead times exceeding 4 d.

The rank diagrams (Fig. 18) reveal biases affecting all forecast ensembles. With the exception of the climatology-based MLP forecasts, an excessively high proportion of observed discharges falls outside the [0.1, 0.9] quantile range of the forecast ensembles. This is partly explained by the biases in the hindcast products and the forecast archives (Fig. A7). However, as these proportions are higher for the LSTM model, it is likely that this model also introduces additional biases when combined with weather forecast ensembles.

This issue certainly deserves further investigation to support a more efficient operational implementation of LSTM-based rainfall–runoff forecasting models. Biases in forecast

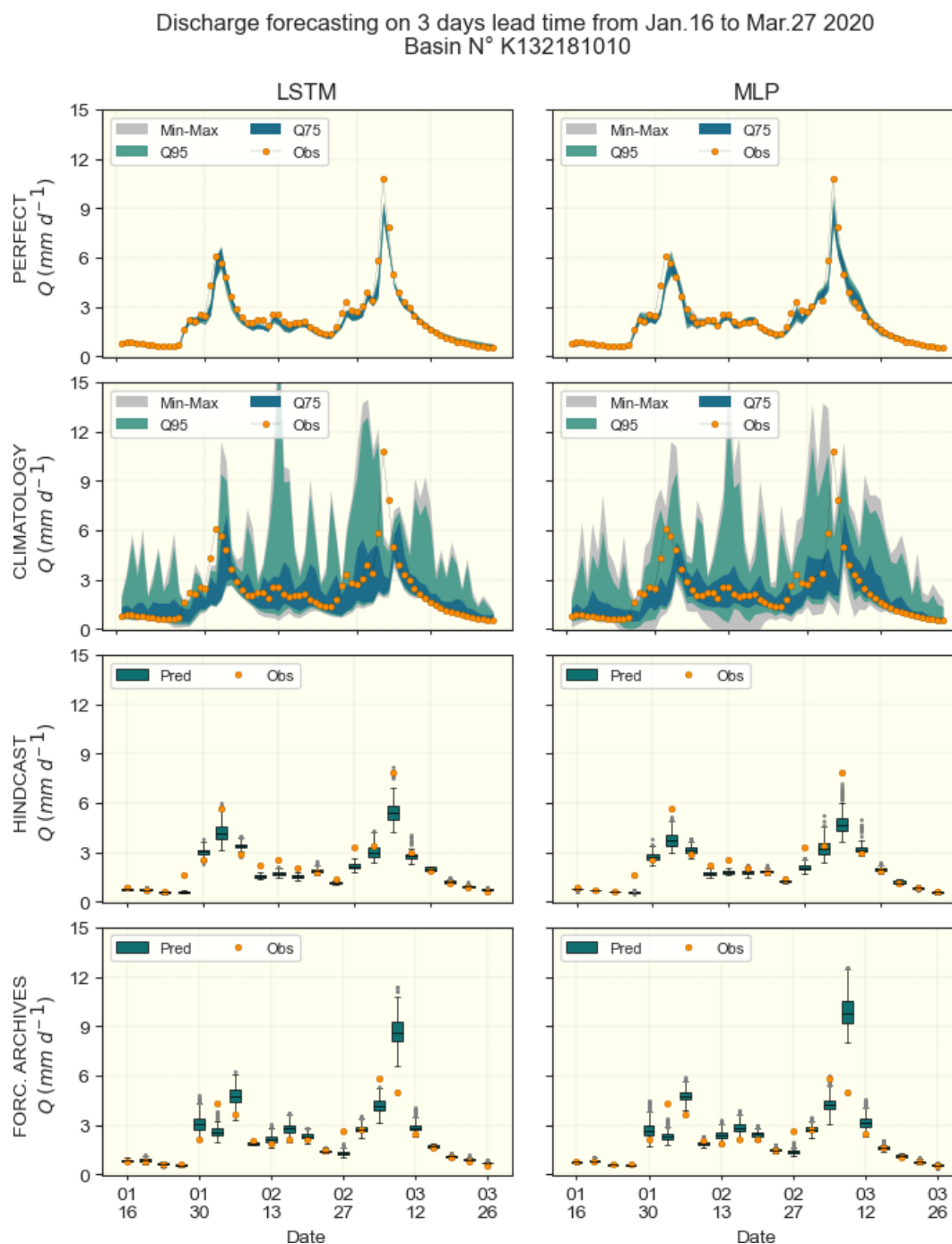


Figure 16. Example of 3 d lead time forecasted hydrograph for the basin K132181010 from 16 January to 27 March 2020. LSTM and DA1 are displayed in columns, while the 4 forecast approaches (Perfect, Climatology, Hindcast and Forecast Archives) are in rows. Given discontinuity in the two last forecast products, they are represented using box-plots.

ensembles reduce the resolution of the forecasts, as the probability of exceedance is less accurately represented.

Consistent with the analysis in Sect. 3, all models and approaches outperform both the random detection and the past-observed discharge ensemble baselines. However, unlike in Sect. 3, this statement clearly holds across all meteorologi-

cal ensembles, lead times, and evaluation metrics (Brier in Fig. 19 and AUC in Fig. B9).

The resolution of the DA1 (MLP) strategy appears higher than that of the LSTM across most tested configurations, with the exception of the Brier scores computed for the low discharge threshold at a 7 d lead time (Fig. 19). In this case, the

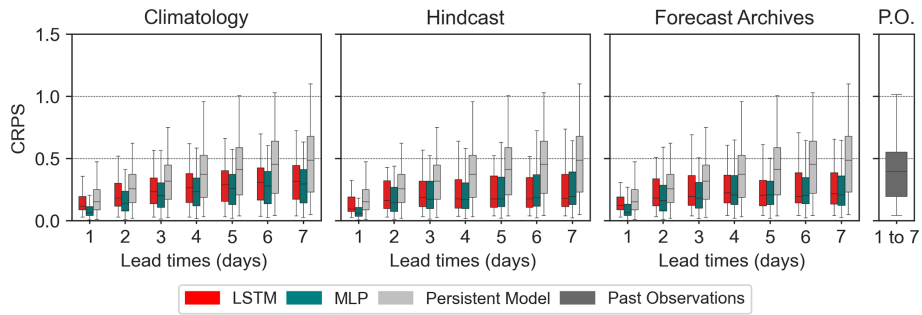


Figure 17. CRPS scores for the LSTM and the DA1 (MLP) with the CAMELS-FR dataset.

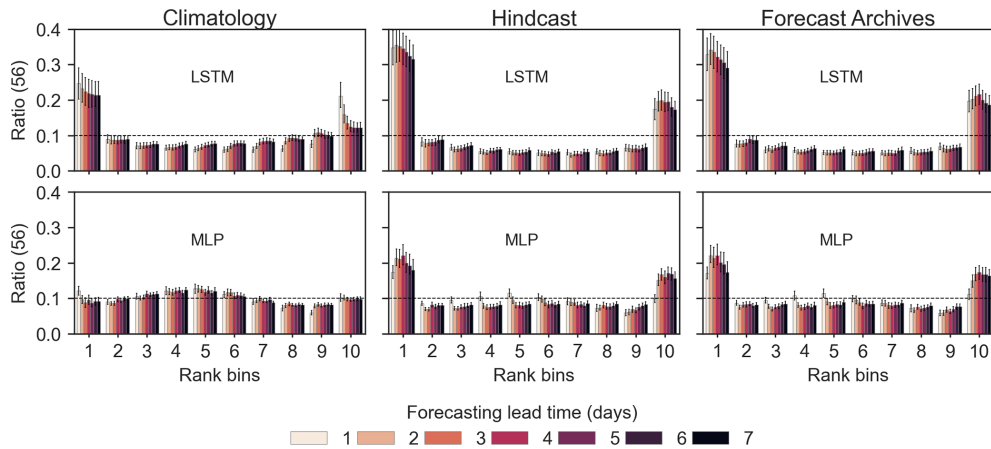


Figure 18. Rank diagrams for the benchmark models and the DA strategies. x axis (10 rank classes), y axis (proportion of observed values in each class), median ratio and error-bars indicating the maximum and minimum ratios for the 56 test basins. Colors indicate the lead times.

