https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

Can we rely on deep-learning hydrological models to assess the quality of precipitation products?

Mohamed Saadi[1], Pierre Brigode[2], Marc Dumont[3], Ludovic Oudin[4], Hélène Roux[1]

[1]Université de Toulouse, Toulouse INP, CNRS, Institut de Mécanique des Fluides de Toulouse, 31400 Toulouse, France
[2]Univ Rennes, CNRS, Géosciences Rennes, 35000 Rennes, France
[3]Institute of Earth Sciences, University of Lausanne, 1015 Lausanne, Switzerland
[4]Sorbonne Université, Université PSL, EPHE, CNRS, METIS, 75005 Paris, France

Correspondence to: Mohamed Saadi (mohamed.saadi@toulouse-inp.fr)

Abstract. Process-based models are commonly used to evaluate the accuracy of precipitation products because they are highly sensitive to errors in precipitation due to mass-balance constraints. In this respect, it is unclear whether to rely on deep-learning models to discriminate between precipitation products, since their larger flexibility may allow them to compensate for the errors in input precipitation. Here, we address this question by comparing a 1-km, radar-based reanalysis (COMEPHORE) against (1) a gauge-based topography-aware product (SAFRAN, 8 km) and (2) a global, model-based reanalysis (ERA5-Land, ~9 km). Over a set of 445 French catchments, we assessed the quality of each precipitation product using three hydrological models: a process-based conceptual model (GR4J), a locally trained, and a regionally trained long short-term memory (LSTM) models. To assess the reliability of the evaluation by LSTM models, we retrained and tested the three hydrological models on shifted and randomly corrupted precipitation inputs to analyze the extent to which LSTM models mobilized their flexibility relatively to GR4J to compensate for the introduced errors. Our key results are: (1) The best model performances were obtained with the radar-based reanalysis COMEPHORE, with comparable performances for SAFRAN and a large deterioration with ERA5-Land, highlighting the need for gauge-based correction. (2) All models ranked the precipitation products similarly. In particular, the flexibility of LSTM models was not sufficient to compensate for the poorer quality of ERA5-Land. (3) All models responded similarly to unrealistic precipitation inputs. However, LSTM models mobilized their flexibility (larger parameter space, absence of mass-balance constraints) to limit the drop in their performances under highly corrupted precipitation inputs. These counterfactual experiments shed new light on the mechanisms by which LSTM-based models may outperform traditional process-based models, and call for efforts to improve precipitation estimates for hydrological applications.

1 Introduction

1.1 Lack of long-term evaluation of radar-based precipitation reanalyses

Accurate precipitation estimates are crucial for hydrological modelling applications at the catchment scale, especially for short-term flood prediction tasks (Hinge et al., 2022; Roux et al., 2020; Saadi et al., 2023) as well as for long-term water

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

management needs (Dumont et al., 2022). Reference precipitation estimates are commonly obtained from rain gauges, but these underestimate true precipitation amounts due to undercatch (Ehsani and Behrangi, 2022) and may miss the spatial variability of precipitation events due to their sparsity (Villarini et al., 2008), especially in mountainous regions and areas with low population density (Kidd et al., 2017). To overcome this shortcoming, spatial fields of precipitation can be obtained from ground-based weather radars, satellites, or using numerical weather prediction models. However, all these precipitation sources suffer from errors: (1) radar-based measurements are affected by instrumental errors (such as beam blockage, signal attenuation, variability of the vertical reflectivity profile, and mosaicking artefacts; see Berne and Krajewski, 2013; Delrieu et al., 2009; Nelson et al., 2016) in addition to uncertainties in the conversion of reflectivity measurements to precipitation estimates (Lee and Zawadzki, 2005); (2) satellite measurements suffer from sub-pixel heterogeneities and difficulties in handling solid hydrometeors (Maggioni and Massari, 2018), and their conversion to precipitation estimates is sensitive to assumptions regarding cloud types and retrieval and classification algorithms (Kidd and Levizzani, 2011; Sun et al., 2018); and (3) model-based precipitation estimates usually lack sufficient spatial resolution for extreme events and depend on imperfect process parametrizations, despite recent progress towards higher-resolution convection-permitting schemes (Caillaud et al., 2021; Prein et al., 2026).

State-of-the-art precipitation products often merge information from two or more sources of precipitation estimates to take advantage of their complementarity, for example by adjusting satellite- or radar-based precipitation estimates to those from rain gauges (Ochoa-Rodriguez et al., 2019; Xie and Xiong, 2011), assimilating ground and satellite observations into weather prediction models (Hersbach et al., 2020; Lopez, 2011), or applying an additional layer of error-aware weighting techniques to existing precipitation datasets (Beck et al., 2017b; Zhang et al., 2026). This diversity of measurement sources and merging techniques makes it difficult to choose the precipitation dataset that best fits the hydrological modelling application at hand. To guide this choice, several benchmarking studies evaluated the accuracy of precipitation products at the country, regional, or continental scale (Beck et al., 2019; Behnke et al., 2016; Dembélé et al., 2020; Essou et al., 2016; Ollivier et al., 2023; Raimonet et al., 2017; Satgé et al., 2020; Tarek et al., 2020) as well as at the global scale (Abbas et al., 2026; Beck et al., 2017a; Gebrechorkos et al., 2024). However, existing benchmarking studies rarely provide a long-term evaluation (order of 10 years) of high-resolution radar-based precipitation reanalyses at the catchment scale, despite their historical and wide use in hydrological applications (e.g., Borga, 2002 and references therein). Radar-based precipitation reanalyses extend over many years now and are more openly accessible for many regions, for example in the United States (US) with the Stage-IV dataset (Lin and Mitchell, 2005), in Europe with the EURADCLIM dataset (Overeem et al., 2023), and in many other countries with their national products (Overeem et al., 2009; Tabary et al., 2012; Winterrath et al., 2018; Wüest et al., 2010; Yu et al., 2020). Recent studies have addressed this evaluation gap, but not satisfactorily from a hydrological modelling standpoint: they either limited the evaluation to rain gauges with no hydrological evaluation (Hammoudeh et al., 2025; Nelson et al., 2016), or to specific catchments, such as those in mountainous regions (Evin et al., 2024). Addressing this gap by providing a country-wide and long-term hydrological evaluation of radar-based precipitation

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

reanalyses is necessary to quantify the added value of their higher spatial resolution in improving the simulation of long-term catchment response as well as extreme flooding events.

1.2 Can we rely on deep-learning models for the evaluation of precipitation products?

Process-based hydrological models are often used in precipitation evaluation studies not only because they provide a more relevant and independent assessment using data other than rain-gauge estimates (namely streamflow; see Abbas et al., 2026; Beck et al., 2017a; Evin et al., 2024; Raimonet et al., 2017), but also because process-based models are highly sensitive to the quality of precipitation input (Arnaud et al., 2011; Huang et al., 2019; Van Beusekom et al., 2022), and their limited flexibility allows for only a limited compensation to errors in precipitation (Oudin et al., 2006; Wang et al., 2023). This implies that the quality of their simulated hydrographs (measured by the agreement with observed streamflow) can be considered as a good proxy for the quality of the precipitation input. Conversely, deep-learning hydrological models based on long short-term memory networks (LSTM; Hochreiter and Schmidhuber, 1997) are comparatively rarely used for the specific task of benchmarking precipitation products, despite their high performances in catchment-scale rainfall-runoff modelling applications (Arsenault et al., 2023; Hashemi et al., 2022; Kratzert et al., 2018; Lees et al., 2021; Loritz et al., 2024; Mai et al., 2022; Nearing et al., 2024). This can be explained by the higher flexibility of LSTM-based model structures, which may allow them to compensate for errors in precipitation inputs compared to process-based models (Beven, 2020), making them less selective with respect to the quality of precipitation input. Frame et al. (2023) showed that even when enforcing mass-conservation constraints on LSTM models (Hoedt et al., 2021), their good performances were less impacted by the precipitation input compared to a process-based model (SAC-SMA). Li et al. (2025) also observed that LSTM models were less sensitive to the precipitation input than the hybrid and differentiable, mass-balance-constrained δHBV model. However, Wilbrand et al. (2023) reached a different conclusion; they showed that replacing a local, country-wide meteorological forcing dataset with a global one resulted in a strong deterioration of LSTM performances at the hourly time step (from a median Nash-Sutcliffe efficiency of 0.71 to 0.54 over 516 catchments), suggesting that LSTM models can be rather selective with respect to the quality of meteorological forcing.

Using an LSTM model to benchmark precipitation products raises the question of whether the resulting ranking of precipitation datasets can be relied on, especially when it conflicts with that of a process-based model. The literature offers two contrasting answers regarding this question:

1. An “optimistic” answer that argues in favour of the LSTM-based ranking, because on top of its good performances, its internal states are able to reflect some learning of intermediate hydrological fluxes (Kratzert et al., 2018; Lees et al., 2022), suggesting that the learned relationships between precipitation and streamflow are physically consistent. In this vein, Frame et al. (2023) argued that LSTM-based models should be more relied on when evaluating precipitation datasets because they have more ability (thanks to the absence of a-priori constraints) to extract information from the precipitation input and thus to learn an effective rainfall-runoff relationship, in contrast to process-based model structures where poorly implemented mass-balance constraints reduce the information content

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

extracted from the precipitation input. A similar argument can be found in climate-change impact studies, which also compared process-based and deep-learning models regarding their sensitivity to changing meteorological inputs. In this context, Martel et al. (2025) concluded that the streamflow projections of LSTM-based models should be more trusted because they performed better than process-based models both in the historical period as well as in analogue catchments (currently under climate conditions that resemble projected future ones for the target catchments), thanks to their ability to learn from a much more diverse set of catchments when regionally trained.

2. An opposite, more cautious or "pessimistic" answer that doubts the ability of LSTM-based models to learn physically consistent relationships between meteorological forcing and streamflow, implying that LSTM-based models should not be relied on more than process-based ones. Yokoo et al. (2022) showed that an LSTM model was able to produce streamflow with zero precipitation input in a snow-dominated catchment by over-relying on correlations with air temperature. Wi and Steinschneider (2024) showed that without additional mass-balance constraints, LSTM models were unable to learn the role of potential evapotranspiration on evaporative water loss, and instead they extrapolated the historical correlation between air temperature and evaporative water loss to the future regardless of the type (temperature-based or energy-budget-based) of the potential evapotranspiration formula.

We therefore note that this is still a puzzling question that may require adopting an approach akin to metamorphic testing (Reichert et al., 2024), under which we propose that both types of models (process-based and LSTM-based) be fed with unrealistic or corrupted precipitation inputs (as in Yokoo et al., 2022). In this case where we expect the performances of any credible model to seriously deteriorate, LSTM-based models should be less relied on when they mobilize their flexibility to provide realistic streamflow simulations to unrealistic precipitation forcing. Such an experiment would provide a more satisfactory answer to the reliability of LSTM models for the evaluation of precipitation products.

1.3 Research questions and outline of the paper

Based on this literature overview, this study addresses the following research questions:

1. What is the added value of a radar-based, high-resolution precipitation dataset in improving long-term hydrological model performances, namely for the simulation of high flows (exceeded less than 10 % of the time)?
2. Does an LSTM-based model agree with a process-based model regarding the assessment of state-of-the-art precipitation products?
3. How do LSTM-based and process-based models respond to corrupted precipitation inputs?

To answer the first and second questions, we compare COMEPHORE, a country-wide, radar-based high-resolution (1 km^2) precipitation dataset (Tabary et al., 2012), against (1) SAFRAN (Vidal et al., 2010), a country-wide, gauge-based but coarser product (64 km^2), and (2) ERA5-Land (Muñoz-Sabater et al., 2021), a global product with similar resolution to SAFRAN but with no gauge adjustment. We evaluate these precipitation datasets over 25 years (1997-2021) using a locally and a regionally trained LSTM models in addition to the conceptual GR4J model (Perrin et al., 2003). To answer the third

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

question, we retrain and test the three models on shifted and randomly corrupted precipitation time series, and analyze the extent to which their test performances are harmed.

The remaining of this manuscript is organized as follows. Section 2 will present the catchment set and collected precipitation datasets. Section 3 will present the tested models and the numerical experiments we carried out. Results are shown and discussed in Sects. 4 and 5, and Section 6 will summarize the take-away messages and provide some future research perspectives.

2 Catchment set

We extracted a set of 445 (see Fig. 1) out of the 654 catchments of the CAMELS-FR dataset (Delaigue et al., 2025) according to the following requirements:

- We discarded snow-dominated catchments for which the fraction of solid precipitation was higher than 10 % (148/654). This choice helped us avoid an additional complexity regarding snow melting processes and the sensitivity of hydrological models to the quality of an additional meteorological forcing (air temperature).
- We removed catchments that were not completely covered by all precipitation products (9/506).
- We computed the rate of missing daily streamflow values over the period 1997-2021, and we removed the catchments with a rate equal to or higher than 30 % (41/497).
- Finally, we removed the catchments with more than 180 days of zero-streamflow values (11/456), in order to focus on non-intermittent catchments and to conveniently apply logarithmic transformations to streamflow (although irrelevant for the results shown in this study).

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

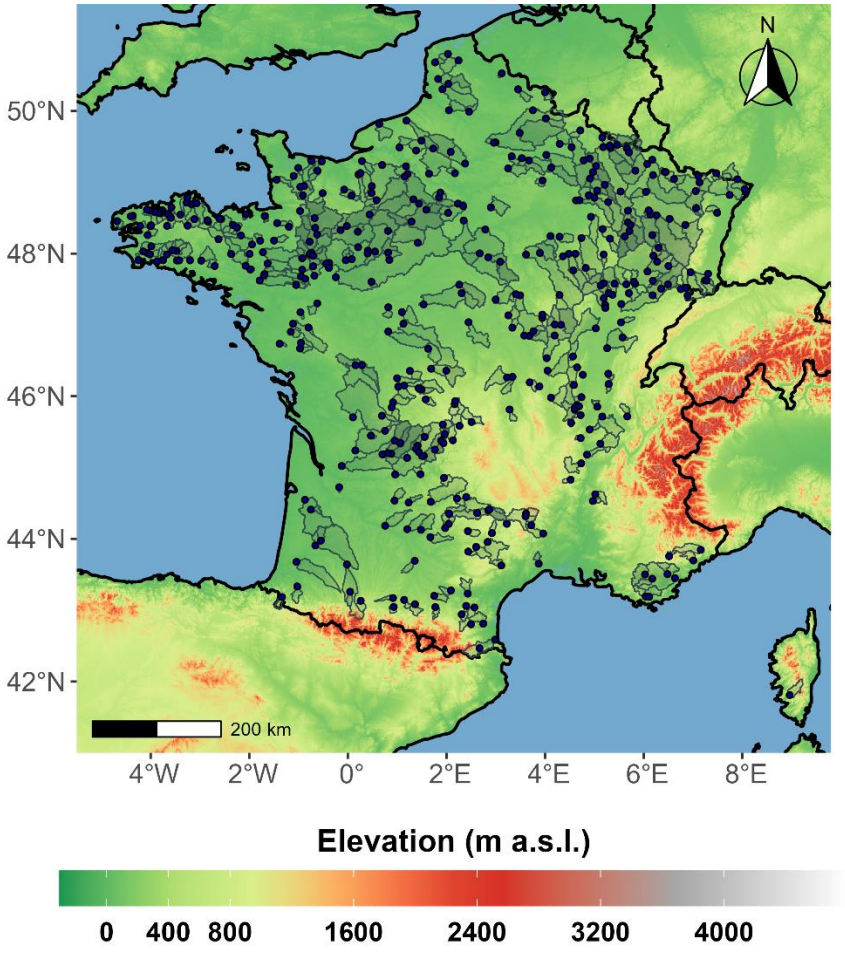


Figure 1: Location of the 445 catchments used in this study. Catchment polygons and outlets (dark blue dots) were provided by the CAMELS-FR dataset. Elevation data were extracted from the SRTM (Shuttle Radar Topography Mission; Farr et al., 2007) dataset available at a resolution of ~90 m (SRTMGL3).

