https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

Streamflow prediction in data-scarce regions with semi-supervised deep learning

Tianlong Jia [1], Guoding Chen [2], Yao Li [3], Xinyu Chang [4,5,1], and Uwe Ehret [1]

[1]Karlsruhe Institute of Technology (KIT), Institute of Water and Environment, Karlsruhe, Germany
[2]Zhejiang Institute of Hydraulics and Estuary (Zhejiang Institute of Marine Planning and Design), Hangzhou 310020, China
[3]LEESU, ENPC, Institut Polytechnique de Paris, Univ Paris Est Creteil, 77455 Marne-la-Vallée, 10 France
[4]Huazhong University of Science and Technology, School of Civil and Hydraulic Engineering, Wuhan 430070, China
[5]Huazhong University of Science and Technology, Hubei Key Laboratory of Digital River Basin Science and Technology, Wuhan 430070, China

Correspondence: Tianlong Jia (tianlong.jia@kit.edu) and Guoding Chen (guoding.chen94@gmail.com)

Abstract.

Deep learning methods have demonstrated great performance in streamflow prediction. However, they typically require large amounts of "labeled" data for supervised learning (SL), including meteorological forcing data paired with corresponding streamflow observations. The data scarcity of streamflow observation limits application of SL models across hydrologically diverse regions worldwide. To address this issue, we propose a two-stage semi-supervised learning (SSL) for streamflow prediction in data-scarce regions, based on the Contrastive Predictive Coding (CPC) method. CPC is a self-supervised learning method, that learns data representations from "unlabeled" data (i.e., meteorological forcing time series without streamflow observations). In the first stage, CPC was used to pre-train an encoder and a Long Short-Term Memory (LSTM) network with a projection head, using a large number of meteorological sequences. In the second stage, we attached a linear layer to the pre-trained encoder and LSTM, and fine-tuned the entire model architecture for streamflow prediction, using labeled data. We developed and evaluated this approach for streamflow prediction in both regional models and single-basin models, using the CAMELS-DE dataset. We assessed the in-domain generalization performances of regional models on 1,265 basins in Germany, used to pre-train and fine-tune models. Moreover, we examined their zero-shot out-of-domain generalization performances, on additional 317 basins from CAMELS-DE, that were not involved in model training. We benchmarked our approach with a baseline SL-trained model. Our results show that the SSL regional models outperforms the SL baseline in both in-domain and zero-shot out-of-domain generalization performance for data-scarce conditions, when less than 10% of one-year labeled sequences are available. SSL models yield significant improvements in median Nash-Sutcliffe Efficiency (NSE) of 0.137 (in-domain) and 0.139 (out-of-domain), with 0.5% of one-year labeled data. Additionally, SSL enhances model ability to predict low flow and floods for data-scarce conditions, reducing the median percent bias of the bottom 30% low flow range (FLV) by 21.047 and the median Mean Absolute Percentage Error of peaks ($MAPE_{peak}$) by 13.933 (out-of-domain), with 1% of one-year labeled data. This improved performance stems from the informative feature representations learned from meteorological forcing inputs though CPC pre-training, that enhances prediction ability across diverse basins under data-scarce conditions. Moreover, the advantages of SSL are more obvious for single-basin models when one-year labeled data

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

is available. These results indicate a promising direction for leveraging SSL to develop hydrological foundation models, that have recently revolutionized artificial intelligence research. Hydrological foundation models involve pre-training on large-scale meteorological forcing datasets using self-supervised learning methods (e.g., CPC), and can be adapted to multiple hydrological tasks by model fine-tuning, e.g., simulation of water temperature and soil moisture.

1 Introduction

Accurate streamflow prediction is essential for effective hydropower generation, flood control, agricultural planning, and water resource management (Ng et al., 2023; Zhong et al., 2024; Jia et al., 2019). In recent years, rapid climate change has intensified water stress and water-related disasters (Li et al., 2023; Nearing et al., 2024). A recent study estimated that floods affected 1.6 billion people worldwide and caused extensive infrastructure damage in 2020 (Rogers et al., 2025). Therefore, improving the accuracy of streamflow prediction is especially important for regions vulnerable to flooding.

Hydrological forecasting models, which predict river streamflow from meteorological forcing and runoff data, are vital for timely warnings and effective flood risk mitigation (Nearing et al., 2021). Conventional streamflow forecasting approaches, including conceptual and physically based models (Ng et al., 2023), require extensive parameter calibration using long-term time series data (Jaiswal et al., 2020; Chen et al., 2023, 2024). These constraints limit their applicability across large, heterogeneous, and ungauged catchments. Currently, deep learning methods, especially the Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), offer an efficient alternative by capturing both short- and long-term temporal dependencies in hydrological time series. LSTM models can learn the underlying relationships between meteorological forcing and streamflow across multiple catchments, through training on large-scale, multi-basin datasets with long observational records, such as the Catchment Attributes and MEteorology for Large-sample Studies (CAMELS) dataset (Addor et al., 2017). Several studies have shown that LSTM models outperform models in rainfall-runoff modelling (Kratzert et al., 2018; Mai et al., 2022) and flood forecasting (Nearing et al., 2024; Zhang et al., 2022). Despite promising results, accurate and robust LSTM-based streamflow prediction requires large amounts of "labeled" data for supervised learning (SL). These labeled datasets consist of meteorological forcing inputs with streamflow records, that act as the ground truth. Several recent studies (Fathi et al., 2025; Staudinger et al., 2025) demonstrated that LSTM performance declines as the training period length decreases, e.g., from 50 to 7 years (Fathi et al., 2025). However, only a small fraction of the world's catchments are gauged (Nearing et al., 2024), leaving most areas with plentiful "unlabeled" meteorological forcing records but sparse streamflow measurements. Such data scarcity constrains the performance of SL models in data-limited regions, and limits their robust generalization across diverse catchments with varying meteorological, hydrological and geographical conditions.

To partially address this limitation, recent studies have proposed transfer learning approaches, and demonstrated their effectiveness for streamflow prediction in data-scarce basins (Ma et al., 2021; Khoshkalam et al., 2023; Ouyang et al., 2025). These approaches generally consist of two stages: (1) pre-training a regional model on a large-scale dataset, including a large number of "base" catchments within a given region to learn general data representations, and (2) transferring the obtained knowledge to a target (local) model, which is subsequently fine-tuned using data from the target basin to capture catchment-specific charac-

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

teristics (Kratzert et al., 2018). During regional model training, catchment attributes (e.g., soil characteristics, and land cover) are incorporated as additional inputs to represent spatial heterogeneity in hydrological processes across different basins. While transfer learning is powerful, its effectiveness declines under certain conditions: (i) when the similarity between the base and target datasets decreases; and (ii) when data availability in the target basins is severely limited. For example, Khoshkalam et al. (2023) employed an LSTM-based hydrological model, that was regionally pre-trained on the CAMELS dataset. Then, they fine-tuned the model using data from eight regions in Canada. The fine-tuned model outperformed locally trained models that were solely trained on Canadian basins. However, they also showed that transfer learning approaches achieve more acceptable performance when the meteorological and physiographic variables are consistent between the base and target datasets than when these variables differ. Similarly, Ouyang et al. (2025) transferred knowledge from U.S. basins to two data-scarce basins in China. The transfer learning approach achieved improvements in Nash-Sutcliffe Efficiency (NSE) of 0.02 and 0.06 compared to a locally trained LSTM model. Nevertheless, the resulting performance (NSE = 0.71 and 0.62) remained constrained by limited data availability in the two target basins, with only seven years of observations.

To overcome the limitations of SL, recent studies in machine learning have increasingly focused on self-supervised and semi-supervised approaches, which offer improved generalization performance while requiring fewer labeled samples for training (Ohri and Kumar, 2021; Zhai et al., 2019). Self-supervised learning leverages unlabeled input data and automatically generates its own labels, enabling models to learn underlying data representations without explicit external labels (Ohri and Kumar, 2021). In recent years, contrastive self-supervised learning has attracted growing interest within the computer vision community (Kumar et al., 2022). These approaches learn data representations by classifying positive samples (i.e., samples with similar representations) against negative samples (i.e., those with dissimilar representations). Semi-supervised learning (SSL) typically combines self-supervised pre-training with a limited number of labeled samples for fine-tuning on specific downstream tasks (Zhai et al., 2019). Recent studies have demonstrated that SSL approaches improve model performance and data efficiency compared with conventional SL methods across various domains, such as transportation (Jiao et al., 2025), sewer defect detection (Yildizli et al., 2026), and microparticles monitoring (Jia et al., 2025b). Although SSL methods have shown considerable promise, SSL methods have not been applied in streamflow prediction and rainfall-runoff modeling at all.

To the best of our knowledge, this study is the first to propose SSL approaches for streamflow prediction based on self-supervised learning methods, aiming to address the challenge of prediction in data-scarce regions. We developed and evaluated SSL models at both the regional and single-basin scales on CAMELS-DE dataset (Loritz et al., 2024). Moreover, we evaluated the ability of SSL methods to generalize to both previously seen (in-domain) and unseen basins (out-of-domain), with varying degrees of data scarcity.

In summary, we conducted three experiments to address the following research questions (Q1)–(Q5):

- Q1: Can SSL methods improve streamflow prediction in data-scarce regions, compared to conventional SL methods?
- Q2: How does the relative performance of SSL and SL approaches depend on the amount of labeled data? Can data availability thresholds be identified to support the selection between SSL and SL models?
- Q3: Are SSL methods more effective for in-domain or out-of-domain streamflow prediction?

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

- Q4: Are SSL methods more effective for regional or single-basin models?
- Q5: Are SSL methods more effective for low flow or flood prediction?

By addressing the above research questions, this study seeks to advance the development of hydrological foundation models, that support multiple hydrological tasks (e.g., simulation of water temperature, groundwater level, and soil moisture).

2 Methodology

2.1 Overview of the semi-supervised learning approach for streamflow prediction

We propose a two-stage semi-supervised learning (SSL) approach for streamflow prediction, based on the Contrastive Predictive Coding (CPC) method (Oord et al., 2018). This approach consists of a self-supervised pre-training stage and supervised fine-tuning stage. Fig. 1 demonstrates the schematic illustration of the SSL approach. In the first stage, the CPC method is sued to pre-train an encoder (e.g., a fully connected neural network) and an LSTM network followed by a projection head (e.g., a linear layer), on a large amount of meteorological forcing data (e.g., precipitation, temperature, and radiation) and catchment attribute data. We further explain the CPC method in Section 2.2. In the second stage, a new model architecture is created by attaching an additional linear layer to the pre-trained encoder and LSTM model. Then, this new model is fine-tuned for a specific streamflow prediction task in a supervised manner, using a limited set of meteorological forcing inputs paired with gauged streamflow data, as well as catchment attribute data. Section 2.3 and 2.4 show detailed descriptions of the self-supervised pre-training and supervised fine-tuning procedures used in this study, respectively.

2.2 Contrastive Predictive Coding

CPC is a self-supervised learning method, that learns data features without manual labels, by predicting the latent representations of future observations from past context. It has demonstrated promising performance in speech recognition, natural language processing, and computer vision applications (Le-Khac et al., 2020).

Fig. 2 shows the schematic illustration of the CPC approach. First, a non-linear encoder network (e.g., fully connected neural network) is used to extract a sequence of the latent representations z from the input sequence x. The input sequence includes meteorological forcing inputs (e.g., precipitation P_t and temperature T_t) at each timestep and static catchment attributes s. Second, the sequence of latent representations z is processed by an LSTM model and a projection head (e.g., a linear layer) to obtain a sequence of context latent representations c, that contains the historical information of the sequence data. Third, a context latent representation c_t is used to predict multiple "future" latent representations z. For simplicity, we only presented a single future latent representation z_{t+k}, where $k > 0$. The prediction is performed through a linear transformation, yielding the predicted latent representation $\hat{z}_{t+k} = W_k c_t$, where W_k denotes the prediction matrix. The sample x_{t+k} is considered as the positive sample corresponding to x_t, while negative samples x_{t^*} are randomly selected from the same input sequence or other sequences in the mini-batch (Henaff, 2020). Finally, the encoder, LSTM model and projection head are trained using the Information Noise Contrastive Estimation (infoNCE) loss function (Oord et al., 2018), that aims to correctly identify the latent

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

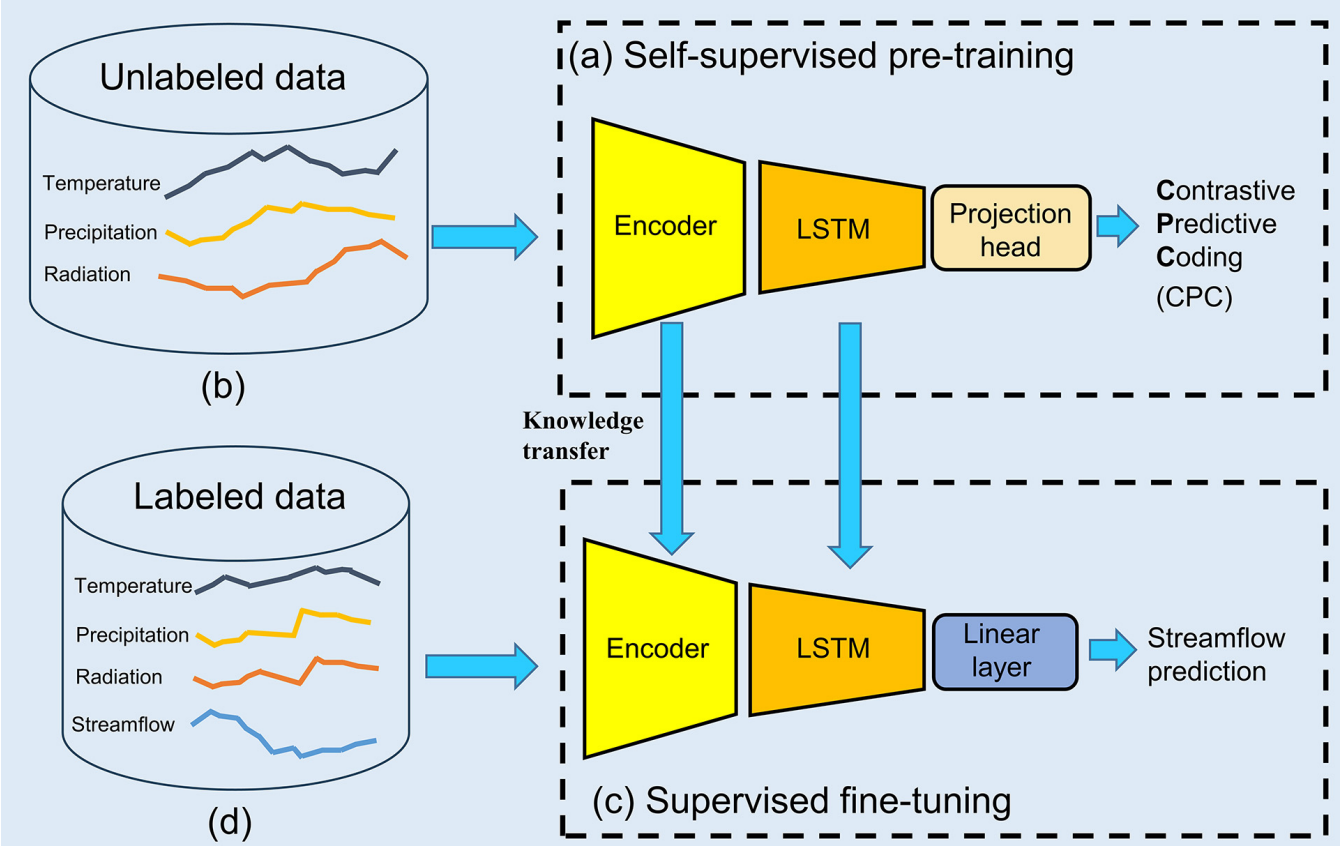

Figure 1. The schematic illustration of the semi-supervised learning (SSL) approach. In the self-supervised pre-training stage (a), the Contrastive Predictive Coding (CPC) method is employed to pre-train an encoder, LSTM model followed by a projection head, using a large amount of meteorological forcing data (e.g., precipitation, temperature, and radiation) and catchment attribute data (b). Then, a linear layer is added to the pre-trained encoder and LSTM model to create a new model architecture. In the supervised learning stage (c), this new model is fine-tuned to conduct a specific streamflow prediction task in a supervised manner, using a limited amount of meteorological forcing data paired with gauged streamflow data, as well as catchment attribute data (d).

