https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

A climate similarity-based transfer learning framework using global Caravan dataset for enhancing streamflow prediction in Ouémé River Basin (Benin, West Africa)

Jérôme Enagnon Ahouandjinou[1,2], Aymar Yaovi Bossa[1], Jean Hounkpe[1], Riccardo Taormina[3]

[1]National Water Institute (INE), University of Abomey-Calavi, Abomey-Calavi, Benin
[2]International Chair in Mathematical Physics and Applications (ICMPA-UNESCO Chair), Université of Abomey-Calavi, Abomey-Calavi, Benin
[3]Department of Water Management, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, the Netherlands
Correspondence to: Jérôme Enagnon Ahouandjinou (jeromeahouandj@gmail.com)

Abstract. Reliable streamflow forecasting is a fundamental component of flood risk management. However, the accuracy of such forecasts in data-scarce river basins remains one of the pressing challenges in most African catchments. While Long Short-Term Memory (LSTM) networks have demonstrated their effectiveness in rainfall-runoff modelling, their data-intensive requirements limit their direct applicability in poorly gauged catchments across sub-Saharan Africa. To address this issue, we propose a climate similarity-based transfer learning framework using the global Caravan hydrological dataset to improve local streamflow forecasting in Ouémé River Basin (ORB). Specifically, tropical-climate catchments are first identified from Caravan global dataset using the Köppen-Geiger climate classification. A composite climate similarity index (CI) is constructed from two normalized climate indices: annual mean precipitation and seasonality index. 200 basins were selected and ranked by CI. From the 200 ranked basins, a fixed subset of 50 basins was selected using stratified sampling across CI quartiles, including 30 basins for validation and 20 basins for independent testing. The remaining 150 basins were then used to define three cumulative pre-training subsets of increasing size (50, 100, and 150 basins), ranked by CI. Different LSTM models are pre-trained on each subset and fine-tuned on the ORB streamflow. The developed transfer learning models are evaluated against the LSTM model trained exclusively on local data as a baseline model. Pre-training improves prediction at all five stations, raising the median Kling–Gupta Efficiency from 0.65 for the local model to 0.75 for the best transfer configuration. The smallest, most climatically similar donor subset gives the best fine-tuning performance; adding less similar catchments does not help and slightly degrades it. The benefit is largest at stations with short or event-poor records and marginal at the well-gauged basin outlet, while peak flows remain underestimated across all configurations. These results show that, for data-scarce tropical basins, the climatic similarity of the transfer basins matters more than their number, and they point to similarity-guided donor selection combined with peak-weighted or hybrid formulations as a practical route to operational flood forecasting.

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

1 Introduction

West Africa is one of the regions where the consequences of climate change on water resources are most visible and most damaging. The Sahel and the Sudano-Guinean zone have experienced a rapid rise in flood frequency and magnitude since the late 1990s, and recent projections under CMIP6 scenarios indicate that flood magnitudes could increase at more than 90 % of West African regions by the end of the 21st century, with some catchments exceeding +45 % (Klutse et al., 2021; Quenum et al., 2021; Diop et al., 2025). In Benin, the Ouémé River Basin (ORB) drains nearly half of the national territory and supports water supply for several hundred thousand people in their activities (agriculture, fishing, livestock farming, and tourism). The catchment has suffered repeated severe floods, including the 2010 and 2020 events that displaced tens of thousands of inhabitants and damaged transport infrastructure across the lower delta (Hounkpè et al., 2016; Amoussou et al., 2024). Against this background, flood early warning systems (FEWS) have emerged as a key tool for reducing the social and economic losses associated with extreme rainfall events. Reliable hydrological forecasting lies at the heart of any effective FEWS, providing the basis for hazard assessment, risk communication, and emergency response planning in a region where adaptive capacity remains limited (Silva et al., 2025; Sorini et al., 2026). Among the various aspects of forecast quality, the accurate prediction of peak discharge values and their timing is particularly important for timely emergency management. Equally important, forecast reliability is directly linked to public trust and compliance with evacuation recommendations (Kreibich et al., 2021).

In the Ouémé River Basin, streamflow prediction has traditionally relied on global conceptual hydrological models (Kodja et al., 2018; Bossa et al., 2024; M. Bodjrènou et al., 2025) and physics-based semi-distributed models (Biao et al., 2016; Hounkpè et al., 2019; Houngue, 2020; R. Bodjrènou et al., 2023), which represent the rainfall–runoff transformation through a small number of physically meaningful parameters calibrated on observed records. These models remain valuable because their structure is interpretable and their parameters are physically constrained. Their predictive accuracy is bounded, however, by the simplifications they make about soil, vegetation, and routing processes; these simplifications produce systematic errors that are particularly visible during high-flow events (Janjić and Tadić, 2023). Moreover, traditional hydrological physics-based models are structurally complex and depend on the careful calibration of a large number of parameters, as well as extensive input data and considerable computational resources in calibration, validation, and prediction. These requirements restrict their practical applicability for real-time flood forecasting.

Over the last decade, deep learning models and in particular Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) have emerged as a competitive alternative for rainfall–runoff modelling. LSTMs learn the rainfall–runoff relation directly from data and can capture nonlinear, time-dependent dynamics that conceptual models tend to smooth out (Frame et al., 2022; Li et al., 2022; Man et al., 2023; Kratzert et al., 2024). When trained on large regional datasets in temperate basins, these models match or exceed the skill of calibrated hydrological models for daily streamflow simulation (Arsenault et al., 2023; Kratzert et al., 2024; Martel et al., 2025). Although deep learning models have demonstrated strong predictive capabilities in hydrology, several limitations are widely acknowledged. In particular, such models tend to underperform when reproducing extreme events (Baste et al., 2025; Song et al., 2025), offer limited interpretability (Choubin et al., 2025), and are highly sensitive to the availability and quality of observational data (Mohammadi et al., 2024; Fathi et al., 2025). This last constraint is especially critical in the African context. Most of the advances reported above rely on the dense, long-term hydroclimatic datasets available in these regions. In contrast, the majority of African basins remain ungauged or poorly monitored, which limits the direct application of LSTM models and points toward transfer learning as a promising alternative.

Transfer learning offers a way around this constraint. The idea is to first train an LSTM on a large, data-rich set of donor catchments, then fine-tune the resulting model on the limited target data. Recent global hydrological datasets such as Caravan (Kratzert et al., 2023), which compiles consistent daily forcings, static attributes, and streamflow records for several thousand basins across continents, have made this strategy practical. Several studies have demonstrated its value. Alzhanov et al. (2025) applied a Caravan-based transfer protocol to the Uba River in Kazakhstan and reported substantial gains over a locally trained LSTM. Yang et al. (2023) introduced a climate-adaptive transfer learning framework for soil moisture estimation in the Qinghai-Tibet Plateau, using Köppen–Geiger classification to filter donors. Ougahi et al. (2026) clustered global basins by hydroclimatic attributes and showed that fine-tuning from the best-matched cluster outperforms a model trained on the full pool. Across these contributions, one conclusion is consistent: pre-training on a global pool helps in data-scarce regions, but the donor set must be assembled with care.

The selection of donor catchments is, in fact, a key methodological question of similarity-based transfer learning. The CAMELS-derived regional LSTMs of Kratzert et al. (2019) and the HYSETS-derived regional LSTM (Arsenault et al., 2023) train on all available basins without any climatic filter, assuming that more diverse training data always improves generalisation. Recent results challenge this assumption. Xu et al. (2023) showed that source catchments selected by similarity in flow regime characteristics outperform an unfiltered global pool for streamflow prediction in data-scarce Chinese catchments. Ougahi et al. (2026) reached a similar conclusion for Central Asia, where indiscriminate inclusion of donors led to negative transfer that is, a fine-tuned model performing worse than a locally trained one. The case for similarity-guided donor selection is therefore growing, but it has not been tested on West African catchments, where the climate regime is markedly different from the temperate and alpine basins that dominate global hydrological datasets. The application of Caravan-based transfer learning to hydrological prediction in West African river basins remains largely unexplored, and the recent work of (Ahouandjinou et al., 2026) on the Ouémé at Bonou explicitly identifies this as a research priority for the region.

This study addresses three gaps in the current literature. First, it presents the first application of Caravan-based transfer learning to the Ouémé River Basin (ORB) and a systematic evaluation of similarity-based donor selection under a tropical climate, a setting that remains largely unexplored. Second, it introduces a composite climate similarity index (CI) built on mean annual precipitation and the seasonality index, with weights calibrated empirically against the high-flow signature of the Ouémé at Bonou, so that donor selection reflects the hydrological behaviour most relevant to flood prediction. Third, it quantifies the effect of donor corpus size on fine-tuning performance across five Ouémé stations, by comparing three

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

cumulative subsets of climatically similar donors (50, 100, and 150 catchments) against a local-only baseline. Taken together, these contributions test the hypothesis that selecting climatically similar donor catchments from a global dataset, rather than relying on local data alone, can substantially improve LSTM streamflow predictions in a poorly gauged tropical basin while clarifying how many donor catchments are needed to achieve this gain.

2. Methods

2.1 Study area.

The study was carried out in the Ouémé River basin (ORB), the largest hydrographic system of the Republic of Benin, located in sub-Saharan West Africa between latitudes 6°25′ N and 12°30′ N and longitudes 1°00′ E and 3°40′ E (Figure 1). Benin is bordered to the north by Niger, to the north-west by Burkina Faso, to the west by Togo, to the east by Nigeria, and to the south by the Gulf of Guinea, and lies entirely within the tropical climatic belt of West Africa. The Ouémé River rises in the Atacora massif of north-western Benin, near the Tanéka classified forest, and flows southward over approximately 510 km before discharging into the Nokoué and Porto-Novo lagoons and ultimately into the Atlantic Ocean (Hounkpè et al., 2019; Bossa et al., 2024; M. Bodjrènou et al., 2025). Its two main tributaries are the Okpara on the eastern flank and the Zou on the western flank. At the Bonou gauge the most downstream station before the delta the basin drains an area of about 49,256 km², which represents close to 41 % of Benin's national territory and the bulk of the country's surface water resources.

The Ouémé basin spans three contrasting climatic zones, all classified as tropical under the Köppen-Geiger system but differing markedly in their rainfall regimes (Odusanya et al., 2021). The northern Ouémé experiences a unimodal regime, with a single rainy season extending from May to October and a pronounced dry season from November to March. The central Ouémé exhibits a transitional regime, characterized by a longer wet period from March to October that is occasionally interrupted by a brief dry spell in August. The southern Ouémé follows a bimodal regime typical of the Guinea coast, with two wet seasons (a long one from March to July and a short one from September to mid-November) separated by two dry seasons (a long one from November to March and a short one in August) (Hounkpè et al., 2019).

This climatic and physiographic heterogeneity makes the Ouémé an informative test bed for evaluating transfer-learning strategies, as it captures within a single hydrographic system the diversity of hydrological regimes that characterize much of poorly gauged tropical West Africa. Furthermore, the basin is recognized as one of the most flood-prone in the region, with several catastrophic events recorded during the late 2000s and early 2010s (Hounkpè et al., 2016), which gives the development of robust streamflow prediction tools both a scientific and an operational rationale.

