AGU ADVANCING EARTH AND SPACE SCIENCES

Water Resources Research

RESEARCH ARTICLE
10.1029/2023WR035876

Deciphering the Mechanism of Better Predictions of Regional LSTM Models in Ungauged Basins

Qiang Yu[1,2], Liguang Jiang[1,3], Raphael Schneider[4], Yi Zheng[1], and Junguo Liu[1,5]

[1]School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen, China, [2]Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, China, [3]Shenzhen Key Laboratory of Precision Measurement and Early Warning Technology for Urban Environmental HealthRisks, School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen, China, [4]Department of Hydrology, Geologic Survey of Denmark and Greenland (GEUS), Copenhagen, Denmark, [5]Henan Provincial Key Laboratory of Hydrosphere and Watershed Water Security, North China University of Water Resources and Electric Power, Zhengzhou, China

Key Points:

- A classification-based training strategy is introduced for the regional long short-term memory (LSTM) model
- The influence of static attributes on the performance of the regional LSTM model in ungauged basins is investigated
- There is a high level of consistency in the enhancement achieved by the two training strategies, either incorporated with static catchment attributes or based on classification

Correspondence to:

L. Jiang,
jianglg@sustech.edu.cn

Citation:

Yu, Q., Jiang, L., Schneider, R., Zheng, Y., & Liu, J. (2024). Deciphering the mechanism of better predictions of regional LSTM models in ungauged basins. *Water Resources Research*, *60*, e2023WR035876. https://doi.org/10.1029/2023WR035876

Received 3 AUG 2023
Accepted 4 JUL 2024

Author Contributions:

Conceptualization: Liguang Jiang
Data curation: Qiang Yu
Formal analysis: Qiang Yu, Liguang Jiang
Investigation: Qiang Yu, Liguang Jiang
Methodology: Qiang Yu, Liguang Jiang
Software: Qiang Yu
Supervision: Liguang Jiang
Validation: Qiang Yu
Visualization: Qiang Yu, Liguang Jiang
Writing – original draft: Qiang Yu
Writing – review & editing: Liguang Jiang, Raphael Schneider, Yi Zheng, Junguo Liu

Abstract Prediction in ungauged basins (PUB) is a concerning hydrological challenge, prompting the development of various regionalization methods to improve prediction accuracy. The long short-term memory (LSTM) model has gained popularity in rainfall-runoff prediction in recent years and has proven applicable in PUB. Prior research indicates that incorporating static attributes in the training of regional LSTM models could improve performance in PUB. However, the underlying reasons for this enhancement have received limited exploration. This study aims to explore the role of static attributes in the training of the regional LSTM model. It is assumed that the regional LSTM model can induce streamflow generation mechanisms with the incorporation of static attributes and apply certain streamflow generation mechanisms to ungauged catchments based on their attributes. To this end, a grouping-based training strategy is proposed, that is, training and validating regional LSTM models on catchments with similar streamflow generation mechanisms within predefined groups. The training strategies of regional LSTM models, either incorporated with static catchment attributes or based on classification, are conducted in 363 catchments. Results demonstrate a high level of consistency in the enhancement achieved by the two training strategies. Specifically, 192 and 216 catchments exhibit enhancement compared to traditionally trained models without inclusion of attributes, with 132 catchments showing improvement under both training strategies. Furthermore, the findings indicate consistent spatial patterns and attribute distributions of enhanced catchments, as well as the notable improvement in reproducing low flow-related hydrological signatures.

1. Introduction

Prediction in ungauged basins (PUB) has been a long-lasting challenge in hydrological fields over the past 2 decades, primarily due to the scarcity of in situ observation data (Guo et al., 2021; Hrachowitz et al., 2013). The main challenge in PUB lies in the transfer of hydrological information from gauged to ungauged catchments, considering the heterogeneity in climate, topography, and geology (Kratzert, Klotz, Shalev, et al., 2019; Nearing et al., 2021). Regionalization methods have been widely employed to address this problem (Guo et al., 2021), primarily by transferring hydrological model parameters from donor (gauged) to target (ungauged) catchments. Regionalization methods can be broadly categorized into three groups: similarity-based, hydrological signatures-based, and regression-based (Guo et al., 2021). Similarity-based methods regionalize by interpolating or averaging model parameters from donor to target catchments, with the selection of donor catchments based on the spatial distance or similarity of catchment attributes such as climatic, topographic, land cover, and soil characteristics. Hydrological signature-based methods involve directly regionalizing hydrological signatures (e.g., flow duration curve, FDC) from donor to target catchments (Atieh et al., 2017), or using hydrological signatures from donor catchments to constrain model parameters in target catchments (Pinheiro & Mauro Naghettini, 2013). Regression-based methods are based on the assumption that catchments with similar attributes exhibit similar hydrological behaviors (Blöschl et al., 2013). These methods establish regression equations to learn the relationship between catchment attributes and hydrological model parameters.

The previously mentioned regionalization methods primarily cater to traditional hydrological models (i.e., physically based and conceptual hydrological models), which rely on physically-informed parameters. In contrast, data-driven models, exemplified by machine learning (ML) models, do not rely on physically-informed

parameters. Notably, the long short-term memory (LSTM) neural network (Hochreiter & Schmidhuber, 1997), an ML model adept at handling time-series data, has been proven applicable in streamflow prediction. Additionally, the LSTM model shows promise as a regional model. In experiments by Kratzert et al. (2018), regional LSTM models trained on hydrological units (HUCs) outperformed the single-catchment trained LSTM models in around 50% of the catchments, using solely meteorological forcing data as model input. Regarding different validation types of regional LSTM models, "in-sample" validation is used here to refer to the regional LSTM models validated on catchments within the training samples, while "out-of-sample" validation refers to the regional LSTM models validated on catchments outside the training samples (i.e., PUB). Previous studies have also confirmed the applicability of the regional LSTM model in PUB. In a comparative analysis by Arsenault et al. (2023), utilizing identical meteorological forcing data and catchment attributes, the LSTM model exhibited superior performance to traditional regionalization methods in over 90% of the catchments. Kratzert, Klotz, Herrnegger, et al. (2019) conducted a k-fold spatial cross-validation ($k = 12$) on 531 catchments and discovered the median Nash-Sutcliffe efficiency (NSE) value of the regional LSTM model exceeded that of a locally trained hydrological model.

Furthermore, with the static catchment attributes as additional input in model training, the regional LSTM model exhibits substantial enhancement in both "in-sample" and "out-of-sample" applications (Feng et al., 2021; Hashemi et al., 2022; Kratzert, Klotz, Shalev, et al., 2019). The reasons behind the improvement of regional LSTM models with the addition of static attributes have been discussed mainly in studies about "in-sample" regional LSTM models. Gauch et al. (2021) reckoned that the LSTM model is able to deduce the relationship between catchment attributes and streamflow patterns. In order to explore the rationale of this phenomenon, Kratzert, Klotz, Shalev, et al. (2019) introduced a novel LSTM cell structure with a separate gate for incorporating static attributes, termed Entity-Aware-LSTM (EA-LSTM). There were found significant variance reductions in hydrological signatures due to clustering based on the gate vector. Kratzert, Klotz, Shalev, et al. (2019) inferred that catchment attributes contain sufficient information to distinguish between diverse rainfall-runoff behaviors. Hashemi et al. (2022) classified catchments into groups that were characterized by homogeneous hydrological regimes based on three hydrological indices, and compared the performance of models trained on catchments within the same hydrological regime with models trained on the complete national catchment data set. The results demonstrated a slight improvement in the homogeneous regime-level training strategy compared to the heterogeneous national-based training strategy. Hashemi et al. (2022) suggested that the model trained on heterogeneous catchments could extract classifications based on static attributes.

