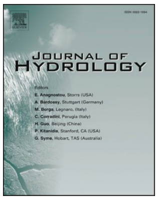
Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol



Research papers

Generalization of an Encoder-Decoder LSTM model for flood prediction in ungauged catchments

Yikui Zhang [a,1,*], Silvan Ragettli [b], Peter Molnar [a], Olga Fink [c], Nadav Peleg [d]

[a] *Institute of Environmental Engineering, ETH Zurich, Zurich, Switzerland*
[b] *Hydrosolutions Ltd., Zurich, Switzerland*
[c] *Laboratory of Intelligent Maintenance and Operations Systems, EPFL, Lausanne, Switzerland*
[d] *Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland*

ARTICLE INFO

Keywords:
Flood prediction
Ungauged catchments
Deep learning
LSTM

ABSTRACT

Flood prediction in ungauged catchments is usually conducted by hydrological models that are parameterized based on nearby and similar gauged catchments. As an alternative to this process-based modelling, deep learning (DL) models have demonstrated their ability for prediction in ungauged catchments (PUB) with high efficiency. Catchment characteristics, the number of gauged catchments, and their level of hydroclimatic heterogeneity in the training dataset used for model regionalization can directly affect the model's performance. Here, we study the generalization ability of a DL model to these factors by applying an Encoder-Decoder Long Short-Term Memory neural network for a 6-hour lead-time runoff prediction in 35 mountainous catchments in China. By varying the available number of catchments and model settings with different training datasets, namely local, regional, and PUB models, we evaluated the generalization ability of our model. We found that both quantity (i.e. number of gauged catchments available) and heterogeneity of the training dataset used for the DL model are important for improving model performance in the PUB context, due to a data synergy effect. The assessment of the sensitivity to catchment characteristics showed that the model performance is mainly correlated to the local hydro-climatic conditions; the more arid the region, the more likely it is to have a poor model performance for prediction in ungauged catchments. The results suggest that the regional ED-LSTM model is a promising method to predict streamflow from rainfall inputs in PUB, and outline the need for preparing a representative training dataset.

1. Introduction

Accurate and computationally efficient hydrological models are necessary for streamflow prediction to issue timely warnings for flash floods (Moore et al., 2005). Physics-based hydrological models are the most robust models that can be used for this purpose. They simulate physical processes in the rainfall-runoff transformation with parameters that represent soil, land surface, and climate properties, that need to be optimized for each geographic location with observations. But most catchments worldwide lack hydrological monitoring data and are considered "ungauged" (Guo et al., 2021), meaning that direct calibration of catchment parameters in these catchments is not possible. For this reason, the problem of prediction in ungauged basins (PUB) has received considerable attention in the hydrological community (Sivapalan et al., 2003).

One solution to the calibration of hydrological models for catchments without available data utilizes the concept of parameter regionalization. The idea is to use parameters calibrated in gauged catchments to predict the model parameters in a target ungauged catchment (Blöschl and Sivapalan 1995). Similarity-based and regression-based methods are widely used for model regionalization (Oudin et al., 2008). For example, Beck et al. (2016) proposed a scheme for the regionalization of model parameters at the global scale based on a similarity approach by selecting 10 gauged catchments with the most similar characteristics as donors for parameter transfer. However, the question of how to identify the selection criteria for choosing the optimal donor catchments remains a challenge that restricts the wide application of this method. Ragettli et al. (2017) applied the

[*] Corresponding author at: Institute of Environmental Engineering, ETH Zurich, Zurich, Switzerland.
E-mail address: yik.zhang@fz-juelich.de (Y. Zhang).
[1] Now at Institute for Bio- and Geosciences, Agrosphere (IBG-3), Research Centre Jülich, Jülich, Germany.

https://doi.org/10.1016/j.jhydrol.2022.128577
Received 19 May 2022; Received in revised form 21 September 2022; Accepted 4 October 2022
Available online 29 October 2022

classification and regression tree (CART) method to explore parameter transferability in the full space of catchment descriptors for the hydrological model and showed that this method can be an effective tool for identifying similarity among catchments. However, it is model dependent and relies on manually identifying the similarity and then transferring the parameter set from a series of pre-defined hydrological models.

An alternative solution to parameterized hydrological models for streamflow simulation in ungauged catchments is the use of data-driven deep learning (DL) models. These models can be directly trained with inputs from meteorological and catchment characteristic data to simulate streamflow without using a physical hydrological model to predefine their similarity. For example, Kratzert et al. (2019) evaluated the ability of a Long Short-Term Memory (LSTM) model for the regionalization of over 500 catchments in the USA. They concluded that data-driven models had a strong capacity to learn non-linear climate-runoff relationships and to achieve model regionalization without identifying pre-defined criteria for similar donor catchments.

The catchment characteristics (e.g. topography, land use) and the climatic training dataset are the two most important factors that affect the model regionalization and the setup of a physical hydrological model in ungauged catchments (Teutschbein et al., 2018; Gong et al., 2021). While the performance of hydrological models in ungauged catchments is sensitive to these two factors, there are only a few studies that evaluated the generalization ability of data-driven DL models. For example, Potdar et al. (2021) predicted flood peak discharge in ungauged catchments based on the gradient boosted trees model (XGBoost) and found that catchment geomorphologic attributes have a higher impact on the prediction skill than climatologic attributes. Gauch et al. (2021) studied the sensitivity of the prediction skill of the LSTM model for daily streamflow in the USA to additional training samples and showed that it is not enough to train data-driven models on a few gauged catchments, but one should strive to use as many catchments as possible. Fang et al. (2022) proposed a concept of ‘data synergy’, pointing out that to achieve higher predictive performance, a representative dataset with large but heterogeneous training samples (i.e. different characteristics of catchments) is needed. However, the generalization of DL models to the PUB problem with respect to the representativeness of the training dataset and catchment characteristics has not been studied thoroughly yet.

This study aims to evaluate the generalization ability of a DL model to predict floods in ungauged catchments considering the above factors. For this purpose, an Encoder-Decoder LSTM (ED-LSTM) neural network was applied to set up a forecast model for a 6-hour lead-time streamflow prediction in 35 mountain catchments in China. Three model setups: (i) a local model for each catchment; (ii) a regional model; and (iii) regional PUB models which differ in the choice of training and testing catchments, were applied. Their performances were compared to the CART regionalization method to evaluate the ability of the DL model to predict streamflow in the ungauged catchments (Ragettli et al., 2017). The analysis of the generalization ability of the PUB models concerning the training dataset and catchment characteristics was conducted by comparing the three model setups. The main purpose is to provide recommendations on the importance of preparing representative training datasets in the context of PUB with DL methods. We aim to answer the question of how ED-LSTM models (and other rainfall-runoff DL-based models) may be used for event-based flood forecasting, and which data requirements have to be present to provide reasonably accurate predictions in ungauged basins.

2. Data and models

2.1. Study area

The study focuses on 35 mountainous catchments (Table S1) located in ten Chinese provinces (Fig. 1). The catchments were classified as northern catchments (17) and southern catchments (18) based on the traditional geographical south-north division of China, which is called the Huai river-Qin mountain line. This line approximates the 0 °C January isotherm and the 800-mm isohyet (Zhao et al., 2015). Mean

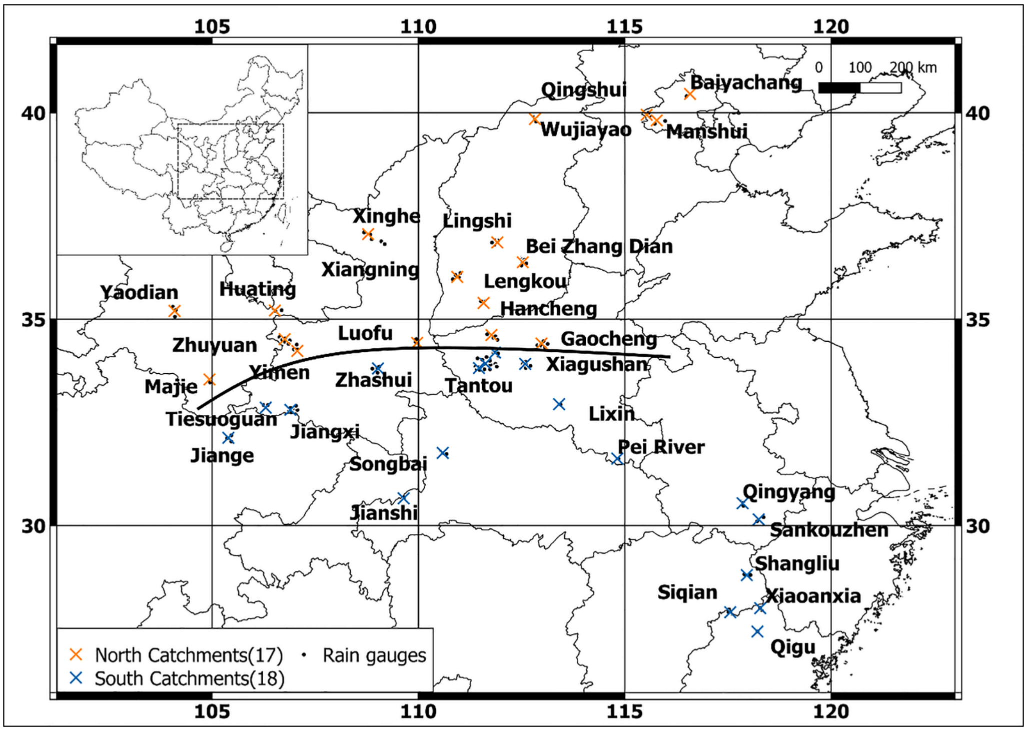


Fig. 1. Locations of the 35 studied catchments are divided into northern (17, orange cross) and southern (18, blue cross) regions, which are corresponding to the location of local hydrological stations. The black solid line represents the south (S) – north (N) division of catchments along the Huai river-Qin mountain line (Ragettli et al., 2017). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

annual precipitation in the north is on average 57 % lower than in the south and mean annual air temperatures are on average 6 °C lower.

