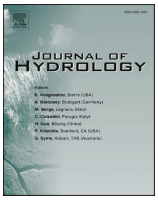
Contents lists available at ScienceDirect

Journal of Hydrology

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



Pooling local climate and donor gauges with deep learning for improved reconstructions of streamflow in ungauged and partially gauged basins

Sungwook Wi [*], Rohini Gupta, Scott Steinschneider

Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, USA

ARTICLE INFO

This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Shreedhar Maskey, Associate Editor

Keywords:
Prediction in ungauged basins (PUB)
Deep learning
Long Short-Term Memory network
Great Lakes

ABSTRACT

Improving the accuracy of streamflow predictions in ungauged basins (PUB) has long been a significant challenge in hydrology. This study hypothesizes that deep learning-based PUB can be enhanced for historical streamflow reconstruction by integrating local climate data from ungauged or partially gauged basins (the target site) with streamflow measurements from nearby gauged basins (donor sites). The rationale is that streamflow records from donor sites offer valuable information for predicting streamflow at the target site. However, in some instances, local weather data may be more readily available, while available donors might be poorly correlated with the target. Therefore, prediction accuracy can be improved by weighting both sources effectively. To test this hypothesis, we conducted a case study using over 200 streamflow gauges in the Great Lakes region. We developed a multi-layer perceptron to estimate rank correlations of streamflow between basins, aiding in the selection of donor sites. These correlations were fed into a Long Short-Term Memory (LSTM) network, along with streamflow data from donor sites and weather data from target sites. We compared this model against two other LSTMs – one trained only with climate data and the other solely with streamflow from donor sites – as well as the average prediction from those two models. Our findings indicate that the integrated approach outperforms the alternatives, particularly for partially gauged and natural ungauged sites. Lastly, we demonstrate the value of the approach for improving lake-wide runoff estimates and use explainable AI to investigate how the model uses both climate and donor streamflow information.

1. Introduction

Improving the accuracy of daily or sub-daily streamflow predictions in ungauged basins (PUB) remains a significant challenge in hydrological research (Sivapalan et al., 2003; Hrachowitz et al., 2013; Blöschl et al., 2019). Various streamflow regionalization methods have been developed to address this issue by utilizing hydrological information from basins with similar characteristics – such as climate, geology, and topography – to estimate streamflow in ungauged or partially gauged basins. One widely adopted approach is data-driven regionalization, which employs statistical, empirical, or machine learning techniques to predict streamflow based on relationships between hydrological responses and catchment characteristics (e.g., Fennessey, 1994; Archfield and Vogel, 2010; Worland et al., 2019a,b). Alternatively, regionalization can be based on hydrological models, where calibrated parameters from gauged basins are transferred to ungauged ones based on spatial proximity and physical similarity (e.g., Choubin et al., 2019; Pool et al., 2021; Singh et al., 2022; Wu et al., 2023).

Despite these efforts, substantial challenges persist in both hydrological model-independent and model parameter regionalization methods. These challenges stem from various factors, including the spatial and temporal heterogeneity in basin characteristics, the non-stationarity of hydrological processes driven by climate change and human activities, uncertainties in model parameters, and limited data availability (Hrachowitz et al., 2013; Razavi and Coulibaly, 2013). Such complexities hinder the generalization of regionalization methods, making accurate predictions in ungauged basins elusive.

With the rise of big data and advances in artificial intelligence (AI), deep learning (DL) models have opened new opportunities for improving streamflow predictions in ungauged basins. Long Short-Term Memory networks (LSTMs) in particular have shown strong generalization to new basins by learning hydrologic patterns from large datasets containing data from hundreds or thousands of gauged basins (Kratzert et al., 2024). Over the past decade, research has demonstrated that LSTMs trained on extensive regional data significantly outperform traditional hydrological models and other statistical approaches in

[*] Corresponding author.
E-mail address: sw2275@cornell.edu (S. Wi).

https://doi.org/10.1016/j.jhydrol.2025.133764
Received 16 January 2025; Received in revised form 9 May 2025; Accepted 22 June 2025
Available online 23 June 2025

predicting streamflow in ungauged basins (e.g., Kratzert et al., 2019; Mai et al., 2022; Arsenault et al., 2023).

Since LSTMs emerged as a state-of-the-art method for streamflow regionalization, research has increasingly focused on refining these models through alternative architectures, innovative training strategies, and additional information sources. For instance, past work has improved streamflow prediction by combining Convolutional Neural Networks (CNNs) with LSTMs to utilize spatial feature extraction with temporal sequence modeling (Guo et al., 2024; Pokharel and Roy, 2024), and also with attention-based LSTMs to dynamically adapt the use of historical patterns (e.g., previous months' snowmelt) for predictions in different times of the year (Alizadeh et al., 2021). Hybrid models that integrate LSTMs with hydrological models (Feng et al., 2024; Jiang et al., 2020) or incorporate physical constraints (Hoedt et al., 2021; Wi and Steinschneider, 2024) enhance the physical consistency of predictions and improve their generalizability across diverse climatic conditions, including those influenced by climate change. Additionally, transfer learning has been shown to improve predictions in under-monitored regions by fine-tuning LSTMs trained on large, high-quality datasets (Ma et al., 2021; Le et al., 2024; Xu et al., 2023), supported by globally available remote sensing products (Wilbrand et al., 2023). Improvements in regionalization have also been achieved through the development of region-specific LSTMs based on watershed clustering using physiographic attributes (He et al., 2024), or by iteratively training LSTMs to all sites and those in specific clusters to improve cluster-specific performance (Ghaneei et al., 2024).

This study focuses on the value of another strategy – data integration – to advance DL-based hydrologic modeling, specifically in the context of historical streamflow estimation. Data integration (also referred to as data fusion) combines multiple, disparate sources of information to improve predictions over what is achievable with a single data source (Dasarathy, 1997). This approach is straightforward with deep learning models but can be challenging with physically based models, which require complex data assimilation schemes to integrate novel data streams (Feng et al., 2020). Past work has shown that data integration can improve hydrologic prediction by combining different precipitation datasets (Kratzert et al., 2021), multi-scale soil moisture measurements (Liu et al., 2022), and the combined use of climate and hydrologic state variables (e.g., streamflow, snow water equivalent, soil moisture). This last data integration strategy, however, has primarily been employed for forward simulations (e.g., streamflow forecasting), using forecasted weather and lagged hydrologic state variables at a target site to improve forecasts at that site (Fang and Shen, 2020; Feng et al., 2020; Meyal et al., 2020; Nearing et al., 2022; Jahangir and Quilty, 2024; Song et al., 2024).

In addition to forward simulations, an equally important goal of PUB is the reconstruction of historical streamflow records. Such reconstruction is crucial for improving our understanding of past water availability and variability, making it an essential component of sustainable water resource management in the face of a changing climate and growing demand (Milly et al., 2005; Vörösmarty et al., 2010). One instance of data integration that has not been adequately explored in this context is the fusion of local climate data at a target site and streamflow measurements at nearby donor sites. When considering streamflow reconstruction (rather than forecasting), streamflow measurements at donor sites have the potential to provide extremely valuable information for streamflow prediction at a target site. While this type of approach is common in the PUB literature (e.g., the QPPQ method promoted by the United States Geological Survey; see Worland et al., 2019a,b), to the authors' knowledge it has not yet been extended to DL-based hydrologic models.

If donor basins are to be used for streamflow regionalization, they must be selected with care (Patil and Stieglitz, 2012). One common approach is to select a donor that is closest to the target basin (Fennessey, 1994; Mohamoud, 2008). Often, improved predictions are achieved by using multiple donor gauges selected based on multiple criteria (e.g., geographic proximity and physical attributes) rather than relying on a single similarity measure (Arsenault and Brissette, 2014; Pool et al., 2021; Shu and Ouarda, 2012; Yang et al., 2018; de Lavenne et al., 2016). Towards this end, one or more donors can be selected to have the highest correlation to the target site, as estimated by regional regressions or kriging that account for multiple similarity measures (Archfield and Vogel, 2010; Yuan, 2013). We leverage this approach in the work presented here.

The central hypothesis of this study is that DL-based PUB can be enhanced for the purpose of historical streamflow reconstruction by integrating local climate data in ungauged basins with streamflow measurements from donor gauges. The underlying concept is that streamflow records from donor basins often provide valuable information to predict streamflow at target sites, but in some instances, local weather data at the target site may offer better predictive capacity when donor basins are distant or poorly correlated with the target site. Therefore, by weighting both sources of information, we aim to achieve more accurate streamflow predictions.

To test the hypothesis, we apply our approach to a case study involving over 200 streamflow gauges in the Great Lakes region. We develop a multi-layer perceptron to estimate inter-basin streamflow correlations and select donor sites. These estimated correlations are also used as inputs to an LSTM, along with streamflow data from the donor sites and local weather data from the ungauged basins. We compare the performance of our integrated model against two other LSTMs – one trained exclusively on climate data and another solely on donor site streamflow data – to isolate the value of each information source and the added value through their integration. We also consider the average prediction from the LSTMs trained only on climate or donor-site streamflow, to compare the benefits of a model ensemble averaging approach (Farmer and Vogel, 2013; Pool et al., 2019; Razavi and Coulibaly, 2013; Swain and Patra, 2017; Waseem et al., 2015) to a single model that integrates all available information sources. We assess all models in two cases, out-of-sample in time and out-of-sample in space, to evaluate how each model performs for partially gauged versus ungauged sites. We also evaluate models separately for sites deemed more natural and those classified as being subject to anthropogenic impacts.

Our results demonstrate utility in the approach that integrates local climate and donor streamflow data, particularly for temporal extrapolation at partially gauged basins and for predictions in ungauged basins with minimal anthropogenic influence. We show that the use of DL for streamflow reconstruction can significantly improve the estimations of lake-wide runoff, with important implications for quantifying the Great Lake water balance. Finally, we utilize explainable AI techniques to show how DL models can balance the use of climate and donor site streamflow information over time and under varying hydrologic conditions.

2. Study area and data

Our study focuses on the Great Lakes region (Fig. 1), which spans over 900,000 km^2 across the United States and Canada (Mai et al., 2022). Encompassing the Great Lakes basin (~765,000 km^2) and the Ottawa River basin (~146,000 km^2), the vast drainage system exhibits a complex array of hydrologic responses shaped by varied geographic features – such as climate, land cover, and topography – as well as significant human impacts, including urbanization, agricultural practices, and water control structures (Kult et al., 2014). The basin's inherent complexity poses challenges for streamflow regionalization, which are further compounded by difficulties in cross-border data access and integration (Fry et al., 2022; Gronewold et al., 2018). As such, the Great Lakes region offers a valuable setting for advancing regionalization techniques, which can improve Great Lakes water level predictions. Enhanced predictions will, in turn, support better water resource management, balancing the needs of economic development, environmental

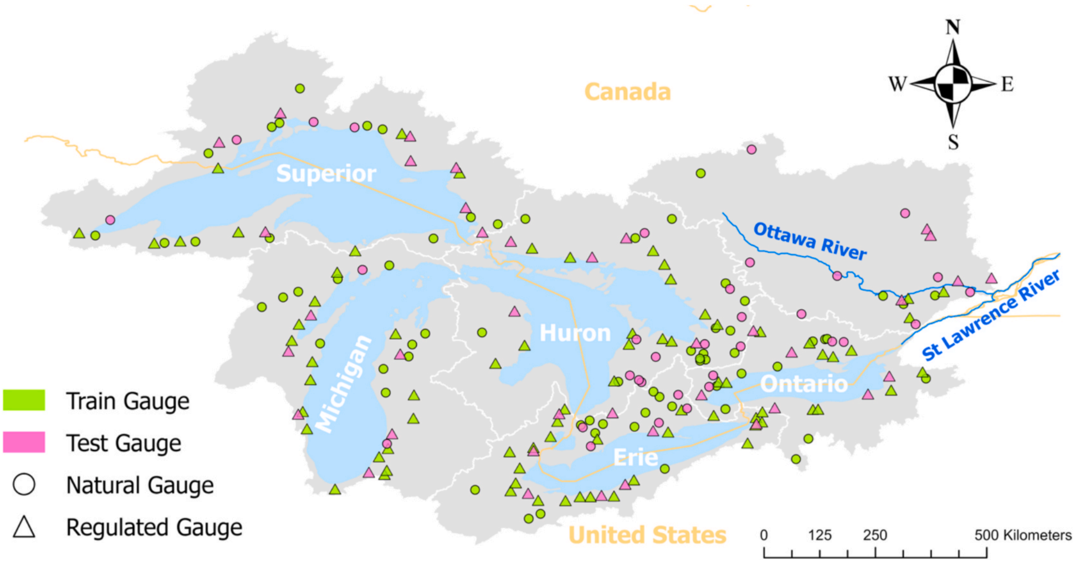


Fig. 1. Study area of the Great Lakes region, including the Great Lakes watersheds and the Ottawa River basin. This study utilizes data from 212 stream gauges (141 for training and 71 for testing) classified as either natural (i.e., low human impact) or regulated (Mai et al., 2022). The training set comprises 66 natural and 75 regulated gauges, while the testing set includes 33 natural and 38 regulated gauges.

stewardship, and community safety in this populous yet largely ungauged area (Gronewold and Rood, 2019; Kult et al., 2014).

This study examines 212 watersheds distributed across the five Great Lake basins and the Ottawa River basin as shown in Fig. 1. Data for these watersheds, sourced from a comprehensive rainfall-runoff model intercomparison in this region (Mai et al., 2022), include daily streamflow records, meteorological forcings, and geophysical attributes. Streamflow measurements, collected by the U.S. Geological Survey and Water Survey Canada, span from January 2000 to December 2017. Each gauging station represents a drainage area of at least 200 km^2, with less than 5 % missing data over the study period. In the experimental setup (described further in the following section), 141 watersheds are designated as training sites and the remaining 71 are used for testing (see Fig. 1). Gauges are classified as being minimally impacted by human activities or regulated by water control structures, regardless of their designation as training or testing sites (see Mai et al., 2022; classification provided in their Supporting Material).

Meteorological forcings are derived from the Regional Deterministic Reanalysis System v2 (Gasset et al., 2021), a 10 km resolution, hourly dataset covering North America. Hourly values of precipitation, net incoming shortwave radiation (Rs), and temperature are aggregated to produce daily basin-wide averages for precipitation and Rs, along with daily minimum and maximum temperatures. The precipitation data, generated through the Canadian Precipitation Analysis, combines surface observations with short-term forecasts from the Regional Deterministic Reforecast System, making it gauge-based and spatially interpolated rather than purely model-driven.

Geophysical attributes of each watershed are sourced from various datasets. Basin-wide elevation and slope statistics come from the HydroSHEDS digital elevation model at 3 arcsec resolution (Lehner et al., 2008). Soil properties, including soil texture classes, are extracted from the Global Soil Dataset for Earth System Models at a 30 arcsec resolution (Shangguan et al., 2014). Land cover data, based on Landsat imagery from 2010 to 2011 at a 30 m resolution, are obtained from the North American Land Change Monitoring System (NALCMS, 2017). These geophysical datasets provide basin-averaged attributes for each watershed, further detailed in Mai et al. (2022) and listed in Table 1.

Table 1
Watershed attributes used as inputs for the deep learning rainfall-runoff models developed in this study (adapted from Wi and Steinschneider, 2024). Attributes marked with an asterisk (*) are used in the multi-layer perceptron model for selecting donor gauges.

