Supplementary material for

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

Introduction

To support the results and conclusions of the study titled "**Semi-supervised Learning Approach to Enhance Predictability in Data-Driven Rainfall-Runoff Models for Hydrologically Data-Scarce Regions**", this file consists of two texts, two tables, and six figures. These elements are specifically utilized in the designated section to reinforce the presented findings:

Text S1
- 2.4.2 The effect of the annealing process on the student model

Text S2

- 2.4.3 Comparison of our proposed framework to the fine-tuning approaches

Tables S1-S2
- 4.1 AR climatology and corresponding water availability over South Korea

Figure S1
- 2.4.2 The effect of the annealing process on the student model

Figures S2- S4
- 3.1 Evaluating semi-supervised learning in data-scarce regions

Figure S5
- 3.2 Evaluating the selection of the annealing process on the student model

Figure S6
- 3.3 Comparing our proposed model with rgn-LSTM-sep and rgn-LSTM-trans

Text S1.

This section offers additional details about the five structures used in $\alpha(\mathbb{t})$ for the comparison purpose. The provided annealing process (i.e., Eq. 12) in the main manuscript is used in our final framework. In addition to this, alternative formulations are developed as variant versions. Each subsequent formulation is applied to one of the five models (rgn-LSTM-vr1, rgn-LSTM-vr2, rgn-LSTM-vr3, rgn-LSTM-vr4, and rgn-LSTM-vr5), respectively.

$$\alpha(\mathbb{t}) = \alpha_0 \qquad \forall \mathbb{t} \qquad \text{Eq. (S1)}$$

$$\alpha(\mathbb{t}) = \begin{cases} 0 & \mathbb{t} \le \mathcal{J} \\ \alpha_0 & \mathcal{J} < \mathbb{t} \end{cases} \qquad \text{Eq. (S2)}$$

$$\alpha(\mathbb{t}) = \begin{cases} 0 & \mathbb{t} \le \mathcal{J}' \\ \frac{\mathbb{t}-\mathcal{J}'}{\mathcal{J}''-\mathcal{J}'}\alpha_0 & \mathcal{J}' < \mathbb{t} \le \mathcal{J}'' \\ \alpha_0 & \mathcal{J}'' < \mathbb{t} \end{cases} \qquad \text{Eq. (S3)}$$

$$\alpha(\mathbb{t}) = \begin{cases} \alpha_0 & \mathbb{t} \le \mathcal{J}' \\ (1 - \frac{\mathbb{t} - \mathcal{J}'}{\mathcal{J}'' - \mathcal{J}'})\alpha_0 & \mathcal{J}' < \mathbb{t} \le \mathcal{J}'' \\ 0 & \mathcal{J}'' < \mathbb{t} \end{cases} \quad \text{Eq. (S4)}$$

$$\alpha(\mathbb{t}) = \begin{cases} 2 \times \alpha_0 & \mathbb{t} \le \mathcal{J} \\ 0 & \mathcal{J} < \mathbb{t} \end{cases} \quad \text{Eq. (S5)}$$

where we utilize α_0 = 1, $\mathcal{J} = 15$, $\mathcal{J}'$ = 5, and $\mathcal{J}''$ = 25 based on the epoch adopted in this study.

The first model, rgn-LSTM-vr1, incorporates a $\mathbb{t}$-invariant structure where labeled and unlabeled data are equally treated. The second model, rgn-LSTM-vr2, highlights the significance of unlabeled data specifically during the latter half of the epoch progression. For the third and fourth models, rgn-LSTM-vr3 and rgn-LSTM-vr4, they employ $\mathbb{t}$-varying structures to gradually amplify or reduce the influence of unlabeled data. Lastly, the final model, rgn-LSTM-vr5, maintains an identical structure to our proposed model while placing further emphasis on the role of unlabeled data in the initial phase of the epoch progression. Additionally, Figure S1 provides a visualization of how each of the $\alpha(\mathbb{t})$ formulations evolves throughout the epoch progression.

