https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

Hydrology and Earth System Sciences
Discussions
Open Access

Semi-supervised learning approach to improve the predictability of data-driven rainfall-runoff model in hydrological data-sparse regions

Sunghyun Yoon[1] and Kuk-Hyun Ahn[2]

June 2023

[1]Assistant Professor, Department of Artificial Intelligence, Kongju National University, Cheon-an, South Korea; e-mail: syoon@kongju.ac.kr

[2]Associate Professor, Department of Civil and Environmental Engineering, Kongju National University, Cheon-an, South Korea; *Corresponding author;* e-mail: ahnkukhyun@kongju.ac.kr

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

Hydrology and Earth System Sciences
Discussions

ABSTRACT

Numerous data-driven models have been introduced to establish reliable predictions in the rainfall-runoff relationship. The majority of these models are trained using a supervised learning (SL) approach, with paired observed samples of climate and streamflow data. However, in practice, the availability of such paired observations is often constrained due to sparse data from streamflow gauges worldwide, which typically covers only a few years. This limited number of paired samples can significantly impede the learning ability of the data-driven model. The semi-supervised learning approach, which is an emerging machine learning paradigm that additionally incorporates unpaired samples, has the potential to be a highly effective method for modeling rainfall-runoff relationships. In this study, we present a novel semi-supervised learning-based framework for rainfall-runoff modeling. Our framework introduces a unique loss function designed to handle two distinct types of samples, namely paired and unpaired samples, effectively during the training process. To validate the effectiveness of the proposed framework, we conducted an extensive set of experiments employing a diverse range of designs, all of which utilized the LSTM network. The experiments are based on 531 basins from the freely available CAMELS dataset, which spans the entire continuous United States. Results indicate that the proposed framework show significantly enhanced performance compared to the baseline models. Results also show that the framework can serve as a viable alternative to the previously developed fully supervised approaches. Lastly, we address potential avenues for enhancing the model and provide an outline of our future research plans in this domain.

Keywords: Long short-term memory (LSTM); Semi-supervised learning; Data-sparse region; Rainfall-runoff modeling; Unpaired samples;

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

Hydrology and Earth System Sciences
Discussions

1. Introduction

Rainfall-runoff modeling is an essential tool for urban planning, land use, flood and water resource management (Nourani et al., 2009). It represents the hydrologic processes involved in converting rainfall into runoff, making it one of the principal interests in hydrological sciences (Beven, 2011; Sitterson et al., 2018). Over time, the modeling of the rainfall-runoff process has evolved from physical-based models such as SHETRAN (Birkinshaw et al., 2010) and VELMA (Mckane et al., 2014) to conceptual models such as Variable Infiltration Capacity (VIC; Liang et al., 1994), Hydrologiska Byråns Vattenbalansavdelning (HBV; Seibert and Vis, 2012), and Sacramento Soil Moisture Accounting (SCA-SMA; Burnash et al., 1973). Data-driven models have also been employed to depict the rainfall-runoff process, with recent studies reporting their ability to outperform traditional models (Hoedt et al., 2021; Lees et al., 2021; Reichstein et al., 2019; Xiang et al., 2020). In this paper, we aim to further improve the predictive ability of the data-driven model, which is currently regarded as the state-of-the-art in hydrologic prediction (Nearing et al., 2021; Shen et al., 2021).

Data-driven models leverage empirical relationships between target and independent variables, offering the advantages of requiring low input, minimal effort for development and application, and moderate computational resources (Abbott, 1999; Chen et al., 2018). Prominent data-driven techniques include genetic programming (Chadalawada et al., 2020), support vector machine (SVM) (Alquraish and Khadr, 2021), random forests (Booker and Woods, 2014), and fuzzy logic (Bartoletti et al., 2018; Kothari and Gharde, 2015). Deep learning (DL) techniques have also gained significant traction for their effectiveness (Roy et al., 2021; Taormina and Chau, 2015; Van et al., 2020; Xie et al., 2021). One standout architecture, the long short-term

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

memory (LSTM; Hochreiter and Schmidhuber, 1997) network, has been specifically designed to simulate time series by incorporating an inductive bias that preserves crucial temporal information over extended periods (Hoedt et al., 2021).

LSTMs have been shown to provide a significant advantage over conventional hydrologic models in rainfall-runoff modeling by a considerable margin (Kratzert et al., 2018; Lees et al., 2021), even at hourly time scales (Gauch et al., 2021a), and for watersheds unseen by the LSTM (Arsenault et al., 2023; Kratzert et al., 2019a). For instance, Kratzert et al. (2019b) demonstrated that when using an LSTM to predict streamflow in 531 basins across the United States (US), it outperformed several different hydrological benchmark models, including SAC-SMA, VIC, and HBV models. In recent years, LSTMs have been utilized to (i) quantify the predictive uncertainty (Klotz et al., 2022; Li et al., 2021), (ii) evaluate the suitability of hydrologic projections under climate change (Wi and Steinschneider, 2022), and (iii) improve the reliability of simulations in hydrologic models as post-processors (Frame et al., 2021; Hunt et al., 2022). Notably, several studies have demonstrated exceptional LSTM performance, especially in situations where abundant data are available (Anderson and Radic, 2021; Gauch et al., 2021b; Lees et al., 2021).

However, acquiring hydrological records with comprehensive long-term coverage is often unattainable in reality. Many regions worldwide face the challenge of limited streamflow gauge networks, resulting in sparse data that typically spans only a few years (Bitew and Gebremichael, 2011; Do et al., 2017). For example, Lee and Ahn (2022) have utilized a limited number of only 27 streamflow gauges to investigate a national-scale hydrologic variability

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

across South Korea. Similarly, in the Tana River basin in Kenya, which covers an expansive area of 95,000 km^2 and serves as a habitat for diverse wildlife species (TNC, 2015), hydrological data from only 26 streamflow gauges spanning a five-year period (February 2015 to January 2020) are available (Leisher et al., 2016). The Global Runoff Data Center (GRDC) provides daily streamflow observations for a significant portion of basins globally, with records typically spanning less than three years in length. The availability of observed streamflow records may also be problematic for data-rich regions due to several situations like gauges installed in recent decades, discontinuation from budgetary constraints, or measuring malfunctioning over an extended period of time (Ahn, 2021). Consequently, many data-driven models have been applied in a local modeling context, wherein a model is trained using data from one or a few basins (e.g., Bowes et al., 2019; Han et al., 2021; Ley et al., 2023; Liang et al., 2018; Xu et al., 2022). The potential limitations associated with sparse records have been discussed, and the need for corrective measures has been addressed (Beven, 2020; Shen, 2018).

In areas where streamflow records are scarce, longer historical climate data records often remain available. However, current methods for training rainfall-runoff models in data-sparse regions typically rely solely on paired recorded samples between climate and streamflow data, known as labeled data in machine learning (ML) terminology. Nonetheless, valuable insights can be gained by incorporating the remaining climate data, referred to as unlabeled data, to improve model performance. In the field of ML, semi-supervised and unsupervised learning are emerging paradigms that utilize unlabeled data to enhance model performance. While semi-supervised learning combines both labeled and unlabeled data to improve performance, unsupervised learning first pre-tunes with unlabeled data before fine-tuning with labeled data

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

(Chen et al., 2022; He et al., 2020). In particular, semi-supervised learning is increasingly recognized as an effective approach with enhanced learning ability (Du et al., 2020; Levatić et al., 2017; Zhou and Zhou, 2021).

While semi-supervised learning has gained popularity in various fields, its formal application in the field of hydrology is still limited. In this study, we introduce a novel framework based on semi-supervised learning for rainfall-runoff modeling, aiming to explore its potential usefulness in structuring hydrological time series modeling problems. Specifically, we present how LSTM models can enhance their predictive performance in regions with limited data, thereby addressing the limitations associated with streamflow observations. In literature, we found two previous studies focusing on improving the modeling performance in data-sparse regions (Ma et al., 2021; Oruche et al., 2021). The approach we propose is notably distinct from those studies in that we do not use any source datasets from other regions. Both of these studies utilize transfer learning, a technique in which a pre-trained model from extensive labeled data from other continents is used to transfer initial weights to a model. In this study, we focus on the dataset obtained from the same region, which is more readily accessible. Summing up, this study seeks to address the following hypotheses using multiple subsets of the continuous United States (CONUS) dataset:

1. The availability of additional climate data, i.e. unlabeled data, could potentially enhance the performance of LSTM models in producing reliable streamflow predictions in diverse modeling scenarios. Therefore, implementing a semi-supervised learning-based framework will be useful.

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

Hydrology and Earth System Sciences
Discussions

2. It would be beneficial to use a semi-supervised learning-based framework that leverages both labeled and unlabeled data, but treats them differently instead of treating them homogeneously. By differentiating between the two datasets and incorporating them into the training process, the model can potentially achieve better performance on unseen data.
3. The joint training of both labeled and unlabeled dataset has the potential to improve model performance in comparison to a separate training approach (i.e., pre-training followed by fine-tuning), which could also be employed to improve modeling performance in data-sparse regions.

Through a series of experiments comparing our proposed semi-supervised learning-based framework to diverse models, we aim to assess the hypotheses mentioned above and gain insight into how LSTM models can enhance performance in data-sparse regions. This exploration will enable us to better understand the benefits of our proposed framework.