Attribute	Description
p_mean	Mean daily precipitation
pet_mean	Mean daily potential evapotranspiration
aridity	Ratio of mean PET to mean precipitation
t_mean	Mean of daily maximum and daily minimum temperature
frac_snow	Fraction of precipitation falling on days with mean daily temperatures below 0 °C
high_prec_freq	Fraction of high-precipitation days (=5 times mean daily precipitation)
high_prec_dur	Average duration of high-precipitation events (number of consecutive days with =5 times mean daily precipitation)
low_prec_freq	Fraction of dry days (<1 mm d-1 daily precipitation)
low_prec_dur	Average duration of dry periods (number of consecutive days with daily precipitation <1 mm d-1)
mean_elev*	Catchment mean elevation
std_elev*	Standard deviation of catchment elevation
mean_slope*	Catchment mean slope
std_slope*	Standard deviation of catchment slope
area_km2*	Catchment area
Temperate-or-sub-polar-needleleaf-forest*	Fraction of land covered by “Temperate-or-sub-polar-needleleaf-forest”
Temperate-or-sub-polar-grassland*	Fraction of land covered by “Temperate-or-sub-polar-grassland”
Temperate-or-sub-polar-shrubland*	Fraction of land covered by “Temperate-or-sub-polar-shrubland”
Temperate-or-sub-polar-grassland*	Fraction of land covered by “Temperate-or-sub-polar-grassland”
Mixed-Forest*	Fraction of land covered by “Mixed-Forest”
Wetland*	Fraction of land covered by “Wetland”
Cropland*	Fraction of land covered by “Cropland”
Barren-Lands*	Fraction of land covered by “Barren-Lands”
Urban-and-Built-up*	Fraction of land covered by “Urban-and-Built-up”
Water*	Fraction of land covered by “Water”
BD*	Soil bulk density (g cm-3)
CLAY*	Soil clay content (% of weight)
GRAV*	Soil gravel content (% of volume)
OC*	Soil organic carbon (% of weight)
SAND*	Soil sand content (% of weight)
SILT*	Soil silt content (% of weight)

3. Methods

Fig. 2 illustrates our experimental design, which is briefly

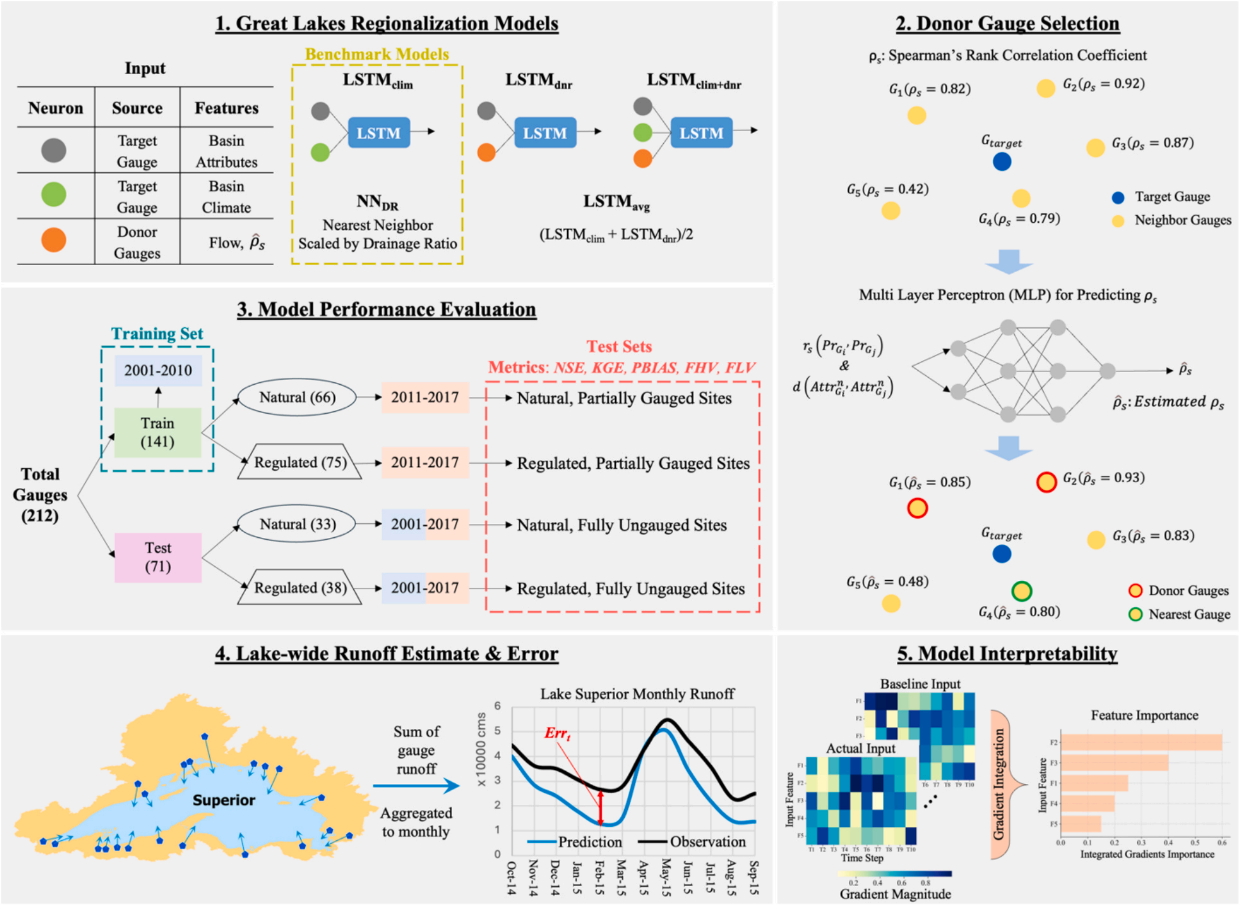

1. Great Lakes Regionalization Models
Input
Neuron
Source
Features
Target Gauge
Basin Attributes
Target Gauge
Basin Climate
Donor Gauges
Flow, $\hat{\rho}_s$
Benchmark Models
LSTM$_{clim}$
LSTM$_{dnr}$
LSTM$_{clim+dnr}$
LSTM
NN$_{DR}$
Nearest Neighbor Scaled by Drainage Ratio
LSTM$_{avg}$
(LSTM$_{clim}$ + LSTM$_{dnr}$)/2
2. Donor Gauge Selection
ρ_s: Spearman's Rank Correlation Coefficient
$G_1(\rho_s = 0.82)$
$G_2(\rho_s = 0.92)$
G_{target}
$G_3(\rho_s = 0.87)$
$G_5(\rho_s = 0.42)$
$G_4(\rho_s = 0.79)$
Target Gauge
Neighbor Gauges
Multi Layer Perceptron (MLP) for Predicting ρ_s
$r_s(Pr_{G_i}, Pr_{G_j})$ & $d(Attr^n_{G_i}, Attr^n_{G_j})$
$\hat{\rho}_s$
$\hat{\rho}_s$: Estimated ρ_s
$G_1(\hat{\rho}_s = 0.85)$
$G_2(\hat{\rho}_s = 0.93)$
G_{target}
$G_3(\hat{\rho}_s = 0.83)$
$G_5(\hat{\rho}_s = 0.48)$
$G_4(\hat{\rho}_s = 0.80)$
Donor Gauges
Nearest Gauge
3. Model Performance Evaluation
Training Set
2001-2010
Train (141)
Total Gauges (212)
Test (71)
Natural (66)
Regulated (75)
Natural (33)
Regulated (38)
2011-2017
2011-2017
2001-2017
2001-2017
Test Sets
Metrics: NSE, KGE, PBIAS, FHV, FLV
Natural, Partially Gauged Sites
Regulated, Partially Gauged Sites
Natural, Fully Ungauged Sites
Regulated, Fully Ungauged Sites
4. Lake-wide Runoff Estimate & Error
Superior
Sum of gauge runoff
Aggregated to monthly
Lake Superior Monthly Runoff
x 10000 cms
0 1 2 3 4 5 6
Err_t
Prediction
Observation
Oct-14 Nov-14 Dec-14 Jan-15 Feb-15 Mar-15 Apr-15 May-15 Jun-15 Jul-15 Aug-15 Sep-15
5. Model Interpretability
Baseline Input
Actual Input
Input Feature
F1 F2 F3 F4 F5
T1 T2 T3 T4 T5 T6 T7 T8 T9 T10
Time Step
Gradient Magnitude
0.2 0.4 0.6 0.8
Gradient Integration
Feature Importance
Input Feature
F2 F3 F1 F4 F5
0.0 0.1 0.2 0.3 0.4 0.5 0.6
Integrated Gradients Importance

Fig. 2. Schematic of the experimental design.

summarized here and described in more detail in the subsections below.

This work develops multiple LSTMs for streamflow regionalization across the Great Lakes region (see step 1 in Fig. 2):

- LSTM$_{\mathrm{Clim}}$: Driven by local climate data, without any donor streamflow inputs.
- LSTM$_{\mathrm{Dnr}}$: Driven by donor streamflow data, excluding climate inputs.
- LSTM$_{\mathrm{Clim+Dnr}}$: Incorporates both local climate data and streamflow data from donor gauges.
- LSTM$_{\mathrm{Avg}}$: An ensemble model that averages predictions from LSTM$_{\mathrm{Clim}}$ and LSTM$_{\mathrm{Dnr}}$.

Each model incorporates as input either climate data at the target site (LSTM$_{\mathrm{Clim}}$), streamflow data at donor sites (LSTM$_{\mathrm{Dnr}}$), or both (LSTM$_{\mathrm{Clim+Dnr}}$). All models also incorporate static basin features (i.e., geophysical attributes) at the target basin as inputs. The LSTM$_{\mathrm{Clim}}$ model is the same as the model developed in the rainfall-runoff model intercomparison of Mai et al. (2022) and again in Wi and Steinschneider (2024) and serves as a benchmark model against which to compare alternative LSTMs that use donor streamflow information. Additionally, we compare our models with another benchmark model, called the nearest-neighbor drainage ratio (NN$_{\mathrm{DR}}$) model, which estimates streamflow for ungauged basins by scaling streamflow data from the closest donor gauge using the drainage area ratio between the ungauged and donor sites. This approach is currently used for historical lake-wide runoff estimation in the Great Lakes (Hunter et al., 2015). Finally, LSTM$_{\mathrm{Avg}}$ represents predictions that are averaged from LSTM$_{\mathrm{Clim}}$ and LSTM$_{\mathrm{Dnr}}$ and so is not trained separately. This model allows us to compare the benefits of model ensemble averaging against a single model that integrates all available information sources (LSTM$_{\mathrm{Clim+Dnr}}$).

To select donor gauges for the LSTM$_{\mathrm{Dnr}}$ and LSTM$_{\mathrm{Clim+Dnr}}$ models (see step 2 in Fig. 2), we train a Multi-Layer Perceptron (MLP) to predict inter-basin streamflow correlations. These predictions are then used to identify donor gauges that are highly correlated with each target gauge, whose streamflow data and correlation estimates are then used as inputs in the LSTM$_{\mathrm{Dnr}}$ and LSTM$_{\mathrm{Clim+Dnr}}$ models.

The LSTM models are trained on data from 141 training basins (serving as donor basins) and evaluated on 71 testing gauges, which are treated as ungauged locations (see step 3 in Fig. 2 and Fig. 1). The period of record is also split into separate training and testing periods. This design allows us to evaluate model performance for the testing period at training sites (i.e., partially gauged sites) and for the training and testing period at testing sites (fully ungauged sites). Additionally, we evaluate models separately for sites under minimal human impact and those affected by regulation or other anthropogenic activity.

We assess lake-wide runoff estimates for six major watersheds in the study domain: the five Great Lakes watersheds and the Ottawa River watershed (see step 4 in Fig. 2). For each lake, we calculate monthly prediction errors from all gauges within the watershed and use these to estimate lake-wide error variance across the different regionalization models.

Finally, we use explainable AI to investigate how the integrated DL model (LSTM$_{\mathrm{Clim+Dnr}}$) balances the use of climate and donor streamflow information when making predictions, and how this balance changes over time and under varying hydrologic conditions (see step 5 in Fig. 2).

3.1. Long Short-Term Memory network (LSTM) for hydrological modeling

The application of LSTM networks in hydrological modeling leverages the model's capacity to retain and recall long-term dependencies in sequential data, enabled by its unique memory cell structure (Hochreiter and Schmidhuber, 1997). This feature makes LSTMs particularly well-suited for modeling rainfall-runoff processes (Kratzert et al., 2018).

An LSTM cell processes input data sequentially, one timestep at a time, while maintaining and updating its internal memory. At each timestep t, the cell receives:

- the current input vector $\boldsymbol{x_t}$ (dimension K),
- the previous hidden state $\boldsymbol{h}_{t-1}$ (dimension D), and
- the previous cell state $\boldsymbol{c}_{t-1}$ (dimension D).

The cell state $\boldsymbol{c}_t$ is updated based on three gating mechanisms: the input gate $\boldsymbol{i}_t$, forget gate $\boldsymbol{f}_t$, and output gate $\boldsymbol{o}_t$. These gates regulate the flow of information into and out of the memory cell, as defined by the following equations:

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \bullet [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_i)(\text{input gate})$$

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}_f \bullet [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_f\right)(\text{forget gate})$$

$$\widetilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_c \bullet [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_c)(\text{candidate cell state})$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o \bullet [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_o)(\text{output gate})$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \widetilde{\boldsymbol{c}}_t(\text{updated cell state})$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot tanh(\boldsymbol{c}_t)(\text{updated hidden state})$$

Here, σ indicates the sigmoid function, tanh is the hyperbolic tangent, and $\odot$ denotes element-wise multiplication. The matrices $\boldsymbol{W}$ and vectors $\boldsymbol{b}$ are learnable weights and biases associated with each gate. Each gate plays a distinct role:

- The input gate $\boldsymbol{i}_t$ determines how much new information from $\widetilde{\boldsymbol{c}}_t$ is added to the cell state.
- The forget gate $\boldsymbol{f}_t$ controls how much of the previous cell state $\boldsymbol{c}_{t-1}$ is retained.
- The output gate $\boldsymbol{o}_t$ regulates the amount of information from the current cell state $\boldsymbol{c}_t$ that is passed to the hidden state $\boldsymbol{h}_t$.

After processing all T timesteps, the final hidden state $\boldsymbol{h}_T$ is passed through a fully connected layer with a single neuron. A ReLU (Rectified Linear Unit) activation function is applied to ensure non-negative streamflow predictions:

$$\boldsymbol{y}_T = \text{ReLU}\left(\boldsymbol{W}_y \boldsymbol{h}_T + \boldsymbol{b}_y\right)$$

We developed three regional LSTM models for daily streamflow predictions across the Great Lakes region: $\text{LSTM}_{\text{Clim}}$, LSTM_{Dnr}, and $\text{LSTM}_{\text{Clim+Dnr}}$, each with a distinct input configuration. The $\text{LSTM}_{\text{Clim}}$ has 39 input features ($K = 39$): 9 climate variables and 30 catchment attributes for the target basin. Climate inputs include basin-averaged daily precipitation, maximum and minimum temperatures, net incoming shortwave radiation, specific humidity, surface air pressure, zonal and meridional wind components, and potential evapotranspiration. Catchment attributes are detailed in Table 1. Climate inputs vary dynamically across all time steps, while catchment attributes remain static within the input sequence $\boldsymbol{x}$.

