Empirical Evidence of the Importance of Data Recency in LSTM-Based Rainfall-Runoff Modeling

Qiutong Yu, Bryan A. Tolson
Department of Civil and Environmental Engineering, University of Waterloo, Canada

Motivation

Deep learning (DL)-based hydrological models, particularly those using Long Short-Term Memory (LSTM) networks, typically require large datasets for effective training. While the benefits of increasing the number of watersheds are well established (Kratzert et al., 2024), the utility of extending the temporal length of training data remains unclear.

Empirical evidence from studies such as Boulmaiz et al. (2020) and Gauch et al. (2021) suggests that longer training periods enhance LSTM performance in rainfall-runoff modeling. However, these studies neglected the influence of **data recency**.

In the context of climate change and anthropogenic interventions, the assumption of stationarity (i.e., that historical patterns reliably represent future conditions) may no longer hold for hydrological systems. Intriguingly, Shen et al. (2022) found that calibrating physically-based hydrologic models to the latest data is superior to calibrating to old data.

This study aims to address two research questions:

(1) As the number of watersheds increases, is it still necessary to train LSTM models on decades of historical observations?

(2) Can LSTM models achieve comparable performance using shorter training periods focused on more recent data?

Model Framework

A standard LSTM neural network was employed to predict daily average streamflow.

- The LSTM models were trained using the open-source library NeuralHydrology (v1.11.0) by Kratzert et al. (2019) for 30 epochs on 4 NVIDIA RTX A6000 GPUs
- The streamflow predictions were obtained as the ensemble average of 10 LSTM models, each trained with a different random seed.
- LSTM hyperparameters were tuned using 5-fold cross-validation on a small sample of watersheds, following a procedure similar to that of Mai et al. (2022).

Performance is assessed using the median Kling-Gupta Efficiency (KGE) for daily average streamflow across all watersheds based on two evaluation settings:

- **Temporal testing**, which uses unseen years at training watersheds.
- **Spatiotemporal testing**, which uses unseen years at unseen/untrained watersheds, indicating the prediction in ungauged basin (PUB) context.

Dataset Overview

This study utilizes the North American scale HYSETS dataset for hydrometeorological modelling (Arsenault et al., 2020) to provide LSTM input variables with continuous daily record from 1950 to 2023. Watersheds were selected only if they had ≤5% missing discharge observations, yielding 1,374 watersheds across North America for experiments.

Among the selected watersheds:

- 1,100 were used for the training set, along with an additional 137 for validation in the machine learning context (to monitor convergence, not used in testing)
- 220 (out of the 1,100 training watersheds) were selected randomly and used for temporal testing.
- A final subset of 137 served as pseudo-ungauged sites for spatiotemporal testing

The number of input variables was limited to enable efficient repetitive training experiments:

- 3 dynamic forcing attributes of daily precipitation, max and min temperature
- 16 static watershed attributes (drainage area, land use composition, elevation, etc.)

Experimental Design

Three experiments were designed to evaluate model testing performance (2011-2023) using different training periods (16 to 61 years in length, from 1950-2010) over a range of training watershed sample sizes (220 to 1100).

Experiment 1: Valuation of adding older data to training with backward-expanding training periods

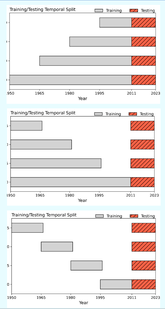
- Training dataset always includes recent period (1995-2010)
- Progressively extend training dataset backward in time with up to three additional 15-year blocks of data

Experiment 2: Valuation of adding newer data to training with forward-expanding training periods

- Training dataset always includes oldest available period (1950-1965)
- Progressively extend training dataset forward in time with up to three additional 15-year blocks of data

Experiment 3: Valuation of choosing newer data *instead* of older data for training with sliding window training periods

- Training spans a fixed length of 16 years, shifting forward in time (e.g., 1950–1965, 1965–1980, 1980–1995...).
- This isolates the effect of training period timing on model performance



Results

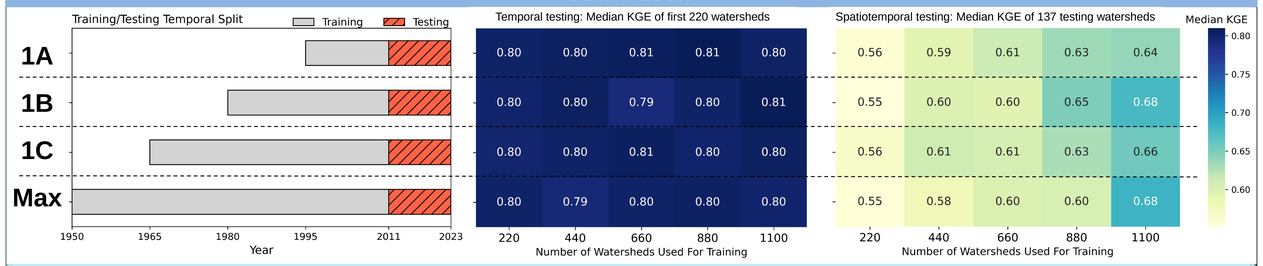


Figure 1. Experiment 1 results for training/testing patterns on left. Median KGE scores for temporal (middle) and spatiotemporal (right) testing. Training dataset size varies spatially (x-axis: number of watersheds) and temporally by progressively adding older data (y-axis).