However, few studies discussed the reasons for the enhancement of regional LSTM models in "out-of-sample" or PUB scenarios. Based on the insights gained from previous studies, it is assumed in this study that the regional LSTM model has the capability to induce streamflow generation mechanisms with the incorporation of static attributes and to apply certain streamflow generation mechanisms to ungauged catchments based on their attributes. Therefore, the overarching goal of this study is to explore the reasons behind the enhanced performance of regional LSTM models in ungauged catchments with the incorporation of static catchment attributes.

The paper follows this structure: Section 2 (Methodology) details the data set, training strategies for the regional LSTM model, grouping methods, and the experimental design. Section 3 (Results) shows the model performance and pattern analysis. Section 4 (Discussion) assesses the distinctiveness of grouping and the consistent performance enhancement across different model training strategies. Section 5 (Conclusion and Perspectives) summarizes the main findings and outlines potential future research directions.

2. Methodology

2.1. Data Set

The catchments and data used in this study were derived from the Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) data set (Addor et al., 2017; A. J. Newman et al., 2015). The CAMELS data set includes daily time series of meteorological forcing data and streamflow observations from 671 catchments. The meteorological forcing data include precipitation, shortwave downward radiation, maximum and minimum temperature, and vapor pressure, all of which serve as input data for training the regional LSTM models.

2.2. LSTM Neural Network and Training Strategies

The LSTM neural network, proposed by Hochreiter and Schmidhuber (1997), is a special type of recurrent neural network (RNN). By replacing the traditional RNN unit with a memory block, the LSTM neural network is able to overcome the weakness of gradient vanishing that troubles traditional RNNs. Its remarkable performance in time series prediction has led to increased adoption of LSTM in hydrology, particularly for rainfall-runoff prediction (Kraft et al., 2022; Kratzert et al., 2018; Tsai et al., 2021).

As mentioned earlier, the essence of PUB is to transfer hydrological information from gauged to ungauged catchments. Hence PUB is widely recognized as an issue of "extrapolation", considering the heterogeneity in climate, topography, and geology (Nearing et al., 2021). However, the difficulty of PUB might be influenced by the presence of adjacent catchments among the donor (gauged) and target (ungauged) catchments. For instance, Bjerre et al. (2022) evaluated model performance in PUB using both spatial cross-validations without adjacency in donor and target catchments and random split-sample validation. Their findings reveal a significant decrease in model performance on the spatial cross-validation (R^2 ranging from 0.13 to 0.61) compared to the random split-sample validation ($R^2 = 0.79$). The experimental design of random sampling is likely to generate spatially adjacent catchments in the training and validation data sets with similar characteristics, thereby facilitating the transfer of hydrological information from gauged to ungauged basins. Therefore, PUB with adjacent basins can be described as "interpolation" (Figure 2a), and the concept of prediction in ungauged regions (PUR) without adjacency in donor and target catchments is recognized as "extrapolation" (Figure 2b) (Feng et al., 2021, 2023). Unsurprisingly, previous studies have consistently shown the somewhat inferior performance of machine learning models in PUR compared to PUB within the same domain (Bjerre et al., 2022; Feng et al., 2021; Yin et al., 2023). Moreover, it is worth noting that catchments from the training data set are typically not evenly distributed across the entire domain (Meyer & Pebesma, 2021). Consequently, The utilization of PUR is a more rigorous assessment of model performance in ungauged catchments.

In this study, the contiguous United States (CONUS) was divided into seven regions, following the methodology of Feng et al. (2021) as depicted in Figure 1a. Each model was trained using data from catchments in six out of the seven regions and evaluated on the remaining holdout region. The traditional training strategy, as outlined by Feng et al. (2021), involved training a single LSTM model using data from all available donor catchments (Figure 2c). Given the black-box nature of the LSTM model, directly verifying the aforementioned assumption could be challenging. Thus, a grouping-based training strategy is proposed to ensure that both donor and target catchments share similar streamflow generation mechanisms. Implementing this training strategy involves partitioning the catchments within the study area into internally similar groups. Each group of catchments in the target region was assigned a specific LSTM model and trained using catchments with the same group from the donor region. For instance, if the target region comprises three groups of catchments (A, B, C), then three models would be trained using similar catchments in the donor region from groups A, B, and C, respectively (Figure 2d). Suppose the performance of the regional LSTM model trained by the traditional strategy with the incorporation of catchment attributes is consistent with the model trained by the grouping-based strategy. In that case, the assumption that the regional LSTM model is capable of inducing streamflow generation mechanisms with the incorporation of catchment attributes can be laterally verified.

2.3. Catchment Grouping

To facilitate the aforementioned grouping-based LSTM model training strategy, catchments were classified with similar streamflow generation mechanisms, following the classification results of Wu et al. (2021a). In this classification, catchments were grouped into eight classes by the fuzzy c-means method based on the similarity of six hydrological signatures. The six hydrological signatures used for catchment classification included (a–c) the Spearman correlation coefficients between event runoff coefficients (RC) and event rainfall intensity (Pint), event rainfall volume (Pvol), and pre-event storage (S), respectively; (d) the characteristic time scale of event runoff response, estimated using a linear-reservoir-based net-rainfall-runoff model (TS); (e) the ratio between base flow and total streamflow (BFI) and (f) the standard deviation of log-scale base flow time series (VARb).

The initial CAMELS data set contains 671 catchments with areas ranging from 4 to 25,000 km^2. Newman et al. (2017) removed catchments (a) larger than 2,000 km^2 and (b) with large discrepancies between different methods of calculating catchment area, thus 531 catchments remain. In the classification of Wu et al. (2021), a smaller subset of 432 catchments was selected based on three rules: (a) the catchment area must be in the range

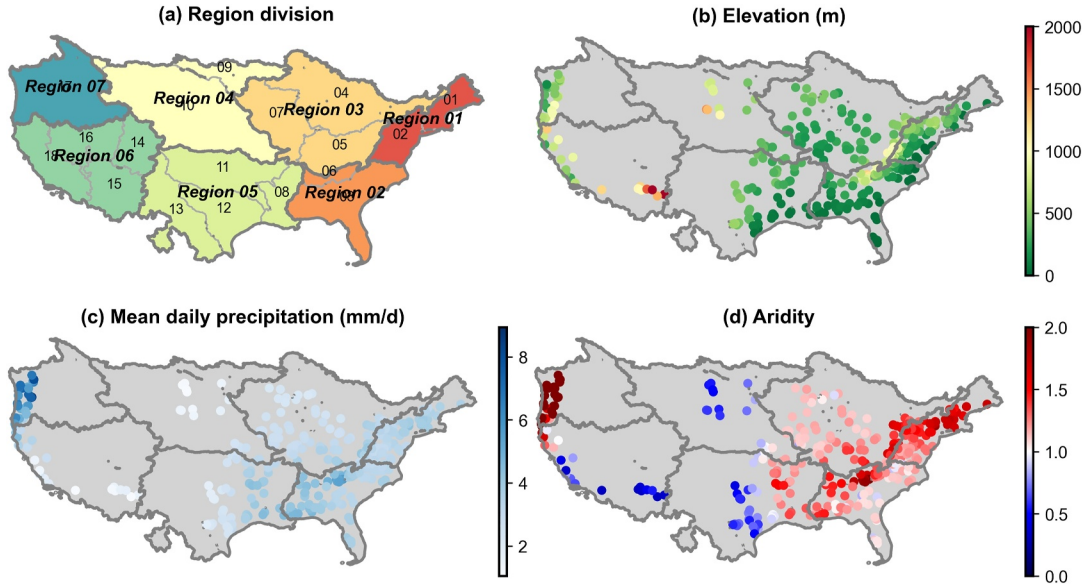

Figure 1. Information and distribution of the 363 CAMELS catchments (represented by catchment centroids here) used in this study. (a) The divided seven regions comprising hydrologic units (polygons with thinner boundaries, labeled with two-digit numbers), (b) the mean elevation, (c) the mean daily precipitation, and (d) the aridity index (PET/P).

from 20 km^2 to 10,000 km^2, (b) there must be less than 30% of precipitation in the catchment that falls as snow, (c) NSE of streamflow predictions by the coupled Snow-17 model and the SAC-SMA model must be at least 0.5 in the catchment (during the period from 1997 to 2014). Based on the selections of both Newman et al. (2017) and Wu et al. (2021), 363 catchments were selected in this study, and their distribution, climatic conditions are shown in Figures 1a–1c. Based on the six hydrological signatures, Wu et al. (2021) conducted catchment classification using the fuzzy c-means algorithm. The 363 catchments from the CAMELS data set were grouped into several classes, each characterized by one or two dominant streamflow generation mechanisms. The catchment classification result is shown in Figure 3a.