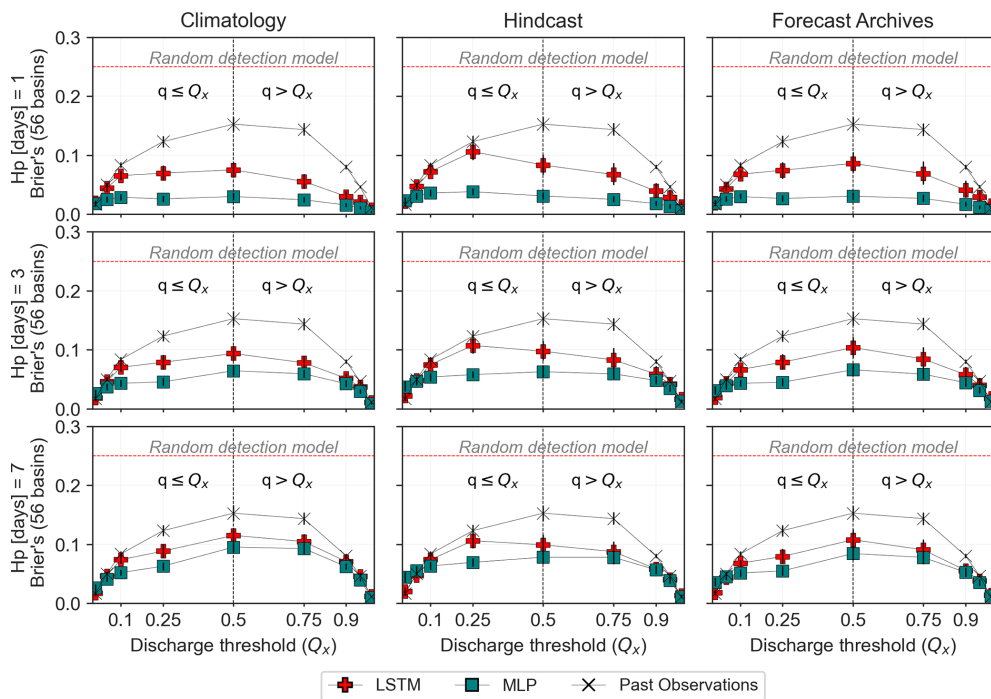


Figure 19. Brier's Scores for the LSTM and the DA1 (MLP) with the CAMELS-FR dataset.

CRPS of the LSTM models appears, on average, lower than that of the DA1, suggesting some consistency across metrics in capturing various properties of the forecasts.

Two specific patterns identified in the AUC analysis in Sect. 3 are also visible in Fig. B9. First, the gap between AUC values based on past-observed (P.O) discharges and those of the forecasting models is particularly pronounced for high-threshold quantiles at a 1 d lead time. Second, for large-discharge thresholds at a 7 d lead time, the AUC scores obtained with hindcast- and archive-based ensemble forecasts clearly exceed those of the climatology-based forecast. This confirms the ability of the weather forecast products to predict significant rainfall events up to one week in advance.

Overall, this extended analysis, which incorporates forecast archives, yields satisfactory results. It confirms the findings of Sect. 3 and reinforces the relevance of the forecasting and discharge assimilation (DA) approaches evaluated in this study. The main findings are as follows: (1) the gain of the DA1 strategy compared to the rainfall–runoff LSTM simulation model is consistently observed, although it is lower for the CAMELS-FR basins, partly due to the initial high performance of the LSTM; (2) the complementarity of the two forecast evaluation frameworks (deterministic vs. ensemble-based) further highlights the importance of ensemble-based evaluation in operational hydrometeorological forecasting. Ensemble-based forecasting also emphasizes the superiority of the DA1 approach over the rainfall–runoff LSTM across the tested lead times and evaluation metrics. Finally, this extended analysis suggests a higher quality of ensemble weather forecast products over the recent period (2018–2021) used to evaluate the DA approaches in the CAMELS-FR basins.

5 Conclusions

This work aimed to evaluate the added value of discharge assimilation (DA) procedures for rainfall–runoff forecasting, particularly in the context of AI-based operational hydrometeorological applications. Three DA strategies are compared against two benchmark models (LSTM and SAC-SMA) that do not incorporate DA. These DA strategies are evaluated under both a traditional perfect weather forecast (deterministic) framework and an ensemble-based forecast framework, using no-skill past observed forcing (climatology), hindcast products, and forecast archives. Additional emphasis is provided through comparisons with both a persistent model and past-observed (P.O) discharge ensembles. The experiments have been conducted on the widely used CAMELS-US dataset and extended to the recently published CAMELS-FR dataset.

While all tested approaches consistently outperform both the persistent model and the P.O baselines, the various DA procedures appear to be globally effective. They generally improve, or at least do not significantly degrade,

the forecasting performance of the benchmark models on which they are based. Within the perfect meteorological forecast evaluation framework, DA approaches consistently improve the SAC-SMA forecasts, while improvements for the LSTM are mainly observed at short lead times and in basins where the benchmark LSTM model initially underperformed. These more limited gains further highlight the strong performance of the LSTM model in rainfall–runoff simulation and forecasting, as already demonstrated in numerous studies (Kratzert et al., 2019; Feng et al., 2020; Hashemi et al., 2022; Nearing et al., 2022; Yang et al., 2025). This behavior is consistently observed across both CAMELS-US and CAMELS-FR basins. Due to the higher hydrological inertia of the CAMELS-FR basins compared to those of the CAMELS-US, the added value of the tested DA strategies remains significant at longer forecasting lead times.

Several interesting insights emerge from the ensemble-based evaluation framework. The DA1 (MLP) approach, which incorporates past observed discharges, appears to outperform the LSTM model across all the tested lead times. This conclusion holds particularly for the assessment criteria characterizing the resolution of the forecasts (Brier's scores and AUC); i.e., the ability to detect in advance the exceedance of a discharge threshold. The LSTM model appears penalized by the limited reliability of its forecast ensembles (biases observed on the rank diagrams). This ensemble evaluation suggests that the performance of the LSTM forecasts could be improved in the future through the implementation of post-processing techniques such as ensemble bias correction.

The tested DA methods are implemented using a relatively simple MLP orchestrator, which already provides satisfactory results. Although this choice aligns with the objective of developing frugal AI solutions, there remains clear potential for improvement by exploring more advanced AI techniques and alternative data assimilation strategies, such as the Ensemble Kalman Filter (Clark et al., 2008) or an autoregressive approach as in Nearing et al. (2022).

It is observed that model performances are globally higher for high observed discharge values than for low flows. This is likely related to the use of the mean squared error loss function during training (Terven et al., 2025). The investigation of alternative loss functions tailored to different flow levels, therefore, represents a promising direction for future research, particularly for the development of AI-based low-flow forecasting models. Moreover, as ensemble discharge forecasts are becoming an operational standard, it may be beneficial to train models directly using ensemble-based metrics, for example, by optimizing the Brier's score for event detection purposes.

Further work could also focus on a more thorough analysis of meteorological and hydrological ensemble spreads, as well as on the application of ensemble bias correction methods to improve the resolution of forecast products.

Appendix A: Data and specificity

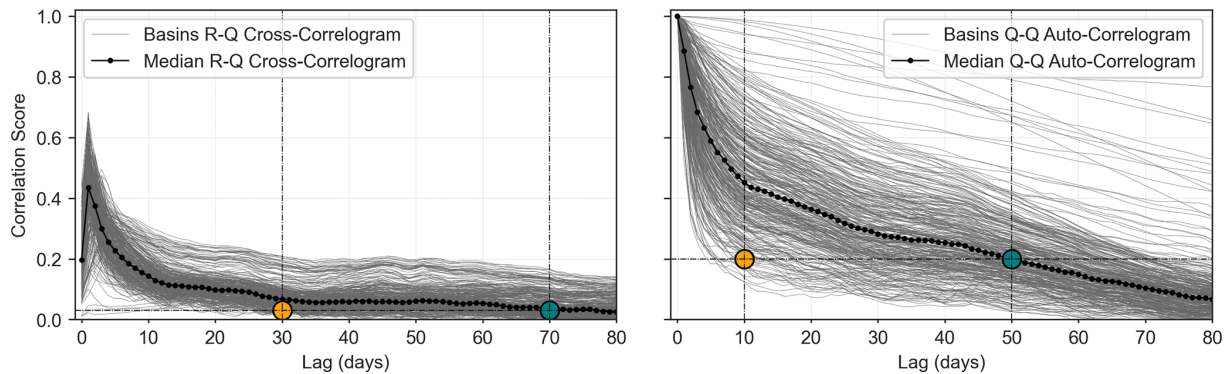


Figure A1. Cross- and auto-correlation analysis between the rainfall and the discharge for the CAMELS-FR dataset. Orange dots denote the position of the n and p used on the CAMELS-US dataset, whereas the teal one indicate the corresponding cross-correlation scores for the CAMELS-FR dataset. This means, following the same approach to setup the size of the input sequences, larger values would have been used on the CAMELS-FR cases.

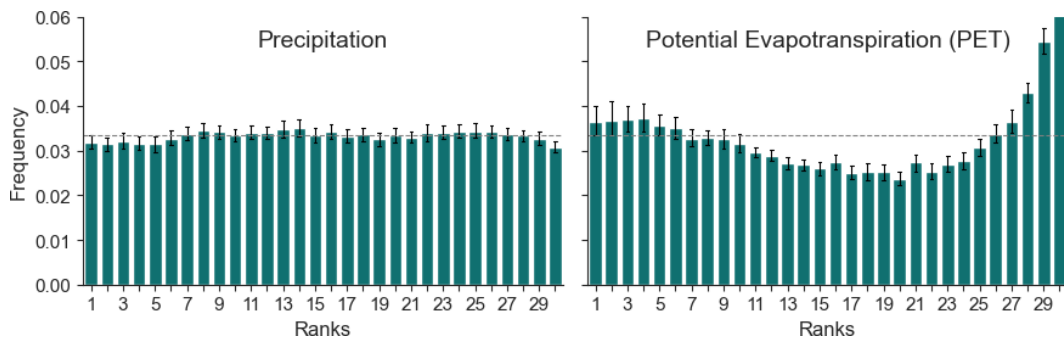


Figure A2. Rank diagrams for Rainfall and PET on CAMELS-FR dataset comparing the test period to the remaining historical observations.