2. Methods and Data

This section begins by introducing the dataset used in this study (section 2.1), followed by an overview of the LSTM model structure (section 2.2) and the proposed framework based on the semi-supervised learning (section 2.3). Finally, we outline the specific experimental designs assessed in this study (section 2.4).

2.1 Dataset

To investigate the effectiveness of semi-supervised learning in analyzing the streamflow network, which covers various geologies and climatic conditions, we use the Catchment Attributes and MEteorology for Large-sample Studies (CAMELS) dataset (Newman et al., 2015). We utilize the dataset due to their abundance of data, which could potentially strengthen the validation of our hypotheses (addressed in section 2.4). The dataset includes the basin-averaged hydrometeorological time series, catchment characteristics and daily streamflow measurements for 671 basins over the CONUS. It is worth noting that the dataset has been widely utilized to facilitate generalization and application of data-driven models for various purposes (e.g., Feng et al., 2020; Gauch et al., 2021b; Kratzert et al., 2019b). We have adopted the same subset of 531 basins as Gauch et al. (2021b) and Kratzert et al. (2019b) (see Figure 1) while excluding 140 basins that display considerable inconsistencies in their calculated watershed boundaries from different methodologies. Consistent with the aforementioned studies, we utilize the Maurer meteorological forcing dataset, which includes daily cumulative precipitation ($PRCP$), maximum and minimum air temperature (T_{max} and T_{min}), short-wave radiation ($SRAD$), and vapor pressure (VP), spatially averaged for each basin. Furthermore, we use 27 of the static catchment characteristics, including topography, climate characteristics, land cover, soil and geology characteristics (Table 1). The spatially aggregated data have been derived from an original gridded dataset that has a resolution of 1/8°. The meteorological forcing and streamflow data are normalized so that all variables for each basin has a mean of zero and unity variance.

2.2 Long Short-Term Memory network

In this work, we utilize a LSTM architecture for the rainfall-runoff modeling. A LSTM network is a type of recurrent neural network designed to model long-term dependencies between input and output data. LSTMs utilize an internal memory state that is updated at each time step by a set of activated functions called *gates* (Hochreiter and Schmidhuber, 1997). The memory cells are comparable to a state vector in a traditional dynamic system model, which leads to the suitability of LSTMs for modeling dynamics such as rainfall-runoff relationships. Compared to vanilla recurrent neural networks, LSTMs are less affected by the vanishing gradient issue that has prevented effective model learning (Hochreiter and Schmidhuber, 1997). Given a raw input sequence $\boldsymbol{x_0} = [x_0^1, x_0^2, \dots, x_0^T]$ with T time steps, where each element $\boldsymbol{x_0^t}$ is a vector containing model input at time step *t*, we specifically employed the following equations for the forward pass through the LSTM:

$$\boldsymbol{x^t} = \mathfrak{D}(GELU(\boldsymbol{W_x x_0^t} + \boldsymbol{b_x})) \qquad \text{Eq. (1)}$$

$$\boldsymbol{i^t} = \sigma(\boldsymbol{W_i x^t} + \boldsymbol{U_i h^{t-1}}) + \boldsymbol{b_i} \qquad \text{Eq. (2)}$$

$$\boldsymbol{f^t} = \sigma(\boldsymbol{W_f x^t} + \boldsymbol{U_f h^{t-1}}) + \boldsymbol{b_f} \qquad \text{Eq. (3)}$$

$$\boldsymbol{g^t} = \tanh(\sigma(\boldsymbol{W_g x^t} + \boldsymbol{U_g h^{t-1}})) + \boldsymbol{b_g} \qquad \text{Eq. (4)}$$

$$\boldsymbol{o^t} = \sigma(\boldsymbol{W_o x^t} + \boldsymbol{U_o h^{t-1}}) + \boldsymbol{b_o} \qquad \text{Eq. (5)}$$

$$\boldsymbol{c^t} = \boldsymbol{g^t} \odot \boldsymbol{i^t} + \boldsymbol{c^{t-1}} \odot \boldsymbol{f^t} \qquad \text{Eq. (6)}$$

$$\boldsymbol{h^t} = tanh(\boldsymbol{c^t}) \odot \boldsymbol{o^t} \qquad \text{Eq. (7)}$$

$$\hat{y}^t = \boldsymbol{W_y} \mathfrak{D}(\boldsymbol{h^t}) + \boldsymbol{b_y} \qquad \text{Eq. (8)}$$

where $\boldsymbol{i^t}$, $\boldsymbol{f^t}$, $\boldsymbol{o^t}$, and $\boldsymbol{g^t}$ are the input gate, forget gate, output gate, and cell input, respectively, at time step *t*. The cell state and recurrent input are denoted by $\boldsymbol{c^t}$ and $\boldsymbol{h^t}$. $GELU$ refers to the Gaussian Error Linear Units (Hendrycks and Gimpel, 2016). Also, two activation functions, sigmoid and hyperbolic tangent, are denoted by σ and tanh. $\boldsymbol{W}$, $\boldsymbol{U}$, and $\boldsymbol{b}$ are learnable parameters for each gate, where subscripts suggest which gate the weight vector is used for,

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

and $\odot$ represents element-wise multiplication. The dropout operator is denoted by $\mathfrak{D}$, which randomly sets some nodes along with corresponding of the network connections to zero in training phase in order to reduce overfitting (Srivastava et al., 2014).

A linear embedding layer (Eq. 1) is utilized to preprocess the inputs before delivering them to the LSTM cell, in order to prevent the possibility of critical inputs being dropped out by the $\mathfrak{D}$ operators and thereby reducing the model's performance.

2.3 Semi-supervised learning-based framework to improve hydrologic prediction

The proposed semi-supervised learning-based framework is an improved version of self-training methods proposed in the ML field (Hinton et al., 2015; Yarowsky, 1995). Self-training methods utilize unlabeled data by imputing predicted labels (called pseudo labels) to the unlabeled data. Specifically, our approach is based on knowledge distillation, where the pseudo labels for unlabeled dataset are generated from a pre-trained teacher model trained on labeled dataset. Student model is trained in supervised manner on both the labeled and (pseudo label-assigned) unlabeled datasets. In this work, both teacher and student models have the same structure, as in self-distillation (Zhang et al., 2019).

Suppose that we are given a set of data $\mathbb{D}$ including labeled data $\mathbb{L}^t = \{(\boldsymbol{x_0}_1^t, y_1^t), (\boldsymbol{x_0}_2^t, y_2^t), \ldots, (\boldsymbol{x_0}_N^t, y_N^t)\}$ at basin $(n = 1, \ldots, N)$ and time $(t = 1, \ldots, T)$ and unlabeled data $\mathbb{U}^t = \{\boldsymbol{x_0}_1^t, \boldsymbol{x_0}_2^t, \ldots, \boldsymbol{x_0}_N^t\}$ at time $(\mathcal{t} = 1, \ldots, \mathcal{T})$. The framework requires two input datasets ($\mathbb{L}^t$ and $\mathbb{U}^t$). The labeled data $\mathbb{L}^t$ is employed to train a teacher LSTM model by minimizing a loss function. The teacher model is then used to estimate streamflow (i.e., pseudo streamflow $\hat{y}_n^t$)

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

on unlabeled data $\mathbb{U}^{t}$. Afterwards, we train a student LSTM model by minimizing the combined loss on both labeled and unlabeled data with pseudo streamflow. Finally, we repeat the process by reinstating the student as a teacher, which generates new pseudo streamflow, allowing us to train a new student. A flowchart of the proposed semi-supervised learning-based framework is presented in Figure 2. The algorithmic procedure for the framework is as follows.

[1] Train the teacher model $f^*(\cdot)$ by minimizing the loss function L on labeled data $\mathbb{L}^t$

$$\frac{1}{N\times T}\sum_{n=1}^{N}\sum_{t=1}^{T} L\left(y_n^t, f^*\left(\boldsymbol{x_0}_n^t, \boldsymbol{\theta}^*\right)\right) \qquad \text{Eq. (9)}$$

[2] Use the teacher model to generate pseudo streamflow for unlabeled data $\mathbb{U}^{t}$

$$\hat{y}_n^t = f^*\left(\boldsymbol{x_0}_n^t, \boldsymbol{\theta}^*\right), \qquad \forall t = 1, \dots, \mathcal{T} \qquad \text{Eq. (10)}$$

[3] Train the student model $f^{**}(\cdot)$ by minimizing the below loss function L to consider the training balance between labeled and unlabeled data

$$\frac{1}{N\times T}\sum_{n=1}^{N}\sum_{t=1}^{T} L\left(y_n^t, f^{**}\left(\boldsymbol{x_0}_n^t, \boldsymbol{\theta}^{**}\right)\right) + \alpha(\mathbb{t})\frac{1}{N\times \mathcal{T}}\sum_{n=1}^{N}\sum_{t=1}^{\mathcal{T}} L\left(\hat{y}_n^t, f^{**}\left(\boldsymbol{x_0}_n^t, \boldsymbol{\theta}^{**}\right)\right) \qquad \text{Eq. (11)}$$

where $\alpha(\mathbb{t})$ is a balance coefficient at epoch $\mathbb{t}$.