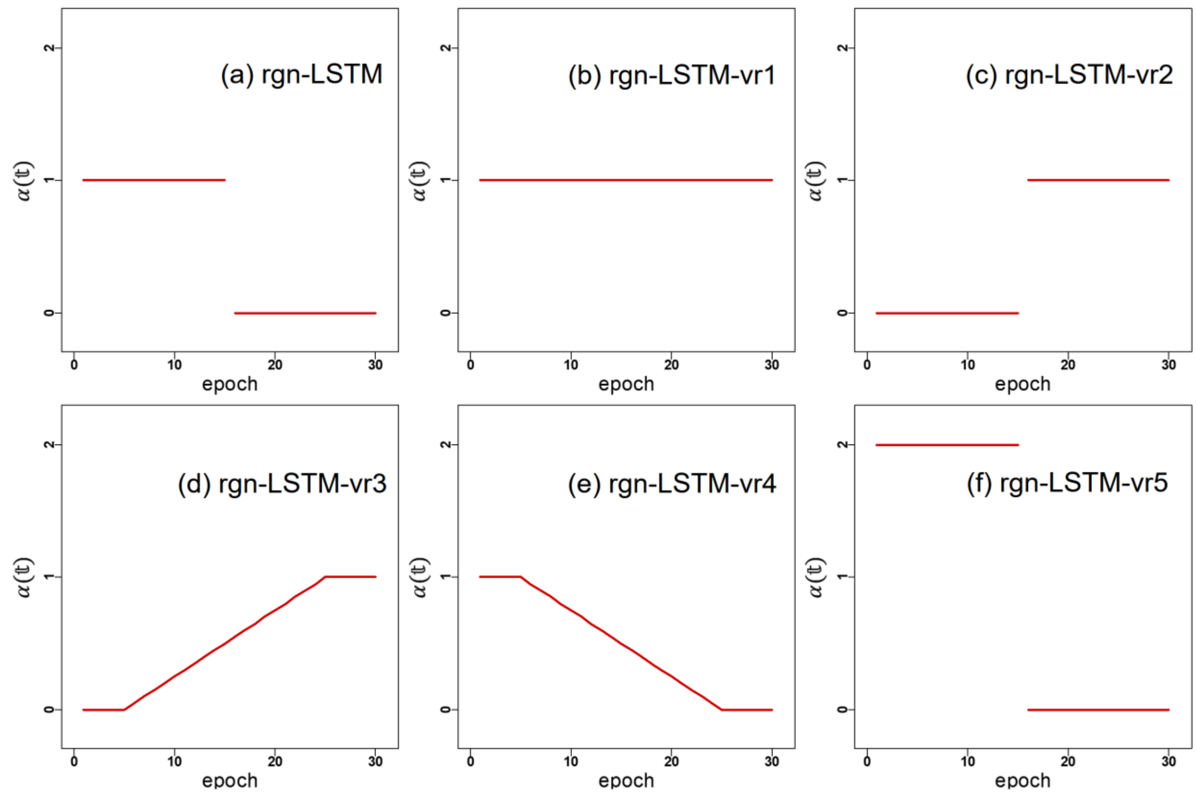

Figure S1. Visualization of how each of the $\alpha(t)$ formulations evolve corresponding to the considered six models (denoted in each plot).

Text S2.

In this section, we present information about the recent technique developed by Ma et al. (2021), which we considered for comparison in our work. Given the similarity in responses required for rainfall-runoff modeling, there is a possibility that their representations in a data-driven model could exhibit similarities. Consequently, training a model with one regional dataset and transferring it to another region becomes possible. To achieve this, Ma et al. (2021) pretrained their models on a data-rich region and then transferred them to data-scarce regions as initial conditions. Following their approach, we also conducted tests using three different combinations (TL-a, TL-b, and TL-c) of transfer learning by controlling weight initialization and freezing. However, we only present the results of TL-c, as it outperformed the other tested models in our analysis (not shown). Our decision to use TL-c aligns with the findings of Ma et al. (2021), who also concluded it to be one of the best options. For our analysis, the regional

LSTM model was pretrained using 44 forcing variables across the 631 basins from the CAMELS-GB dataset (see Table S1). For the pertaining model, we developed the model using a 10-year dataset spanning from October 1, 1980, to September 30, 1990 (as the training period). Additionally, we validated the model's performance using a separate 3-year dataset covering the period from October 1, 1990, to September 30, 1993.

Table S1 List of the climate and basin attributes from CAMELS-GB dataset.