For each catchment, we collected three catchment-average precipitation time series:

1. The daily precipitation estimates from the SAFRAN reanalysis provided by Météo-France (Quintana-Seguí et al., 2008; Vidal et al., 2010), which are readily available in the CAMELS-FR dataset. Precipitation estimates in SAFRAN were produced by an optimal interpolation technique performed over a set of 615 climatologically homogeneous zones (400-1000 km^2) using more than 4000 rain gauges (Le Moigne, 2024, pp. 34-35). To account for orographic effects on precipitation, the interpolation was performed for several elevation tranches at a vertical step of 300 m, where a precipitation gradient of +2 mm/year per 1 m of elevation was applied. Precipitation estimates over the 615 zones were then projected on a regular grid with a cell size of 64 km^2 (8 km). This dataset was evaluated against rain gauges by Vidal et al. (2010) and using hydrological models over 931 French catchments by Raimonet et al. (2017).

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

2. The daily precipitation series from the hourly COMEPHORE dataset provided by Météo-France (Tabary et al., 2012), from which we computed the catchment-average precipitation depth at each hour using the catchment polygon, then summed the hourly precipitation depths for each day d between 06:00 UTC of day d and 06:00 UTC of day $d + 1$. COMEPHORE is a high-resolution (1 km^2) radar-based precipitation reanalysis, which uses maps of precipitation estimates computed from pre-processed radar reflectivity images via a Marshall-Palmer relationship (Marshall and Palmer, 1948). These precipitation estimates were aggregated to the daily time step and then adjusted to match daily precipitation depths from approximately 4200 daily and hourly rain gauges. The resulting calibration factors were interpolated over each 1-km^2 pixel using kriging with external drift. These calibration factors were then used to correct hourly radar accumulations, and the final hourly precipitation estimates were computed from hourly rain gauges and corrected hourly radar accumulations using kriging with external drift. COMEPHORE has been increasingly used in hydro-climatological modelling studies especially for the study of extreme precipitation and high flows (Astagneau et al., 2022; Caillaud et al., 2021; Lobligeois et al., 2014; Thébault et al., 2024; Vigoureux et al., 2024) and the hydrology of urbanized catchments (Saadi et al., 2021). It was recently evaluated over mountainous catchments in the French Alps against other precipitation datasets using two hydrological models by Evin et al. (2024).
3. The daily precipitation time series from the hourly ERA5-Land dataset provided by the ECMWF (Muñoz-Sabater et al., 2021), from which we computed the daily catchment-average precipitation depths using the same procedure as with COMEPHORE. ERA5-Land was produced by a disaggregation of the model-based reanalysis ERA5 (Hersbach et al., 2020) to a resolution of 0.1° (around 78-92 km^2 for our catchment set). To reduce bias in precipitation estimates, ERA5 relies heavily on assimilating satellite observations to improve the estimation of water content in the atmosphere. Ground-based precipitation estimates are assimilated only over the conterminous US (Lopez, 2011). ERA5-Land does not perform an additional correction of precipitation estimates after disaggregation due to constraints of real-time availability (Muñoz-Sabater et al., 2021). ERA5-Land is the reference dataset used for meteorological forcing in Caravan (Kratzert et al., 2023), and it has been evaluated against several national CAMELS datasets using hydrological models by Clerc-Schwarzenbach et al. (2024) and over the French Alps by Evin et al. (2024).

Additionally, we used the time series of daily streamflow and potential evapotranspiration (Oudin et al., 2005) already provided by the CAMELS-FR dataset. Finally, for regional training, we used a set of 17 static catchment attributes (Table 1) to characterize the catchment set in terms of climate (5 attributes), morphology (4 attributes), land use (3 attributes), and soil and subsoil properties (5 attributes). This catchment set covers the majority of metropolitan France except the mountainous regions (Massif Central, Pyrenees, Alps, and Jura; see Fig. 1), and is subject to Mediterranean climate in the south-east, oceanic climate close to the western and northern coasts, and continental climate in the center and to the east. Catchment areas vary between 7 km^2 to ~1.1 10^4 km^2, with 359/445 catchments (80 %) having an area higher than 92 km^2 (the largest cell size of ERA5-Land). Despite removing a significant number of mountainous catchments, the elevation gradient is still

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

quite important with mean catchment elevations between 51 and 1069 m. Land cover is mostly dominated by agricultural and forest areas, with a larger proportion of sand content in the topsoil and subsoil (median at 41 %) compared to silt and clay contents (median at 34 % and 24 %, respectively).

Table 1: Summary of static, catchment attributes and their distributions across the catchment set.

Name (unit)	Min	Median	Max
Aridity index[a] (-)	0.38	0.75	1.28
Number of rainy days[b] (-)	1370	2740	3560
Ratio[c] of 99th precipitation percentile to average precipitation P_{99}/P_m (-)	7.74	9.08	19.00
Seasonality index using potential evapotranspiration[d] (-)	0.27	0.42	0.55
Asynchronicity between the annual precipitation and potential evapotranspiration[d] (-)	0	0.3	0.57
Area[d] (km^2)	7	228	10,790
Mean catchment elevation[d] (m)	51	246	1069
Mean catchment slope[d] (°)	0.88	3.26	19.81
Drainage density[d] (km/km^2)	0	0.45	1.05
Percentage of urban areas[e] (%)	0	2.3	21.4
Percentage of forest areas[f] (%)	0	23.5	90.6
Percentage of agricultural areas[g] (%)	1.34	70.6	97.5
Mean catchment logarithmic permeability[d] ($log_{10}(m^2)$)	-13.42	-11.49	-8.61
Mean catchment porosity[d] (-)	0	0.07	0.23
Clay content[h] (%)	4.05	23.36	43.4
Sand content[h] (%)	22.29	41	87.5
Silt content[h] (%)	8.43	33.98	46.65

[a]Aridity index was computed as the ratio of average potential evapotranspiration to average precipitation (from COMEPHORE) over the years 1997-2021.
[b]A day was considered to be rainy (or wet) when the corresponding daily precipitation depth exceeded 2 mm. This statistic was computed over the years 1997-2021 from the COMEPHORE times series.
[c]Both the 99th percentile and the average precipitation depth were computed using the COMEPHORE time series at the daily time step.
[d]The values for these attributes were taken directly from the CAMELS-FR dataset.
[e]We took the average for the years 1990 and 2018 of the proportion of catchment area covered by the CORINE classes 1.1 (Urban fabric), 1.2.1 (Industrial or commercial units), and 1.2.2 (Road and rail networks and associated land).
[f]We took the average for years 1990 and 2018 of the proportion of catchment area covered by all the CORINE classes under 3.1 (Forest).
[g]We took the average for years 1990 and 2018 of the proportion of catchment area covered by all the CORINE classes under 2 (Agricultural areas).
[h]We took the mean content of the top and subsoil layers for each texture group provided in the CAMELS-FR dataset.

200

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

3 Models and numerical experiments

3.1 Model structures and training details

To compare the sensitivity of deep-learning models to the quality of precipitation estimates with that of process-based, conceptual models, we trained and tested the following lumped, daily hydrological models:

1. A locally trained LSTM model (LSTM-L) that contains one layer with 128 LSTM cells, connected to a dropout layer with a dropout rate at 0.4 to limit the risk of overfitting (Srivastava et al., 2014), and finally connected to the output layer (1 neuron) via a dense layer of 128 neurons activated using a softplus function to limit the output of the LSTM model to nonnegative values (as in Wi and Steinschneider, 2024). The softplus function is an alternative to the rectified linear unit (ReLU) that is differentiable everywhere and preserves nonzero gradients when the input is negative. LSTM-L is trained, validated, and tested at each catchment using sequences of 365 days of meteorological inputs (precipitation and potential evapotranspiration) to produce the single streamflow value at the 365^{th} day. We limited the meteorological inputs to precipitation and potential evapotranspiration because these were the only ones used to feed the process-based alternative (similar to Lees et al., 2021). All input values were standardized by removing the mean and dividing by the standard deviation computed using the training sequences only. Note that this standardization does not allow for testing the effect of systematic biases in precipitation inputs, as will be discussed later (Sect. 3.2.2).
2. A regionally trained LSTM model (LSTM-R) that shares the same structure and sequence length with LSTM-L, in order to quantify the contribution of multi-catchment, regional training (Kratzert et al., 2024). This regionally trained LSTM model was trained and validated on a concatenation of training and validation sequences from all the 445 catchments, and then tested at each catchment individually. In addition to the meteorological forcing features used for LSTM-L, we fed LSTM-R with the set of 17 static attributes (Table 1) by repeating, for each catchment, the value of each static attribute to match the length of the dynamic inputs. We assumed that these static attributes have little interaction with the dynamic inputs and serve only as proxies to map each catchment distinctively to the model space (see, for example, the study of Li et al., 2022). All input features (dynamic as well as static) were standardized using mean and standard-deviation values computed over the concatenation of all training sequences from the 445 catchments.
3. The daily conceptual model GR4J (Perrin et al., 2003), which is a lumped hydrological model that links precipitation and potential evapotranspiration at the catchment scale to streamflow. We chose this model due to its parsimony and its demonstrated good performances over many catchments in France and worldwide (Gu et al., 2023; Raimonet et al., 2017). This model contains four trainable parameters for each catchment: the capacity of the soil-moisture accounting store X_1 (mm), the groundwater exchange coefficient X_2 (mm), the capacity of the nonlinear routing store X_3 (mm), and the time base of unit hydrographs X_4 (d). Simply put, X_1 and X_2 control the mass balance in the catchment, with higher X_1 values and negative values of X_2 leading to more water losses due to

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

evapotranspiration and groundwater-surface water exchanges. X_3 and X_4 control the flashiness and rapidity of catchment response, with higher values of X_3 and X_4 indicating longer recessions and slower catchment responses.

To define the training, validation, and test periods, the first year (1997) was discarded to serve as a warm-up year to obtain a less arbitrary initialization of the GR4J stores. Starting from 1998 onwards, the available streamflow values were divided into three sets: (1) 50 % of the available values served as training data for all models; (2) the subsequent 20% were used for validation of the LSTM models only; and finally (3) the remaining 30 % were used to test all models. The number of training days ranged between 3024 (more than 8 years) to 4382 days (~12 years), with 429/445 catchments having at least ten years' worth of data for training. The number of test days ranged between 1815 (~5 years) to 2630 (more than 7 years), with 371/445 catchments having at least seven years' worth of data for test. All models were trained to minimize the following average Nash-Sutcliffe efficiency (Kratzert et al., 2019; Nash and Sutcliffe, 1970):

$$\mathcal{L}(\theta) = \frac{1}{N_T}\sum_{c=1}^{C}\sum_{t=1}^{N(c)} w(c)\left(Q_{sim}(c,t;\theta) - Q_{obs}(c,t)\right)^2 \quad (1)$$

where θ are trained model parameters, $Q_{sim}(c,t;\theta)$ and $Q_{obs}(c,t)$ are the simulated and observed streamflow for catchment c at day t, $w(c)$ is taken as the inverse of the variance of streamflow observations over the training period for the catchment c, $N(c)$ is the number of time steps for catchment c, and $N_T = \sum_{c=1}^{C} N(c)$, where C is the total number of catchments. Note that in the case of local training ($C = 1$), minimizing $\mathcal{L}(\theta)$ is strictly equivalent to maximizing the traditional Nash-Sutcliffe efficiency (NSE). We trained GR4J to minimize this loss function using a Python implementation of the default step-by-step calibration algorithm in the airGR R package (Coron et al., 2023; Edijatno et al., 1999), which starts by a global search using a Latin hypercube sampling technique to find an initial optimal point, from which a gradient descent algorithm is carried out to estimate the final optimal parameter set. For LSTM-L and LSTM-R, we carried out our training experiments using classes and functions from PyTorch (Paszke et al., 2019) by randomly initializing the weights and biases of each network and using Adam as an optimizer (Kingma and Ba, 2017) with a batch size at 256 and an initial learning rate at 0.001. We then decreased the learning rate to 0.0005 after the 10th epoch and then to 0.0001 after the 30th epoch. We stopped the training when $\mathcal{L}(\theta)$ did not improve over the validation set by at least 0.005 after 40 epochs. We chose this number of epochs to give enough time to the LSTM models to train, a choice that is consistent with the total number of epochs used to train LSTM models in many reference studies (e.g., Frame et al., 2023). For the LSTM-R model, the order in which the catchments were passed to the model was picked randomly at each epoch. The best model (LSTM-L or LSTM-R) on the validation set was then saved to simulate streamflow on the test set. Note that due to computational constraints, the training of LSTM-L and LSTM-R was not repeated to account for uncertainties arising from the random initialization of weights and biases, and the results should be interpreted as single realizations. However, other tests that we carried out with different choices (for example, a different loss function or a higher hidden size) showed limited changes in the distributions of test performances that did not significantly impact the final conclusions.

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

3.2 Numerical experiments

3.2.1 Experiment 1: Evaluation of state-of-the-art precipitation products over France

In this first experiment, we aimed at evaluating the added value of COMEPHORE for lumped hydrological modelling at the catchment scale compared to SAFRAN and ERA5-Land. Comparing COMEPHORE against SAFRAN will show the added value of both the higher spatial resolution of COMEPHORE and the use of radar information to estimate the catchment-scale precipitation. Comparing SAFRAN to ERA5-Land will isolate the added value of using information from rain gauges. Finally, comparing COMEPHORE to ERA5-Land will show the cumulative added value of higher spatial resolution, radar information, and the application of gauge adjustment.

In a first step, we quantified the extent to which the three catchment-average time series were different over the period 1997-2021 according to three statistics: (1) the long-term average of each precipitation input $\mu(P)$, which indicates the total amount of input water at the catchment scale according to the precipitation dataset P, (2) the long-term standard deviation $sd(P)$, which serves to compare the precipitation time series in terms of their temporal variability or regularity (for example, number of wet/dry days), and (3) the 99th percentile of daily precipitation $P_{99}(P)$, which serves to compare the precipitation time series regarding the catchment-scale daily extreme events. Taking COMEPHORE as a reference, we computed the following relative error for both SAFRAN and ERA5-Land for each statistic $s \in \{\mu(P), sd(P), P_{99}(P)\}$:

$$r[s](P) = \frac{s(P)}{s(P_{\text{ref}})} - 1 \quad (2)$$

where P is either SAFRAN or ERA5-Land, and P_{ref} is COMEPHORE. Positive $r[s]$ values indicate an overestimation of the statistic s in SAFRAN/ERA5-Land compared to COMEPHORE, and negative values an underestimation.