representation of the positive sample x_{t+k}, from a set of negative latent representations. Specifically, the similarity between the prediction $W_k c_t$ and the latent representation z_{t+k} of the positive sample is maximized, while its similarity to the latent representation z_{t^*} of negative samples is minimized. The loss function is defined, as follows:

$$\mathcal{L}_N = -\mathbb{E}_X \left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)} \right] \quad (1)$$

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

where N denotes the total number of samples, including one positive and $N-1$ negative samples, given a set $X = \{x_1, x_2, \ldots, x_N\}$. The function $f_k(x_{t+k}, c_t)$ quantifies the agreement between the sample x_{t+k} and the latent representation c_t. It is computed as follows:

$$f_k(x_{t+k}, c_t) = \exp\left(z_{t+k}^{\top} W_k c_t\right) \tag{2}$$

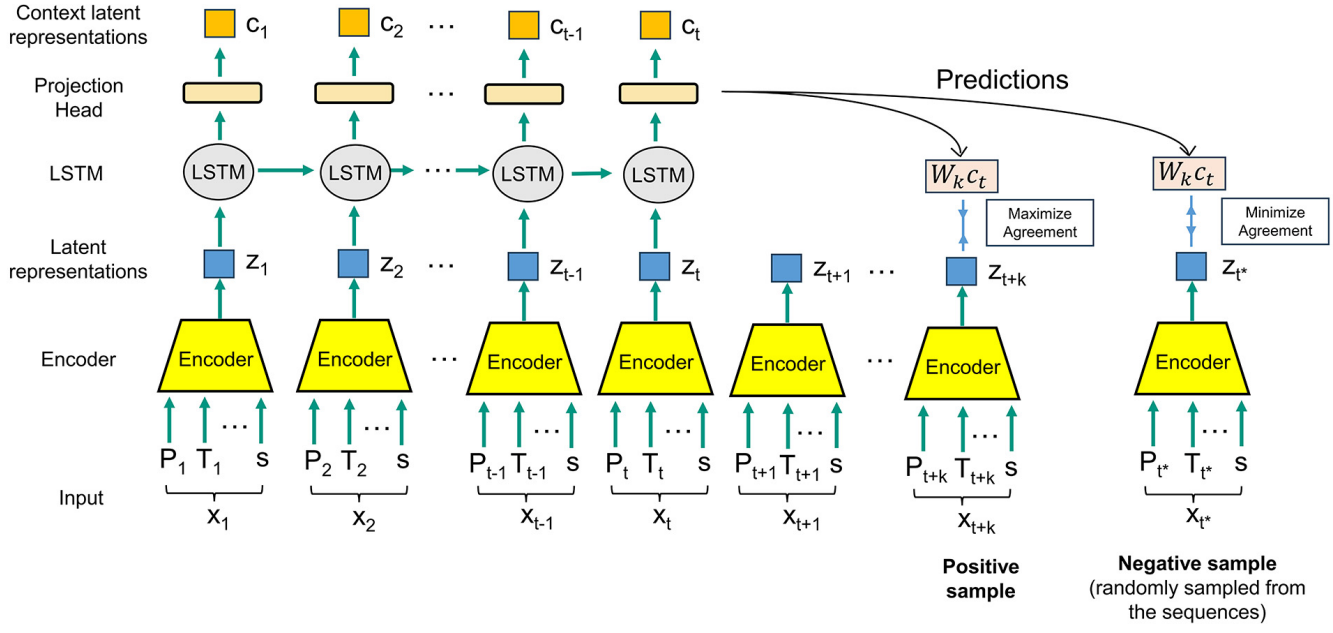

Figure 2. The schematic illustration of the Contrastive Predictive Coding (CPC) approach. First, a non-linear encoder extracts latent representations z from the input sequence x. Second, the latent representations z are processed by an LSTM model and a projection head to produce context latent representations c. Third, a context latent representation c_t is used to predict "future" latent representations z_{t+k} ($k > 0$) via a linear transformation $\hat{z}_{t+k} = W_k c_t$. The true future sample x_{t+k} serves as the positive sample, while negative samples are randomly chosen from the same or other sequences within the mini-batch. The encoder, LSTM model, and projection head are trained using a loss function to maximize agreement between the prediction $W_k c_t$ and the latent representation z_{t+k} of the positive sample, and minimize agreement with latent representations z_{t^*} of negative samples.

2.3 CPC pre-training

In this study, we employed an encoder including four fully connected layers, each with 128 neurons and ReLU activation. Inspired by Oord et al. (2018), we used a single-layer LSTM with 128 hidden units, and a projection head consisting a linear layer with 128 neurons without bias. During the CPC pre-training process, we assigned one positive sample and 128 negative

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

samples to each input x_t. These negative samples were randomly selected from both the same input sequence and other sequences within the same batch.

2.4 Fine-tuning for streamflow prediction

We constructed a new model architecture, by attaching a linear output layer with 128 neurons (including bias) to the pre-trained encoder and LSTM. Inspired by Jia et al. (2023) and Ouyang et al. (2025), we did not freeze any layer of this architecture during fine-tuning. Thus, all model weights were updated using a labeled dataset. While conventional LSTM models for streamflow prediction typically do not include an encoder to pre-process input data, we integrated the encoder into the SSL models, to transfer its feature knowledge learned from unlabeled data to streamflow prediction tasks.

3 Experiments

We designed three experiments to evaluate the potential of SSL for streamflow prediction in both *regional models* (jointly trained on multiple basins) and *single-basin models* (trained on each individual basin). We performed daily streamflow prediction at time step t, using meteorological forcing from the preceding one-year period (i.e., 365 days, from t-364 to t). The models produce a single predicted streamflow value, without generating a probability distribution.

We evaluated *in-domain* and zero-shot *out-of-domain* generalization capability of regional models in Experiment 1 and 2, respectively. In-domain generalization refers to the model performance on previously unseen data from the same basins but from time periods different from those used for training. In contrast, out-of-domain generalization refers to performance on unseen data from entirely different basins (i.e., ungauged basins). Zero-shot out-of-domain generalization further describes models' capability to predict streamflow for previously unseen basins without access to any training data from those basins (Jia et al., 2025a). Such capability is particularly important for large-scale, global streamflow and flood prediction, as it allows forecasting across multiple basins with diverse meteorological conditions and catchment characteristics, without requiring well-measured, location-specific streamflow data for further model refinement. More importantly, we assessed how the prediction performance varies with the amount of streamflow data available for fine-tuning. This analysis is particularly critical for evaluating the practical applicability of models in data-scarce basins and countries. In addition to evaluating overall model performance, we also assessed low flow and flood prediction accuracy in these two experiments. Furthermore, we compared the prediction results with those obtained from a SL benchmark using the same LSTM architecture in all experiments. Thus, we address Q1, Q2 and Q5 (see the end of Section 1) through both Experiment 1 and 2, while we address Q3 through a comparison of the results from both experiments.

In Experiment 3, we evaluated the in-domain generalization capability of multiple single-basin models. Thus, we further address Q1 and Q5 through this experiment, and address Q4 through a comparison of the results from this experiment and Experiment 1.

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

3.1 Dataset

In this study, we developed and validated models using the large-sample hydrological dataset CAMELS-DE (Loritz et al., 2024). This dataset comprises river streamflow and meteorological forcing time series for 1582 basins in Germany, spanning about 70 years (January 1951 - December 2020). Following the LSTM experimental setup described in Loritz et al. (2024), we used five dynamic (time-varying) input features, including (1) mean precipitation, (2) precipitation standard deviation, (3) mean global radiation, (4) mean minimum temperature, and (5) mean maximum temperature. Observed catchment-specific discharge is the target variable. Additionally, we included 17 basin-specific static (time-invariant) attributes as input features, as detailed in Appendix A.

Table 1 summarizes the subsets used to develop regional models in Experiment 1 and 2. In Experiment 1, we randomly selected 80% of basins (1,265 basins) from the CAMELS-DE dataset for developing SSL regional models. Fig. 3 (a) shows the spatial distribution of streamflow gauging stations within these basins. We extracted meteorological forcing time series without streamflow data using a sliding window approach from these basins over a 28-year period (1970-10-01 to 1998-12-31). We performed the sliding window approach, with a window length of 365 days and a stride of 30 days. This configuration results in a total of 435,160 sequences with an 11-month overlap between consecutive sequences. We selected this stride value to generate a sufficiently large yet computationally feasible dataset, given to the substantial size of CAMELS-DE dataset and the limited available computational resources. Next, we incorporated basin static attributes into the sequences, resulting in inputs that combine dynamic meteorological forcing (i.e., 365 days of five meteorological variables) with static features (i.e., 17 catchment attributes). Then, we randomly split these sequences into the $\text{Pre-train}_{\text{CPC}}$ (391,644 sequences) and Val_{CPC} (43,516 sequences) subsets at a ratio of 9:1. We used these two subsets for CPC pre-training and CPC validation, respectively.

Table 1. The subsets for model development in Experiment 1 and 2.

Learning method	Training dataset				Validation dataset				Test dataset			
	Name	Time period	No. sequences	Streamflow used	Name	Time period	No. sequences	Streamflow used	Name	Time period	No. sequences	Streamflow used
Self-supervised	$\text{Pre-train}_{\text{CPC}}$	1970-1998	391,644	No	Val_{CPC}	1970-1998	43,516	No				
Semi-supervised and supervised	$\text{Train}_{100\%}$	1999	411,469	Yes	$\text{Validation}_{\text{regional}}$	1965-1970	2,309,890	Yes	Test_{in}	2000-2020	9,703,815	Yes
	$\text{Train}_{80\%}$		329,175									
	$\text{Train}_{60\%}$		246,881									
	$\text{Train}_{40\%}$		164,587									
	$\text{Train}_{20\%}$		82,293									
	$\text{Train}_{10\%}$		41,146									
	$\text{Train}_{5\%}$		20,573						Test_{out}	2000-2020	2,431,707	Yes
	$\text{Train}_{1\%}$		4,114									
	$\text{Train}_{0.5\%}$		2,057									

We extracted meteorological forcing time series with corresponding streamflow data from different time periods to generate the $\text{Train}_{100\%}$ (1998-12-31 to 1999-12-31), $\text{Validation}_{\text{regional}}$ (1965-10-01 to 1970-09-30), and Test_{in} (2000-01-01 to 2020-

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

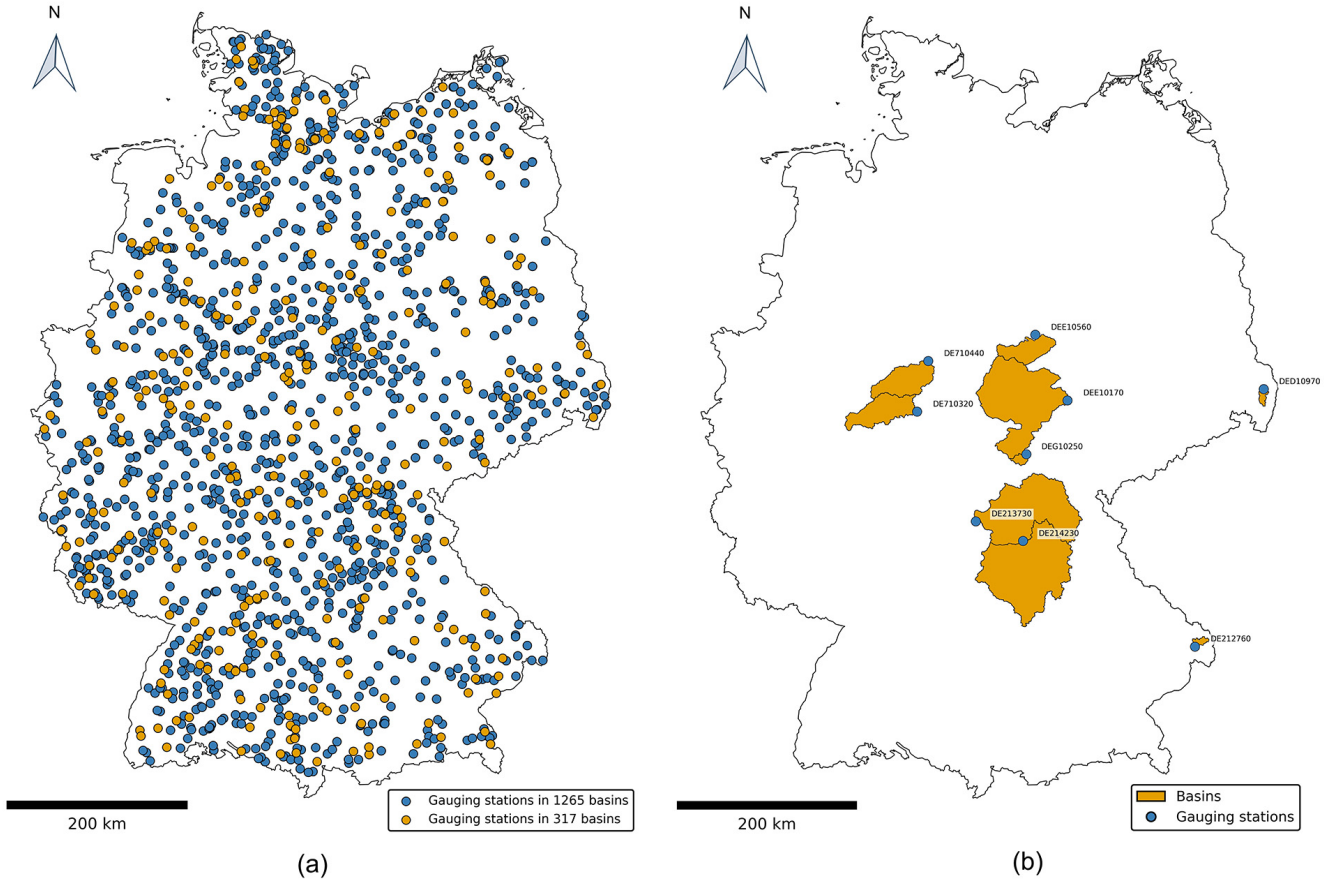

Figure 3. Spatial distribution of streamflow gauging stations: (a) across the 1,265 basins used for regional model development and the 317 basins used to evaluate out-of-domain generalization performance in Experiment 1 and 2, and (b) across the nine basins selected for Experiment 3.