	Variable name	Description	Unit
Climate forcing	precipitation	Catchment daily averaged precipitation	mm
	temperature	Catchment daily averaged temperature	℃
	humidity	Catchment daily averaged specific humidity	℃
	shortwave_rad	Catchment daily averaged downward shortwave radiation	W/m^2
	longwave_rad	Catchment daily averaged long-wave radiation	W/m^2
	windspeed	Catchment daily averaged wind speed	m/s
Basin attributes	p_mean	Catchment mean daily precipitation	mm
	pet_mean	Catchment mean daily potential evapotranspiration	mm
	aridity	Ratio of catchment mean PET to mean precipitation	-
	p_seasonality	Seasonality and timing of precipitation	-
	inter_high_perc	Significant intergranular flow – high productivity	%
	q_mean	Mean daily discharges	mm
	runoff_ratio	Ratio of mean daily discharge to mean daily precipitation	-
	stream_elas	Streamflow precipitation elasticity	-
	baseflow_index	Ratio of mean daily base flow to daily discharge	-
	Q5	5% flow quantile	mm
	Q95	95% flow quantile	mm
	dwood_perc	percentage cover of deciduous woodland	%
	ewood_perc	percentage cover of evergreen woodland	%
	grass_perc	percentage cover of grass and pasture	%
	shrub_perc	percentage cover of medium-scale vegetation	%
	crop_perc	percentage cover of crops	%
	urban_perc	percentage cover of suburban and urban	%
	inwater_perc	percentage cover of inland water	%
	bares_perc	percentage cover of bare soil and rocks	%
	sand_perc	percentage sand	%

silt_perc	percentage silt	%
clay_perc	percentage clay	%
organic_perc	percentage organic content	%
bulkdens	bulk density	g/cm^3
tawc	total available water content	mm
porosity_cosby	saturated water content estimated using a pedo-transfer function based on sand and clay fractions	-
porosity_hypres	saturated water content estimated using a pedo-transfer function based on silt, clay and organic fractions, bulk density, and topsoil	-
conductivity_cosby	estimated using a pedo-transfer function based on sand and clay fractions	cm/h
conductivity_hypres	estimated using a pedo-transfer function based on sand and clay fractions	cm/h
root_depth	depth available for roots	m
soil_depth_pelletier	depth to bedrock	m
gauge_lat	gauge latitude	degree
gauge_lon	gauge longitude	degree
gauge_elev	gauge elevation	ma.s.l.
area	catchmentarea	km^2
dpsbar	catchment mean drainage path slope	m/km
elev_mean	catchment mean elevation	ma.s.l.
elev_min	catchment minimum elevation	ma.s.l.

To ensure the accurate reproduction of the results reported in Ma et al. (2021), we implemented their regional LSTM model using the CAMELS-GB dataset. For the training scenarios, we utilized 666 basins for the 1-year scenario and 668 basins for the 5-year scenario, following the train and test evaluation scheme outlined by Ma et al. (2021). Specifically, in the 1-year (5-year) training scenario, the models were trained from January 1, 2004, to January 1, 2005 (January 1, 2000, to January 1, 2005), and subsequently tested from January 1, 2005, to January 1, 2010 (January 1, 2005, to January 1, 2010). It is worth noting that the basin selection in our study differs slightly from that of Ma et al. (2021), who employed 667 basins in both scenarios. However, a significant majority of the basins overlap, and the performance statistics for the test phase in our study (see Table S2) exhibit similarities with the results reported in Ma et al. (2021)

(their Table S3). Based on these outcomes, we employed a regional LSTM model applied to basins across England for our comparative analysis.

Table S2 Validating the results in Ma et al. (2021) by developing the regional LSTM models.

Utilized data	Temporal scenario	NSE_{mean}	Ensemble NSE_{mean}
CAMELS-GB	1-year training	0.728	0.706
	5-year training	0.830	0.804