The LSTM_{Dnr} takes as input the same 30 catchment attributes as $\text{LSTM}_{\text{Clim}}$, but it does not use the dynamic climate inputs. Instead, the LSTM_{Dnr} model takes as input standardized streamflow data from M donor sites, as well as estimated rank correlations between streamflow at each donor site and the target site (leading to a total of $2M + 30$ input features). More detail on the standardized donor streamflow, estimated rank correlations, and selection of M donors is provided in Section 3.2. Both streamflow and correlation inputs for each donor site serve as static features, since donor streamflow measurements are provided only for the final time step (T) rather than across all time steps $1,\ldots,T$.

The $\text{LSTM}_{\text{Clim+Dnr}}$ model combines the inputs from the $\text{LSTM}_{\text{Clim}}$ and LSTM_{Dnr} models, for a total of $2M + 39$ inputs. Before training, all input features for all models are standardized by subtracting the mean and dividing by the standard deviation across all training sites in the training period, except for donor gauge streamflow measurements, which were pre-standardized (see Section 3.2). Observed streamflow, while not standardized, is divided by drainage area to represent flow in millimeters.

Each LSTM model is trained on 141 training sites from 2001 to 2010 and evaluated across these sites for the period 2011–2017 to assess out-of-sample performance in time. Performance is also evaluated on 71 test gauges over the entire 2001–2017 period (out-of-sample in space). The models were trained by minimizing the mean-squared error (MSE) across training watersheds:

$$MSE = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{T_n}\sum_{t=1}^{T_n}\left(\widehat{Q}_{n,t} - Q_{n,t}\right)^2$$

where N is the number of training watersheds and T_n is the number of samples in the n^{th} watershed. $\widehat{Q}_{n,t}$ and $Q_{n,t}$ are, respectively, the predicted and observed streamflow (in mm) for basin n and day t.

The final model architectures are determined using 5-fold cross-validation to tune hyperparameters such as LSTM size, learning rate, mini-batch size, sequence length, dropout rate, epochs, and donor gauge count M (for applicable models). Based on this cross-validation, each model was optimized with an input sequence of one year (i.e., $T = 365$ days), one LSTM layer with 256 neurons ($D = 256$), a mini-batch size of 64, learning rate of 0.0005, drop-out rate of 0.4, and $M = 5$ donor gauges. The full cross-validation results and rationale for these selections are provided in the Supporting Information (Text S1, Figs. S1–S5). The $\text{LSTM}_{\text{Clim}}$ models were trained over 30 epochs, while $\text{LSTM}_{\text{Clim+Dnr}}$ and LSTM_{Dnr} converged faster and were trained for 10 epochs to prevent overfitting. Network weights were tuned using the ADAM optimizer (Kingma and Ba, 2015). To address uncertainty in model training, each model was trained 10 times with different random initializations. The daily streamflow predictions for each model represent the ensemble average across these 10 trials.

Finally, we constructed an ensemble model, LSTM_{Avg}, which averages the daily predictions of $\text{LSTM}_{\text{Clim}}$ and LSTM_{Dnr}, assigning equal weights (0.5) to each model's output to produce a composite streamflow prediction.

3.2. Multi-Layer Perceptron (MLP) for selecting donor gauges

We employ a MLP to select donor gauges that will inform the LSTM_{Dnr} and $\text{LSTM}_{\text{Clim+Dnr}}$ models (step 2 in Fig. 2). The MLP is used to estimate the Spearman's rank correlation for daily streamflow between pairs of stream gauges, with the goal of identifying donor gauges that are strongly correlated to target sites (similar to Archfield and Vogel (2010) and Yuan (2013)). The use of correlation to select donor sites helps identify hydrologically similar donor basins, even those located beyond the immediate vicinity of the target site. Spearman's rank correlation is adopted to capture monotonic relationships in flow patterns while mitigating the influence of infrequent extreme values. Initial testing showed that the use of Spearman correlations led to better performance of the LSTM_{Dnr} and $\text{LSTM}_{\text{Clim+Dnr}}$ models compared to Pearson correlations. Consistent with geostatistical approaches such as Top-kriging (Skøien and Blöschl, 2007; de Lavenne et al., 2016), we incorporate multiple donor gauges to enhance prediction robustness and mitigate the risk associated with relying on a single donor.

The MLP model is trained on streamflow data from the 141 training gauges during the training period (2001–2010). The model is structured with an input layer, two hidden layers, and a fully connected output layer containing a single neuron, designed to output an estimated rank correlation in daily streamflow between pairs of training gauges. For each pair of sites, inputs to the MLP include: 1) the Euclidean distance of physiographic characteristics between watersheds (see Table 1 for a list of these characteristics); 2) the Euclidean distance between watershed centroids; and 3) the Spearman's rank correlation of daily precipitation between the two sites. Cross validation over the training period and training sites was used to optimize model hyperparameters (the number of neurons and activation functions). The final MLP architecture includes two hidden layers with 35 neurons each, using a hyperbolic tangent function in the first layer and a sigmoid activation function in the second.

Once trained, the MLP model is extended for donor gauge selection across all 212 gauges (141 training and 71 testing gauges). For each streamflow gauge, *M* donor gauges are selected from the training set based on the highest MLP-estimated correlations. As mentioned in Section 3.1, each selected donor gauge contributes two input features to the regional LSTM_{Dnr} and $\text{LSTM}_{\text{Clim+Dnr}}$ models: (1) MLP-estimated correlation, serving as a weight indicating the strength of the relationship, and (2) donor gauge streamflow data for the day on which a prediction is generated for the target basin. Importantly, instead of using raw donor streamflow values as inputs, the donor gauge data are standardized via quantile mapping. That is, each donor streamflow value is passed through the empirical flow duration curve at the donor site to derive the associated non-exceedance probability, which is then passed through the quantile function of the standard normal distribution to produce a z-score. This process reduces the impact of outliers at donor gauges that might otherwise mislead predictions for the target basin.

Rather than relying on a single donor gauge, the MLP model enables the identification of multiple high-correlation donor gauges. This multi-donor approach enhances predictive stability by incorporating diverse hydrological responses, ultimately improving the accuracy of regional streamflow predictions (Qi et al., 2021; Yang et al., 2018). As mentioned in Section 3.1, the number of donors *M* is a hyperparameter of the LSTM_{Dnr} and $\text{LSTM}_{\text{Clim+Dnr}}$ models that is selected through cross-validation on the training set.

3.3. Model performance evaluation

As noted previously, 141 basins are designated as training sites, and the remaining 71 basins are used for testing. The training period spans January 2001 to December 2010, while the testing period extends from January 2011 to December 2017. This configuration supports both temporal and spatial out-of-sample evaluation. Additionally, we evaluate model performance for sites under minimal human impact separately from those affected by regulation or other anthropogenic activity. This setup results in four groups of target sites for model evaluation:

- Natural, partially gauged sites: evaluate model performance at natural training gauges for the test period of 2011–2017.
- Regulated, partially gauged sites: evaluate model performance at regulated training gauges for the test period of 2011–2017.
- Natural, fully ungauged sites: evaluate model performance at natural test gauges over the entire period of 2001–2017.
- Regulated, fully ungauged sites: evaluate model performance at regulated test gauges over the entire period of 2001–2017.

Following previous intercomparison studies (Frame et al., 2022; Mai et al., 2022), we use several metrics for model evaluation, including: the Nash-Sutcliffe Efficiency (NSE; (Nash and Sutcliffe, 1970)); Kling-Gupta Efficiency (KGE; (Gupta et al., 2009)); absolute percent bias (PBIAS); peak flow PBIAS (FHV; (Yilmaz et al., 2008)), focusing on the top 2 % of flow values; and low flow PBIAS (FLV; (Yilmaz et al., 2008)), focusing on the bottom 30 % of flow values. Each metric is calculated independently for the four modes of model evaluation at each site. For all LSTM regionalization models, performance results are derived from the ensemble mean across 10 separate training trials.

3.4. Lake-wide runoff estimate and error evaluation

To evaluate each model's performance in predicting the Great Lakes water balance, we analyze errors in lake-wide runoff predictions at a monthly timescale. We focus on monthly rather than daily values because this is the timestep most often used to analyze Great Lakes water balance variability (see (O'Brien et al., 2024)). We estimate lake-wide runoff for each calendar month across the six major watersheds in the Great Lakes region—the five Great Lakes watersheds and the Ottawa River watershed. This monthly assessment enables us to identify potential seasonal patterns or trends, highlighting any systematic biases or temporal variations in model accuracy.

For each regionalization model, we aggregate streamflow predictions from all gauges within each of the six major watersheds to obtain total daily runoff estimates on a lake-wide scale, covering the period from 2001 to 2017. We do not remove nested gauges when summing flows to a daily, lake-wide total. Daily lake-wide estimates are then compiled into monthly totals for each watershed. Monthly error calculations are performed by comparing model predictions with observed streamflow data, also aggregated from gauges, to provide a direct assessment of model accuracy.

To quantify uncertainty in monthly totals at the lake-wide scale, we derive an aggregated variance of monthly lake-wide runoff error, calculated using the variance of individual site errors and their covariances:

$$Var\left(\sum_{i=1}^{N} Err_i\right) = \sum_{i,j=1}^{N} Cov\left(Err_i, Err_j\right) = \sum_{i=1}^{N} Var(Err_i) + \sum_{i\neq j} Cov\left(Err_i, Err_j\right)$$

Here, Err_i represents the monthly runoff error for each stream gauge within one of the Great Lakes watersheds. This aggregation process adheres to the principles of variance for sums of random variables.

We compare the variance of monthly, lake-wide runoff errors across all models to determine how differences between models – which were trained at the daily scale for individual rivers and streams – ultimately propagate into estimates of the monthly water balance across the Great Lakes. Understanding these differences will help identify the level of model complexity needed to improve Great Lakes water level estimation, with implications for regional resource management and water availability.

3.5. Model interpretability

To enhance model interpretability, we applied the Integrated Gradients (IG) method (Sundararajan et al., 2017), a widely adopted technique in explainable AI. IG estimates the contribution of each input feature to a model's prediction by integrating the gradients of the model's output with respect to its input along a straight path from a baseline (typically an all-zero or neutral input) to the actual input. As emphasized by Sundararajan et al., (2017), effective attribution methods should satisfy two key properties: sensitivity and implementation variance. Sensitivity ensures that if changing an input affects the output, the input should receive a non-zero attribution. Implementation variance guarantees that two functionally equivalent models (i.e., models that produce identical outputs for all inputs) yield identical attributions. IG satisfies both criteria, making it a theoretically sound and reliable method for feature attribution.

In our study, we use IG to compute attribution scores for all input features across all prediction days in the 2001–2017 simulation period for each of the 212 Great Lakes basins. We explore the average attribution scores across all inputs, as well as their seasonal patterns and co-variability, with a particular focus on climate and donor streamflow inputs in the $\mathrm{LSTM}_{\mathrm{Clim+Dnr}}$ model. This analysis provides insights into how the model integrates and prioritizes climate versus donor information over time and under varying hydrologic conditions.

4. Results

4.1. Donor gauge selection by MLP

The MLP model was trained on data from 141 gauges (2001–2010) to estimate Spearman's rank correlations (ρ_s) between gauge pairs. We evaluate its performance in two aspects: (1) the accuracy of MLP-predicted correlations ($\widehat{\rho}_s$) compared to observed correlations (ρ_s); and (2) the quality of donor gauges selected based on $\widehat{\rho}_s$ (termed 'estimated donors') versus those selected using ρ_s (termed 'optimal donors'). Note that in practice, donors for ungauged locations can only be selected based on $\widehat{\rho}_s$. However, we compare these donors to those that would be selected based on perfect information (i.e., using ρ_s) to determine the degree to which errors in $\widehat{\rho}_s$ lead to the selection of sub-optimal donors.

Fig. 3 presents scatter plots of ρ_s and $\widehat{\rho}_s$ for the four evaluation groups of target sites (described in Section 3.3), with black asterisks (*) marking the five donors selected for each target site based on the highest $\widehat{\rho}_s$. Fig. S6 presents a similar figure but with selected donors chosen based on ρ_s. For training sites, predicted correlations closely aligned with observed values (Fig. 3a and b). The MLP model maintains high accuracy for natural testing sites (Fig. 3c), although performance slightly declines compared to the training dataset. However, for regulated testing sites (Fig. 3d), performance deteriorates, with much larger spread between ρ_s and $\widehat{\rho}_s$ and a notable overestimation bias.

Donors selected for natural sites demonstrate strong correlations to the target sites, as shown by the clustering of donors in the upper-right corners of Fig. 3a and c. This indicates that natural sites often have available highly correlated donor sites, and estimated correlations can be used to identify these donors. Supporting evidence from Figs. S6a and S6c (see the Supporting Information) shows similar clustering for optimal donors, suggesting that the estimated donors closely approximated the optimal ones. For natural training and testing sites, the donors selected using estimated correlations match those selected using observed correlations 78 % and 70 % of the time, respectively.

For regulated training sites, estimated donors largely aligned with optimal donors (see Fig. 3b and Fig. S6b), indicating limited impacts from using $\widehat{\rho}_s$ for donor selection. Here, estimated donors match optimal donors 82 % of the time. However, regulated testing sites reveal notable errors, including the selection of poorly correlated donors due to significant overestimations of $\widehat{\rho}_s$ (Fig. 3d and Fig. S6d). Unlike natural sites, regulated sites exhibit a broader range of donor correlation values, with many falling below 0.6 for both estimated and optimal donors. This suggests that high-correlation donors are more often unavailable for regulated sites. In addition, the degree of mismatch between estimated and optimal donors grows to 34 %.