The catchment areas range from 14 to 1693 km^2, whereas the mean catchment size is 278 km^2. Hourly hydrological and meteorological data are available from stream and rain gauges located within or in close vicinity to the catchments (Fig. 1). In addition, the county weather stations record daily maximum and minimum 2-meter air temperature. On average, 11 years of data are available per catchment, with 1 to 7 storm events occurring per year between April and October. We consider storm events as days with total precipitation greater than 5 mm d^{-1}, following the definition by Ragettli et al. (2017). Each of the catchments is characterized by several static variables (Kratzert et al., 2019) describing their climatic, vegetation, soil, and topographical properties (Table S3).

2.2. Long Short-Term memory (LSTM) network

LSTM networks are a type of recurrent neural network that can learn time dependencies in time series data (Hochreiter and Schmidhuber 1997). It has cell and hidden states which can account for the long-short term memory effects. Therefore, it is a good choice for modeling time series of runoff as it can account for a range of time-dependent delays, like seasonality and natural annual variability cycles (long-term, months to years), and the immediate rainfall-runoff response (short-term, minutes to hours).

Equations 1 to 6 provide the mathematical formulation of the LSTM at each time step:

$f\,[t] = \sigma[W_f \bullet x[t] + U_f \bullet h[t-1] + b_f]$ (1).

$i[t] = \sigma[W_i \bullet x[t] + U_i \bullet h[t-1] + b_i]$ (2).

$c[t] = \tanh[W_g \bullet x[t] + U_g \bullet h[t-1] + b_g]$ (3).

$o[t] = \sigma[W_o \bullet x[t] + U_o \bullet h[t-1] + b_o]$ (4).

$c[t] = f\,[t] \times c[t-1] + i[t] \times c[t]$ (5).

$h[t] = o[t] \times \tanh(c[t])$ (6).

where f[t], i[t], c[t], o[t], and h[t] represent the forget gate, input gate, cell state, output gate, and hidden state at each time step respectively; W, U, and b are the weights and bias term of the neural network. The σ (sigmoid function) and tanh are two activation functions and x[t] are the inputs.

The LSTM cell has three gates maintaining and adjusting its cell state and hidden state (Fig. 2), including a forget gate (Eq. 1), an input gate (Eq. 2), and an output gate (Eq. 4). Each gate has a sigmoid function that adds non-linearity to the linear combination of the input x[t] and hidden state from last time step h[t-1]. Eqs. 3 and 5 represent how much input and last hidden state is contributed to the cell state c[t]. Finally, the output h[t] of the current time step is calculated from the output gate and cell state shown in Eq. 6. Cell state and hidden state are then passed to the next time step.

The Encoder-Decoder (ED) (Fig. 2) structure has been used in the field of sequence-to-sequence prediction problems, especially for language translation (Cho et al., 2014). The encoder and decoder enable the model to operate on different input and output time steps. The ED-LSTM model consists of two LSTM networks in both the encoder and the decoder parts. The application of the ED structure in LSTMs can efficiently improve the performance of ahead-time prediction in the field of hydrology since the existence of the encoding and decoding architecture can eliminate the restriction on the length of input and output sequences (Kao et al., 2020). In this case, the input and output sequences do not necessarily have the same time steps. The input and output can be flexible to embed different types of input data. For example, catchment static variables and rainfall time series can be used as input to the model at the same time.

Unlike the basic LSTM structure, the encoder layer only outputs the

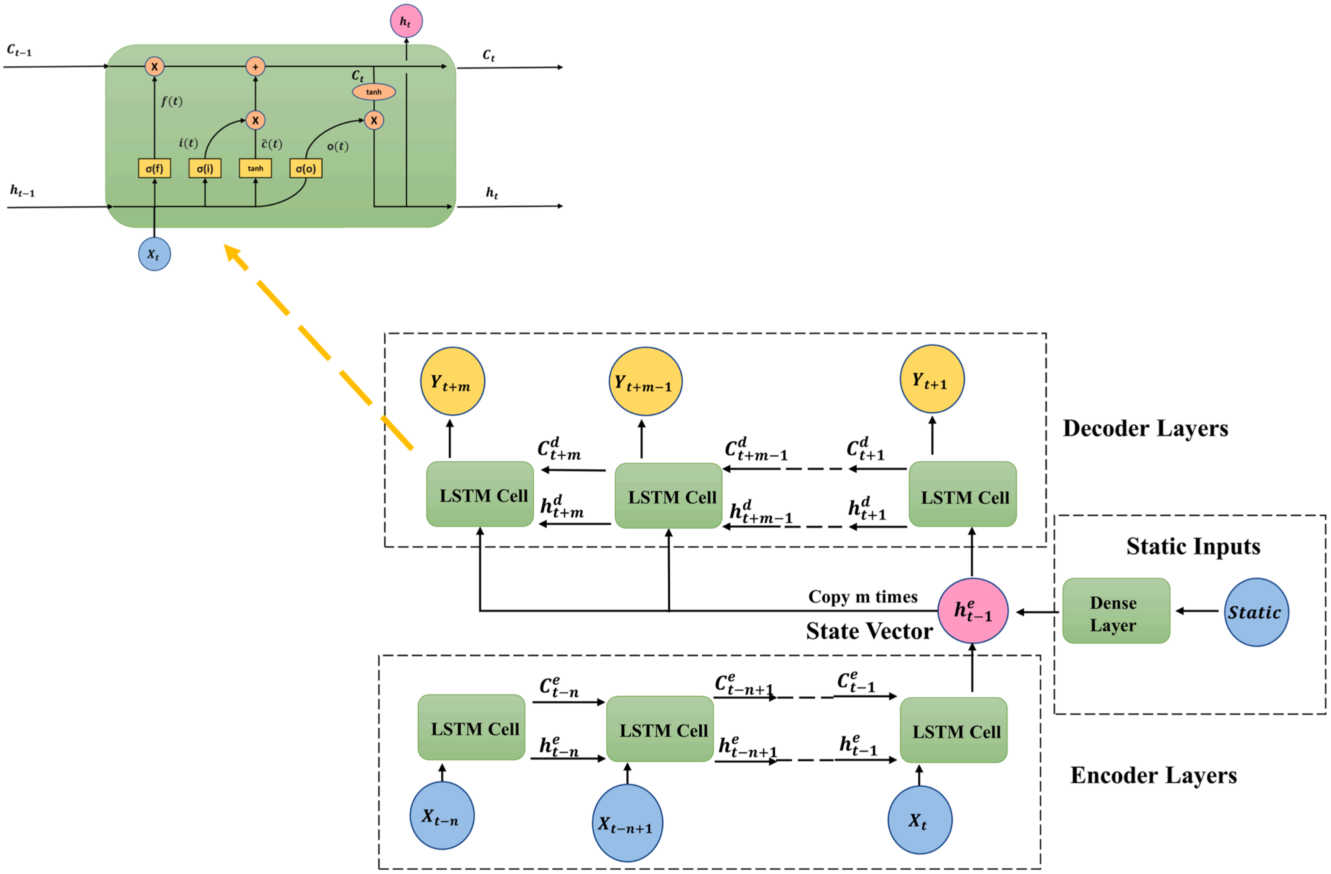


Fig. 2. Configuration of the LSTM cell: σ and tanh represent sigmoid and tanh activation functions, respectively. X represents the input while C and h are the cell state and hidden state at the current time step.

hidden state from the last cell. Then, it is copied as input for each LSTM cell in the decoder layer. It contains information collected from the input sequence at each time step. Thus, it could be effective to use this structure to improve long-term dependencies for longer time step prediction than in the regular LSTM. In this study, five additional dense layers were set after the LSTM decoder layer for better decoding of the sequence output at each time step.