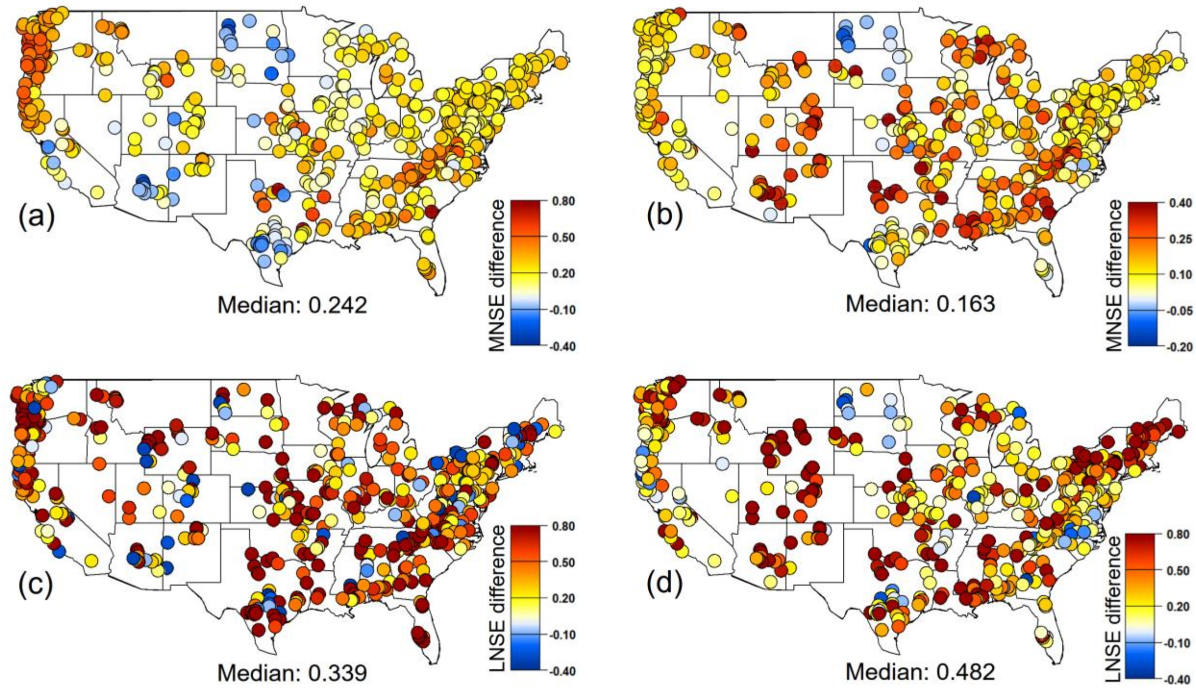

Figure S2. Difference of (a and b) MNSE and (c and d) LMSE results of idv-LSTM compared to their baseline models for (left column) single and (right column) multi-year training scenarios. The color maps are limited for enhanced visualization (see each subplot).