The first category of streamflow generation mechanism corresponds to infiltration excess overland flow (hereinafter also referred to as Horton overland flow, HOF). The second category includes two specific mechanisms: saturation excess overland flow (SOF) and subsurface flow (referred to as SSF1). The third category refers to lateral preferential flow (referred to as SSF2) and the fourth category is the delayed-response flow from groundwater storage (referred to as GWF). The six hydrological signatures and the dominant streamflow generation mechanisms of each class are shown in Table 1.

However, the hydrological signatures used in Wu's classification were derived from historical streamflow records, which are often unavailable in ungauged catchments. In addition, can the regional LSTM models benefit from groups partitioned by static attributes? To investigate the feasibility of achieving similar outcomes, a clustering method was applied using the same static attributes as input in the traditional LSTM model training strategy. Based on the 24 static attributes related to topography, climate, land cover, and soil outlined in Table A1, a k-means clustering analysis was conducted to categorize the 363 catchments in the study area into five clusters (Figure 3b).

2.4. Experimental Design

Four sets of experiments were conducted to test the assumption in this study, and the description of models in each experiment is outlined in Table 2. The regional LSTM models employing the traditional training method are denoted as TRA-models ("TRA" for "traditional"). These are further categorized into TRA-NA ("NA" for "no attributes") and TRA-WA ("WA" for "with attributes") models based on the inclusion or exclusion of static attributes as model input. The model using the classification-based training strategy is called CLA-NA ("CLA"

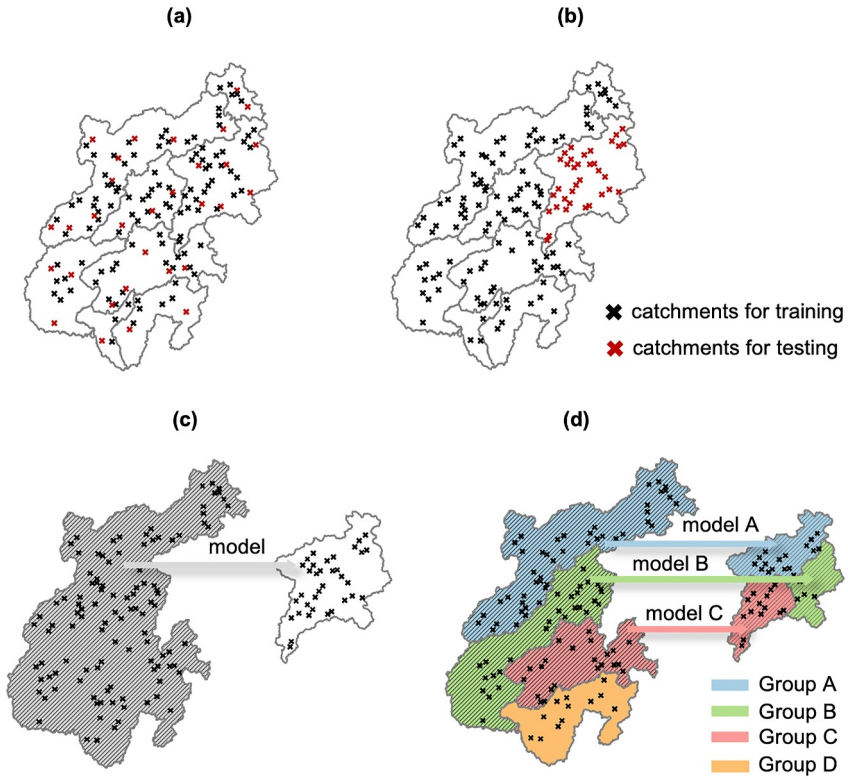

Figure 2. An illustration of the difference between PUB (a) and PUR (b) and the difference between traditional training strategy (c) and group-based training strategy (d). Note that the shaded area in (c–d) indicates catchments required for model training.

for "classification"), and the model using the cluster-based training strategy is labeled as CLU-NA ("CLU" for "cluster").

Since the primary distinction between the TRA-WA and TRA-NA models lies in the inclusion of static catchment attributes as model input, the role of these attributes in the regional LSTM model training can be, therefore, analyzed by comparing the performance differences between the two models. The CLA-NA and TRA-NA models have the same input data, and the difference lies in the fact that the donor and target catchments in the CLA-NA model come from the same class with similar streamflow generation mechanisms. By comparing the consistency of performance between TRA-WA and CLA-NA models as well as the consistency in the enhancement of these two models relative to TRA-NA, it can be determined whether the incorporation of static attributes achieves consistent effectiveness compared to the classification of streamflow generation mechanisms. In addition, evaluating the performance of the CLU-NA model enables an assessment of the effectiveness of the attribute-based clustering.

In this study, the open-source code developed by Kratzert, Klotz, Shalev, et al. (2019) was employed to train the LSTM models. Each model consisted of a single LSTM layer with 256 hidden nodes, and the sequence length was

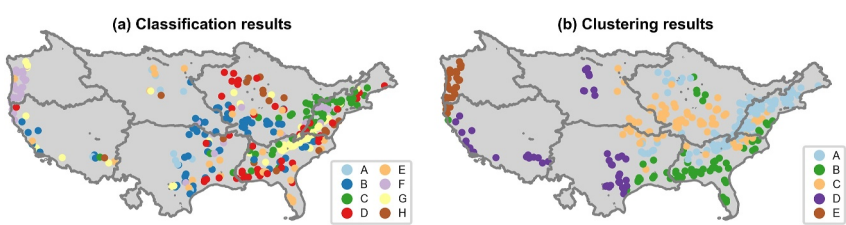

Figure 3. (a) Catchment classification results based on streamflow generation mechanism and (b) catchment clustering results based on k-means clustering analysis.

Table 1
The Qualitative Descriptions of Six Hydrological Signatures and the Inferred Dominant Streamflow Generation Mechanisms for Each Class (Wu et al., 2021a)

	Streamflow signatures				Base flow signatures		Mechanism			
		RC								
Class	Pint	Pvol	S	TS	BFI	VARb	HOF	SOF/SSF1	SSF2	GWF
A				Small	Small	Large	○			
B			○	Small	Small	Large		○		
C			○	Small	Medium	Medium		○		○
D			○	Large	Medium	Medium				○
E			○	Large	Medium	Large				○
F		○		Medium	Medium	Medium			○	○
G		○		Small	Large	Small			○	○
H				Large	Large	Small				○

Note. (Note that the circle under "RC" indicates the correlation coefficient between the signature and RC is larger than 0.5, and the correlation is the largest among the three streamflow signatures. The circle under "Mechanism" means the mechanism is inferred as the dominant mechanism of the class).

set as 270 days. The static attributes are repeated over time and then concatenated with the meteorological forcing input at each timestep. In addition, the prediction results of three lumped conceptual hydrological models (SAC-SMA, VIC, and HBV) were used as benchmarks. The SAC-SMA and VIC models were locally calibrated (Kratzert, Klotz, Shalev, et al., 2019; A. J. Newman et al., 2017), while the HBV model employed globally regionalized parameters from Beck et al. (2020). In this study, the predictions provided by the regional HBV model were considered as PUB, as the parameter transfer function in Beck et al. (2020) was calibrated using globally sampled catchments that partially intersected with our study domain. Considering model uncertainty, five iterations of training and testing were conducted for each LSTM model in this study, and the median values of these five predictions were used to represent the model's performance. The training and testing periods aligned with those of Kratzert, Klotz, Shalev, et al. (2019), covering water years from 2000 to 2008 and 1990 to 1999, respectively. The NSE coefficient (NSE) was employed as the evaluation metric to assess the performance of the aforementioned models.