Overall, the MLP model effectively estimates correlations and identified hydrologically relevant donors for natural sites. At these locations, estimated donors are highly correlated to the target sites, their estimated

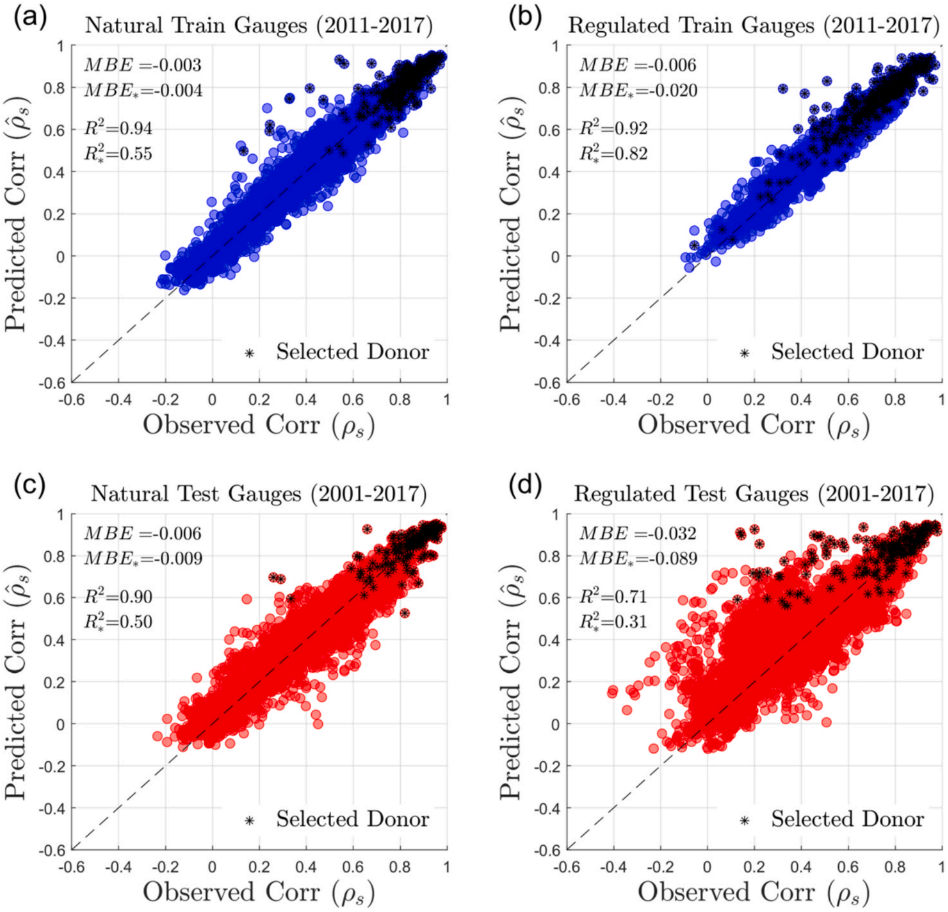


Fig. 3. Comparison of observed (ρ_s) and MLP-predicted correlations ($\widehat{\rho}_s$) for (a) natural training sites from 2011 to 2017, (b) regulated training sites from 2011 to 2017, (c) natural testing sites from 2001 to 2017, and (d) regulated testing sites from 2001 to 2017. Black asterisks (*) indicate donor gauges selected based on $\widehat{\rho}_s$ (i.e., estimated donors). Mean bias error (MBE) and the coefficient of determination (R^2) are presented for both the entire dataset and the subset of selected donors (marked with *). Positive (negative) MBE indicates an underestimation (overestimation) bias.

correlations are similar to the true correlations, and differences between estimated and optimal donors are small. In regulated basins, however, the use of donors poses larger challenges. In some instances, there are few if any donors that are highly correlated to the target site, making it difficult to identify hydrologically meaningful donors even if correlations with the target are known. This challenge likely stems from anthropogenic impacts causing unique hydrological behavior, and is further exacerbated by the need to estimate correlations between donors and regulated target sites, leading to further error in donor selection. The consequences of these outcomes on streamflow reconstructions are shown next.

4.2. Model performance evaluation

Fig. 4 compares the predictive performance of five regional rainfall-runoff models developed for the Great Lakes region, with each panel corresponding to one of four evaluation groups. Model performance is assessed using the distribution of Nash-Sutcliffe Efficiency (NSE) across sites, with distinct lines representing each model. Recall that all results shown are out-of-sample in time (for the training sites) or out-of-sample in space (for the testing sites).

The $\mathrm{NN_{DR}}$ model consistently and substantially underperforms across all evaluation groups and is therefore excluded from detailed comparisons. The analysis focuses instead on the LSTM models. At natural training sites (Fig. 4a), models that utilize donor information ($\mathrm{LSTM_{Dnr}}$ and $\mathrm{LSTM_{Clim+Dnr}}$) outperform those that do not ($\mathrm{LSTM_{Clim}}$). This highlights the value of hydrologically relevant donors in enhancing model accuracy, particularly when donors are selected based on strong correlations with target sites (see Fig. 3a). At regulated training sites (Fig. 4b), the $\mathrm{LSTM_{Clim+Dnr}}$ remains the top-performing model, but its advantage over $\mathrm{LSTM_{Clim}}$ diminishes somewhat due to the inclusion of lower-quality donors at regulated sites (see Fig. 3b). It is also noteworthy that at the regulated training sites, $\mathrm{LSTM_{Dnr}}$ performs slightly worse than $\mathrm{LSTM_{Clim}}$ even though $\mathrm{LSTM_{Clim+Dnr}}$ outperforms $\mathrm{LSTM_{Clim}}$, suggesting that the $\mathrm{LSTM_{Clim+Dnr}}$ model learned how to effectively combine climate and donor information to enhance predictive skill over models that only use one of those two information sources. Similar results are seen for $\mathrm{LSTM_{Avg}}$ (i.e., when predictions from $\mathrm{LSTM_{Clim}}$ and $\mathrm{LSTM_{Dnr}}$ are averaged).

At natural testing sites (Fig. 4c), $\mathrm{LSTM_{Clim+Dnr}}$ and $\mathrm{LSTM_{Avg}}$ perform slightly better than the other models, demonstrating the value of high-quality donors (see Fig. 3c) in improving predictions at fully ungauged sites. However, the degree of improvement of $\mathrm{LSTM_{Clim+Dnr}}$ and $\mathrm{LSTM_{Avg}}$ over $\mathrm{LSTM_{Clim}}$ is smaller than seen for the training sites. Conversely, at regulated testing sites (Fig. 4d), $\mathrm{LSTM_{Clim}}$ emerges as the best-performing model. Donor-informed models struggle in this scenario, as donor gauges for regulated sites often exhibit overestimated correlations and low hydrological relevance (see Fig. 3d). These poor-quality donor inputs mislead the models, resulting in degraded hydrologic prediction.

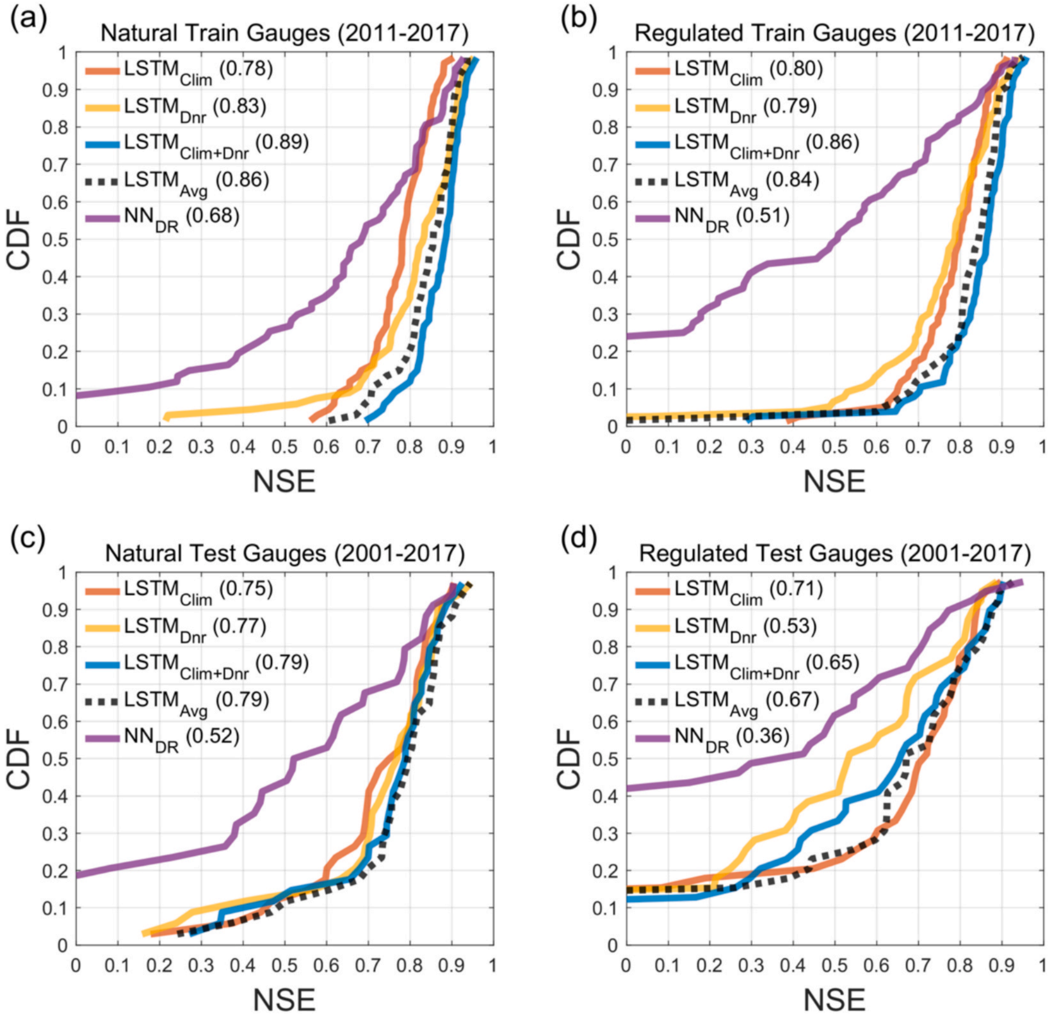


Fig. 4. Comparison of model performance (NSE) across four evaluation site groups, presented as cumulative distribution functions (CDFs). Median NSE values for each model are noted in parentheses.

A comprehensive comparison of model performance across all metrics is presented in Table 2. At natural and regulated training sites, $\mathrm{LSTM_{Clim+Dnr}}$ consistently outperforms other models across most metrics, except for FLV, where $\mathrm{LSTM_{Avg}}$ is the leading performer. At natural testing sites, the best-performing model varies depending on the metric. For instance, $\mathrm{LSTM_{Clim+Dnr}}$ and $\mathrm{LSTM_{Avg}}$ perform best in NSE, while $\mathrm{LSTM_{Clim}}$ performs best for KGE, and $\mathrm{LSTM_{Avg}}$ shows a slight advantage in PBIAS. Interestingly, despite its overall poor performance, $\mathrm{NN_{DR}}$ achieves the smallest FLV bias at natural testing sites. At regulated testing sites, $\mathrm{LSTM_{Clim}}$ demonstrates superior performance across all metrics, showcasing its robustness when reliable donor inputs are unavailable.

Overall, these results highlight that the integrated $\mathrm{LSTM_{Clim+Dnr}}$ model effectively balances donor and local climate inputs, achieving strong performance at minimally regulated sites and especially at partially gauged locations. However, the performance of donor-based models, including $\mathrm{LSTM_{Clim+Dnr}}$, degrades for regulated testing sites. $\mathrm{LSTM_{Clim}}$ proves to be the optimal model for these locations. Meanwhile, $\mathrm{LSTM_{Avg}}$ emerges as a promising alternative, offering competitive performance across different evaluation groups and metrics.

4.3. Donor-informed LSTMs versus $\mathrm{LSTM_{Clim}}$

Building on the findings above, we examine in more detail how donor quality – whether estimated or optimal – affects model performance. The goal is to determine the levels of correlation required for donor-informed models, such as $\mathrm{LSTM_{Dnr}}$ and $\mathrm{LSTM_{Clim+Dnr}}$, to enhance predictive skill, as well as identifying thresholds below which donor inputs become counterproductive, resulting in underperformance relative to $\mathrm{LSTM_{Clim}}$, which excludes donor data entirely.

Fig. 5 presents a scatter plot comparing the NSE values of $\mathrm{LSTM_{Clim+Dnr}}$ models informed by estimated and optimal donors. The observed Spearman's rank correlation (ρ_s) of the first (i.e., highest correlated) donor among the five optimal donors for each target site is categorized into four groups, represented by different colors. The results demonstrate that using optimal donors leads to only modest improvements in $\mathrm{LSTM_{Clim+Dnr}}$ performance compared to using estimated donors, with the most noticeable improvements observed at the regulated test sites. Similar conclusions are drawn in Fig. S7 for all donor-based models, which re-evaluates the results of Fig. 4 using optimal donors. Fig. 5 also reinforces the results from Fig. 3, showing that highly correlated optimal donors ($\rho_s > 0.8$) are more readily available for natural gauges. In contrast, regulated gauges often have optimal donors with weaker correlations, typically below 0.8 and sometimes even below 0.7.

A clear relationship emerges between donor quality and model performance. When the highest optimal donor correlation is 0.8 or below, the NSE values for $\mathrm{LSTM_{Clim+Dnr}}$ rarely reach the levels achieved with correlations of 0.9 or higher (represented by dark blue-colored points in Fig. 5). This demonstrates that high-quality donor correlations are critical for maximizing the predictive skill of donor-informed models. Similar patterns are observed for $\mathrm{LSTM_{Dnr}}$ in Fig. S8, which mirrors the results of $\mathrm{LSTM_{Clim+Dnr}}$ in Fig. 5.

Further analysis focuses on evaluating the conditions under which donor-informed models underperform relative to $\mathrm{LSTM_{Clim}}$. To do this, Fig. 6 compares the NSEs of $\mathrm{LSTM_{Clim+OptDnr}}$ (using optimal donors) and $\mathrm{LSTM_{Clim}}$ across the four evaluation groups. The use of optimal donors for $\mathrm{LSTM_{Clim+OptDnr}}$ in this comparison eliminates uncertainties stemming from estimated donors. The Spearman's rank correlation of the highest-correlated donor is again used to categorize results by color. Also, the comparison of NSE distributions across all models using optimal donors is provided in Fig. S7.

The results in Fig. 6 reveal that $\mathrm{LSTM_{Clim+OptDnr}}$ largely outperforms $\mathrm{LSTM_{Clim}}$ when donor correlations exceed 0.9. For sites where the highest donor correlations fall between 0.8–0.9, $\mathrm{LSTM_{Clim+OptDnr}}$ often outperforms $\mathrm{LSTM_{Clim}}$, but there are instances when $\mathrm{LSTM_{Clim}}$ substantially outperforms $\mathrm{LSTM_{Clim+OptDnr}}$. This becomes even more common when the highest donor correlations fall between 0.7–0.8. At regulated sites with very low donor correlations ($\rho_s < 0.7$), $\mathrm{LSTM_{Clim+OptDnr}}$ struggles, producing highly inaccurate predictions (as evidenced by the red-colored circles in Fig. 6b and 6d, where NSE values often fall below zero). In these challenging conditions, even $\mathrm{LSTM_{Clim}}$, which relies solely on local climate inputs, fails to achieve high predictive skill, with NSE values often lower than those of $\mathrm{LSTM_{Clim+OptDnr}}$. This highlights that local climate inputs alone are insufficient for challenging regulated sites. Again, similar patterns are observed for $\mathrm{LSTM_{OptDnr}}$ in Fig. S9, which mirrors the results of $\mathrm{LSTM_{Clim+OptDnr}}$ in Fig. 6.

Together, the results in Figs. 5 and 6 (and Figs. S7–S9) underscore two important points: 1) the MLP donor selection algorithm is generally effective at identifying high-quality donors in most cases, such that the performance of donor-based models is similar when using estimated or optimal donors; and 2) the availability of high-quality donors is a more significant limiting factor in improving the predictive accuracy of donor-based models than our ability to estimate correlations and select donors.

To complement the results above, Fig. 7 displays the spatial distribution of NSE values for daily streamflow predictions across 141 training sites (2011–2017) and 71 test sites (2001–2017). Across the training gauges, all LSTM models demonstrate high accuracy in predicting daily streamflow, consistent with the findings discussed earlier. However, the maps reveal the locations of a few training sites where model performance is poor (NSE < 0.5, shown in reddish colors). For instance, all models perform poorly at two training sites – one in the Lake Superior watershed and another in the Lake Michigan watershed – with NSE values below 0.5.