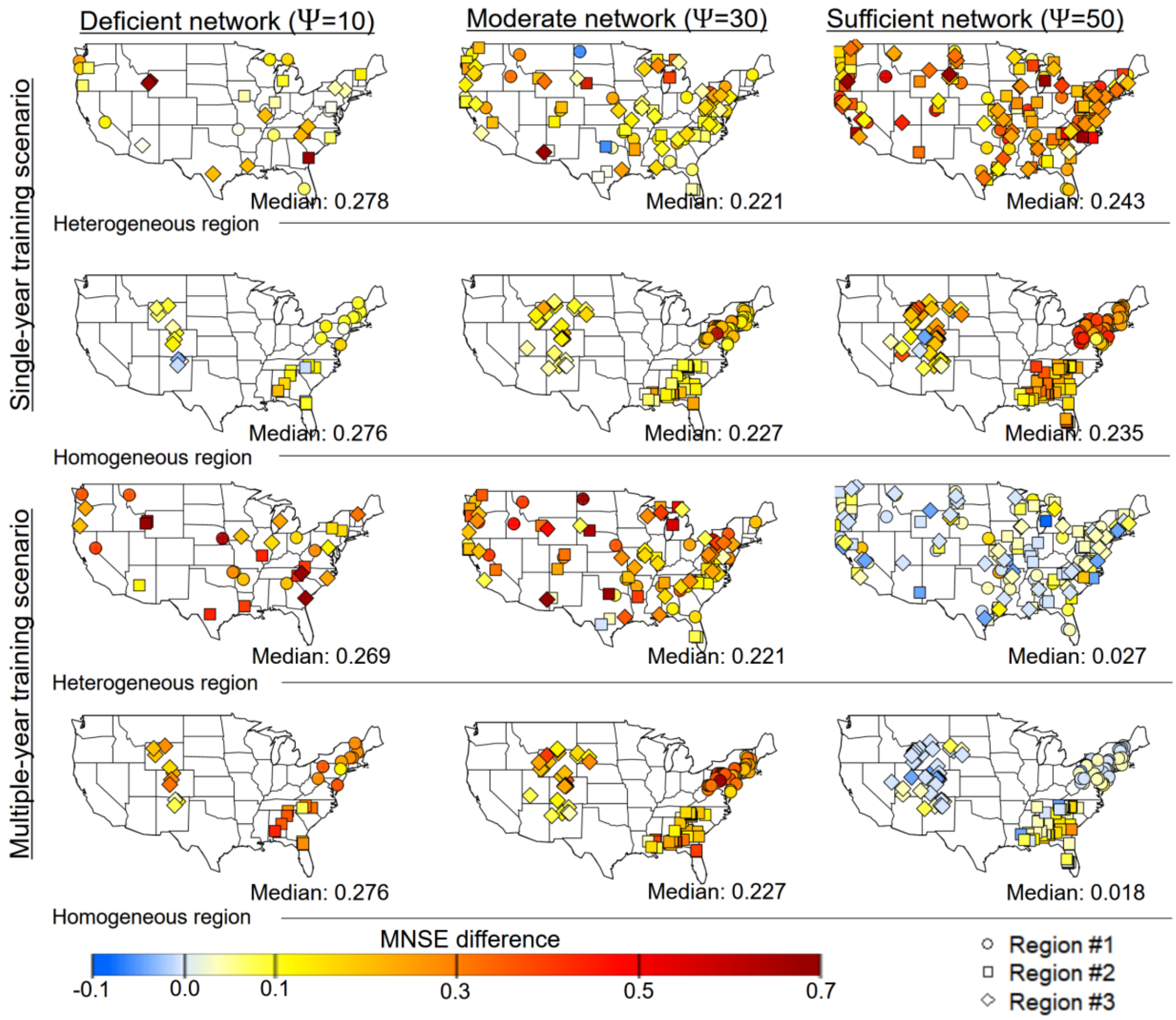

Figure S3 Difference of MNSE results of rgn-LSTM compared to their baseline models across experimental factors including three defined regions, two training scenarios, and three basin densities in network. Here, the median MNSE differences across basins in three defined regions are presented in each plot.