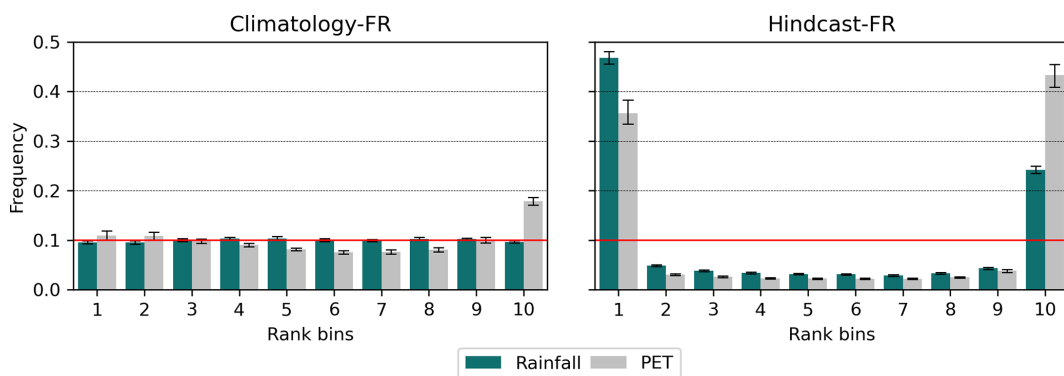


Figure A3. Rank diagrams for the daily precipitation and PET for the climatological ensembles (left panel) and Hindcast (right panel) products for the CAMELS-FR dataset. Plots correspond to 1989–2017 and evaluated for the test period 2017–2021. The error-bars represent variability for the 56 tested basins, the red line denotes the expected uniform distribution

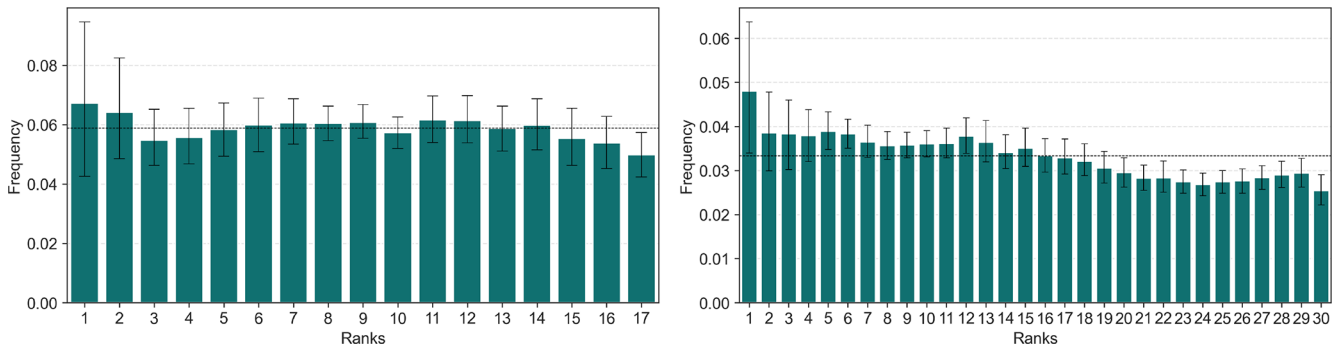


Figure A4. Rank diagrams of the test period against the remaining data for the discharge (discharge climatology) for CAMELS-US (left panel) and CAMELS-FR (right panel) datasets.

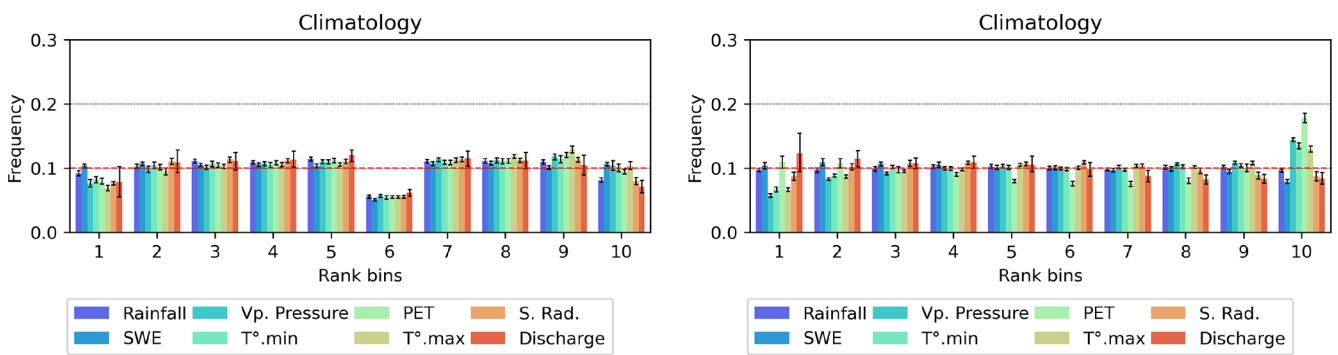


Figure A5. Dispersion analysis of the climatology for all the features in both CAMELS-US (left panel) and CAMELS-FR (right panel) datasets. For easier visualization, the 18 and 29 members the two datasets have been forced to be displayed on 10 classes per graphic.

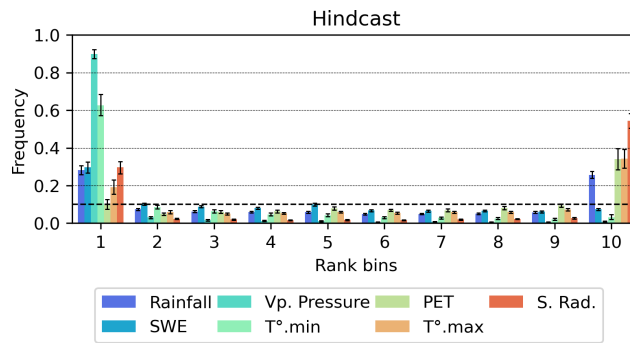


Figure A6. Dispersion analysis of the forecast products on the CAMELS-US case.

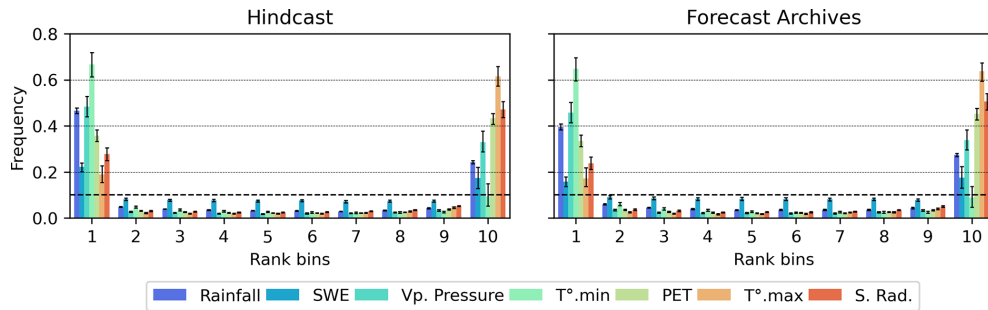


Figure A7. Dispersion analysis of the forecast products on the CAMELS-FR case.

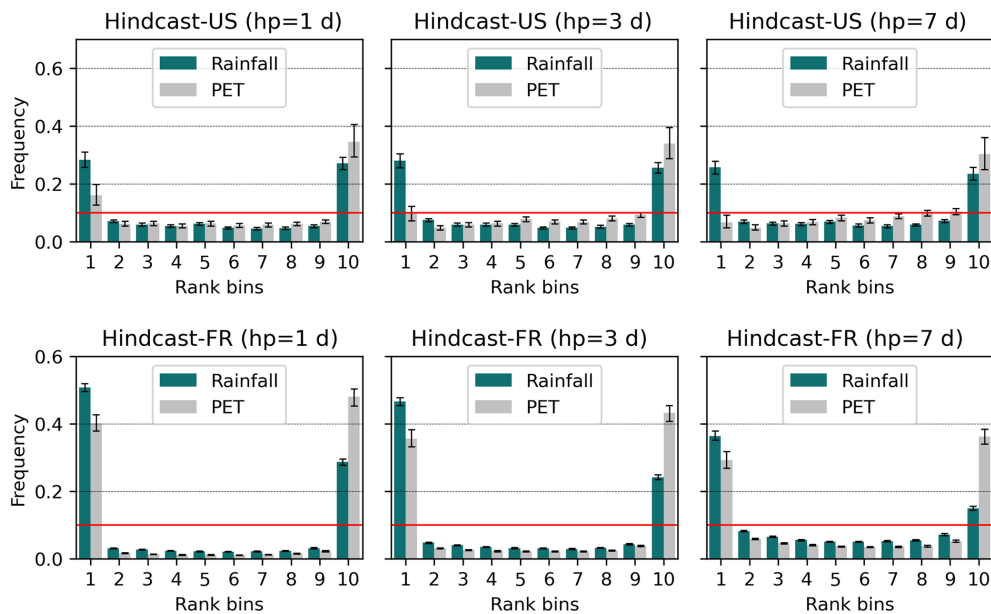


Figure A8. Rank diagrams for daily precipitation and PET for the Hindcast-based ensemble, for lead times 1, 3, and 7 d for both US-basins (top panels) and FR-basins (bottom panels). The plots correspond to the evaluation of the respective test-period within the respective forecast data. The error bars represent variability across the 56 basins considered, and the red line denotes the expected uniform distribution. For ease comparison, the ensembles have been condensed into 10 classes from 32 and 10 members, respectively. Under-dispersion trend of the hindcast products appears diminished within increasing lead times.