3. Experiment setup

3.1. Input data composition

The ED-LSTM model experiments described in the next section require dynamic and static inputs. The dynamic input data includes the following climate variables: (i) hourly precipitation; (ii) hourly streamflow; and (iii) maximum/minimum 2-m daily air temperature. The dynamic inputs were divided into an observation phase and a prediction phase (Fig. 3). The first phase includes 24-hour precipitation, temperature, and streamflow data, computed from the hourly observed data before the prediction phase. The second phase contains the 6-hour precipitation and temperature as driving data for the streamflow forecast. Also, previously observed streamflow was used as the dynamic input because it is known to improve forecast (e.g. Song et al., 2020). Daily maximum and minimum temperature data are used to better account for snow-induced streamflow processes (e.g. Xiang et al., 2020).

We reduced the number of static catchment attributes from a total of 27 (Table S2) to 14 (Table S3) using a principal component (PC) analysis, preserving only the uncorrelated attributes that represent best the natural clusters in each category of catchment characteristics (Singh et al., 2014). The absolute PC scores (Table S4) of each selected attribute were taken to represent the uncorrelated patterns rather than the individual catchment characteristics (Ragettli et al., 2017).

3.2. Numerical experiments

Three different ED-LSTM numerical experiments were set up: (i) local models, i.e. a unique ED-LSTM setup for each of the catchments; (ii) a regional model simulating all catchments at once; and (iii) regional PUB models – models that include both gauged and ungauged catchments in different combinations.

In the first experiment, ED-LSTM models were trained and tested on individual catchments without using data from other catchments. In total, 35 local models were set up. Only the dynamic variables were applied as input data and no static variables were involved in this setup. Preliminary tests were conducted on the catchment with the longest training samples to select the optimal ED-LSTM model hyperparameters.

In the second experiment, we set up a regional model, namely-one model for all catchments. This ED-LSTM model was trained and tested using an ensemble of events from all catchments together. The regional model was applied with both the dynamic and static variables as inputs. Kratzert et al. (2018) demonstrated that the LSTM model can perform better with a regional model setting than with a setup for individual catchments (as in the first experiment) as more data is available for model training (i.e. additional rainfall-runoff interactions are available for the LSTM to learn from). This experiment aims to find out how much can the flood warning prediction ability benefit from a large training dataset.

In the last experiment, the regional ED-LSTM setup was applied for prediction in the context of ungauged catchments (PUB models). We conducted two tests here, first to explore the ED-LSTM generalization ability to the number of gauged catchments used in the training, and second to the characteristics of the catchment. To explore the first question, we followed a similar setup as for the regional model. However, we trained the model with fewer catchments using the k-fold validation strategy, which enables us to test the model performance on unseen catchments in the training process. The k-fold validation is commonly used for model parameter selection (e.g. Chang et al., 2015)

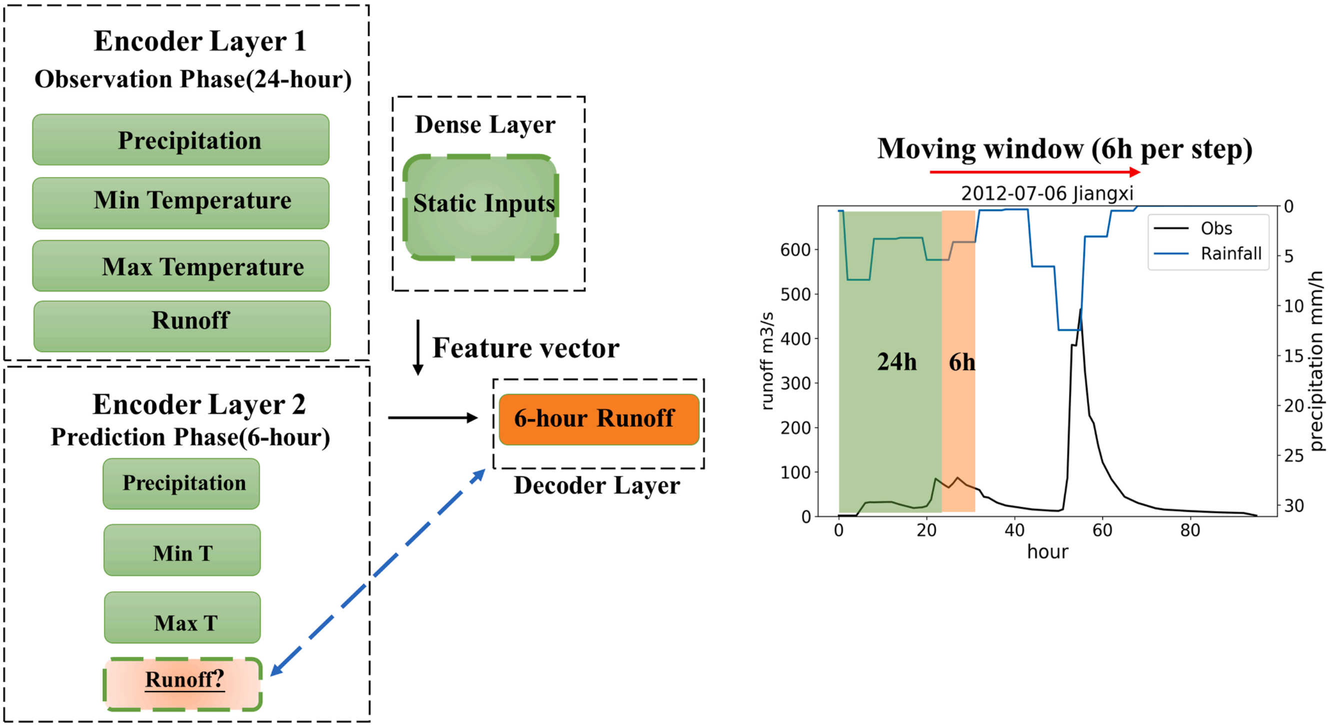


Fig. 3. Input composition and prediction process of the ED-LSTM structure for one event. Two encoder layers incorporated previous 24-hour data in the observation phase and 6-hour pseudo forecasting meteorological data as driving forces (all of these data were observations). The decoder layer outputs the predicted 6-hour runoff corresponding to a 6-hour driving force in the prediction phase. The dense layer functions as an embedding layer for feeding the catchment static variables. For one event, the two phases consist of a moving window and each step moves them forward in 6-hour increments.

but here we used it as a tool for an 'out-of-sample prediction': The 35 catchments were split randomly into *k* groups (namely 'folds') of approximately equal size; catchments from *k-1* groups were used to train the model, and then the model was tested on the remaining single group of catchments as ungauged catchments. This procedure is repeated *k* times so that out-of-sample predictions are made available to all catchments (Kratzert et al., 2019). As the catchments are heterogeneous, the number of valid training samples varies. At first, 10-fold validation was adopted to train 10 models with the same model structure and each model was applied for the prediction in three ungauged catchments. This means that the training process was repeated 10 times with different 32 catchments to cover all catchments (Pub1, Table S4).

To evaluate the generalization ability of the PUB modeling to the catchment characteristics such as climate and topography, k-means clustering methods were applied for classifying and grouping the catchments based on the 27 catchment static variables listed in Table S2 including climatic, topographic, vegetation and catchment drainage properties. Based on the silhouette scores (Fig. S1), we found that five clusters are required to group the catchments by their attributes. We averaged the model performance based on each cluster.

Fang et al. (2022) hypothesized that DL-based models will have a better prediction skill in the context of PUB if a regional model is not trained on a relatively small and hydrologically homogenous dataset (e.g. few catchments but sharing similar hydrological characteristics) but rather on a larger and heterogeneous sample (e.g. multiple catchments with varying characteristics). To test this effect, additional experiments were conducted: the 'PUB' model was applied to each of the 5 clusters to create a new 'PUB' model for the smaller sub-regions represented by the clusters. The leave-one-out scheme was used so each sub-regional PUB model was trained on N-1 catchments and tested on a specific catchment; the 'PUB' model was then applied to either the north (17 catchments in "dry and cold" climate) or the south (18, "wet and warm") catchments. To examine the effects of regionalization performance on the sample size (e.g. Gong et al., 2021), the PUB experiment was iterated for a different number of catchments in training ranging from 18 to 30 based on fold numbers from 2 to 10 (Pub2 to Pub6, Table S4). As a reference for the quality of the ED-LSTM model predictions in the PUB mode, we used simulations for the 35 catchments by the PRMS hydrological model presented in Ragettli et al. (2017). Note that Ragettli et al. (2017) used two CART methods to emulate parameter regionalization in ungauged catchments. However, we used only the results of their classification tree as our reference, as the other CART method resulted in a very similar model performance.