The suitability of $\alpha(\mathbb{t})$ significantly impacts the performance of the student model. When $\alpha(\mathbb{t})$ is high, the loss function is primarily influenced by $\mathbb{U}^{t}$, whereas a small value allows the benefits from $\mathbb{L}^t$ to become more apparent. Consequently, to mitigate the risk of ending up in poor local optima, we employ an annealing process that incorporates a $\mathbb{t}$-varying $\alpha(\mathbb{t})$ as follows:

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

Hydrology and Earth System Sciences
Open Access
Discussions

$$\alpha(\mathfrak{t}) = \begin{cases} \alpha_0 & \mathfrak{t} \leq \mathcal{J} \\ 0 & \mathcal{J} < \mathfrak{t} \end{cases} \qquad \text{Eq. (12)}$$

where this study utilizes α_0 = 1 and $\mathcal{J} = 15$ based on the epoch adopted in this study.

[4] Further train the student model as a teacher model and go back to step [2]. Our experiment involves two iterations where the student assumes the role of the teacher, but it may be beneficial to conduct additional iterations.

2.4 Experiments

To depict the situation in hydrological data-sparse regions, this study considers two dimensions for each research hypothesis. First, the Ψ subset of basins, rather than considering all 531 basins, are utilized under two differently defined regions (heterogeneous and homogeneous regions). The approach is adopted since, in data-scarce regions, the numbers of the streamflow gauge are also limited in reality. To explore the performance in heterogeneous regions, the Ψ subset of basins over the CONUS are randomly selected, and the model performance is investigated. The purpose of this analysis is to emulate the diverse environmental factors that exist across the subspace of the considered area. This analysis is repeated 3 times to match the experiment trial conducted in heterogeneous regions. For the analysis of homogeneous areas, this study employs three regions, namely the North Atlantic, Southwest, and Southern Rockies regions (Figure 1). The purpose of this analysis is to take into account a broad spectrum of environmental conditions while reproducing a homogeneous environmental situation within the target region. The North Atlantic region comprises 84 basins that are moderately affected

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

Hydrology and Earth System Sciences
Discussions

by snow accumulation and melting processes and are predominantly covered by dense forests (with an average forest coverage of 89%). In contrast, the Southwest region comprises 66 basins with relatively flat topography and lesser snow influence compared to other regions. Lastly, the Southern Rockies region (50 basins) is significantly influenced by snow process and consists relatively arid catchments (aridity index 1.71 and average annual precipitation 700 mm/year). The three cases consider scenarios in which Ψ takes on values of 10, 30, and 50, representing situations where the basin density network is relatively deficient, moderate, and sufficient, respectively.

Second, we consider two training scenarios, single and multi-year training scenarios, to represent the available data in data-scare regions. It is worth mentioning that in the data-scare regions, streamflow records often have a restricted length. For the single-year training scenario, all models are trained from October 1, 1988 to September 30, 1989 and validated from October 1, 1989 to September 30, 1990. For unlabeled extended data, we employ data from October 1, 1983 to September 30, 1988. The models are then evaluated over 12 years (October 1, 1996 to September 30, 2008). For the multi-year training scenario, all models are trained for 3 years (October 1, 1988 to September 30, 1991) and validated over 2 years (October 1, 1991 to September 30, 1993). Afterwards, we use the same data in unlabeled and evaluation periods adopted for the single-year training scenario.

2.4.1 The effect of semi-supervised learning on individual and regional setting

With the first experiment, this study evaluates our proposed framework if it bolsters the ability of rainfall-runoff modeling. In particular, we hypothesize that implementing a semi-supervised learning-based framework would yield benefits in a diverse model setting. To confirm this

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

hypothesis, we run this experiment with individual and regional model setting. For the individual setting, we train one network separately for each basin (hereafter idv-LSTM). On the other hand, for the regional setting, we train a regional scale single network using all data across multiple basins while allowing the network to learn more general pattern of the input-to-output relationship (hereafter rgn-LSTM). Although the benefits of the proposed framework may be expected in idv-LSTM due to increased learning data, it is unclear whether there would be additional benefits in a rgn-LSTM. Previous studies have shown that LSTM predictions are reliable when the model is trained over a large set of basins and that regional models already learn more general patterns from a diverse set of basins (e.g, Gauch et al., 2021b). Therefore, it remains to be seen whether the proposed framework would offer additional benefits beyond those already achieved by regional models trained on diverse basin data. In addition, we will evaluate the performance of a regional model in this experiment by increasing the amount of training data, specifically by including a larger number of basins in rgn-LSTM. This will allow us to determine the maximum number of basins for which the proposed framework offers additional benefits, once its effectiveness in a regional setting is confirmed. To establish a comparison, we obtain simulation results from a LSTM using a standard train-validation-testing framework. These results are then used as the baseline for evaluating the performance of our proposed semi-supervised learning-based framework.

2.4.2 The effect of the annealing process on the student model

The objective of the second set of experiments is to examine the impact of the annealing process (Eq. 12) adopted in the proposed framework. In simpler terms, we hypothesize that utilizing an imbalanced-based cost function with both labeled and unlabeled data would enhance the accuracy of the model. The rationale behind this is that by accounting for pseudo streamflow

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

and exploiting their impact, it enables the data-driven model to effectively learn the underlying hydrologic response to input variables.

To investigate this, we employ variants of the rgn-LSTM model. Specifically, we additionally develop five different versions of the rgn-LSTM model within a semi-supervised learning-based framework. The first additional model (rgn-LSTM-vr1) treats the equivalent data for $\mathbb{L}^t$ and $\mathbb{U}^t$ by replacing $\alpha(\mathbb{t})$ with a value of 1 and does not provide a distinguishing weight. The remaining four models (rgn-LSTM-vr2, rgn-LSTM-vr3, rgn-LSTM-vr4, and rgn-LSTM-vr5) use $\mathbb{t}$-varying weight but adopt different formulations from Eq. 12. They are designed to amplify, slowly increase, or slowly decrease the influence of $\mathbb{U}^t$. The specific $\alpha(\mathbb{t})$ configurations are presented in the supporting information (see Text S1). Also, Figure S3 shows how the $\mathbb{t}$-varying weight is changed given increases to the epoch for each rgn-LSTM model.

2.4.3 Comparison of our proposed framework to the separate training approaches

Previous studies have suggested the separate training approach as a means of improving neural network models (Anderson and Radic, 2021; Read et al., 2019). In this approach, a model is first pre-trained on a specific dataset to learn general patterns and relationships between input and output data. The model is then fine-tuned on an additional dataset to learn more specific behaviors and improve its performance on a particular task. This process allows the model to adapt to the nuances of the task at hand, and has been shown to be effective in the ML field as well (George et al., 2017; Yosinski et al., 2014).

For our third experiment, we aim to determine whether our proposed framework, which incorporates joint training using both labeled and unlabeled data, can achieve better results compared to the separate training approaches. To accomplish this, we introduce two additional rgn-LSTM models: rgn-LSTM-sep and rgn-LSTM-trans. The rgn-LSTM-sep model initially leverages unlabeled data $\mathbb{U}^t$ to capture the underlying patterns of runoff generating processes. Subsequently, it undergoes fine-tuning on labeled data $\mathbb{L}^t$ to refine its performance, specifically considering the delicate input and output relationships within specific basins. On the other hand, the rgn-LSTM-trans model incorporates the recent technique proposed by Ma et al. (2021), which utilizes Transfer Learning (Thrun and Pratt, 1998) (see Supplemental information for a brief description of the method). They adopt a methodology wherein the models are initially pretrained on a region abundant in data (known as the source region). These pretrained models are subsequently transferred to data-scarce regions to overcome the limitations of local observations. For this study, we employ the CAMELS-GB dataset, which is a comprehensive dataset for Great Britain based on the CAMELS framework (Coxon et al., 2020), as our source dataset. The dataset is selected because the CAMELS-GB basins exhibit a wide range of hydrological conditions, analogous to the conditions found in our study basins. To be specific, the rgn-LSTM-trans model is pretrained using 44 climate and basin attributes from the CAMELS-GB dataset (as shown in Table S1).

2.4.4 Evaluation metrics and hyperparameters

To evaluate the modeling performance for each experiment, we run all models with four random seeds and use the average estimated streamflow obtained from the resulting ensemble members. The first metric used to assess the performance is the Nash-Sutcliffe efficiency (NSE) coefficient (Nash and Sutcliffe, 1970), which is calculated for each basin. Also, we utilize two metrics to evaluate the model's performance for both extreme flows: the modified Nash–

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

Sutcliffe efficiency (MNSE) and the logged transformed Nash-Sutcliffe efficiency (LNSE). These metrics specifically focus on the performance of the model for high and low flows, respectively (Ahn et al., 2016; Muleta, 2011). It is important to note that all of the metrics reported in the manuscript are calculated based on the evaluation period.

Hyperparameters including learning rate, hidden states, length of input sequence, dropout rate, epochs, and numbers of LSTM layer are configurations of LSTM model and thus yield varying degrees of influence on the model's performance (Bengio et al., 2017). To avoid potential bias in performance evaluation that may favor our proposed framework, we choose to adopt the same hyperparameter configurations used in previous studies (Kratzert et al., 2021, 2019b), rather than determining new ones for this study. Finally, all model configurations are trained using the mean squared error (MSE) metric similar to the previous work.