3. Results

The NSE values for all seven models (SAC-SMA, VIC, HBV, TRA-NA, TRA-WA, CLA-NA, and CLU-NA) were computed across the 363 catchments from the CAMELS data set. The median NSE values of the TRA-NA, TRA-WA, CLA-NA, and CLU-NA models are 0.568, 0.599, 0.616, and 0.570, respectively. The performance of the four LSTM models in PUR is notably better than that of the regional HBV model (0.465) in PUB; such performance is comparable to that of the locally calibrated models, namely SAC-SMA (0.606) and VIC (0.553). However, the box lengths of the locally calibrated models are considerably shorter than those of the LSTM

Table 2
Description of Experimental Design

Model name	Description
TRA-NA	Using the traditional training strategy, without the addition of static catchment attributes
TRA-WA	Using the traditional training strategy, with the addition of static catchment attributes
CLA-NA	Using a classification-based training strategy, without the addition of static catchment attributes
CLU-NA	Using a cluster-based training strategy, without the addition of static catchment attributes

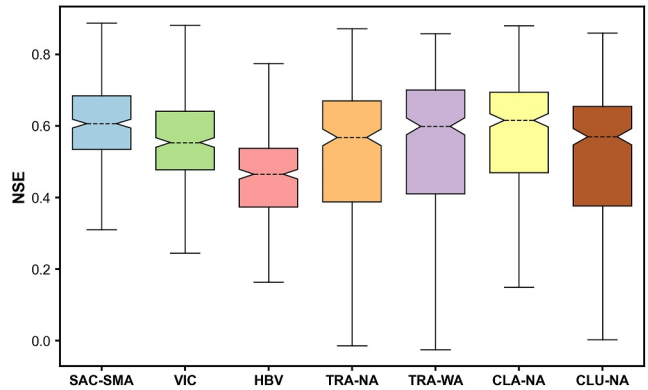

Figure 4. Nash-Sutcliffe efficiency values of models on 363 catchments from the CAMELS data set (Note: the outliers are not marked in this figure).

models. Additionally, the lower whiskers and lower quartiles of the locally calibrated models exceed those of the LSTM models (Figure 4). These results indicate that the performance of regional LSTM models in PUR is not yet robust, which is consistent with the findings of Feng et al. (2021). In the evaluation of different regional LSTM models, the performance of TRA-WA and CLA-NA models is enhanced in PUR compared to the TRA-NA model, with CLA-NA demonstrating a slightly greater enhancement. However, the performance of the CLU-NA model exhibited marginal improvement relative to the TRA-NA model.

To elucidate the geographical distribution of catchments where the TRA-WA and CLA-NA models outperform the TRA-NA model, scatter plots depicting the NSE differences of the TRA-WA and CLA-NA models against the TRA-NA model for each catchment are presented in Figure 5. A positive NSE difference indicates improved performance by the TRA-WA and CLA-NA models, with darker colors denoting larger NSE differences. The median NSE differences between the two models and the TRA-NA model are 0.009 and 0.020, respectively. The spatial patterns of catchments manifesting enhanced NSE values for both the TRA-WA and CLA-NA models relative to the TRA-NA model demonstrate similarities. Statistically, 192 and 216 catchments exhibit enhancement in NSE with TRA-WA and CLA-NA models in comparison to the TRA-NA model. Additionally, 132 catchments demonstrate enhancement with both TRA-WA and CLA-NA models.

For further analysis of the characterization of catchments enhanced by the TRA-WA and CLA-NA models relative to the TRA-NA model, catchments were categorized into two groups based on whether the NSE enhancement of these two models relative to the TRA-NA model is remarkable (NSE difference exceeds 0.05). Eight catchment attributes were chosen from the total of 24 employed in the TRA-WA model to analyze the climatic, soil, and geological characteristics in catchments with or without remarkable enhancement. The

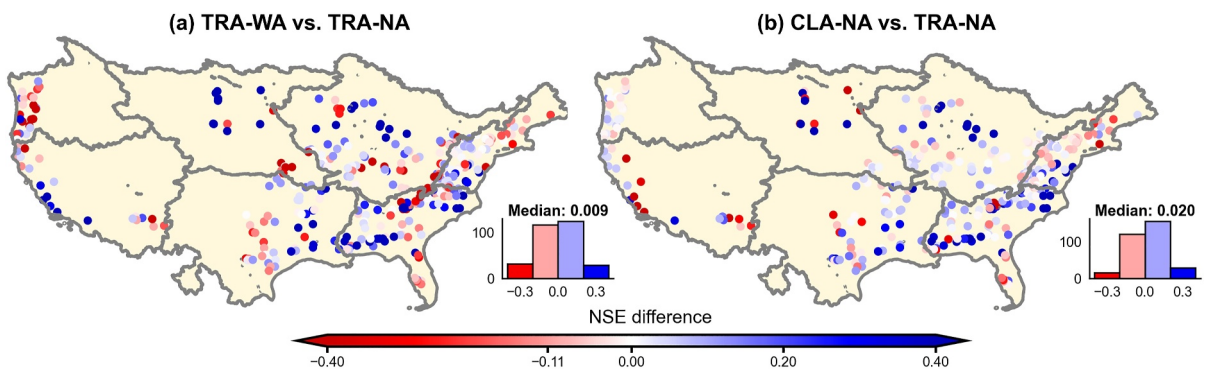

Figure 5. The Nash-Sutcliffe efficiency (NSE) difference (a) between the TRA-WA and TRA-NA models, and (b) between the CLA-NA and TRA-NA models in the validation period. (Note: Blue colors (NSE difference >0) indicate that the TRA-WA/CLA-NA model outperforms TRA-NA, while red colors (NSE difference <0) indicate the opposite; the color shade reflects the magnitude of the NSE difference, that is, the darker the color, the greater the NSE difference).

10.1029/2023WR035876

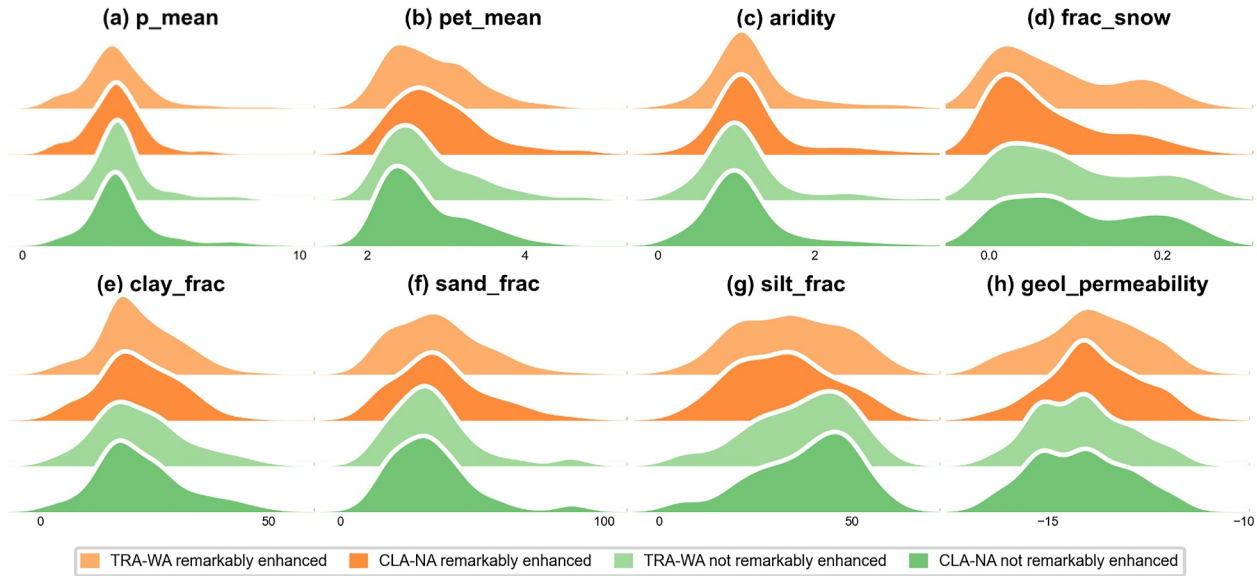

Figure 6. Distribution of attributes for catchments where TRA-WA and CLA-NA models manifest remarkable (>0.05) and non-remarkable (<0.05) Nash-Sutcliffe efficiency enhancement compared to the TRA-NA model.