12-31) subset. All sequences in these subsets consist of dynamic meteorological forcing (i.e., 365 days of 5 meteorological variables), static features (i.e., 17 basin attributes), and the daily streamflow value on the final day of each sequence. We used these subsets for model fine-tuning, validation and in-domain performance evaluation, respectively. To systematically evaluate the impact of streamflow data availability on model performance, we created nine additional smaller fine-tuning subsets by progressively reducing the number of training sequences down to 0.5% ($Train_{80\%}$ to $Train_{0.5\%}$), representing varying levels of data scarcity. To achieve this, we performed random sampling of labeled sequences from the $Train_{100\%}$ subset, as this strategy has been demonstrated to outperform targeted sampling approaches, e.g., high-flow-only sampling strategy (Heudorfer and Loritz, 2026). In Experiment 2, we extracted meteorological forcing time series with corresponding streamflow data from the remaining 20% of basins (317 basins, see Fig. 3 (a)) to create the $Test_{out}$ subset (2000-01-01 to 2020-12-31), that was used to evaluate models' out-of-domain generalization performances.

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

Table 2 presents the basin-specific subsets used to develop nine single-basin models in Experiment 3. We randomly selected nine basins with varying basin areas from the 80% subset of basins used in Experiment 1 (see Fig. 3 (b)). We extracted meteorological forcing time series without streamflow data from each basin for the period 1970-10-01 to 1998-12-31, using the sliding window approach with a stride of 7 days. That results in a relatively large number of samples (1,474 sequences) per basin. Following the subset construction procedure described above, we divided 1,474 sequences into a pre-training subset (1,327 sequences, Pre-train_1 to Pre-train_9) and a validation subset (147 sequences, Validation_1 to Validation_9) for self-supervised learning. Additionally, we generated fine-tuning (Train_1 to Train_9), validation (Val_1 to Val_9), and test (Test_1 to Test_9) subsets containing both meteorological forcing time series and streamflow data for the development and evaluation of single-basin models, as detailed in Table 2.

Table 2. The basin-specific subsets for model development in Experiment 3.

Basin	Basin area (km^2)	Pre-training dataset[1]			Fine-tuning dataset			Validation dataset			Test dataset		
		Name	Time period	No. sequences	Name	Time period	No. sequences	Name	Time period	No. sequences	Name	Time period	No. sequences
DE214230	6,980	Pre-train_1			Train_1			Val_1			Test_1		
DE213730	12,704	Pre-train_2			Train_2			Val_2			Test_2		
DE213730	12,704	Pre-train_2			Train_2			Val_2			Test_2		
DEE10170	6,224	Pre-train_3			Train_3			Val_3			Test_3		
DE710320	1,799	Pre-train_4	1970-1998	1,327	Train_4	1999	366	Val_4	1965-1970	1,826	Test_4	2000-2020	7,671
DE710440	1,762	Pre-train_5			Train_5			Val_5			Test_5		
DEE10560	1,207	Pre-train_6			Train_6			Val_6			Test_6		
DE212760	120	Pre-train_7			Train_7			Val_7			Test_7		
DEG10250	163	Pre-train_8			Train_8			Val_8			Test_8		
DED10970	126	Pre-train_9			Train_9			Val_9			Test_9		

[1] Here, we reported only the details of the pre-training datasets. The details of the validation subsets used in self-supervised learning process are provided in Section 3.1.

3.2 Implementation of experiments

We evaluated the in-domain and out-of-domain performance of regional models in Experiment 1 and 2, respectively. In addition, we investigated how model performance varies with the amount of streamflow data available. This analysis provides insights into the effectiveness of SSL methods in countries or regions with limited gauged streamflow data, but with abundant meteorological forcing data and sufficient computational resources for extensive hyperparameter tuning.

For developing SSL regional models, we pre-trained the encoder, LSTM, and projection head from scratch on the $\text{Pre-train}_{\text{CPC}}$ subset for 100 epochs using the CPC method, given our limited computational resources. Each pre-training epoch required around 1 hour under the training configuration and hardware described in Section 3.4. We computed the infoNCE loss on the Val_{CPC} subset, and saved the model weights yielding the minimum validation loss. These weights were subsequently transferred to streamflow prediction tasks for fine-tuning. Appendix B shows the loss at each pre-training epoch. Furthermore, we fine-tuned a new model architecture, built upon the pre-trained encoder and LSTM, using all 10 available subsets for SL,

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

ranging from $\text{Train}_{100\%}$ to $\text{Train}_{0.5\%}$. We performed model validation on the $\text{Validation}_{\text{regional}}$ subset. Finally, we selected the SSL model that achieved the highest validation accuracy, measured by the Nash–Sutcliffe Efficiency (NSE; see Section 3.3). Then, we assessed its in-domain and zero-shot out-of-domain generalization performance on the Test_{in} and Test_{out} subset, respectively.

To facilitate efficient pre-training, we standardized the input variables (i.e., meteorological forcings and catchment attributes) by subtracting the mean and dividing by the standard deviation (Kratzert et al., 2018). The normalization statistics were computed on the $\text{Pre-train}_{\text{CPC}}$ subset, as detailed in Appendix A. During the supervised fine-tuning stage, we applied the same standardization procedure to the labeled data, including meteorological forcings, catchment attributes, and observed streamflow data. In this case, the statistics were calculated from the $\text{Train}_{100\%}$ subset, and then used to pre-process the input data during validation and testing. The model outputs were finally transformed back using the corresponding streamflow normalization statistics to obtain streamflow predictions in physical units.

We compared the performance of the SSL regional models with baseline SL regional models, developed with the same supervised fine-tuning step (see Fig. 1 (b)), but without the CPC pre-training step (see Fig. 1 (a)). The baseline models include the same LSTM architecture used in SSL models, followed by a linear output layer. These models were supervised trained from scratch on meteorological forcing time series paired with streamflow data (i.e., $\text{Train}_{100\%}$ to $\text{Train}_{0.5\%}$). We validated and tested the baseline models on the same subsets used for SSL model development.

In Experiment 3, we evaluated the in-domain performance of nine single-basin models using SSL and baseline SL approaches. We developed these models following the same procedures as in Experiment 1, except that basin-specific datasets were used. For developing SSL single-basin models, we pre-trained and validated models on the corresponding basin-specific subsets (Pre-train_1 to Pre-train_9, and Validation_1 to Validation_9, respectively). Then, we fine-tuned, validated, and tested each model using the corresponding basin-specific subsets (Train_1 to Train_9, Val_1 to Val_9, and Test_1 to Test_9, respectively). For comparison, we supervised trained, validated and tested nine baseline SL models using the same subsets for SSL model development.

3.3 Performance evaluation

To evaluate the overall performance for a given basin b, we used several widely used metrics: (1) Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970), (2) Mean Squared Error (MSE), (3) Root Mean Square Error (RMSE), and (4) Kling-Gupta Efficiency (KGE) (Gupta et al., 2009). While these metrics provide a comprehensive assessment of overall model performance, they may obscure model deficiencies during critical hydrological conditions, such as floods. To more rigorously assess model performance under extreme flow conditions, we further considered additional metrics: (1) the percent bias of the top 2% peak-flow range (FHV) (Yilmaz et al., 2008), (2) the percent bias of the bottom 30% low-flow range (FLV) (Yilmaz et al., 2008), (3) peak timing error (Kratzert et al., 2021), and (4) the Mean Absolute Percentage Error of peaks ($\text{MAPE}_{\text{peak}}$) (Kratzert et al., 2022).

The FHV metric quantifies the difference in streamflow volume within the top 2% peak flow range between observed and predicted values, while FLV measures that within the bottom 30% low flow range. The peak-timing error and $\text{MAPE}_{\text{peak}}$ metric

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

measures the lag, and the average absolute percentage difference in streamflow volume between observed and predicted peaks, respectively. Among these metrics, FLV specifically evaluates model performance in low flow conditions, while the other three metrics assess performances under peak flow (flood) conditions. Detailed definitions and formulations of all metrics are provided in Appendix A.

When fine-tuning regional models, we used the basin-averaged NSE loss function (Kratzert et al., 2019), since it mitigates bias toward large and humid basins (with high flow variance) during the optimization process, and avoids poor model performance on small and arid basins (with low flow variance). The basin-averaged NSE is defined as follows:

$$\text{basin-averaged NSE} = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{N}\frac{(Q_{pred,i} - Q_{obs,i})^2}{(s_b + \epsilon)^2} \tag{3}$$

where B is the number of basins. s_b represents the standard deviation of observed streamflow in basin b, computed over the training period. ϵ is a numerical stabilizer (ϵ = 0.1).

During the validation and testing stages, we separately calculated the eight evaluation metrics described above for each basin (e.g., NSE and FHV). We reported and analyzed the median values of these metrics across multiple basins in Section 4.

3.4 Training setup and procedure

We trained and tested all models on a NVIDIA A100-40 GPU (40 GB). During the CPC pre-training stage, we set the positive sample corresponding to x_t as the input at the next timestep x_{t+1} (see Fig. 2). Thus, the prediction task involves using the context latent representation c_t to predict the latent representation of the subsequent timestep z_{t+1}. The latent representations learned through CPC pre-training are effective for the streamflow prediction task, that also involves next-step prediction. This alignment can enhance the streamflow prediction accuracy of the SSL models. We pre-trained models using CPC for 100 epochs, using the Adam optimizer (Kingma, 2014) with a batch size of 32, a fixed learning rate of 0.0001, and a weight decay of 0.001. For fine-tuning or training SSL and baseline SL models, we followed the LSTM experimental setup described in Loritz et al. (2024), including the Adam optimizer, a number of epochs of 20, a batch size of 256, a learning rate of 0.001, initial forget gate bias of 3, and dropout rate of 0.4. To mitigate the effect of randomization in the training procedure, we fine-tuned or trained SSL and SL regional models 10 times with different random seed. All metric values reported in Section 4.1 and 4.2 represent the median values calculated from these runs.

4 Results and Discussion

4.1 Experiment 1: In-domain generalization performance of regional models

Fig. 4 illustrates the in-domain generalization performance of regional models in NSE on the Test_{in} subset, using the SSL and SL methods. Additional evaluation metrics are summarized in Table 3. When limited labeled data is available (i.e., $\text{Train}_{0.5\%}$ to $\text{Train}_{5\%}$ subsets), the SSL method consistently outperforms the SL baseline, achieving an median NSE ranging from 0.331

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

to 0.551. This yields a significant improvement of up to 0.137 in median NSE, compared to the baseline model (median NSE=0.194∼0.528). Moreover, the advantage of SSL becomes more obvious as the amount of labeled data decreases, with the median NSE improvement increasing from 0.024 for $Train_{5\%}$ to 0.137 for $Train_{0.5\%}$. The SSL method achieves superior performance using only 0.5% of one-year labeled sequences (median NSE = 0.331, 2,057 meteorological forcing input sequences with streamflow observations), compared to the baseline SL method using 1% of one-year labeled sequences (median NSE = 0.327, 4,114 sequences with streamflow observations). These results demonstrate that transferring data representations learned through CPC pre-training on long-term meteorological forcing data (28 years, 1970-1998), is more effective for streamflow prediction tasks, than training an LSTM model from scratch for data-scarce scenarios, when less than 10% of one-year labeled sequences are available. This advantage arises from data representations learned through CPC pre-training, that capture diverse meteorological conditions (e.g. low, mean, and high precipitation or extremes) across multiple temporal scales (e.g. event, season, years). These representations enhance model ability to capture hydrological processes across diverse meteorological conditions, when streamflow observations are limited.

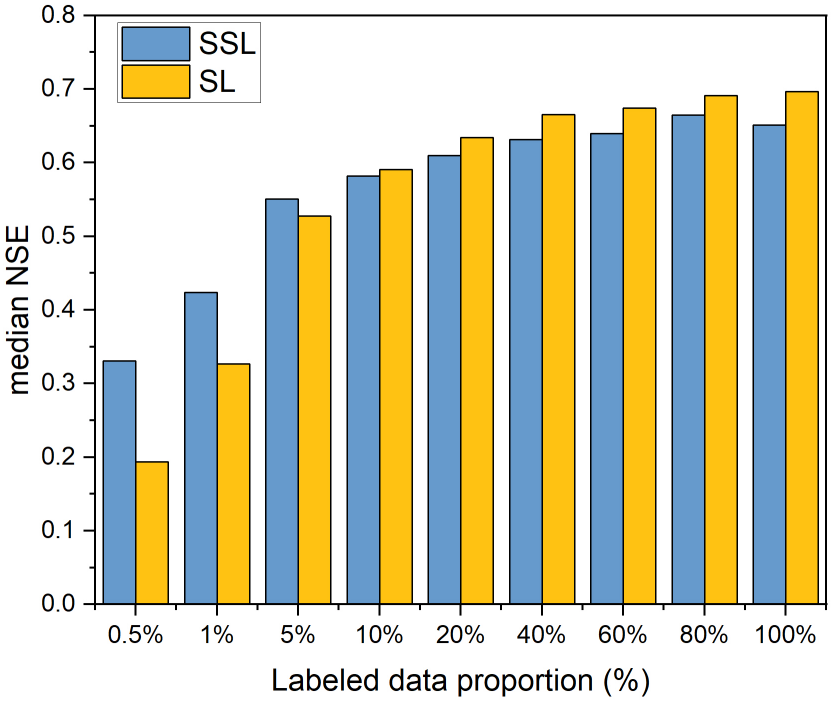

Figure 4. Median Nash-Sutcliffe Efficiency (NSE) achieved by regional models on the $Test_{in}$ subset, using the semi-supervised learning (SSL) and baseline supervised learning (SL) method, with different proportion of labeled data for model fine-tuning or training. The full labeled dataset (100%) includes 411,469 labeled sequences over a one-year period from 1,265 basins in the CAMELS-DE dataset.

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

Table 3. In-domain generalization performance of regional models using the semi-supervised learning (SSL) and supervised learning (SL) method across multiple metrics, fine-tuned on different amounts of labeled data. Median metric values are reported across basins in the $Test_{in}$ subset.

Training dataset	MSE		RMSE		KGE		FHV		FLV		Peak-timing error		$MAPE_{peak}$	
	SSL	SL	SSL	SL	SSL	SL	SSL	SL	SSL	SL	SSL	SL	SSL	SL
$Train_{0.5\%}$	**0.428**	0.548	**0.654**	0.740	**0.364**	0.272	**-45.461**	-59.452	**21.883**	37.753	**0.703**	0.976	**54.308**	66.974
$Train_{1\%}$	**0.367**	0.437	**0.606**	0.661	**0.477**	0.362	**-33.418**	-51.589	**13.734**	30.017	**0.558**	0.898	**46.972**	60.956
$Train_{5\%}$	**0.290**	0.294	**0.539**	0.543	**0.567**	0.546	**-20.718**	-30.361	**39.532**	44.003	**0.433**	0.486	**42.581**	46.151
$Train_{10\%}$	0.277	**0.260**	0.526	**0.510**	0.606	**0.609**	**-19.138**	-21.567	**31.647**	35.980	**0.410**	**0.410**	41.300	**41.287**
$Train_{20\%}$	0.252	**0.239**	0.502	**0.488**	0.624	**0.659**	-19.032	**-17.287**	33.816	**33.368**	0.406	**0.400**	**39.055**	39.622
$Train_{40\%}$	0.236	**0.217**	0.485	**0.466**	0.661	**0.688**	-17.754	**-16.540**	**30.696**	39.130	**0.391**	0.398	38.091	**37.937**
$Train_{60\%}$	0.238	**0.205**	0.487	**0.453**	0.671	**0.704**	-17.454	**-16.938**	**28.747**	33.737	**0.390**	**0.390**	37.865	**37.730**
$Train_{80\%}$	0.220	**0.200**	0.469	**0.448**	0.682	**0.718**	-17.893	**-14.833**	35.068	**34.821**	**0.377**	0.393	**36.380**	36.704
$Train_{100\%}$	0.227	**0.198**	0.477	**0.445**	0.689	**0.714**	-16.079	**-15.782**	33.424	**29.007**	**0.373**	0.390	37.018	**36.555**

The SL baseline surpasses the SSL approach, obtaining a median NSE between 0.591 and 0.697, when larger amount of labeled data is available (i.e., $Train_{10\%}$ to $Train_{100\%}$ subsets). In such scenarios, the SSL method exhibits a slight performance degradation, with a decrease in median NSE ranging from 0.009 to 0.045, compared to the SL baseline. Additionally, the performance gap increases as more labeled data become available, with the decrease growing from 0.009 for $Train_{10\%}$ to 0.045 for $Train_{100\%}$. Nevertheless, despite this reduced performance, the SSL approach demonstrates greater efficiency during the fine-tuning process compared to the training process of SL baseline. Fig. 5 shows the median NSE of the SSL and baseline SL regional models on the $Validation_{regional}$ subset at each epoch. The results show that the SSL achieves higher validation accuracy than the SL benchmark during the early stages of fine-tuning or training, while the SL models perform comparably to or slightly better than the SSL methods as fine-tuning or training progresses (up to 20 epochs). Notably, after only one epoch, the SSL methods yield a substantial improvement in median NSE, ranging from 2.3% to 13.6%, compared to the benchmark. The SSL method improves model training efficiency by initializing model parameters with pre-trained weights on the relevant meteorological forcing data. These pre-trained weights provide a more informative starting point than random initialization, facilitating faster convergence and reducing the risk of overfitting. This advantage of the SSL method is especially essential to obtain a robust model with limited computational resources for model development.