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

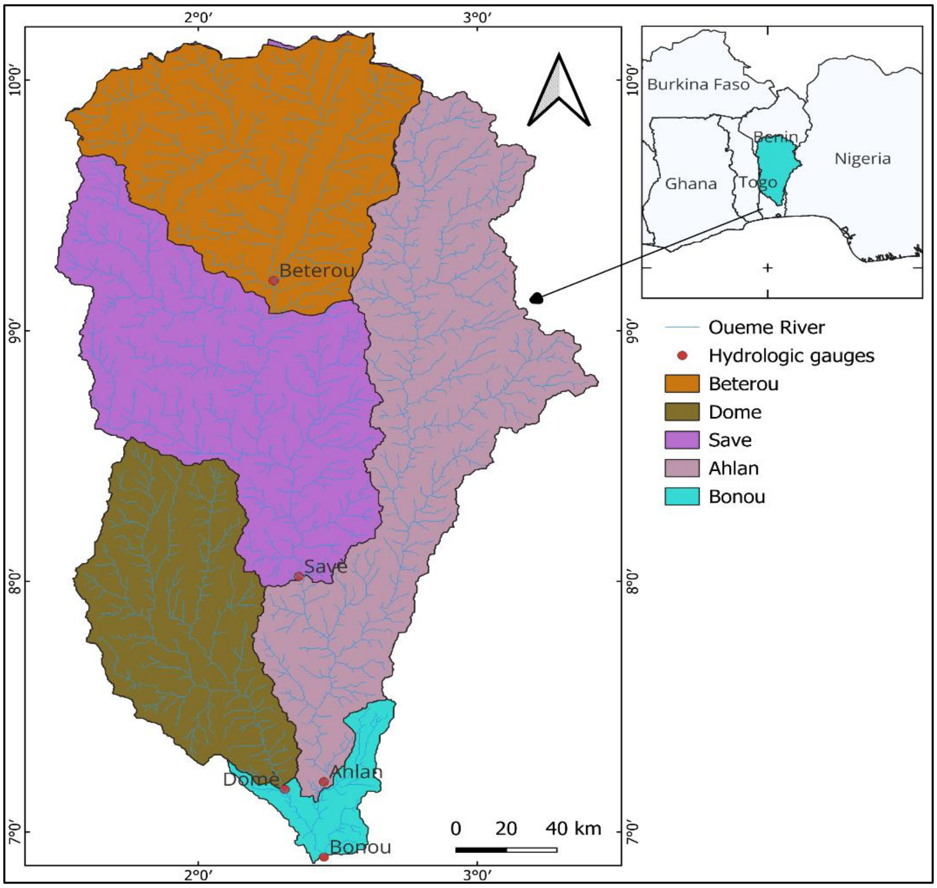

Figure 1: Study area

2.2 Dataset used

2.2.1 Caravan dataset

This study uses the Caravan dataset (Kratzert et al., 2023b), a globally standardized large-sample hydrology dataset specifically developed to support comparative hydrological analyses and deep learning applications. Caravan was introduced to overcome the lack of harmonization among existing regional CAMELS datasets and to provide a unified framework for global rainfall–runoff modeling. The dataset aggregates and standardizes hydrometeorological information from several regional large-sample hydrology datasets, including CAMELS-US, CAMELS-AUS, CAMELS-BR, CAMELS-CL, CAMELS-GB, HYSETS, and LamaH-CE. Caravan dataset covers the 1981-2020 and provides three main categories of information: (i) daily meteorological forcing data, (ii) observed streamflow time series, and (iii) static catchment attributes describing physiographic, climatic, hydrological, geological, and land-surface characteristics. The meteorological forcings are primarily derived from the ERA5-Land reanalysis product, while streamflow observations are collected from national and regional hydrological monitoring agencies. The dataset has been widely used for benchmarking LSTM-based rainfall–

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

runoff models, transfer learning approaches, flood forecasting systems, and global hydrological regionalization analyses (Alzhanov et al., 2025; Ougahi et al., 2026). Its large spatial diversity and rich set of static catchment descriptors make it particularly appropriate for investigating hydrological similarity and improving prediction in data-scarce basins.

In this study, we used precipitation, air temperature, evapotranspiration, and streamflow discharge as hydroclimate forcing data. We selected 23 statics catchments attributes from the Caravan dataset. These attributes include long-term precipitation and evapotranspiration metrics, runoff conditions, aridity and moisture indices, precipitation seasonality and extremes, groundwater and soil properties, elevation, slope, and basin area.

2.2.2 Local dataset

The local dataset of the Ouémé River Basin (ORB) is built following the Caravan methodological framework (Kratzert et al., 2023) to ensure structural compatibility between the local and donor data for transfer learning. The extraction of the gridded meteorological forcings and catchment attributes from global sources is the most computationally demanding step of this workflow, and it is carried out here through Google Earth Engine (GEE), a cloud-based platform that hosts a broad collection of geospatial datasets and makes them accessible through a Python application programming interface (API). The shapefiles of the five gauged sub-catchments of the ORB are passed to a GEE Python routine that returns, for each catchment, the static descriptors derived from HydroATLAS (Linke et al., 2019)and the daily meteorological forcings from ERA5-Land (Muñoz-Sabater et al., 2021). Observed streamflow data are collected from the in-situ records of the Direction Générale de l'Eau (DG-Eau) of Benin at the five stations of the basin: Beterou, Savè, Bonou, Domè, and Ahlan. The records cover the period 1996-2020 and consist of daily discharge values in cubic meters per second (m^3/s). To match the unit convention of the Caravan streamflow variable, discharge is converted to millimeters per day (mm/day) as follows:

$$Q_{mm/day} = \frac{Q_{m^3/s} \times 86400}{S \times 1000}, \tag{1}$$

where $Q_{m^3/s}$ is the observed discharge in cubic meters per second and S is the drainage area of the station in square meters. The discharge record was split into three non-overlapping periods: 1996-2012 for training, 2013-2015 for validation, and 2016–2020 for testing.

2.3 Caravan watersheds selection

Source domain selection is a critical step in the proposed framework, as the choice of source catchments can strongly influence transfer learning performance in the target basin and may even lead to negative transfer effects. When all catchments from a large-scale hydrological dataset are indiscriminately used, it becomes difficult to determine which basins contribute positively or negatively to model generalization. Given that climate is recognized as a key driver of streamflow dynamics, particularly through the strong linkage between precipitation and river discharge (Berg et al., 2018), climate

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

similarity was adopted as the guiding criterion for source domain selection. More specifically, tropical-climate catchments were extracted from the Caravan dataset using the Köppen–Geiger climate classification developed by Beck et al. (2018) and updated by Rossi et al. (2026), as the Ouémé River Basin (ORB) is situated within a tropical climate region.

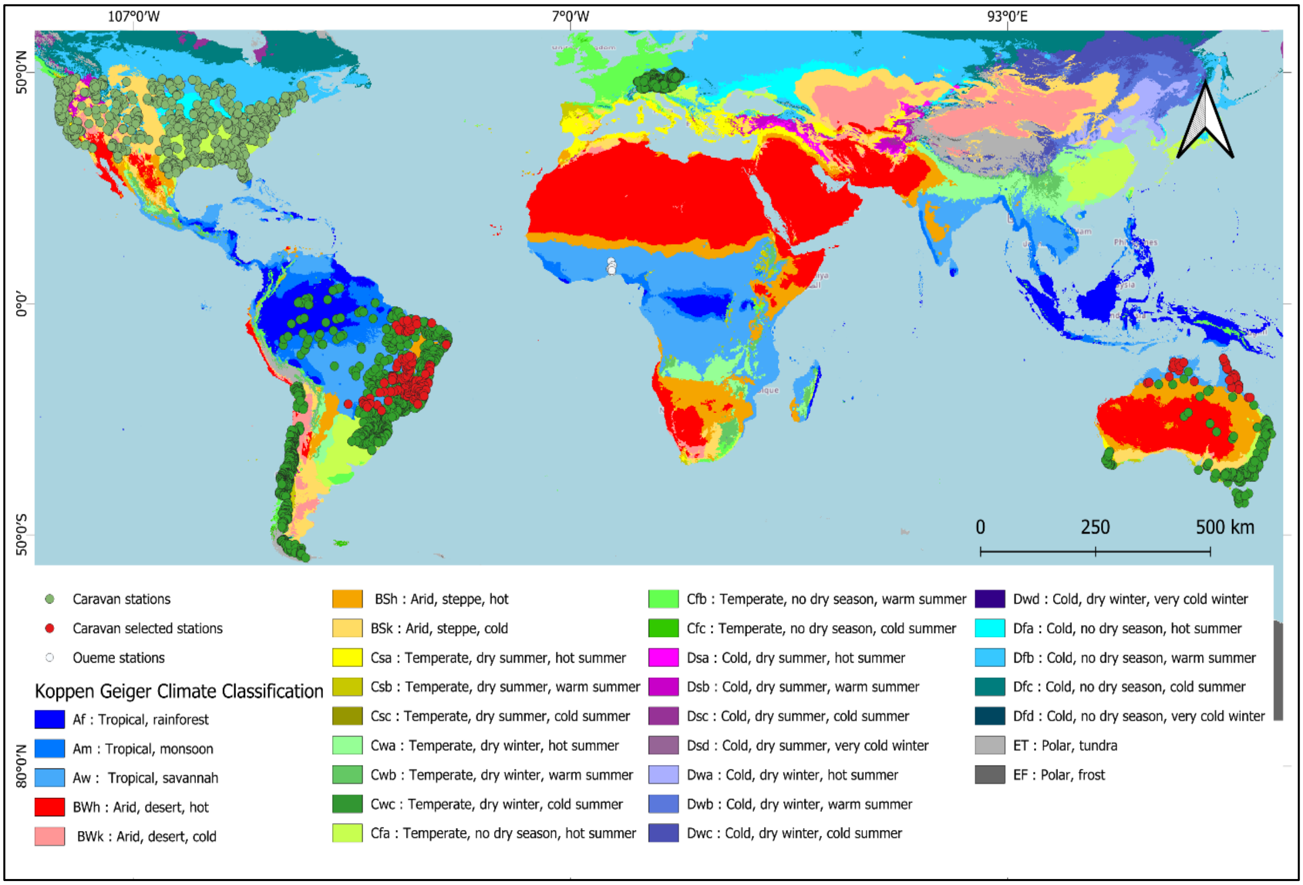

Figure 2: Global distribution of the selected basins in the Caravan dataset

A total of 200 catchments were selected in tropical climate domain in the global Caravan dataset (Figure 2). Secondly, the seasonality index (SI) and annual mean precipitation (P) were calculated for each selected catchment and normalized with respect to the ORB climate index, yielding the normalized indices ΔSI and ΔP, respectively. The Spearman correlations were computed between each normalized climate index and the annual 95th-percentile flow (FHV) observed in ORB at Bonou outlet, so that similarity is defined relative to the flood response behaviour of the target basin. This non-parametric measure captures monotonic but not necessarily linear relationships between climate indices and streamflow response, while remaining robust to the heterogeneity and outliers inherent in a multi-continental tropical basin sample. Then, a composite climate similarity index (CI) is constructed as follows:

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

$$\rho_i = 1 - \frac{6\sum_{i=1}^{n} d_i}{n(n^2-1)}, \quad (2)$$

$$w_i = \frac{\rho_i}{\sum_{i=1}^{n} \rho_i}, \quad (3)$$

$$CI = \sum_{i=1}^{n} w_i . x_i , \quad (4)$$

where ρ_i is the Spearman rank correlation coefficient associated with the ith climate index; d_i represents the difference between the ranks of paired observations for the ith climate index; n is the total number of observations; w_i is the weight assigned to the ith climate index; x_i denotes the normalized value of the ith climate index; and CI refers to the composite climate indices.