Table 2
Performance metrics (median values) for all models across evaluation groups. Metrics for percentage bias (PBIAS, FHV, FLV) are reported as absolute values. Best-performing models are highlighted in bold.

Model	Metrics									
	NSE	KGE	PBIAS	FHV	FLV	NSE	KGE	PBIAS	FHV	FLV
	Natural Train Gauges (2011–2017)					Regulated Train Gauges (2011–2017)				
$\mathrm{LSTM_{Clim}}$	0.78	0.81	7.46	22.63	26.13	0.80	0.82	6.68	22.05	19.17
$\mathrm{LSTM_{Dnr}}$	0.83	0.83	5.89	17.30	21.96	0.79	0.79	6.90	25.64	19.42
$\mathrm{LSTM_{Clim+Dnr}}$	**0.89**	**0.89**	**4.83**	**12.19**	16.47	**0.86**	**0.87**	**4.83**	**14.86**	16.23
$\mathrm{LSTM_{Avg}}$	0.86	0.82	7.60	21.40	**15.32**	0.84	0.80	6.89	23.60	**16.00**
$\mathrm{NN_{DR}}$	0.68	0.69	11.86	23.13	40.20	0.51	0.59	11.65	22.97	39.54
	Natural Test Gauges (2001–2017)					Regulated Test Gauges (2001–2017)				
$\mathrm{LSTM_{Clim}}$	0.75	**0.79**	8.95	19.72	36.72	**0.71**	**0.72**	**10.82**	**21.36**	**26.38**
$\mathrm{LSTM_{Dnr}}$	0.77	0.73	13.99	22.16	41.54	0.53	0.60	20.62	31.80	43.22
$\mathrm{LSTM_{Clim+Dnr}}$	**0.79**	0.77	10.36	**19.05**	39.71	0.65	0.64	15.58	22.00	30.60
$\mathrm{LSTM_{Avg}}$	**0.79**	0.78	**8.54**	19.85	30.59	0.67	0.67	14.47	26.31	28.51
$\mathrm{NN_{DR}}$	0.52	0.66	16.32	19.73	**30.05**	0.36	0.41	19.31	35.21	47.53

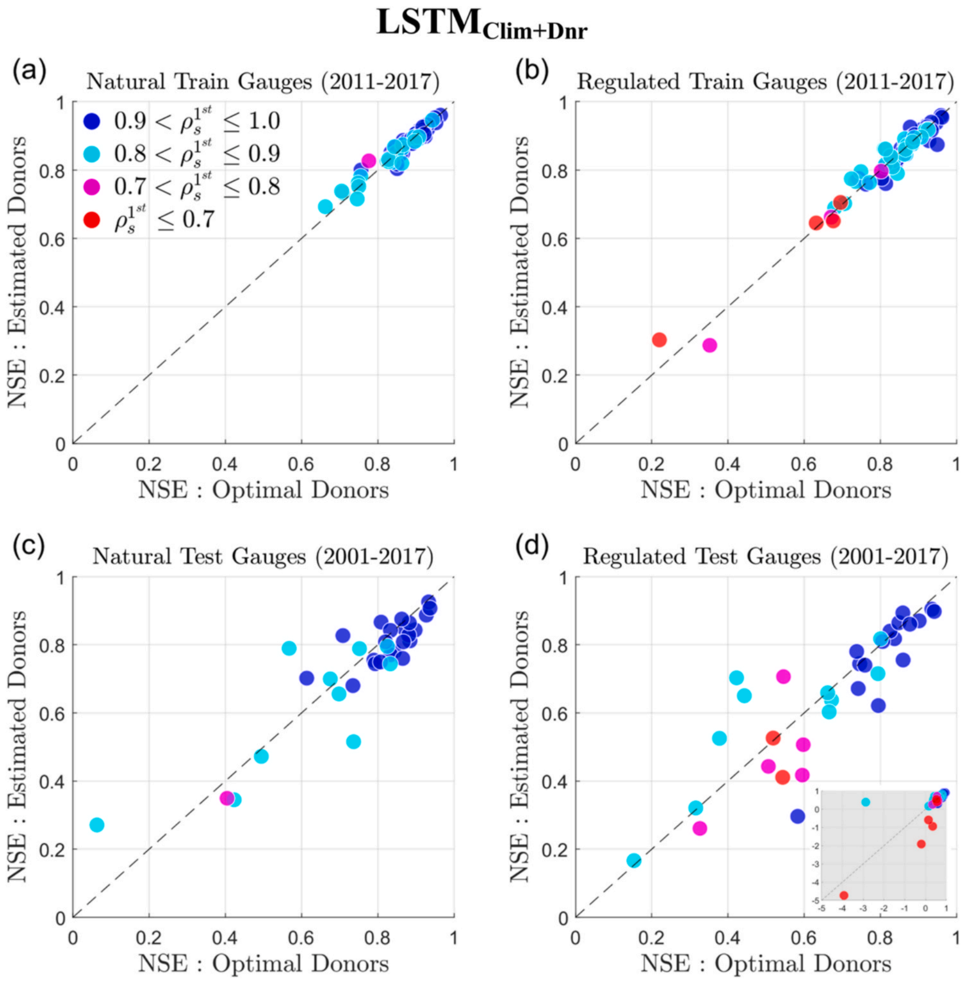


Fig. 5. Comparison of the Nash Sutcliffe Efficiency (NSE) at target sites for $\text{LSTM}_{\text{Clim+Dnr}}$ when informed by optimal donors versus estimated donors. For each target site, the highest Spearman's rank correlation of the optimal donors ($\rho_s^{1^{st}}$) is visually represented with different colors. The inset in panel (d) displays the full range of NSE values for that group of target sites.

More importantly, the maps highlight the spatial distribution of performance for fully ungauged test sites, where predictions are more challenging. A clear lake-wide pattern emerges: all LSTM models perform better at fully ungauged sites in the Lake Huron, Lake Michigan, Lake Erie, and Lake Ontario watersheds (with few sites showing NSE < 0.5) than in the Lake Superior and Ottawa River basins. The Ottawa River basin, in particular, presents the greatest challenge, as it contains the highest proportion of test sites with NSE values below 0.5 across all models.

The poor performance of donor-informed models, such as LSTM_{Dnr} and $\text{LSTM}_{\text{Clim+Dnr}}$, in the Ottawa River basin is noteworthy. This basin has only seven training gauges, all of which are relatively small (drainage areas between 258–3,811 km^2) and most located in the south. In contrast, many of the test sites in the basin are larger (drainage areas between 246–90,900 km^2) and situated further north. These differences (among others) made it difficult to identify hydrologically relevant donors with high correlations ($\rho_s > 0.9$) for the test sites. Consequently, these models underperform relative to $\text{LSTM}_{\text{Clim}}$, which relies solely on local climate data.

Hydrographs from selected test sites (Fig. 8) illustrate these dynamics. At regulated sites such as 02LC008 and 02LC029 in the Ottawa River basin (Fig. 8a and 8b; also see Fig. S10), donor-informed models failed to accurately predict peak flow events during the snowmelt season in terms of timing and magnitude. For 02LC008, this failure resulted from poorly correlated donor gauges, despite alignment between estimated and optimal donors (see Table S1). At 02LC029, both poor donor correlations and misalignment between estimated and optimal donors contributed to model underperformance (see Fig. 8b and Table S1). Conversely, in natural basins located in the Lake Erie and Huron watersheds, high-quality donors enabled donor-informed models to outperform climate-only models (Fig. 8c and 8d; also see Fig. S10). Notably, donor-based models (particularly LSTM_{Dnr}) significantly surpassed $\text{LSTM}_{\text{Clim}}$ and avoided overestimation of peak flow events.

4.4. Lake-wide monthly runoff estimate and error

The lake-wide runoff predictions for six major watersheds – Superior (SUP), Huron (HUR), Michigan (MIC), Erie (ERI), Ontario (ONT), and Ottawa (OTT) – were assessed using aggregated monthly runoff estimates derived from gauge-level predictions. These estimates were then compared against observed data, also aggregated to monthly values and summed across gauge-level observations. Results are shown in Fig. 9. Notably, the observed data in Fig. 9 are shown only when data were available for all gauges in a given lake watershed. This restriction limits the number of lake-wide observations available for comparison with the model-based estimates.

A clear discrepancy emerges between observations and predictions from NN_{DR} across several of the Great Lakes watersheds. Among all models, NN_{DR} demonstrates the poorest performance in predicting lake-wide monthly runoff. This aligns with earlier findings, where NN_{DR} showed the lowest predictive skill for daily streamflow at individual gauges. The accumulation of these poor individual predictions

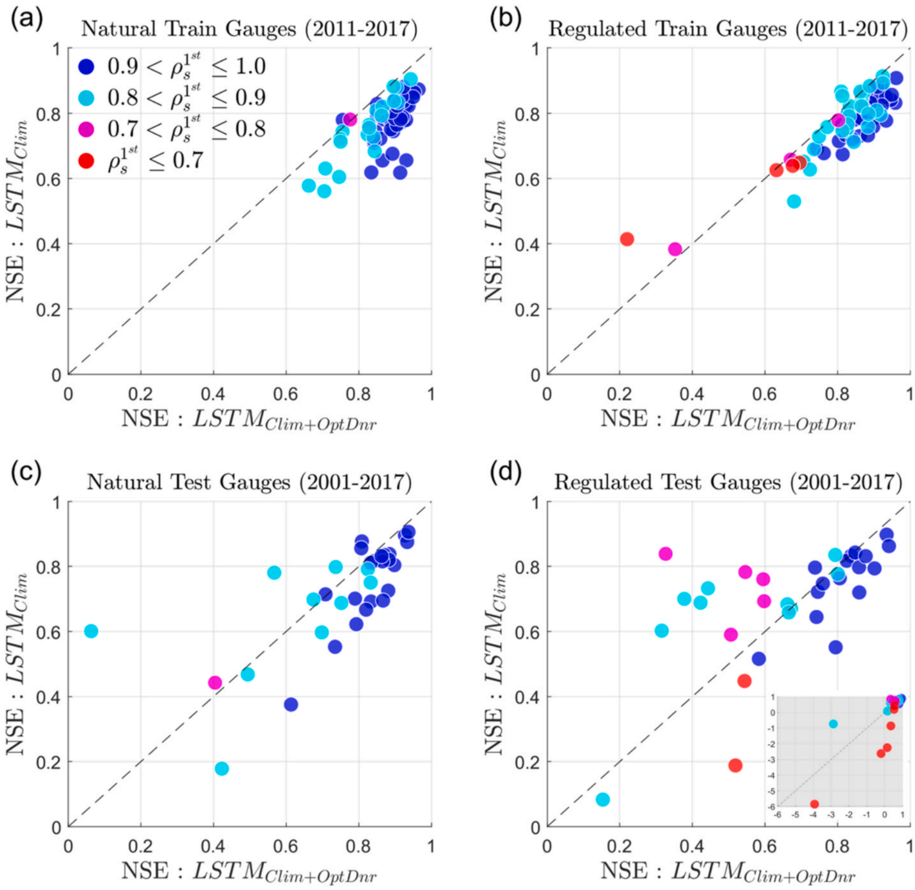


Fig. 6. Comparison of the Nash Sutcliffe Efficiency (NSE) between $\text{LSTM}_{\text{Clim+OptDnr}}$ (with optimal donors, denoted as OptDnr) and $\text{LSTM}_{\text{Clim}}$ at target sites. Similar to Fig. 5, the highest Spearman's rank correlation of the optimal donors ($\rho_s^{1^{st}}$) is represented with color coding for each target site. The inset in panel (d) highlights the full range of NSE values for that group of target sites.

ultimately degrades NN_{DR}'s monthly lake-wide runoff estimates. Despite its general underperformance, NN_{DR}'s predictions for HUR, MIC, and ERI watersheds appear closer to those of other LSTM models. However, for SUP, ONT, and OTT, NN_{DR}'s performance is distinctly worse, largely due to its consistent overestimation of monthly runoff.

In contrast, the discrepancies observed among LSTM models during daily runoff predictions at individual gauges diminish when aggregated to the monthly scale. Across all watersheds, it is challenging to distinguish the performance of different LSTM models visually, except for the Ottawa River Basin (OTT). For OTT, monthly runoff predictions vary more noticeably among the LSTM models, with $\text{LSTM}_{\text{Clim}}$ showing a tendency to estimate lower monthly runoff relative to the other donor-informed LSTMs. Limited observations in OTT indicate that LSTM_{Dnr} performs better at capturing peak monthly runoff compared to other LSTM models. In general, the LSTM models yield highly accurate lake-wide monthly runoff predictions for HUR, MIC, ERI, and ONT, closely matching observed data. However, their predictive capabilities are more limited for SUP, particularly during the dry season, as evidenced by their underestimation of runoff during this period (Fig. 9).

The poorer performance of all LSTM models in SUP and OTT watersheds can be attributed to spatial variations in predictive skill, as revealed in Fig. 7. Across all models, the test gauges in SUP and OTT show the weakest performance, which subsequently affects the lake-wide monthly runoff estimates. For SUP, the relatively sparse distribution of training gauges, despite the watershed's large size, may contribute to poor predictions at ungauged sites. A similar issue arises in OTT, where only seven training gauges are available, most of which are concentrated in the south. This sparse distribution poses challenges for donor-informed LSTMs, which rely on high-quality donor gauges with strong hydrological correlations to the target sites. However, the $\text{LSTM}_{\text{Clim}}$ also struggled the most in these regions.

Table 3 provides a detailed analysis of error variances in the lake-wide monthly runoff estimates for each watershed (described in Section 3.4), presented as standard deviations. Note that these lake-wide error standard devation estimates are derived from aggregating the monthly error variances and covariances at and between individual sites, allowing them to better account for missing observations across sites compared to the results shown in Fig. 9. Consistent with the time series analysis, NN_{DR} exhibits the highest error standard deviation among all five models, by a considerable margin, across all watersheds. In comparison, the LSTM models exhibit significantly lower error standard deviations. Among these, $\text{LSTM}_{\text{Clim+Dnr}}$ consistently achieves some of the lowest error standard deviation values across all six watersheds. Although the differences among the LSTM models are subtle, the standard deviation of error for LSTM_{Dnr} is notably lower than that of $\text{LSTM}_{\text{Clim}}$ in all watersheds except ONT. Additionally, the error standard deviation of LSTM_{Dnr} is comparable to that of LSTM_{Avg}. Across all models, the largest error standard deviation occurs in OTT, followed by SUP, as anticipated from the time series analysis of monthly runoff predictions. Overall, these findings emphasize the potential of advanced LSTM models to improve monthly water balance predictions for the Great Lakes watersheds.

4.5. Model interpretability

To facilitate the interpretability of $\text{LSTM}_{\text{Clim+Dnr}}$, we employed the Integrated Gradients (IG) method (Sundararajan et al., 2017) to quantify the contribution of each input feature to daily streamflow predictions.