3. Results

3.1 Evaluating semi-supervised learning in data-scarce regions

In this section, we assess the effectiveness of the proposed semi-supervised learning-based framework in enhancing streamflow predictions. Figures 3 and 4 illustrate the spatial distribution of the NSE difference for both idv-LSTM and rgn-LSTM cases, respectively, during the evaluation period in comparison to the baseline models. Figures S2, S3 and S4 present the differences in other metrics (MNSE and LNSE) for the idv-LSTM and rgn-LSTM settings. In each figure, the red color indicates that our proposed framework outperforms the baseline models in terms of prediction accuracy, while the blue color indicates that our proposed framework underperforms the baseline models. Additionally, Table 2 provides a summary of the median performance across all experiments, encompassing the three evaluation metrics. Notably, our framework exhibits improvements across all three metrics, underscoring its effectiveness. The single-year training scenario in idv-LSTM stands out by yielding the most significant benefits, with a notable improvement of 0.390 in median NSE and 0.238 in MNSE. Similarly, the multi-year training scenario in idv-LSTM exhibits substantial improvements, where our semi-supervised learning approach yields remarkable improvements of 0.165 in median NSE and 0.374 in LNSE. The results indicate that our proposed framework offers greater advantages when addressing regions with limited data availability, particularly in data-scarce areas where the available data is relatively smaller.

Moreover, our proposed framework delivers substantial benefits in the context of rgn-LSTM. When considering the single-year training scenario for the deficient network (i.e., $\Psi = 10$) in rgn-LSTM across all six heterogeneous and homogeneous regions, it demonstrates a remarkable improvement of 0.371 in median NSE, 0.275 in MNSE, and 0.560 in LNSE. Similarly, in the case of the sufficient network (i.e., $\Psi = 50$), the multi-year training scenario yields an improvement of 0.023 in median NSE, 0.027 in MNSE, and 0.062 in LNSE. The results reveal several notable insights. Firstly, similar to idv-LSTM, our framework demonstrates increased effectiveness when dealing with insufficient records. This highlights its utility in situations where data availability is significantly limited. We also note that, for some basins particularly in the Southern Rockies region, the baseline model performs better than the models trained by our framework (see Figure 4). The performance declines may be related to frequently having zero discharge in observation. Having zero values for a high percentage of the training samples seems to be a difficult information for the teacher model to learn and to reproduce this hydrological behavior and affect the performance of the student model. However, we observe that the median of a metric is still positive, indicating that the

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

models trained by our framework performs are effective. Next, the efficacy of our framework extends to relatively large streamflow networks, as evidenced by our results in a network comprising 50 basins. Our proposed framework offers additional benefits that surpass those achieved by regional models trained on diverse basin data, even when rgn-LSTM has already learned general patterns from a diverse set of basins. This is particularly relevant in data-scarce regions where some streamflow stations may be available with limited records. Consequently, the findings from this analysis provide valuable insights that can guide the practical implementation of our framework in real-world applications, addressing the challenges posed by data scarcity in streamflow prediction.

It is important to highlight that the effectiveness of our proposed framework is especially pronounced when using a separate network for each basin (idv-LSTM). Also, there is an expectation that rgn-LSTM would still exhibit improvement when utilizing a semi-supervised learning-based framework. This suggests that employing a single setting for our remaining assessment is acceptable. Furthermore, as previously mentioned, it is probable that some streamflow stations are available even in data-scarce regions. This suggests that conducting an analysis by combining data from those stations with regional models trained on multiple basin data would offer a more realistic evaluation. Therefore, for the remaining analysis, we will adopt rgn-LSTM particularly with the moderate density network.

3.2 Evaluating the selection of the annealing process on the student model

In the first experiment, we confirm the benefits of a semi-supervised learning-based framework in enhancing streamflow predictions. We now analyze the performance of six rgn-LSTM models (see Figure S1) to explore the appropriateness of the annealing process (Eq. 12) adopted

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

in the proposed framework. Figures 5 and S5 show that results of rgn-LSTM obtained over the evaluation period with the single and multi-year training scenarios, respectively.

Those figures present two noteworthy observations. First, the performance of the rgn-LSTM-vr1 model is notably lower compared to the other models. Specifically, significant declines in performance are observed for the single-year training scenario, while its performance remains similar to the other models for the multi-year training scenario. However, even in the multi-year training scenario, lower performance is evident in LNSE particularly for the heterogeneous region. These findings suggest that incorporating an imbalanced-based cost function between labeled and unlabeled data enhances the model's predictive capabilities. Next, the models employing structures that diminish the influence of unlabeled data (e.g., rgn-LSTM-vr5 and rgn-LSTM) show better results compared to the models that amplify the role of unlabeled data (e.g., rgn-LSTM-vr2 and rgn-LSTM-vr4) particularly in the single-year training scenarios. There could be multiple factors contributing to this disparity, but our inference is that the outperformance may be attributed to the low quality in learning of the teacher model due to insufficient data. The low quality for the teacher model potentially affects the quality of the unlabeled data. By leveraging the expanded training data that includes unlabeled data, the student model can gain a rough understanding of streamflow modeling. This initial exploration proves beneficial, allowing the model to converge quickly and reducing the chances of overfitting. Subsequently, the network undergoes fine-tuning using high-quality labeled data on the latter part of the epoch progression. Therefore, the models employing structures to diminish the influence of unlabeled data would be beneficial. Our inference is also supported by the multi-year training scenario. While rgn-LSTM remains competitive, its superiority becomes less apparent due to the improved learning of the teacher model resulting from the

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

expanded training samples. Summing up, by differentiating between the two unlabeled and labeled datasets, the model can potentially achieve better performance in a semi-supervised learning-based framework.

3.3 Comparing our proposed model with rgn-LSTM-sep and rgn-LSTM-trans

Finally, we compare our semi-supervised learning-based framework with two separate training approaches, namely rgn-LSTM-sep and rgn-LSTM-trans. Figures 6 and S6 present the spatial distribution of the NSE, MNSE, and LNSE metric differences for the three regions in the heterogeneous and homogeneous regions, respectively. The figures illustrate the relative performance of our framework's models compared to the two fine-tuning models. Here, the utilization of the red color highlights instances where our proposed framework surpasses a separate training approach in terms of prediction accuracy in the evaluation period.

Based on the comparison between rgn-LSTM and rgn-LSTM-sep, the benefits of utilizing a semi-supervised learning approach over relying solely on weight initialization using unlabeled data (corresponding to pre-training) are evident. For example, the single-year training scenario in heterogeneous region yields notable benefits, with an improvement of 0.024 in median NSE and 0.031 in MNSE. Similarly, the multi-year training scenario also show substantial improvements, where our semi-supervised learning approach yields remarkable improvements of 0.020 in median NSE and 0.044 in MNSE. Particularly, the most significant improvement is observed in mean LNSEs for both scenarios, with our proposed framework achieving a noteworthy improvement of 0.077 and 0.090 when compared to rgn-LSTM-sep. The results indicate that the joint training of both labeled and unlabeled datasets leads to better performance than the separate training approach that utilizes weight initialization only with

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

unlabeled data.

When comparing rgn-LSTM and rgn-LSTM-trans, we observe slight differences in contrast to the results between rgn-LSTM and rgn-LSTM-sep, suggesting that those approaches (rgn-LSTM and rgn-LSTM-trans) provide reliable predictions in the rainfall-runoff relationship. However, rgn-LSTM tends to exhibit higher prediction accuracy overall compared to rgn-LSTM-trans. This is especially evident when considering the MNSE and LNSE metrics, highlighting the effectiveness of a semi-supervised learning-based framework in more accurately representing local extreme flows. One possible explanation for the performance difference between rgn-LSTM and rgn-LSTM-trans is that transfer learning can be effective when the dataset in both the source and target regions are sufficiently similar. However, rgn-LSTM-trans is based on pre-trained knowledge from the source region, utilizing 44 forcing variables that substantially differ from the attributes employed in this study (refer to Table 1). It is commonly referred to as negative transfer learning (Torrey and Shavlik, 2010). Our inference is supported by the outperformance of rgn-LSTM in the Southwest region, which exhibits conditions fairly different from those of Great Britain (see Figure S6). It is also worth noting in this comparison that rgn-LSTM may have a disadvantage due to the fact that rgn-LSTM-trans is trained using a 10-year labeled dataset, which is nearly double the additional data used in the training process for rgn-LSTM. Taken together, we therefore consider a semi-supervised learning-based approach a useful and complementary approach to the transfer learning approach, but would caution against using it as a replacement for bolstering the ability of rainfall-runoff modeling in all cases.

4. Discussion

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

4.1 Impact of the performance of teacher model

The predictive capability of the teacher model is vital within the proposed semi-supervised learning-based framework. This is because the teacher model is employed to generate pseudo streamflow on unlabeled data. Consequently, an enhanced performance of the teacher model is anticipated to result in greater improvements within the framework. Figure 7 shows the accuracy improvements obtained in the semi-supervised learning-based framework relative to the baselines when they are compared to the performance of the teacher model. The figure presents results for two model settings, idv-LSTM and rgn-LSTM with the moderate density network. It is important to note that similar patterns are observed in the other results (not shown). Interestingly, the impact of the performance of the teacher model is different from our expectation. While there are slight variations in each plot, the anticipated improvement (indicated by the red lines) generally follows an upward trend, reaching its peak around a NSE value of 0.4, and subsequently experiencing a decline in improvement.

These findings suggest that achieving higher performance in the teacher model does not necessarily translate into greater improvement within the framework. One potential explanation is that the involvement of numerous latent processes in the rainfall-runoff process. These processes include factors such as subsurface interactions (e.g., aquifer dynamics and transmissivity). Due to the complexity of these confounding factors, it becomes challenging for the network to further capture the entire runoff generation process especially in the basins well trained by the student network. Instead, our analysis shows that the proposed framework exhibits its highest effectiveness in the study basins when the network achieves a moderate level of accuracy, specifically around an NSE value of approximately 0.4 when the baseline network is applied.