In terms of additional metrics, the SSL approach consistently outperforms the baseline method across all additional metrics, when only limited labeled data is available ($Train_{0.5\%}$ to $Train_{5\%}$, see Table 3). For such cases of very limited data availability, the SSL method significantly decreases the median MSE by up to 0.12 and the median RMSE by up to 0.086, while increasing the median KGE by up to 0.115, compared to the baseline method. The advantages of CPC pre-training are particularly obvious under both low-flow and flood conditions. Specifically, the SSL method achieves superior performance in FLV, peak-

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

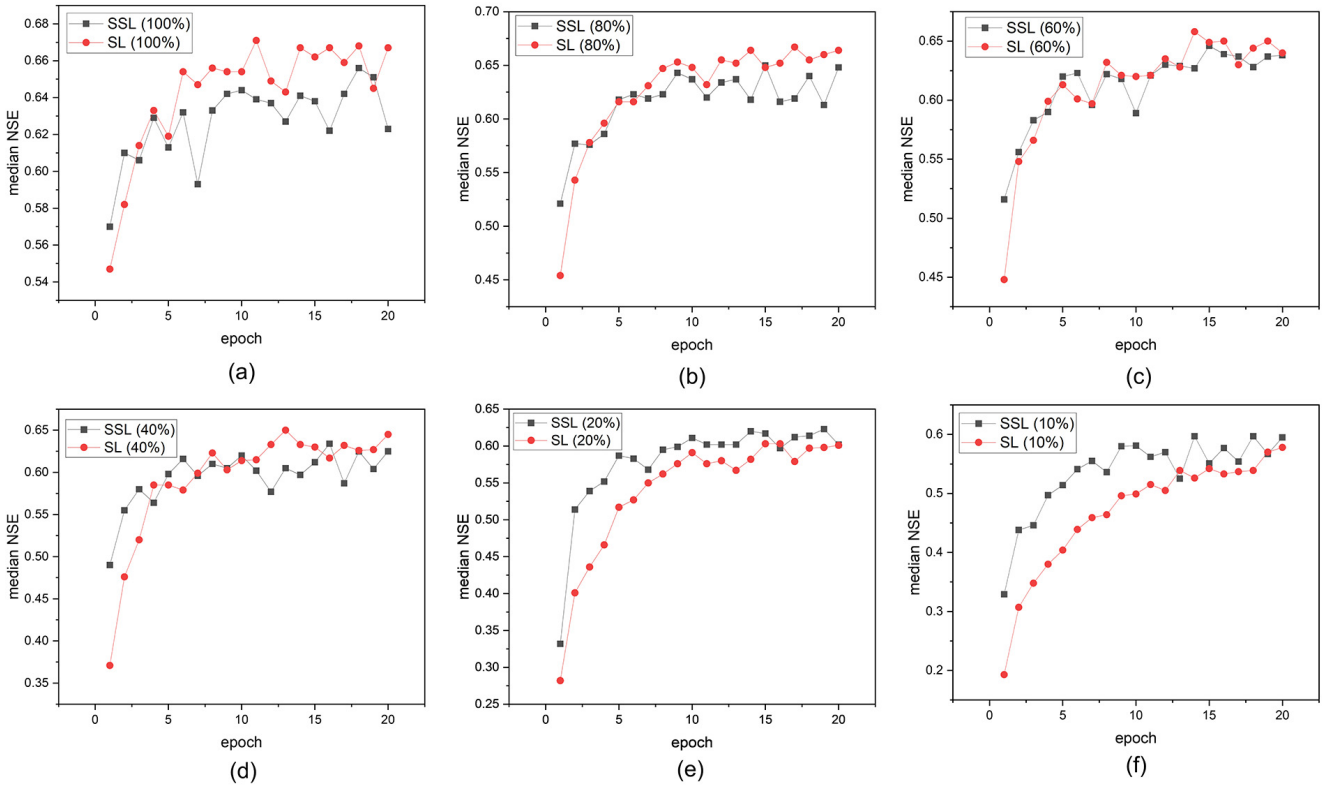

Figure 5. Median Nash–Sutcliffe Efficiency (NSE) achieved by regional models on the Validation$_{regional}$ subset, using the semi-supervised learning (SSL) and baseline supervised learning (SL) method (random seed=110), with different proportion of labeled data available, ranging from (a) 100% to (f) 10%.

timing error, and MAPE$_{peak}$ in more than half of the tested cases, including several scenarios with relatively larger amounts of labeled data available. It decreases the median FLV by up to 16.283, the median peak-timing error by up to 0.34, and the median MAPE$_{peak}$ by up to 13.984, compared against the SL baseline. While the SSL model does not explicitly learn the fundamental relationships between meteorological forcing time series and streamflow in the CPC pre-training stage, it captures data representations of extreme meteorological events by CPC pre-training on a large amount of historical meteorological forcing data (e.g., 28 years). These representations are effective to improve low flow and flood prediction accuracy, when the amount of labeled data is limited (e.g., less than 10% of one-year labeled sequences).

Fig. 6 illustrates streamflow predictions from the SSL and SL methods for a flood event in basin DE212760, (i.e., one of the top five highest streamflow events). The SSL method outperforms the SL baseline in predicting flood magnitude, when less than 40% of labeled data is available. While both methods underestimate the flood peak on 2019-03-16 under these data-scarce scenarios, the SSL predictions are consistently higher than the SL predictions, that are more closely aligned with the observations. Thus, the SSL method achieves lower MAPE$_{peak}$ values, compared to the SL baseline (see Table 3). Additionally,

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

streamflow predictions progressively improve, and become closer to the observations as the amount of labeled data increases from 0.5% to 20%, regardless of learning approach used. When more labeled data become available ($Train_{40\%}$ to $Train_{100\%}$), the SL method surpasses the SSL method. The SL predictions exceed the SSL predictions, and are more closely aligned with the observations. Finally, the SL predictions nearly match the observed streamflow, as the amount of labeled data increases to 100%. More streamflow prediction results for other flood events can be found in Appendix B.

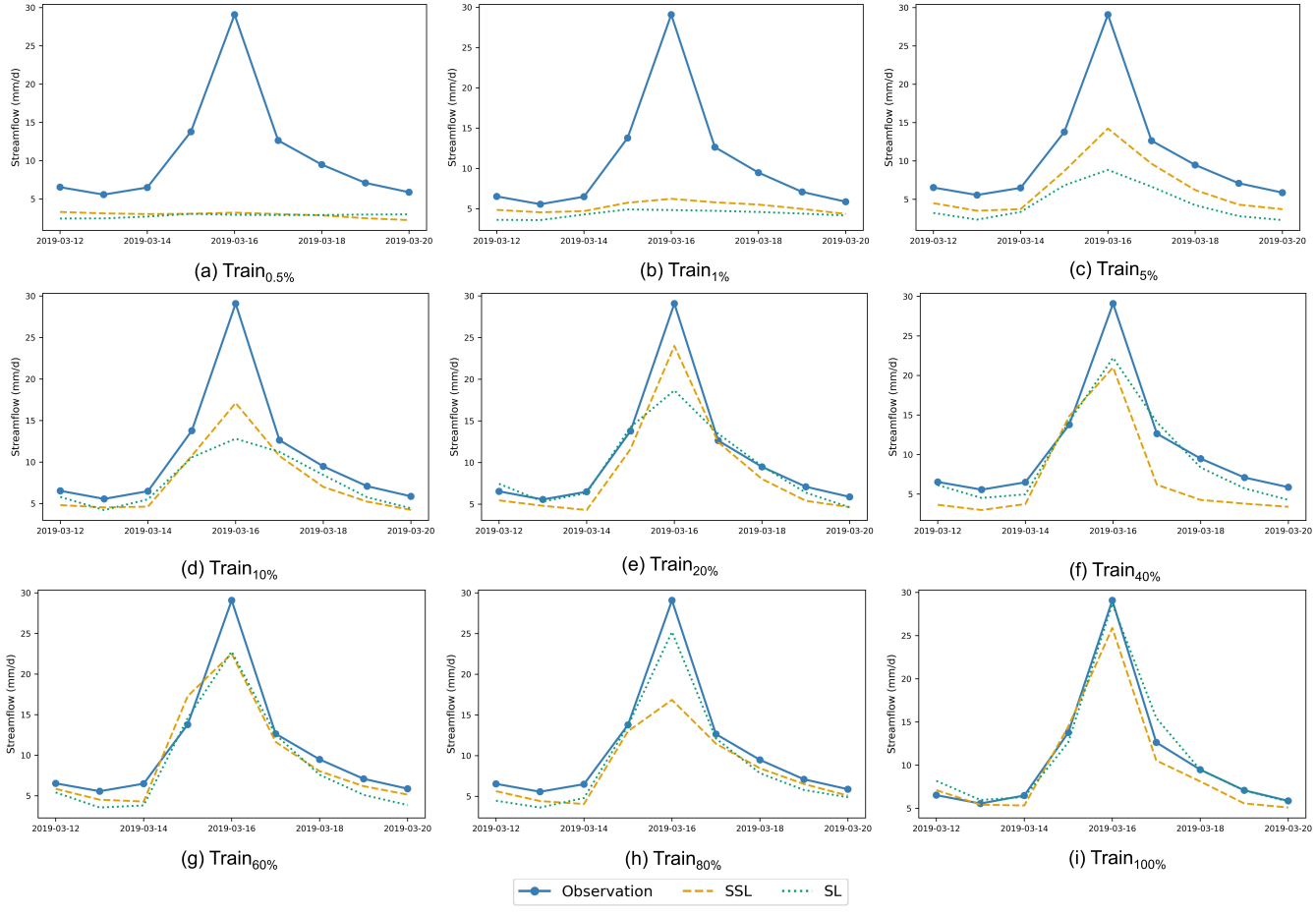

Figure 6. A peak flow event in DE212760 basin, with the observed streamflow values, and the corresponding streamflow predicted by the semi-supervised learning (SSL) approach and baseline supervised learning (SL) method (random seed=110), with varying amount of labeled data available ($Train_{0.5\%}$ to $Train_{100\%}$).

4.2 Experiment 2: Out-of-domain generalization performance of regional models

Fig. 7 demonstrates the out-of-domain generalization performance of regional models in NSE on the $Test_{out}$ subset, using the SSL and SL methods. Additional evaluation metrics are summarized in Table 4. Consistent with the findings from Experiment

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

1, the SSL method achieves better out-of-domain generalization performance than the SL baseline, when limited labeled data is available (i.e., $Train_{0.5\%}$ to $Train_{5\%}$ subsets). This similarity may be attributed to the spatial embedding of the previously unseen 20% of catchments within the remaining 80% (see Fig. 3 (a)). The SSL method achieves a median NSE ranging from 0.318 to 0.526, yielding a significant improvement of up to 0.139 compared to the SL baseline (median NSE=0.179∼0.505), under the data-scarce conditions, where less than 10% of one-year labeled sequences are available. In contrast, when larger amounts of labeled data is available (i.e., $Train_{10\%}$ to $Train_{100\%}$ subsets), the SL baseline surpasses the SSL approach, achieving median NSE values between 0.57 and 0.633. In these data-rich scenarios, the SSL method exhibits a slight performance degradation, with a decrease in median NSE ranging from 0.013 to 0.041, compared to the SL baseline. In this experiment, the median NSE across the 20% of basins is slightly lower than that observed across the 80% of basins in Experiment 1, regardless of deep learning method employed or the amount of labeled data available, while these two group basins exhibit spatial embedding relationships. This difference arises because the 20% of basins are previously unseen by the models, that were exclusively trained on the 80% of basins.

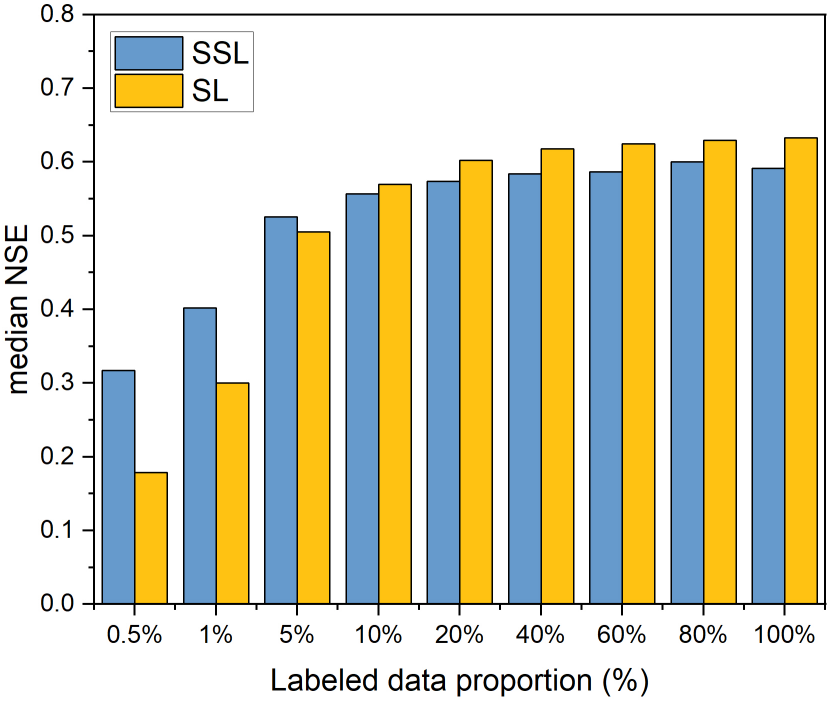

Figure 7. Median Nash-Sutcliffe Efficiency (NSE) achieved by regional models on the $Test_{out}$ subset, using the semi-supervised learning (SSL) and baseline supervised learning (SL) method, with different proportion of labeled data available. The full labeled dataset (100%) includes 411,469 labeled sequences over a one-year period from 1,265 basins in the CAMELS-DE dataset.