Following the computation of CI for each of the 200 tropical catchments identified from the Caravan dataset, basins were ranked in ascending order of CI, such that higher-ranked basins exhibit greater climatic similarity to the ORB. We constructed four cumulative subsets (50, 100, 150, and 200 basins), where each larger set strictly encompasses all basins of the preceding one, ensuring a systematic and monotonic increase in corpus size while preserving CI based ordering. To prevent data leakage and ensure an unbiased evaluation of transfer learning performance, a fixed set of 50 basins were selected through stratified random sampling across CI quartiles, ensuring that the validation and test set spans the full range of climatic similarity rather than being biased toward either highly similar or dissimilar catchments. This stratification guarantees representativeness of the evaluation set and avoids over-optimistic performance estimates that would arise from a purely random holdout.

The remaining 150 basins were subsequently re-ranked in ascending order of CI and used to construct three pre-training subsets (50, 100, and 150 basins) corresponding respectively to the 50, 100, and 150 most CI-ranked available basins. This cumulative design allows a controlled assessment of the effect of pre-training corpus size and climatic similarity on fine-tuned streamflow forecasting performance over the ORB, with the 50 fixed basins serving exclusively as the validation and test set throughout all experiments (Figure 3).

as the validation and test set throughout all experiments (Figure 3).

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

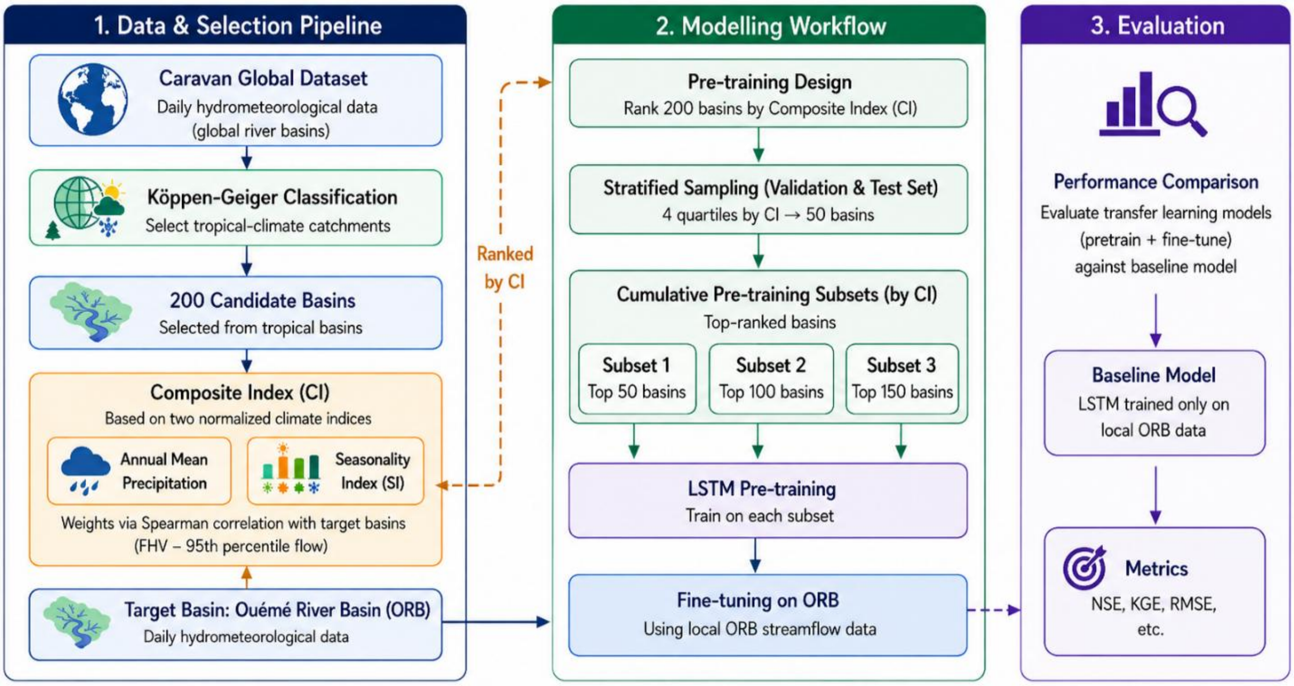

Figure 3: Methodologic flowchart

2.4 LSTM model setup, training, validation and test

Hydrology-based rainfall–runoff modelling increasingly relies on deep learning approaches capable of capturing complex temporal dependencies in hydroclimatic processes. Among these approaches, Long Short-Term Memory networks have demonstrated strong performance in streamflow simulation due to their ability to learn long-term temporal relationships between meteorological inputs and hydrological responses. This characteristic is particularly relevant for rainfall–runoff processes, where catchment behaviour is influenced by antecedent hydroclimatic conditions accumulated over extended periods (Kratzert et al., 2018: Feng et al., 2021). Unlike conventional recurrent neural networks that rely on a single hidden state, LSTM architectures incorporate a memory cell regulated by three gating mechanisms: the input gate, forget gate, and output gate. These gates control the integration of new information, the retention or removal of previous states, and the transmission of internal memory to subsequent network layers. Such a structure effectively alleviates the vanishing-gradient problem commonly encountered in standard recurrent networks and enables the model to represent hydrologically meaningful memory processes, including soil moisture storage, groundwater recession, and seasonal catchment dynamics (Lees et al., 2022).

All LSTM models were implemented within the NeuralHydrology framework (Kratzert et al., 2018) and trained using a set of hyperparameters across pre-training and fine-tuning stages. Model hyperparameters, including learning rate, dropout rate, batch size and hidden size were tuned manually to optimize predictive performance. A GPU-accelerated LSTM architecture (cudalstm), which learns nonlinear input-output relationships through trainable weight matrices updated during backpropagation was used. The hidden state size was set to 256, with a dropout rate of 0.4 applied during training to mitigate overfitting. Models were optimized using the Adam optimizer with an initial learning rate of 0.0005 over 50 epochs and a batch size of 256. Mean Squared Error (MSE) was used as the training loss function. For each group of basins (50, 100 and 150 basins), an independent LSTM was pre-trained on the 1996-2012 period. The evaluation of each model relied exclusively on the 50 basins fixed through stratified sampling, the validation was performed on the 2013-2015 period with 30 basins, while the test period spans 2016-2020 with 20 basins. This strict temporal separation prevents data leakage and yields an unbiased estimate of generalization performance. These periods were selected to maximize the use of the available streamflow observations.

Each pre-trained model was subsequently fine-tuned on the five target stations of the Ouémé River Basin. The three transfer-learning models were fine-tuned on the ORB after pretraining on 50, 100, and 150 Caravan catchments, and are referred to as LSTM1, LSTM2, and LSTM3, respectively. The fine-tuning models retained the same architecture and hyperparameters used during pretraining, except for the learning rate. In this study, the learning rate was increased from 0.0005 to 0.001 to accelerate adaptation to the ORB. Although fine-tuning is commonly performed with a lower learning rate to preserve pretrained representations, this higher value was selected based on empirical performance during validation. During fine-tuning, all layers of the pre-trained network were frozen except the regression head, a single linear layer mapping the LSTM hidden state (size 256) to the predicted streamflow. Only the parameters of this layer (256 weights and 1 bias) were updated, while the LSTM input, recurrent, and gate weights retained their pre-trained values, to adapt the model's predictions to the hydrological regime of the Ouémé River Basin. Fine-tuning followed the same temporal partitioning as pre-training, with the 1996-2012 period used for training, 2013-2015 for validation, and 2016-2020 for testing. We also trained LSTM model exclusively on local data as baseline model using the same architecture as the fine-tuned model. This model is named LSTM4.

2.5 Evaluation Metrics

Quantitative metrics are required to assess how closely a predictive model reproduces observed behaviour, and they form a standard component of model evaluation in hydrology and machine learning. In this study, four complementary metrics are used. The Kling-Gupta Efficiency (KGE) decomposes the overall fit into correlation, variability ratio, and bias components, which makes it well suited to diagnosing the origin of streamflow errors (Gupta et al., 2009). The Nash-Sutcliffe Efficiency (NSE) measures the agreement between observed and simulated values by comparing the residual variance to the variance of the observations (Nash and Sutcliffe, 1970). The coefficient of determination (R^2) (Barber et al., 2020), defined here as the square of the Pearson correlation, quantifies the proportion of observed variance that is linearly explained by the simulation. The Root Mean Square Error (RMSE) (Hodson, 2022) measures the average magnitude of the errors in the units of the target variable, giving more weight to large deviations than to small ones. The KGE is a standard hydrology metric that jointly

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

accounts for correlation, bias, and variability between simulated and observed streamflow. KGE metric is calculated as follows:

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}\,, \tag{5}$$

Here, r represents the Pearson correlation coefficient between observed and simulated streamflow, α is the ratio of their standard deviations, and β is the ratio of their means. A KGE of 1 denote perfect model-observation agreement; values near 0 or below denotes poor performance.

$$NSE = 1 - \frac{\sum_{i=1}^{n}(Q_{obs,i} - Q_{sim,i})^2}{\sum_{i=1}^{n}(Q_{obs,i} - \bar{Q}_{obs})^2}\,, \tag{6}$$

$$R^2 = \left(\frac{\sum_{i=1}^{n}(Q_{obs,i}-\bar{Q}_{obs})(Q_{sim,i}-\bar{Q}_{sim})}{\sqrt{\sum_{i=1}^{n}(Q_{obs,i}-\bar{Q}_{obs})^2}\sqrt{\sum_{i=1}^{n}(Q_{sim,i}-\bar{Q}_{sim})^2}}\right)^2, \tag{7}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_{obs,i} - Q_{sim,i})^2}\,, \tag{8}$$

Here, $Q_{obs,i}$ and $Q_{sim,i}$ represents the observed and simulated of streamflow discharge at time step i; $\bar{Q}_{obs}$ denotes the mean value of observed discharge; $\bar{Q}_{sim}$ is the mean value of simulated discharge, and n is the total count of events included.

3 Results

3.1 Construction and validation of the composite climate similarity index

Figure 5a shows the distributions of the two normalized climate deviation indices ΔP and ΔSI computed for the 200 tropical catchments selected from the global Caravan dataset. Both distributions are right-skewed, with the majority of catchments exhibiting relatively low deviations from the Ouémé reference values, and a limited number of outliers representing climatically distant regimes at the margins of the tropical Köppen-Geiger classification. The ΔP distribution is more compact (median = 0.17), indicating that annual precipitation is broadly comparable across the selected tropical catchments, while ΔSI displays greater spread (median = 0.21), reflecting substantial variability in seasonal flow concentration among tropical basins worldwide. This higher discriminant power of ΔSI across the donor pool provides a first indication that seasonality captures hydrologically relevant differences between catchments that annual precipitation alone cannot resolve.