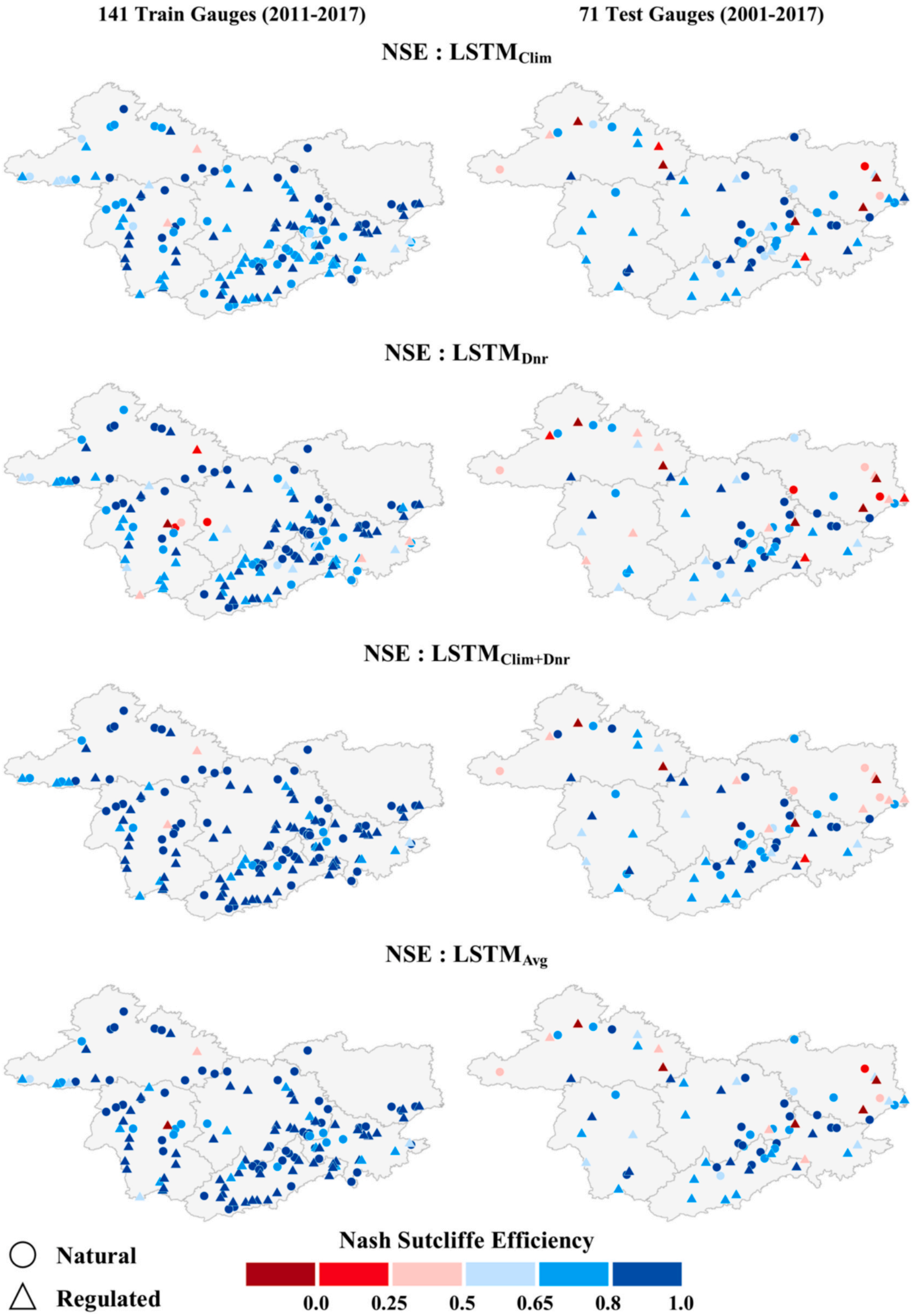


Fig. 7. Nash Sutcliffe Efficiency (NSE) of LSTM daily streamflow predictions across 141 training sites for the period 2011–2017 (left column) and 71 testing sites for the period 2001–2017 (right column).

Average attribution scores across all sites and days (Fig. S11) reveal that precipitation is the most important individual feature overall, although donor streamflow information—particularly from the most correlated donors—plays a significant role in enhancing predictions. To investigate climate and donor streamflow attributions in more detail, we compare the ratio of IG-based feature attributions between precipitation and donor streamflow to the correlation values associated with those donors (Fig. 10). As the ratios of IG attribution scores decrease, this suggests more importance is being placed on donor streamflow relative to precipitation for model prediction. Fig. 10 shows a consistent negative trend across all five donors, suggesting that as donor correlations increase, the relative influence of donor streamflow increases. This trend

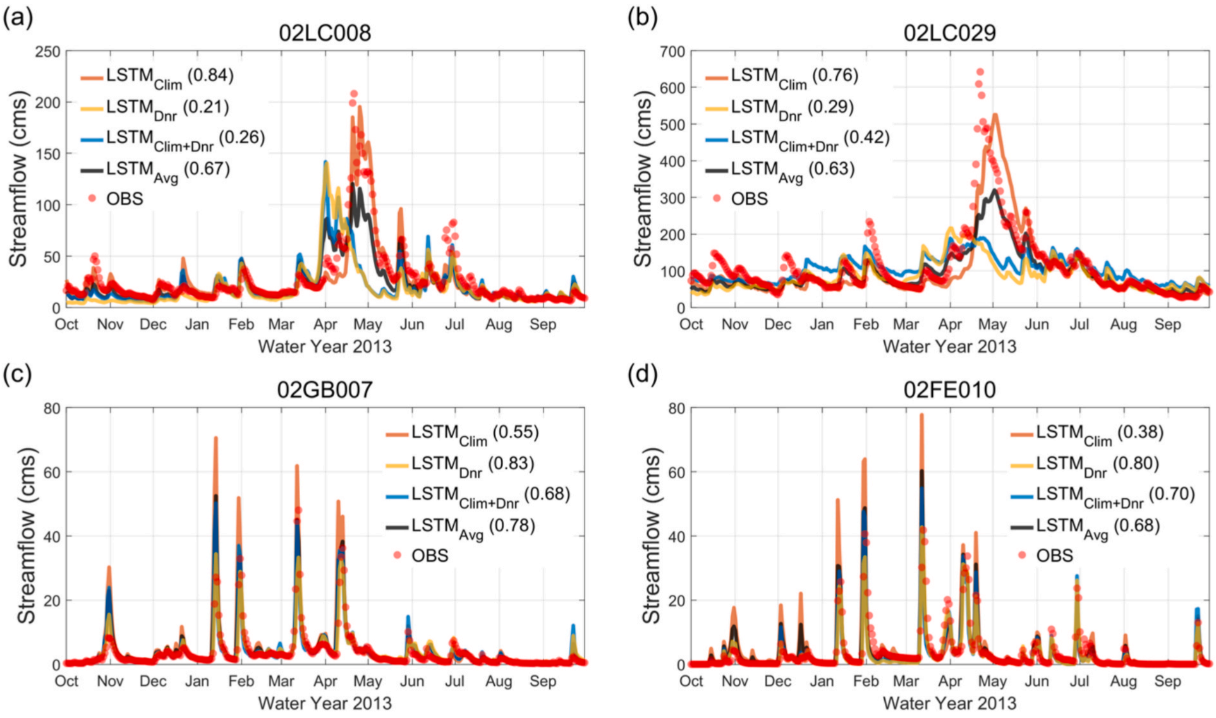


Fig. 8. Example hydrographs for two regulated test sites (a and b) and two natural target test sites (c and d) for the water year of 2013. NSE values for each model are noted in parentheses. Site locations are marked in Fig. S10.

indicates that the model is appropriately weighing donor information based on its hydrologic relevance.

The use of donor streamflow in the LSTM$_{Clim+Dnr}$ model is not static through time. Rather, it varies seasonally and with hydrologic condition. Fig. 11 illustrates the daily attribution scores for precipitation and the first donor's streamflow, averaged across all 212 basins for each day of the year. These results reveal the temporal dynamics of feature importance over the annual cycle and highlight that donor streamflow contributions peak during the spring snowmelt season (March–May). This might indicate that donor site streamflow, which integrates the melt of snowpack that has accumulated over the last several months, represents information that is particularly useful to transfer to target sites. Conversely, contributions from precipitation modestly decline in the winter and summer months, when precipitation is less likely to lead to immediate runoff responses because of snow processes and depleted soil moisture storage, respectively. These nuanced temporal patterns illustrate the model's ability to adjust its internal weighting of climate and flow inputs in ways that reflect seasonal hydrologic processes across a diverse set of watersheds. Collectively, these findings support the notion that the LSTM$_{Clim+Dnr}$ model not only improves predictive accuracy but also aligns with hydrologic process understanding, offering a path forward for explainable deep learning in hydrology (Xu et al., 2024).

5. Discussion and conclusion

This study highlights the potential of integrating local climate data with donor gauge streamflow measurements using deep learning models to enhance streamflow reconstructions in ungauged and partially gauged basins. Several DL models were developed and compared across natural and regulated sites in the Great Lakes region, including those that use climate data (LSTM$_{Clim}$), donor streamflow data (LSTM$_{Dnr}$), or both (LSTM$_{Clim+dnr}$; LSTM$_{Avg}$). The integrated LSTM$_{Clim+Dnr}$ model consistently outperformed single-source models (LSTM$_{Clim}$ or LSTM$_{Dnr}$) in basins with high-quality donor gauges. This improvement underscores the value of leveraging donor streamflow information when strong hydrological correlations exist between target and donor basins. By weighting inputs from both data sources, LSTM$_{Clim+Dnr}$ effectively balances local and regional hydrological signals, achieving superior performance in temporal extrapolation at partially gauged basins and predictions at natural, ungauged sites. Similar benefits were observed using LSTM$_{Avg}$, although this requires the development and averaging of two separate models.

The success of donor-informed models (LSTM$_{Clim+Dnr}$ and LSTM$_{Dnr}$) hinges on two key factors: availability of high-quality donors and the ability to select those donors based on estimated rank correlations. Poor donor quality, reflected in low observed correlations with target sites, had a greater impact on predictive accuracy than donor misalignment caused by suboptimal selection. Even when optimal donors were identified, the improvements over using estimated donors were generally small, underscoring that the availability of highly correlated donors is critical. This condition was observed more frequently in basins with denser gauging networks (Lake Huron, Michigan, Erie, and Ontario), as compared to the Lake Superior and Ottawa River basins, where available donors were sparse.

Importantly, this study identified a threshold for donor usefulness: when Spearman's rank correlations between donor and target basins exceeded 0.9, donor-informed models consistently outperformed climate-only models. Below this threshold, the advantage of integrating donor data diminished, and in some cases, performance deteriorated. For regulated testing sites in particular, the selection of poorly correlated donors – due to a lack of suitable donors and significant overestimations of $\widehat{\rho}_s$ – may mislead models that rely on donor information (i.e., LSTM$_{Dnr}$ and LSTM$_{Clim+Dnr}$), potentially degrading predictive performance. Consequently, LSTM$_{Clim}$ proved to be the optimal model in scenarios where donor quality was questionable.

One of the study's key practical contributions is the demonstrated ability of LSTM$_{Clim+Dnr}$ to enhance lake-wide monthly runoff estimates, a crucial metric for quantifying the Great Lakes water balance. The model reduced error variance in monthly runoff estimates compared to single-source models across all six Great Lakes watersheds. These improved runoff estimates could be used as inputs to statistical models designed to resolve the full water balance of the Great Lakes (Do et al.,

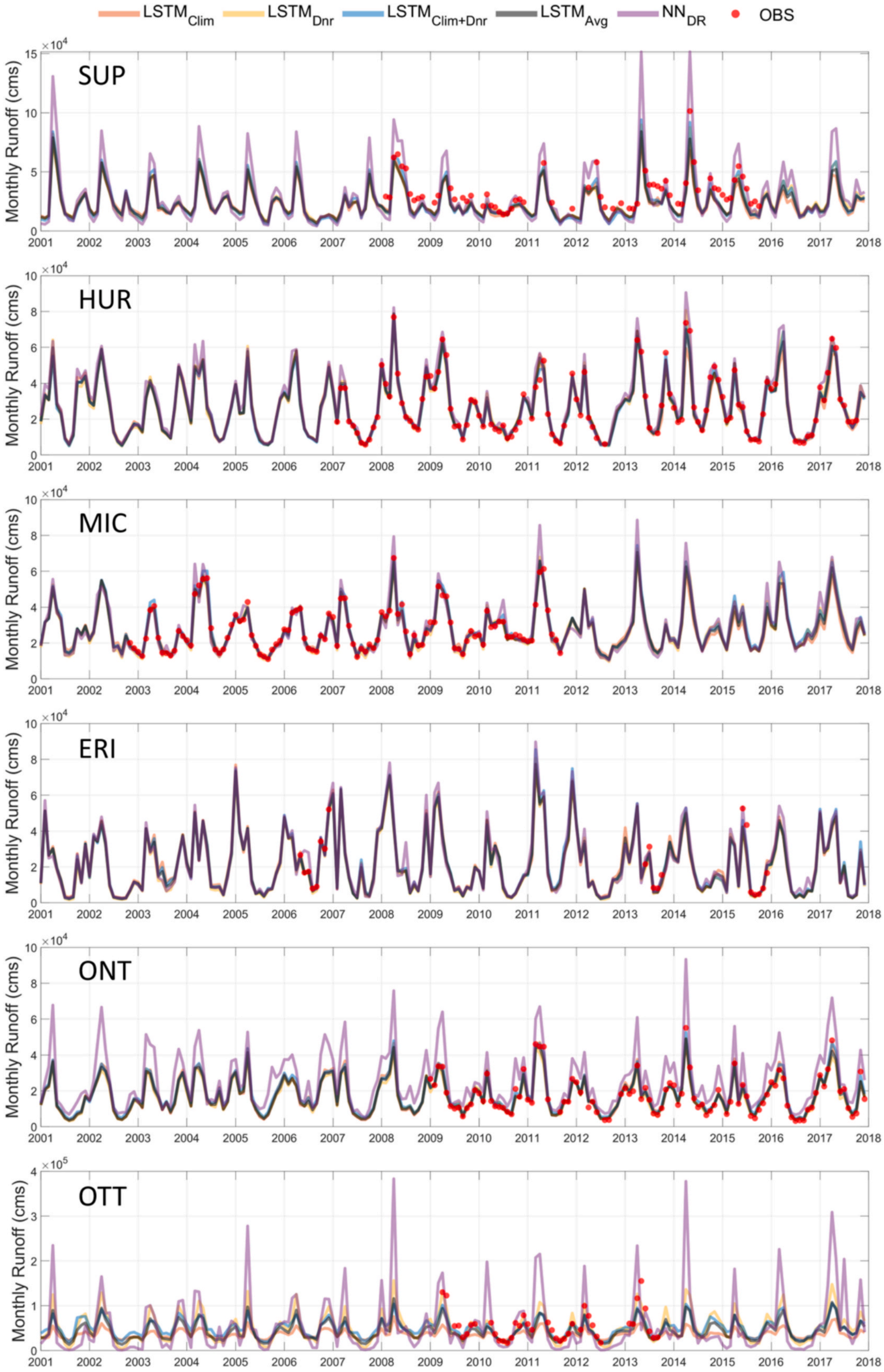


Fig. 9. Lake-wide monthly runoff predictions for six Great Lakes watersheds: Superior (SUP), Huron (HUR), Michigan (MIC), Erie (ERI), Ontario (ONT), and Ottawa (OTT).

Table 3
Standard deviation of lake-wide monthly runoff error; unit is 1000 cms. The lowest error standard deviation for each watershed is highlighted in bold.