distributions of 8 representative attributes are depicted in Figure 6. This representation highlights catchments that exhibited remarkable enhancement compared to those that did not. The shapes of the attribute distributions in both remarkably enhanced (in orange) and non-remarkably enhanced (in green) catchments were similar between the two models. Figures 6a–6d illustrate the distribution of four climate indices, revealing that the mean daily potential evapotranspiration (PET) distributions in remarkably enhanced catchments are consistent and broader in comparison to those without remarkable enhancement. This pattern contrasts with the distributions of the fraction of precipitation falling as snow. Figures 6e–6h depict the distribution of four soil and geological characteristics, emphasizing the consistency of soil compositions and permeability in enhanced catchments, which diverge from those in catchments without remarkable enhancement.

4. Discussion

4.1. Distinctiveness in Catchment Groupings

As depicted in Figure 4, it is evident that the CLA-NA model exhibits a notable enhancement in PUR performance compared to the TRA-NA model, while CLU-NA does not. In order to analyze the factors contributing to the dissimilarities in outcomes between these two catchment grouping methods, the distinctiveness of the grouping by the two methods was assessed. High distinctiveness is inferred if a model trained on a specific group of catchments demonstrates significantly superior performance within the same group when compared to others. For this analysis, two five-fold cross-validation experiments were executed using LSTM models trained solely on meteorological forcing data. Catchments within each group were randomly divided into five subsets, and during each validation, models were trained on four subsets of the group (donor class/cluster) and validated on the remaining subset, as well as the catchments in other groups (target classes/clusters). NSE values for the five validations were computed, and the median NSE values were considered as the final results, illustrated in a heatmap (Figure 7). In essence, if the NSE value on the diagonal is markedly higher than other values in the same row, it indicates a group with high distinctiveness.

As illustrated in Figure 7a, among the eight classes proposed by Wu et al., models trained on five of these catchments (D, E, F, G, H) exhibited optimal validation within the same class, with the NSEs for the remaining classes (A, B, C, D) not deviating from the optimal NSEs by more than 0.05. It is worth noting that models trained in class A displayed poor performance in all other classes, and conversely, models trained in other classes exhibited suboptimal validation in class A. Two plausible explanations for this outcome are the limited number of catchments (i.e., 10) in class A, predominantly situated in arid regions, and substantial

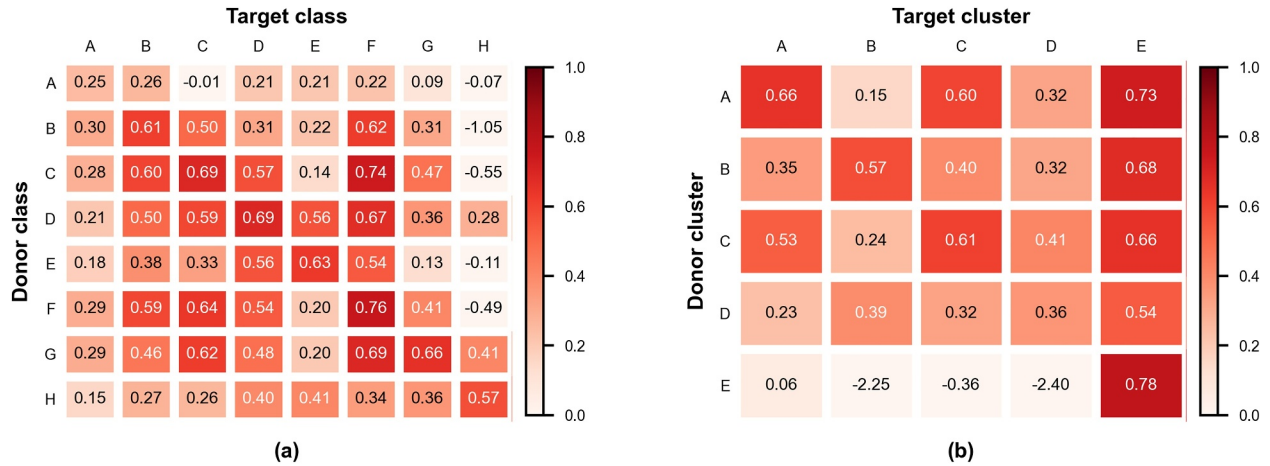

Figure 7. Nash-Sutcliffe efficiency (NSE) values of k-fold cross-validation based on (a) classification by Wu, (b) k-means clustering (Note: donor class means class used for model training while target class/cluster for model validation; therefore, NSE values in each row represent the performance of the other classes/clusters on the target class/cluster; NSE values in each column represent the performance of the donor class/cluster on the other classes/clusters).

differences in dominant streamflow generation mechanisms between class A catchments and those in other classes. In contrast, within the k-means clustering based on static catchment attributes, only cluster E demonstrated optimal NSEs verified within the same cluster, while the remaining optimal values were no less than 0.05 than the NSEs verified within the same cluster. Consequently, it can be inferred that groups classified based on hydrological signatures exhibit higher distinctiveness than those clustered based on static catchment attributes.

Nevertheless, there are some limitations to the classification according to Wu et al. For instance, there exist arid catchments dominated by other streamflow generation mechanisms like SOF, subsurface flow (SSF1), and/or groundwater storage (GWF) in the CAMELS data set. However, the samples of these catchments are insufficient to be divided into separate classes. As a consequence, these particular catchments may become anomalous samples in some classes. Another interesting finding is that models trained on class B, C, D, E, and F demonstrated median NSE values exceeding 0.5 on catchments not only within their own classes but also from other classes. Taking class D and E as examples, both classes exhibit high correlation coefficients between event runoff coefficients (RC) and pre-event storage (S), along with a large time scale of event runoff response (TS) and a medium ratio between base flow and total streamflow (BFI) (Table 1). The only distinction between the two classes in the six hydrological signatures for classification lies in the standard deviation of the log-scale base flow time series (VARb). Specifically speaking, catchments in class D have a medium VARb, whereas catchments in class E exhibit a large VARb. Consequently, catchments in both the two classes are dominated by groundwater flow (GWF) and lack dominant surface flow generation mechanisms (i.e., HOF, SOF/SSF1, and SSF2). This similarity in streamflow generation mechanisms between different classes accounts for the favorable performance of classes D and E on each other and lowers the distinctiveness between the two classes.

4.2. Consistency in Performance Enhancement

As previously mentioned, it is challenging to verify the assumption in this study directly. Observing the consistency in the performance of TRA-WA and CLA-NA models compared to the TRA-NA model is the most feasible way to support this assumption. In general, the outcomes revealed a consistent spatial distribution of enhanced catchments by the TRA-WA and CLA-NA models in comparison to the TRA-NA model (Figure 5). Moreover, the distributions of catchment attributes in the remarkably and non-remarkably enhanced catchments demonstrate a high degree of consistency between the two models (Figure 6).