In a second step, we calibrated each model (LSTM-L, LSTM-R, and GR4J) using precipitation input from each product (COMEPHORE, SAFRAN, ERA5-Land), and kept the same precipitation product between the training and the test periods. We evaluated the performances of each hydrological model over the test period in order not only to measure the quality of precipitation product in terms of the agreement of model simulations with streamflow observations, but also to measure the quality of precipitation product in terms of model robustness and temporal stability of estimated model parameters. In particular, we analyzed whether the performances of LSTM-L and LSTM-R were less sensitive to the precipitation product compared to GR4J, given their larger flexibility. We limited our test evaluation to four metrics: (1) NSE, to quantify the overall agreement between model simulations and observations (ideal value at 1); (2) the correlation coefficient between the simulated and the observed streamflow values, to quantify the agreement between the temporal dynamics of simulations and observations (ideal value at 1); (3) the relative bias, which corresponds to applying Eq. 2 to long-term averages of simulated and observed streamflow time series (ideal value at 0), in order to quantify whether the long-term streamflow volume is underestimated (negative relative bias) or overestimated (positive relative bias); and finally (4) the relative mean absolute errors restricted to the top 10 % values of observed streamflow over the test period (MAE_{10}), computed as

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

$$\text{Relative MAE}_{10} = \frac{\sum_{t:Q_{obs}(t)\geq Q_{obs,90}} |Q_{obs}(t) - Q_{sim}(t)|}{\sum_{t:Q_{obs}(t)\geq Q_{obs,90}} Q_{obs}(t)} \quad (3)$$

where $Q_{obs,90}$ is the 90th percentile of observed streamflow over the test period, and $Q_{obs}(t)$ and $Q_{sim}(t)$ represent the observed and simulated streamflow at day t.

In a third step, we analyzed the robustness of model performances under contrasting precipitation products between the training period and the test period. Specifically, we compared the option of (1) using a good-quality precipitation product for training and a worse-quality precipitation product for test against (2) using a bad-quality precipitation product for training and a better-quality precipitation product for test. We also analyzed whether LSTM-L and LSTM-R were better at handling this contrast in precipitation quality between the training and test periods compared to GR4J. To do this, we looked at the evolution of test performances using NSE for nine scenarios, corresponding to three precipitation products in training and three precipitation products in test. Three scenarios correspond to the case where the precipitation product is kept between the training and the test (for example, COMEPHORE to COMEPHORE). The remaining six scenarios correspond to contrasting precipitation products between the training and the test periods (COMEPHORE to SAFRAN and vice versa, COMEPHORE to ERA5-Land and vice versa, and SAFRAN to ERA5-Land and vice versa). Note that for LSTM-L and LSTM-R, the precipitation product used during the validation phase is the same as the one used during the training phase.

3.2.2 Experiment 2: Sensitivity of data-driven and process-based models to shifts and random errors in input precipitation

In order to decide which model assessment to rely on when the three models disagree, we conducted a second experiment in which we retrained each model on unrealistic precipitation inputs: (1) shifted precipitation, and (2) randomly corrupted precipitation. We did not test systematically biased precipitation inputs because our implemented LSTM models are insensitive to this corruption due to standardization of meteorological inputs. A similar approach has been implemented by Oudin et al. (2006) for two lumped conceptual models (GR4J and TOPMO), which showed that the storage components were extremely challenged by severely corrupted precipitation inputs, leading to a large deterioration of their performances. Here, we analyze whether this is also the case for the data-driven models LSTM-L and LSTM-R.

First, we trained and tested all three models using a family of unrealistic precipitation time series constructed by shifting the COMEPHORE time series to the past or to the future according to:

$$\hat{P}(d) = P(d - \Delta) \quad (4)$$

where $\hat{P}(d)$ and $P(d)$ are the shifted and the original precipitation time series at day d, and Δ represents the shift, which is negative when precipitation is shifted to the past and positive when shifted to the future. We tested values for $\Delta \in \{-3, -1, +1, +3\}$ days, where we expect the model performances to significantly deteriorate especially for positive temporal shifts and less so for negative temporal shifts, since the needed information on precipitation to reproduce a streamflow event is completely lost with positive shifts, whereas it is still accessible for negative shifts.

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

Second, we retrained and tested all models using an additional family of unrealistic precipitation inputs produced by randomly corrupting the COMEPHORE precipitation input at the catchment scale using a lognormal multiplicative error with autocorrelation in time:

$$\tilde{P}(d) = \kappa(d;\sigma,\rho)\, P(d) \quad (5)$$

where $\tilde{P}(d)$ and $P(d)$ are the corrupted and the original precipitation depths at day d, and $\kappa(d;\sigma,\rho) = \exp\left(-\frac{\sigma^2}{2} + \sigma\eta(d)\right)$ represents the corruption ration, taken as a random variable following a lognormal distribution with an expectancy $\mathbb{E}(\kappa) = 1$ so that the total volume of precipitation is not impacted by the random corruption. σ controls the intensity of the random corruption, with higher values indicating stronger corruption. The temporal persistence is implemented by choosing $\eta(d)$ to be a gaussian autoregressive variable with null expectancy and unit variance such as $\eta(d) = \rho\eta(d-1) + \epsilon(d)$, where $\epsilon \sim \mathcal{N}(0, 1-\rho^2)$ and ρ controls the temporal correlation (see Fig. A1 in Appendix A for an illustration of the role of both σ and ρ). The multiplicative error model of Eq. 5 is commonly chosen in the study of the role of input uncertainty on the predictive uncertainty of hydrological models (e.g., Kavetski et al., 2006), and it was found to be adequate to characterize the error structure of rainfall errors, except for the tails (McMillan et al., 2011). We tested two values of $\sigma \in \{0.3, 0.8\}$ to analyze the impact of the intensity of the errors on model performances, and we expect the performances of GR4J to substantially deteriorate for $\sigma = 0.8$ according to the results of Oudin et al. (2006). For temporal persistence, we tested an uncorrelated case ($\rho = 0$) and a strongly persistent case ($\rho = 0.6$; similar to Moulin et al., 2009) to evaluate the effect of the temporal persistence of errors on model performances. Here, we expect the stores of GR4J to be more challenged when an underestimation/overestimation of precipitation persists in time, leading to a stronger deterioration in model performances. While the response of GR4J to these corruptions of precipitation inputs is a priori known, it is not well clear whether the LSTM models would be able to learn to compensate for these introduced errors during the training phase and transfer well to the test phase (i.e., keeping the same parameter values between training and test), namely for the case of autocorrelated errors.

4 Results

4.1 Results for experiment 1: Evaluation of state-of-the-art precipitation products over France

Despite similar spatial resolutions of SAFRAN (8 km) and ERA5-Land (8.8-9.6 km), we found a stronger agreement between SAFRAN and COMEPHORE than between ERA5-Land and COMEPHORE, as highlighted by Fig. 2. Long-term catchment-scale precipitation averages were very similar between SAFRAN and COMEPHORE except for a handful of catchments (Fig. 2a), which is reflected by relative bias values that were centred around zero (Fig. 2c). For ERA5-Land, the points show a larger spread (Fig. 2b), with ERA-Land producing significantly lower precipitation amounts especially at smaller catchments (Fig. 2b), and catchments located close to the coastal regions (see Fig. A2). However, this underestimation was compensated by an overestimation of precipitation amounts at catchments located farther from the

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

coastal regions (Fig. A2), leading to a country-wide, median relative bias of zero (Fig. 2c). Relative errors in variability show that the catchment-scale precipitation time series of both SAFRAN and ERA5-Land had smaller standard-deviation values compared to COMEPHORE (Fig. 2d), suggesting more rainy days in SAFRAN and ERA5-Land compared to COMEPHORE. However, the underestimation of standard deviation was much stronger with ERA5-Land (median at -0.12) than with SAFRAN (median at -0.03), indicating that contrasts in spatial resolution alone do not explain the differences with COMEPHORE. Similarly, both SAFRAN and ERA5-Land underestimated the extreme catchment-scale daily precipitation (99th percentile; Fig. 2e), which is partly (but not fully) attributed to their coarser spatial resolutions that may have smoothed extreme precipitation values.

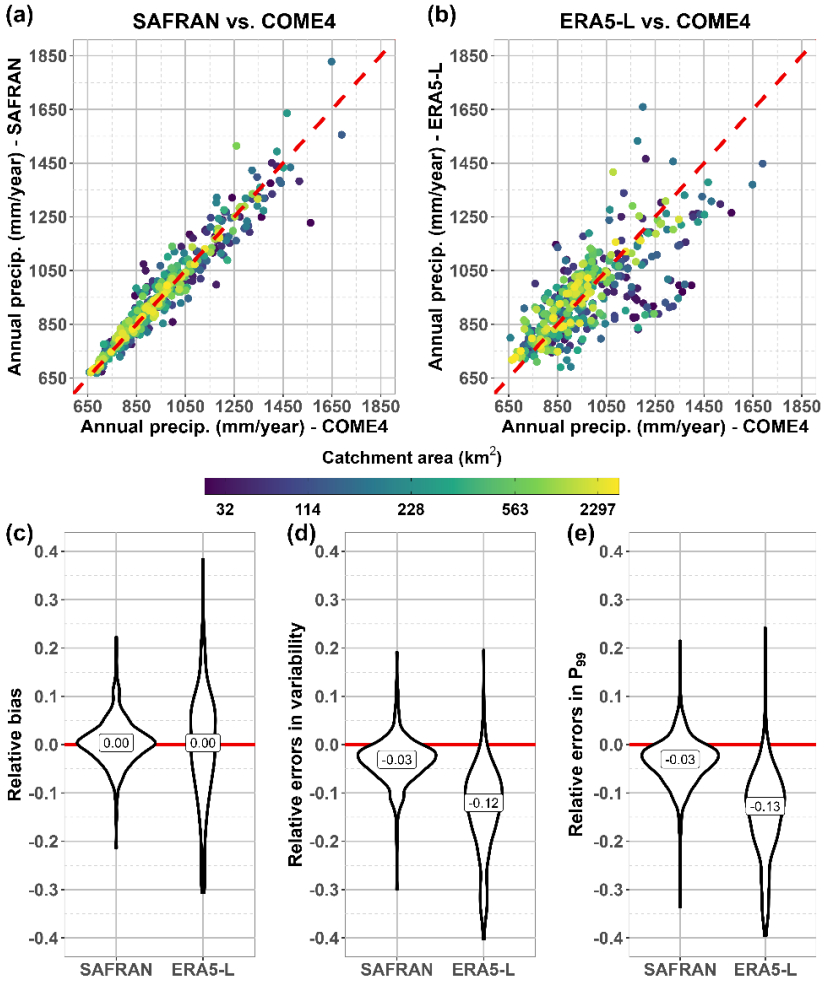


Figure 2: Comparison of long-term (1997-2021) and catchment-scale precipitation averages (a) of SAFRAN and COMEPHORE (COME4), and (b) of ERA5-Land (ERA5-L) and COMEPHORE (COME4), where each dot represents one of the 445 catchments, with colors varying according to the catchment area. Distributions of relative errors for SAFRAN and ERA5-Land with COMEPHORE as a reference in terms of (c) long-term averages (relative bias), (d) long-term standard deviations (variability), and (e) 99th percentile of daily precipitation (P_{99}). Each distribution contains 445 points.

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

All three hydrological models (GR4J, LSTM-L, and LSTM-R) showed a limited added value of COMEPHORE's higher spatial resolution compared to SAFRAN at the catchment scale, despite scoring the best overall performances with COMEPHORE. In addition, they all registered a significant drop in training and test performances when fed with ERA5-Land instead of COMEPHORE or SAFRAN, as shown by Table 2 and Fig. 3. This highlights that, in particular, the flexibility of LSTM models was not helpful in maintaining the same level of performances for all precipitation products. In detail:

- The best training performances (Table 2) were obtained with COMEPHORE (median NSE at 0.89-0.92) and SAFRAN (median NSE at 0.88-0.92). These performances dropped by 0.09-0.13 in median NSE when taking the precipitation from ERA5-Land as input during the training period (median NSE at 0.78-0.82).
- The degradation in overall model performances with ERA5-Land is more striking over the test period (Fig. 3a). The best median NSE values were obtained with COMEPHORE (0.86-0.88) and SAFRAN (0.84-0.87), whereas they dropped by up to 0.15 with ERA5-Land (0.70-0.77), with a larger drop in performances with LSTM models than with GR4J.
- The correlation scores (Fig. 3b) and errors at high flows (top 10 %; Fig. 3d) indicate that ERA5-Land adversely affected the capacity of all models in producing timely and accurate streamflow simulations, especially at the event scale. Median correlation scores dropped from 0.92-0.95 with COMEPHORE and SAFRAN to 0.85-0.89 with ERA5-Land (Fig. 3b), and median relative MAE_{10} raised from 0.20-0.23 to 0.27-0.31 (Fig. 3d). However, median relative bias values suggest that all models had more ability in reproducing observed water volumes with ERA5-Land (median values around zero) than with COMEPHORE or SAFRAN (median values from -0.03 to -0.09; Fig. 3c).

Table 2: Median model performances over the training period expressed using the Nash-Sutcliffe efficiency (NSE) for each precipitation dataset. Each median was computed from a sample of 445 values corresponding to the 445 catchments.