	SUP	HUR	MIC	ERI	ONT	OTT
LSTM$_{\text{Clim}}$	4.23	3.33	2.57	2.84	1.83	23.25
LSTM$_{\text{Dnr}}$	4.06	2.10	1.74	2.51	2.46	19.50
LSTM$_{\text{Clim+Dnr}}$	**3.26**	**1.76**	**1.44**	**1.93**	**1.81**	**16.26**
LSTM$_{\text{Avg}}$	3.45	2.35	1.82	2.38	1.93	21.22
NN$_{\text{DR}}$	11.01	3.13	5.34	3.20	6.60	51.23

2020), thereby improving the accuracy of historical estimates of other water balance terms (e.g., over-lake precipitation and evaporation). Such improvements, particularly if extended back several decades, could help managers in the Great Lakes better understand emerging trends in the Great Lakes water balance, which has significant implications for the trajectories of long-term water levels (Gronewold et al., 2021).

Despite these successes, several limitations warrant further investigation. One area of improvement involves expanding the training datasets to include a broader range of regulated basins. Addressing the challenges posed by regulated sites requires incorporating more training examples with low-correlation donor-target pairs. The limited training set of 141 gauges used in this study, with only 7 regulated gauges supported by optimal donors with correlations below 0.8, likely constrained the model's ability to generalize in these situations. Expanding the training dataset to include additional regulated basins with low-correlation donors would enhance model robustness and generalizability. Similarly, expanding the methodology to other geographic regions with diverse climatic and hydrological conditions, such as the CAMELS basins across the contiguous US (Addor et al., 2017), could further improve its generalizability and help identify region-specific

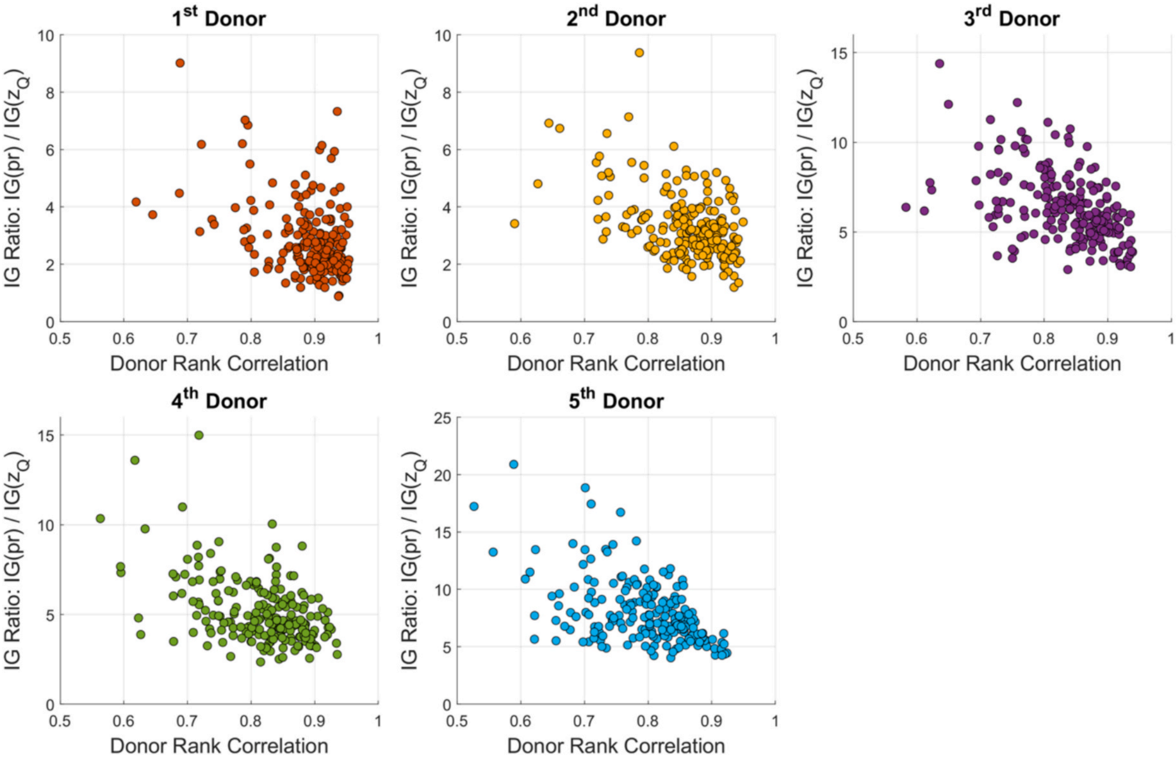


Fig. 10. The relationship between donor correlation and the ratio of IG-based attribution scores for precipitation and donor streamflow across 212 Great Lakes basins for each of five donors. The ratios being presented are averaged across all days of record for each site. Lower ratios indicate greater influence of donor streamflow relative to precipitation.

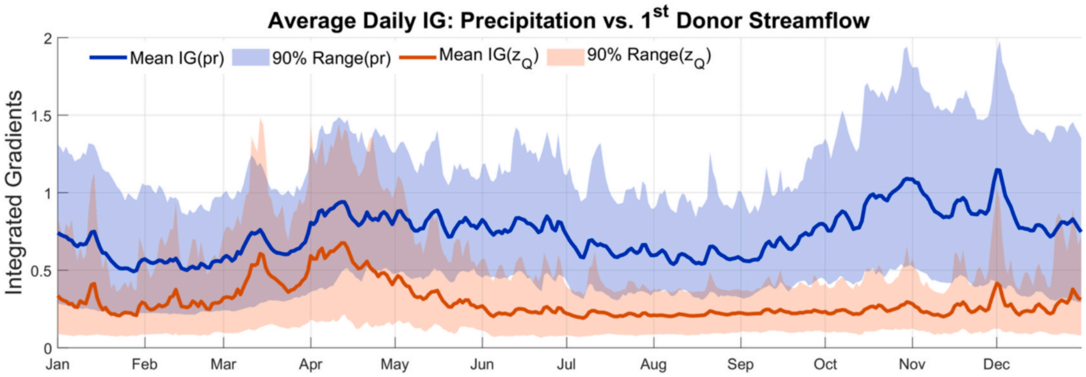


Fig. 11. Average day-of-year, IG-based attribution scores for precipitation and streamflow from the first (i.e., most correlated) donor for the LSTM$_{\text{Clim+Dnr}}$ model, computed over the 2001–2017 simulation period. Solid lines indicate daily mean attribution scores across all 212 Great Lakes basins; shaded bands represent the 90 % inter-basin range.

factors influencing performance.

The donor selection process presents another area for refinement (Villalba et al., 2021; Wang et al., 2013). Future work could explore alternative learning frameworks, such as graph-based models (Villalba et al., 2021), to further improve donor selection in human-impacted basins. Similarly, the performance of donor-informed LSTM models may benefit from the inclusion of additional similarity metrics, including those that reflect the type and degree of anthropogenic activity (e.g., regulation indices, land use intensity, network connectivity) (Ouyang et al., 2021; Tursun et al., 2024a). Including regulation-specific inputs could help the model learn when to downweight or decouple unreliable donor information, improving performance in regulated environments (Tursun et al., 2024b). Additionally, the use of remotely sensed runoff data, even if biased and limited in length (e.g., SWOT; (Fu et al., 2024), could significantly improve the estimation of rank correlations between donor and target sites at both natural and regulated sites.

Uncertainty quantification is another critical area for future research (S. Liu et al., 2023). This study did not explicitly address uncertainties in model structure and input data, such as errors in meteorological forcings or streamflow measurements, nor did it attempt to estimate the uncertainty in reconstructed streamflow values. Adopting techniques like ensemble-based methods (Li et al., 2022) or Monte Carlo dropout (Klotz et al., 2022) for this purpose could enhance the hydrological predictions by highlighting when their uncertainty is too high for practical use.

Ultimately, the methods introduced in this work – integrating local climate and regional hydrologic signals in a deep learning framework – mark a significant advancement in addressing the Prediction in Ungauged Basins challenge. These methods provide enhanced accuracy for site-specific predictions and large-scale water balance assessments. The results have important implications for water resource management, which relies on historical streamflow reconstructions across large regions to support decision-making in areas such as reservoir management, flood forecasting, and climate change signal detection and adaptation (Kayastha et al., 2022; O'Brien et al., 2024). As the class of DL models presented in the work is further refined, there is significant potential for applying it across large regions with streamflow gauging networks of moderate density, ultimately creating a state-of-the-art daily streamflow reconstruction product to support a wide range of water resource studies.

CRediT authorship contribution statement

Sungwook Wi: Writing – original draft, Validation, Methodology, Data curation, Writing – review & editing, Visualization, Software, Investigation, Conceptualization, Resources, Formal analysis. **Rohini Gupta:** Methodology, Validation, Resources, Writing – review & editing, Formal analysis, Software. **Scott Steinschneider:** Validation, Resources, Methodology, Conceptualization, Writing – original draft, Supervision, Investigation, Formal analysis, Writing – review & editing, Project administration, Funding acquisition.

Declaration of competing interest

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

Acknowledgements

This research was supported by the U.S. National Science Foundation (NSF) through the Global Center for Climate Change and Transboundary Waters (GCTW) (award no. 2330317), with partial support also provided through NSF award no. 2144332.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2025.133764.

Data availability

The code and data used for this study is available at https://doi.org/10.5281/zenodo.14610023. All data used to train and evaluate the models are available at https://doi.org/10.20383/103.0598.

References

Addor, N., Newman, A.J., Mizukami, N., Clark, M.P., 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrol. Earth Syst. Sci.

Alizadeh, B., Ghaderi Bafti, A., Kamangir, H., Zhang, Y., Wright, D.B., Franz, K.J., 2021. A novel attention-based LSTM cell post-processor coupled with bayesian optimization for streamflow prediction. J. Hydrol. 601, 126526. https://doi.org/10.1016/j.jhydrol.2021.126526.

Archfield, S.A., Vogel, R.M., 2010. Map correlation method: selection of a reference streamgage to estimate daily streamflow at ungaged catchments. Water Resour. Res. 46 (10), 2009WR008481. https://doi.org/10.1029/2009WR008481.

Arsenault, R., Brissette, F.P., 2014. Continuous streamflow prediction in ungauged basins: the effects of equifinality and parameter set selection on uncertainty in regionalization approaches. Water Resour. Res. 50 (7), 6135–6153. https://doi.org/10.1002/2013WR014898.

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 (1), 139–157. https://doi.org/10.5194/hess-27-139-2023.

Blöschl, G., Bierkens, M.F.P., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J.W., McDonnell, J.J., Savenije, H.H.G., Sivapalan, M., Stumpp, C., Toth, E., Volpi, E., Carr, G., Lupton, C., Salinas, J., Széles, B., Viglione, A., Aksoy, H., Zhang, Y., 2019. Twenty-three unsolved problems in hydrology (UPH) – a community perspective. Hydrol. Sci. J. 64 (10), 1141–1158. https://doi.org/10.1080/02626667.2019.1620507.

Choubin, B., Solaimani, K., Rezanezhad, F., Habibnejad Roshan, M., Malekian, A., Shamshirband, S., 2019. Streamflow regionalization using a similarity approach in ungauged basins: application of the geo-environmental signatures in the Karkheh River Basin Iran. Catena 182, 104128. https://doi.org/10.1016/j.catena.2019.104128.

Dasarathy, B.V., 1997. Sensor fusion potential exploitation-innovative architectures and illustrative applications. Proc. IEEE 85 (1), 24–38. https://doi.org/10.1109/5.554206.

de Lavenne, A., Skøien, J.O., Cudennec, C., Curie, F., Moatar, F., 2016. Transferring measured discharge time series: Large-scale comparison of Top-kriging to geomorphology-based inverse modeling. Water Resour. Res. 52, 5555. https://doi.org/10.1002/2016WR018716.

Do, H.X., Smith, J.P., Fry, L.M., Gronewold, A.D., 2020. Seventy-year long record of monthly water balance estimates for Earth's largest lake system. Sci. Data 7 (1), 276. https://doi.org/10.1038/s41597-020-00613-z.

Fang, K., Shen, C., 2020. Near-real-time forecast of satellite-based soil moisture using long short-term memory with an adaptive data integration kernel. J. Hydrometeorol. 21 (3), 399–413. https://doi.org/10.1175/JHM-D-19-0169.1.

Farmer, W.H., Vogel, R.M., 2013. Performance-weighted methods for estimating monthly streamflow at ungauged sites. J. Hydrol. 477, 240–250. https://doi.org/10.1016/j.jhydrol.2012.11.032.

Feng, D., Beck, H., De Bruijn, J., Sahu, R.K., Satoh, Y., Wada, Y., Liu, J., Pan, M., Lawson, K., Shen, C., 2024. Deep dive into hydrologic simulations at global scale: Harnessing the power of deep learning and physics-informed differentiable models (δ HBV-globe1.0-hydroDL). Geosci. Model Dev. 17 (18), 7181–7198. https://doi.org/10.5194/gmd-17-7181-2024.

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 (9), e2019WR026793. https://doi.org/10.1029/2019WR026793.

Fennessey, N.M., 1994. A Hydro-Climatological Model of Daily Streamflow for the Northeastern United States. Tufts University [Doctoral dissertation].

Frame, J.M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L.M., Gupta, H.V., Nearing, G.S., 2022. Deep learning rainfall–runoff predictions of extreme events. Hydrol. Earth Syst. Sci. 26 (13), 3377–3392. https://doi.org/10.5194/hess-26-3377-2022.

Fry, L.M., Gronewold, A.D., Seglenieks, F., Minallah, S., Apps, D., Ferguson, J., 2022. Navigating Great Lakes hydroclimate data. Front. Water 4, 803869. https://doi.org/10.3389/frwa.2022.803869.

Fu, L., Pavelsky, T., Cretaux, J., Morrow, R., Farrar, J.T., Vaze, P., Sengenes, P., Vinogradova-Shiffer, N., Sylvestre-Baron, A., Picot, N., Dibarboure, G., 2024. The surface water and ocean topography mission: a breakthrough in radar remote sensing of the ocean and land surface water. Geophys. Res. Lett. 51 (4), e2023GL107652. https://doi.org/10.1029/2023GL107652.

Gasset, N., Fortin, V., Dimitrijevic, M., Carrera, M., Bilodeau, B., Muncaster, R., Gaborit, É., Roy, G., Pentcheva, N., Bulat, M., Wang, X., Pavlovic, R., Lespinas, F., Khedhaouiria, D., Mai, J., 2021. A 10 km north American precipitation and land-surface reanalysis based on the GEM atmospheric model. Hydrol. Earth Syst. Sci. 25 (9), 4917–4945. https://doi.org/10.5194/hess-25-4917-2021.

Ghaneei, P., Foroumandi, E., Moradkhani, H., 2024. Enhancing streamflow prediction in ungauged basins using a nonlinear knowledge-based framework and deep learning. Water Resour. Res. 60 (11), e2024WR037152. https://doi.org/10.1029/2024WR037152.

Gronewold, A.D., Do, H.X., Mei, Y., Stow, C.A., 2021. A tug-of-war within the hydrologic cycle of a continental freshwater basin. Geophys. Res. Lett. 48 (4), e2020GL090374. https://doi.org/10.1029/2020GL090374.

Gronewold, A.D., Fortin, V., Caldwell, R., Noel, J., 2018. Resolving hydrometeorological data discontinuities along an international border. Bull. Am. Meteorol. Soc. 99 (5), 899–910. https://doi.org/10.1175/BAMS-D-16-0060.1.