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

Table 4. Out-of-domain generalization performance of regional models using the semi-supervised learning (SSL) and supervised learning (SL) method across multiple metrics, fine-tuned on different amounts of labeled data. Median metric values are reported across basins in the $Test_{out}$ subset.

Training dataset	MSE		RMSE		KGE		FHV		FLV		Peak-timing error		$MAPE_{peak}$	
	SSL	SL	SSL	SL	SSL	SL	SSL	SL	SSL	SL	SSL	SL	SSL	SL
$Train_{0.5\%}$	**0.456**	0.586	**0.675**	0.766	**0.356**	0.235	**-45.308**	-60.828	**23.942**	39.816	**0.702**	0.952	**55.529**	68.363
$Train_{1\%}$	**0.391**	0.455	**0.625**	0.675	**0.459**	0.335	**-33.012**	-51.307	**13.648**	34.695	**0.578**	0.889	**47.887**	61.879
$Train_{5\%}$	**0.318**	0.328	**0.564**	0.573	**0.548**	0.510	**-17.079**	-29.996	**37.762**	42.403	**0.432**	0.506	**43.793**	47.445
$Train_{10\%}$	0.318	**0.297**	0.564	**0.545**	0.571	**0.573**	**-15.797**	-21.423	**33.443**	35.866	**0.423**	0.424	43.080	**42.768**
$Train_{20\%}$	0.293	**0.276**	0.542	**0.525**	0.572	**0.615**	**-15.995**	-16.224	37.561	**35.295**	0.424	**0.414**	**40.962**	41.583
$Train_{40\%}$	0.287	**0.263**	0.536	**0.513**	0.594	**0.622**	-17.158	**-15.291**	**33.855**	41.693	0.422	**0.406**	41.148	**40.269**
$Train_{60\%}$	0.296	**0.261**	0.544	**0.511**	0.600	**0.627**	-15.972	**-15.378**	**30.437**	36.167	0.405	**0.398**	41.056	**40.477**
$Train_{80\%}$	0.276	**0.259**	0.526	**0.509**	0.613	**0.639**	-16.132	**-13.065**	36.095	**35.774**	**0.400**	0.401	**39.770**	39.910
$Train_{100\%}$	0.278	**0.255**	0.527	**0.505**	0.609	**0.641**	**-14.320**	-14.535	34.610	**30.243**	**0.385**	0.400	**39.766**	39.992

Consistent with the findings from Experiment 1, the SSL approach consistently outperforms the baseline method across all additional metrics, when only limited labeled data is available ($Train_{0.5\%}$ to $Train_{5\%}$). For such cases of very limited data availability, the SSL method significantly reduces the median MSE by up to 0.131 and the median RMSE by up to 0.091, while improving the median KGE by up to 0.124, compared to the SL baseline. For low flow and flood prediction in unseen basins, the SSL method exhibits superior out-of-domain generalization performance across FLV, FHV, peak-timing error, and $MAPE_{peak}$ in more than half of the tested scenarios, including cases with relatively larger amounts of labeled data available. For example, the SSL method reduces the median FLV by up to 21.047, the median peak-timing error by up to 0.311, and the median $MAPE_{peak}$ by up to 13.933, compared with the SL baseline.

4.3 Experiment 3: Performance of single-basin models

Fig. 8 shows the NSE values obtained by each single-basin model using the SSL and baseline SL method. Additional evaluation metrics are summarized in Table 5. The SSL method performs best in most cases, obtaining an NSE ranging from to -0.012 to 0.503. That yields a significant improvement in mean NSE of 0.172 (with a maximum of 0.39), compared to the baseline SL method (NSE = -0.138∼0.170). For basin DE710320, the SSL method yields a slight decrease in NSE by 0.036, compared to the SL baseline. Across all basins, the baseline SL method achieves NSE values of below 0.2, whereas the SSL method obtains NSE values of above 0.2 for at least four basins. Moreover, the baseline SL method performs extremely poorly for four basins, with an NSE of zero or below, including DEE10170 (0.008), DE710440 (-0.015), DEE10560 (-0.138), and DED10970

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

(-0.006). In contrast, the SSL method achieves NSE values above 0.1 for all these four basins, representing a substantial improvement in model performance.

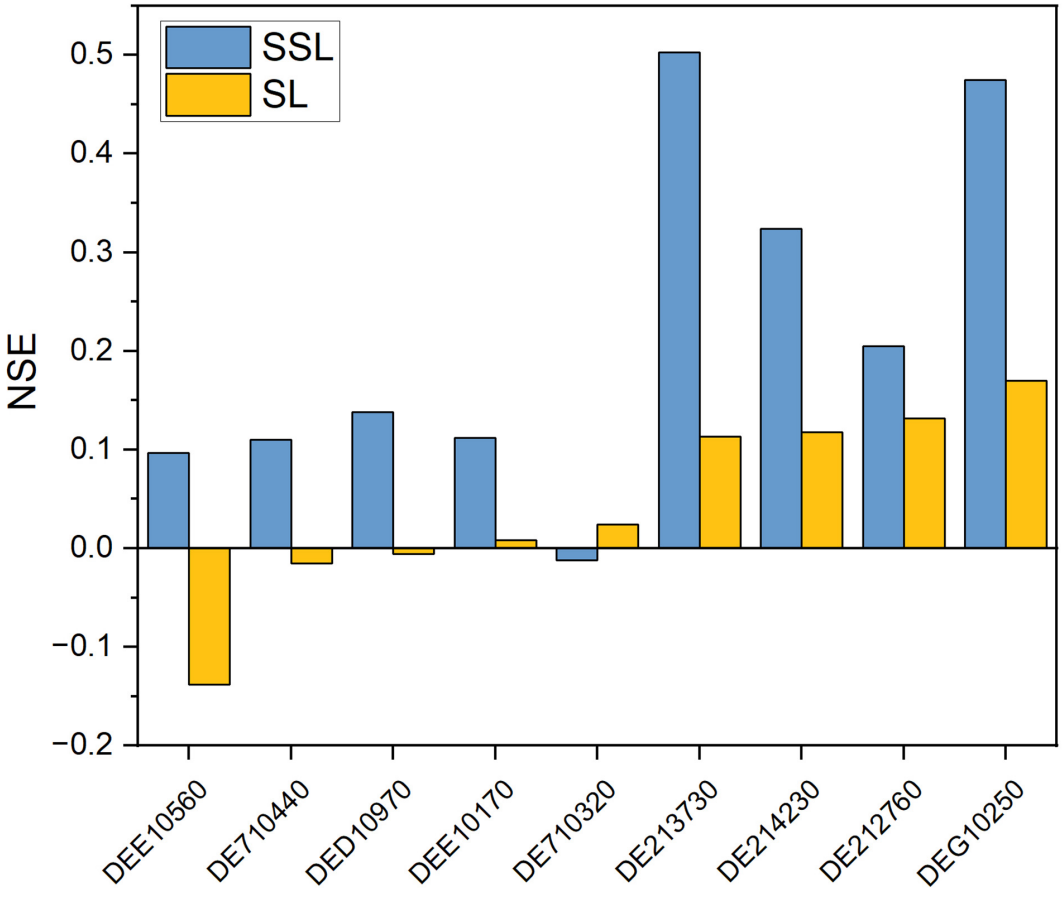

Figure 8. Nash–Sutcliffe Efficiency (NSE) values for each single-basin model using the semi-supervised learning (SSL) and baseline supervised learning (SL) approach. These models were fine-tuned (SSL) or trained (SL) on one-year labeled sequences from each basin. Basins are ordered from left to right according to the NSE values achieved by the SL baseline models.

The performances across these basins achieved by single-basin models are substantially lower than those of the regional models developed in Experiment 1, regardless of whether SSL methods (NSE=0.363∼0.74) or SL method (NSE=0.248∼0.799) were used (see Appendix B). This is not surprising, as regional models were trained on significantly more labeled data (28 years of data from 1,265 basins), than single-basin models (only one year of data from each of nine selected basins). This comparison highlights the challenge of developing robust single-basin models under data-scarce conditions, that cannot leverage shared knowledge from a large number of basins (Kratzert et al., 2024). Under such conditions, we recommend developing and

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

Table 5. Performance comparison between the semi-supervised learning (SSL) and supervised learning (SL) method across multiple metrics for each single-basin model.

Basin	MSE		RMSE		KGE		FHV		FLV		Peak-timing error		$MAPE_{peak}$	
	SSL	SL	SSL	SL	SSL	SL	SSL	SL	SSL	SL	SSL	SL	SSL	SL
DE214230	**0.139**	0.182	**0.373**	0.427	**0.461**	0.199	**-41.370**	-61.338	**5.970**	67.202	**0.689**	1.822	**40.641**	54.776
DE213730	**0.187**	0.334	**0.433**	0.578	**0.569**	0.372	**-40.543**	-57.509	**-27.947**	75.234	**0.892**	2.270	**44.016**	56.204
DEE10170	**0.071**	0.079	**0.266**	0.281	0.433	**0.436**	**-36.084**	-43.982	15.353	**8.794**	**1.552**	2.310	**39.119**	45.345
DE710320	0.795	**0.767**	0.892	**0.876**	**0.194**	0.017	**-42.586**	-65.090	-701.921	**71.958**	**1.596**	1.872	**58.175**	60.836
DE710440	**0.218**	0.249	**0.467**	0.499	**0.509**	0.452	-29.359	**-28.338**	-325.385	**-34.034**	**0.917**	1.667	**51.201**	59.103
DEE10560	**0.148**	0.186	**0.385**	0.432	**0.496**	0.449	**-41.104**	-42.754	-1673.791	**-1498.062**	**0.906**	1.938	**55.334**	56.896
DE212760	**3.099**	3.387	**1.760**	1.840	**0.501**	0.254	**-46.854**	-61.152	**-1635.611**	-1704.724	**0.630**	1.609	**66.089**	80.054
DEG10250	**1.438**	2.273	**1.199**	1.508	**0.509**	0.384	**-48.446**	-63.341	-771.178	**-629.363**	**0.946**	1.730	**54.200**	67.869
DED10970	**0.436**	0.509	**0.661**	0.714	**0.109**	-0.085	**-69.151**	-74.451	**22.399**	76.106	**1.053**	1.684	**64.092**	76.197

applying regional models rather than single-basin models for accurate streamflow prediction, provided that unlabeled and/or labeled data from a large number of basins is available.

Consistent with the comparison based on the NSE metric, the SSL method outperforms the SL baseline in most cases across the additional metrics, with the exception of FLV (see Table 5). In terms of overall performance, the SSL method significantly reduces MSE by up to 0.835, and RMSE by up to 0.309, compared to the SL baseline. Moreover, it yields a significant improvement in KGE of up to 0.262. For flood-related metrics, the SSL method significantly decreases the peak-timing error by up to 1.378, and $MAPE_{peak}$ by up to 14.135, compared to the SL baseline. Consistent with the findings from Experiment 1 and 2, the SSL method again demonstrates an enhanced capability to capture flood dynamics in this experiment, particularly with respect to peak timing and peak streamflow magnitude. For low flow prediction, the SSL method does not show significantly obvious performance improvement across the nine tested basins. While the SSL method achieves FLV values closer to zero than the SL baseline in four basins, it performs worse in the remaining five basins, suggesting limited and basin-dependent gains for low flow predictions.

Fig. 9 shows four flood events across different basins (i.e., one of the top five highest streamflow events for each basin), with additional examples provided in Appendix B. These results further demonstrate that the SSL method significantly improves both the accuracy and timeliness of flood prediction compared to the SL baseline. In basin DE212760 (Fig. 9 (a)), the observed streamflow exhibits a sharp flood peak, increasing from 3.95 mm/d on 2002-08-11 to 25.49 mm/d on 2002-08-12. The SL method substantially underestimates the peak (0.35 vs 25.49 mm/d), and incorrectly predicts this rapid rise in streamflow as a slight drop from 0.49 to 0.35 mm/d. Additionally, it fails to capture the subsequent recession phase of the hydrograph, resulting in a relatively flat hydrograph (0.46-0.47 mm/d). In contrast, the SSL method captures a noticeable increase in streamflow from 1.72 to 2.20 mm/d during the peak period, better reflecting the temporal dynamics of the flood. While the SSL method also

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

underestimates the peak magnitude (2.20 vs 25.49 mm/d), due to the limitations discussed in Section 4.1, it more accurately captures both the shape and timing of the hydrograph, compared to the SL baseline. Furthermore, it accurately captures the subsequent recession phase of the hydrograph, with streamflow decreasing from 2.20 to 0.70 mm/d. A similar pattern can be also observed for flood events in other basins (see Fig. 9 (b-d)). The observed streamflow rapidly rises to a peak within one or two days. However, the SL model generates nearly constant streamflow predictions or even a slight decline, failing to predict this sharp increase in streamflow. In contrast, the SSL method captures the increasing trend, the flow peak, and the recession of the event at the correct time. While both methods significantly underestimate streamflow, the SSL predictions are more closely aligned with the observations, resulting in lower $MAPE_{peak}$ values than those of the SL baseline (see Table 5).

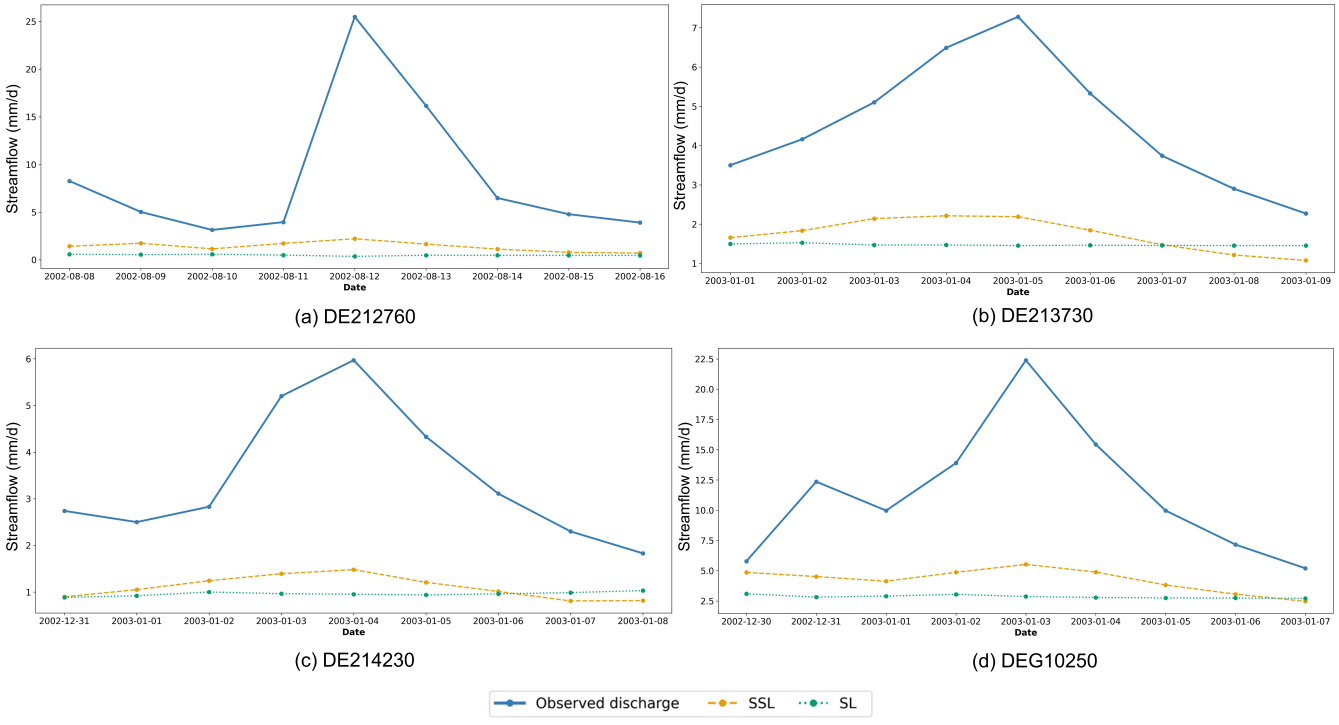

Figure 9. Examples of peak flow events across different basins, with the observed daily streamflow, and the corresponding streamflow predicted by single-basin models using the semi-supervised learning (SSL) approach and baseline supervised learning (SL) method.