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

4.2 Applicability of the semi-supervised learning-based framework

While semi-supervised learning holds significant promise, further exploration is encouraged to assess its applicability, particularly in the context of time series tasks. The successful application of semi-supervised learning has predominantly been observed in computer vision tasks (Chen et al., 2021; Yang et al., 2019), but it has also demonstrated success in other machine learning domains (Wang et al., 2021; Zhu et al., 2021). However, the success has rarely been extended to time series related tasks (Dai et al., 2023). This scarcity of success in time series tasks further underscores the significance and value of the present study.

Furthermore, the improvement of hydrological modeling initiatives has been dependent on both sufficient data collection and enhancements in the model's algorithm to an equal degree. The sparsity and inconsistency of the meteorological dataset additionally result in low performance in the streamflow prediction and create a problematic situation to implement our proposed framework. In developing countries, the situation arises due to the insufficient availability of equipment used for monitoring meteorological data, such as precipitation and air temperature. To be specific, the lack of sufficient data for tracking meteorological information in African countries contributes to the encountered situation. Although initiatives like The trans-African hydrometeorological observatory (TAHMO) have been launched (van de Giesen et al., 2014), there is still a significant gap in data availability and coverage. As a potential solution to mitigate the situation, we can consider the utilization of reanalysis-based climate data, such as the global dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). Additionally, employing approaches like statistical downscaling of these global

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

datasets, as demonstrated in the studies by Voropay et al. (2021) and Xie et al. (2022), could prove effective in mitigating the challenges presented by the limited meteorological data.

4.3 Future work

This paper utilizes a semi-supervised learning approach to improve the predictive ability in rainfall-runoff modeling while addressing the limitations associated with streamflow observations. The proposed framework utilizes predicted streamflow estimated by a teacher model as pseudo-labels, indicating high-quality pseudo-labels is important for the performance of the student model. However, this study does not address the issue of uncertainty associated with these pseudo-labels. One potential solution is to employ Bayesian neural networks (BNNs; Kendall and Gal, 2017), which effectively handle input data noise, known as aleatoric uncertainty, by incorporating its impact into the loss function. This utilization of BNNs as a heteroscedastic modeling technique may be useful to reduce prediction variance and enhance the quality of pseudo-labels obtained. Our team intends to explore this approach in the near future as part of our ongoing research efforts.

5. Conclusions

The science of hydrology has primarily evolved by leveraging established physical and empirical relationships to comprehend the complex dynamics of rainfall-runoff interactions. Although significant progress has been made in harnessing data-driven models to enhance insights and intuition derived from abundant hydrological dataset, a fundamental obstacle remains due to the scarcity of available data.

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

In this study, we developed a semi-supervised learning-based framework to mitigate the challenges associated with predicting streamflow in regions with limited data availability. The framework enables a data-driven model to enhance its training dataset by incorporating additional climate data, even in scenarios with limited paired records of climate and streamflow data. This is achieved through the generation of pseudo streamflow data. In particular, we introduced a novel loss function for the student model, designed to effectively distinguish the contributions of labeled and unlabeled data to the loss function during the training process. Through a range of diverse experimental designs, we conducted extensive validation to demonstrate the substantial efficacy of the proposed framework in comparison to a simple baseline model. Lastly, we conducted a thorough comparison between our proposed framework and two separate training approaches, affirming the effectiveness of our framework. We firmly believe that the value of this framework is immense, as it capitalizes on the availability of longer historical climate data records, including the utilization of global climate datasets. This is particularly advantageous in regions where streamflow records are scarce, as it facilitates the extraction of valuable insights from the wealth of accessible climate data.

Code availability

The code is available upon the request to the corresponding author.

Author contribution

Sunghyun Yoon carried out data acquisition, formal analysis, investigation, visualization, and writing. Kuk-Hyun Ahn conceptualized the study, designed the methodology, performed formal analysis, participated in writing the draft, and served as supervisor of the study.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00208210).

References

Abbott, M.B., 1999. Introducing hydroinformatics. J. Hydroinformatics 1, 3–19.

Ahn, K.-H., 2021. Streamflow estimation at partially gaged sites using multiple-dependence conditions via vine copulas. Hydrol. Earth Syst. Sci. 25, 4319–4333.

Ahn, K.-H., Steinschneider, S., Palmer, R., 2016. A hierarchical Bayesian model for regionalized seasonal forecasts of low flows in the northeastern United States. Water Resour. Res.

Alquraish, M.M., Khadr, M., 2021. Remote-Sensing-Based Streamflow Forecasting Using Artificial Neural Network and Support Vector Machine Models. Remote Sens. 13, 4147.

Anderson, S., Radic, V., 2021. Evaluation and interpretation of convolutional-recurrent networks for regional hydrological modelling. Hydrol Earth Syst Sci Discusspreprint Httpsdoi Org105194hess-2021-113 Rev.

Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F., Mai, J., 2023. Continuous streamflow prediction in ungauged basins: long short-term memory neural networks clearly outperform traditional hydrological models. Hydrol. Earth Syst. Sci. 27, 139–157.

Bartoletti, N., Casagli, F., Marsili-Libelli, S., Nardi, A., Palandri, L., 2018. Data-driven rainfall/runoff modelling based on a neuro-fuzzy inference system. Environ. Model. Softw. 106, 35–47.

Bengio, Y., Goodfellow, I., Courville, A., 2017. Deep learning. MIT press Cambridge, MA, USA.

Beven, K., 2020. Deep learning, hydrological processes and the uniqueness of place. Hydrol. Process. 34, 3608–3613.

Beven, K.J., 2011. Rainfall-runoff modelling: the primer. John Wiley & Sons.

Birkinshaw, S.J., James, P., Ewen, J., 2010. Graphical user interface for rapid set-up of SHETRAN physically-based river catchment model. Environ. Model. Softw. 25, 609–610.
Bitew, M.M., Gebremichael, M., 2011. Assessment of satellite rainfall products for streamflow simulation in medium watersheds of the Ethiopian highlands. Hydrol. Earth Syst. Sci. 15, 1147–1155.
Booker, D., Woods, R., 2014. Comparing and combining physically-based and empirically-based approaches for estimating the hydrology of ungauged catchments. J. Hydrol. 508, 227–239.
Bowes, B.D., Sadler, J.M., Morsy, M.M., Behl, M., Goodall, J.L., 2019. Forecasting groundwater table in a flood prone coastal city with long short-term memory and recurrent neural networks. Water 11, 1098.
Burnash, R.J., Ferral, R.L., McGuire, R.A., 1973. A generalized streamflow simulation system, conceptual modeling for digital computers.
Chadalawada, J., Herath, H., Babovic, V., 2020. Hydrologically informed machine learning for rainfall-runoff modeling: A genetic programming-based toolkit for automatic model induction. Water Resour. Res. 56, e2019WR026933.
Chen, I.-T., Chang, L.-C., Chang, F.-J., 2018. Exploring the spatio-temporal interrelation between groundwater and surface water by using the self-organizing maps. J. Hydrol. 556, 131–142.
Chen, X., Yuan, Y., Zeng, G., Wang, J., 2021. Semi-supervised semantic segmentation with cross pseudo supervision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2613–2622.
Chen, Y., Mancini, M., Zhu, X., Akata, Z., 2022. Semi-supervised and unsupervised deep visual learning: A survey. IEEE Trans. Pattern Anal. Mach. Intell.
Coxon, G., Addor, N., Bloomfield, J., Freer, J., Fry, M., Hannaford, J., Howden, N., Lane, R., Lewis, M., Robinson, E., others, 2020. Catchment attributes and hydro-meteorological timeseries for 671 catchments across Great Britain (CAMELS-GB).
Dai, W., Li, X., Cheng, K.-T., 2023. Semi-Supervised Deep Regression with Uncertainty Consistency and Variational Model Ensembling via Bayesian Neural Networks. ArXiv Prepr. ArXiv230207579.
Do, H.X., Westra, S., Leonard, M., 2017. A global-scale investigation of trends in annual maximum streamflow. J. Hydrol. 552, 28–43.
Du, F., Zhu, A.-X., Liu, J., Yang, L., 2020. Predictive mapping with small field sample data using semi-supervised machine learning. Trans. GIS 24, 315–331.
Feng, D., Fang, K., Shen, C., 2020. Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. Water Resour. Res. 56, e2019WR026793.
Frame, J.M., Kratzert, F., Raney, A., Rahman, M., Salas, F.R., Nearing, G.S., 2021. Post-processing the national water model with long short-term memory networks for streamflow predictions and model diagnostics. JAWRA J. Am. Water Resour. Assoc. 57, 885–905.
Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., Hochreiter, S., 2021a. Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network. Hydrol. Earth Syst. Sci. 25, 2045–2062.
Gauch, M., Mai, J., Lin, J., 2021b. The proper care and feeding of CAMELS: How limited training data affects streamflow prediction. Environ. Model. Softw. 135, 104926.