The training strategy of the ED-LSTM model was as follows. In the first step, we determined the hyperparameters (e.g. learning rate, batch size, cell numbers) based on a grid search. The hydrometeorological dataset was divided into training, validation, and testing sets (50 %, 25 %, and 25 %, respectively) using the local model. Afterward, the dataset was split into training and testing sets for training the local and regional experiments (75 % and 25 %, respectively). For the PUB models, all data in 'gauged' catchments were used for training while the events in 'ungauged' catchments were used for testing. All ED-LSTM models had 256 memory cells in both the encoder and decoder layer, with a dropout rate of 0.4 based on the results of the hyperparameter grid search. There were 128, 64, 32, 16, and 1 cells in the five dense layers after the LSTM layer, and the optimal batch size was 32.

3.3. Evaluation metrics

The evaluation of the model performance aimed to (i) assess the capacity of the model to reproduce an overall streamflow fit at the event scale; and (ii) evaluate its ability to correctly identify streamflow extremes, i.e. the peak flow which is important for flood warning. The Nash-Sutcliffe efficiency (NSE, Nash and Sutcliffe 1970) metric was used to assess the overall streamflow fit:

$$NSE = 1 - \frac{\sum (sim^t - obs^t)^2}{\sum (obs^t - obs^m)^2} \tag{7}$$

where *sim* and *obs* are the predicted and observed streamflow, *t* indicates a given time step and *m* refers to the mean. NSE ranges from -infinity to 1, with 1 being a perfect match. NSE values larger than 0.5 can be considered as a satisfying prediction capacity while the value of 0 signifies that the prediction is as good as the mean of the observations (Moriasi et al., 2007).

Flood frequency analysis was used for the quantification of flood warning performance. The cumulative distribution function of the Generalized Extreme Value distribution was applied to estimate the return periods of the observed and simulated hourly streamflow peaks (see Ragettli et al. 2017, for example). For each storm event, we determined if the maximum hourly streamflow exceeded a reference flood quantile of a given return period. We consider the 2-year return period to represent common high streamflow and the 10-year return period to represent a severe flood.

We identified three cases for flood prediction performance (following Javelle et al., 2016): (i) 'hit' (*H*) – when both simulated and observed streamflow exceeded the flow threshold corresponding to a certain return period indicating high flow; (ii) 'miss' (*M*) – when the simulated streamflow was below the threshold and failed to agree with the high flow detected by observations; and (iii) 'false alert' (*FA*) – when the simulated streamflow indicated a high flow but the observed streamflow did not (see the contingency table in Table S5). Moreover, to assess the temporal accuracy of reproducing the peak flow, a 2-hour condition was added to the evaluation, which means that if the simulated peak flow had a 2-hour shift compared to the observed peak, the prediction was also identified as a miss. Three contingency scores were computed for evaluating the flood warning ability: (i) the Probability of Detection (*POD*, Eq. (8)) which is the fraction of correct event predictions (hits) in all observed high flow events; (ii) the Success Rate (*SR*, Eq. (9)) which is the fraction of hits in the total number of all high flow event predictions; (iii) and the Critical Success Index (*CSI*, Eq. (10)) which is the fraction of hits in the total number of event predictions plus the number of missed observations.

$$POD = \frac{H}{H + M} \tag{8}$$

$$SR = \frac{H}{H + FA} \tag{9}$$

$$CSI = \frac{H}{H + M + FA} \tag{10}$$

4. Results

4.1. Evaluation of the model performance

First, we compared the NSE values between the observed and predicted streamflow in the three experiments (Fig. 4). We qualitatively divided the NSE values into three performance groups: poor (NSE $\leq$ 0), average (0 $<$ NSE $<$ 0.5), and above average (NSE $\geq$ 0.5). The training of local models resulted in 14 catchments with above average NSE values and 12 catchments with poor NSE values, while in the regional model, most catchments resulted in an above average performance and only three catchments had poor performance (Fig. 4a). Evaluating the performance of the PUB models in the north and south areas separately (Fig. 4b and c), it becomes apparent that the models have better prediction ability in southern catchments, with a considerably higher number of models with good performance (4 on average in the northern catchments in comparison to 12 on average in the southern ones). Compared to Ragettli et al. 2017′s results, the PUB model performed better than the CART-based method in the northern catchments, as 2 (1)

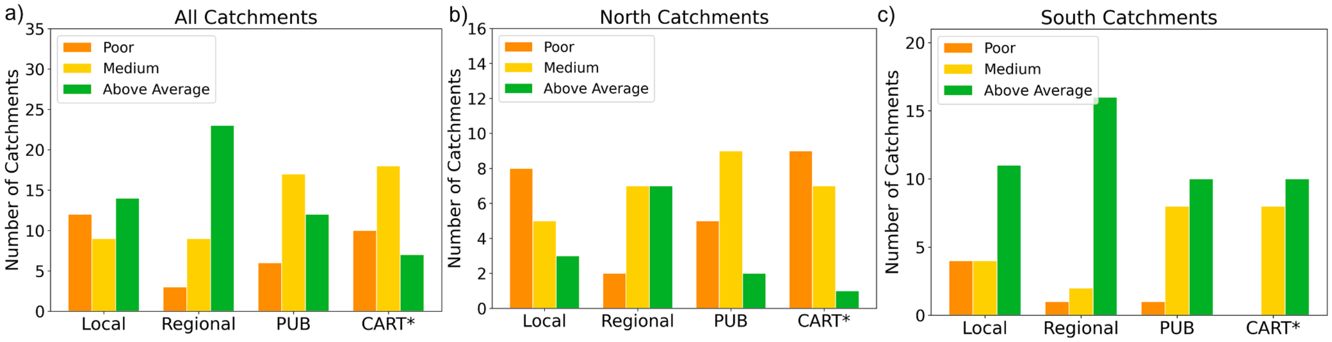


Fig. 4. Summary of the model performance classified by NSE values (poor, NSE < 0; average, 0 < NSE < 0.5; and above average, NSE greater than 0.5). The NSE values for the classification tree method ('CART') are from the hydrological model in Ragettli et al. (2017).

catchments resulted in above average performance and 5 (9) catchments in poor performance for the PUB (CART-based) model (Fig. 4a and b). In southern catchments, the performance distribution of the two methods is nearly identical (Fig. 4c).

The results of the contingency scores, evaluating the flood warning performance, are presented in Table 1. For the 2-year return period events, the regional and the PUB models had the best performances (equally) for detecting streamflow peaks (POD values of 0.85, SR values of 0.82, and CSI values of 0.72, considering all catchments). For prediction in ungauged catchments, PUB models outperform the CART-based regionalization methods by 12 % on the probability of detection considering all catchments. Similarly, PUB models have a better success rate and critical success index in comparison to the CART-based regionalization. For the 10-year return period events, all models' contingency scores decreased by at least 10 %. In line with the 2-year return period predictions, the regional and PUB models showed the best performance (POD of 0.7, considering all catchments). Again, the performance of the PUB models was 10 % higher than CART-based methods. The PUB model achieved the best performance for SR and CSI scores instead of being equal to the regional model. Results indicate that the model detection performances were affected by the climate, as in wetter climate (i.e. the southern catchments) better prediction abilities were observed than in the drier regions (i.e. the northern catchments). The generalization ability increases with the prediction of higher streamflow from an order of 8 % difference between the southern and northern catchments for the 2-year return period to ~30 % for the 10-year return period (exception is PUB, 10-year return period).

4.2. *Generalization ability to different training dataset*

Fig. 5 shows the overall performance of PUB models that were trained on different sub-regions. The homogeneous (similar climate) dataset did not improve but rather impair the PUB model performance (Fig. 5a): while the median NSE value of the model trained on southern catchments was similar to the result trained on the global dataset (around 0.5), for the northern sub-region the model performance was significantly worse than the result when trained on all catchments. This is even more evident when the models are trained based on the climate clusters (Fig. 5b), where the overall performance has decreased significantly compared to the median NSE when the models are trained on the entire dataset. For clusters 1 and 3, the median has dropped significantly from around 0.35 to −0.1 and −0.7 (respectively). In addition, the variation of the models' NSE skill (i.e. the box plots) increases in all models trained on the cluster-based (homogenous) datasets in comparison to the training with the entire region datasets, with negative lower quartile and lower whisker NSE values reaching −1 for clusters 1, 3, and 4.

The model generalization ability to the number of catchments used for training is presented in Fig. S4. The median NSE of PUB models did not vary notably for models trained on 32 to 24 catchments but when the number of catchments used in the training dataset was below 18, the median NSE declined from 0.3 to only 0.15 (Fig. Sa) and this decline trend is consistent for even smaller number of catchments. In contrast to the NSE results, the POD scores of the PUB models do not degrade with the decreasing number of training catchments and all PUB models demonstrate good flood warning capability with median POD scores higher than 0.8 (Fig. S4b).