Appendix B: Hydrograms

Figure B2 provides an illustration of the ROC curves based on which the AUC values have been calculated, as well as the variability of the ROC curve shapes across the 56 test basins. One ROC curve and one AUC value are computed for each basin and each forecasting method tested.

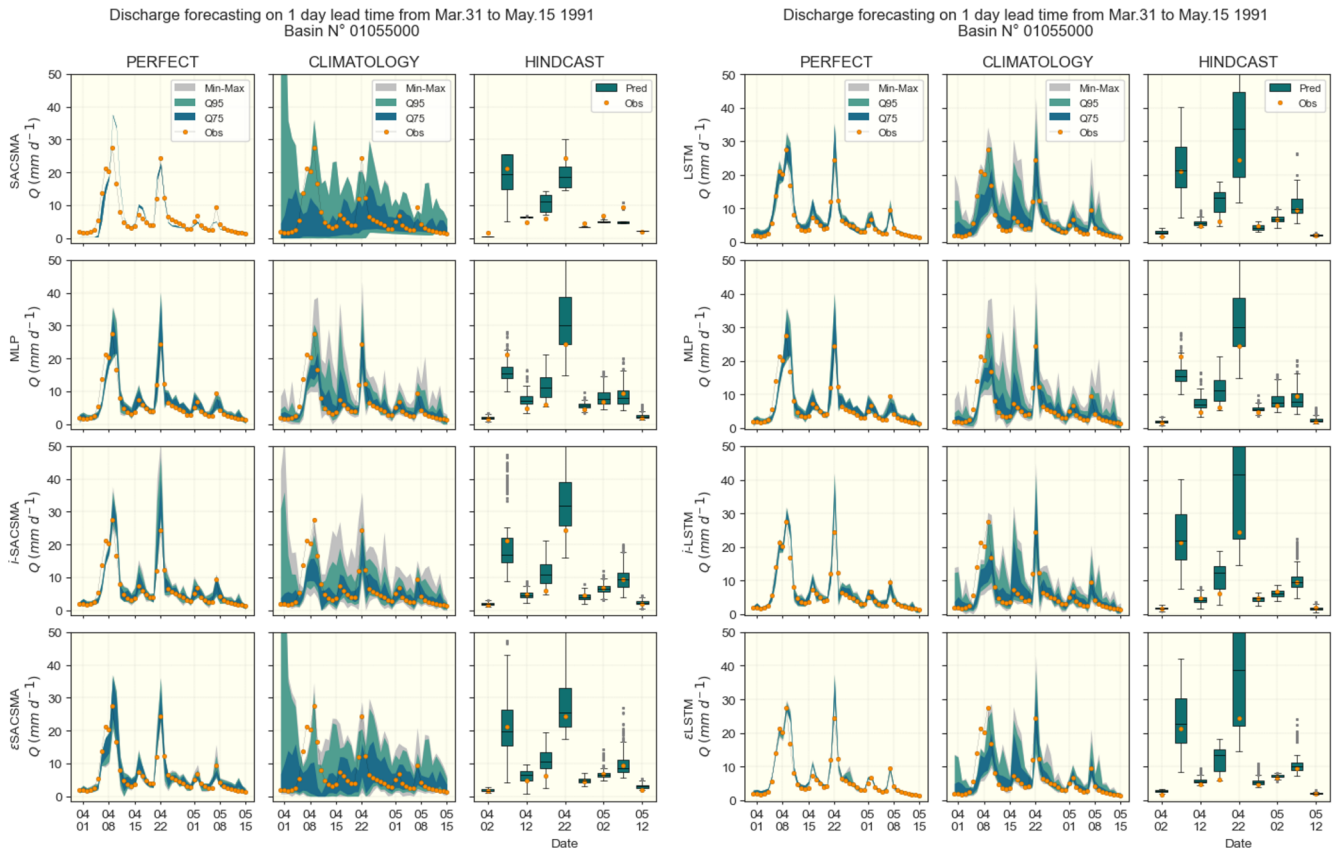


Figure B1. Example of hydrograph for 1 d lead times on the CAMELS-US dataset for both SACS-SMA (left panels) and LSTM (right panels) cases for basin No. 01055000.

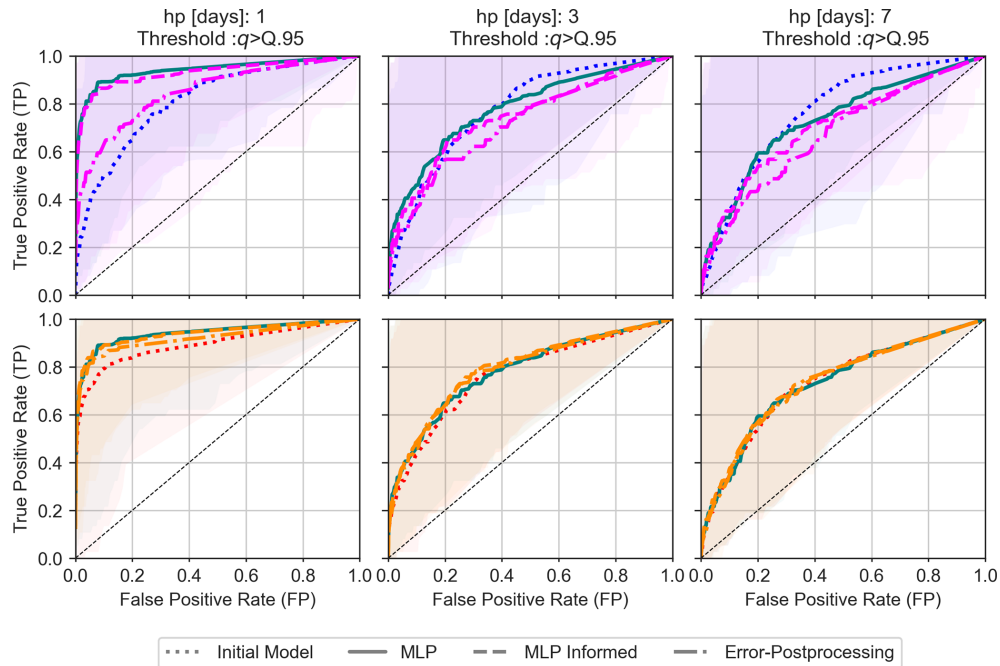


Figure B2. ROC curves for flood detection ($q \geq Q_{95}$) for 1, 3 and 7 d lead times. Results are style-coded: *MLP Simple* (green solid, DA-1), *MLP informed by benchmark* (dashed, DA-2), *Benchmark ePP* (dot-dashed, DA-3), *initial Benchmark* (dotted). Benchmark cases are color-coded: *SACSMA* (blue to pink, first row), *LSTM* (red to orange, second row). Halos show the variability across the 56 basins around the median values.

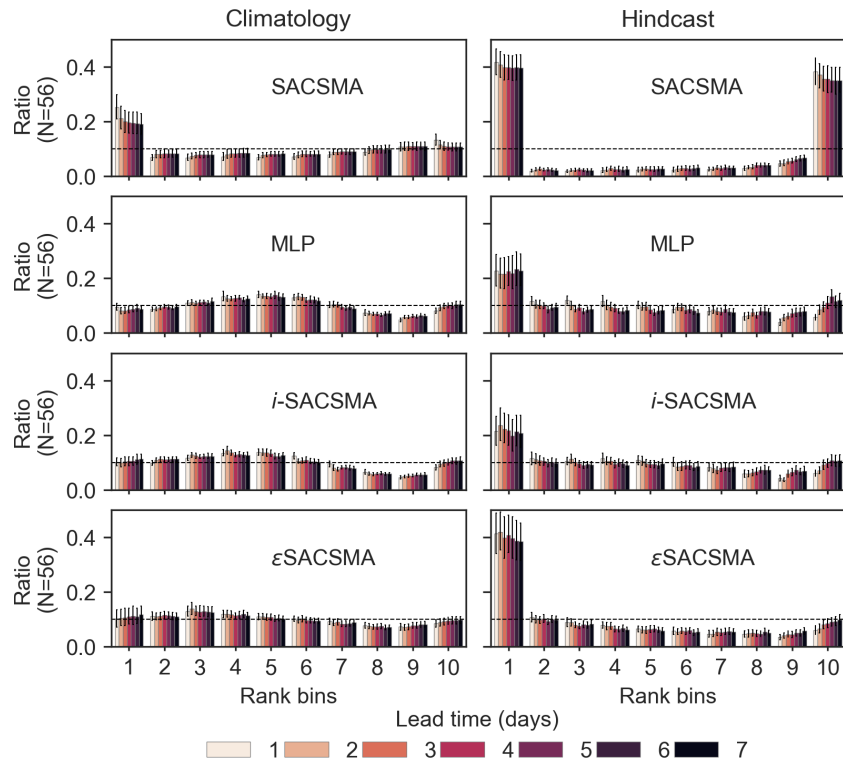


Figure B3. Rank diagrams for the benchmark SACCSMA-cases and the DA strategies. x axis (10 rank classes), y axis (proportion of observed values in each class), median ratio and error-bars indicating the distributions of the 56 basins. Colors indicate the lead times.