Model	Precipitation dataset		
	COMEPHORE	SAFRAN	ERA5-Land
GR4J	0.89	0.88	0.78
LSTM-L	0.92	0.91	0.82
LSTM-R	0.90	0.92	0.79

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

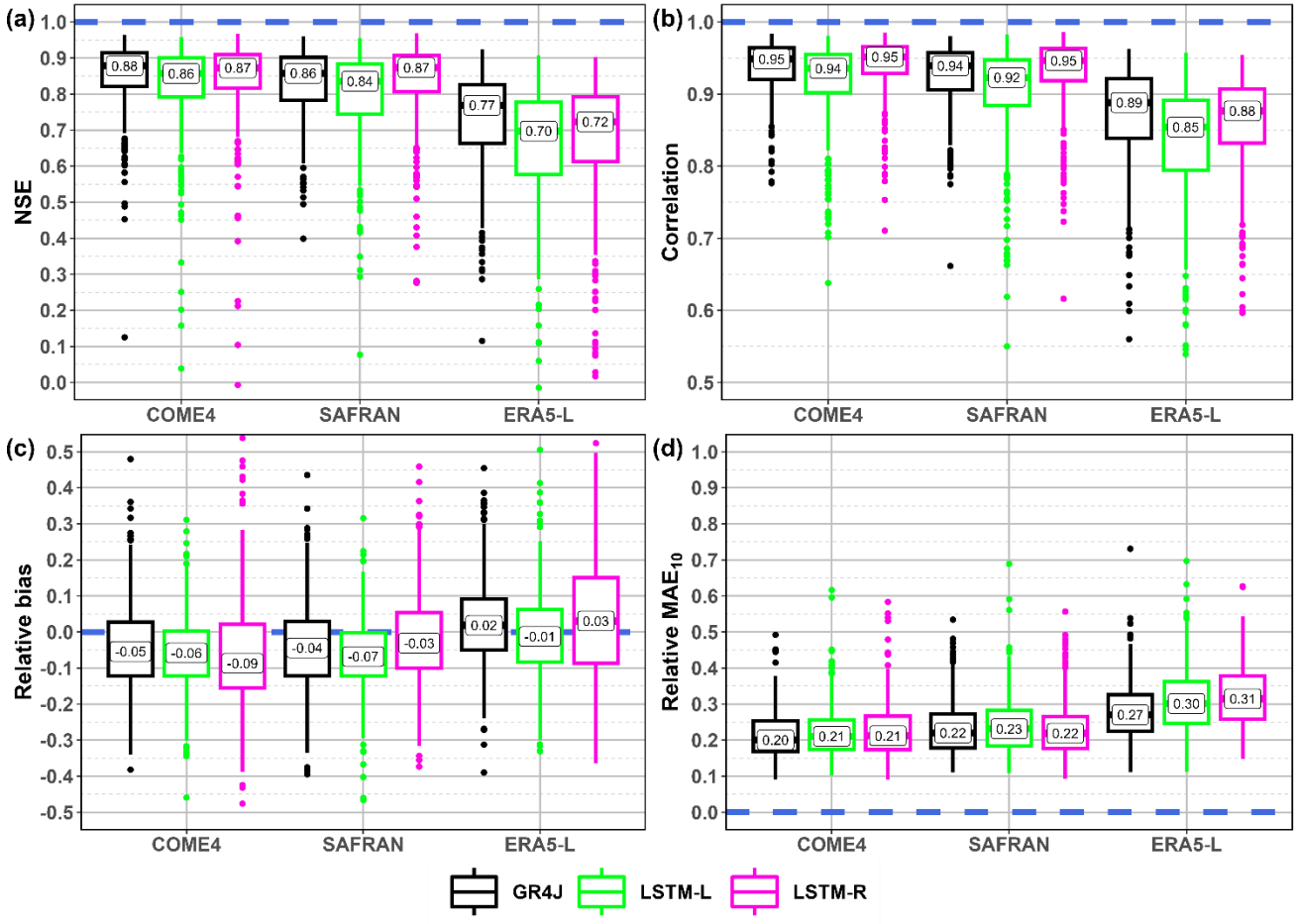


Figure 3: Test results for the three hydrological models (GR4J, LSTM-L, LSTM-R) fed with catchment-average precipitation time series from three datasets (COME4 for COMEPHORE, SAFRAN, and ERA5-L for ERA5-Land) with the following evaluation metrics: (a) Nash-Sutcliffe efficiency (NSE), (b) correlation, (c) relative bias, and (d) relative MAE_{10}. Each distribution contains 445 points.

For model performances over the test period, the quality of the precipitation product chosen for test was more impactful than the quality of the precipitation product chosen for training, as highlighted by Fig. 4. In other words, all models transferred better from a poor-quality precipitation product in training to a good-quality precipitation product in test than from a good-quality precipitation product in training to a poor-quality precipitation product in test. Changing the precipitation product from COMEPHORE to SAFRAN and vice versa between the training and test periods had limited impact on test performances for all models. Training the models on ERA5-Land and testing them on COMEPHORE or SAFRAN led to better test performances than keeping ERA5-Land for test, with the best gains in median NSE obtained by LSTM-L (from 0.70 to 0.76-0.78). Conversely, training the models on COMEPHORE or SAFRAN and testing them on ERA5-Land led to test performances that were usually worse than keeping ERA5-Land for both training and test, with

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

significant drops in performances for GR4J (from 0.77 to 0.68) and LSTM-R (0.72 to 0.60-0.64) and limited gains for LSTM-L (from 0.70 to 0.72).

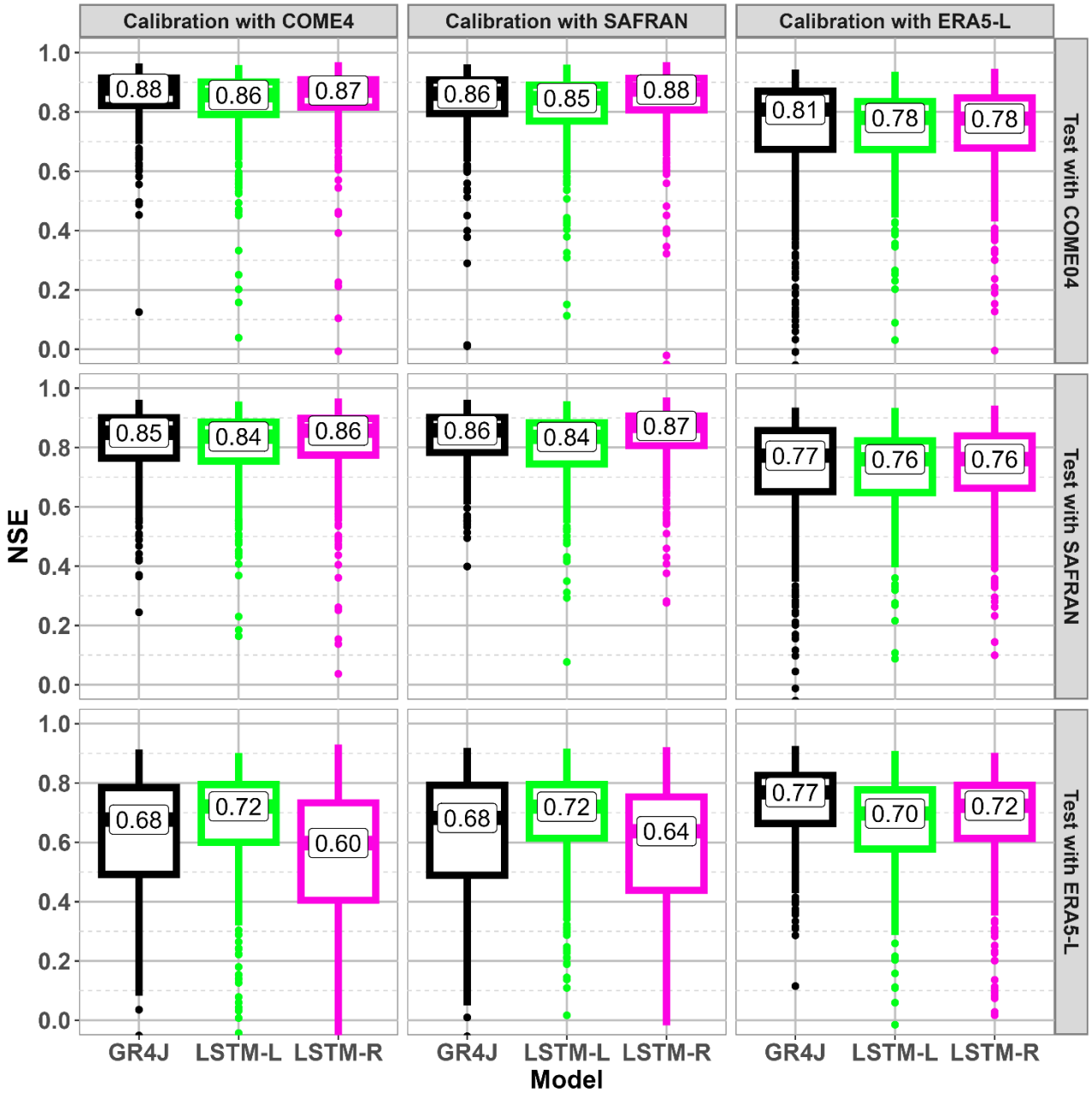


Figure 4: Distribution of Nash-Sutcliffe efficiency over the test period for all combinations of pairs of precipitation products used for training and test (COME4 for COMEPHORE, SAFRAN, and ERA5-L for ERA5-Land). Columns indicate the precipitation product used during the training period, and rows indicate the precipitation product used during the test period. Note that the distributions on the diagonals correspond to keeping the same precipitation product between the training and the test periods, i.e. to performance values shown in Fig. 3a.

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

4.2 Results for experiment 2: Sensitivity of data-driven and process-based models to shifts and random errors in input precipitation

The test performances of all three models (GR4J, LSTM-L, and LSTM-R) significantly deteriorated when trained and tested using randomly corrupted or shifted precipitation inputs to the future, as shown in Fig. 5. However, while the three models responded quite similarly to shifts in precipitation inputs (Figs. 5a and 5c), the pace at which the performances deteriorated with increasing intensity and persistence of random corruption was faster with GR4J than with the LSTM models (Figs. 5b and 5d). More specifically:

- Shifting precipitation inputs to the past resulted in limited deterioration of model performances, with the largest drop in median NSE from 0.86 to 0.81 and increase in median MAE_{10} from 0.21 to 0.25 for LSTM-L (Figs. 5a and 5c). Shifting the precipitation inputs one day to the future had also limited impacts on model performances, with even a slight gain in median NSE and MAE_{10} for LSTM-R. However, shifting the precipitation inputs three days to the future resulted in significant deterioration of model performances as expected, with all three models reacting in a similar way. In particular, the larger flexibility of LSTM-L and LSTM-R did not compensate for the loss in information induced by shifting the precipitation events three days to the future. Finally, the sensitivity of model performances to shifts in precipitation inputs depended on the catchment size (see Fig. A3), with lower deterioration of model performances for larger catchments (which likely have longer response times) than for medium or smaller catchments.
- Model performances significantly deteriorated with increasing intensity σ and temporal persistence ρ (Figs. 5b and 5d). With a mild random corruption ($\sigma = 0.3$), median NSE dropped similarly for all models from 0.86-0.88 to 0.78-0.81 with no persistence and to 0.73-0.76 with a strong persistence ($\rho = 0.6$). However, with a more intense random corruption ($\sigma = 0.8$), median performances of GR4J dropped to 0.41 and 0.27 for $\rho = 0$ and $\rho = 0.6$, respectively, whereas they dropped only to 0.59 and 0.50 respectively for LSTM-L and to 0.62 and 0.57 respectively for LSTM-R (Fig. 5b). Similar patterns of performance deterioration were also observed for high flows (Fig. 5d), suggesting that model performance concerned mostly the catchment response during rainfall events. These observations highlight three points: (1) with persistent, autocorrelated errors, all models struggled more than with uncorrelated errors (that tend more to compensate during an event), (2) LSTM models somewhat mobilized their larger flexibility (larger parameter space, and absence of mass-balance constraints) to limit the deterioration in model performances, and (3) regional training helped the LSTM models to cope better with random errors in precipitation, but not at high flows (Fig. 5d).

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

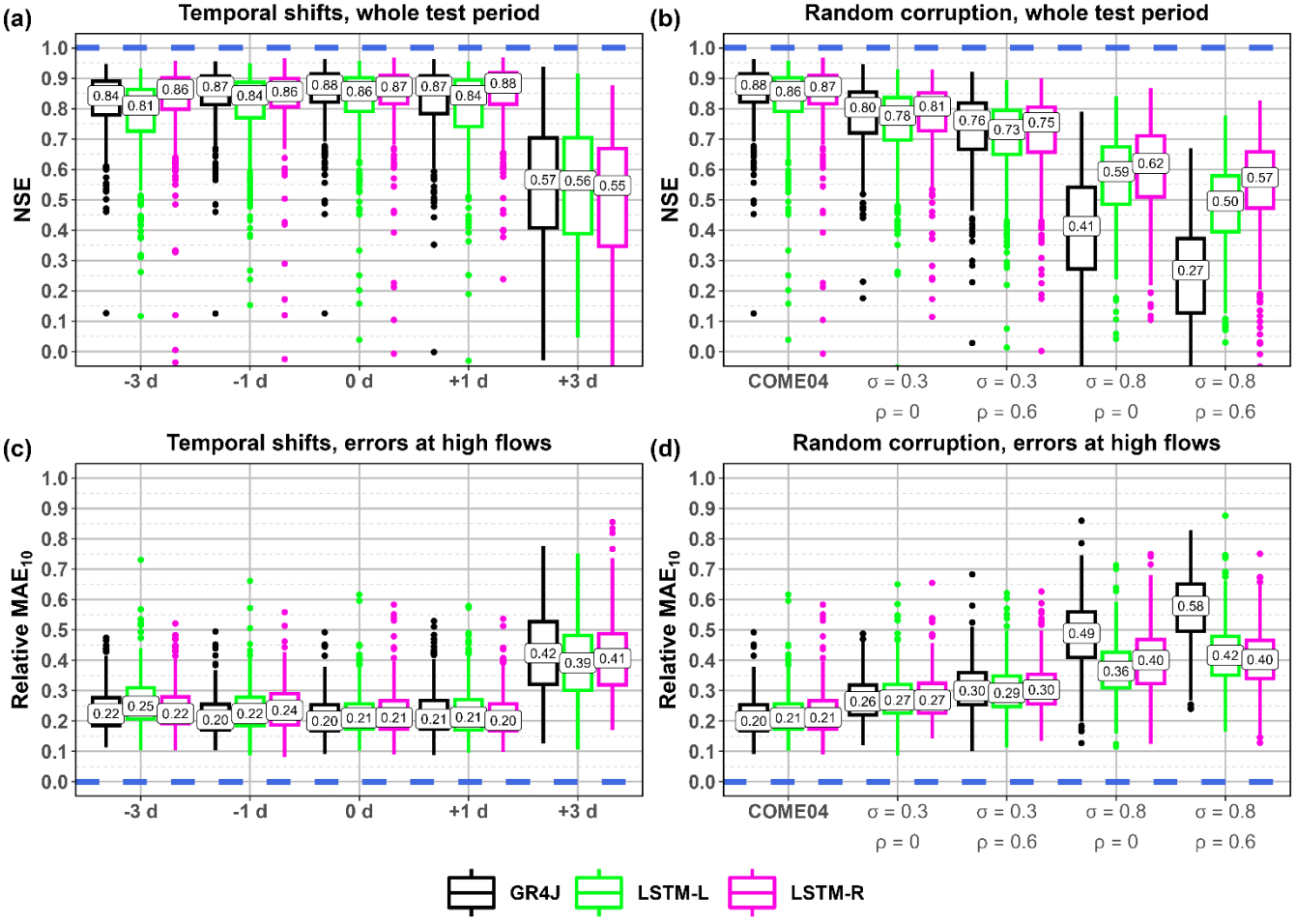


Figure 5: Test performances with Nash-Sutcliffe efficiency (NSE) and relative MAE_{10} as evaluation metrics for models trained and tested using shifted (a and c) and randomly corrupted precipitation inputs (b and d). For shifted precipitation inputs, values indicate the length of the shift in days, with negative values indicating shifts to the past and positive values indicating shifts to the future. For randomly corrupted precipitation inputs, σ indicate the intensity, and ρ the temporal correlation (or persistence).

To investigate the discrepancies in the pace of performance deterioration between GR4J and LSTM models due to random corruption of precipitation, two examples with representative performances for GR4J, LSTM-L, and LSTM-R are shown in Fig. 6 for the most intense and persistent corruption ($\sigma = 0.8$ and $\rho = 0.6$). These two representative examples provide us with elements to understand when GR4J performances dropped significantly while LSTM models succeeded at limiting the deterioration. Note that GR4J reacted to random corruption by increasing the buffering capacity of its production and non-linear transfer stores (X_1 and X_3) and taking increasingly negative X_2 values, as shown in Fig. A4. This reaction attempts at counteracting the effect of large corruption ratios κ during precipitation events, by losing the exceeding water to evapotranspiration or groundwater and smoothing the slow-flow contribution to total streamflow. Keeping this in mind, we observe that

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

- in the first case (Fig. 6a), GR4J stores were not at the maximum capacity allowed by the optimization algorithm, and so the model simply responded to unrealistically extreme precipitation depths by producing unrealistically high peak flows under mass-balance constraints, resulting in deteriorated model performances. In contrast, LSTM-L and LSTM-R produced peak flow estimates that were lower than the observed peak flow especially for the event with the largest peak flow during the test period, limiting the deterioration in model performances compared to GR4J.
- In the second case (Fig. 6b), GR4J stores were at maximum capacity, and consequently model simulations were highly smoothed, resulting in a deteriorated NSE. In contrast, the flexibility of LSTM allowed for reproducing the observed dynamics, but it did not succeed in reproducing the largest observed peak flow, leading to a fair but significantly better NSE than GR4J over the test period.