However, model performance at an individual catchment may not always be improved when trained on a large dataset. As shown in Fig. 6, four performance categories can be distinguished: (i) performance is similar for all models (e.g. Yimen catchment); (ii) poor performance in the local model but satisfactory in others (e.g. Shangliu catchment); (iii) poor performance in the PUB model but satisfactory in others (e.g. Pei River); and (iv) random performance variation with model setup (e.g. Qigu catchment). Most of the catchments (12) fall into category (ii) performance, followed by 10 catchments with category (i) performance.

Table 1
Contingency scores (POD –probability of detection; SR – success rate; CSI – critical success index) evaluating high streamflow predictability (A – all catchments; S – southern catchments; N – northern catchments). Bold numbers represent the highest contingency score for each area.

	POD			SR			CSI		
	A	S	N	A	S	N	A	S	N
2-year flood events									
Local	0.73	0.74	0.72	0.80	**0.87**	0.73	0.62	0.66	0.57
Regional	**0.86**	**0.89**	**0.82**	**0.83**	0.84	**0.82**	**0.73**	**0.76**	0.69
PUB	0.85	0.88	0.80	0.81	0.83	0.79	0.71	0.75	0.66
CART	0.73	0.76	0.70	0.73	0.80	0.66	0.58	0.64	0.51
10-year flood events									
Local	0.50	0.67	0.36	0.48	0.50	0.46	0.33	0.40	0.25
Regional	0.69	0.67	0.71	0.64	0.62	0.67	0.50	0.47	0.53
PUB	**0.71**	0.65	**0.78**	**0.73**	**0.72**	**0.73**	**0.56**	0.52	**0.61**
CART	0.62	**0.74**	0.45	0.56	0.65	0.44	0.42	0.54	0.29

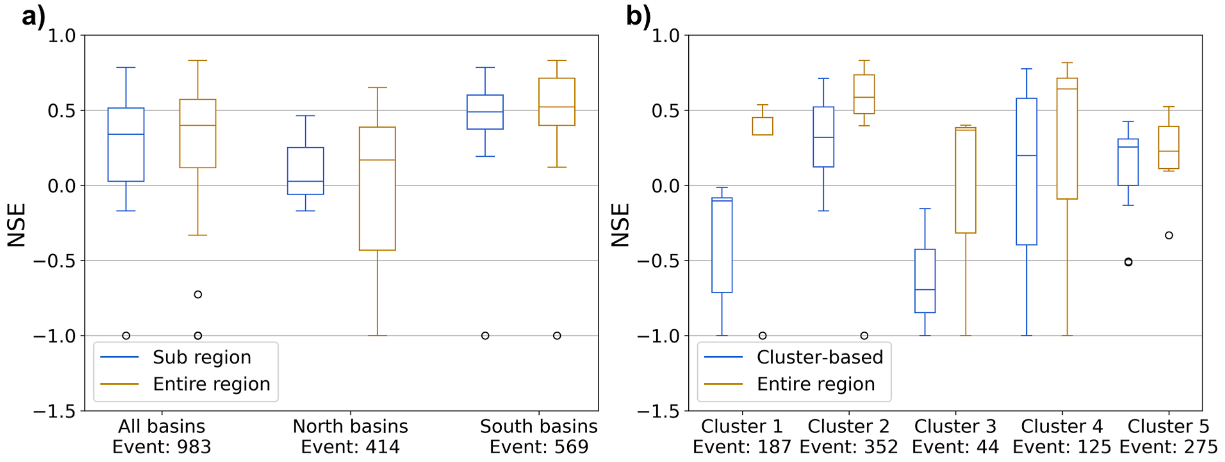


Fig. 5. Boxplot of NSE for PUB models, comparing the models trained on sub-regional areas: (a) north and south regions; (b) hydroclimatic clusters. The line inside the box shows the median and the box includes the lower and upper quartile values (the 25th-75th percentile range). The circles represent the outliers. The number of flood events (samples) that are valid for modeling training for each sub-region is listed below each label. For better visualization, NSE values lower than −1 are set to be −1.

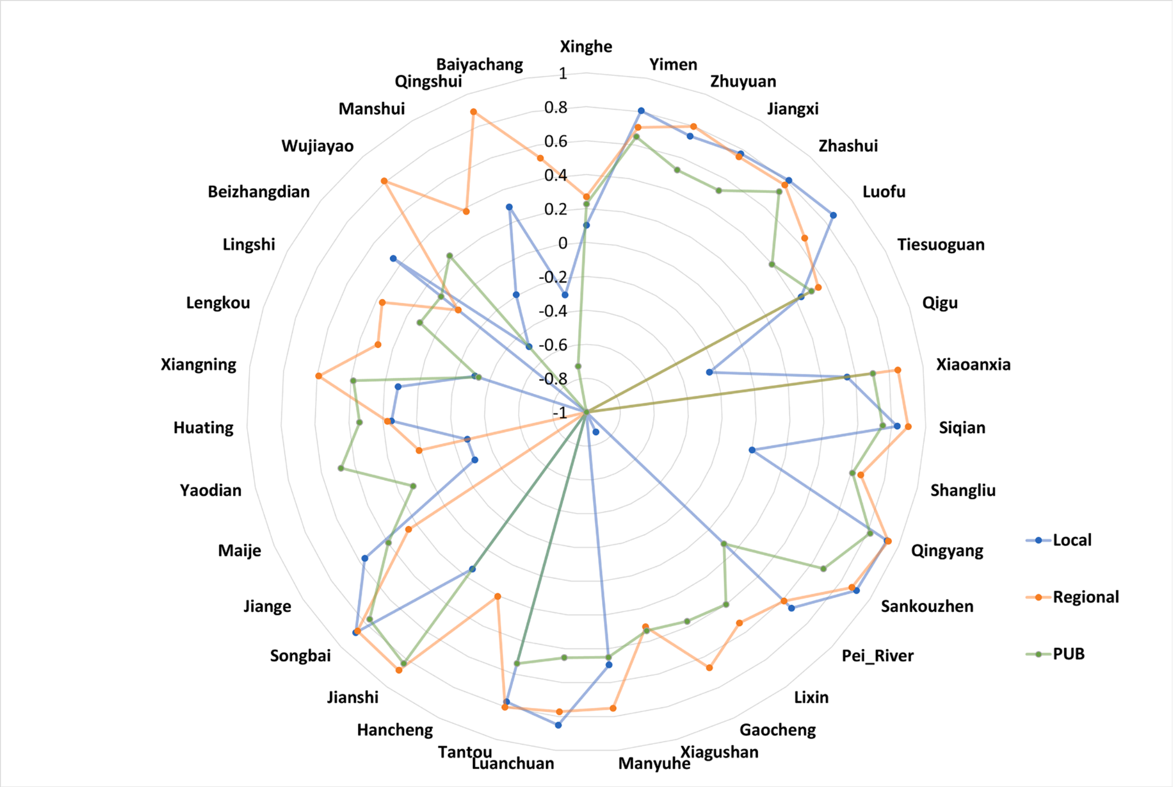


Fig. 6. Radar plot for showing median NSE at 35 catchments for all events of local, regional, and PUB experiments.

But still, 8 catchments fall into category (iii) while 5 resulted in category (iv), with a poorer performance even though they were trained using a larger dataset.

4.3. Generalization ability to catchment characteristics

The classification of the 35 catchments into five classes based on the k-means clustering method is presented in Fig. 7a. Of the northern catchments, 11 (81 %) were grouped into cluster 5. The southern catchments were classified into clusters 2 to 4; catchments from the southwest were mostly clustered in cluster 2, while catchments from the southeast region were grouped in cluster 4. The catchments in cluster 1 and cluster 3 are mainly located in central China in Henan province.