George, D., Shen, H., Huerta, E., 2017. Deep Transfer Learning: A new deep learning glitch classification method for advanced LIGO. ArXiv Prepr. ArXiv170607446.
Han, H., Choi, C., Jung, J., Kim, H.S., 2021. Deep learning with long short term memory based sequence-to-sequence model for rainfall-runoff simulation. Water 13, 437.
He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738.
Hendrycks, D., Gimpel, K., 2016. Gaussian error linear units (gelus). ArXiv Prepr. ArXiv160608415.
Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. ArXiv Prepr. ArXiv150302531.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.
Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G.S., Hochreiter, S., Klambauer, G., 2021. Mc-lstm: Mass-conserving lstm, in: International Conference on Machine Learning. PMLR, pp. 4275–4286.
Hunt, K.M., Matthews, G.R., Pappenberger, F., Prudhomme, C., 2022. Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States. Hydrol. Earth Syst. Sci. 26, 5449–5472.
Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. 30.
Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., Nearing, G., 2022. Uncertainty estimation with deep learning for rainfall–runoff modeling. Hydrol. Earth Syst. Sci. 26, 1673–1693.
Kothari, M., Gharde, K., 2015. Application of ANN and fuzzy logic algorithms for streamflow modelling of Savitri catchment. J. Earth Syst. Sci. 124, 933–943.
Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall–runoff modelling using long short-term memory (LSTM) networks. Hydrol. Earth Syst. Sci. 22, 6005–6022.
Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., Hochreiter, S., Nearing, G.S., 2019a. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. Water Resour. Res. 55, 11344–11354.
Kratzert, F., Klotz, D., Hochreiter, S., Nearing, G.S., 2021. A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling. Hydrol. Earth Syst. Sci. 25, 2685–2703.
Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., Nearing, G., 2019b. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrol. Earth Syst. Sci. 23, 5089–5110.
Lee, D.-G., Ahn, K.-H., 2022. Assessment of Suitable Gridded Climate Datasets for Large-Scale Hydrological Modelling over South Korea. Remote Sens. 14, 3535.
Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., Dadson, S.J., 2021. Benchmarking Data-Driven Rainfall-Runoff Models in Great Britain: A comparison of LSTM-based models with four lumped conceptual models. Hydrol. Earth Syst. Sci.
Leisher, C., Makau, J., Kihara, F., Kariuki, A., Sowles, J., Courtemanch, D., Njugi, G., Apse, C., 2016. Upper Tana-Nairobi Water Fund Monitoring and Evaluation Plan. Nat. Conserv. IFAD CIAT GEF TRICOKEN.
Levatić, J., Ceci, M., Kocev, D., Džeroski, S., 2017. Self-training for multi-target regression with tree ensembles. Knowl.-Based Syst. 123, 41–60.

Ley, A., Bormann, H., Casper, M., 2023. Intercomparing LSTM and RNN to a Conceptual Hydrological Model for a Low-Land River with a Focus on the Flow Duration Curve. Water 15, 505.
Li, D., Marshall, L., Liang, Z., Sharma, A., Zhou, Y., 2021. Bayesian LSTM with stochastic variational inference for estimating model uncertainty in process-based hydrological models. Water Resour. Res. 57, e2021WR029772.
Liang, C., Li, H., Lei, M., Du, Q., 2018. Dongting lake water level forecast and its relationship with the three gorges dam based on a long short-term memory network. Water 10, 1389.
Liang, X., Lettenmaier, D.P., Wood, E.F., Burges, S.J., 1994. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. J. Geophys. Res. Atmospheres 99, 14415–14428.
Ma, K., Feng, D., Lawson, K., Tsai, W.-P., Liang, C., Huang, X., Sharma, A., Shen, C., 2021. Transferring hydrologic data across continents–leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. Water Resour. Res. 57, e2020WR028600.
Mckane, R., Brookes, A., Djang, K., Stieglitz, M., Abdelnour, A., Halama, J., Pettus, P., Phillips, D., 2014. VELMA Version 2.0 User Manual and Technical Documentation. Corvallis Or. Httpswww Epa Govsitesproductionfiles2016-01documentsvelma2 0usermanual Pdflast Accessed 1006 17.
Muleta, M.K., 2011. Model performance sensitivity to objective function during automated calibrations. J. Hydrol. Eng. 17, 756–767.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. J. Hydrol. 10, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6
Nearing, G.S., Kratzert, F., Sampson, A.K., Pelissier, C.S., Klotz, D., Frame, J.M., Prieto, C., Gupta, H.V., 2021. What role does hydrological science play in the age of machine learning? Water Resour. Res. 57, e2020WR028091.
Newman, A., Clark, M., Sampson, K., Wood, A., Hay, L., Bock, A., Viger, R., Blodgett, D., Brekke, L., Arnold, J., others, 2015. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. Hydrol. Earth Syst. Sci. 19, 209–223.
Nourani, V., Komasi, M., Mano, A., 2009. A multivariate ANN-wavelet approach for rainfall–runoff modeling. Water Resour. Manag. 23, 2877–2894.
Oruche, R., Egede, L., Baker, T., O'Donncha, F., 2021. Transfer learning to improve streamflow forecasts in data sparse regions. ArXiv Prepr. ArXiv211203088.
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204.
Roy, B., Singh, M.P., Kaloop, M.R., Kumar, D., Hu, J.-W., Kumar, R., Hwang, W.-S., 2021. Data-Driven Approach for Rainfall-Runoff Modelling Using Equilibrium Optimizer Coupled Extreme Learning Machine and Deep Neural Network. Appl. Sci. 11, 6238.
Seibert, J., Vis, M.J., 2012. Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. Hydrol. Earth Syst. Sci. 16, 3315–3325.
Shen, C., 2018. A transdisciplinary review of deep learning research and its relevance for water resources scientists. Water Resour. Res. 54, 8558–8593.
Shen, C., Chen, X., Laloy, E., 2021. Broadening the use of machine learning in hydrology. Front. Water.

Sitterson, J., Knightes, C., Parmar, R., Wolfe, K., Avant, B., Muche, M., 2018. An overview of rainfall-runoff model types.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.
Taormina, R., Chau, K.-W., 2015. Data-driven input variable selection for rainfall–runoff modeling using binary-coded particle swarm optimization and Extreme Learning Machines. J. Hydrol. 529, 1617–1632.
Thrun, S., Pratt, L., 1998. Learning to learn: Introduction and overview. Learn. Learn 3–17.
TNC, 2015. Upper Tana-Nairobi water fund business case.
Torrey, L., Shavlik, J., 2010. Transfer learning, in: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques. IGI global, pp. 242–264.
van de Giesen, N., Hut, R., Selker, J., 2014. The trans-African hydro-meteorological observatory (TAHMO). Wiley Interdiscip. Rev. Water 1, 341–348.
Van, S.P., Le, H.M., Thanh, D.V., Dang, T.D., Loc, H.H., Anh, D.T., 2020. Deep learning convolutional neural network in rainfall–runoff modelling. J. Hydroinformatics 22, 541–561.
Voropay, N., Ryazanova, A., Dyukarev, E., 2021. High-resolution bias-corrected precipitation data over South Siberia, Russia. Atmospheric Res. 254, 105528.
Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., Dupoux, E., 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. ArXiv Prepr. ArXiv210100390.
Wi, S., Steinschneider, S., 2022. Assessing the physical realism of deep learning hydrologic model projections under climate change. Water Resour. Res. 58, e2022WR032123.
Xiang, Z., Yan, J., Demir, I., 2020. A rainfall-runoff model with LSTM-based sequence-to-sequence learning. Water Resour. Res. 56, e2019WR025326.
Xie, K., Liu, P., Zhang, J., Han, D., Wang, G., Shen, C., 2021. Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships. J. Hydrol. 603, 127043.
Xie, W., Yi, S., Leng, C., Xia, D., Li, M., Zhong, Z., Ye, J., 2022. The evaluation of IMERG and ERA5-Land daily precipitation over China with considering the influence of gauge data bias. Sci. Rep. 12, 8085.
Xu, Y., Hu, C., Wu, Q., Jian, S., Li, Z., Chen, Y., Zhang, G., Zhang, Z., Wang, S., 2022. Research on particle swarm optimization in LSTM neural networks for rainfall-runoff simulation. J. Hydrol. 608, 127553.
Yang, T.-Y., Chen, Y.-T., Lin, Y.-Y., Chuang, Y.-Y., 2019. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1087–1096.
Yarowsky, D., 1995. Unsupervised word sense disambiguation rivaling supervised methods, in: 33rd Annual Meeting of the Association for Computational Linguistics. pp. 189–196.
Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? Adv. Neural Inf. Process. Syst. 27.
Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., Ma, K., 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation, in: Proceedings

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

Hydrology and Earth System Sciences
Discussions

of the IEEE/CVF International Conference on Computer Vision. pp. 3713–3722.
Zhou, Z.-H., Zhou, Z.-H., 2021. Semi-supervised learning. Mach. Learn. 315–341.
Zhu, Y., Shareghi, E., Li, Y., Reichart, R., Korhonen, A., 2021. Combining Deep Generative Models and Multi-lingual Pretraining for Semi-supervised Document Classification. ArXiv Prepr. ArXiv210110717.

List of Figures

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

Hydrology and Earth System Sciences
Open Access
EGU
Discussions

period and the improved NSEs between the student models and baseline models during the evaluation period under two model setting: (a) idv-LSTM and (b) rgn-LSTM with the moderate density network, in multi-year training scenarios. Additionally, the expected improved NSEs corresponding to the NSE of the teacher model are depicted, along with a Lowess fit represented by a red line.