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

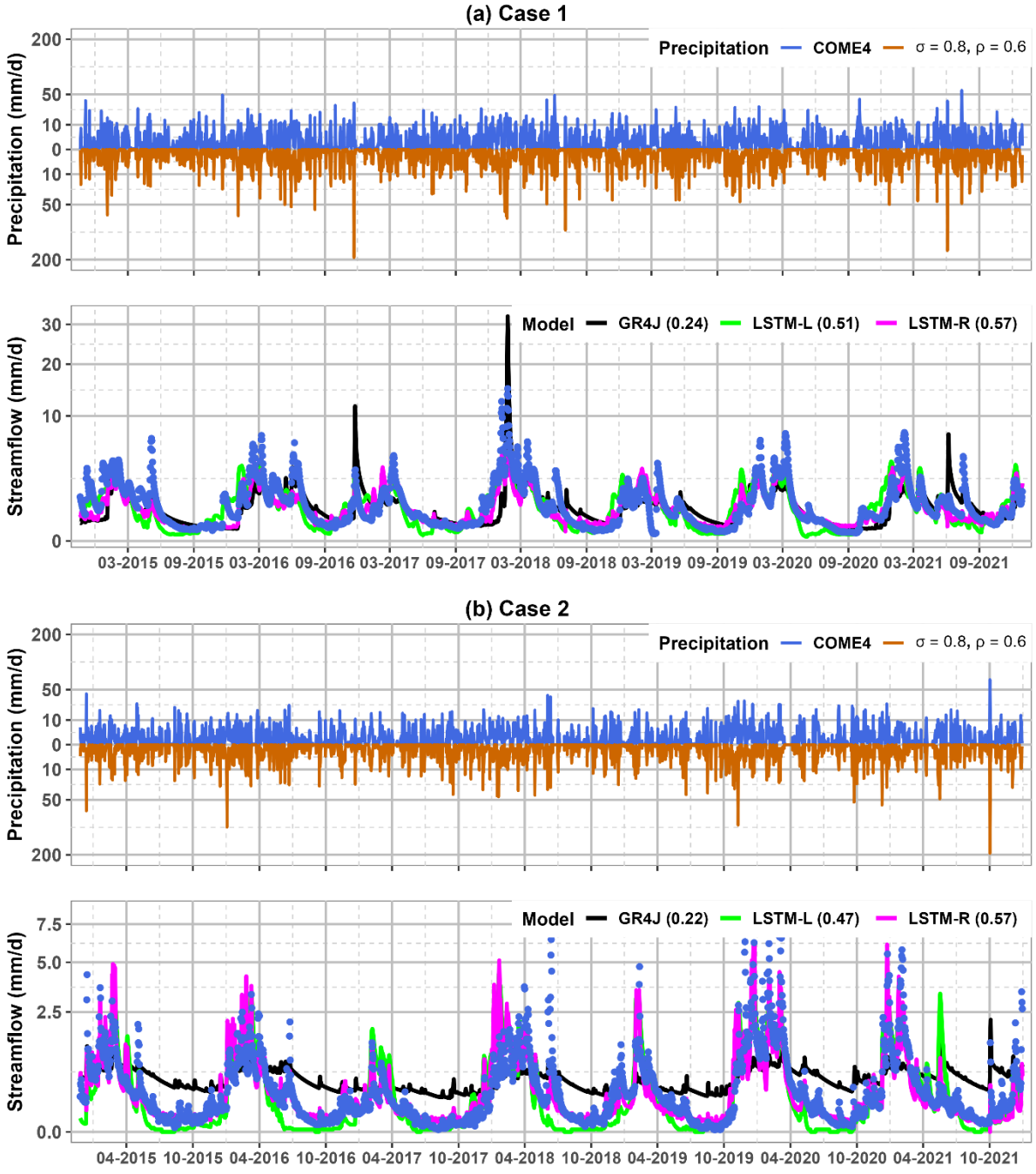


Figure 6: Examples of model response to randomly corrupted precipitation ($\sigma = 0.8, \rho = 0.6$) for two representative catchments with model performances that are close to median performances for GR4J, LSTM-L, and LSTM-R according to Nash-Sutcliffe efficiency (see Figure 5b). Blue dots represent the observed streamflow. Case 1 represents the model response for the catchment drained by the Aube river at Bar-sur-Aube (station code: H120101001, area: 1300 km^2). Case 2 represents the catchment drained by the Oudon river at Châtelais (station code: M377181010, area: 730 km^2). Note that vertical axes were transformed using a square-root function to enhance readability.

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

5 Discussion

5.1 Added value of the radar-based precipitation reanalysis

Our results showed that the radar-based reanalysis COMEPHORE was the best precipitation product in terms of overall model performances over the whole test period as well as at high flows (Figs. 3a, 3b, and 3d), which confirms and completes the results obtained by Evin et al. (2024) for mountainous catchments in the French Alps with two distributed hydrological models. Our results with lumped models at the catchment scale highlight the gains in accuracy brought by the use of radar reflectivity fields to better estimate the catchment-scale precipitation depths, which are especially beneficial for catchment response to extreme precipitation events with high spatial variability (Emmanuel et al., 2017; Lobligeois et al., 2014). However, gains in model performances with respect to the coarser SAFRAN dataset seem modest (see Fig. 3), which can be explained by the catchment-scale averaging that might have masked the benefits of finer spatial resolution, especially for large catchments with lumped models. The large drop in model performances with ERA5-Land compared to COMEPHORE and SAFRAN is consistent with the results of Evin et al. (2024) and Clerc-Schwarzenbach et al. (2024), and underlines the critical importance of using information from rain gauges to produce accurate precipitation estimates from a hydrological standpoint (Abbas et al., 2026; Beck et al., 2019). It can be attributed to ERA5-Land overestimating the number of wet days (Fig. 2d; Bandhauer et al., 2022) and underestimating extreme events (Fig. 2e). Comparing SAFRAN and ERA5-Land also shows that having a similar spatial resolution does not imply having a similar quality (Behnke et al., 2016; Raimonet et al., 2017). However, the long-term averages of streamflow were better reproduced with ERA5-Land than with COMEPHORE or SAFRAN (Fig. 3c). This highlights that the ranking of precipitation products should rely on complementary evaluation metrics to fully inspect the merits of each product. Obtaining satisfactory performances according to all metrics can be achieved by leveraging the complementarity between precipitation thanks to ensemble methods (Li et al., 2025) or learnable merging of precipitation products (Sawadekar et al., 2025).

5.2 Precipitation quality has greater influence than model choice

Our comparison framework involved contrasting precipitation products as well as contrasting modelling approaches, in which we found that the differences between precipitation products dominated the differences between process-based (GR4J) and LSTM models (LSTM-L and LSTM-R). This was illustrated by (1) the relatively lower contrasts between GR4J and LSTM models for each precipitation product, compared to the changes in the performances of each model across precipitation datasets (Fig. 3), and (2) the large deterioration of model performances during the test period when using a poor-quality precipitation dataset (Fig. 4), despite being calibrated using a good-quality precipitation dataset. The results of Fig. 3 suggest that with good-quality precipitation products that benefit from a high-density gauge network, process-based conceptual models can reach performances that are as good as those of deep-learning models, implying that efforts to improve the hydrological prediction should primarily target the quality and the uncertainties in meteorological and hydrological data (Beven, 2019; Kim et al., 2026; Sperna Weiland et al., 2015). This conclusion is also supported by the

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

inability of locally or regionally trained LSTM models to maintain good performances with ERA5-Land, a result that has been also observed by Wilbrand et al. (2023) over US catchments with ERA5 as input. These results are not completely in line with the conclusions of Frame et al. (2023) or Kratzert et al. (2021), who argued that the LSTM models were skilful enough across all US precipitation datasets to the point of completely dominating the process-based alternative, leading Frame et al. (2023) to even question the suitability of mass-conservation constraints for hydrological modelling at the catchment scale. While we acknowledge that modelling and process-representation choices are a dominant factor behind the differences in model skill, our results suggest that efforts should first focus on obtaining and using a high-quality precipitation product before drawing conclusions about the adequacy of current hydrological theory and process-based modelling for hydrological prediction.

The asymmetric model transferability between good-quality and bad-quality precipitation products from the training to the test period shown in Fig. 4 echoes the results of Bárdossy and Das (2008), who showed that models had better transferability results when calibrated on a low-density gauge network and tested on a high-density gauge network than vice versa. This was explained in the case of Bárdossy and Das (2008) by the fact that low-density networks may have missed precipitation events in the test phase that could not be compensated by the model parameters, and a similar argument could be made here for ERA5-Land, which does not use information from rain gauges. Our results imply that while models allow for greater flexibility in choosing the precipitation product for their training, the skill of their application in an independent period depends critically on the quality of the precipitation input. They also suggest an empirical approach to compare two precipitation products, by evaluating the ability of hydrological models to transfer between the two from the training to the test phase, assuming better transferability from a bad-quality precipitation product in training to a good-quality precipitation product in test.

5.3 When to rely on LSTM models?

LSTM models showed a similar sensitivity to unrealistic precipitation inputs (randomly corrupted or shifted, Fig. 5), which reassured us that LSTM models cannot magically compensate for losses in information in precipitation, and thus they can be relied upon for the evaluation of precipitation datasets. However, for the case of randomly corrupted precipitation (Figs. 5b and 5d), they somewhat mobilized their flexibility features (larger parameter space, and absence of mass-conservation constraints) to limit the deterioration in NSE over the test period, especially with the regionally trained model. For example, instead of responding to an unrealistically high precipitation event in Fig. 6a, LSTM-L responded with a rather lower peak flow. This feature is particularly malignant because LSTM models prioritized performance over physical consistency in this case, in contrast to GR4J. This tempers the arguments made by Martel et al. (2025) and invite us to rely more on the process-based models than on LSTM ones (especially for the case of simulating future extreme events). Conversely, when the lower flexibility of GR4J was extremely challenged in the case of Fig. 6b, LSTM-L was able to at least reproduce the dynamics, perhaps because the information on dry/wet days was still preserved in the corrupted precipitation input. This highlights a case where the flexibility of LSTM models could be benign compared with the relative inability of GR4J in extremely

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

challenging catchments. This analysis calls for using physical constraints to tame the flexibility of LSTM models, for example as proposed by Wi and Steinschneider (2024) with constraints on both precipitation and evaporative demand, or by benefiting the physical consistency and interpretability of process-based models from the flexibility of LSTMs (Feng et al., 2024; Li et al., 2025; Song et al., 2026; Tsai et al., 2021), although recent studies reported that the flexibility of LSTMs may take over the constraints of process-based models in such hybrid approaches (Álvarez Chaves et al., 2026).

5.4 Study limitations

Our study suffers from several limitations. First, our methodological setup did not explicitly handle the uncertainties emanating from the random initializations of LSTM parameters. Despite using an independent validation period to halt the training and limit the dependency on initialization, the interplay between parameter initialization and test results can only be assessed using several training seeds, which we could not do due to computational constraints. However, we repeated the whole experiment with a hidden size of 256, and despite some variability in median performances, our conclusions were not significantly impacted. In addition, we found comparable test results with SAFRAN to those of Hashemi et al. (2022) and with ERA5-Land to those of Puche et al. (2025) for the French catchments, suggesting that our single-realization performances of LSTM models may not be far from the case where many initializations would be used. Second, due to the fact that our LSTM results should only be interpreted as a single realization with respect to random initializations, we could not confidently discuss the discrepancies in model performance and robustness between GR4J and LSTM models (see Table 2 and Figs. 3-4), despite some previous work pointing to the lower robustness of LSTM models compared to process-based ones (Natel de Moura et al., 2022). Third, we could not analyze whether adding mass-conservation constraints to the LSTM models could impact their sensitivity to corrupted precipitation inputs, a limitation that is left for future work, in which the impact of systematic errors could also be analyzed (Oudin et al., 2006). Finally, our evaluation should be extended to mountainous, snow-dominated catchments to compare the topography-aware SAFRAN dataset with COMEPHORE at the daily time step, or to the scale of several countries to provide a large-scale assessment of the added value of radar-based precipitation reanalyses. The extension of our sensitivity analysis to snow-dominated catchments would shed more light on the effect of corrupting other meteorological forcing variables (namely air temperature) in addition to precipitation in order to complete/extend the results of Yokoo et al. (2022) and Reichert et al. (2024).

6 Take-away messages and future research perspectives

We provided a catchment-scale assessment of three contrasting precipitation datasets over France: the radar-based reanalysis COMEPHORE (1 km^2), the gauge-based gridded dataset SAFRAN (64 km^2), and the model-based reanalysis ERA5-Land (78-92 km^2). Over 445 French catchments, we trained and tested GR4J against locally and regionally trained LSTM models using the state-of-the-art precipitation products, in addition to unrealistic precipitation inputs to assess their ability in learning physically consistent rainfall-runoff relationships. Our key findings are as follows:

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

1. The best performances were obtained by the radar-based reanalysis COMEPHORE, with SAFRAN scoring comparable performances. However, all models showed a large drop in performances with ERA5-Land, highlighting the need for gauge-based correction to obtain hydrologically accurate precipitation estimates.
2. LSTM models provided a similar ranking of precipitation datasets to that of GR4J. In particular, their flexibility did not compensate for the relatively lower quality of ERA5-Land.
3. All models showed a significant drop in performances with randomly corrupted and shifted precipitation inputs to the future. However, LSTM models mobilized their flexibility and benefited from regional training to limit the deterioration of performances induced by random corruption.

Future work could extend this analysis to evaluate the added value of radar-based reanalyses at the European scale (Hammoudeh et al., 2025; Overeem et al., 2023) or at US catchments. Including mass-conservative LSTM models (Frame et al., 2023; Wi and Steinschneider, 2024) will shed more light on the benefits of mass-conservation constraints in increasing the reliability of LSTM models. Finally, extending our sensitivity experiment to other meteorological forcing variables (especially on snow-dominated catchments) and to a combination of systematic and random errors will provide a more complete picture regarding the sensitivity of process-based and deep-learning models to contrasting types of errors.