The empirical calibration of the CI weights is presented in Figure 5b, which shows the distributions of Spearman rank correlation coefficients between each annual climate index and the observed high-flow volume (Q95) at the Bonou outlet over the study period. Both ρ_P and ρ_SI are predominantly positive across the catchment ensemble, confirming that deviations in annual precipitation and seasonality are systematically and coherently related to high-flow behaviour in the Ouémé basin. The median Spearman coefficient for precipitation (ρ_P = 0.58) exceeds that of the seasonality index (ρ_SI = 0.38), establishing an empirical hierarchy that directly informs the weight assignment. This data-driven calibration ensures

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

that the CI reflects the relative hydrological importance of each climate descriptor for the target basin, rather than relying on arbitrary a priori assumptions.

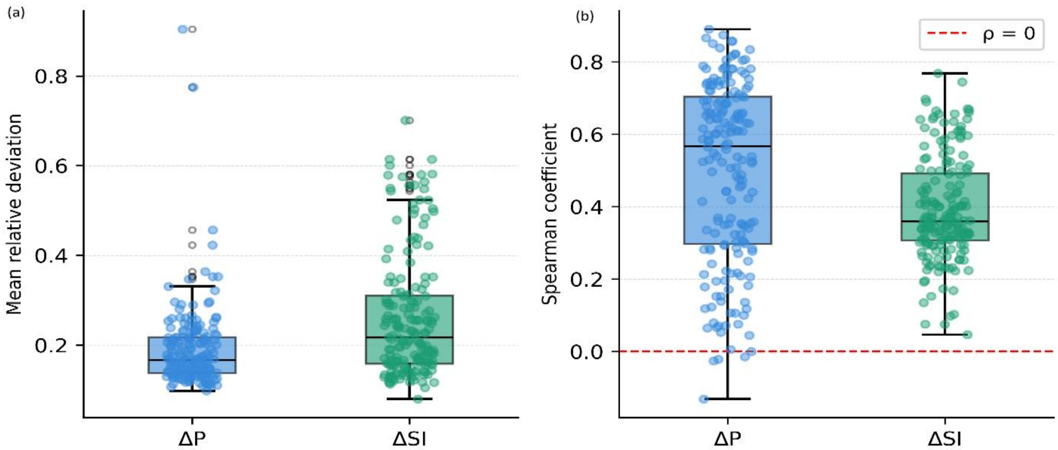

Figure 4: The distributions of the two normalized climate deviation indices ΔP and ΔSI (a) and empirical calibration of the CI weights of these indices (b).

Figure 5 illustrates the relationship between the composite index (CI) and each of its constituent components across the 200 donor catchments. Colors indicate membership in the cumulative pre-training subsets, with green representing the first 50 catchments, blue the first 100 catchments, orange the first 150 catchments, and red the complete set of 200 catchments. The CI is strongly correlated with ΔP ($r = 0.94$, $p < 0.001$), reflecting the dominant role of the precipitation deviation in structuring the composite index, and moderately correlated with ΔSI ($r = 0.60$, $p < 0.001$), confirming the complementary and non-redundant contribution of seasonality. The spatial segregation of subsets in both scatter plots is consistent with the cumulative selection strategy: the 50 basins catchments cluster tightly in the lower-left quadrant of both panels, representing the most climatically similar donors, while 100 and 150 catchments extend progressively towards higher deviation values. This ordered structure confirms that the CI-based ranking successfully identifies donors of decreasing climatic proximity to the Ouémé basin, thereby providing a reproducible and physically grounded framework for donor selection in transfer learning applications across data-scarce tropical regions.

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

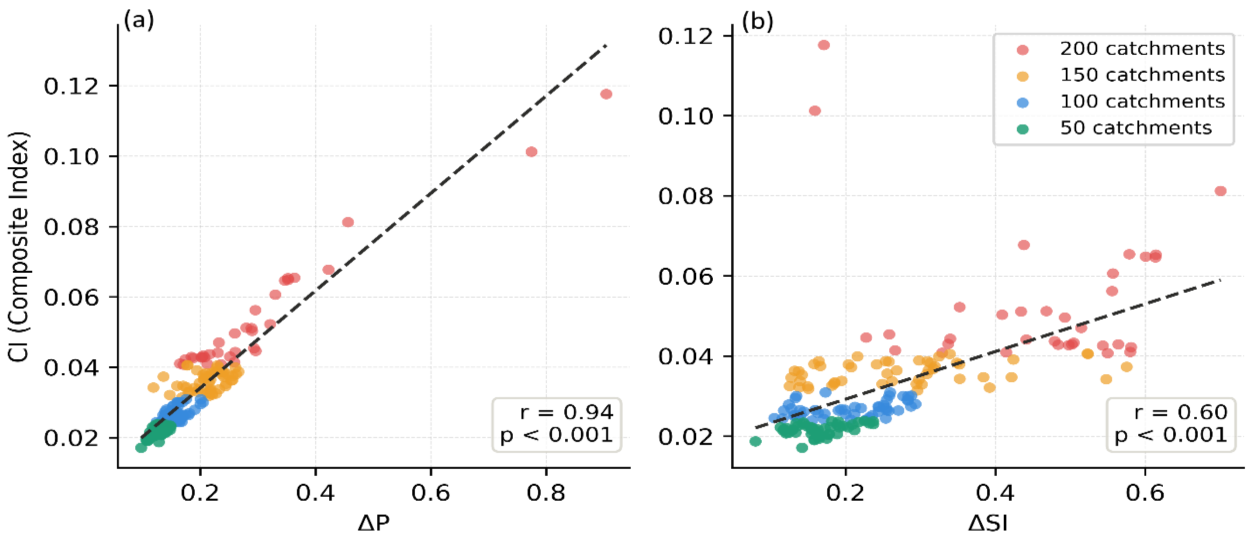

Figure 5: Relationship between the composite index CI and each climate index: (a) annual mean precipitation; (b) seasonality index.

3.2 Evaluation of pre-trained models in training, validation and testing period results

Figure 6 shows the KGE, R^2, and RMSE distributions across training, validation, and test periods for the three pre-trained LSTM configurations. All models achieve high training performance, with median KGE values exceeding 0.95 for LSTM1, 0.85 for LSTM2 and 0.84 for LSTM3. This reflects the capacity of LSTM networks to fit complex rainfall–runoff dynamics when trained on sufficiently large datasets. The pre-trained LSTMs are completely blind to the test period basins, making the test results a genuine measure of out-of-sample generalisation. A consistent decline in performance from training to test is observed across all three configurations and all three metrics, which is expected behaviour for data-driven models. However, this gap narrows progressively as corpus size increases: LSTM1 shows the largest training-to-test drop in KGE (approximately 0.28 units), while LSTM2 and LSTM3 display smaller and comparable declines. Comparing the three configurations in the test period, performance improves monotonically with corpus size across all metrics. Median KGE decreases from 0.67 for LSTM1 to 0.61 for LSTM3, median R^2 rises from 0.66 to 0.68, and median RMSE decreases from 0.49 to 0.31 mm day^{-1}. Although the absolute gains between consecutive configurations are modest, the interquartile range narrows from LSTM1 to LSTM3, indicating that larger pre-training corpora produce not only marginally better but also more consistent generalisation across unseen basins. These results confirm that corpus size positively influences predictive skill at the global Caravan scale.

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

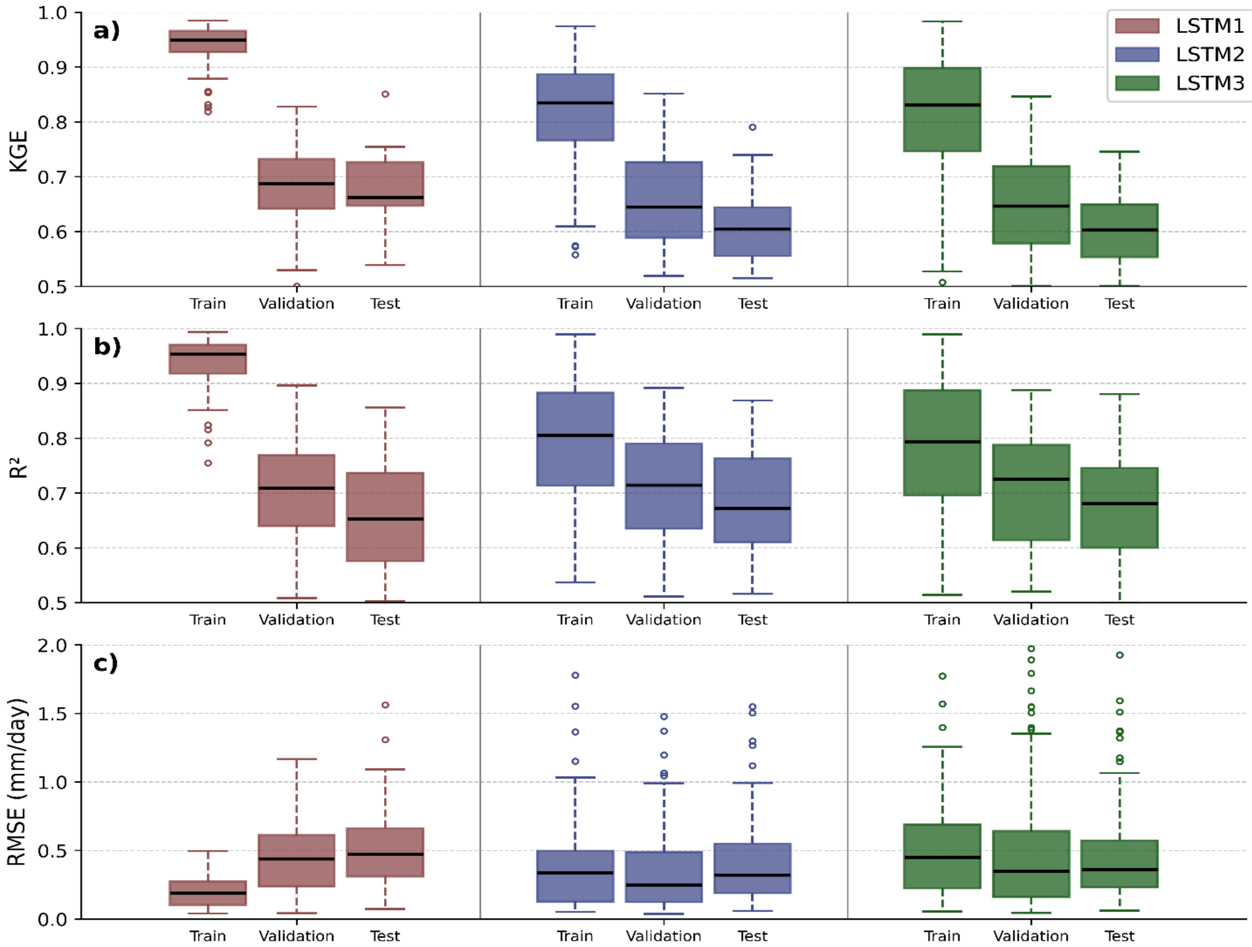

Figure 6. Boxplot distributions of model performance metrics across training, validation, and test periods for the three pre-trained LSTM configurations: LSTM1 (50 basins), LSTM2 (100 basins), and LSTM3 (150 basins). Panels show (a) Kling-Gupta Efficiency (KGE), (b) coefficient of determination (R^2), and (c) root mean square error (RMSE, mm day^{-1}).