List of Tables

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

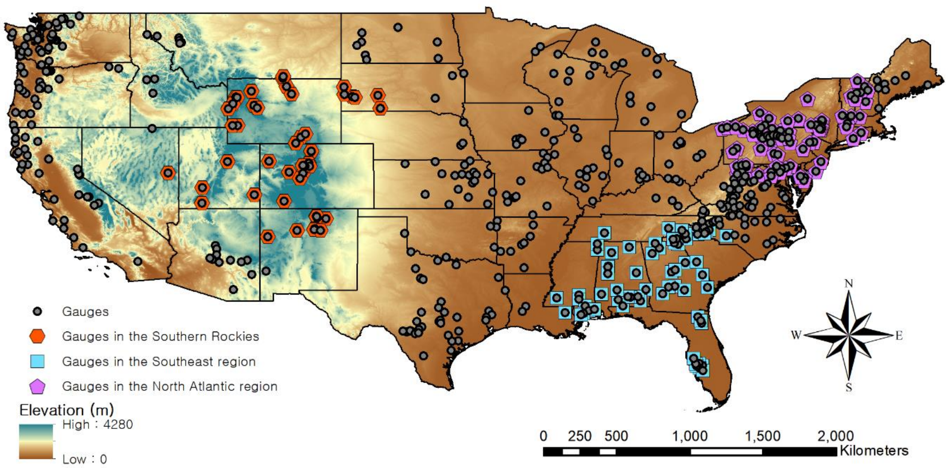

Figure 1 Locations of the selected 531 gauges in the United States. The gauge stations across three homogenous regions are also illustrated.

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

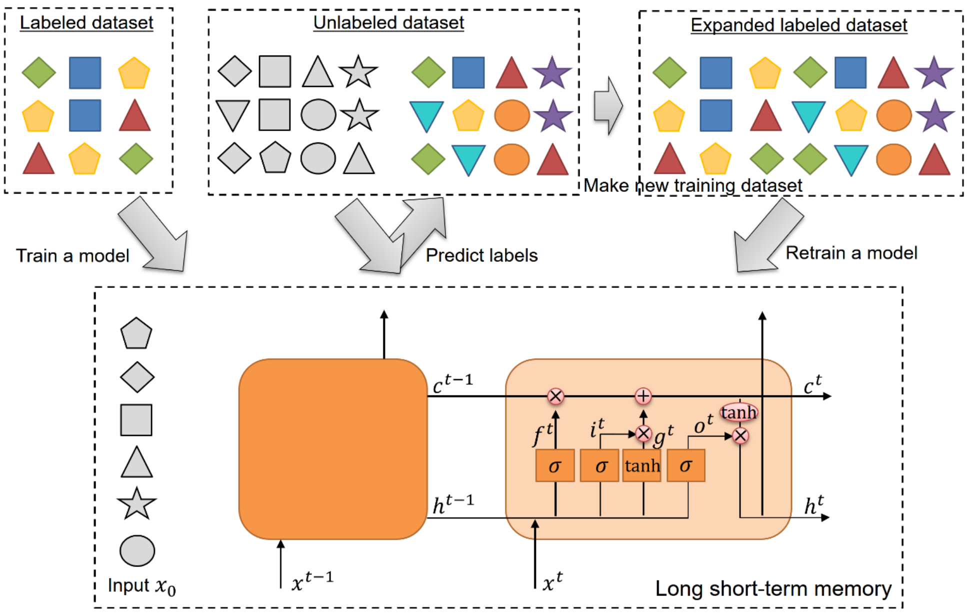

Figure 2 Overview of the semi-supervised learning-based framework proposed in this study. Here, the colored geometric shapes represent the recorded samples, pairing climate forcing data with streamflow.

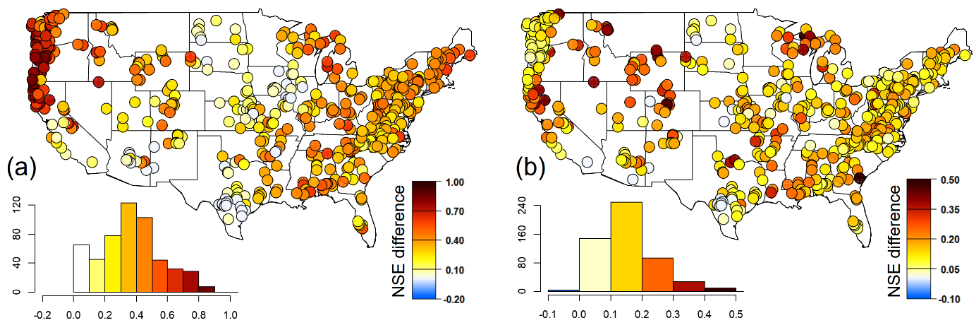

Figure 3 Difference of NSE results of idv-LSTM compared to their baseline models for (a) single and (b) multi-year training scenarios. The color maps are limited to [-0.2, 1.0] for the single training scenario and [-0.1, 0.5] for the multi-year training scenario for enhanced visualization.

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

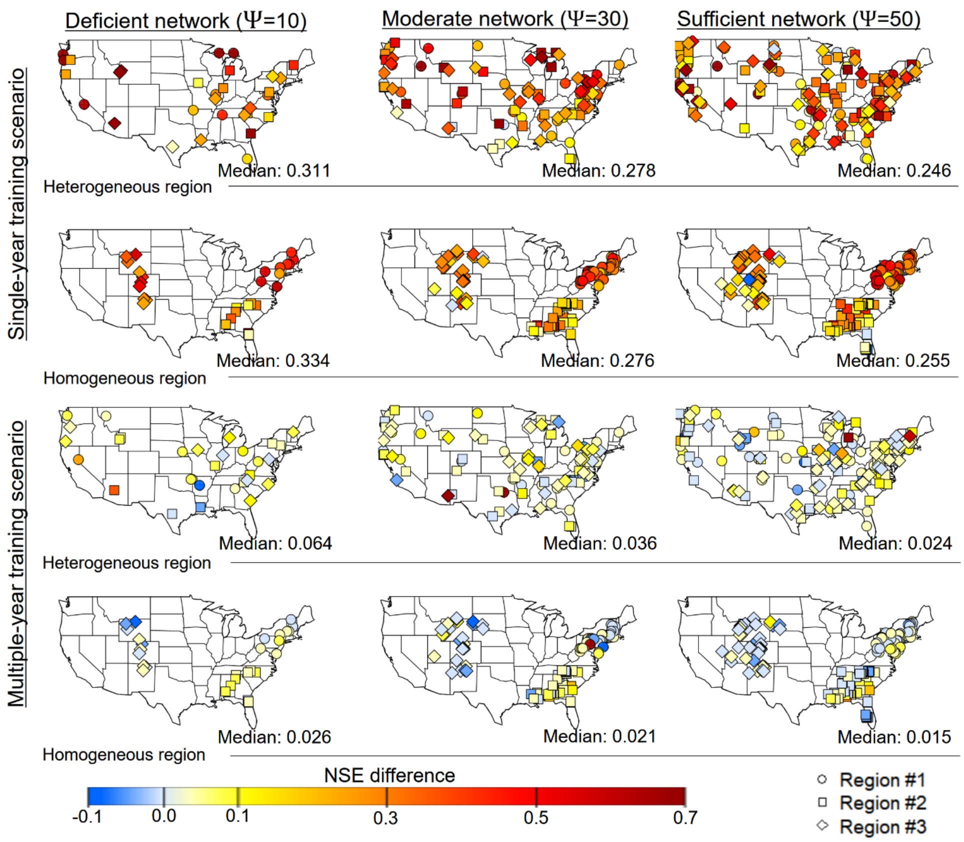

Figure 4 Difference of NSE results of rgn-LSTM compared to their baseline models across experimental factors including three defined regions, two training scenarios, and three basin densities in network. Here, the median NSE differences across basins in three defined regions are presented in each plot.

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

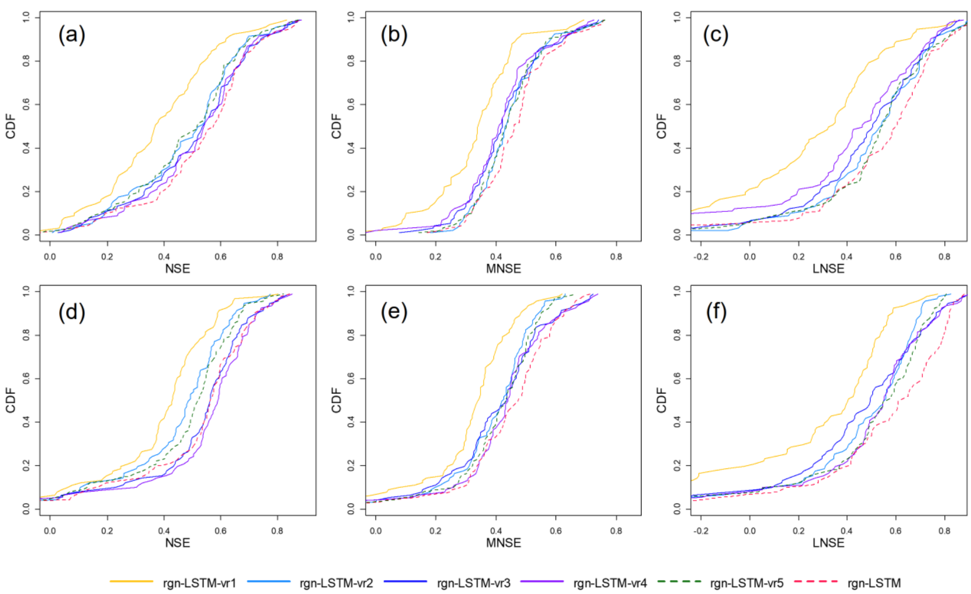

Figure 5 Cumulative density functions of the results of the annealing process on rgn-LSTM with the single-year training scenarios obtained for basins across (a), (b), (c) heterogeneous region; and (d), (e), (f) homogeneous region. Here, three metrics, namely NSE (first column), MNSE (second column), and LNSE (last column), are utilized.