Appendix A: Additional useful figures

In this appendix, we provide four supplementary and useful figures:

- Figure A1 illustrates the role of the intensity σ and the temporal persistence ρ in controlling the variability of the corruption ratio $\kappa(d;\ \sigma, \rho)$. Comparing Figs. A1a-c against A1b-d shows that σ controls the variance of κ (recall that $\mathbb{V}[\kappa] = e^{\sigma^2} - 1$). The temporal correlation ρ controls the persistence of the errors, which is illustrated by Figs. A1c and A1d, where overestimation and underestimation periods tend to last longer.
- Figure A2 shows the spatial distribution of relative biases in long-term precipitation amounts between ERA5-Land and COMEPHORE, with COMEPHORE taken as the reference. This figure complements the results shown in Fig. 2, and suggests that the country-wide median relative bias of zero (Fig. 2c) masks underlying spatial patterns, with underestimations located near the coasts and overestimations mostly occurring farther inland.
- Figure A3 shows that temporal shifts in precipitation inputs are more harmful to the skill of hydrological models over small and medium catchments, highlighting the effect of catchment size, and likely the effect of longer response times. However, all models suffered with shifts of three days to the future regardless of the catchment size.
- Figure A4 shows the evolution of the parameters of GR4J with increasing intensity and persistence of random corruption. This figure highlights the reaction of GR4J to random corruption by increasing the buffering capacity of the production and transfer stores (increasing X_1 and X_3) and losing more water to groundwater by taking increasingly negative values for X_2.

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

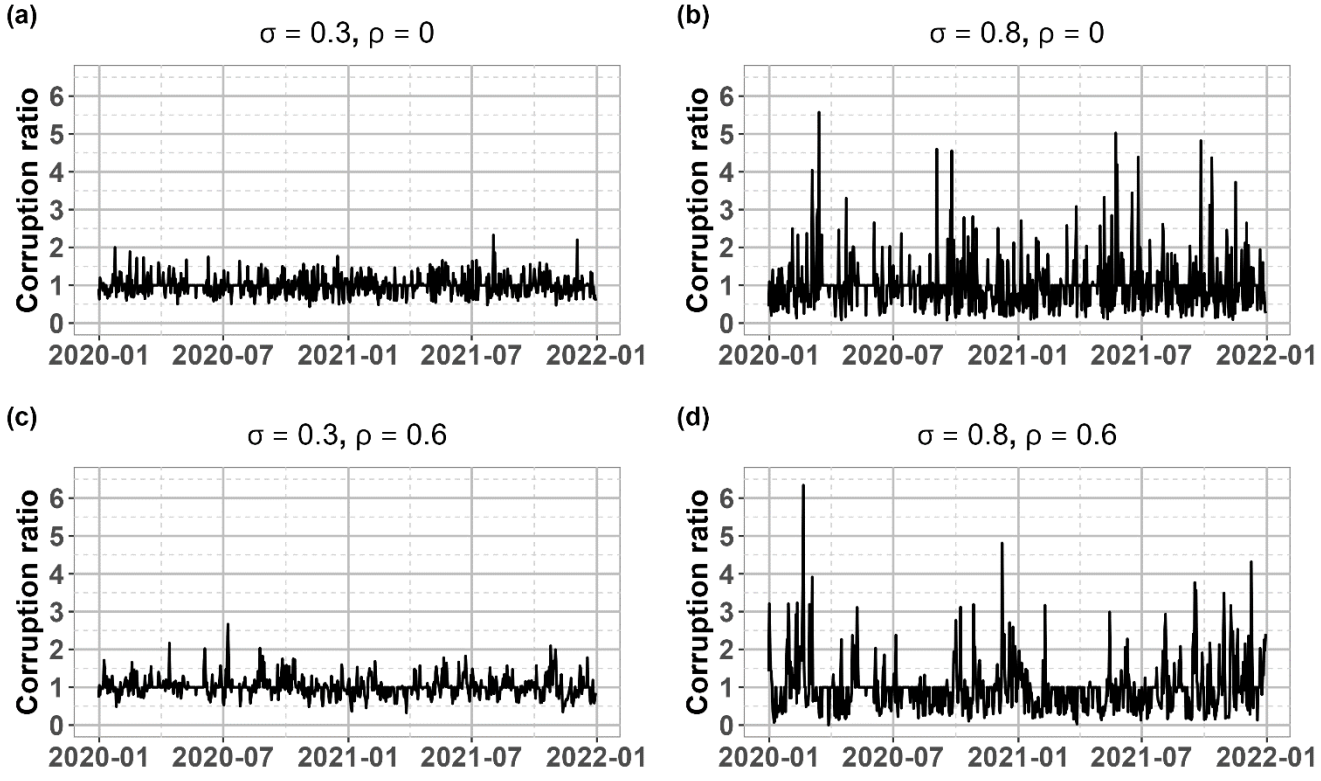


Figure A1: Illustration of the role of the intensity σ and the correlation ρ on the corruption ratio κ, i.e., the ratio between the corrupted and the original daily precipitation depth. Note that for dry days, the ratios are defined to be equal to 1.

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

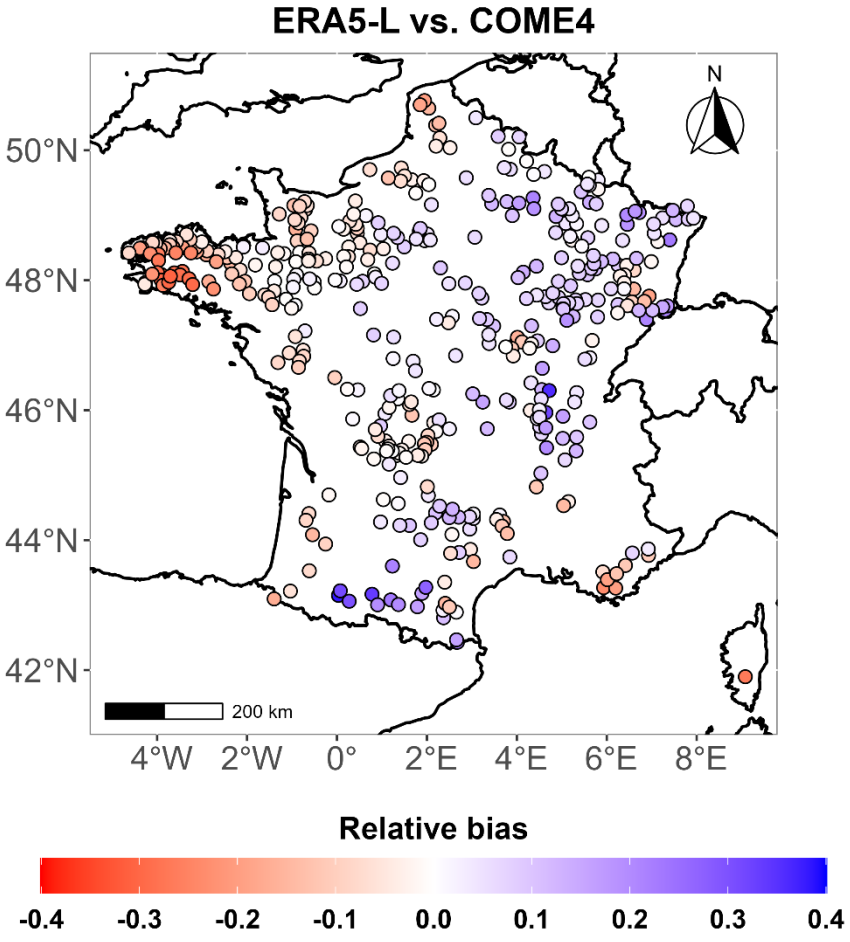
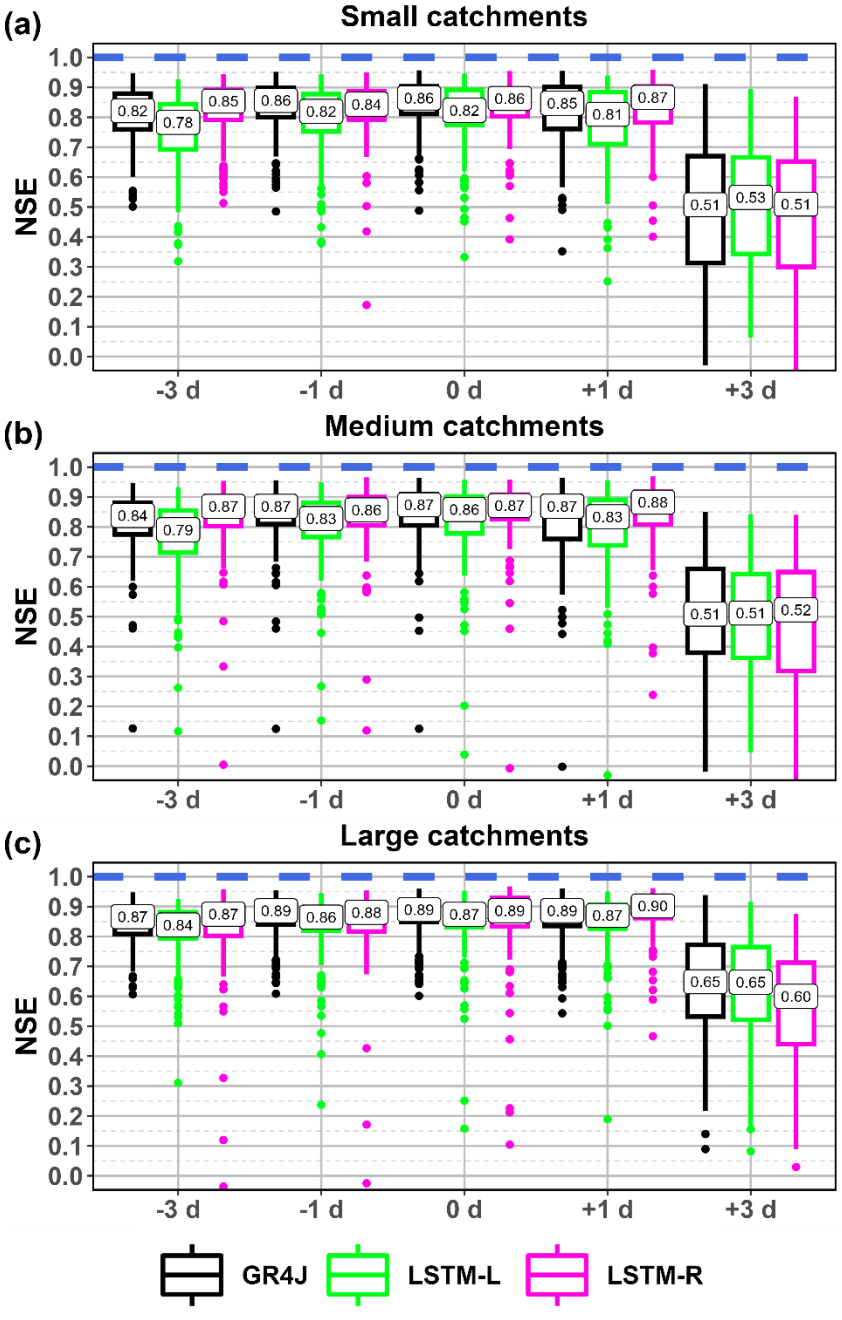


Figure A2: Spatial distribution of the relative bias in long-term precipitation amounts from ERA5-Land (ERA5-L) with precipitation amounts from COMEPHORE (COME-4) as a reference. Each dot corresponds to the location of the centroid of each catchment polygon.

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.



605

Figure A3: Test performances with Nash-Sutcliffe efficiency (NSE) as an evaluation metric for models trained and tested using shifted precipitation for (a) small catchments (<137 km^2), (b) medium catchments (between 137-380 km^2), and (c) large catchments (>380 km^2). The definition of small, medium, and large catchments was made using the 33rd and 67th percentiles of catchment area so as to have an equivalent number of catchments per each set.

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

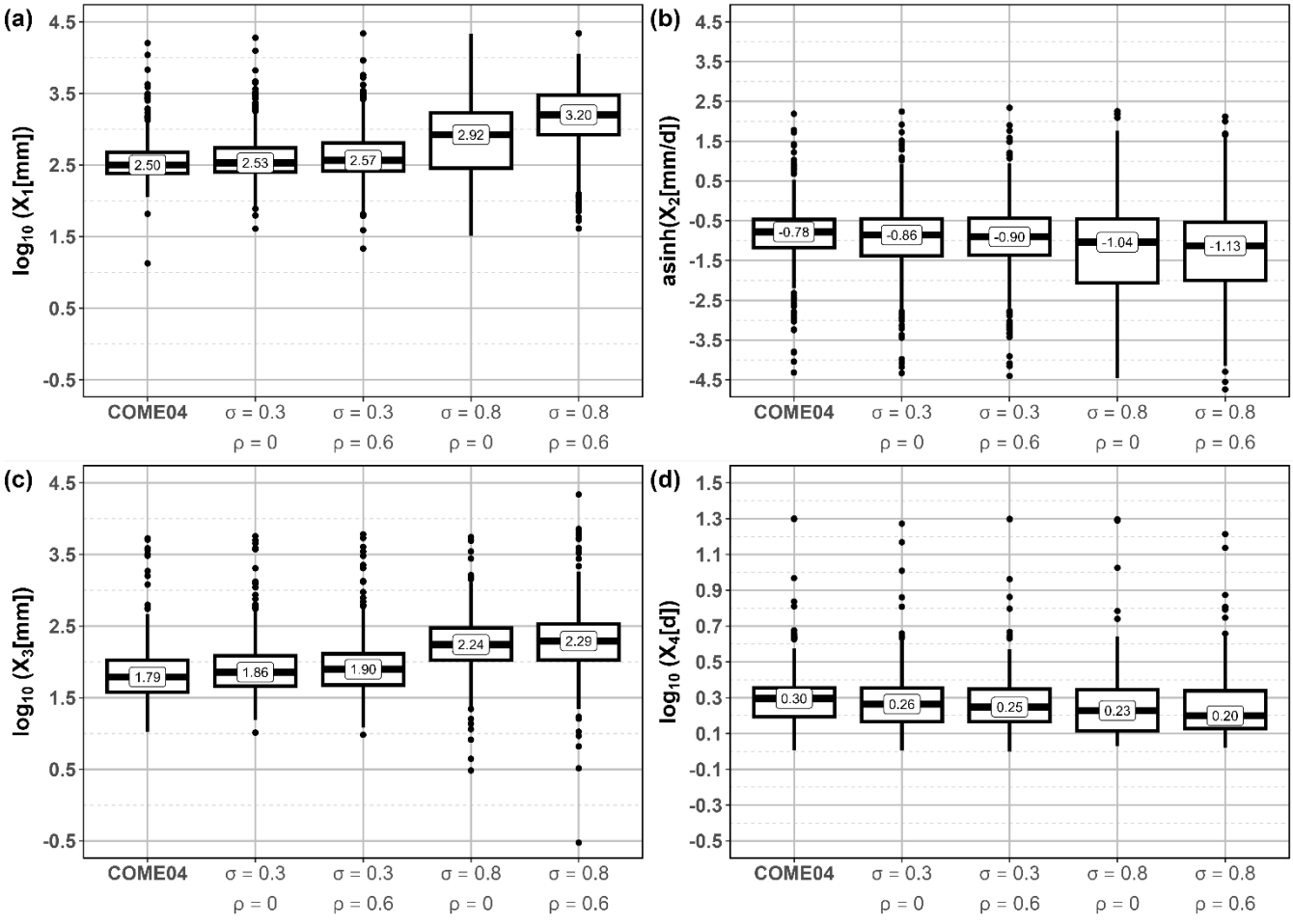


Figure A4: Evolution of GR4J parameters with increasing corruption of input precipitation. Note that all models including GR4J were retrained for each precipitation input.