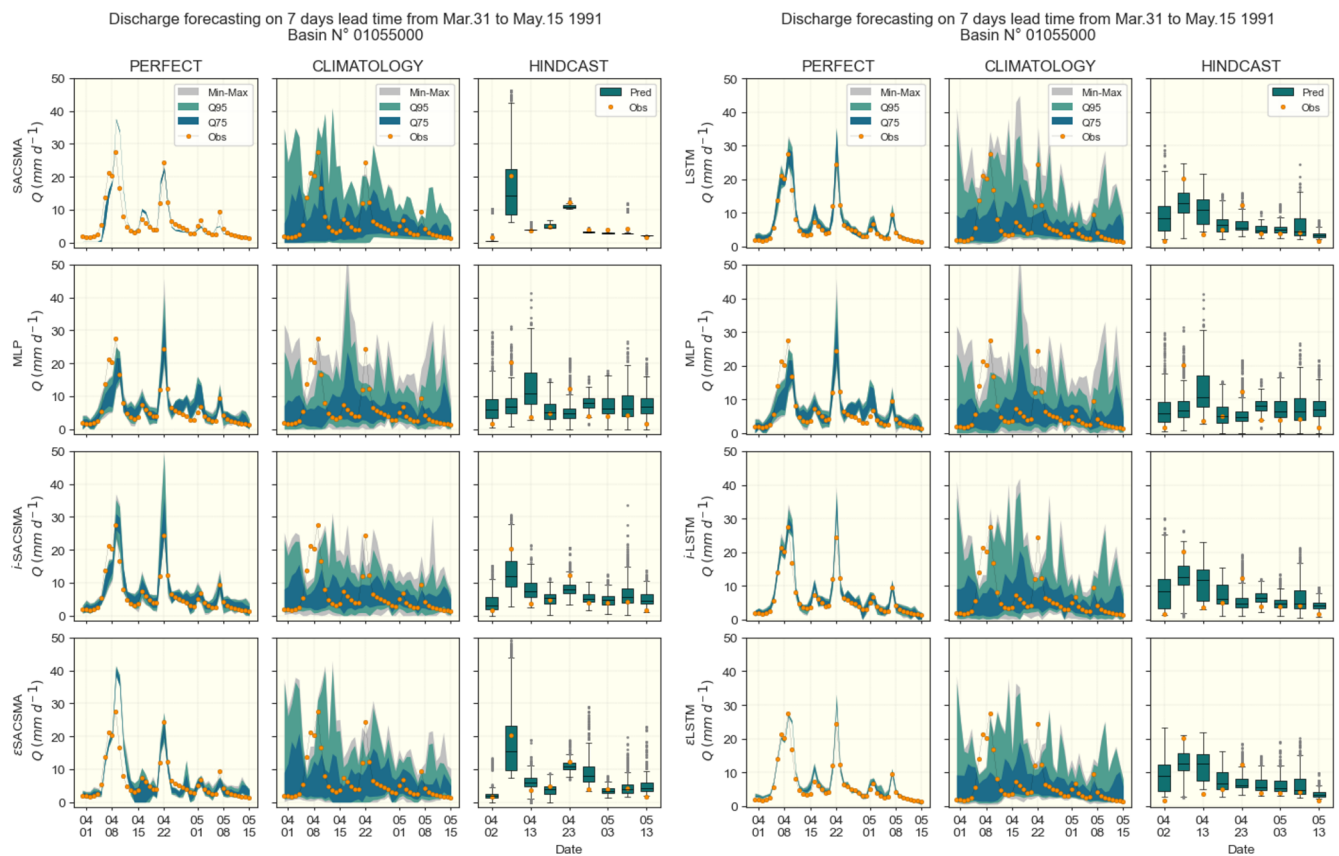


Figure B4. Example of hydrograph for 7 d lead times on the CAMELS-US dataset for both SACS-SMA (left panels) and LSTM (right panels) cases for basin No. 01055000.

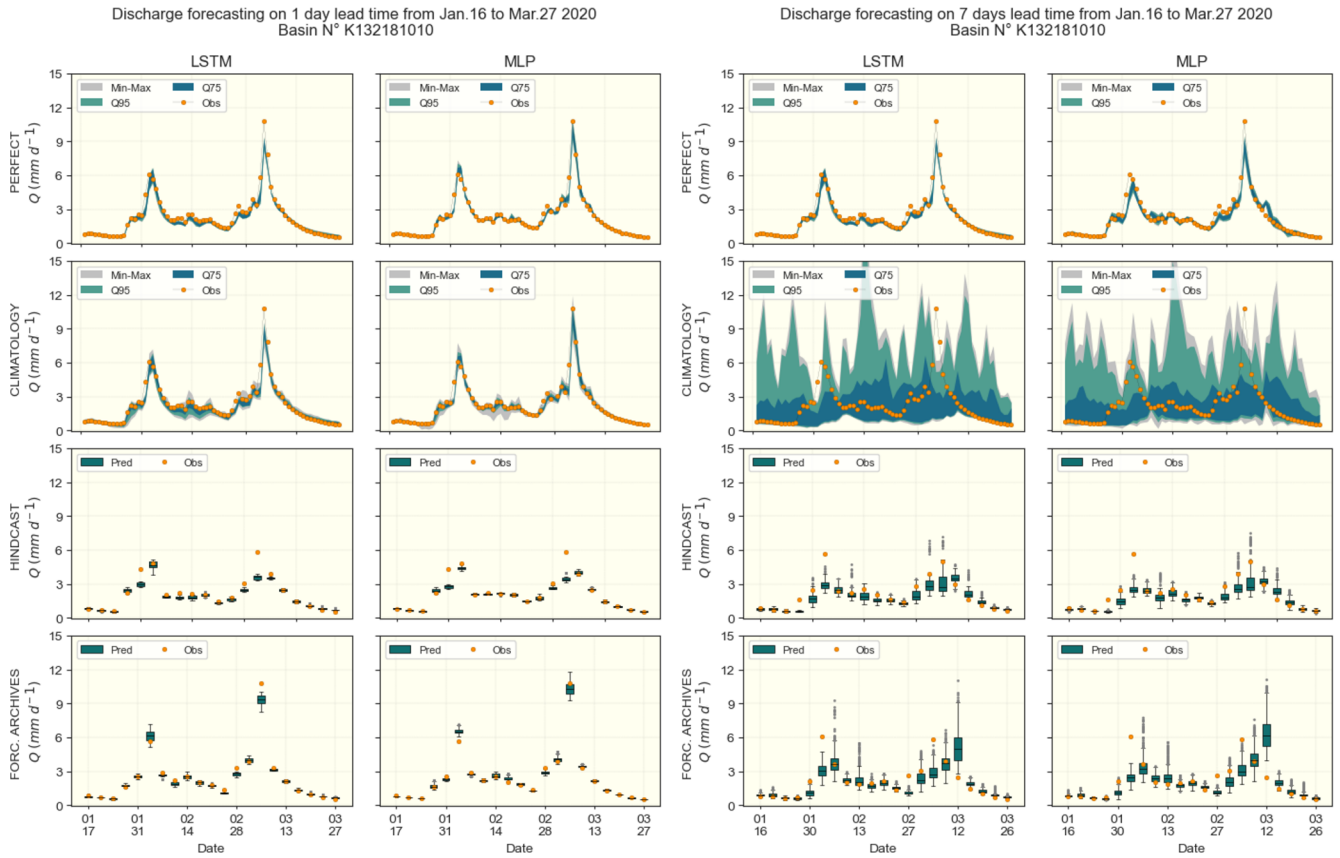


Figure B5. Example of hydrograph for 1 and 7 d lead times on the CAMELS-FR dataset for the basin K132181010.

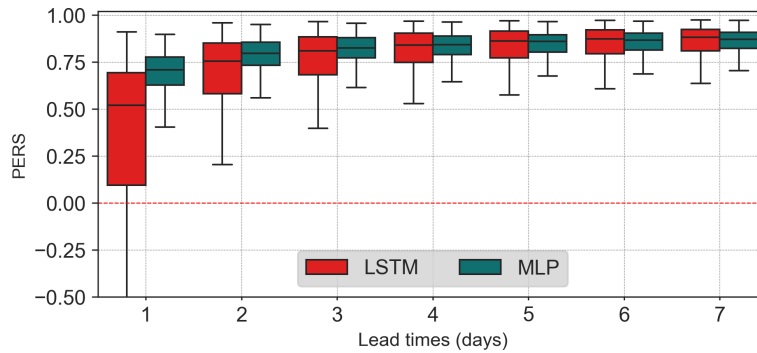


Figure B6. Persistence scores for LSTM and the MLP (DA1) on the CAMELS-FR dataset.