Gronewold, A.D., Rood, R.B., 2019. Recent water level changes across Earth's largest lake system and implications for future variability. J. Great Lakes Res. 45 (1), 1–3. https://doi.org/10.1016/j.jglr.2018.10.012.

Guo, Q., He, Z., Wang, Z., 2024. Monthly climate prediction using deep convolutional neural network and long short-term memory. Sci. Rep. 14 (1), 17748. https://doi.org/10.1038/s41598-024-68906-6.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377 (1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.

He, M., Jiang, S., Ren, L., Cui, H., Qin, T., Du, S., Zhu, Y., Fang, X., Xu, C.-Y., 2024. Streamflow prediction in ungauged catchments through use of catchment classification and deep learning. J. Hydrol. 639, 131638. https://doi.org/10.1016/j.jhydrol.2024.131638.

Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G., Hochreiter, S., Klambauer, G., 2021. MC-LSTM: Mass-Conserving LSTM.

Hrachowitz, M., Savenije, H.H.G., Blöschl, G., McDonnell, J.J., Sivapalan, M., Pomeroy, J.W., Arheimer, B., Blume, T., Clark, M.P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H.V., Hughes, D.A., Hut, R.W., Montanari, A., Pande, S., Tetzlaff, D., Cudennec, C., 2013. A decade of predictions in ungauged basins (PUB)—a review. Hydrol. Sci. J. 58 (6), 1198–1255. https://doi.org/10.1080/02626667.2013.803183.

Hunter, T.S., Clites, A.H., Campbell, K.B., Gronewold, A.D., 2015. Development and application of a North American Great Lakes hydrometeorological database — part I: precipitation, evaporation, runoff, and air temperature. J. Great Lakes Res. 41 (1), 65–77. https://doi.org/10.1016/j.jglr.2014.12.006.

Jahangir, M.S., Quilty, J., 2024. Generative deep learning for probabilistic streamflow forecasting: Conditional variational auto-encoder. J. Hydrol. 629, 130498. https://doi.org/10.1016/j.jhydrol.2023.130498.

Jiang, S., Zheng, Y., Solomatine, D., 2020. Improving AI system awareness of geoscience knowledge: symbiotic integration of physical approaches and deep learning. Geophys. Res. Lett. 47 (13), e2020GL088229. https://doi.org/10.1029/2020GL088229.

Kayastha, M.B., Ye, X., Huang, C., Xue, P., 2022. Future rise of the Great Lakes water levels under climate change. J. Hydrol. 612, 128205. https://doi.org/10.1016/j.jhydrol.2022.128205.

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 (6), 1673–1693. https://doi.org/10.5194/hess-26-1673-2022.

Kratzert, F., Gauch, M., Klotz, D., Nearing, G., 2024. HESS opinions: never train a long short-term memory (LSTM) network on a single basin. Hydrol. Earth Syst. Sci. 28 (17), 4187–4201. https://doi.org/10.5194/hess-28-4187-2024.

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 (11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., Hochreiter, S., Nearing, G.S., 2019. Toward improved predictions in ungauged basins: exploiting the power of machine learning. Water Resour. Res. 55 (12), 11344–11354. https://doi.org/10.1029/2019WR026065.

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 (5), 2685–2703. https://doi.org/10.5194/hess-25-2685-2021.

Kult, J.M., Fry, L.M., Gronewold, A.D., Choi, W., 2014. Regionalization of hydrologic response in the Great Lakes basin: considerations of temporal scales of analysis. J. Hydrol. 519, 2224–2237. https://doi.org/10.1016/j.jhydrol.2014.09.083.

Le, M.-H., Kim, H., Do, H.X., Beling, P.A., Lakshmi, V., 2024. A framework on utilizing of publicly availability stream gauges datasets and deep learning in estimating monthly basin-scale runoff in ungauged regions. Adv. Water Resour. 188, 104694. https://doi.org/10.1016/j.advwatres.2024.104694.

Lehner, B., Verdin, K., Jarvis, A., 2008. New global hydrography derived from spaceborne elevation data. Eos Trans. AGU 89 (10), 93–94. https://doi.org/10.1029/2008EO100001.

Li, D., Marshall, L., Liang, Z., Sharma, A., 2022. Hydrologic multi-model ensemble predictions using variational Bayesian deep learning. J. Hydrol. 604, 127221. https://doi.org/10.1016/j.jhydrol.2021.127221.

Liu, J., Rahmani, F., Lawson, K., Shen, C., 2022. A multiscale deep learning model for soil moisture integrating satellite and in situ data. Geophys. Res. Lett. 49 (7), e2021GL096847. https://doi.org/10.1029/2021GL096847.

Liu, S., Lu, D., Painter, S.L., Griffiths, N.A., Pierce, E.M., 2023. Uncertainty quantification of machine learning models to improve streamflow prediction under changing climate and environmental conditions. Front. Water 5, 1150126. https://doi.org/10.3389/frwa.2023.1150126.

Ma, K., Feng, D., Lawson, K., Tsai, W., 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 (5), e2020WR028600. https://doi.org/10.1029/2020WR028600.

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., Waddell, J.W., 2022. The Great Lakes runoff intercomparison project phase 4: the Great Lakes (GRIP-GL). Hydrol. Earth Syst. Sci. 26 (13), 3537–3572. https://doi.org/10.5194/hess-26-3537-2022.

Meyal, A.Y., Versteeg, R., Alper, E., Johnson, D., Rodzianko, A., Franklin, M., Wainwright, H., 2020. Automated cloud based long short-term memory neural network based SWE prediction. Front. Water 2, 574917. https://doi.org/10.3389/frwa.2020.574917.

Milly, P.C.D., Dunne, K.A., Vecchia, A.V., 2005. Global pattern of trends in streamflow and water availability in a changing climate. Nature 438 (7066), 347–350. https://doi.org/10.1038/nature04312.

Mohamoud, Y.M., 2008. Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves. Hydrol. Sci. J. 53 (4), 706–724. https://doi.org/10.1623/hysj.53.4.706.

NALCMS (2017). NALCMS website, http://www.cec.org/north-american-environmental-atlas/land-cover-2010-landsat-30m/ (last access: 4 January 2025).

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — a discussion of principles. J. Hydrol. 10 (3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

Nearing, G.S., Klotz, D., Frame, J.M., Gauch, M., Gilon, O., Kratzert, F., Sampson, A.K., Shalev, G., Nevo, S., 2022. Technical note: data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks. Hydrol. Earth Syst. Sci. 26 (21), 5493–5513. https://doi.org/10.5194/hess-26-5493-2022.

O'Brien, N.L., Seglenieks, F., Fry, L.M., Fielder, D., Temgoua, A.G.T., Bruxer, J., Fortin, V., Durnford, D., Gronewold, A.D., 2024. Historical datasets (1950–2022) of monthly water balance components for the Laurentian Great Lakes. Sci. Data 11 (1), 1243. https://doi.org/10.1038/s41597-024-03994-7.

Ouyang, W., Lawson, K., Feng, D., Ye, L., Zhang, C., Shen, C., 2021. Continental-scale streamflow modeling of basins with reservoirs: towards a coherent deep-learning-based strategy. J. Hydrol. 599, 126455. https://doi.org/10.1016/j.jhydrol.2021.126455.

Patil, S., Stieglitz, M., 2012. Controls on hydrologic similarity: Role of nearby gauged catchments for prediction at an ungauged catchment. Hydrol. Earth Syst. Sci. 16 (2), 551–562. https://doi.org/10.5194/hess-16-551-2012.

Pokharel, S., Roy, T., 2024. A parsimonious setup for streamflow forecasting using CNN-LSTM. J. Hydroinf. 26 (11), 2751–2761. https://doi.org/10.2166/hydro.2024.114.

Pool, S., Vis, M., Seibert, J., 2021. Regionalization for ungauged catchments—lessons learned from a comparative large-sample study. Water Resour. Res. 57 (10), e2021WR030437. https://doi.org/10.1029/2021WR030437.

Pool, S., Viviroli, D., Seibert, J., 2019. Value of a limited number of discharge observations for improving regionalization: a large-sample study across the United States. Water Resour. Res. 55 (1), 363–377. https://doi.org/10.1029/2018WR023855.

Qi, W., Chen, J., Li, L., Xu, C.-Y., Xiang, Y., Zhang, S., Wang, H.-M., 2021. Impact of the number of donor catchments and the efficiency threshold on regionalization performance of hydrological models. J. Hydrol. 601, 126680. https://doi.org/10.1016/j.jhydrol.2021.126680.

Razavi, T., Coulibaly, P., 2013. Streamflow prediction in ungauged basins: review of regionalization methods. J. Hydrol. Eng. 18 (8), 958–975. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690.

Shangguan, W., Dai, Y., Duan, Q., Liu, B., Yuan, H., 2014. A global soil data set for earth system modeling. J. Adv. Model. Earth Syst. 6 (1), 249–263. https://doi.org/10.1002/2013MS000293.

Shu, C., Ouarda, T.B.M.J., 2012. Improved methods for daily streamflow estimates at ungauged sites. Water Resour. Res. 48 (2), 2011WR011501. https://doi.org/10.1029/2011WR011501.

Singh, L., Mishra, P.K., Pingale, S.M., Khare, D., Thakur, H.P., 2022. Streamflow regionalisation of an ungauged catchment with machine learning approaches. Hydrol. Sci. J. 67 (6), 886–897. https://doi.org/10.1080/02626667.2022.2049271.

Sivapalan, M., Takeuchi, K., Franks, S.W., Gupta, V.K., Karambiri, H., Lakshmi, V., Liang, X., McDONNELL, J.J., Mendiondo, E.M., O'Connell, P.E., Oki, T., Pomeroy, J. W., Schertzer, D., Uhlenbrook, S., Zehe, E., 2003. IAHS decade on predictions in ungauged basins (PUB), 2003–2012: shaping an exciting future for the hydrological sciences. Hydrol. Sci. J. 48 (6), 857–880. https://doi.org/10.1623/hysj.48.6.857.51421.

Skøien, J.O., Blöschl, G., 2007. Spatiotemporal topological kriging of runoff time series. Water Resour. Res. 43, W09419. https://doi.org/10.1029/2006WR005760.

Song, Y., Tsai, W.-P., Gluck, J., Rhoades, A., Zarzycki, C., McCrary, R., Lawson, K., Shen, C., 2024. LSTM-based data integration to improve snow water equivalent prediction and diagnose error sources. J. Hydrometeorol. 25 (1), 223–237. https://doi.org/10.1175/JHM-D-22-0220.1.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks. In Proceeding of the 34th International Conference on Machine Learning (ICML), 70, 3319–3328.

Swain, J.B., Patra, K.C., 2017. Streamflow estimation in ungauged catchments using regionalization techniques. J. Hydrol. 554, 420–433. https://doi.org/10.1016/j.jhydrol.2017.08.054.

Tursun, A., Xie, X., Wang, Y., Liu, Y., Peng, D., Zheng, B., 2024a. Enhancing streamflow simulation in large and human-regulated basins: Long short-term memory with multiscale attributes. J. Hydrol. 630, 130771. https://doi.org/10.1016/j.jhydrol.2024.130771.

Tursun, A., Xie, X., Wang, Y., Peng, D., Liu, Y., Zheng, B., Wu, X., Nie, C., 2024b. Streamflow prediction in human-regulated catchments using multiscale deep learning modeling with anthropogenic similarities. Water Resour. Res. 60 (9), e2023WR036853. https://doi.org/10.1029/2023WR036853.

Villalba, G.A., Liang, X., Liang, Y., 2021. Selection of multiple donor gauges via graphical lasso for estimation of daily streamflow time series. Water Resour. Res. 57 (5), e2020WR028936. https://doi.org/10.1029/2020WR028936.

Vörösmarty, C.J., McIntyre, P.B., Gessner, M.O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S.E., Sullivan, C.A., Liermann, C.R., Davies, P.M., 2010. Global threats to human water security and river biodiversity. Nature 467 (7315), 555–561. https://doi.org/10.1038/nature09440.

Wang, W., Hu, S., Li, Y., Cao, S., 2013. How to select a reference basin in the ungauged regions. J. Hydrol. Eng. 18 (8), 941–947. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000680.

Waseem, M., Ajmal, M., Kim, T.-W., 2015. Ensemble hydrological prediction of streamflow percentile at ungauged basins in Pakistan. J. Hydrol. 525, 130–137. https://doi.org/10.1016/j.jhydrol.2015.03.042.

Wi, S., Steinschneider, S., 2024. On the need for physical constraints in deep learning rainfall–runoff projections under climate change: a sensitivity analysis to warming and shifts in potential evapotranspiration. Hydrol. Earth Syst. Sci. 28 (3), 479–503. https://doi.org/10.5194/hess-28-479-2024.

Wilbrand, K., Taormina, R., Ten Veldhuis, M.-C., Visser, M., Hrachowitz, M., Nuttall, J., Dahm, R., 2023. Predicting streamflow with LSTM networks using global datasets. Front. Water 5, 1166124. https://doi.org/10.3389/frwa.2023.1166124.

Worland, S.C., Steinschneider, S., Asquith, W., Knight, R., Wieczorek, M., 2019a. Prediction and inference of flow duration curves using multioutput neural networks. Water Resour. Res. 55 (8), 6850–6868. https://doi.org/10.1029/2018WR024463.

Worland, S.C., Steinschneider, S., Farmer, W., Asquith, W., Knight, R., 2019b. Copula theory as a generalized framework for flow-duration curve based streamflow estimates in ungaged and partially gaged catchments. Water Resour. Res. 55 (11), 9378–9397. https://doi.org/10.1029/2019WR025138.

Wu, H., Zhang, J., Bao, Z., Wang, G., Wang, W., Yang, Y., Wang, J., 2023. Runoff modeling in ungauged catchments using machine learning algorithm-based model parameters regionalization methodology. Engineering 28, 93–104. https://doi.org/10.1016/j.eng.2021.12.014.

Xu, Y., Lin, K., Hu, C., Wang, S., Wu, Q., Zhang, L., Ran, G., 2023. Deep transfer learning based on transformer for flood forecasting in data-sparse basins. J. Hydrol. 625, 129956. https://doi.org/10.1016/j.jhydrol.2023.129956.

Xu, Y., Lin, K., Hu, C., Wang, S., Wu, Q., Zhang, J., Xiao, M., Luo, Y., 2024. Interpretable machine learning on large samples for supporting runoff estimation in ungauged basins. J. Hydrol. 639, 131598. https://doi.org/10.1016/j.jhydrol.2024.131598.

Yang, X., Magnusson, J., Rizzi, J., Xu, C.-Y., 2018. Runoff prediction in ungauged catchments in Norway: comparison of regionalization approaches. Hydrol. Res. 49 (2), 487–505. https://doi.org/10.2166/nh.2017.071.

Yilmaz, K.K., Gupta, H.V., Wagener, T., 2008. A process-based diagnostic approach to model evaluation: application to the NWS distributed hydrologic model. Water Resour. Res. 44 (9), 2007WR006716. https://doi.org/10.1029/2007WR006716.

Yuan, L.L., 2013. Using correlation of daily flows to identify index gauges for ungauged streams. Water Resour. Res. 49 (1), 604–613. https://doi.org/10.1002/wrcr.20070.