Code and data availability

The CAMELS-FR dataset (Delaigue et al., 2025) is accessible at https://doi.org/10.57745/WH7FJR (last access: 19 October 2026). COMEPHORE can be downloaded from https://www.data.gouv.fr/datasets/reanalyses-comephore (last access: 19 October 2026). ERA5-Land can be downloaded from https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=download (last access: 19 October 2026). GR4J can be trained using the airGR R package (Coron et al., 2023), accessible at https://cran.r-project.org/web/packages/airGR/index.html (last access: 19 October 2026). Preprocessed time series for the catchment set, scripts used to train and test the LSTM models, as well as training and test results for GR4J, LSTM-L, and LSTM-R can be accessed at www.doi.org/10.5281/zenodo.23470931 (Saadi, 2026). Any questions regarding the code or the data are to be addressed to the corresponding author (mohamed.saadi@toulouse-inp.fr).

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

Author contributions

MS prepared the data, conceived the experiments with suggestions from LO, ran the experiments, and wrote the original draft. All co-authors discussed and provided suggestions regarding the results and contributed to the final version of the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

The first author (MS) would like to thank Patrick Le Moigne (Météo-France) for providing useful explanations regarding the SAFRAN product.

References

Abbas, A., Yang, Y., Pan, M., Tramblay, Y., Shen, C., Ji, H., Gebrechorkos, S. H., Pappenberger, F., Pyo, J., Feng, D., Huffman, G., Nguyen, P., Massari, C., Brocca, L., Tan, J., and Beck, H. E.: Comprehensive Global Assessment of 24 Gridded Precipitation Datasets Across 18 428 Catchments Using Hydrological Modeling, Hydrol. Earth Syst. Sci., 30, 3399–3423, https://doi.org/10.5194/hess-30-3399-2026, 2026.

Álvarez Chaves, M., Acuña Espinoza, E., Ehret, U., and Guthke, A.: When physics gets in the way: an entropy-based evaluation of conceptual constraints in hybrid hydrological models, Hydrology and Earth System Sciences, 30, 629–658, https://doi.org/10.5194/hess-30-629-2026, 2026.

Arnaud, P., Lavabre, J., Fouchier, C., Diss, S., and Javelle, P.: Sensitivity of hydrological models to uncertainty in rainfall input, Hydrological Sciences Journal, 56, 397–410, https://doi.org/10.1080/02626667.2011.563742, 2011.

Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F., and Mai, J.: Continuous streamflow prediction in ungauged basins: long short-term memory neural networks clearly outperform traditional hydrological models, Hydrology and Earth System Sciences, 27, 139–157, https://doi.org/10.5194/hess-27-139-2023, 2023.

Astagneau, P. C., Bourgin, F., Andréassian, V., and Perrin, C.: Catchment response to intense rainfall: Evaluating modelling hypotheses, Hydrological Processes, 36, e14676, https://doi.org/10.1002/hyp.14676, 2022.

Bandhauer, M., Isotta, F., Lakatos, M., Lussana, C., Båserud, L., Izsák, B., Szentes, O., Tveito, O. E., and Frei, C.: Evaluation of daily precipitation analyses in E-OBS (v19.0e) and ERA5 by comparison to regional high-resolution datasets in European regions, International Journal of Climatology, 42, 727–747, https://doi.org/10.1002/joc.7269, 2022.

Bárdossy, A. and Das, T.: Influence of rainfall observation network on model calibration and application, Hydrology and Earth System Sciences, 12, 77–89, https://doi.org/10.5194/hess-12-77-2008, 2008.

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.



Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A. I. J. M., Weedon, G. P., Brocca, L., Pappenberger, F., Huffman, G. J., and Wood, E. F.: Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling, Hydrology and Earth System Sciences, 21, 6201–6217, https://doi.org/10.5194/hess-21-6201-2017, 2017a.

Beck, H. E., van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., and de Roo, A.: MSWEP: 3-hourly 0.25° global gridded precipitation (1979-2015) by merging gauge, satellite, and reanalysis data, Hydrology and Earth System Sciences, 21, 589–615, https://doi.org/10.5194/hess-21-589-2017, 2017b.

Beck, H. E., Pan, M., Roy, T., Weedon, G. P., Pappenberger, F., van Dijk, A. I. J. M., Huffman, G. J., Adler, R. F., and Wood, E. F.: Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS, Hydrology and Earth System Sciences, 23, 207–224, https://doi.org/10.5194/hess-23-207-2019, 2019.

Behnke, R., Vavrus, S., Allstadt, A., Albright, T., Thogmartin, W. E., and Radeloff, V. C.: Evaluation of downscaled, gridded climate data for the conterminous United States, Ecological Applications, 26, 1338–1351, https://doi.org/10.1002/15-1061, 2016.

Berne, A. and Krajewski, W. F.: Radar for hydrology: Unfulfilled promise or unrecognized potential?, Advances in Water Resources, 51, 357–366, https://doi.org/10.1016/j.advwatres.2012.05.005, 2013.

Beven, K.: How to make advances in hydrological modelling, Hydrology Research, 50, 1481–1494, https://doi.org/10.2166/nh.2019.134, 2019.

Beven, K.: Deep learning, hydrological processes and the uniqueness of place, Hydrological Processes, 34, 3608–3613, https://doi.org/10.1002/hyp.13805, 2020.

Borga, M.: Accuracy of radar rainfall estimates for streamflow simulation, Journal of Hydrology, 267, 26–39, https://doi.org/10.1016/S0022-1694(02)00137-3, 2002.

Caillaud, C., Somot, S., Alias, A., Bernard-Bouissières, I., Fumière, Q., Laurantin, O., Seity, Y., and Ducrocq, V.: Modelling Mediterranean heavy precipitation events at climate scale: an object-oriented evaluation of the CNRM-AROME convection-permitting regional climate model, Clim Dyn, 56, 1717–1752, https://doi.org/10.1007/s00382-020-05558-y, 2021.

Clerc-Schwarzenbach, F., Selleri, G., Neri, M., Toth, E., van Meerveld, I., and Seibert, J.: Large-sample hydrology – a few camels or a whole caravan?, Hydrology and Earth System Sciences, 28, 4219–4237, https://doi.org/10.5194/hess-28-4219-2024, 2024.

Coron, L., Delaigue, O., Thirel, G., Dorchies, D., Perrin, C., and Michel, C.: airGR: Suite of GR Hydrological Models for Precipitation-Runoff Modelling, https://doi.org/10.15454/EX11NA, 2023.

Delaigue, 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 System Science Data, 17, 1461–1479, https://doi.org/10.5194/essd-17-1461-2025, 2025.

Delrieu, G., Braud, I., Berne, A., Borga, M., Boudevillain, B., Fabry, F., Freer, J., Gaume, E., Nakakita, E., Seed, A., Tabary, P., and Uijlenhoet, R.: Weather radar and hydrology, Advances in Water Resources, 32, 969–974, https://doi.org/10.1016/j.advwatres.2009.03.006, 2009.

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

Dembélé, M., Schaefli, B., van de Giesen, N., and Mariéthoz, G.: Suitability of 17 gridded rainfall and temperature datasets for large-scale hydrological modelling in West Africa, Hydrology and Earth System Sciences, 24, 5379–5406, https://doi.org/10.5194/hess-24-5379-2020, 2020.

Dumont, M., Saadi, M., Oudin, L., Lachassagne, P., Nugraha, B., Fadillah, A., Bonjour, J.-L., Muhammad, A., Hendarmawan, Dörfliger, N., and Plagnes, V.: Assessing rainfall global products reliability for water resource management in a tropical volcanic mountainous catchment, Journal of Hydrology: Regional Studies, 40, 101037, https://doi.org/10.1016/j.ejrh.2022.101037, 2022.

Edijatno, Nascimento, N. de O., Yang, X., Makhlouf, Z., and Michel, C.: GR3J: a daily watershed model with three free parameters, Hydrological Sciences Journal, 44, 263–277, https://doi.org/10.1080/02626669909492221, 1999.

Ehsani, M. R. and Behrangi, A.: A comparison of correction factors for the systematic gauge-measurement errors to improve the global land precipitation estimate, Journal of Hydrology, 610, 127884, https://doi.org/10.1016/j.jhydrol.2022.127884, 2022.

Emmanuel, I., Payrastre, O., Andrieu, H., and Zuber, F.: A method for assessing the influence of rainfall spatial variability on hydrograph modeling. First case study in the Cevennes Region, southern France, Journal of Hydrology, 555, 314–322, https://doi.org/10.1016/j.jhydrol.2017.10.011, 2017.

Essou, G. R. C., Arsenault, R., and Brissette, F. P.: Comparison of climate datasets for lumped hydrological modeling over the continental United States, Journal of Hydrology, 537, 334–345, https://doi.org/10.1016/j.jhydrol.2016.03.063, 2016.

Evin, G., Le Lay, M., Fouchier, C., Penot, D., Colleoni, F., Mas, A., Garambois, P.-A., and Laurantin, O.: Evaluation of hydrological models on small mountainous catchments: impact of the meteorological forcings, Hydrology and Earth System Sciences, 28, 261–281, https://doi.org/10.5194/hess-28-261-2024, 2024.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D.: The Shuttle Radar Topography Mission, Reviews of Geophysics, 45, https://doi.org/10.1029/2005RG000183, 2007.

Feng, D., Beck, H., de Bruijn, J., Sahu, R. K., Satoh, Y., Wada, Y., Liu, J., Pan, M., Lawson, K., and Shen, C.: Deep dive into hydrologic simulations at global scale: harnessing the power of deep learning and physics-informed differentiable models (δHBV-globe1.0-hydroDL), Geoscientific Model Development, 17, 7181–7198, https://doi.org/10.5194/gmd-17-7181-2024, 2024.

Frame, J. M., Kratzert, F., Gupta, H. V., Ullrich, P., and Nearing, G. S.: On strictly enforced mass conservation constraints for modelling the Rainfall-Runoff process, Hydrological Processes, 37, e14847, https://doi.org/10.1002/hyp.14847, 2023.

Gebrechorkos, S. H., Leyland, J., Dadson, S. J., Cohen, S., Slater, L., Wortmann, M., Ashworth, P. J., Bennett, G. L., Boothroyd, R., Cloke, H., Delorme, P., Griffith, H., Hardy, R., Hawker, L., McLelland, S., Neal, J., Nicholas, A., Tatem, A. J., Vahidi, E., Liu, Y., Sheffield, J., Parsons, D. R., and Darby, S. E.: Global-scale evaluation of precipitation datasets for hydrological modelling, Hydrology and Earth System Sciences, 28, 3099–3118, https://doi.org/10.5194/hess-28-3099-2024, 2024.

Gu, L., Yin, J., Wang, S., Chen, J., Qin, H., Yan, X., He, S., and Zhao, T.: How well do the multi-satellite and atmospheric reanalysis products perform in hydrological modelling, Journal of Hydrology, 617, 128920, https://doi.org/10.1016/j.jhydrol.2022.128920, 2023.

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

Hammoudeh, S., Goergen, K., Belleflamme, A., Giles, J. A., Trömel, S., and Kollet, S.: Evaluating precipitation products for water resources hydrologic modeling over Germany, Front. Earth Sci., 13, https://doi.org/10.3389/feart.2025.1548557, 2025.

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?, Hydrology and Earth System Sciences, 26, 5793–5816, https://doi.org/10.5194/hess-26-5793-2022, 2022.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hinge, G., Hamouda, M. A., Long, D., and Mohamed, M. M.: Hydrologic utility of satellite precipitation products in flood prediction: A meta-data analysis and lessons learnt, Journal of Hydrology, 612, 128103, https://doi.org/10.1016/j.jhydrol.2022.128103, 2022.

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

Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G. S., Hochreiter, S., and Klambauer, G.: MC-LSTM: Mass-Conserving LSTM, in: Proceedings of the 38th International Conference on Machine Learning, International Conference on Machine Learning, 4275–4286, 2021.

Huang, Y., Bárdossy, A., and Zhang, K.: Sensitivity of hydrological models to temporal and spatial resolutions of rainfall data, Hydrology and Earth System Sciences, 23, 2647–2663, https://doi.org/10.5194/hess-23-2647-2019, 2019.

Kavetski, D., Kuczera, G., and Franks, S. W.: Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, Water Resources Research, 42, https://doi.org/10.1029/2005WR004368, 2006.

Kidd, C. and Levizzani, V.: Status of satellite precipitation retrievals, Hydrology and Earth System Sciences, 15, 1109–1116, https://doi.org/10.5194/hess-15-1109-2011, 2011.

Kidd, C., Becker, A., Huffman, G. J., Muller, C. L., Joe, P., Skofronick-Jackson, G., and Kirschbaum, D. B.: So, How Much of the Earth's Surface Is Covered by Rain Gauges?, Bulletin of the American Meteorological Society, 98, 69–78, https://doi.org/10.1175/BAMS-D-14-00283.1, 2017.

Kim, S.-H., Hwang, J., Devineni, N., and Sankarasubramanian, A.: Do Deep-Learning Models Perform Better Than Process-Based Models? A Diagnostic Evaluation Using Synthetic Streamflow Simulation Under Various Sources of Uncertainties, Water Resources Research, 62, e2025WR040439, https://doi.org/10.1029/2025WR040439, 2026.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, https://doi.org/10.48550/arXiv.1412.6980, 2017.

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

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

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

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

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., and Matias, Y.: Caravan - A global community dataset for large-sample hydrology, Sci Data, 10, 61, https://doi.org/10.1038/s41597-023-01975-w, 2023.

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

Le Moigne, P.: Modélisation des Processus de Surface pour les Modèles de Prévision Numérique du Temps et de Climat, et Réanalyses de Surface pour les Etudes d'Impact en Hydrologie, Habilitation à diriger des recherches, Institut national polytechnique de Toulouse (INPT), https://doi.org/https://theses.hal.science/tel-04548056, 2024.

Lee, G. W. and Zawadzki, I.: Variability of Drop Size Distributions: Time-Scale Dependence of the Variability and Its Effects on Rain Estimation, Journal of Applied Meteorology and Climatology, 44, 241–255, https://doi.org/10.1175/JAM2183.1, 2005.

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

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Kumar Sahu, R., Greve, P., Slater, L., and Dadson, S. J.: Hydrological concept formation inside long short-term memory (LSTM) networks, Hydrology and Earth System Sciences, 26, 3079–3101, https://doi.org/10.5194/hess-26-3079-2022, 2022.

Li, P., Song, Y., Pan, M., Lawson, K., and Shen, C.: Ensembling differentiable process-based and data-driven models with diverse meteorological forcing datasets to advance streamflow simulation, Hydrology and Earth System Sciences, 29, 6829–6861, https://doi.org/10.5194/hess-29-6829-2025, 2025.

Li, X., Khandelwal, A., Jia, X., Cutler, K., Ghosh, R., Renganathan, A., Xu, S., Tayal, K., Nieber, J., Duffy, C., Steinbach, M., and Kumar, V.: Regionalization in a Global Hydrologic Deep Learning Model: From Physical Descriptors to Random Vectors, Water Resources Research, 58, e2021WR031794, https://doi.org/10.1029/2021WR031794, 2022.

Lin, Y. and Mitchell, K. E.: The NCEP stage II/IV hourly precipitation analyses: Development and applications, in: Proceedings of the 19th conference on Hydrology, 9–13, 2005.