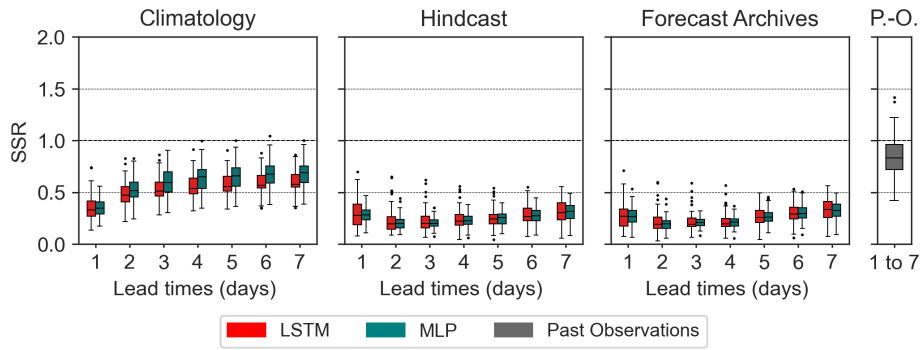


Figure B7. SSR for the LSTM and the DA1 (MLP) with the CAMELS-FR dataset.

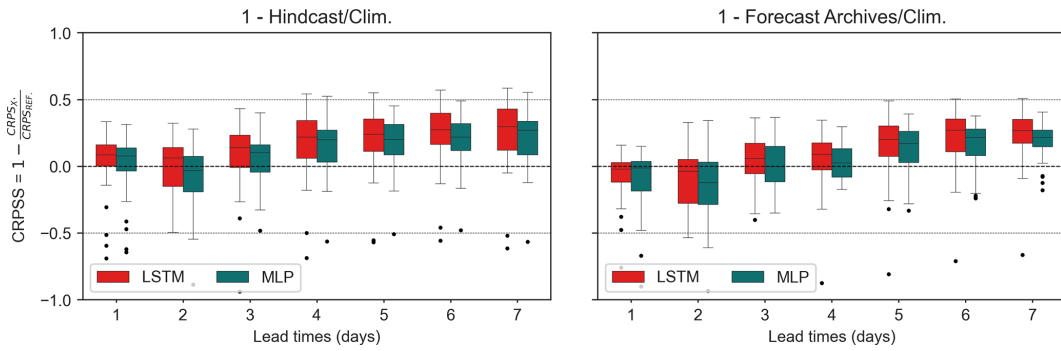


Figure B8. CRPSS of forecast products against the Climatology-based scenario for the CAMELS-FR dataset.

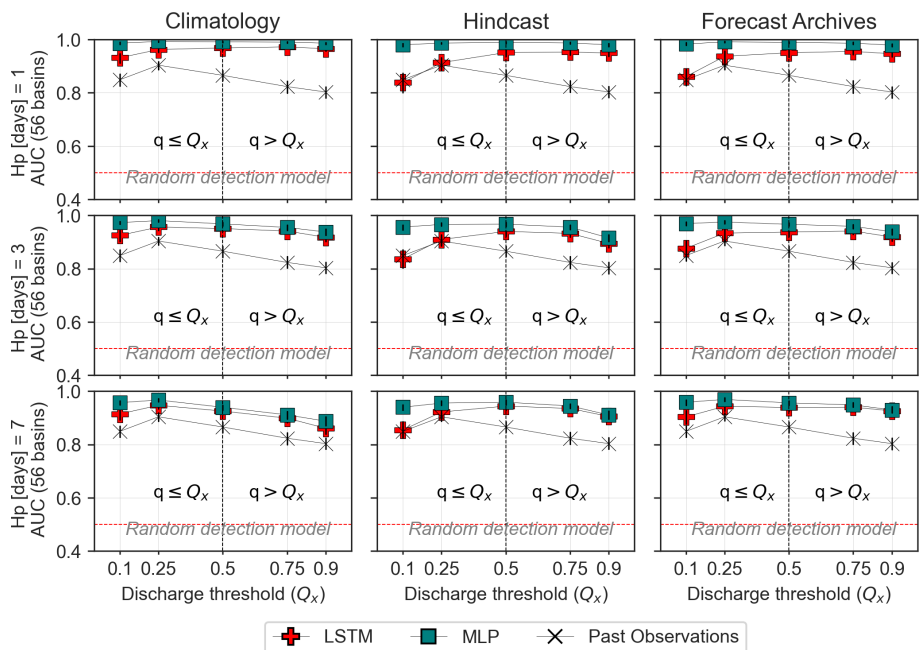


Figure B9. AUC scores for the LSTM and the DA1 (MLP) with the CAMELS-FR dataset.

Code and data availability. All data used in this study are drawn from the CAMELS-US (<https://gdex.ucar.edu/dataset/camels.html>, last access: 17 April 2025) and CAMELS-FR (<https://entrepot.recherche.data.gouv.fr/dataverse/CAMELS-FR>, last access: 17 April 2025) datasets. The processed version of these datasets supporting this study is made available on Zenodo (<https://doi.org/10.5281/zenodo.19825677>), including the necessary instructions to ensure reproducibility. The benchmark models used in this study (LSTM and SAC-SMA) are described in their original publications, which should be consulted for methodological details. Their adapted code versions used in this work are publicly released on Zenodo (SACSMA: <https://doi.org/10.5281/zenodo.20379006>; LSTM: <https://doi.org/10.5281/zenodo.20379019>). The code for the MLP-based data assimilation (DA) implementation is available at <https://doi.org/10.5281/zenodo.20415493>.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Author contributions. All the indicated authors contributed to the realization and the discussions of this study. BSF and EG carried out the experiments and the analysis of the scientific relevance of the results. BSF developed the model code, performed the simulations and post-processed the results. FS participated in the deployment of the SAC-SMA model, including the post-processing of the results. NA and DT contributed in the discussion for the operationalization of the models as the aQuasys partners.

Disclaimer. The paper is written in LaTeX using **Overleaf**. **Writefull** and **ChatGPT** have been used for rephrasing and minor corrections. The experiments are primarily based on the CAMELS-US and CAMELS-FR datasets and implemented using open-source software and programming languages, including Python 3.9, scikit-learn, PyTorch, NumPy, and pandas.

Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. The authors bear the ultimate responsibility for providing appropriate place names. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements. The authors would like to thank *Gustave Eiffel University* and *aQuasys Company* for initiating the *Anticipation, Planification et Pilotage des Prélèvements Agricoles (A3P)* project and the AiQua LabCom. We are grateful to the Neural-Hydrology team for making their regional LSTM code publicly available, as well as to the authors of the SAC-SMA model. We also thank the contributors of the CAMELS-US and CAMELS-FR datasets for their significant contributions to the community. We acknowledge the *Groupement Ligérien pour le Calcul Intensif Distribu   (GLiCID)* for providing the computing resources. Finally, we thank Micha  l Savary, Pierre Nicolle, Reyhaneh Hashemi, Zo   Jack

and Otis Cooper for their support, including preliminary proofreading and grammar checking.

Financial support. This research has been supported by the Agence Nationale de la Recherche (ANR) under the AiQua LabCom (grant no. ANR-24-LCV2-0015-01) and Bpifrance under the A3P project (grant no. DOS0231020/00).

Review statement. This paper was edited by Ralf Loritz and reviewed by two anonymous referees.

References

- Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, *Hydrol. Earth Syst. Sci.*, 21, 5293–5313, <https://doi.org/10.5194/hess-21-5293-2017>, 2017.
- Anctil, F., Michel, C., Perrin, C., and Andr  assian, V.: A soil moisture index as an auxiliary ANN input for stream flow forecasting, *J. Hydrol.*, 286, 155–167, <https://doi.org/10.1016/j.jhydrol.2003.09.006>, 2004.
- Anon: Anaconda Software Distribution, <https://www.anaconda.com> (last access: 31 July 2024), 2020.
- Atmaja, B. T. and Akagi, M.: Deep Multilayer Perceptrons for Dimensional Speech Emotion Recognition, arXiv [preprint], <https://doi.org/10.48550/arXiv.2004.02355>, 2020.
- Bell, R., Spring, A., Brady, R., Andrew, Squire, D., Blackwood, Z., Sitter, M. C., and Chegini, T.: xarray-contrib/xskillscore: Release v0.0.23, Zenodo [data set], <https://doi.org/10.5281/zenodo.5173153>, 2021.
- Boucher, M.-A., Quilty, J., and Adamowski, J.: Data Assimilation for Streamflow Forecasting Using Extreme Learning Machines and Multilayer Perceptrons, *Water Resour. Res.*, 56, e2019WR026226, <https://doi.org/10.1029/2019WR026226>, 2020.
- Bourgin, F., Ramos, M. H., Thirel, G., and Andr  assian, V.: Investigating the interactions between data assimilation and post-processing in hydrological ensemble forecasting, *J. Hydrol.*, 519, 2775–2784, <https://doi.org/10.1016/j.jhydrol.2014.07.054>, 2014.
- Bradley, A. A. and Schwartz, S. S.: Summary Verification Measures and Their Interpretation for Ensemble Forecasts, *Mon. Weather Rev.*, 139, 3075–3089, <https://doi.org/10.1175/2010MWR3305.1>, 2011.
- Brier, G. W.: Verification of forecasts expressed in terms of probability, *Mon. Weather Rev.*, 78, 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2), 1950.
- Buizza, R., Houtekamer, P. L., Pellerin, G., Toth, Z., Zhu, Y., and Wei, M.: A Comparison of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems, *Mon. Weather Rev.*, 133, 1076–1097, <https://doi.org/10.1175/MWR2905.1>, 2005.
- Chevillon, G.: Direct multi-step estimation and forecasting, *J. Econ. Surv.*, 21, 746–785, <https://doi.org/10.1111/j.1467-6419.2007.00518.x>, 2007.
- Clark, M. P., Rupp, D. E., Woods, R. A., Zheng, X., Ibbitt, R. P., Slater, A. G., Schmidt, J., and Uddstrom, M. J.: Hydrological data assimilation with the ensemble Kalman fil-