Figure 8 depicts the NSE difference between TRA-WA and TRA-NA models in comparison to that between CLA-NA and TRA-NA models over all catchments and for each class of catchments. Notably, the correlation in

10.1029/2023WR035876

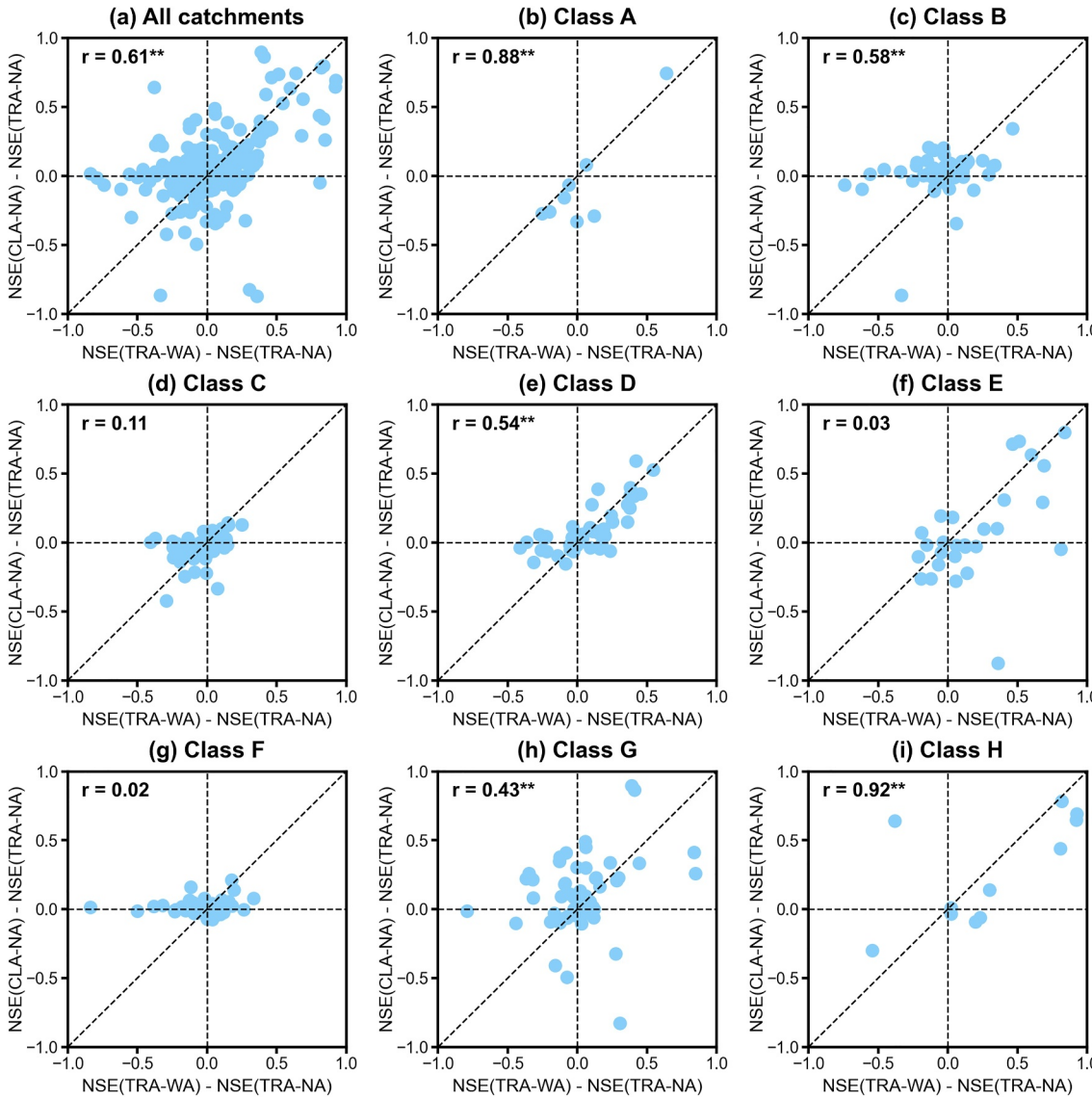

Figure 8. Nash-Sutcliffe efficiency difference between TRA-WA and TRA-NA models in comparison to that between CLA-NA and TRA-NA models over all catchments (a) and for each class of catchments (b–i).

NSE differences between the TRA-WA and TRA-NA models, as well as the NSE differences between the CLA-NA and TRA-NA models on all 363 catchments, is significant, with correlation coefficients of 0.61. For individual classes, excluding C, E, and F, the NSE differences between TRA-WA and TRA-NA models exhibit a noteworthy correlation with those of the CLA-NA model. Specifically, correlation coefficients surpass 0.5 for classes A, B, D, and H.

In order to evaluate the performance of the three regional LSTM models in PUR in depth, hydrological signatures of both simulated and observed streamflows (refer to Table 3) were computed. Additionally, to assess the effectiveness of the three models in reproducing hydrological signatures, the Pearson correlation coefficients between simulated and observed streamflow signatures were also calculated, as shown in Figure 9.

Figures 4–9a1 and 4–9d1 illustrate the performance of the three models in reproducing high flow-related signatures, specifically the 95th streamflow percentile (Q95) and high flow frequency (HFF). Consistently

Table 3
Description of Hydrological Signatures

Hydrological signature	Description	Unit
Q5	5-th streamflow percentile	mm/d
Q95	95-th streamflow percentile	mm/d
BFI	Baseflow index	-
HFF	High flow frequency	-
LFF	Low flow frequency	-
CV	Coefficient of variation	-
FI	Richards-Baker flashiness index	-

high correlation coefficients between the simulated and observed Q95 values (ranging from 0.92 to 0.94) across all three models manifest their capability to reproduce high flow volumes. Notably, the CLA-NA model outperforms in reproducing HFF, achieving a correlation coefficient of 0.66. In contrast, the TRA models exhibit lower correlation coefficients (0.28 and 0.34). It is significant to highlight that the TRA-WA model improves the underestimation of HFF for catchments in classes A and B observed in the TRA-NA model with the incorporation of static attributes. This improvement aligns with the performance of the classification-based CLA-NA model.

Regarding hydrological signatures characterizing low flows—specifically, the 5th streamflow percentile (Q5), baseflow index (BFI), and low flow frequency (LFF)—it is evident that the TRA-WA model and the CLA-NA model exhibit improved reproducibility compared to the TRA-NA model. Notably, the enhancement is particularly significant in the case of the CLA-NA model. As for the reproduction of BFI, the correlation coefficients show a progression from an insignificant 0.09 in TRA-NA to 0.27 in TRA-WA, culminating in a notable 0.73 in CLA-NA. Concurrently, the scatter plots gradually converge toward the 45-degree line. From the perspective of classification, the TRA-NA model's tendency to overestimate BFI in classes A and B and underestimate it in classes G and H is mitigated by both the inclusion of catchment attributes and the classification-based training strategy. This finding aligns with Feng et al. (2021), where the incorporation of the FDC signature improves the performance of LSTM in simulating low flows. Feng et al. (2021) inferred that LSTM appears to implicitly extract features from the FDC to emulate the BFI. Such extraction of FDC characteristics is fundamentally similar to our assumption of inducing streamflow generation mechanisms. This similarity explains why these two training strategies could enhance the reproduction of low flow-related signatures. The improvement in reproduction of the low flow-related signatures is beneficial in enhancing performance in groundwater-dominant catchments (e.g., class H) where the LSTM model demonstrates lower effectiveness (Yao et al., 2023). Similar enhancements are observed in the Richards-Baker flashiness index (FI), a signature quantifying the response time of runoff from the onset of a rainfall event to the return to base flow conditions, along with the coefficient of variation (CV) in streamflow.