Lobligeois, F., Andréassian, V., Perrin, C., Tabary, P., and Loumagne, C.: When does higher spatial resolution rainfall information improve streamflow simulation? An evaluation using 3620 flood events, Hydrology and Earth System Sciences, 18, 575–594, https://doi.org/10.5194/hess-18-575-2014, 2014.

Lopez, P.: Direct 4D-Var Assimilation of NCEP Stage IV Radar and Gauge Precipitation Data at ECMWF, Monthly Weather Review, 139, 2098–2116, https://doi.org/10.1175/2010MWR3565.1, 2011.

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

Loritz, R., Dolich, A., Acuña Espinoza, E., Ebeling, P., Guse, B., Götte, J., Hassler, S. K., Hauffe, C., Heidbüchel, I., Kiesel, J., Mälicke, M., Müller-Thomy, H., Stölzle, M., and Tarasova, L.: CAMELS-DE: hydro-meteorological time series and attributes for 1582 catchments in Germany, Earth System Science Data, 16, 5625–5642, https://doi.org/10.5194/essd-16-5625-2024, 2024.

Maggioni, V. and Massari, C.: On the performance of satellite precipitation products in riverine flood modeling: A review, Journal of Hydrology, 558, 214–224, https://doi.org/10.1016/j.jhydrol.2018.01.039, 2018.

Mai, J., Shen, H., Tolson, B. A., Gaborit, É., Arsenault, R., Craig, J. R., Fortin, V., Fry, L. M., Gauch, M., Klotz, D., Kratzert, F., O'Brien, N., Princz, D. G., Rasiya Koya, S., Roy, T., Seglenieks, F., Shrestha, N. K., Temgoua, A. G. T., Vionnet, V., and Waddell, J. W.: The Great Lakes Runoff Intercomparison Project Phase 4: the Great Lakes (GRIP-GL), Hydrology and Earth System Sciences, 26, 3537–3572, https://doi.org/10.5194/hess-26-3537-2022, 2022.

Marshall, J. S. and Palmer, W. M. K.: The distribution of raindrops with size, Journal of the Atmospheric Sciences, 5, 165–166, https://doi.org/10.1175/1520-0469(1948)005%3C0165:TDORWS%3E2.0.CO;2, 1948.

Martel, J.-L., Brissette, F., Arsenault, R., Turcotte, R., Castañeda-Gonzalez, M., Armstrong, W., Mailhot, E., Pelletier-Dumont, J., Rondeau-Genesse, G., and Caron, L.-P.: Assessing the adequacy of traditional hydrological models for climate change impact studies: a case for long short-term memory (LSTM) neural networks, Hydrology and Earth System Sciences, 29, 2811–2836, https://doi.org/10.5194/hess-29-2811-2025, 2025.

McMillan, H., Jackson, B., Clark, M., Kavetski, D., and Woods, R.: Rainfall uncertainty in hydrological modelling: An evaluation of multiplicative error models, Journal of Hydrology, 400, 83–94, https://doi.org/10.1016/j.jhydrol.2011.01.026, 2011.

Moulin, L., Gaume, E., and Obled, C.: Uncertainties on mean areal precipitation: assessment and impact on streamflow simulations, Hydrology and Earth System Sciences, 13, 99–114, https://doi.org/10.5194/hess-13-99-2009, 2009.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth System Science Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

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

Natel de Moura, C., Seibert, J., and Detzel, D. H. M.: Evaluating the long short-term memory (LSTM) network for discharge prediction under changing climate conditions, Hydrology Research, 53, 657–667, https://doi.org/10.2166/nh.2022.044, 2022.

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

Nelson, B. R., Prat, O. P., Seo, D.-J., and Habib, E.: Assessment and Implications of NCEP Stage IV Quantitative Precipitation Estimates for Product Intercomparisons, Weather and Forecasting, 31, 371–394, https://doi.org/10.1175/WAF-D-14-00112.1, 2016.

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

Ochoa-Rodriguez, S., Wang, L.-P., Willems, P., and Onof, C.: A Review of Radar-Rain Gauge Data Merging Methods and Their Potential for Urban Hydrological Applications, Water Resources Research, 55, 6356–6391, https://doi.org/10.1029/2018WR023332, 2019.

Ollivier, C. C., Carrière, S. D., Heath, T., Olioso, A., Rabefitia, Z., Rakoto, H., Oudin, L., and Satgé, F.: Ensemble precipitation estimates based on an assessment of 21 gridded precipitation datasets to improve precipitation estimations across Madagascar, Journal of Hydrology: Regional Studies, 47, 101400, https://doi.org/10.1016/j.ejrh.2023.101400, 2023.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., and Loumagne, C.: Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling, Journal of Hydrology, 303, 290–306, https://doi.org/10.1016/j.jhydrol.2004.08.026, 2005.

Oudin, L., Perrin, C., Mathevet, T., Andréassian, V., and Michel, C.: Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models, Journal of Hydrology, 320, 62–83, https://doi.org/10.1016/j.jhydrol.2005.07.016, 2006.

Overeem, A., Holleman, I., and Buishand, A.: Derivation of a 10-Year Radar-Based Climatology of Rainfall, Journal of Applied Meteorology and Climatology, 48, 1448–1463, https://doi.org/10.1175/2009JAMC1954.1, 2009.

Overeem, A., van den Besselaar, E., van der Schrier, G., Meirink, J. F., van der Plas, E., and Leijnse, H.: EURADCLIM: the European climatological high-resolution gauge-adjusted radar precipitation dataset, Earth System Science Data, 15, 1441–1464, https://doi.org/10.5194/essd-15-1441-2023, 2023.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, https://doi.org/10.48550/arXiv.1912.01703, 2019.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275–289, https://doi.org/10.1016/S0022-1694(03)00225-7, 2003.

Prein, A. F., Pothapakula, P. K., Zeman, C., Lalonde, M., Rixen, M., Dipankar, A., Leclair, M., and Jocksch, A.: From single storms to large-scale waves: a multi-year kilometer-scale global simulation, Geoscientific Model Development, 19, 5277–5303, https://doi.org/10.5194/gmd-19-5277-2026, 2026.

Puche, M., Troin, M., and Fox, D. M.: Assessing temporal and spatial generalization of LSTMs for streamflow modeling in French watersheds with and without European training data, Journal of Hydrology: Regional Studies, 62, 103022, https://doi.org/10.1016/j.ejrh.2025.103022, 2025.

Quintana-Seguí, P., Moigne, P. L., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchisteguy, L., and Morel, S.: Analysis of Near-Surface Atmospheric Variables: Validation of the SAFRAN Analysis over France, Journal of Applied Meteorology and Climatology, 47, 92–107, https://doi.org/10.1175/2007JAMC1636.1, 2008.

Raimonet, M., Oudin, L., Thieu, V., Silvestre, M., Vautard, R., Rabouille, C., and Moigne, P. L.: Evaluation of Gridded Meteorological Datasets for Hydrological Modeling, Journal of Hydrometeorology, 18, 3027–3041, https://doi.org/10.1175/JHM-D-17-0018.1, 2017.

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

Reichert, P., Ma, K., Höge, M., Fenicia, F., Baity-Jesi, M., Feng, D., and Shen, C.: Metamorphic testing of machine learning and conceptual hydrologic models, Hydrology and Earth System Sciences, 28, 2505–2529, https://doi.org/10.5194/hess-28-2505-2024, 2024.

Roux, H., Amengual, A., Romero, R., Bladé, E., and Sanz-Ramos, M.: Evaluation of two hydrometeorological ensemble strategies for flash-flood forecasting over a catchment of the eastern Pyrenees, Natural Hazards and Earth System Sciences, 20, 425–450, https://doi.org/10.5194/nhess-20-425-2020, 2020.

Saadi, M.: Scripts and data for the paper "Can we rely on deep-learning hydrological models to assess the quality of precipitation products?," https://doi.org/10.5281/zenodo.23470931, 2026.

Saadi, M., Oudin, L., and Ribstein, P.: Physically consistent conceptual rainfall–runoff model for urbanized catchments, Journal of Hydrology, 599, 126394, https://doi.org/10.1016/j.jhydrol.2021.126394, 2021.

Saadi, M., Furusho-Percot, C., Belleflamme, A., Chen, J.-Y., Trömel, S., and Kollet, S.: How uncertain are precipitation and peak flow estimates for the July 2021 flooding event?, Natural Hazards and Earth System Sciences, 23, 159–177, https://doi.org/10.5194/nhess-23-159-2023, 2023.

Satgé, F., Defrance, D., Sultan, B., Bonnet, M.-P., Seyler, F., Rouché, N., Pierron, F., and Paturel, J.-E.: Evaluation of 23 gridded precipitation datasets across West Africa, Journal of Hydrology, 581, 124412, https://doi.org/10.1016/j.jhydrol.2019.124412, 2020.

Sawadekar, K., Song, Y., Pan, M., Beck, H., McCrary, R., Ullrich, P., Lawson, K., and Shen, C.: Improving differentiable hydrologic modeling with interpretable forcing fusion, Journal of Hydrology, 659, 133320, https://doi.org/10.1016/j.jhydrol.2025.133320, 2025.

Song, Y., Sawadekar, K., Frame, J. M., Pan, M., Clark, M. P., Knoben, W. J. M., Wood, A. W., Lawson, K. E., Patel, T., and Shen, C.: Physics-Informed, Differentiable Hydrologic Models for Capturing Unseen Extreme Events, Water Resources Research, 62, e2025WR040414, https://doi.org/10.1029/2025WR040414, 2026.

Sperna Weiland, F. C., Vrugt, J. A., van Beek, R. (L.) P. H., Weerts, A. H., and Bierkens, M. F. P.: Significant uncertainty in global scale hydrological modeling from precipitation data errors, Journal of Hydrology, 529, 1095–1115, https://doi.org/10.1016/j.jhydrol.2015.08.061, 2015.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research, 15, 1929–1958, 2014.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., and Hsu, K.-L.: A Review of Global Precipitation Data Sets: Data Sources, Estimation, and Intercomparisons, Reviews of Geophysics, 56, 79–107, https://doi.org/10.1002/2017RG000574, 2018.

Tabary, P., Dupuy, P., L'Henaff, G., Gueguen, C., Moulin, L., Laurantin, O., Merlier, C., and Soubeyroux, J.-M.: A 10-year (1997–2006) reanalysis of Quantitative Precipitation Estimation over France: methodology and first results, IAHS Publication, 351, 255–260, 2012.

Tarek, M., Brissette, F. P., and Arsenault, R.: Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America, Hydrology and Earth System Sciences, 24, 2527–2544, https://doi.org/10.5194/hess-24-2527-2020, 2020.

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

Thébault, C., Perrin, C., Andréassian, V., Thirel, G., Legrand, S., and Delaigue, O.: Multi-model approach in a variable spatial framework for streamflow simulation, Hydrology and Earth System Sciences, 28, 1539–1566, https://doi.org/10.5194/hess-28-1539-2024, 2024.

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., and Shen, C.: From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling, Nat Commun, 12, 5988, https://doi.org/10.1038/s41467-021-26107-z, 2021.

Van Beusekom, A. E., Hay, L. E., Bennett, A. R., Choi, Y.-D., Clark, M. P., Goodall, J. L., Li, Z., Maghami, I., Nijssen, B., and Wood, A. W.: Hydrologic Model Sensitivity to Temporal Aggregation of Meteorological Forcing Data: A Case Study for the Contiguous United States, Journal of Hydrometeorology, 23, 167–183, https://doi.org/10.1175/JHM-D-21-0111.1, 2022.

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., and Soubeyroux, J.-M.: A 50-year high-resolution atmospheric reanalysis over France with the Safran system, International Journal of Climatology, 30, 1627–1644, https://doi.org/10.1002/joc.2003, 2010.

Vigoureux, S., Brigode, P., Ramos, M.-H., Poggio, J., Dreyfus, R., Moreau, E., Laroche, C., and Tric, E.: Spatio-temporal characteristics of heavy precipitation events observed over the last decade on the eastern French Mediterranean coastal area, Journal of Hydrology: Regional Studies, 56, 101988, https://doi.org/10.1016/j.ejrh.2024.101988, 2024.

Villarini, G., Mandapaka, P. V., Krajewski, W. F., and Moore, R. J.: Rainfall and sampling uncertainties: A rain gauge perspective, Journal of Geophysical Research: Atmospheres, 113, https://doi.org/10.1029/2007JD009214, 2008.

Wang, J., Zhuo, L., Han, D., Liu, Y., and Rico-Ramirez, M. A.: Hydrological Model Adaptability to Rainfall Inputs of Varied Quality, Water Resources Research, 59, e2022WR032484, https://doi.org/10.1029/2022WR032484, 2023.

Wi, S. and Steinschneider, S.: On the need for physical constraints in deep learning rainfall–runoff projections under climate change: a sensitivity analysis to warming and shifts in potential evapotranspiration, Hydrology and Earth System Sciences, 28, 479–503, https://doi.org/10.5194/hess-28-479-2024, 2024.

Wilbrand, K., Taormina, R., ten Veldhuis, M.-C., Visser, M., Hrachowitz, M., Nuttall, J., and Dahm, R.: Predicting streamflow with LSTM networks using global datasets, Front. Water, 5, https://doi.org/10.3389/frwa.2023.1166124, 2023.

Winterrath, T., Brendel, C., Hafer, M., Junghänel, T., Klameth, A., Lengfeld, K., Walawender, E., Weigl, E., and Becker, A.: Radar climatology (RADKLIM) version 2017.002; gridded precipitation data for Germany: Radar-based gauge-adjusted one-hour precipitation sum (RW), 2018.

Wüest, M., Frei, C., Altenhoff, A., Hagen, M., Litschi, M., and Schär, C.: A gridded hourly precipitation dataset for Switzerland using rain-gauge analysis and radar-based disaggregation, International Journal of Climatology, 30, 1764–1775, https://doi.org/10.1002/joc.2025, 2010.

Xie, P. and Xiong, A.-Y.: A conceptual model for constructing high-resolution gauge-satellite merged precipitation analyses, Journal of Geophysical Research: Atmospheres, 116, https://doi.org/10.1029/2011JD016118, 2011.

Yokoo, K., Ishida, K., Ercan, A., Tu, T., Nagasato, T., Kiyama, M., and Amagasaki, M.: Capabilities of deep learning models on learning physical relationships: Case of rainfall-runoff modeling with LSTM, Science of The Total Environment, 802, 149876, https://doi.org/10.1016/j.scitotenv.2021.149876, 2022.

https://doi.org/10.5194/egusphere-2026-6443
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.



940 Yu, J., Li, X.-F., Lewis, E., Blenkinsop, S., and Fowler, H. J.: UKGrsHP: a UK high-resolution gauge–radar–satellite merged hourly precipitation analysis dataset, Clim Dyn, 54, 2919–2940, https://doi.org/10.1007/s00382-020-05144-2, 2020.

Zhang, Y., Wang, Q. J., Frost, A. J., Pudashine, J., Trewin, B., Velasco-Forero, C., Su, C.-H., and Villani, V.: A localisation strategy for scalable and efficient multi-source precipitation blending at high resolutions over large domains, Journal of Hydrology, 677, 135831, https://doi.org/10.1016/j.jhydrol.2026.135831, 2026.

945