The five clusters represent different climatological and hydrological conditions (Table 2). Cluster 5 climate is arid to semi-arid, with a ratio of annual potential evapotranspiration to precipitation (PET/P) lower than 1. Moreover, the catchments in cluster 5 are mainly located in high mountain areas in northern latitudes (Fig. 7a), and, thus, are also colder on average compared to catchments in other clusters. The catchments in cluster 3 are located in lower, flatter, and warmer areas. The topography attributes of cluster 4 are similar to cluster 3. However, the latter is more humid. The catchments in cluster 2 are located in high elevation warm and humid areas and are located in southwest China (Fig. 7a). The climatic and topographical characteristics of the catchments in cluster 1

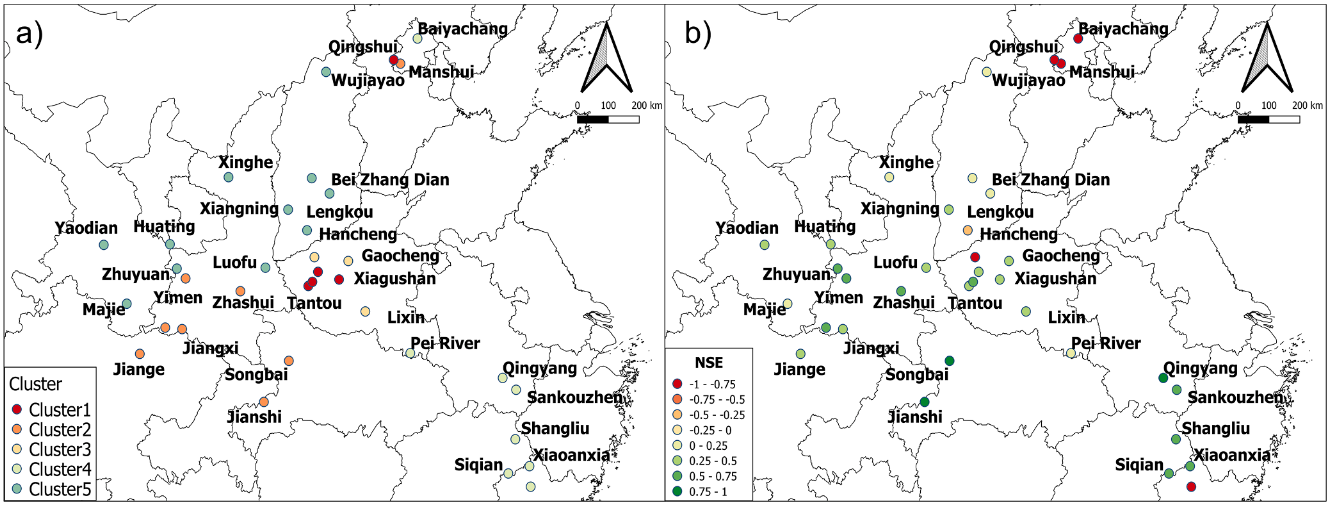


Fig. 7. (a) Classification of the 35 catchments into 5 clusters. (b) Mean NSE values of the PUB model.

Table 2
Statistics of main mean climatic and hydrological catchment characteristics and model performance (NSE and POD scores for 2-year return period flood) for the PUB model for each of the clusters presented in Fig. 7.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Number of catchments	5	8	3	8	11
Precipitation (mm/yr)	895	1144	883	1857	545
T (℃)	9.9	10.4	16.1	14.3	7.1
PET/P	1.27	0.91	1.33	0.58	1.74
Area (m2)	339	229	253	103	271
Elevation (m)	1080	1163	490	519	1482.
h-gradient (%)	0.20	0.77	0.001	0.17	0.13
NSE	0.45	0.57	0.36	0.64	0.22
POD (2-yr)	0.80	0.76	0.82	0.91	0.76

are intermediate of the clusters and the three catchments associated with this cluster are found in central China (Fig. 7a).

The models are sensitive to the catchment characteristics as the model performances vary between clusters. The NSE scores of the clustered catchments in the PUB model in clusters 2 and 4 (both above 0.5) are the highest among the five clusters, while cluster 5 has the lowest median NSE with 0.23 (Table 2 and Fig. 7b). A declining trend in NSE values is also observed from south to north (Fig. 7b). The POD scores, however, show no remarkable differences in flood warning capability between clusters, with the highest value of 0.91 in cluster 4 and the lowest values of 0.76 in clusters 2 and 5 (Table 2).

A negative correlation was found between PET/P (used as climate proxy) and NSE (Fig. 8a), indicating that the DL model is more likely to perform well in wetter areas. However, such a relationship cannot be observed between POD scores and PET/P (Fig. 8b). No correlations of either NSE or POD were found with topographic attributes. For example, neither NSE nor POD shows a clear correlation with the h-gradient (used as a proxy for topographic steepness, Fig. 8c and d).

5. Discussion

5.1. Use of the ED-LSTM in the PUB context

The PUB model using the regional ED-LSTM structure showed a higher skill in predicting high streamflow than the conceptual hydrological model (Fig. 4 and Table 1). Even in catchments with poor prediction performance (i.e. NSE $<$ 0.5), where the streamflow is not perfectly simulated by the ED-LSTM model, decent performances for flood warning were obtained (i.e. high POD values). It is further evident in the results that the PUB model is not sensitive to the number of training catchments and always keeps a good flood warning skill (Fig. S5). This implies that while the deep-learning models do not always learn the streamflow dynamic (i.e. the time series as a whole) well, they can still extrapolate the extreme streamflow events – this is further seen when plotting the correlation between NSE and POD (Fig. S2). It also implies that there is a high similarity in the intense observed rainfall between the catchments, which triggers flood peaks of similar magnitude. There may be a potential to use deep learning for flood prediction in ungauged catchments, even if the number of gauged catchments is small. In other words, the ED-LSTM appears to be a reliable flood warning model in ungauged catchments since it does not require successfully capturing the entire event hydrograph and the timing of the peak but rather solely forecasting the magnitude of the peak.

Given a similar size of training data, the prediction ability of an LSTM model is often much better than physics-based models (Kratzert et al., 2018; Frame et al., 2022). While the physically-based regionalization method selects and transfers the optimal parameter set from a gauged catchment to an ungauged catchment (Yang et al., 2018), more flexibility is found with the ED-LSTM PUB approach as the model adapts to different dynamic patterns rather than be limited to a fixed set of parameters obtained from a donor catchment. These processes of the proposed approach are spontaneous and do not require the identification of any criteria for selecting donor catchments, resulting in a potentially better performance of the ED-LSTM in PUB predictions. For example, here we obtained POD values of above 0.8 and 0.7 for the prediction of high streamflow at 2- and 10-year return periods (Table 1), while in previous studies (using various physically-based methods but for different locations, climates, and time scales) obtained values lower than 0.48 (2-year return period, France, Javelle et al., 2016), 0.61 (2-year return period, 6 catchments in France and Italy, Norbiato et al., 2008), and 0.38 (10-year return period, Pakistan, Kim et al., 2018). We conclude that the proposed ED-LSTM PUB model can be considered an applicable tool for issuing fast and reasonably accurate flood warnings in ungauged catchments.

5.2. Model sensitivity to the training dataset

We found that the overall PUB model performance is better when training under a larger and more climate-heterogeneous dataset rather than using a smaller and regional-focused (i.e. climate-homogenous) dataset, as implied from the slight performance decrease comparing the south-north based PUB models (Fig. 5a) and the meaningful

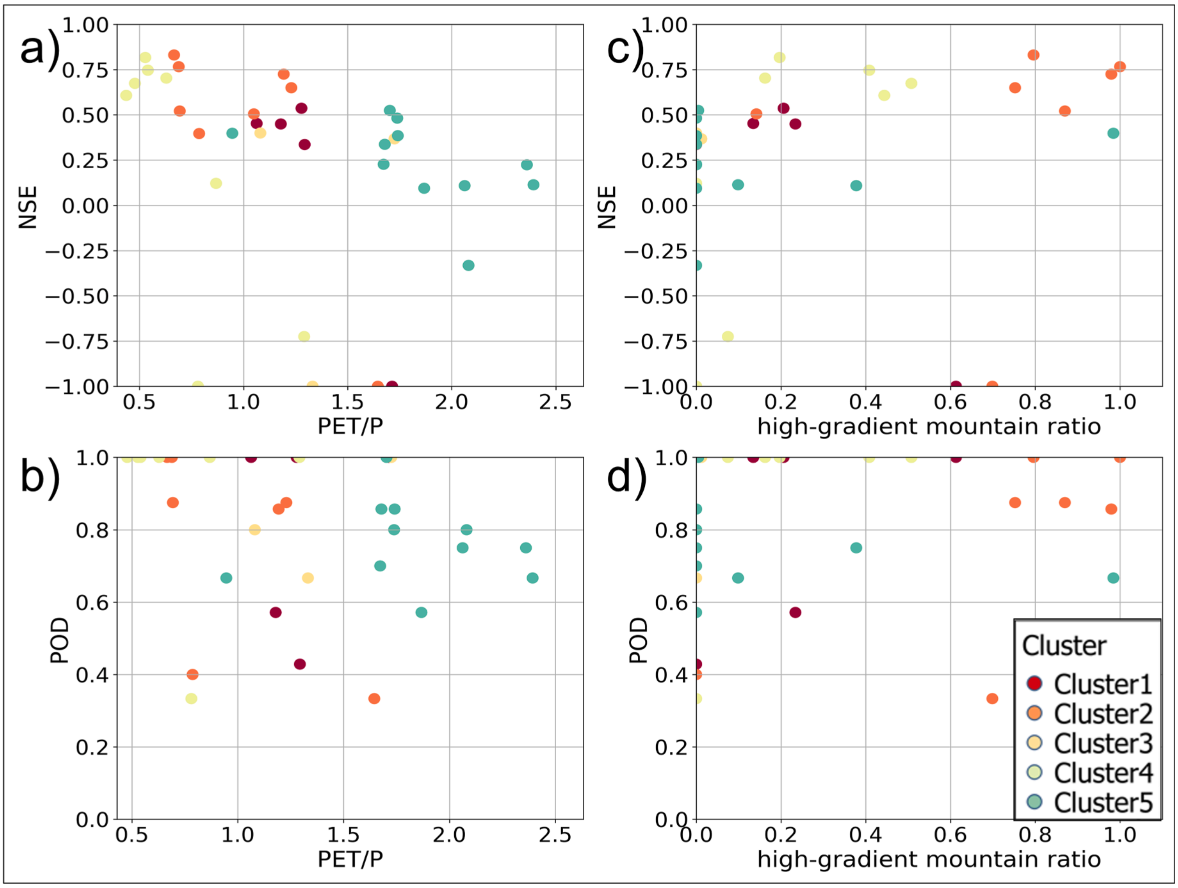


Fig. 8. (a) Model performance (NSE) of the PUB model as a function of climate condition (PET/P); (b) the same but for POD of 2-year return period events instead of NSE; (c) and (d) are the same as (a) and (b) but model performance is plotted as a function of topography (h-gradient). Each dot represents a catchment, and the color shows the clustering group of the catchment as shown in Fig. 7.