3.3 Performance of the fine-tuned models on the Ouémé River Basin

Before fine-tuning, each pre-trained model was applied directly to the Ouémé stations over the test period (2016-2020), using the CARAVAN-trained weights without any local adjustment. This zero-shot evaluation isolates what climate-based pre-training alone transfers to the target basin (Fig. 7a). The outcome depends strongly on the size of the donor corpus. The smallest and most similar set (LSTM1, 50 basins) already yields usable predictions at all five stations, with a median KGE of 0.59 and positive values everywhere, from 0.40 at Beterou to 0.81 at Savè. Skill then collapses as less similar catchments are added: the median KGE falls to 0.15 for LSTM2 (100 basins) and to −1.92 for LSTM3 (150 basins), and the number of stations with positive KGE drops from five to three and then to one. At LSTM3, several stations show large positive high-flow biases and strongly negative efficiencies, which shows that a model pre-trained on a broader, less similar corpus does

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

not represent the Ouémé regime without local correction. Direct transfer is therefore useful only when the donor basins are closely matched to the target in climate.

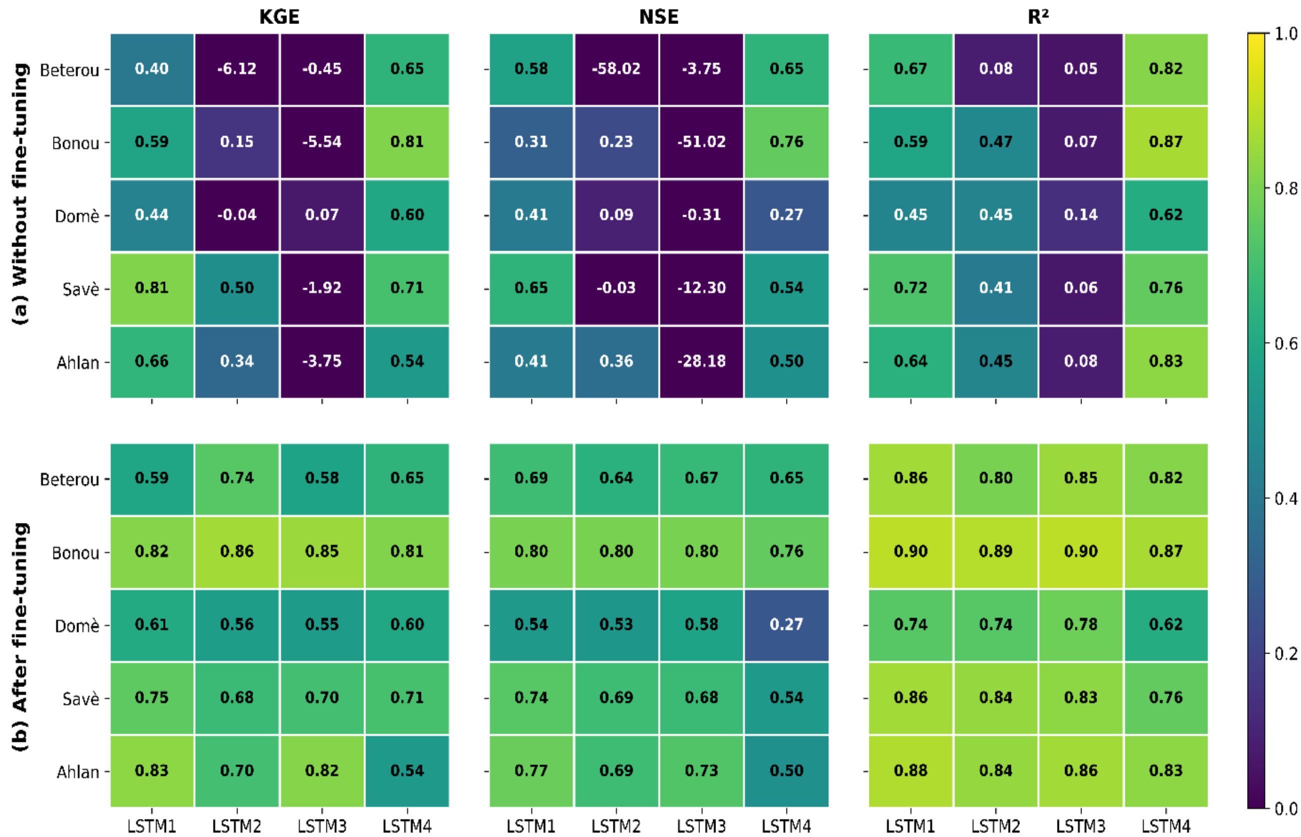

Figure 7. Kling-Gupta Efficiency (KGE), Nash-Sutcliffe Efficiency (NSE), and coefficient of determination (R^2) at the five Ouémé gauging stations (Beterou, Bonou, Domè, Savè, Ahlan) for four LSTM configurations, evaluated over the independent test period (2016-2020). (a) Pre-trained models applied directly to the Ouémé without fine-tuning, shown together with the local baseline (LSTM4) for reference; (b) the same models after fine-tuning on the five local stations. LSTM1, LSTM2, and LSTM3 are pre-trained on the 50, 100, and 150 donor basins of highest climatic similarity to the Ouémé; LSTM4 is trained from scratch on local data only and is identical in both rows, as it has no pre-trained (zero-shot) version. Cell colours are clipped to the [0, 1] range for readability; the numerical values, including negative ones, are printed in each cell.

Adapting each pre-trained model to the five Ouémé stations lifts performance almost everywhere and, more importantly, removes the strong dependence on corpus size seen in the zero-shot case (Fig. 7b). The first result show that the three pre-trained models clearly outperform the local baseline. Across the five stations, the median KGE over the test period is 0.75 for LSTM1, 0.70 for LSTM2, 0.70 for LSTM3, and 0.65 for LSTM4; the gap is wider for NSE, at 0.74, 0.69, 0.68, and 0.54 respectively. The improvement is largest at the data-poor stations. At Ahlan, KGE rises from 0.54 with the local model to 0.83 with LSTM1 and NSE from 0.50 to 0.77; at Domè, NSE increases from 0.27 with the local model to 0.54–0.58 after

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

pre-training. This consistent gain confirms that the climate-similarity-based transfer framework exploits the global CARAVAN dataset to compensate for the short in-situ records typical of West African tropical catchments.

The second result concerns the ranking among the three pre-trained models. LSTM1, pre-trained on the 50 most similar donor basins, gives the best median performance on all three metrics (KGE = 0.75, NSE = 0.74, R^2 = 0.86), while LSTM2 and LSTM3 settle slightly lower (median KGE = 0.70; median NSE = 0.69 and 0.68). Enlarging the donor corpus beyond the most similar group therefore does not improve fine-tuning performance and slightly reduces it. The 50-basin set appears to capture the climatic and hydrological characteristics most representative of the Ouémé, whereas the basins added in LSTM2 and LSTM3 introduce redundant or less relevant information.

Station-level responses refine this reading. Bonou is the most stable station: after fine-tuning, KGE remains above 0.80 and R^2 close to 0.90 for all four models, including the local one, so the added value of transfer learning there is marginal. Domè, Savè, and Ahlan show a clear preference for the smallest corpus, LSTM1 giving the best KGE at each (0.61, 0.75, and 0.83); these are also the stations where the framework helps most, especially Ahlan, where the gain over the local baseline is largest (ΔKGE = +0.29). Beterou is the only exception, its best KGE being reached with LSTM2 (0.74) rather than LSTM1 a behaviour we return to in the Discussion.

These results indicate that climatic similarity governs how much a model transfers before any local training: the more similar the donors, the more skill is available without adaptation. Second, fine-tuning is not a marginal refinement but a necessary step that compensates for the inclusion of less similar donors, which is why the large differences seen without fine-tuning (Fig. 7a) largely disappear after it (Fig. 7b). The relationship between donor-corpus size and fine-tuned performance is not monotonic, and the smallest, most similar corpus gives the best overall results.

Figure 8 presents the observed and simulated daily streamflow at the three Ouémé stations Domè (a), Ahlan (b), and Bonou (c) during the 2016-2020 test period.

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

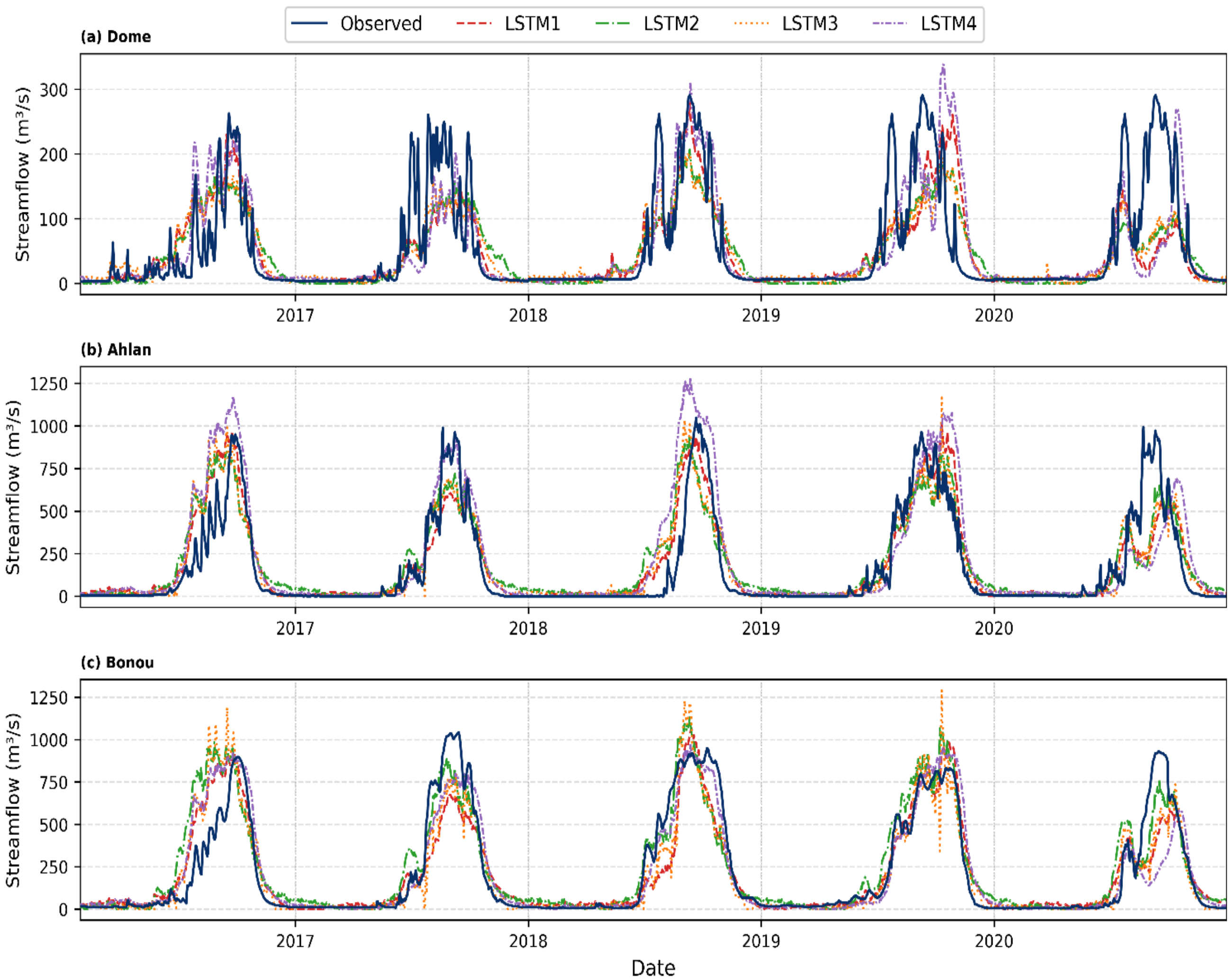

Figure 8: Observed and simulated daily streamflow at the three Ouémé stations Domè (a), Ahlan (b), and Bonou (c) during the 2016-2020 test period.