4.4 Limitations and future works

In this study, we did not use the maximum amount of unlabeled data available from basins over a 28-year period in CAMELS-DE dataset. Additionally, we did not perform hyperparameter fine-tuning for CPC pre-training stage (e.g., the number of pre-training epochs), due to limited computational resources (1 hour per pre-training epoch). Despite these constraints, the proposed approach achieves competitive performance, demonstrating the potential effectiveness of the methodology for streamflow pre-

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

diction in data-scarce regions. Further performance improvement is expected through scaling both the pre-training dataset, and the computational effort devoted to hyperparameter optimization. Previous studies in other domains have reported noticeable enhancement in SSL model performance with larger self-supervised pre-training datasets and longer pre-training time, such as when increasing unlabeled data from 1.2 million to 1 billion images (Goyal et al., 2022), and extending pre-training from 50 to 200 epochs (Yildizli et al., 2026).

4.5 Implications for streamflow prediction

Beyond the quantitative performance improvements demonstrated in this study, the proposed SSL methods have broader methodological implications for streamflow prediction, that are summarized in three key aspects:

1. The SSL methods show the potential to advance the development of daily streamflow prediction models for countries and regions with limited (high-quality) hydrological data, e.g., China. Lin et al. (2023) reported that hydrological data in China are not widely shared and are difficult to obtain. While some data are digitally available through national data platforms, their quality is often poor, with issues such as temporal discontinuities and sparse spatial coverage. Consequently, developing robust regional models for China remains challenging. Therefore, the SSL methods hold substantial potential for application in such data-constrained settings. In contrast, our results indicate that Germany cannot be considered a "data-scarce" country for the proposed SSL methods under the configuration tested in this study (see Section 4.1). This finding may also extend to other countries or regions with a sufficient volume of hydro-meteorological forcing time series data, e.g., more than 400,000 training sequences in the $\text{Train}_{100\%}$ subset (see Table 1).

2. The SSL methods may further enhance the development of hourly streamflow prediction or flood forecasting models. Under rapidly changing climate conditions, high-resolution early warning systems are essential for reducing flood-related deaths and economic losses. While the large-sample hydrological datasets at daily resolution have recently become available for several countries (e.g., CAMELS-DE for Germany, and CAMELS-DK for Denmark (Liu et al., 2025)), open-access datasets containing hourly streamflow observations remain limited in most regions. This limitation hinders the development of data-driven models for hourly flood forecasting. We believe that the SSL methods could address this issue, by reducing the reliance on large amounts of hourly streamflow observations for model training.

3. The SSL methods offer new perspectives for applying transfer learning approaches in streamflow prediction. To date, a systematic evaluation of different transfer learning strategies for streamflow prediction is still lacking. When developing models for target basins with limited streamflow observations, existing studies often transfer knowledge from irrelevant basins, with different meteorological, hydrological and geographical conditions (Ouyang et al., 2025; Khoshkalam et al., 2023). In contrast, the SSL methods enable the direct transfer of knowledge from the data-scarce target basins themselves. Such a strategy may enhance the prediction performance for target basins by leveraging more relevant data, and reducing domain mismatch.

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

4.6 Towards foundation models in hydrology

Foundation models have revolutionized artificial intelligence (AI) across various domains, including natural language processing (e.g., GPT (Achiam et al., 2023)), computer vision (e.g., Segment Anything Model (Kirillov et al., 2023)) and multi-modal understanding (e.g., Deepseek-vl2 (Wu et al., 2024)). Compared to traditional SL models, that rely on predefined physical structures or specific data annotations, foundation models represent a paradigm shift in AI, enabling a deeper understanding of underlying processes. They learn broad and adaptable data representations from large-scale unlabeled datasets using self-supervised learning methods and large model architectures, such as Transformers (Vaswani et al., 2017). These rich representations enable effective adaptation to a wide range of downstream tasks through task-specific fine-tuning, and improve models' out-of-domain generalization capability. The OpenAI GPT series, which underpins the development of ChatGPT, exemplifies this paradigm shift in natural language processing (Achiam et al., 2023). A similar example of a foundation model is AlphaEarth, developed by Google DeepMind (Brown et al., 2025), which has already shown promising results in earth observation.

This paradigm shift opens new possibilities for developing robust hydrological foundation models for multiple hydrological tasks, such as simulating climate-driven variables, including water temperature, groundwater level, and soil moisture. These models are typically pre-trained on large-scale meteorological forcing datasets using self-supervised learning methods (e.g., CPC), and can be adapted to various hydrological tasks through task-specific fine-tuning. Traditional data-driven hydrological modeling typically follows a single-task paradigm, where a separate model is developed for each target. However, the foundation model can serve as a unified "all-in-one" base model for multiple climate-driven variables. By leveraging shared representations across tasks, this model can provide a generalizable, scalable, and computationally efficient solution, that could fundamentally shift the current paradigm of hydrological prediction, and benefit academic researchers and operational practitioners in the fields of hydrology, water resources management, earth observation, and environmental science. Recently, Wang et al. (2025) have applied self-supervised learning methods for soil moisture prediction and reported promising results. This success further supports the feasibility of developing foundation models for hydrology.

5 Conclusions

Accurate streamflow prediction is critical for flood forecasting and water resource management, particularly under intensifying climate extremes and increasing flood hazards worldwide. Recently, deep learning models have demonstrated great performance in streamflow prediction, when trained on large-scale datasets. However, achieving robust models requires large amounts of "labeled" training data for supervised learning (SL), including meteorological forcing input sequences paired with corresponding streamflow observations. This limits model performance for data-scarce basins. To overcome this challenge, we propose a two-stage semi-supervised learning (SSL) method for streamflow prediction, based on the Contrastive Predictive Coding (CPC) method. CPC is a contrastive self-supervised learning approach, that learns data representations from "unlabeled" data (i.e., meteorological forcing input sequences without streamflow observation), by contrasting correct future predictions against incorrect ones. We conducted three experiments using the CAMEL-DE dataset to address the main research

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

question: can SSL methods improve streamflow prediction in data-scarce regions, compared to conventional SL methods, with respect to (1) regional and single-basin models, (2) in-domain and out-of-domain performance, and (3) low flow and flood prediction?

The main findings of these experiments are as follows:

1. The SSL method improves performance of regional models compared to the baseline SL method for data-scarce conditions, when less than 10% of one-year labeled sequences are available for training. It yields a significant improvement in the median Nash–Sutcliffe Efficiency (NSE) across basins of 0.137 (with 0.5% of one-year labeled data), when evaluated on previously seen (in-domain) basins from the training dataset. More importantly, SSL also demonstrates similar improvements in NSE (of 0.139) with 0.5% of labeled data, when tested on previously unseen (out-of-domain) basins from the training dataset. This is mainly due to the extraction of informative data representations through CPC pre-training on long-term meteorological forcing data. However, the SL baseline slightly outperforms the SSL method under data-rich conditions, when $\geq$ 10% of one-year labeled sequences are available.

2. For low flow and flood prediction in both seen and unseen basins, the SSL regional models outperform the SL baseline under data-scarce conditions, when less than 10% of one-year labeled sequences are available. When tested on unseen basins, the SSL method reduces the median percent bias of the bottom 30% low flow range (FLV) by 21.047, the median peak-timing error by 0.331. and the median mean Absolute Percentage Error of peaks ($MAPE_{peak}$) by 13.933, compared to the SL baseline with 1% of one-year labeled data. Owing to the data representations of extreme meteorological events learned through CPC pre-training on large volumes of historical meteorological forcing data, the SSL method more effectively captures low flow and flood dynamics.

3. The SSL method is more effective for single-basin models than regional models under data-scarce conditions. The SSL single-basin models achieve significant improvements in mean NSE of 0.172 (with a maximum of 0.39) across nine tested basins, compared to the SL single-basin models, when one-year labeled sequences are available for each basin. Additionally, the SSL method enhances the ability of single-basin models to capture both low flow conditions and flood dynamics, particularly with respect of peak timing and peak streamflow magnitude.

Based on the above findings, we conclude that SSL methods improve streamflow prediction in data-scarce regions, compared to conventional SL methods across all considered aspects, including (1) both regional and single-basin models, (2) both in-domain and out-of-domain performance, and (3) both low flow and flood prediction.

The aim of this study was not to develop a model intended for operational deployment in streamflow prediction. Rather, we aimed to provide clear evidence that data representations learned through SSL can enhance models' generalization capability, when only limited streamflow observations are available in data-scarce basins or regions. Furthermore, this study outlines a potential pathway toward the development of hydrological foundation models based on the proposed SSL approaches. Foundation models have revolutionized machine learning applications in numerous fields. By scaling our method with significantly larger and more diverse datasets, these models could be extended to perform multi-variable prediction tasks, e.g., simulation of water temperature, groundwater level, and soil moisture.

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

Code availability. The code for this study is available on [*Zenodo repository made available after acceptance*].

Data availability. All the data generated for this study is available on [*Zenodo repository made available after acceptance*]. The CAMELS-DE dataset is available on https://doi.org/10.5281/zenodo.13837553 (Loritz et al., 2024).

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

Appendix A: Experimental Setup

A1 Model settings

Table A1. Meteorological variables, target variables and basin-specific static attributes used in model development, and their normalization statistics computed on different subsets.

	Variable or attribute name (following Loritz et al. (2024))	Description (following Loritz et al. (2024))	Unit	Pre-train$_{regional}$ subset		Train$_{100\%}$ subset	
				mean	standard deviation	mean	standard deviation
Meteorological variable	precipitation_mean	Observed interpolated spatial mean of the daily precipitation	mm d^{-1}	2.3	4.7	2.6	4.9
	precipitation_stdev	Observed interpolated spatial standard deviation of the daily precipitation	mm d^{-1}	0.5	1.0	0.6	1.0
	radiation_global_mean	Observed interpolated spatial mean of the global radiation	W m^{2}	114.3	85.2	114.2	85.3
	temperature_min	Observed interpolated spatial mean daily minimum temperatures	°C	4.3	6.6	5.2	6.3
	temperature_max	Observed interpolated spatial mean daily maximum temperatures	°C	12.2	8.5	13.1	8.3
Target variable	discharge_spec	Observed catchment-specific discharge	mm d^{-1}			1.1	1.8
Catchment-specific static attributes	area	catchment area	km^{2}	342.5	778.1	342.5	778.1
	elev_mean	mean elevation in the catchment	ma.s.l.	322.4	237.2	322.4	237.2
	clay_0_30cm_mean	weight percent of clay particles (<0:002 mm) in the fine earth fraction at depths of 0–30 cm	wt%	21.0	7.4	21.0	7.4
	sand_0_30cm_mean	weight percent of silt particles ($\geq$ 0:002 mm and $\leq$ 0:05/0.063 mm) in the fine earth fraction at depths of 0–30 cm	wt%	36.1	18.8	36.1	18.8
	silt_0_30cm_mean	weight percent of silt particles ($\geq$ 0:002 mm and $\leq$ 0:05/0.063 mm) in the fine earth fraction at depths of 0–30 cm	wt%	37.7	11.7	37.7	11.7
	artificial_surfaces_perc s	areal coverage of artificial surfaces	%	6.8	6.5	6.8	6.5
	agricultural_areas_perc	areal coverage of agricultural areas	%	56.0	21.9	56.0	21.9
	forests_and_seminatural_areas_perc	areal coverage of forests and semi-natural areas	%	36.6	22.3	36.6	22.3
	wetlands_perc	areal coverage of wetlands	%	0.1	0.6	0.1	0.6
	water_bodies_perc	areal coverage of water bodies	%	0.5	1.7	0.5	1.7
	p_mean	long-term mean of daily precipitation from 1951 to 2020	mm d^{-1}	2.3	0.6	2.3	0.6
	p_seasonality	seasonality and timing of precipitation		0.1	0.1	0.1	0.1
	frac_snow	fraction of precipitation falling as snow, i.e. while mean air temperature is <0 °C		0.1	0.0	0.1	0.0
	high_prec_freq	frequency of high-precipitation days ($\geq$ 5 times mean daily precipitation)	d yr^{-1}	16.8	1.3	16.8	1.3
	low_prec_freq	frequency of dry days (<1 mm d^{-1})	d yr^{-1}	227.0	16.2	227.0	16.2
	high_prec_dur	mean duration of high precipitation events (number of consecutive days $\geq$ 5 times mean daily precipitation)	d	1.2	0.0	1.2	0.0
	low_prec_dur	mean duration of dry periods (number of consecutive days <1 mm d^{-1} mean daily precipitation)	d	4.0	0.1	4.0	0.1

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

A2 Evaluation metrics

A2.1 Overall performance

Table A2 shows the evaluation metrics used in this study, including their value ranges and optimal values. We evaluated overall performances using four metrics: (1) Nash–Sutcliffe Efficiency (NSE), (2) Mean Square Error (MSE), (3) Root Mean Square Error (RMSE), and (4) Kling-Gupta Efficiency (KGE), that are defined as follows:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N} \left(Q_{\mathrm{obs},i} - Q_{\mathrm{pred},i}\right)^2}{\sum_{i=1}^{N} \left(Q_{\mathrm{obs},i} - \overline{Q}_{\mathrm{obs}}\right)^2} \tag{A1}$$

$$\mathrm{MSE} = \frac{1}{N} \sum_{i=1}^{N} (Q_{\mathrm{obs},i} - Q_{\mathrm{pred},i})^2 \tag{A2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (Q_{\mathrm{obs},i} - Q_{\mathrm{pred},i})^2} \tag{A3}$$

where N is the number of samples (days) per basin b. $Q_{obs,i}$ and $Q_{pred,i}$ denote the observed streamflow and predicted streamflow at time step i, respectively. $\overline{Q}_{\mathrm{obs}}$ is the mean observed streamflow over the N samples.

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{A4}$$

$$\alpha = \frac{\sigma(Q_{\mathrm{pred}})}{\sigma(Q_{\mathrm{obs}})} \tag{A5}$$

$$\beta = \frac{\overline{Q}_{\mathrm{pred}}}{\overline{Q}_{\mathrm{obs}}} \tag{A6}$$

where r is the pearson correlation coefficient between the observed and predicted streamflow. α and β are defined as the ratios of predicted to observed standard deviation (variability) and mean (bias), respectively.

A2.2 Performance under extreme flow conditions

To assess model performance in predicting extreme flows, we used four additional metrics: (1) the percent bias of the top 2% peak flow range (FHV), (2) the percent bias of the bottom 30% low flow range (FLV), (3) peak-timing error, and (4) Mean Absolute Percentage Error of peaks ($\mathrm{MAPE}_{\mathrm{peak}}$), that are defined as follows:

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

Table A2. Summary of performance metrics used for model evaluation, including value ranges and optimal values.