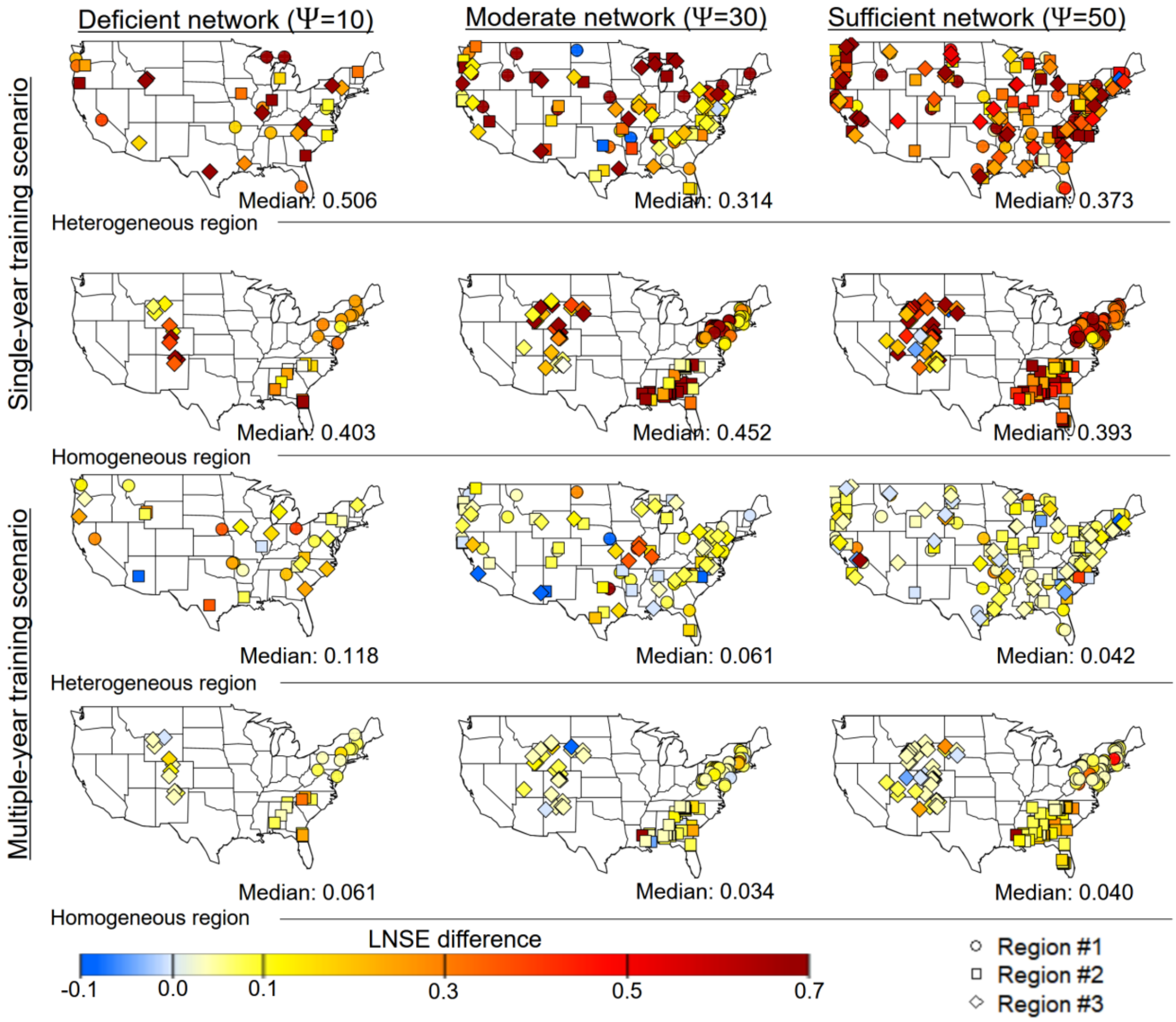

Figure S4 Difference of LNSE results of rgn-LSTM compared to their baseline models across experimental factors including three defined regions, two training scenarios, and three basin densities in network. Here, the median LNSE differences across basins in three defined regions are presented in each plot.

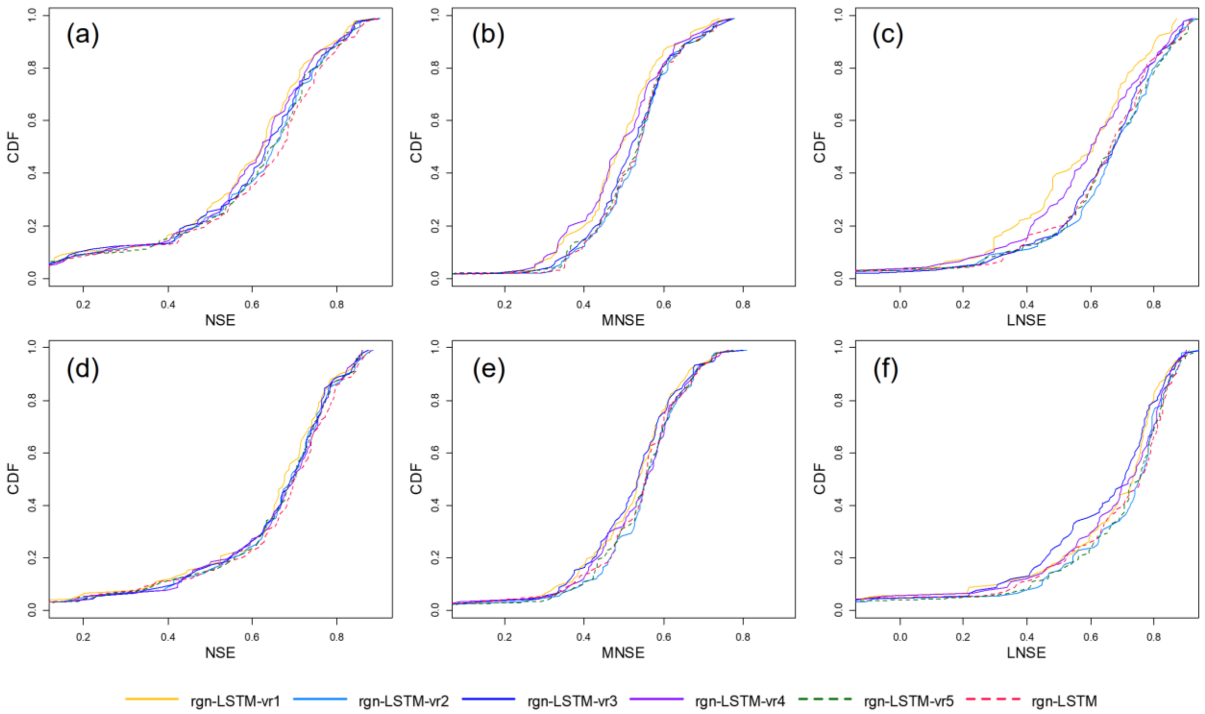

Figure S5 Cumulative density functions of the results of the annealing process on rgn-LSTM with the multi-year training scenarios obtained for basins across (a), (b), (c) heterogeneous region; and (d), (e), (f) homogeneous region. Here, three metrics, namely NSE (first column), MNSE (second column), and LNSE (last column), are utilized.