Overall, the four LSTM configurations successfully reproduce the main seasonal dynamics at all stations. The onset of the rainy season, the occurrence of the annual flood peak, and the prolonged low-flow period between October and June are consistently captured. At Domè and Ahlan, the simulated dry-season discharge decreases to near-zero values, in agreement with the observed intermittent flow conditions characteristic of the Sudano-Sahelian climate. In addition, the timing of both the rising and recession limbs is generally well reproduced, indicating that the models effectively learned the dominant rainfall–runoff response of the basin.

The main discrepancies are related to the magnitude and sharpness of extreme flood events. At Domè (Fig8. a), the observed hydrograph is characterized by abrupt multi-peak flood episodes, particularly during 2018 and 2020, whereas all four models

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

tend to smooth these events and underestimate the highest discharges. Among the configurations, LSTM4 (local training baseline) reproduces the 2019 flood peak more closely but substantially overestimates the 2020 event, suggesting reduced robustness when the model is trained only on local data with limited extreme-event representation. At Ahlan (Fig8. b), the pre-trained models reproduce the 2016 and 2017 flood peaks reasonably well but underestimate the 2018 flood event. In contrast, LSTM4 overestimates this peak by nearly 200 m³/s. This behavior is consistent with the lower KGE and NSE values previously reported for LSTM4 at this station, confirming a less stable generalization capability compared to the pre-trained configurations. At Bonou (Fig8. c), the agreement between observed and simulated streamflow is generally strong for all configurations. Only minor deviations are observed, including a slightly premature flood peak in 2017 for several models and short-duration artificial spikes in 2018 generated by LSTM3 and LSTM4 that are not present in the observations.

The three pre-trained models (LSTM1, LSTM2, and LSTM3) exhibit very similar hydrographs across all stations, and the differences among them remain relatively small compared to their differences with the locally trained LSTM4 model. This observation supports the performance metrics discussed in Section 3.3 and suggests that, once transfer learning is applied, increasing the size of the donor-basin corpus has only a limited impact on model performance. Across all configurations, the largest residual error remains the underestimation of extreme peak flows, especially at Domè where the hydrograph exhibits the strongest short-term variability.

4 Discussion

4.1 Quality over quantity in the donor corpus for similarity-based transfer learning

The clearest finding of this study is that, among the donor-set sizes tested, the smallest and most climatically similar donor set provides the best transfer-learning performance across the five Ouémé stations. Compared with the locally trained baseline model (LSTM4), similarity-based pre-training using the 50-catchment subset (LSTM1) substantially improves model performance, with the median KGE increasing from 0.65 to 0.75, the median NSE from 0.54 to 0.74, and the median R^2 from 0.82 to 0.86. Increasing the donor pool size to 100 (LSTM2) and 150 catchments (LSTM3) does not lead to further improvements. Instead, the median KGE decreases to approximately 0.70, accompanied by similar reductions in NSE. These results suggest that, in this experimental setting, transfer learning benefits more from donor similarity than from simply increasing the number of donor catchments.

This finding somewhat contrasts with the common expectation in deep-learning-based hydrology that larger donor datasets generally improve model generalization (Arsenault et al., 2023; Fang et al., 2021; Kratzert et al., 2019; Wilbrand et al., 2023). In our case, we interpret this behaviour as the result of two competing mechanisms during pre-training. The first is a generalization effect, whereby larger and more diverse datasets stabilize model parameters and reduce overfitting during fine-tuning. The second is a specificity effect, whereby climatically similar donor catchments encode hydrological responses that are closer to those of the target basin and therefore require fewer adjustments during fine-tuning. In this study, expanding the donor set from 50 to 150 catchments progressively introduces basins with larger differences in annual

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

precipitation and seasonality relative to the Ouémé River Basin. Under the present training setup, this additional diversity appears to provide little benefit for the Ouémé fine-tuning task and may even introduce a mild domain-shift penalty. The performance of the lowest-CI donor subset provides preliminary support for the proposed climate similarity index as a donor-ranking criterion.

The CI combines normalized differences in annual precipitation and seasonality index, weighted according to the Spearman correlations between these climate descriptors and the observed Q95 at the Ouémé outlet. The fact that the lowest-CI subset consistently produces the best fine-tuning performance provides an internal validation of the proposed similarity framework: the same index used to define climatic similarity prior to training also explains the transfer learning performance obtained after fine-tuning. However, this should be interpreted as an encouraging internal consistency check rather than a general validation of the similarity framework, which would require testing across additional target basins and donor-selection strategies.

Nevertheless, several limitations should be acknowledged. First, the optimal donor corpus size remains uncertain, as only three subset sizes (50, 100, and 150 catchments) were evaluated, and the true optimum may lie below 50 catchments. Second, the CI was calibrated using a single reference basin (Ouémé at Bonou), and its applicability to other tropical river basins remains to be assessed. Third, the similarity framework relies exclusively on two climatic descriptors, namely annual precipitation and the seasonality index. Additional hydroclimatic characteristics, such as rainfall intermittency, baseflow conditions, or vegetation seasonality, could further improve donor ranking and transferability assessment. Fourth, performance differences between the 50-, 100-, and 150-catchment donor sets may also partly reflect modelling and training choices, including model size, fine-tuning learning rate, pretraining duration, sampling strategy, and normalization. Although a hyperparameter search was conducted, it was not fully exhaustive due to computational limitations; therefore, the observed differences should be interpreted as resulting from both donor similarity and the specific training configuration used. These limitations suggest that similarity-based transfer learning should be viewed as a continuous optimization problem involving several interacting factors, among which donor corpus size represents only one component.

4.2 Station-level heterogeneity of the transfer benefit

Figure 9 highlights a strong spatial variability in transfer-learning benefits across the five Ouémé stations. The transfer gain, ranges from only +0.01 at Domè to +0.29 at Ahlan. Three main behaviors can be identified. First, Ahlan shows a strong dependence on pre-training, with substantial improvements obtained for all transferred models, regardless of donor corpus size. Second, Bonou exhibits a saturated regime, where all configurations achieve similarly high performance and transfer learning provides only marginal gains. Third, Domè and Savè clearly favor the smallest and most climatically similar donor set (LSTM1), while performance decreases as less similar basins are added. Beterou represents an exception, as the intermediate donor corpus (LSTM2) yields the best performance, whereas LSTM1 performs worse than the local baseline. Overall, these results indicate that transfer learning effectiveness is strongly site-dependent and that increasing donor basin number does not systematically improve predictive skill.

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

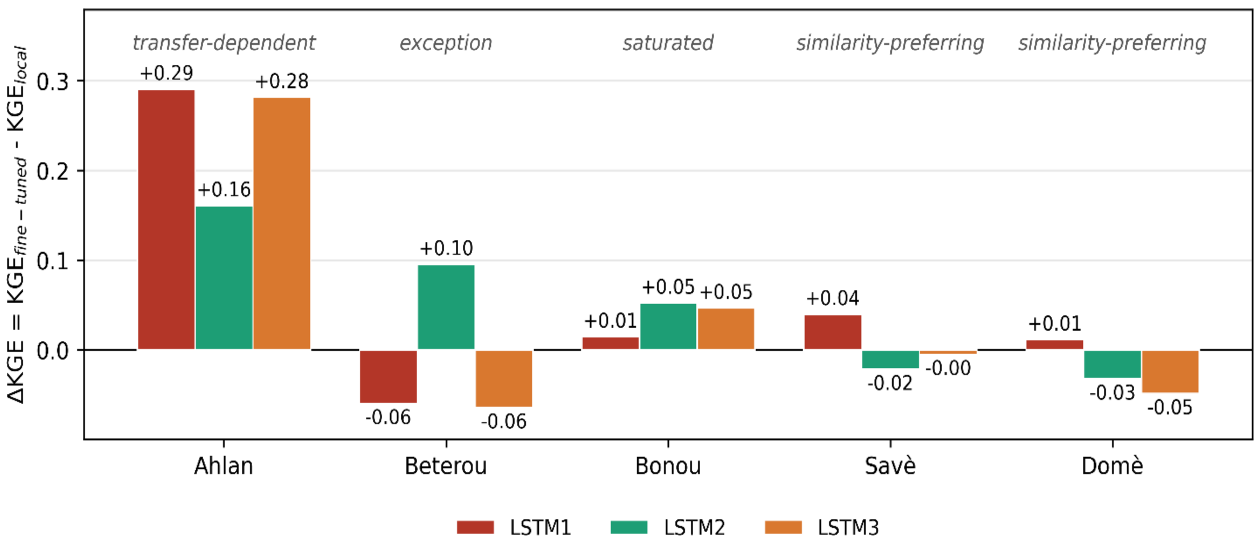

Figure 9: Spatial variability of transfer-learning benefits across the Ouémé stations

These differences can be related to two hydrological controls. The first is the length and richness of the local record relative to the events to be predicted. Ahlan and Domè have shorter usable records with few large floods in the fine-tuning period, which limits the local LSTM and leaves room for pre-training to supply the missing event diversity. This interpretation is consistent with the data-availability threshold reported for single-basin LSTMs (Ayzel and Heistermann, 2021)and with the observation of (Martel et al., 2025) that the value of donor data is highest where local extreme events are scarce. Bonou, the basin outlet, integrates the upstream signal and benefits from a longer, more consistent record; its local LSTM is already close to the achievable performance, so pre-training adds little. The second control is the position of the station within the basin. Bonou and Ahlan are downstream stations where the streamflow signal is well integrated and the climatic forcing is averaged over a large drainage area, whereas Domè and Savè are intermediate stations whose response is closer to the forcing scale represented in the LSTM1 model.

As the most upstream station of the basin, Beterou exhibits a distinct response pattern that likely reflects specific hydrological controls rather than anomalous behavior. The improvement in KGE observed when expanding the donor corpus from 50 to 100 catchments (+0.15) suggests that Beterou benefits from additional regional hydrological information beyond climatic similarity alone. This may indicate that processes such as groundwater contributions or sub-grid heterogeneity, which are not fully represented in the strictly climate-similar 50 basins set, are better captured in the larger donor corpus. A comparable behaviour was reported by Y (Yao et al., 2022), who found that transfer-learning LSTMs were least effective in headwater catchments with strong groundwater contributions, where climate forcing is a weaker predictor of streamflow. The Beterou case therefore does not invalidate the CI ranking; it indicates that climatic similarity is a necessary but not sufficient criterion for optimal donor selection at the station scale, and that physiographic descriptors may need to complement it for upstream catchments.