Metric	Full name	Value range	Optimal value
NSE	Nash–Sutcliffe Efficiency	$(-\infty, 1]$	1
MSE	Mean Squared Error	$[0, +\infty)$	0
RMSE	Root Mean Squared Error	$[0, +\infty)$	0
KGE	Kling–Gupta Efficiency	$(-\infty, 1]$	1
FHV	The percent bias of the top 2% peak flow range	$(-\infty, +\infty)$	0
FLV	The percent bias of the bottom 30% low flow range	$(-\infty, +\infty)$	0
Peak-timing error		$[0, +\infty)$	0
$\text{MAPE}_{\text{peak}}$	Mean Absolute Percentage Error of peaks	$[0, +\infty)$	0

$$\text{FHV} = \frac{\sum_{h=1}^{H}(Q_{\text{pred},h} - Q_{\text{obs},h})}{\sum_{h=1}^{H} Q_{\text{obs},h}} \times 100 \tag{A7}$$

$$\text{FLV} = -1 \times \frac{\sum_{l=1}^{L}[\log(Q_{\text{pred},l}) - \log(Q_{\text{pred},L})] - \sum_{l=1}^{L}[\log(Q_{\text{obs},l}) - \log(Q_{\text{obs},L})]}{\sum_{l=1}^{L}[\log(Q_{\text{obs},l}) - \log(Q_{\text{obs},L})]} \times 100 \tag{A8}$$

where $h = 1, 2, \ldots, H$ denotes the index of the top 2% peak flows. $l = 1, 2, \ldots, L$ is the index of the bottom 30% low flows. L represents the index of the minimum flow.

Peak-timing error metric measures the lag between observed and predicted peaks. First, the most significant peaks in the observed time series are heuristically identified. Second, peaks with topographic prominence smaller than the standard deviation of the series are discarded. Third, the smallest remaining peaks are removed until all retained peaks are separated by ≥ 100 time steps. For each retained observed peak, the corresponding predicted peak is defined as the maximum streamflow within a three-day temporal window. Finally, we calculate peak-timing error as the mean absolute difference between the times of the observed and predicted peaks.

Mean Absolute Percentage Error of peaks ($\text{MAPE}_{\text{peak}}$) quantifies the average absolute percentage difference in streamflow volume between observed and predicted peaks. The observed and predicted peaks are identified following the same procedure described for the peak-timing error metric. $\text{MAPE}_{\text{peak}}$ is defined as follows:

$$\text{MAPE}_{\text{peak}} = \frac{1}{P}\sum_{p=1}^{P}\left|\frac{Q_{\text{pred},p} - Q_{\text{obs},p}}{Q_{\text{obs},p}}\right| \times 100 \tag{A9}$$

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

where P is the number of peaks.

Appendix B: Experiment results

B1 Validation loss on the Val_{CPC} dataset

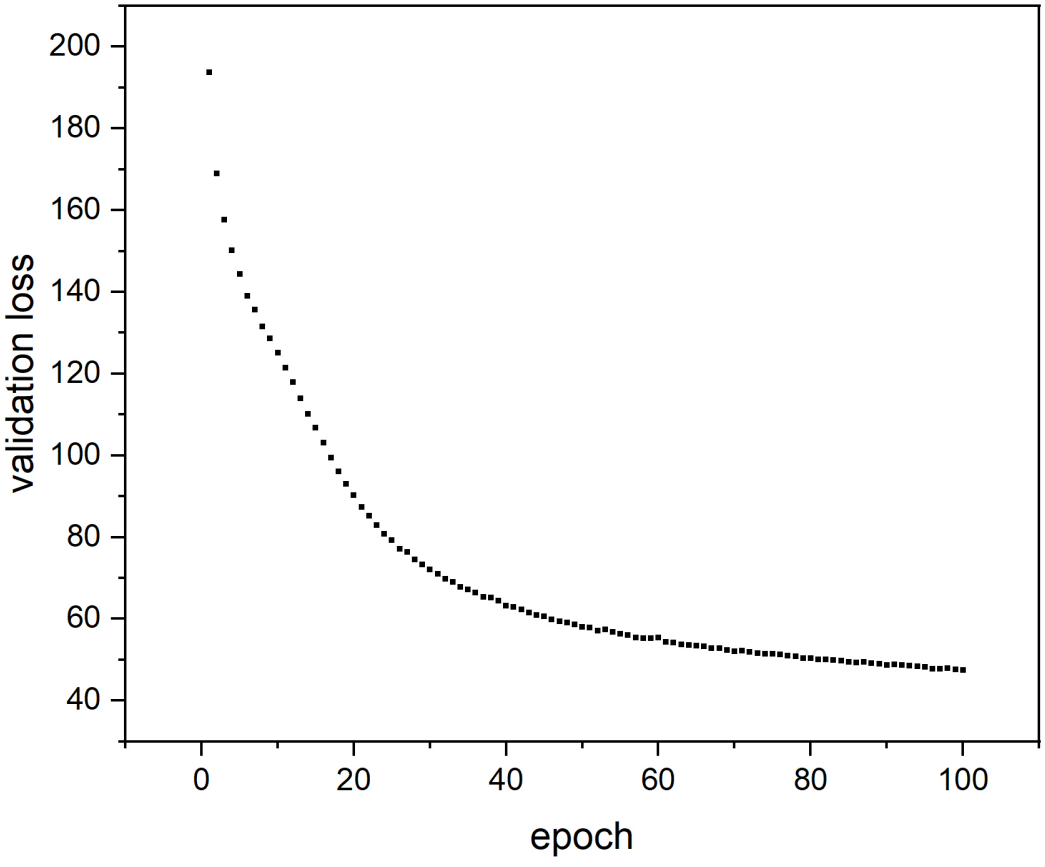

Figure B1. The infoNCE loss on the Val_{CPC} dataset per CPC pre-training epoch.

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

B2 Cumulative NSE distribution of different methods in Experiment 1

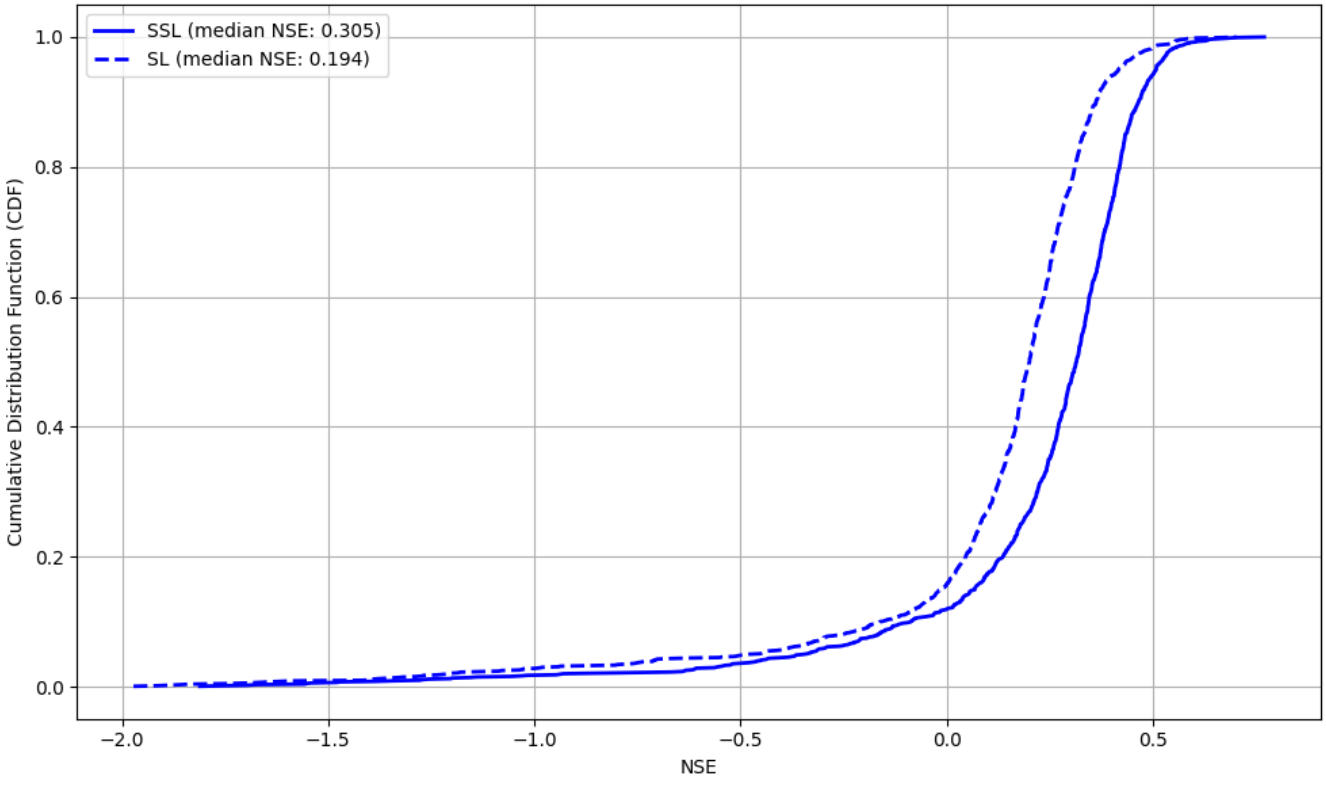

Figure B2. Cumulative NSE distribution of the semi-supervised learning (SSL) and baseline supervised learning (SL) models (random seed=110) evaluated on the Test_{in} subset. Both models were fine-tuned or trained on the $\text{Train}_{0.5\%}$ subset. We only showed NSE values between -2 and 1 in this figure.

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

B3 Streamflow prediction results of SSL and SL regional models for a flood event in Experiment 1

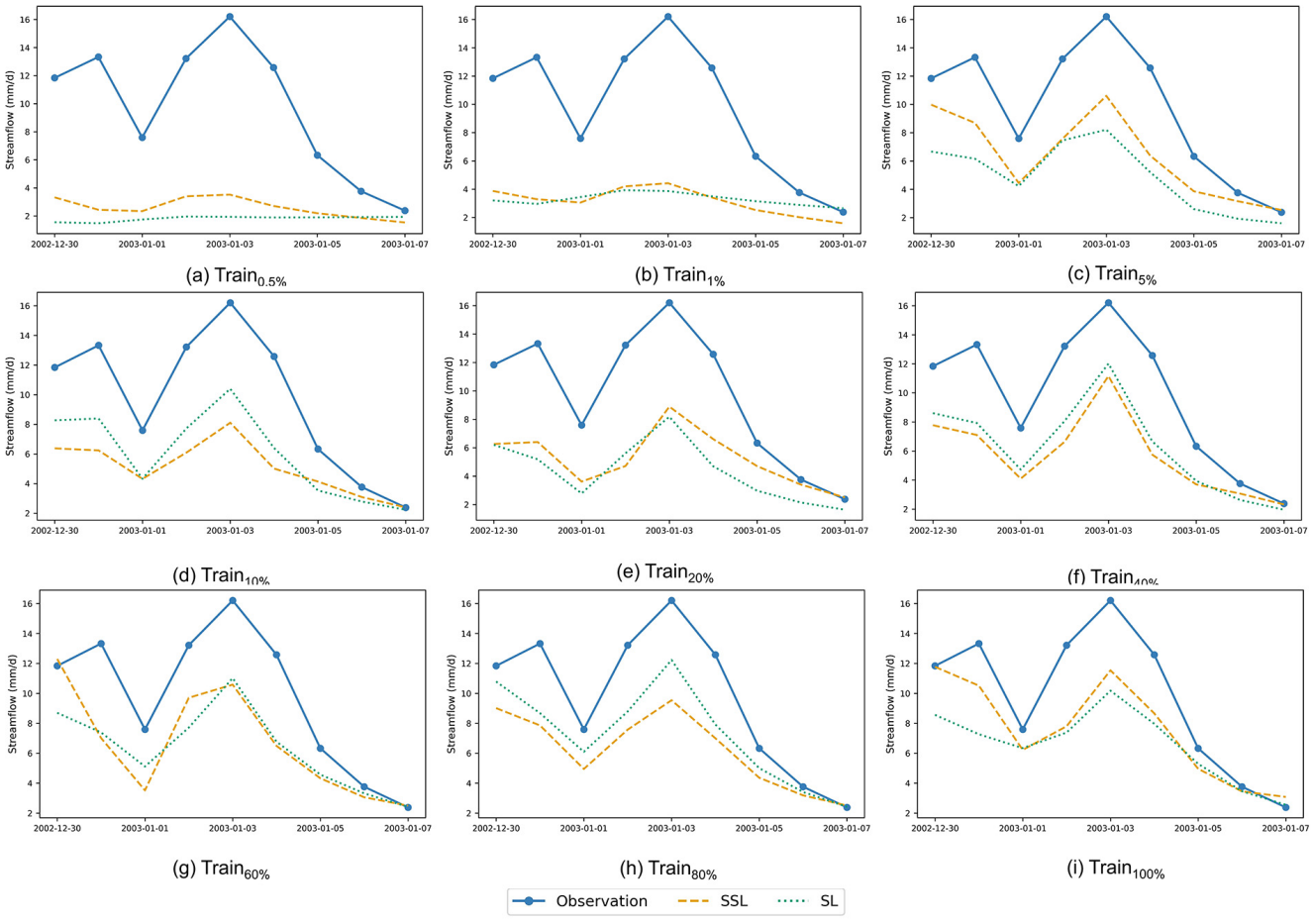

Figure B3. A peak flow event in DE710280 basin, with the observed streamflow values, and the corresponding streamflow predicted by the semi-supervised learning (SSL) approach and baseline supervised learning (SL) method (random seed=110), with varying amount of labeled data available ($Train_{0.5\%}$ to $Train_{100\%}$).

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

B4 Regional and single-basin model performance on nine selected basins in Experiment 3

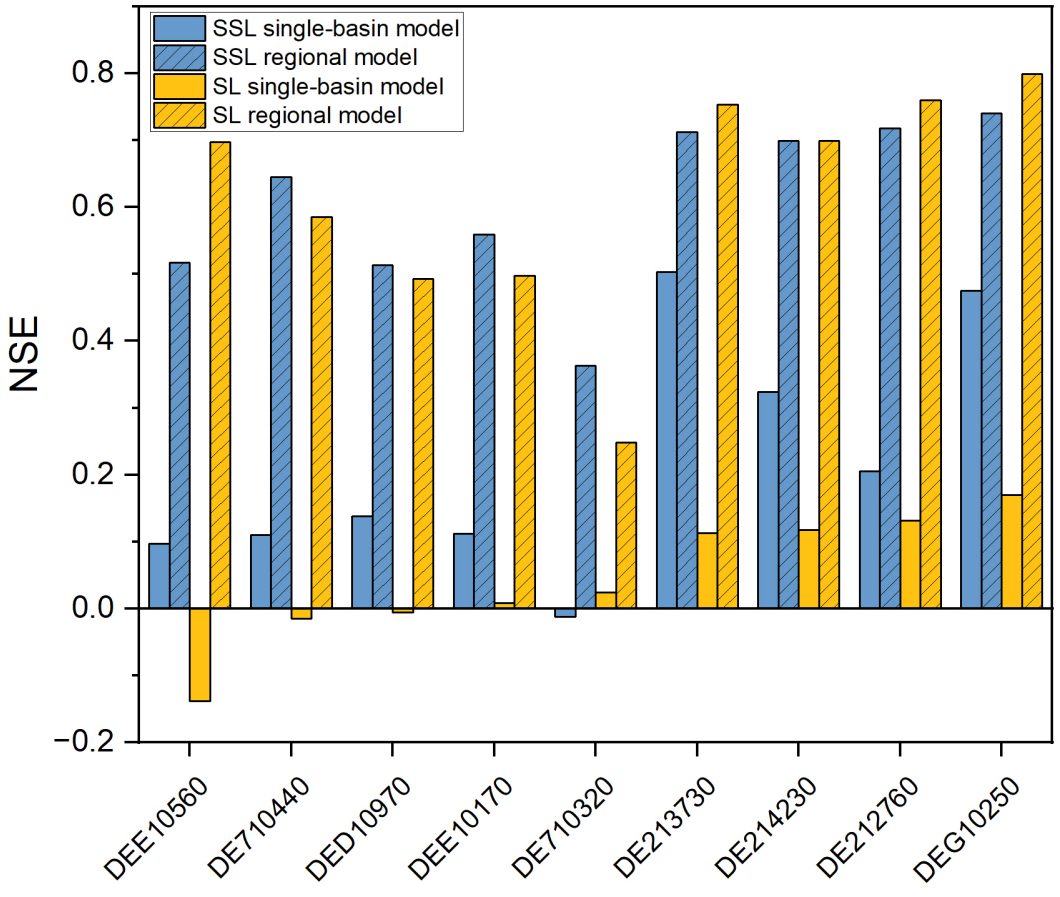

Figure B4. Nash–Sutcliffe Efficiency (NSE) values for regional and single-basin models using the semi-supervised learning (SSL) and baseline supervised learning (SL) approach. Regional models were fine-tuned (SSL) or trained (SL) on one-year labeled sequences from 1,265 basins, while single-basin models were fine-tuned or trained on those from each of nine selected basins (random seed=110).