performance decline of the cluster-based PUB models (Fig. 5b and Fig. S4). Since the characteristics of the catchments in the different clusters vary significantly, the models trained solely on the data from the catchments in each of the single clusters are not able to generalize well when applied to catchments from clusters with other characteristics. This behavior is expected and is in line with the hypothesis proposed by Fang et al. (2022) that in the context of PUB, the DL models can still benefit from the data synergy effect provided by the modest diversity in the training data. A more heterogeneous dataset may increase the probability of covering the relevant conditions for a new catchment to increase the representativeness of various rainfall-runoff processes.

The sensitivity of the ED-LSTM model to simulate a time series of streamflow concerning the number of training events (or catchments) is non-linear but with a positive correlation for NSE scores as shown in Fig. 5 and Fig. S4. The decrease in the ED-LSTM NSE performance with decreasing number of catchments used for the training (Fig. S4) can be potentially explained by the fact that catchments that are not well represented by the training dataset are more sensitive to the change in training dataset size. This is evident in the results, as the upper quartile (the well-represented catchments) in PUB1, PUB2, and PUB3 remain the same, while the lower quartile drops significantly (Fig. S4).

The results suggest that both quantity and diversity of the training dataset for DL models are equally important for improving PUB model performance. We stress that future applications of machine learning models in PUB context should ensure a representative dataset with a sufficiently large number of training samples and catchments to properly include the impacts emerging from catchment characteristics on overall PUB model performance.

However, some of our results challenge the hypothesis that DL models' performance is not compromised by additional information for streamflow simulation, even when they appear to have different hydroclimatic conditions (Fang et al., 2022). The performances of several catchments, especially for those with sufficiently representative training data, were also decreasing (Fig. 6) from local to the PUB model. A possible explanation is the lack of data on the pseudo-ungauged catchments when conducting out-of-sample prediction. It means that the hydrological responses in these catchments are unique and the enlarged dataset still fundamentally lacks critical inputs so, in the PUB context, the prediction results were much poorer than the performance of the local model. Meanwhile, the performance decrease also happens between local and regional models due to the addition of some poorly-performing catchments without sufficiently representative training data, which have irregular hydrological responses in comparison to the well-performing catchments, to the enlarged dataset. This even happens in a single catchment. For example, the flood hydrographs measured in the Xinghe catchment show strong non-stationaries, whereas for similar rainfall intensities the resulting flood waves are very different, which finally results in a very low prediction skill even for a local model. The diverse hydrological behaviors, in this case, can impair the learning. This drop in performance of some catchments in the regionalization of LSTM models is also reported and discussed by Kratzert et al. (2019) and Hashemi et al. (2021). However, we did not quantify to what extent the heterogeneity may harm the model performance in this study. Such analyses are left for further research. Nevertheless, when adding new data, the drop in performance of the well-performed catchments is minor compared to the benefits of increasing the performance of those poorly-represented catchments.

5.3. Model sensitivity to the catchment characteristics

Previous studies have found that the prediction of streamflow in ungauged catchments tends to be better in humid areas for either physics-based or data-driven models (Ragettli et al., 2017; Kratzert et al., 2018; Feng et al., 2020; Lees et al., 2021) and we confirm this finding with our results (e.g. the NSE – PET/P relation, Fig. 8 and Table 1). There is a physical explanation for this finding: in dry regions, runoff generation is more likely to occur due to infiltration excess. Soil infiltration capacity has large spatial variability in 35 catchments so different soil saturation states can result in different streamflow magnitude and timing for the same rainfall intensity. This can explain the variability in streamflow responses within one catchment, for example, the Xinghe catchment described in section 5.2. Hence, it is likely that streamflow simulation in dry PUB areas can be improved if the machine learning model is trained to learn the interaction between rainfall and streamflow with a larger sample of different rainfall-streamflow-soil moisture conditions. Alternatively, deep-learning models can be improved to simulate runoff infiltration excess by introducing physical laws, as discussed in the next section.

We found that the model performance is only sensitive to climatic variables and not to topographic variables (Fig. 8) when considering NSE as model evaluation metrics. This is in agreement with the statement of Addor et al. (2018) and Stein et al. (2021), who concluded that streamflow behavior across regions is most strongly influenced by climate attributes for flood prediction in ungauged catchments. Our findings, thus, support the need to incorporate more dynamic and static climatic variables that can increase the representativeness of the dataset when setting regional machine-learning models for flood warnings.

5.4. Limitations and future development

The analyses we conducted are limited by the small number of events that were available for the model training and evaluation. For example, we used the 10-year return period as the representative of a high streamflow event but in 11 of the catchments, the number of events is not sufficient to estimate the 10-year return period with high accuracy. Another limitation is that the models were trained by event-based data. If continuous streamflow data is available and used instead, the model can potentially learn better the hydrological patterns, such as the streamflow seasonality and event-antecedent soil moisture conditions (which has the potential to improve predictions in dry climates, see the previous section). A complete observation time series covering 5 to 10 years would likely be sufficient to represent the non-stationary behind the hydrological dynamics (O et al., 2020) but it was not available to us for this study. Extending the data for the training of the models will result in a better prediction, but will not change the main conclusions of this work, namely, the outperformance of DL models in comparison to a physics-based hydrological model in predicting floods in general and in the context of ungauged catchments in particular.

The results of our experiments imply that it is essential to focus on developing model structures that can be adaptable also in dry regions. This can be done by incorporating governing equations or physical constraints into “hydrological” machine learning models. An example is the development of the Mass-Conserved DL model that incorporates conservation law into the loss function (Hoedt et al., 2021). Another alternative is the development of physically-informed hybrid models that embed the hydrological dynamics into the recurrent neural network architecture (e.g. Jiang et al., 2020). These types of models can better capture the interaction between soil-rainfall-evaporation and streamflow, and be more readily generalized beyond the regimes covered with the training data (Khandelwal et al., 2020). Future applications of DL models in the field of hydrology should be combined with such hydrological knowledge and physics.

6. Conclusions

We applied the Encoder-Decoder LSTM model to predict rainfall-runoff events and flood peaks in ungauged catchments. The model outperformed conventional hydrological-model regionalization methods. The most considerable improvement in the model predictive ability was observed in the poorly represented catchments. By evaluating the generalization ability, i.e. the applicability of the model across many catchments and conditions, we found that the performance of the ED-LSTM model was not only sensitive to the number of samples used for the training of the model but also the representativeness (climate-heterogeneity level) of the dataset. Also, the DL regional model still suffers from issues of model adaptability – although the ED-LSTM model reliably predicts the occurrence of rare events also in arid regions, it is more likely to have a poor model performance for predicting streamflow in arid catchments than in humid catchments. Surprisingly, we discovered that the catchment topographic attributes, such as elevation and gradient, did not improve the model performance when added as static variables in the model setup. We conclude that, compared to conventional methods, the regional ED-LSTM model is a promising method for hydrological modeling in ungauged catchments, and our results could be an important reference for further studies of DL-based hydrological modeling with a rather limited amount of data to set a representative training dataset.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The code of Encoder-decoder LSTM modeling is available through GitHub (https://github.com/yikuizh/edlstm_flood_prediction). Rainfall and runoff data from ground stations in China that were used for this study are not freely available for academic or commercial use (contact SR for further details).

Acknowledgments

The training of the ED-LSTM models was done on Google Colab, a free online platform for deep learning study courtesy of Google. NP acknowledges the support of the Swiss National Science Foundation (SNSF), Grant 194649 (“Rainfall and floods in future cities”).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2022.128577.

References

Addor, N., Nearing, G., Prieto, C., Newman, A.J., Le Vine, N., Clark, M.P., 2018. A Ranking of Hydrological Signatures Based on Their Predictability in Space. Water Resources Research 54 (11), 8792–8812. https://doi.org/10.1029/2018WR022606.