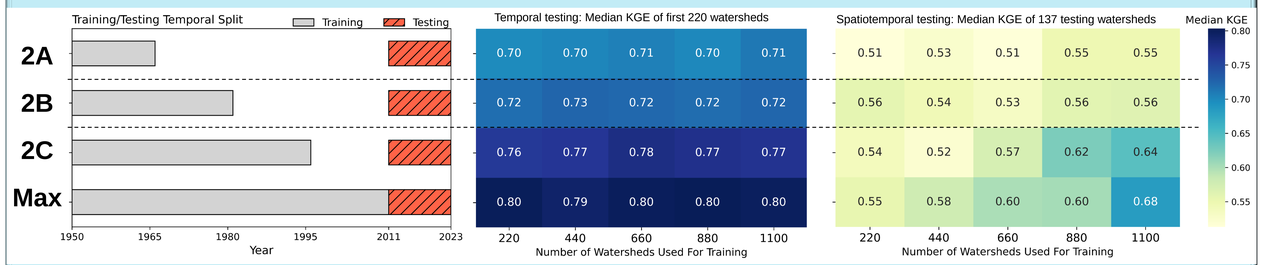


Figure 2. Experiment 2 results for training/testing patterns on left. Median KGE scores for temporal (middle) and spatiotemporal (right) testing. The training dataset size varies spatially (x-axis: number of watersheds) and temporally by progressively adding newer data (y-axis).

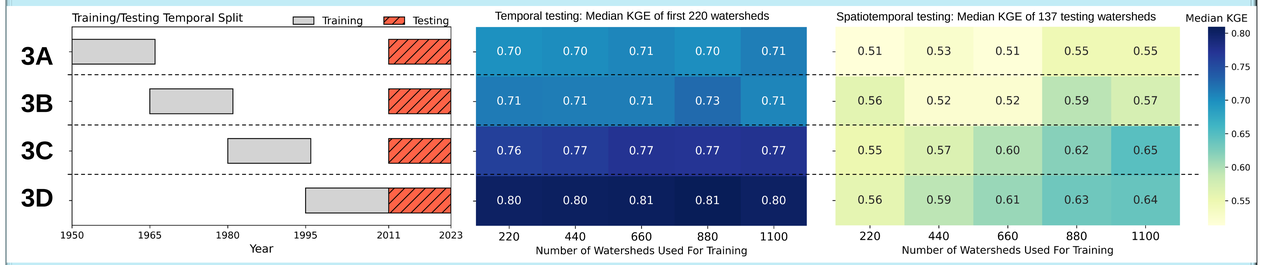


Figure 3. Experiment 3 results for training/testing patterns on left. Median KGE scores for temporal (middle) and spatiotemporal (right) testing. The training dataset size varies spatially (x-axis: number of watersheds) and temporally by shifting the training period while keeping the period length constant.

References

Arsenault, R., Brissette, F., Martel, JL. et al. A comprehensive, multisource database for hydrometeorological modeling of 14,425 North American watersheds. Scientific Data, 7, 243, https://doi.org/10.1038/s41597-020-00583-2, 2020.

Boulmaiz, T., Guermoui, M., and Boutaghane, H.: Impact of training data size on the LSTM performances for rainfall–runoff modeling, Model Earth Systems and Environment, 6, 2153–2164, https://doi.org/10.1007/S40808-020-00830-W/FIGURES/9, 2020.

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

Kratzert, F., Gauch, M., Nearing, G., & Klotz, D. NeuralHydrology---A Python library for Deep Learning research in hydrology. Journal of Open Source Software, 7(71), 4050, https://doi.org/10.21105/joss.04050, 2022.

Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS Opinions: Never train an LSTM on a single basin, Hydrology and Earth System Science, 28, 4187–4201, https://doi.org/10.5194/hess-28-4187-2024, 2024.

Mai, J., Shen, H., Tolson, B. A., Gaborit, É., Arsenault, R., Craig, J. R., Fortin, V., Fry, L. M., Gauch, M., Klotz, D., Kratzert, F., O'Brien, N., Princz, D. G., Rasiya Koya, S., Roy, T., Seglenieks, F., Shrestha, N. K., Temgoua, A. G. T., Vionnet, V., and Waddell, J. W.: The Great Lakes Runoff Intercomparison Project Phase 4: the Great Lakes (GRIP-GL), Hydrology and Earth System Science, 26, 3537–3572, https://doi.org/10.5194/hess-26-3537-2022, 2022.

Shen, H., Tolson, B. A., and Mai, J.: Time to Update the Split-Sample Approach in Hydrological Model Calibration, Water Resources Research, 58, e2021WR031523, https://doi.org/10.1029/2021WR031523, 2022.

Discussion & Conclusions

Experiment 1: Results indicate that models trained with shorter but more recent data often perform as well as those trained on extended historical records, regardless of the number of watersheds.

Experiment 2: Results clearly show that adding more recent data improves model performance, regardless of the testing context. In contrast, models trained solely on older data (e.g., pre-1980) perform poorly.

Contrasting results between Exp.1 and Exp.2 (e.g., 1A vs 2A) highlights that *the effect of data recency is larger than the effect of training period length.*

Experiment 3: Results clearly show that models trained with recent data consistently outperform those trained with older data, despite having the same training length. This trend persists equally in both PUB and non-PUB contexts, aligning with findings from Shen et al. (2022).

Does increasing number of watersheds (from 220 to 1,100) improve model performance?

- True in PUB context but conditional on the selection of training period. Significant improvement when trained with latest data (i.e., most recent 16-year data); Little to no improvement when trained with only old data, regardless of watershed count (2A, 2B, 3A, 3B).
- False in non-PUB context. Training with 220 watersheds sufficed to achieve near-optimal performance across all training periods.

Does extending the training period length (from 16 to 61 years) improve model performance?

- Adding distant-past data, especially those older than 30 years, is not beneficial.
- Moreover, training model on extended data records leads to noticeably increased training time.

The temporal recency of training dataset is a critical factor when training LSTM for rainfall-runoff modelling. Ongoing work will systematically evaluate the generalizability of these findings for using LSTM-based modelling in North America.