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

B5 Streamflow prediction results of SSL and SL single-basin models for a flood event in Experiment 3

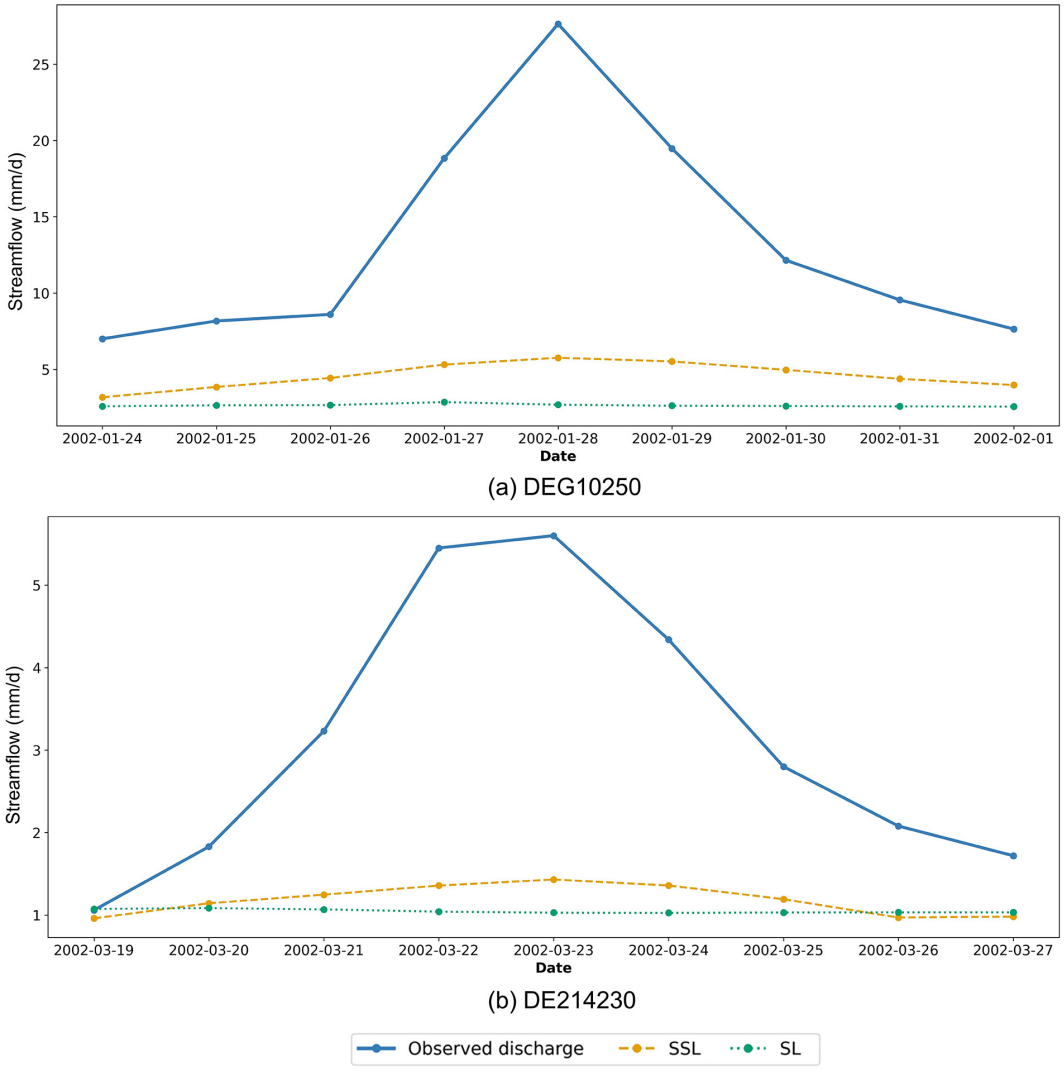

Figure B5. Examples of peak flow events across different basins, with the observed daily streamflow (blue point), and the corresponding streamflow predicted by single-basin models using the semi-supervised learning (SSL) approach and baseline supervised learning (SL) method (random seed=110).

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

Author contributions. **Tianlong Jia**: Conceptualization, Methodology, Software, Validation, Investigation, Writing (original draft preparation), Writing (review and editing), Formal Analysis, Data Curation, Visualization, Project administration, Funding Acquisition. **Guoding Chen**: Methodology, Writing (review and editing), Validation, Visualization, Funding Acquisition. **Yao Li**: Writing (review and editing), Validation, Visualization. **Xinyu Chang**: Writing (review and editing), Visualization. **Uwe Ehret**: Conceptualization, Methodology, Resources, Writing (review and editing), Project administration, Funding Acquisition.

Competing interests. The authors declare that they have no conflict of interest.

Disclaimer. We acknowledge the use of ChatGPT (version 5.3) for assistance with writing style refinement. All methods, results and analyses are solely the work of the authors.

Acknowledgements. Tianlong Jia acknowledges the financial support from (1) the Federal Ministry of Research, Technology and Space (BMFTR) through the "KI-HopE-De" project (No. 16IS24088A), (2) the Open Research Fund of State Key Laboratory of Efficient Utilization of Agricultural Water Resources (Grant No. SKLAWR-KF-2026-12), and (3) the Open Research Fund of Hubei Key Laboratory of Digital River Basin Science and Technology. The work of Guoding Chen was supported by Zhejiang Provincial Natural Science Foundation of China (Grant No. LQN25D010003), and Special Support Fund of Institutes of the Zhejiang Institute of Hydraulics and Estuary (Grant No. ZIHE25Q005). We would like to thank Yiru Jiao, Lijun Wang, Fedor Scholz, Jean-Paul Brede, and Ralf Loritz for fruitful scientific discussions. This work is supported by the Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT partition.

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report, arXiv preprint arXiv:2303.08774, 2023.

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrology and Earth System Sciences, 21, 5293–5313, 2017.

Brown, C. F., Kazmierski, M. R., Pasquarella, V. J., Rucklidge, W. J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., et al.: Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data, arXiv preprint arXiv:2507.22291, 2025.

Chen, G., Zhang, K., Wang, S., Xia, Y., and Chao, L.: iHydroSlide3D v1. 0: an advanced hydrological–geotechnical model for hydrological simulation and three-dimensional landslide prediction, Geoscientific Model Development, 16, 2915–2937, 2023.

Chen, G., Zhang, K., Wang, S., and Jia, T.: PHyL v1. 0: A parallel, flexible, and advanced software for hydrological and slope stability modeling at a regional scale, Environmental Modelling & Software, 172, 105 882, 2024.

Fathi, M. M., Al Mehedi, M. A., Smith, V., Fernandes, A. M., Hren, M. T., and Terry Jr, D. O.: Evaluation of LSTM vs. conceptual models for hourly rainfall runoff simulations with varied training period lengths, Scientific Reports, 15, 15 820, 2025.

Goyal, P., Duval, Q., Seessel, I., Caron, M., Misra, I., Sagun, L., Joulin, A., and Bojanowski, P.: Vision models are more robust and fair when pretrained on uncurated images without supervision, arXiv preprint arXiv:2202.08360, 2022.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of hydrology, 377, 80–91, 2009.

Henaff, O.: Data-efficient image recognition with contrastive predictive coding, in: International conference on machine learning, pp. 4182–4192, PMLR, 2020.

Heudorfer, B. and Loritz, R.: Is smart sampling worth it? Impact of training data selection on the performance of LSTMs in streamflow prediction, Hydrology Research, p. nh2026119, 2026.

Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural computation, 9, 1735–1780, 1997.

Jaiswal, R., Ali, S., and Bharti, B.: Comparative evaluation of conceptual and physical rainfall–runoff models, Applied water science, 10, 48, 2020.

Jia, T., Qin, H., Yan, D., Zhang, Z., Liu, B., Li, C., Wang, J., and Zhou, J.: Short-term multi-objective optimal operation of reservoirs to maximize the benefits of hydropower and navigation, Water, 11, 1272, 2019.

Jia, T., Vallendar, A. J., de Vries, R., Kapelan, Z., and Taormina, R.: Advancing deep learning-based detection of floating litter using a novel open dataset, Frontiers in Water, 5, 1298 465, 2023.

Jia, T., Taormina, R., de Vries, R., Kapelan, Z., van Emmerik, T. H., Vriend, P., and Okkerman, I.: A semi-supervised learning-based framework for quantifying litter fluxes in river systems, Water Research, p. 124833, 2025a.

Jia, T., Yu, J., Sun, A., Wu, Y., Zhang, S., and Peng, Z.: Semi-supervised learning-based identification of the attachment between sludge and microparticles in wastewater treatment, Journal of Environmental Management, 375, 124 268, 2025b.

Jiao, Y., van Cranenburgh, S., Calvert, S., and van Lint, H.: Structure-preserving contrastive learning for spatial time series, Artificial Intelligence for Transportation, 3, 100 031, 2025.

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

Khoshkalam, Y., Rousseau, A. N., Rahmani, F., Shen, C., and Abbasnezhadi, K.: Applying transfer learning techniques to enhance the accuracy of streamflow prediction produced by long Short-term memory networks with data integration, Journal of Hydrology, 622, 129 682, 2023.

Kingma, D. P.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al.: Segment anything, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 4015–4026, 2023.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall-runoff modelling using long short-term memory (LSTM) networks, Hydrology and Earth System Sciences, 22, 6005–6022, 2018.

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

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

Kratzert, F., Gauch, M., Nearing, G., and Klotz, D.: NeuralHydrology — A Python library for Deep Learning research in hydrology, Journal of Open Source Software, 7, 4050, https://doi.org/10.21105/joss.04050, 2022.

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

Kumar, P., Rawat, P., and Chauhan, S.: Contrastive self-supervised learning: review, progress, challenges and future research directions, International Journal of Multimedia Information Retrieval, 11, 461–488, 2022.

Le-Khac, P. H., Healy, G., and Smeaton, A. F.: Contrastive representation learning: A framework and review, Ieee Access, 8, 193 907–193 934, 2020.

Li, Y., Osei, F. B., Hu, T., and Stein, A.: Urban flood susceptibility mapping based on social media data in Chengdu city, China, Sustainable Cities and Society, 88, 104 307, 2023.

Lin, J., Bryan, B. A., Zhou, X., Lin, P., Do, H. X., Gao, L., Gu, X., Liu, Z., Wan, L., Tong, S., et al.: Making China's water data accessible, usable and shareable, Nature water, 1, 328–335, 2023.

Liu, J., Koch, J., Stisen, S., Troldborg, L., Højberg, A. L., Thodsen, H., Hansen, M. F., and Schneider, R. J.: CAMELS-DK: hydrometeorological time series and landscape attributes for 3330 Danish catchments with streamflow observations from 304 gauged stations, Earth System Science Data, 17, 1551–1572, 2025.

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

Ma, K., Feng, D., Lawson, K., Tsai, W.-P., Liang, C., Huang, X., Sharma, A., and Shen, C.: Transferring hydrologic data across continents–Leveraging data-rich regions to improve hydrologic prediction in data-sparse regions, Water Resources Research, 57, e2020WR028 600, 2021.

Mai, J., Shen, H., Tolson, B. A., Gaborit, É., Arsenault, R., Craig, J. R., Fortin, V., Fry, L. M., Gauch, M., Klotz, D., et al.: The great lakes runoff intercomparison project phase 4: the great lakes (GRIP-GL), Hydrology and Earth System Sciences Discussions, 2022, 1–54, 2022.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I—A discussion of principles, Journal of hydrology, 10, 282–290, 1970.

https://doi.org/10.5194/egusphere-2026-1637
Preprint. Discussion started: 12 May 2026

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., et al.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, 2024.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What role does hydrological science play in the age of machine learning?, Water Resources Research, 57, e2020WR028 091, 2021.

Ng, K., Huang, Y., Koo, C., Chong, K., El-Shafie, A., and Ahmed, A. N.: A review of hybrid deep learning applications for streamflow forecasting, Journal of Hydrology, 625, 130 141, 2023.

Ohri, K. and Kumar, M.: Review on self-supervised image recognition using deep neural networks, Knowledge-Based Systems, 224, 107 090, 2021.

Oord, A. v. d., Li, Y., and Vinyals, O.: Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748, 2018.

Ouyang, W., Zhang, C., Ye, L., Zhang, H., Meng, Z., and Chu, J.: Dive into transfer-learning for daily rainfall-runoff modeling in data-limited basins, Journal of Hydrology, 657, 133 063, 2025.

Rogers, J. S., Maneta, M. P., Sain, S. R., Madaus, L. E., and Hacker, J. P.: The role of climate and population change in global flood exposure and vulnerability, Nature Communications, 16, 1287, 2025.

Staudinger, M., Herzog, A., Loritz, R., Houska, T., Pool, S., Spieler, D., Wagner, P. D., Mai, J., Kiesel, J., Thober, S., et al.: How well do process-based and data-driven hydrological models learn from limited discharge data?, Hydrology and Earth System Sciences, 29, 5005–5029, 2025.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I.: Attention is all you need, Advances in neural information processing systems, 30, 2017.

Wang, L., Shi, L., Reimers, C., Wang, Y., He, L., Wang, Y., Reichstein, M., and Jiang, S.: A self-supervised deep learning model for enhanced generalization in soil moisture prediction, Journal of Hydrology, p. 133974, 2025.

Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, arXiv preprint arXiv:2412.10302, 2024.

Yildizli, T., Jia, T., Langeveld, J., and Taormina, R.: Self-supervised learning for multi-label sewer defect classification, Automation in Construction, 182, 106 751, 2026.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water resources research, 44, 2008.

Zhai, X., Oliver, A., Kolesnikov, A., and Beyer, L.: S4l: Self-supervised semi-supervised learning, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 1476–1485, 2019.

Zhang, Y., Ragettli, S., Molnar, P., Fink, O., and Peleg, N.: Generalization of an Encoder-Decoder LSTM model for flood prediction in ungauged catchments, Journal of Hydrology, 614, 128 577, 2022.

Zhong, L., Lei, H., and Yang, J.: Development of a distributed physics-informed deep learning hydrological model for data-scarce regions, Water Resources Research, 60, e2023WR036 333, 2024.