- ter: Use of streamflow observations to update states in a distributed hydrological model, *Adv. Water Resour.*, 31, 1309–1324, <https://doi.org/10.1016/j.advwatres.2008.06.005>, 2008.
- Corradini, C., Melone, F., and Ubertini, L.: A semi-distributed adaptive model for real-time flood forecasting, *J. Am. Water Resour. Assoc.*, 22, 1031–1038, 1986.
- Crochemore, L., Ramos, M.-H., Pappenberger, F., and Perrin, C.: Seasonal streamflow forecasting by conditioning climatology with precipitation indices, *Hydrol. Earth Syst. Sci.*, 21, 1573–1591, <https://doi.org/10.5194/hess-21-1573-2017>, 2017.
- Day, G. N.: Extended Streamflow Forecasting Using NWS-RFS, *J. Water Resour. Plan. Manage.*, 111, 157–170, [https://doi.org/10.1061/\(ASCE\)0733-9496\(1985\)111:2\(157\)](https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157)), 1985.
- Delaigne, O., Guimarães, G. M., Brigode, P., Génot, B., Perrin, C., Soubeyroux, J.-M., Janet, B., Addor, N., and Andréassian, V.: CAMELS-FR dataset: a large-sample hydroclimatic dataset for France to explore hydrological diversity and support model benchmarking, *Earth Syst. Sci. Data*, 17, 1461–1479, <https://doi.org/10.5194/essd-17-1461-2025>, 2025.
- Fang, Z., Wang, Y., Peng, L., and Hong, H.: Predicting flood susceptibility using LSTM neural networks, *J. Hydrol.*, 594, 125734, <https://doi.org/10.1016/j.jhydrol.2020.125734>, 2021.
- 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 Resour. Res.*, 56, e2019WR026793, <https://doi.org/10.1029/2019WR026793>, 2020.
- 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, *J. Hydrol.*, 377, 80–91, <https://doi.org/10.1016/j.jhydrol.2009.08.003>, 2009.
- Hamill, T. M.: Interpretation of Rank Histograms for Verifying Ensemble Forecasts, *Mon. Weather Rev.*, 129, 550–560, [https://doi.org/10.1175/1520-0493\(2001\)129<0550:IORHFV>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2), 2001.
- Harold, B., Barb, B., Beth, E., Chris, F., Johannes, J., Ian, J., Tieh-Yong, K., Paul, R., and David, S.: WWRP/WGNE Joint Working Group on Forecast Verification Research, <https://www.cawcr.gov.au/projects/verification/> (last access: 13 December 2024), 2015.
- Hashemi, R., Brigode, P., Garambois, P.-A., and Javelle, P.: How can we benefit from regime information to make more effective use of long short-term memory (LSTM) runoff models?, *Hydrol. Earth Syst. Sci.*, 26, 5793–5816, <https://doi.org/10.5194/hess-26-5793-2022>, 2022.
- Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, *Weather Forecast.*, 15, 559–570, [https://doi.org/10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2), 2000.
- Hidalgo, J. and Jouglu, R.: On the use of local weather types classification to improve climate understanding: An application on the urban climate of Toulouse, *PLOS ONE*, 13, e0208138, <https://doi.org/10.1371/journal.pone.0208138>, 2018.
- Hudson, D., Alves, O., Hendon, H. H., Lim, E.-P., Liu, G., Luo, J.-J., MacLachlan, C., Marshall, A. G., Shi, L., Wang, G., Wedd, R., Young, G., Zhao, M., and Zhou, X.: Corrigendum to: ACCESS-S1: The new Bureau of Meteorology multi-week to seasonal prediction system, *J. South. Hemis. Earth Syst. Sci.*, 70, 393, https://doi.org/10.1071/ES17009_CO, 2020.
- Hunter, J. D.: Matplotlib: A 2D Graphics Environment, *Comput. Sci. Eng.*, 9, 90–95, <https://doi.org/10.1109/MCSE.2007.55>, 2007.
- Husic, A., Al-Aamery, N., and Fox, J. F.: Simulating hydrologic pathway contributions in fluvial and karst settings: An evaluation of conceptual, physically-based, and deep learning modeling approaches, *J. Hydrol. X*, 17, 100134, <https://doi.org/10.1016/j.hydroa.2022.100134>, 2022.
- Jeannin, P.-Y., Artigue, G., Butscher, C., Chang, Y., Charlier, J.-B., Duran, L., Gill, L., Hartmann, A., Johannet, A., Jourde, H., Kavousi, A., Liesch, T., Liu, Y., Lüthi, M., Malard, A., Mazzilli, N., Pardo-Igúzquiza, E., Thiéry, D., Reimann, T., Schuler, P., Wöhling, T., and Wunsch, A.: Karst modelling challenge 1: Results of hydrological modelling, *J. Hydrol.*, 600, 126508, <https://doi.org/10.1016/j.jhydrol.2021.126508>, 2021.
- JetBrains: PyCharm, <https://www.jetbrains.com/pycharm/> (last access: 20 March 2026), 2024.
- Kitanidis, P. K. and Bras, R. L.: Real-time forecasting with a conceptual hydrologic model: 2. Applications and results, *Water Resour. Res.*, 16, 1034–1044, <https://doi.org/10.1029/WR016i006p01034>, 1980.
- Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., and Willing, C.: Jupyter Notebooks – a publishing format for reproducible computational workflows, in: Positioning and Power in Academic Publishing: Players, Agents and Agendas, IOS Press, 87–90, <https://doi.org/10.3233/978-1-61499-649-1-87>, 2016.
- Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks, *Hydrol. Earth Syst. Sci.*, 22, 6005–6022, <https://doi.org/10.5194/hess-22-6005-2018>, 2018.
- Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, *Hydrol. Earth Syst. Sci.*, 23, 5089–5110, <https://doi.org/10.5194/hess-23-5089-2019>, 2019.
- Lai, T. L., Gross, S. T., and Shen, D. B.: Evaluating probability forecasts, *Ann. Stat.*, 39, 2356–2382, 2011.
- Li, H., Zhang, C., Chu, W., Shen, D., and Li, R.: A process-driven deep learning hydrological model for daily rainfall-runoff simulation, *J. Hydrol.*, 637, 131434, <https://doi.org/10.1016/j.jhydrol.2024.131434>, 2024.
- Liu, X. and Wang, W.: Deep Time Series Forecasting Models: A Comprehensive Survey, *Mathematics*, 12, <https://doi.org/10.3390/math12101504>, 2024.
- Mangin, A.: Pour une meilleure connaissance des systèmes hydrologiques à partir des analyses corrélatoire et spectrale, *J. Hydrol.*, 67, 25–43, [https://doi.org/10.1016/0022-1694\(84\)90230-0](https://doi.org/10.1016/0022-1694(84)90230-0), 1984.
- Matheson, J. E. and Winkler, R. L.: Scoring Rules for Continuous Probability Distributions, *Manage. Sci.*, 22, 1087–1096, <https://doi.org/10.1287/mnsc.22.10.1087>, 1976.
- McKinney, W.: Data Structures for Statistical Computing in Python, in: Proceedings of the 9th Python in Science Conference, edited by: v. D. W. Stefan and Jarrod, M., Austin, Texas, USA, 56–61, <https://doi.org/10.25080/Majora-92bf1922-00a>, 2010.