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

Hydrology and Earth System Sciences
Open Access
Discussions

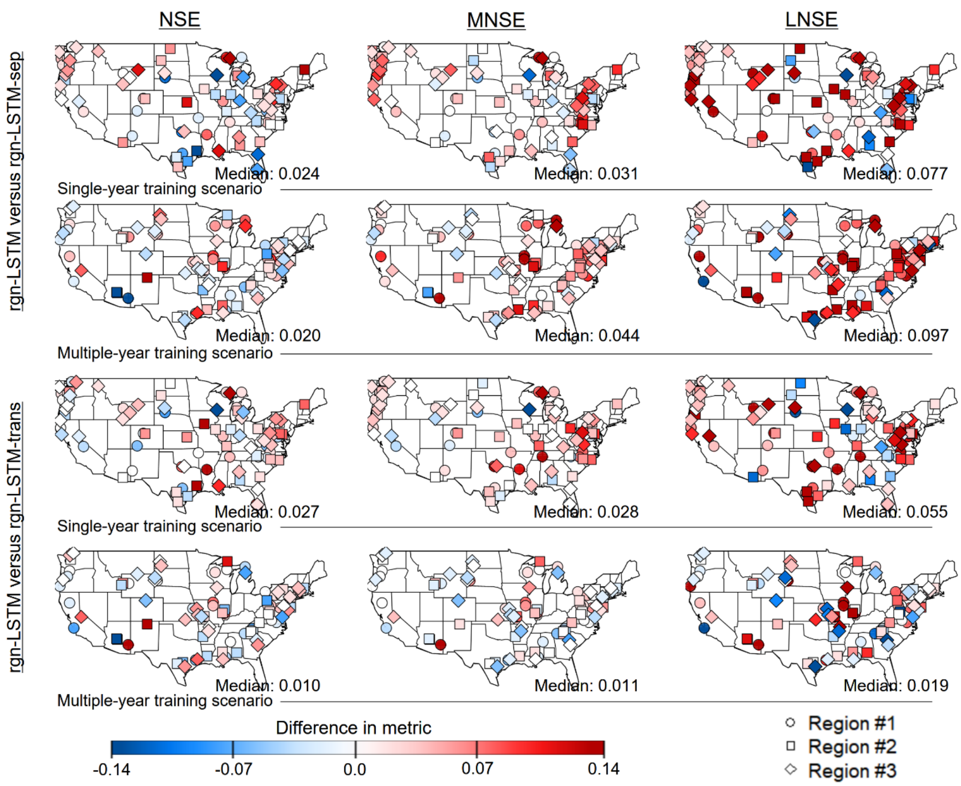

Figure 6 Difference of performance in the three metrics, NSE (first column), MNSE (second column), and LNSE (third column), of rgn-LSTM compared to the two fine-tuning approaches (rgn-LSTM-sep and rgn-LSTM-trans) across three basin networks in heterogeneous regions. Here, the median NSE differences across basins in three defined regions are presented in each plot.

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

Hydrology and Earth System Sciences
Open Access
Discussions

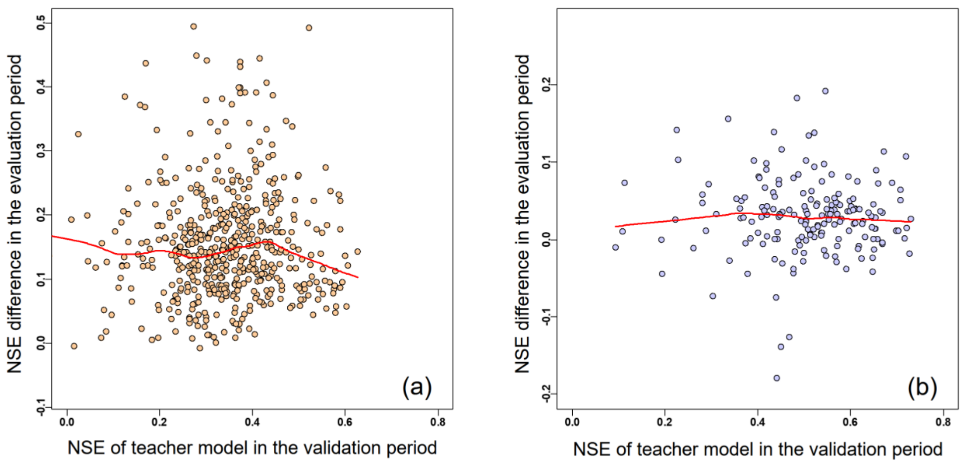

Figure 7 Plots comparing the NSEs achieved by the teacher model during the validation period and the improved NSEs between the student models and baseline models during the evaluation period under two model setting: (a) idv-LSTM and (b) rgn-LSTM with the moderate density network, in multi-year training scenarios. Additionally, the expected improved NSEs corresponding to the NSE of the teacher model are depicted, along with a Lowess fit represented by a red line.

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

Table 1 List of the climate and basin attributes used in this study.

	Variable name	Description	Unit
Climate forcing	PRCP	Precipitation	mm
	T_{max}	Maximum air temperature	℃
	T_{min}	Minimum air temperature	℃
	SRAD	Short-wave radiation	W/m^2
	VP	Vapor pressure	Pa
Basin attributes	p_mean	Catchment mean daily precipitation	mm
	pet_mean	Catchment mean daily potential evapotranspiration	mm
	p_seasonality	Seasonality and timing of precipitation	-
	frac_snow	Fraction of precipitation falling as snow	-
	aridity	Ratio of catchment mean PET to mean precipitation	-
	high_prec_freq	Frequency of high precipitation days (≥ 5 × p_mean)	Days
	high_prec_dur	Average duration of high precipitation events (number of consecutive days ≥ 5 × p_mean)	Days
	low_prec_freq	Frequency of dry days (< 1 mm/day)	Days
	low_prec_dur	Average duration of dry periods (number of consecutive days <1 mm/day)	Days
	soil_depth_pelletier	Depth to bedrock (maximum 50m)	m
	soil_depth_statsgo	Soil depth (maximum 1.5m)	m
	soil_porosity	Volumetric porosity	-
	soil_conductivity	Saturated hydraulic conductivity	cm/hr
	max_water_content	Maximum water content of the soil	m
	sand_frac	Fraction of sand	%
	silt_frac	Fraction of silt	%
	clay_frac	Fraction of clay	%
	carbonate_rocks_frac	Fraction of the catchment area characterized as "Carbonate sedimentary rocks"	%
	geol_permeability	Subsurface permeability (log10)	-
	elev_mean	Catchment mean elevation	m
	slope_mean	Catchment mean slope	m/km
	area_gauges	Catchment area	km^2
	frac_forest	Forest fraction	%
	lai_max	Maximum monthly mean of the leaf area index	-
	lai_diff	Difference between the maximum and mimumum monthly mean of	-

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

		the leaf area index	
	gvf_max	Maximum monthly mean of the green vegetation fraction	-
	gvf_diff	Difference between the maximum and mimumum monthly mean of the green vegetation fraction	-

Table 2 Median performance of the models trained under our proposed framework (baseline models in the parenthesis) for two training scenarios.

Model	No. of basins	Training scenario	NSE_{median}		$MNSE_{median}$		$LNSE_{median}$	
			Heterogeneous region	Homogeneous region	Heterogeneous region	Homogeneous region	Heterogeneous region	Homogeneous region
idv-LSTM	531 basins	Single-year training	0.392 (0.002)		0.314 (0.076)		0.047 (-0.193)	
		Multi-year training	0.576 (0.411)		0.441 (0.301)		0.418 (0.044)	
rgn-LSTM	30 basins in 3 experimental trials	Single-year training	0.521 (0.175)	0.531 (0.145)	0.452 (0.206)	0.421 (0.176)	0.491 (-0.634)	0.624 (0.070)
		Multi-year training	0.619 (0.563)	0.650 (0.627)	0.482 (0.446)	0.536 (0.481)	0.595 (0.461)	0.700 (0.620)
rgn-LSTM	90 basins in 3 experimental trials	Single-year training	0.570 (0.213)	0.538 (0.241)	0.483 (0.259)	0.463 (0.226)	0.580 (-0.312)	0.609 (0.022)
		Multi-year training	0.667 (0.608)	0.689 (0.673)	0.549 (0.492)	0.553 (0.515)	0.666 (0.577)	0.755 (0.672)
rgn-LSTM	150 basins in 3	Single-year training	0.567 (0.263)	0.542 (0.246)	0.506 (0.241)	0.467 (0.230)	0.631 (0.166)	0.636 (0.135)

https://doi.org/10.5194/hess-2023-148
Preprint. Discussion started: 18 July 2023

experimental trials	Multi-year training	0.675 (0.638)	0.704 (0.676)	0.555 (0.532)	0.579 (0.543)	0.691 (0.645)	0.750 (0.668)