4.3 Implications for operational flood forecasting in data-scarce tropical basins

This work is relevant for flood-prone basins such as the Ouémé River Basin, where improved streamflow prediction can support hydrological monitoring and risk-informed decision-making under sparse observational conditions. Against this objective, the most important result is also the most sobering. All four configurations, including the best similarity-based model (LSTM1), underestimate the highest observed peaks at Domè and Ahlan in 2018 and 2020. The pre-training improves the overall fit and the timing of the flood wave, but it does not remove the tendency of the LSTM to smooth the sharpest flood peaks. This behaviour is consistent with the broader deep-learning hydrology literature, where LSTMs reproduce the central part of the flow distribution well but lose accuracy at the upper tail (Frame et al., 2022b; Martel et al., 2025). For flood forecasting, this is the part of the distribution that matters most.

The operational consequence depends on how the prediction is used. Underestimating a flood peak and overestimating it do not carry the same cost: a missed peak translates into a missed or under-scaled warning, whereas an overestimated peak leads at worst to a false alarm. The local baseline (LSTM4) tends to overshoot the Ahlan peaks while the pre-trained models tend to undershoot them, which means the choice of configuration is not neutral from a risk-management standpoint. A model selected on aggregate metrics such as median KGE may not be the model that minimizes the operational cost of flood underprediction, and the selection criterion for an early-warning application should be weighted towards high-flow performance rather than overall fit.

Four avenues can be explored to reduce the residual peak error in the West African context. The first is to give more weight to high-flow events during fine-tuning, either through a peak-weighted loss function or through resampling of the monsoon-season records, which has been shown to improve neural-network high-flow accuracy without degrading the rest of the hydrograph (Snieder et al., 2021) . The second is to combine the LSTM with a locally calibrated conceptual model in a residual or hybrid framework, an approach that already improved peak simulation on the Ouémé at Bonou when GR4J was coupled with machine learning (Ahouandjinou et al., 2026) and that mirrors the gain reported by Martel et al. (2025) when hydrological-model outputs were added to LSTM inputs. The third is to explore more physically guided LSTM variants or hybrid process–deep learning architectures, as recent studies suggest that explicitly representing hydrological processes can improve the prediction of extreme events (Song et al., 2026). The fourth is to address the quality of the meteorological forcing: ERA5-Land precipitation is known to have substantial biases over West Africa, and satellite-based products such as CHIRPS or TAMSAT may better capture the convective rainfall that drives the largest floods (Dembele et al., 2020). These four avenues are complementary and could be combined within the transfer-learning framework presented here.

For the operationalization, a similarity-based fine-tuned LSTM such as LSTM1 is a usable tool for daily streamflow monitoring, seasonal water-resource planning, and the detection of the onset and recession of the annual flood, where it clearly outperforms a locally trained model at the data-poorest stations. It is not yet a reliable tool for the magnitude of

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

design floods or rare return periods, where the peak underestimation remains a barrier and where a hybrid or peak-weighted formulation would be required. This conclusion echoes the operational trajectory of large-scale deep-learning flood models, which have reached skilful lead times for ungauged basins worldwide while still being deployed alongside, rather than in place of, established forecasting systems (Nearing et al., 2024). Similarity-based transfer learning is best seen as a way to extend that capability to data-scarce tropical basins, not as a standalone replacement for the hydrological reasoning that underpins flood risk management.

5 Conclusion

This study introduced a composite climate-similarity index to guide donor selection for transfer learning in a data-scarce tropical basin, and tested it on five stations of the Ouémé River Basin in Benin. Pre-training an LSTM on climatically similar CARAVAN catchments, then fine-tuning it locally, improved daily streamflow prediction at every station relative to a locally trained model.

The central result is that similarity outweighs volume: the smallest, most similar donor set (50 catchments) gave the best fine-tuning performance, and adding less similar catchments did not help. This runs against the assumption that larger donor pools are always preferable, and it points to donor selection, not donor quantity as the lever that matters most when transferring global knowledge to poorly gauged African basins. The framework is already useful for daily monitoring and flood-onset detection, though not yet for the magnitude of the largest floods, which all configurations underestimate. Three directions would strengthen it: testing donor sets smaller than 50 catchments, calibrating the index on several reference basins rather than one, and adding physiographic descriptors to capture controls such as the crystalline-basement geology of the upper Ouémé that climate alone cannot resolve.

Beyond the Ouémé, the approach offers a transferable and low-cost template for streamflow prediction across data-scarce tropical regions, where global hydrological datasets carry the most value precisely where local networks are weakest.

Appendix A: Catchment descriptors

This study relies on a set of static catchment descriptors that help the LSTM models learn the relationship between the meteorological time series and streamflow. To this end, 23 static attributes were extracted for each catchment from the HydroATLAS database (Linke et al., 2019) and from the ERA5-Land climatic indices provided within the CARAVAN dataset (Kratzert et al., 2023). A summary of these attributes is presented in Table A1.

Table A1. Static catchment attributes used as input features for the LSTM models. Attributes are derived from the HydroATLAS database and from ERA5-Land climatic indices provided within the CARAVAN dataset.

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

Attribute	**Description**	**Unit**	**Source**
aet_mm_syr	Actual evapotranspiration, annual average	$\mathrm{mm\ yr^{-1}}$	HydroATLAS
pet_mm_syr	Potential evapotranspiration, annual average	$\mathrm{mm\ yr^{-1}}$	HydroATLAS
dis_m3_pyr	Natural discharge, annual average	$\mathrm{m^3\ s^{-1}}$	HydroATLAS
run_mm_syr	Land surface runoff, annual average	$\mathrm{mm\ yr^{-1}}$	HydroATLAS
inu_pc_smx	Inundation extent, maximum	%	HydroATLAS
rev_mc_usu	Reservoir volume, total upstream	million $\mathrm{m^3}$	HydroATLAS
gwt_cm_sav	Groundwater table depth, average	cm	HydroATLAS
ele_mt_sav	Elevation, average	m	HydroATLAS
slp_dg_sav	Terrain slope, average	degrees × 10	HydroATLAS
sgr_dk_sav	Stream gradient, average	$\mathrm{dm\ km^{-1}}$	HydroATLAS
pre_mm_syr	Precipitation, annual average	$\mathrm{mm\ yr^{-1}}$	HydroATLAS
ari_ix_sav	Global aridity index, average	–	HydroATLAS
tmp_dc_syr	Air temperature, annual average	°C × 10	HydroATLAS
area	Catchment drainage area	$\mathrm{km^2}$	Caravan
aridity_index	Aridity index (PET / P)	–	ERA5-Land (Caravan)
high_prec_dur	Mean duration of high-precipitation events	days	ERA5-Land (Caravan)
high_prec_freq	Frequency of high-precipitation events	$\mathrm{days\ yr^{-1}}$	ERA5-Land (Caravan)
low_prec_dur	Mean duration of low-precipitation (dry) events	days	ERA5-Land (Caravan)
low_prec_freq	Frequency of low-precipitation (dry) events	$\mathrm{days\ yr^{-1}}$	ERA5-Land (Caravan)
moisture_index	Moisture index	–	ERA5-Land (Caravan)
p_mean	Mean daily precipitation	$\mathrm{mm\ day^{-1}}$	ERA5-Land (Caravan)
pet_mean	Mean daily potential evapotranspiration	$\mathrm{mm\ day^{-1}}$	ERA5-Land (Caravan)
seasonality	Precipitation seasonality index	–	ERA5-Land (Caravan)

Code and data Availability Statement

Catchment descriptors and meteorological forcing data from ERA5-Land were obtained using the CARAVAN framework https://github.com/kratzert/Caravan/ (Kratzert et al., 2023). We implemented the LSTM model using the NeuralHydrology Python package, which is freely available at https://github.com/neuralhydrology/neuralhydrology (Kratzert et al., 2022). Data processing and the codes used in this study are available at https://doi.org/10.5281/zenodo.21813084 and https://github.com/Enagnon97/TL-ORB.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Author contributions

Jérôme Enagnon Ahouandjinou: Conceptualization, Methodology, Software, Validation, Data curation, Writing-Original draft preparation, Visualization, Writing- Review & Editing. Aymar Yaovi Bossa: Supervision, Methodology, Writing- Review & Editing. Jean Hounkpe: Supervision, Writing- Review & Editing. Riccardo Taormina: Conceptualization, Supervision, Writing- Review & Editing.

References

Ahouandjinou, J. E., Bossa, A. Y., and Hounkpe, J.: Residual-based hybrid modeling combining GR4J and machine learning for streamflow prediction in data-scarce catchment: case of the Ouémé catchment at Bonou (Benin, West Africa), Proceedings of IAHS, 389, 9–15, https://doi.org/10.5194/PIAHS-389-9-2026, 2026.

Alzhanov, A., Nugumanova, A., and Moreido, V.: Transfer learning using the global Caravan dataset for developing a local river streamflow prediction model, Environmental Modelling & Software, 194, 106691, https://doi.org/10.1016/J.ENVSOFT.2025.106691, 2025.

Amoussou, E., Amoussou, F. T., Bossa, A. Y., Kodja, D. J., Totin Vodounon, H. S., Houndenou, C., Borrell Estupina, V., Paturel, J. E., Mahe, G., Cudennec, C., and Boko, M.: Use of the HEC RAS model for the analysis of exceptional floods in the Oueme basin, Proceedings of the International Association of Hydrological Sciences, 385, 141–146, https://doi.org/10.5194/PIAHS-385-141-2024, 2024.

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, Hydrol. Earth Syst. Sci., 27, 139–157, https://doi.org/10.5194/HESS-27-139-2023, 2023.

Ayzel, G. and Heistermann, M.: The effect of calibration data length on the performance of a conceptual hydrological model versus LSTM and GRU: A case study for six basins from the CAMELS dataset, Comput. Geosci., 149, 104708, https://doi.org/10.1016/J.CAGEO.2021.104708, 2021.

Barber, C., Lamontagne, J. R., and Vogel, R. M.: Improved estimators of correlation and R2 for skewed hydrologic data, Hydrological Sciences Journal, 65, 87–101, https://doi.org/10.1080/02626667.2019.1686639, 2020.

Baste, S., Klotz, D., Acuña Espinoza, E., Bardossy, A., and Loritz, R.: Unveiling the limits of deep learning models in hydrological extrapolation tasks, Hydrol. Earth Syst. Sci., 29, 5871–5891, https://doi.org/10.5194/HESS-29-5871-2025, 2025.

Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., and Wood, E. F.: Present and future köppen-geiger climate classification maps at 1-km resolution, Sci. Data, 5, 180214-, https://doi.org/10.1038/SDATA.2018.214;SUBJMETA, 2018.

Berg, P., Donnelly, C., and Gustafsson, D.: Near-real-time adjusted reanalysis forcing data for hydrology, Hydrol. Earth Syst. Sci., 22, 989–1000, https://doi.org/10.5194/HESS-22-989-2018, 2018.

Biao, E. I., Alamou, E. A., and Afouda, A.: Improving rainfall–runoff modelling through the control of uncertainties under increasing climate variability in the Ouémé River basin (Benin, West Africa), Hydrological Sciences Journal, 61, 2902–2915, https://doi.org/10.1080/02626667.2016.1164315;SUBPAGE:STRING:FULL, 2016.

Bodjrènou, M., Peng, K., Affévé, D. J., Hounkpè, J., Donnou, H. E. V., Adounkpè, J., and Akpo, A. B.: Assessment of Satellite-Based Rainfall Products for Flood Modeling in the Ouémé River Basin in Benin (West Africa), Hydrology, 12, https://doi.org/10.3390/hydrology12040071, 2025.