5. Conclusion and Perspectives

Previous studies have confirmed that the incorporation of static catchment attributes could enhance the performance of regional LSTM models in streamflow prediction in ungauged catchments. However, the underlying reasons for this enhancement remain unclear. This study assumes that the regional LSTM model possesses the capability to induce streamflow generation mechanisms with the incorporation of static attributes and to apply certain streamflow generation mechanisms to catchments in ungauged catchments based on their attributes. To test this assumption, the performance of regional LSTM models trained either with the inclusion of static attributes or based on classification was assessed through two different training strategies across 363 catchments from the CAMELS data set. Notably, all experiments were conducted in the framework of PUR to ensure a rigorous evaluation of model performance. The key conclusions drawn from the findings of this study are outlined below:

10.1029/2023WR035876

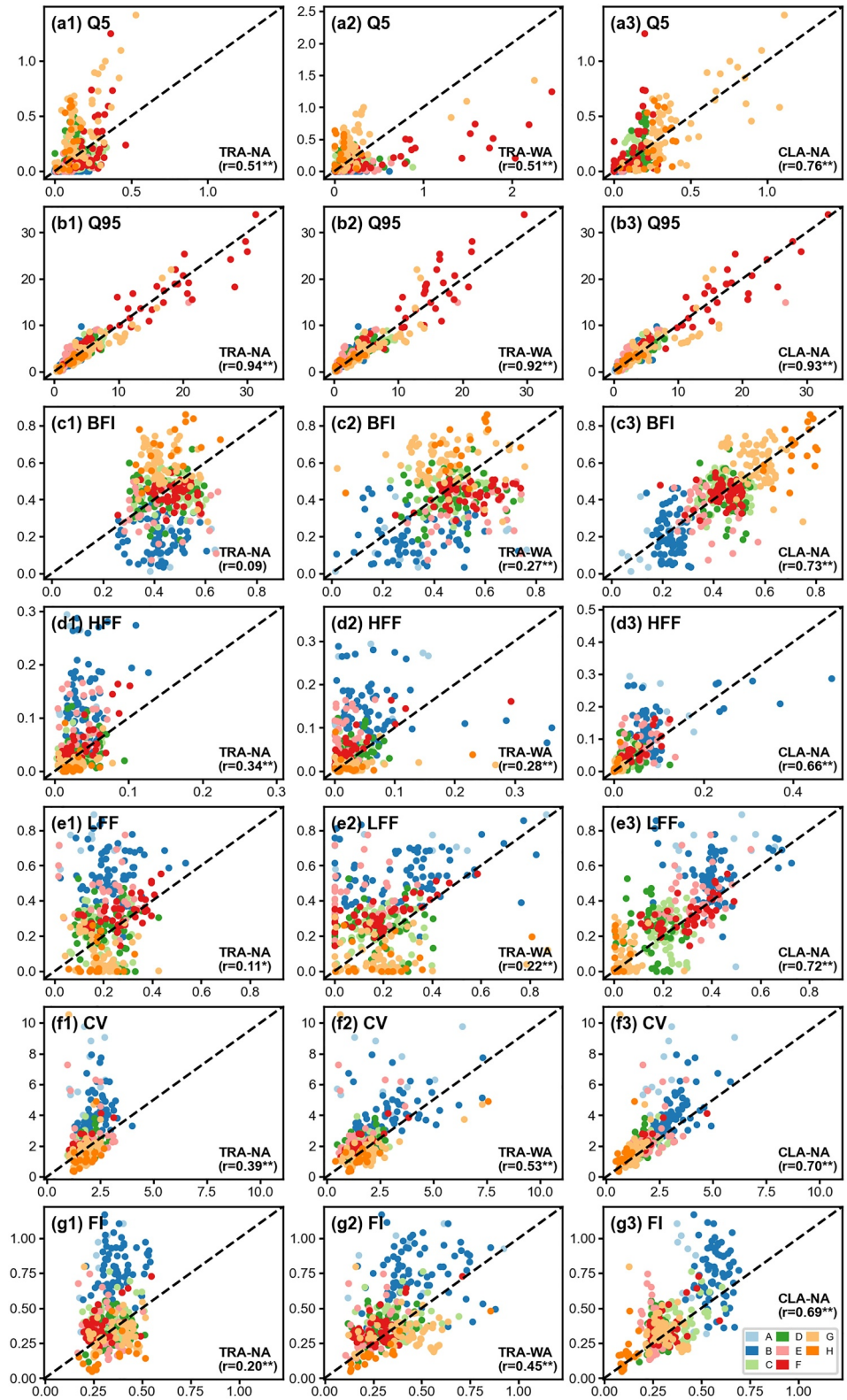

Figure 9. Scatter plots of hydrological signatures simulated by the TRA-NA, TRA-WA, and CLA-NA models versus observed signatures (Note: Plots a1–g1, a2–g2 and a3–g3 reflect the comparison of the TRA-NA, TRA-WA, and CLA-NA models in terms of the same signature; Units of different hydrological signatures please refer to Table 3; * and ** behind the correlation coefficients indicate the statistical significance at the level of 0.05 and 0.01, respectively).

The two regional LSTM model training strategies, involving the incorporation of static catchment attributes and classification-based training based on groups with similar streamflow generation mechanisms, have been demonstrated effective in enhancing model performance in ungauged catchments. The consistency of enhancement across both training strategies is evident, especially in the spatial pattern and attribute distributions of catchments with enhancement. In addition, there has been consistent improvement in the reproduction of low flow-related hydrological signatures in the two training strategies. Notably, the classification-based training strategy exhibited a slightly superior enhancement. Consequently, the results in this study suggest that the regional LSTM model is capable of inducing streamflow generation mechanisms with the incorporation of static attributes and applying certain streamflow generation mechanisms to ungauged catchments based on their attributes.

This study contributes to the comprehension of how static attributes influence the regional LSTM model's performance in ungauged basins. Future research may include exploring innovative LSTM cell structures, such as the entity-aware LSTM proposed by Kratzert, Klotz, Shalev, et al. (2019), to examine how the LSTM model utilizes static attributes. Additionally, employing machine learning models to identify the most suitable streamflow generation mechanisms in ungauged basins could be a promising direction (Chadalawada et al., 2020). Another direction involves unraveling the internal mechanisms of LSTM models incorporating attributes through explainable AI (Jiang et al., 2022). The explicit revelation of streamflow prediction mechanisms or other hydrological characteristics in ungauged basins based on gauged data could significantly enhance the accuracy of predictions in ungauged basins.

Appendix A

Table A1 describes the static catchment attributes incorporated in the training of regional LSTM models.

Table A1
Description of Static Catchment Attributes (Addor et al., 2017)

Type	Attribute	Description	Unit
Topography and location	elev_mean	Catchment mean elevation	meter above sea level
	slope_mean	Catchment mean slope	m/km
	area_gages2	Catchment area (GAGESII estimate)	km^2
Climate indices	p_mean	Mean daily precipitation	mm/day
	pet_mean	Mean daily PET [estimated by N15 using Priestley-Taylor formulation calibrated for each catchment]	mm/day
	Aridity	Aridity (PET/P, ratio of mean PET [estimated by N15 using Priestley-Taylor formulation calibrated for each catchment] to mean precipitation)	-
	frac_snow_daily	Fraction of precipitation falling as snow (i.e., on days colder than 0°C)	-
	high_prec_freq	Frequency of high precipitation days (≥5 times mean daily precipitation)	days/year
	high_prec_dur	Average duration of high precipitation events (number of consecutive days≥5 times mean daily precipitation)	days
	low_prec_freq	Frequency of dry days (<1 mm/day)	days/year
	low_prec_dur	Average duration of dry periods (number of consecutive days <1 mm/day)	days
Land cover characteristics	forest_frac	Forest fraction	-
	lai_max	Maximum monthly mean of the leaf area index (based on 12 monthly means)	-
	lai_diff	Difference between the maximum and minimum monthly mean of the leaf area index (based on 12 monthly means)	-
	gvf_max	Maximum monthly mean of the green vegetation fraction (based on 12 monthly means)	-
	gvf_diff	Difference between the maximum and minimum monthly mean of the green vegetation fraction (based on 12 monthly means)	-