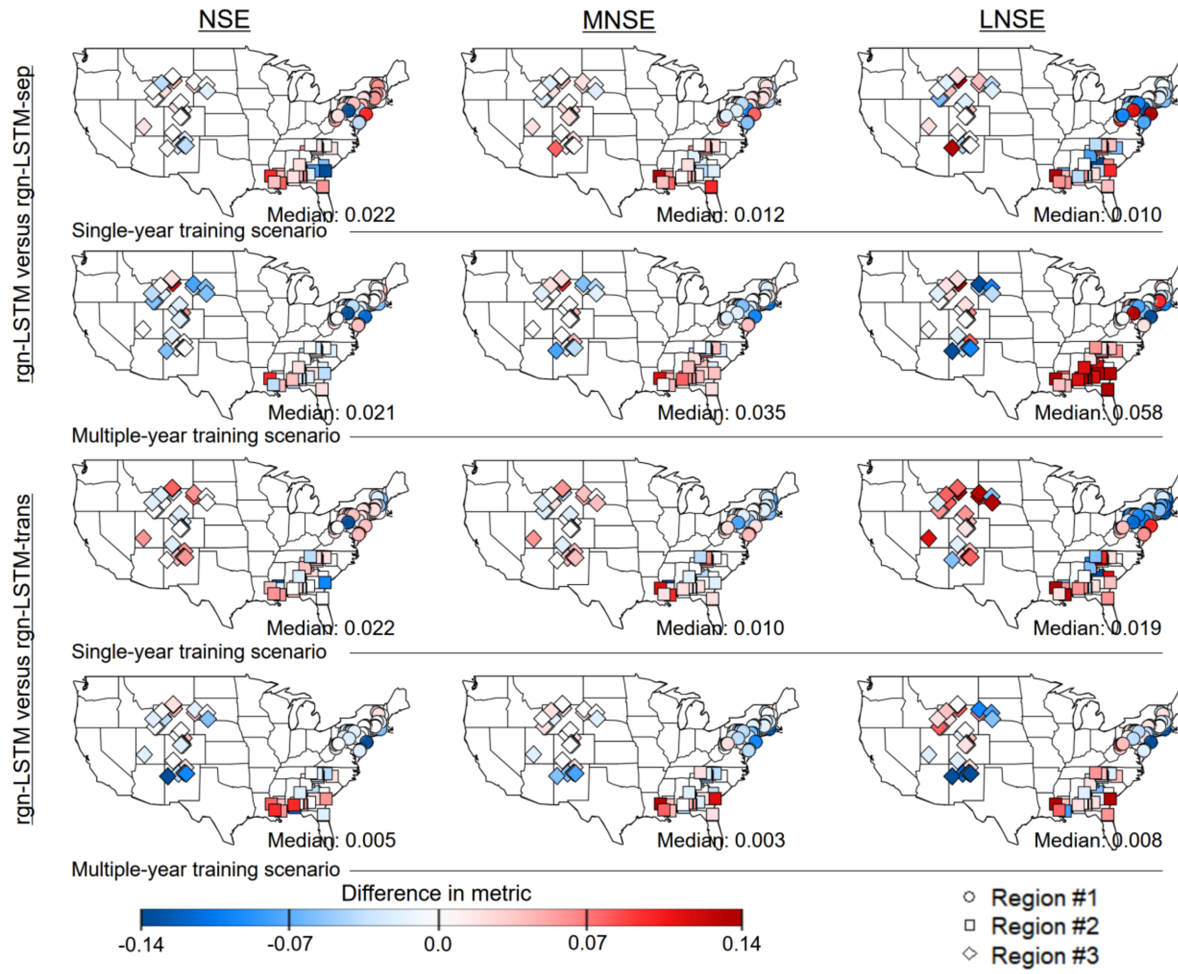

Figure S6 Difference of performance in the three metrics, NSE (first column), MNSE (second column), and LNSE (third column), of rgn-LSTM compared to the two fine-tuning approaches (rgn-LSTM-sep and rgn-LSTM-trans) across three basin networks in homogeneous regions. Here, the median NSE differences across basins in three defined regions are presented in each plot.