- Murphy, A. H.: What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting, *Weather Forecast.*, 8, 281–293, [https://doi.org/10.1175/1520-0434\(1993\)008<0281:WIAGFA>2.0.CO;2](https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2), 1993.
- Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – A discussion of principles, *J. Hydrol.*, 10, 282–290, [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6), 1970.
- Nearing, G. S., Klotz, D., Frame, J. M., Gauch, M., Gilon, O., Kratzert, F., Sampson, A. K., Shalev, G., and Nevo, S.: Technical note: Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks, *Hydrol. Earth Syst. Sci.*, 26, 5493–5513, <https://doi.org/10.5194/hess-26-5493-2022>, 2022.
- Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., and Nearing, G.: Benchmarking of a Physically Based Hydrologic Model, *J. Hydrometeorol.*, 18, 2215–2225, <https://doi.org/10.1175/JHM-D-16-0284.1>, 2017.
- Oliveira, D. D., Rampinelli, M., Tozatto, G. Z., Andreão, R. V., and Müller, S. M. T.: Forecasting vehicular traffic flow using MLP and LSTM, *Neural Comput. Appl.*, 33, 17245–17256, <https://doi.org/10.1007/s00521-021-06315-w>, 2021.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, *CoRR*, abs/1201.0490, arXiv [preprint], <https://doi.org/10.48550/arXiv.1201.0490>, 2012.
- Pelletier, A. and Andréassian, V.: An underground view of surface hydrology: what can piezometers tell us about river floods and droughts?, *Comptes Rendus. Géoscience*, 355, 271–280, <https://doi.org/10.5802/crgeos.195>, 2024.
- Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, *J. Hydrol.*, 279, 275–289, [https://doi.org/10.1016/S0022-1694\(03\)00225-7](https://doi.org/10.1016/S0022-1694(03)00225-7), 2003.
- Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A., Ellison, J., Fiszeder, P., Frances, P. H., Frazier, D. T., Gilliland, M., Gönül, M. S., Goodwin, P., Grossi, L., Grushka-Cockayne, Y., Guidolin, M., Guidolin, M., Gunter, U., Guo, X., Guseo, R., Harvey, N., Hendry, D. F., Hollyman, R., Januschowski, T., Jeon, J., Jose, V. R. R., Kang, Y., Koehler, A. B., Kolassa, S., Kourentzes, N., Leva, S., Li, F., Litsiou, K., Makridakis, S., Martin, G. M., Martinez, A. B., Meeran, S., Modis, T., Nikolopoulos, K., Önköl, D., Paccagnini, A., Panagiotelis, A., Panapakidis, I., Pavía, J. M., Pedio, M., Pedregal, D. J., Pinson, P., Ramos, P., Rapach, D. E., Reade, J. J., Rostami-Tabar, B., Rubaszek, M., Sermpinis, G., Shang, H. L., Spiliotis, E., Syntetos, A. A., Talagala, P. D., Talagala, T. S., Tashman, L., Thomakos, D., Thorarindottir, T., Todini, E., Trapero Arenas, J. R., Wang, X., Winkler, R. L., Yusupova, A., and Ziel, F.: Forecasting: theory and practice, *Int. J. Forecast.*, 38, 705–871, <https://doi.org/10.1016/j.ijforecast.2021.11.001>, 2022.
- Philip, S., Kew, S., van Oldenborgh, G. J., Otto, F., Vautard, R., van der Wiel, K., King, A., Lott, F., Arrighi, J., Singh, R., and van Aalst, M.: A protocol for probabilistic extreme event attribution analyses, *Adv. Stat. Climatol. Meteorol. Oceanogr.*, 6, 177–203, <https://doi.org/10.5194/ascmo-6-177-2020>, 2020.
- Piazzì, G., Thirel, G., Perrin, C., and Delaigue, O.: Sequential Data Assimilation for Streamflow Forecasting: Assessing the Sensitivity to Uncertainties and Updated Variables of a Conceptual Hydrological Model at Basin Scale, *Water Resour. Res.*, 57, <https://doi.org/10.1029/2020WR028390>, 2021.
- Pözl, A., Blaschke, A. P., Komma, J., Farnleitner, A. H., and Derx, J.: Transformer Versus LSTM: A Comparison of Deep Learning Models for Karst Spring Discharge Forecasting, *Water Resour. Res.*, 60, e2022WR032602, <https://doi.org/10.1029/2022WR032602>, 2024.
- Rahbar, A., Mirarabi, A., Nakhaei, M., Talkhabi, M., and Jamali, M.: A Comparative Analysis of Data-Driven Models (SVR, ANFIS, and ANNs) for Daily Karst Spring Discharge Prediction, *Water Resour. Res.*, 36, 589–609, <https://doi.org/10.1007/s11269-021-03041-9>, 2022.
- Rentschler, J., Avner, P., Marconcini, M., Su, R., Strano, E., Voudoukas, M., and Hallegatte, S.: Global evidence of rapid urban growth in flood zones since 1985, *Nature*, 622, 87–92, <https://doi.org/10.1038/s41586-023-06468-9>, 2023.
- Rosenblatt, F.: The perceptron: A probabilistic model for information storage and organization in the brain., *Psycholog. Rev.*, 65, 386–408, <https://doi.org/10.1037/h0042519>, 1958.
- Saint Fleur, B. E., Artigue, G., Johannet, A., and Pistre, S.: Deep Multilayer Perceptron for Knowledge Extraction: Understanding the Gardon de Mialet Flash Floods Modeling, in: *Theory and Applications of Time Series Analysis*, edited by: Valenzuela, O., Rojas, F., Herrera, L. J., Pomares, H., and Rojas, I., Springer International Publishing, Cham, 333–348, ISBN 978-3-030-56219-9, 2020.
- Saint-Fleur, B. E., Allier, S., Lassara, E., Rivet, A., Artigue, G., Pistre, S., and Johannet, A.: Towards a better consideration of rainfall and hydrological spatial features by a deep neural network model to improve flash floods forecasting: case study on the Gardon basin, France, *Model. Earth Syst. Environ.*, 9, 3693–3708, <https://doi.org/10.1007/s40808-022-01650-w>, 2023.
- Schiermeier, Q.: Droughts, heatwaves and floods: How to tell when climate change is to blame, *Nature*, 560, 20–22, <https://doi.org/10.1038/d41586-018-05849-9>, 2018.
- Seillier-Moisewitsch, F. and Dawid, A. P.: On Testing the Validity of Sequential Probability Forecasts, *J. Am. Stat. Assoc.*, 88, 355–359, <https://doi.org/10.2307/2290731>, 1993.
- Slater, L. J., Villarini, G., and Bradley, A. A.: Evaluation of the skill of North-American Multi-Model Ensemble (NMME) Global Climate Models in predicting average and extreme precipitation and temperature over the continental USA, *Clim. Dynam.*, 53, 7381–7396, <https://doi.org/10.1007/s00382-016-3286-1>, 2019.
- Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of probabilistic prediction systems, PhD thesis, Shinfield Park, Reading, 1997.
- Teräsvirta, T., Tjøstheim, D., and Granger, C. W. J.: *Modelling Nonlinear Economic Time Series*, Oxford University Press, ISBN 9780199587148, <https://doi.org/10.1093/acprof:oso/9780199587148.001.0001>, 2010.
- Terven, J., Cordova-Esparza, D.-M., Romero-González, J.-A., Ramírez-Pedraza, A., and Chávez-Urbiola, E. A.: A comprehensive survey of loss functions and metrics in deep learning, *Artif. Intel. Rev.*, 58, 195, <https://doi.org/10.1007/s10462-025-11198-7>, 2025.

- van der Walt, S., Colbert, S. C., and Varoquaux, G.: The NumPy Array: A Structure for Efficient Numerical Computation, *Comput. Sci. Eng.*, 13, 22–30, <https://doi.org/10.1109/MCSE.2011.37>, 2011.
- van Rossum, G.: Python tutorial, CWI – Centrum voor Wiskunde en Informatica, Amsterdam, the Netherlands, 1995.
- Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H.-S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mustrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A. W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D.-J., Xiao, H., Zaripov, R., and Zhang, L.: The Subseasonal to Seasonal (S2S) Prediction Project Database, *B. Am. Meteorol. Soc.*, 98, 163–173, <https://doi.org/10.1175/BAMS-D-16-0017.1>, 2017.
- Waskom, M. L.: seaborn: statistical data visualization, *J. Open Sour. Softw.*, 6, <https://doi.org/10.21105/joss.03021>, 2021.
- Werbos, P.: Beyond regression: New tools for prediction and analysis in the behavioral sciences, PhD thesis, Committee on Applied Mathematics, Harvard University, Cambridge, MA, <https://gwrn.net/doc/ai/nn/1974-werbos.pdf> (last access: 5 June 2026), 1974.
- Werbos, P.: Backpropagation: Past and future, in: IEEE 1988 International Conference on Neural Networks, 343–353, <https://doi.org/10.1109/ICNN.1988.23866>, 1988.
- Whitaker, J. S. and Loughe, A. F.: The Relationship between Ensemble Spread and Ensemble Mean Skill, *Mon. Weather Rev.*, 126, 3292–3302, [https://doi.org/10.1175/1520-0493\(1998\)126<3292:TRBESA>2.0.CO;2](https://doi.org/10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2), 1998.
- Wunsch, A., Liesch, T., and Broda, S.: Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX), *Hydrol. Earth Syst. Sci.*, 25, 1671–1687, <https://doi.org/10.5194/hess-25-1671-2021>, 2021.
- Yang, C., Yuan, H., and Su, X.: Bias correction of ensemble precipitation forecasts in the improvement of summer streamflow prediction skill, *J. Hydrol.*, 588, 124955, <https://doi.org/10.1016/j.jhydrol.2020.124955>, 2020.
- Yang, Y., Pan, M., Feng, D., Xiao, M., Dixon, T., Hartman, R., Shen, C., Song, Y., Sengupta, A., Delle Monache, L., and Ralph, F. M.: Improving streamflow simulation through machine learning-powered data integration and its potential for forecasting in the Western U.S., *Hydrol. Earth Syst. Sci.*, 29, 5453–5476, <https://doi.org/10.5194/hess-29-5453-2025>, 2025.
- Zalachori, I., Ramos, M.-H., Garçon, R., Mathevet, T., and Gailhard, J.: Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies, *Adv. Sci. Res.*, 8, 135–141, <https://doi.org/10.5194/asr-8-135-2012>, 2012.