Table A1
Continued

Type	Attribute	Description	Unit
Soil characteristics	soil_depth_pelletier	Depth to bedrock (maximum 50 m)	m
	soil_porosity	Volumetric porosity (saturated volumetric water content estimated using a multiple linear regression based on sand and clay fraction for the layers marked as USDA soil texture class and a default value [0.9] for layers marked as organic material, layers marked as water, bedrock and "other" were excluded)	-
	soil_conductivity	Saturated hydraulic conductivity (estimated using a multiple linear regression based on sand and clay fraction for the layers marked as USDA soil texture class and a default value [36 cm/hr] for layers marked as organic material, layers marked as water, bedrock and "other" were excluded)	cm/hr
	max_water_content	Maximum water content (combination of porosity and soil_depth_statgso, layers marked as water, bedrock and "other" were excluded)	m
	sand_frac	Sand fraction (of the soil material smaller than 2 mm, layers marked as oragnic material, water, bedrock and "other" were excluded)	%
	silt_frac	Silt fraction (of the soil material smaller than 2 mm, layers marked as oragnic material, water, bedrock and "other" were excluded)	%
	clay_frac	Clay fraction (of the soil material smaller than 2 mm, layers marked as oragnic material, water, bedrock and "other" were excluded)	%
Geological characteristics	geol_permeability	Subsurface permeability (log10)	m^2

Data Availability Statement

All data used in this study are publicly available. The CAMELS data set (Newman et al., 2022) and the catchment classification results (Wu et al., 2021b) can be downloaded online.

Acknowledgments
This work was partially supported by the Shenzhen Key Laboratory of Precision Measurement and Early Warning Technology for Urban Environmental Health Risks (ZDSYS20220606100604008), High-level University Special Fund (G03050K001), the SUSTech Research Start-up Grants (Y01296129; Y01296229), and the Open Research Fund of Henan Provincial Key Laboratory of Hydrosphere and Watershed Water Security (HWWSF202303). The computation of this work was supported by the Center for Computational Science and Engineering at the Southern University of Science and Technology. We thank the editors and reviewers for providing critical comments that helped further improve the manuscript.

References

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F., & Mai, J. (2023). Continuous streamflow prediction in ungauged basins: Long short-term memory neural networks clearly outperform traditional hydrological models. *Hydrology and Earth System Sciences*, *27*(1), 139–157. https://doi.org/10.5194/hess-27-139-2023

Atieh, M., Taylor, G., Sattar, A., & Gharabaghi, B. (2017). Prediction of flow duration curves for ungauged basins. *Journal of Hydrology*, *545*, 383–394. https://doi.org/10.1016/j.jhydrol.2016.12.048

Beck, H. E., Pan, M., Lin, P., Seibert, J., Dijk, A. I. J. M., & Wood, E. F. (2020). Global fully distributed parameter regionalization based on observed streamflow from 4,229 headwater catchments. *Journal of Geophysical Research: Atmospheres*, *125*(17), e2019JD031485. https://doi.org/10.1029/2019JD031485

Bjerre, E., Fienen, M. N., Schneider, R., Koch, J., & Højberg, A. L. (2022). Assessing spatial transferability of a random forest metamodel for predicting drainage fraction. *Journal of Hydrology*, *612*, 128177. https://doi.org/10.1016/j.jhydrol.2022.128177

Blöschl, G., Bloschl, G., Sivapalan, M., Wagener, T., Savenije, H., & Viglione, A. (2013). *Runoff prediction in ungauged basins: Synthesis across processes, places and scales*. Cambridge University Press.

Chadalawada, J., Herath, H. M. V. V., & Babovic, V. (2020). Hydrologically informed machine learning for rainfall-runoff modeling: A genetic programming-based toolkit for automatic model induction. *Water Resources Research*, *56*(4), e2019WR026933. https://doi.org/10.1029/2019WR026933

Feng, D., Beck, H., Lawson, K., & Shen, C. (2023). The suitability of differentiable, physics-informed machine learning hydrologic models for ungauged regions and climate change impact assessment. *Hydrology and Earth System Sciences*, *27*(12), 2357–2373. https://doi.org/10.5194/hess-27-2357-2023

Feng, D., Lawson, K., & Shen, C. (2021). Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data. *Geophysical Research Letters*, *48*(14), e2021GL092999. https://doi.org/10.1029/2021GL092999

Gauch, M., Mai, J., & Lin, J. (2021). The proper care and feeding of CAMELS: How limited training data affects streamflow prediction. *Environmental Modelling & Software*, *135*, 104926. https://doi.org/10.1016/j.envsoft.2020.104926

Guo, Y., Zhang, Y., Zhang, L., & Wang, Z. (2021). Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review. *WIREs Water*, *8*(1), e1487. https://doi.org/10.1002/wat2.1487

Hashemi, R., Brigode, P., Garambois, P.-A., & Javelle, P. (2022). 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*(22), 5793–5816. https://doi.org/10.5194/hess-26-5793-2022

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., et al. (2013). A decade of Predictions in Ungauged Basins (PUB)—A review. *Hydrological Sciences Journal*, *58*(6), 1198–1255. https://doi.org/10.1080/02626667.2013.803183

Jiang, S., Zheng, Y., Wang, C., & Babovic, V. (2022). Uncovering flooding mechanisms across the contiguous United States through interpretive deep learning on representative catchments. *Water Resources Research*, *58*(1), e2021WR030185. https://doi.org/10.1029/2021WR030185

Kraft, B., Jung, M., Körner, M., Koirala, S., & Reichstein, M. (2022). Towards hybrid modeling of the global hydrological cycle. *Hydrology and Earth System Sciences*, *26*(6), 1579–1614. https://doi.org/10.5194/hess-26-1579-2022

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019a). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, *55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019b). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, *12*(9), 1620–1633. https://doi.org/10.1111/2041-210X.13650

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research*, *57*(3), e2020WR028091. https://doi.org/10.1029/2020WR028091

Newman, A., Sampson, K., Clark, M., Bock, A., Viger, R. J., Blodgett, D., et al. (2022). CAMELS: Catchment Attributes and MEteorology for Large-sample Studies (1.2) [Dataset]. *UCAR/NCAR - GDEX*. https://gdex.ucar.edu/dataset/camels.html

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. https://doi.org/10.5194/hess-19-209-2015

Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., & Nearing, G. (2017). Benchmarking of a physically based hydrologic model. *Journal of Hydrometeorology*, *18*(8), 2215–2225. https://doi.org/10.1175/JHM-D-16-0284.1

Pinheiro, V. B., & Mauro, N. (2013). Calibration of the parameters of a rainfall-runoff model in ungauged basins using synthetic flow duration curves as estimated by regional analysis. *Journal of Hydrologic Engineering*, *18*(12), 1617–1626. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000737

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., et al. (2021). From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature Communications*, *12*(1), 5988. https://doi.org/10.1038/s41467-021-26107-z

Wu, S., Zhao, J., Wang, H., & Sivapalan, M. (2021a). Regional patterns and physical controls of streamflow generation across the conterminous United States. *Water Resources Research*, *57*(6), e2020WR028086. https://doi.org/10.1029/2020WR028086

Wu, S., Zhao, J., Wang, H., & Sivapalan, M. (2021b). Regional patterns and physical controls of streamflow generation across the conterminous United States [Dataset]. *Zenodo*, *57*(6). https://doi.org/10.5281/zenodo.4904012

Yao, Y., Zhao, Y., Li, X., Feng, D., Shen, C., Liu, C., et al. (2023). Can transfer learning improve hydrological predictions in the alpine regions? *Journal of Hydrology*, *625*, 130038. https://doi.org/10.1016/j.jhydrol.2023.130038

Yin, H., Zhu, W., Zhang, X., Xing, Y., Xia, R., Liu, J., & Zhang, Y. (2023). Runoff predictions in new-gauged basins using two transformer-based models. *Journal of Hydrology*, *622*, 129684. https://doi.org/10.1016/j.jhydrol.2023.129684