Bodjrènou, R., Sintondji, L. O., and Comandan, F.: Hydrological modeling with physics-based models in the oueme basin: Issues and perspectives for simulation optimization, J. Hydrol. Reg. Stud., 48, 101448, https://doi.org/10.1016/J.EJRH.2023.101448, 2023.

Bossa, A. Y., Kpossou, M. A. O., Hounkpè, J., Badou, F. D., Bossa, A. Y., Kpossou, M. A. O., Hounkpè, J., and Badou, F. D.: Multi-Model Approach for Assessing the Influence of Calibration Criteria on the Water Balance in Ouémé Basin, J. Water Resour. Prot., 16, 207–218, https://doi.org/10.4236/JWARP.2024.163012, 2024.

Choubin, B., Jaafari, A., Henareh, J., Karimi, O., and Sajedi Hosseini, F.: Explainable artificial intelligence (XAI) for interpreting predictive models and key variables in flood susceptibility, Results in Engineering, 27, 105976, https://doi.org/10.1016/J.RINENG.2025.105976, 2025.

Dembele, M., Schaefli, B., and Van De Giesen, N.: Suitability of 17 gridded rainfall and temperature datasets for large-scale hydrological modelling in West Africa, Hydrol. Earth Syst. Sci., 24, 5379–5406, https://doi.org/10.5194/HESS-24-5379-2020, 2020.

Diop, S. B., Ekolu, J., Tramblay, Y., Dieppois, B., Grimaldi, S., Bodian, A., Blanchet, J., Rameshwaran, P., Salamon, P., and Sultan, B.: Climate change impacts on floods in West Africa: New insight from two large-scale hydrological models, Natural Hazards and Earth System Sciences, 25, 3161–3184, https://doi.org/10.5194/NHESS-25-3161-2025, 2025.

Fang, Z., Wang, Y., Peng, L., and Hong, H.: Predicting flood susceptibility using LSTM neural networks, J. Hydrol. (Amst)., 594, 125734, https://doi.org/10.1016/J.JHYDROL.2020.125734, 2021.

Fathi, M. M., Al Mehedi, M. A., Smith, V., Fernandes, A. M., Hren, M. T., and Terry, D. O.: Evaluation of LSTM vs. conceptual models for hourly rainfall runoff simulations with varied training period lengths, Scientific Reports 2025 15:1, 15, 15820-, https://doi.org/10.1038/s41598-025-96577-4, 2025.

Feng, D., Lawson, K., and Shen, C.: Mitigating Prediction Error of Deep Learning Streamflow Models in Large Data-Sparse Regions With Ensemble Modeling and Soft Data, Geophys. Res. Lett., 48, e2021GL092999, https://doi.org/10.1029/2021GL092999, 2021.

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shelev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S.: Deep learning rainfall-runoff predictions of extreme events, Hydrol. Earth Syst. Sci., 26, 3377–3392, https://doi.org/10.5194/HESS-26-3377-2022, 2022a.

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shelev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S.: Deep learning rainfall-runoff predictions of extreme events, Hydrol. Earth Syst. Sci., 26, 3377–3392, https://doi.org/10.5194/HESS-26-3377-2022, 2022b.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol. (Amst)., 377, 80–91, https://doi.org/10.1016/J.JHYDROL.2009.08.003, 2009.

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

Hodson, T. O.: Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not, Geosci. Model Dev., 15, 5481–5487, https://doi.org/10.5194/GMD-15-5481-2022, 2022.

Houngue, R.: Climate Change Impacts on Hydrodynamic Functioning of Oueme Delta (Benin), 2020.

Hounkpè, J., Diekkrüger, B., Badou, D. F., and Afouda, A. A.: Change in Heavy Rainfall Characteristics over the Ouémé River Basin, Benin Republic, West Africa, Climate 2016, Vol. 4, 4, https://doi.org/10.3390/CLI4010015, 2016.

Hounkpè, J., Diekkrüger, B., Afouda, A. A., and Sintondji, L. O. C.: Land use change increases flood hazard: a multi-modelling approach to assess change in flood characteristics driven by socio-economic land use change scenarios, Natural Hazards, 98, 1021–1050, https://doi.org/10.1007/S11069-018-3557-8/FIGURES/10, 2019.

Janjić, J. and Tadić, L.: Fields of Application of SWAT Hydrological Model—A Review, Earth 2023, Vol. 4, Pages 331-344, 4, 331–344, https://doi.org/10.3390/earth4020018, 2023.

Japhet Kodja, D., Mahé, G., Amoussou, E., Boko, M., Paturel Assessment, J., and Emmanuel Paturel, J.: Assessment of the Performance of Rainfall-Runoff Model GR4J to Simulate Streamflow in Ouémé Watershed at Bonou's outlet (West Africa, https://doi.org/10.20944/preprints201803.0090.v1, 2018.

Klutse, N. A. B., Quagraine, K. A., Nkrumah, F., Quagraine, K. T., Berkoh-Oforiwaa, R., Dzrobi, J. F., and Sylla, M. B.: The Climatic Analysis of Summer Monsoon Extreme Precipitation Events over West Africa in CMIP6 Simulations, Earth Systems and Environment 2021 5:1, 5, 25–41, https://doi.org/10.1007/s41748-021-00203-y, 2021.

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

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, Water Resour. Res., 55, 11344–11354, https://doi.org/10.1029/2019WR026065, 2019.

https://doi.org/10.5194/egusphere-2026-4749
Preprint. Discussion started: 28 August 2026

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., and Matias, Y.: Caravan - A global community dataset for large-sample hydrology, Scientific Data 2023 10:1, 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, Hydrol. Earth Syst. Sci, 28, 4187–4201, 2024.

Kratzert, F., Gauch, M., Nearing, G., and Klotz, D.: NeuralHydrology — A Python library for Deep Learning research in hydrology, J. Open Source Softw., https://doi.org/10.21105/joss.04050, n.d.

Kreibich, H., Hudson, P., Merz, B., Kreibich, H., Hudson, P., and Merz, B.: Knowing What to Do Substantially Improves the Effectiveness of Flood Early Warning, Bull. Am. Meteorol. Soc., 102, E1450–E1463, https://doi.org/10.1175/BAMS-D-20-0262.1, 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, Hydrol. Earth Syst. Sci., 26, 3079–3101, https://doi.org/10.5194/HESS-26-3079-2022, 2022.

Li, X., Xu, W., Ren, M., Jiang, Y., and Fu, G.: Hybrid CNN-LSTM models for river flow prediction, iwaponline.comX Li, W Xu, M Ren, Y Jiang, G FuWater Supply, 2022•iwaponline.com, 22, 4902–4920, https://doi.org/10.2166/WS.2022.170, 2022.

Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H., Tan, F., and Thieme, M.: Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution, Scientific Data 2019 6:1, 6, 283-, https://doi.org/10.1038/s41597-019-0300-6, 2019.

Man, Y., Yang, Q., Shao, J., Wang, G., Bai, L., and Xue, Y.: Enhanced LSTM Model for Daily Runoff Prediction in the Upper Huai River Basin, China, Engineering, 24, 229–238, https://doi.org/10.1016/J.ENG.2021.12.022, 2023.

Martel, J. L., Arsenault, R., Turcotte, R., Castañeda-Gonzalez, M., Brissette, F., Armstrong, W., Mailhot, E., Pelletier-Dumont, J., Lachance-Cloutier, S., Rondeau-Genesse, G., and Caron, L. P.: Exploring the ability of LSTM-based hydrological models to simulate streamflow time series for flood frequency analysis, Hydrol. Earth Syst. Sci., 29, 4951–4968, https://doi.org/10.5194/HESS-29-4951-2025, 2025.

Mohammadi, M., Gato-Trinidad, S., and Kuok, K. K.: The limitation of machine learning methods for water supply and demand forecasting: A case study for Greater Melbourne, Australia, Water Supply, 24, 3848–3861, https://doi.org/10.2166/WS.2024.225, 2024.

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 Syst. Sci. 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, J. Hydrol. (Amst)., 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., 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 2024 627:8004, 627, 559–563, https://doi.org/10.1038/s41586-024-07145-1, 2024.

Ougahi, J. H. and Rowan, J. S.: Investigating Deep Learning Knowledge Transfer in Streamflow Prediction From Global to Local Catchment, Water Resour. Res., 62, e2025WR041194, https://doi.org/10.1029/2025WR041194;WGROUP:STRING:PUBLICATION, 2026.

Quenum, G. M. L. D., Nkrumah, F., Klutse, N. A. B., and Sylla, M. B.: Spatiotemporal Changes in Temperature and Precipitation in West Africa. Part I: Analysis with the CMIP6 Historical Dataset, Water 2021, Vol. 13, 13, https://doi.org/10.3390/w13243506, 2021.

Rossi, E., Maiorano, A., and Gobbi, A.: Update of the Köppen–Geiger climate classification for EFSA PLH risk assessment, EFSA Journal, 24, e9859, https://doi.org/10.2903/J.EFSA.2026.9859;SUBPAGE:STRING:FULL, 2026.

Silva, A. F. R., Eleutério, J. C., Apel, H., and Kreibich, H.: Assessing the impact of early warning and evacuation on human losses during the 2021 Ahr Valley flood in Germany using agent-based modelling, Natural Hazards and Earth System Sciences, 25, 1501–1520, https://doi.org/10.5194/NHESS-25-1501-2025, 2025.

Snieder, E., Abogadil, K., and Khan, U. T.: Resampling and ensemble techniques for improving ANN-based high-flow forecast accuracy, Hydrol. Earth Syst. Sci., 25, 2543–2566, https://doi.org/10.5194/HESS-25-2543-2021, 2021.

Song, Y., Hong, J., Liu, F., Liu, J., Chen, Y., Li, Z., Su, J., Hu, S., and Fu, J.: Deep Learning-Assisted Diagnosis of Malignant Cerebral Edema Following Endovascular Thrombectomy, Acad. Radiol., 32, 3237–3247, https://doi.org/10.1016/J.ACRA.2025.02.021, 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 Resour. Res., 62, https://doi.org/10.1029/2025WR040414, 2026.

Sorini, G., Blanchet, J., Panthou, G., Vischel, T., and Tramblay, Y.: Typology of flood trends across West Africa, EGUsphere, 1–26, https://doi.org/10.5194/EGUSPHERE-2026-1176, 2026.

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, Frontiers in Water, 5, https://doi.org/10.3389/frwa.2023.1166124, 2023.

Xu, Y., Lin, K., Hu, C., Wang, S., Wu, Q., Zhang, L., and Ran, G.: Deep transfer learning based on transformer for flood forecasting in data-sparse basins, J. Hydrol. (Amst)., 625, 129956, https://doi.org/10.1016/J.JHYDROL.2023.129956, 2023.

Yang, S., Zhang, Y., and Zhang, Z.: Runoff Prediction Based on Dynamic Spatiotemporal Graph Neural Network, Water 2023, Vol. 15, 15, https://doi.org/10.3390/W15132463, 2023.

Yao, S., Kang, Q., Zhou, M. C., Rawa, M. J., and Abusorrah, A.: A survey of transfer learning for machinery diagnostics and prognostics, Artificial Intelligence Review 2022 56:4, 56, 2871–2922, https://doi.org/10.1007/S10462-022-10230-4, 2022.