Beck, H.E., van Dijk, A.I.J.M., de Roo, A.d., Miralles, D.G., McVicar, T.R., Schellekens, J., Adrian Bruijnzeel, L., 2016. Global-Scale Regionalization of Hydrologic Model Parameters. Water Resources Research 52 (5), 3599–3622. https://doi.org/10.1002/2015WR018247.

Blöschl, G., Sivapalan, M., 1995. Scale Issues in Hydrological Modelling: A Review. Hydrological Processes 9 (3–4), 251–290. https://doi.org/10.1002/hyp.3360090305.

Chang, F.-J., Tsai, Y.-H., Chen, P.-A., Coynel, A., Vachaud, G., 2015. Modeling Water Quality in an Urban River Using Hydrological Factors – Data Driven Approaches. Journal of Environmental Management 151 (March), 87–96. https://doi.org/10.1016/j.jenvman.2014.12.014.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical

Methods in Natural Language Processing (EMNLP) 1724–1734. https://doi.org/10.3115/v1/D14-1179.

Fang, K., Kifer, D., Lawson, K., Feng, D., Shen, C., 2022. The data synergy effects of time-series deep learning models in hydrology. Water Resources Research 58. https://doi.org/10.1029/2021WR029583 e2021WR029583.

Feng, D., Fang, K., Shen, C., 2020. Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. Water Resources Research 56. https://doi.org/10.1029/2019WR026793 e2019WR026793.

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

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 (January), 104926. https://doi.org/10.1016/j.envsoft.2020.104926.

Gong, J., Yao, C., Li, Z., Chen, Y., Huang, Y., Tong, B., 2021. Improving the Flood Forecasting Capability of the Xinanjiang Model for Small- and Medium-Sized Ungauged Catchments in South China. Natural Hazards 106 (3), 2077–2109. https://doi.org/10.1007/s11069-021-04531-0.

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, e1487. https://doi.org/10.1002/wat2.1487.

Hashemi, R., Brigode, P., Garambois, P.-A., Javelle, P., 2021. How can regime characteristics of catchments help in training of local and regional LSTM-based runoff models? Hydrol. Earth Syst. Sci. Discuss. https://doi.org/10.5194/hess-2021-511 [preprint].

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.

Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G., Hochreiter, S., Klambauer, G., 2021. Mc-lstm: Mass-conserving lstm. arXiv. https://doi.org/10.48550/arXiv.2101.05186.

Javelle, P., Organde, D., Demargne, J., Saint-Martin, C., de Saint-Aubin, C., Garandeau, L., Janet, B., Lang, M., Klijn, F., Samuels, P., 2016. Setting up a French national flash flood warning system for ungauged catchments based on the AIGA method. E3S Web Conf. 7, 18010–18021. https://doi.org/10.1051/e3sconf/20160718010.

Jiang, S., Zheng, Y., Solomatine, D., 2020. Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning. Geophysical Research Letters 46. https://doi.org/10.1029/2020GL088229 e2020GL088229.

Kao, I.-F., Zhou, Y., Chang, L.-C., Chang, F.-J., 2020. Exploring a Long Short-Term Memory Based Encoder-Decoder Framework for Multi-Step-Ahead Flood Forecasting. Journal of Hydrology 583, 124631. https://doi.org/10.1016/j.jhydrol.2020.124631.

Khandelwal, A., Xu, S., Li, X., Jia, X., Stienbach, M., Duffy, C., Nieber, J., Kumar, V., 2020. Physics guided machine learning methods for hydrology. arXiv. https://doi.org/10.48550/arXiv.2012.02854.

Kim, S., Paik, K., Johnson, F.M., Sharma, A., 2018. Building a Flood-Warning Framework for Ungauged Locations Using Low Resolution, Open-Access Remotely Sensed Surface Soil Moisture, Precipitation, Soil, and Topographic Information. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11 (2), 375–387. https://doi.org/10.1109/JSTARS.2018.2790409.

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., 2019. 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.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., Dadson, S.J., 2021. Benchmarking Data-Driven Rainfall-Runoff Models in Great Britain: A Comparison of Long Short-Term Memory (LSTM)-Based Models with Four Lumped Conceptual Models. Hydrology and Earth System Sciences 25 (10), 5517–5534. https://doi.org/10.5194/hess-25-5517-2021.

Moore, R.J., Bell, V.A., Jones, D.A., 2005. Forecasting for Flood Warning. Comptes Rendus Geoscience 337 (1), 203–217. https://doi.org/10.1016/j.crte.2004.10.017.

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. Transactions of the ASABE 50 (3), 885–900. https://doi.org/10.13031/2013.23153.

Nash, J.E., Sutcliffe, J.V., 1970. River Flow Forecasting through Conceptual Models Part I — A Discussion of Principles. Journal of Hydrology 10 (3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

Norbiato, D., Borga, M., Esposti, S.D., Gaume, E., Anquetin, S., 2008. Flash Flood Warning Based on Rainfall Thresholds and Soil Moisture Conditions: An Assessment for Gauged and Ungauged Basins. Journal of Hydrology 362 (3), 274–290. https://doi.org/10.1016/j.jhydrol.2008.08.023.

O, S., Dutra, E., Orth, R., 2020. Robustness of Process-Based versus Data-Driven Modeling in Changing Climatic Conditions. Journal of Hydrometeorology 21 (9), 1929–1944. https://doi.org/10.1175/JHM-D-20-0072.1.

Oudin, L., Andréassian, V., Perrin, C., Michel, C., Le Moine, N., 2008. Spatial Proximity, Physical Similarity, Regression and Ungaged Catchments: A Comparison of Regionalization Approaches Based on 913 French Catchments. Water Resources Research 44, (3). https://doi.org/10.1029/2007WR006240.

Potdar, A.S., Kirstetter, P., Woods, D., Saharia, M., 2021. Toward Predicting Flood Event Peak Discharge in Ungauged Basins by Learning Universal Hydrological Behaviors with Machine Learning. Journal of Hydrometeorology 22 (11), 2971–2982. https://doi.org/10.1175/JHM-D-20-0302.1.

Ragettli, S., Zhou, J., Wang, H., Liu, C., Guo, L., 2017. Modeling Flash Floods in Ungauged Mountain Catchments of China: A Decision Tree Learning Approach for Parameter Regionalization. Journal of Hydrology 555 (December), 330–346. https://doi.org/10.1016/j.jhydrol.2017.10.031.

Singh, R., Archfield, S.A., Wagener, T., 2014. Identifying Dominant Controls on Hydrologic Parameter Transfer from Gauged to Ungauged Catchments – A Comparative Hydrology Approach. Journal of Hydrology 517 (September), 985–996. https://doi.org/10.1016/j.jhydrol.2014.06.030.

Sivapalan, M., Takeuchi, K., Franks, S.W., Gupta, V.K., Karambiri, H., Lakshmi, V., Liang, X., McDONNELL, J.J., Mendiondo, E.M., O'connell, P.E., Oki, T., Pomeroy, J.W., Schertzer, D., Uhlenbrook, S., Zehe, E., 2003. IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an Exciting Future for the Hydrological Sciences. Hydrological Sciences Journal 48 (6), 857–880. https://doi.org/10.1623/hysj.48.6.857.51421.

Song, T., Ding, W., Jian, W.u., Liu, H., Zhou, H., Chu, J., 2020. Flash Flood Forecasting Based on Long Short-Term Memory Networks. Water 12 (1), 109. https://doi.org/10.3390/w12010109.

Stein, L., Clark, M.P., Knoben, W.J.M., Pianosi, F., Woods, R.A., 2021. How do climate and catchment attributes influence flood generating processes? A large-sample study for 671 catchments across the contiguous USA. Water Resources Research 57. https://doi.org/10.1029/2020WR028300 e2020WR028300.

Teutschbein, C., Grabs, T., Laudon, H., Karlsen, R.H., Bishop, K., 2018. Simulating Streamflow in Ungauged Basins under a Changing Climate: The Importance of Landscape Characteristics. Journal of Hydrology 561 (June), 160–178. https://doi.org/10.1016/j.jhydrol.2018.03.060.

Xiang, Z., Yan, J., Demir, I., 2020. A rainfall-runoff model with LSTM-based sequence-to-sequence learning. Water Resources Research 56. https://doi.org/10.1029/2019WR025326 e2019WR025326.

Yang, X., Magnusson, J., Rizzi, J., Chong-Yu, X.u., 2018. Runoff Prediction in Ungauged Catchments in Norway: Comparison of Regionalization Approaches. Hydrology Research 49 (2), 487–505. https://doi.org/10.2166/nh.2017.071.

Zhao, Z., Li, S., Liu, J., Peng, J., Wang, Y., 2015. The Distance Decay of Similarity in Climate Variation and Vegetation Dynamics. Environmental Earth Sciences 73 (8), 4659–4670. https://doi.org/10.1007/s12665-014-3751-2.