https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

Generalization of Deep Learning Models to Ungauged Glacierized Basins: Evidence from Alpine, Patagonian, and North American Catchments

Meelisha Maharjan[1], Colin J. Gleason[1], Casey Brown[1]

[1]University of Massachusetts, Amherst, 01002 Massachusetts, USA

Correspondence to: Meelisha Maharjan (mmaharjan@umass.edu)

Abstract. Glacierized high-mountain basins supply water to approximately two billion people yet remain among the most data-scarce hydrologic regions globally, making truly ungauged streamflow prediction a critical challenge. Deep learning (DL) offers a promising alternative to traditional regionalization, but fundamental questions remain about when and why DL models generalize to a target domain that is not merely ungauged but hydro-climatically distinct from the training data. We address two questions: under what training data do DL models generalize reliably to completely ungauged glacierized basins? And how does model architecture, including physics-informed DL, modulate sensitivity to these conditions? We systematically evaluate three architectures — Long Short-Term Memory networks (LSTM), Graph Neural Networks (GNN), and differentiable HBV (δHBV) — across four experiments that control for training dataset size, hydroclimatic representativeness, and inclusion of basins with glaciers, using 2,845 basins from the Caravan global dataset with 283 target glacierized basins. We perform 100-trial repeated K-fold cross-validation by holding out glacierized basins as test basins strictly in space and time. Hydroclimatic representativeness of training data- the degree to which training basins cover the target glacierized regime consistently dominates both training data size and architecture choice as the primary determinant of generalization skill. Including glacierized catchments in training provides the strongest representativeness signal, with all three architectures achieving median NSE between 0.66 and 0.71. When glacierized catchments are excluded, LSTM median NSE falls to −1.43 in the most dissimilar partition; larger dataset size only partially improves skill (median NSE −0.96), confirming that dataset size cannot substitute for representativeness. Non-glacierized mountain catchments partially improve skill, demonstrating that partial hydroclimatic representativeness – through inclusion of non-glacierized mountain basins - contributes to model performance in glacierized basins. Architecture differences are secondary: δHBV and GNN show greater resilience under data scarcity due to structural constraints, but no architecture compensates for lack of hydroclimatic representativeness in training data. These findings reframe model selection for ungauged glacierized basins, highlighting the importance of representative training data and the potential limits of "out of sample in landscape" performance of DL models, specifically for DL deployment in climate impact assessments of high-mountain water towers.

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

1 Introduction

Almost two billion people rely on water supply originating from high-mountain glacierized basins (Immerzeel et al., 2019; Viviroli et al., 2020). The relatively high altitude and slope of mountains generate orographic precipitation, which commonly takes the form of snow and ice due to lower temperature (Immerzeel et al., 2019; Viviroli and Weingartner, 2004). The existence of glaciers, snow or glacier-ice on catchments enables these catchments to function as natural reservoirs, releasing stored water during late summer long after snowmelt pulse has receded (Fountain and Tangborn, 1985; Frenierre and Mark, 2014; Jansson et al., 2003; Kaser et al., 2010; Koboltschnig and Schöner, 2011; Moore et al., 2009; Singh, 2011; Stahl and Moore, 2006). Even if glacier fraction is low, snow and glacier melt processes contribute to streamflow volume and timing variability (Casassa et al., 2009; Jansson et al., 2003). This buffering capacity of glaciated mountain basins is important for downstream regions, especially in arid and semi-arid regions as they maintain relatively constant supply during the dry and hot season (Immerzeel et al., 2019; Viviroli et al., 2003). However, with changing climate, the glaciers are accelerating their retreat (Bolch et al., 2012; Marta et al., 2021; Rabatel et al., 2013; Rounce et al., 2023; Vargo et al., 2020) accompanied by shorter snow cover duration (Ackroyd et al., 2021; Brown and Mote, 2009; Singh et al., 2016), affecting the interseasonal storage capacity of the mountains (Huss et al., 2017). Improving hydrological prediction in these regions is both scientifically pertinent and of direct societal consequence.

A prominent method for learning about the hydrology of glacierized catchments is process-based hydrological modelling (Bocchiola et al., 2010; Gurtz et al., 2003; Huss et al., 2008; Mejía-Veintimilla et al., 2019; Viviroli et al., 2009; Zappa et al., 2000; Zhang et al., 2013). Accurate simulation of runoff in a glacierized catchment requires adequate representation of the hydrological cycle to accurately simulate melt from snow and glacier in addition to the mass balance components of a non-glacierized basin (Azam et al., 2021; van Tiel et al., 2020). Various glacial components like glacial lakes, reservoirs and permafrost function as additional storage either at the surface or under the ground regulating the overall flow from the basins (Fang et al., 2018; Shafeeque et al., 2020; van Tiel et al., 2020). The accumulation and melting of snow in the glaciers are a function of various physiographic and climatic factors of the basin. Moreover, shortwave radiation flux plays a significant role, both spatially and temporally, in the accumulation and melt process (Azam et al., 2019, 2014; Litt et al., 2019; Schaner et al., 2012). Numerous methods from a simple regression approach to distributed modeling have been applied to capture these unique characteristics of glacierized basins. van Tiel et al. (2020) reviewed 145 different glacio-hydrological modeling studies – applied to a wide variety of glacierized catchments around the world and reported major challenges in representing snow accumulation and redistribution, temperature-index and energy balance melt processes, and the role of glacial lakes, reservoirs, and permafrost as additional storage components. The performance of the models evaluated depended critically on the availability of observations and measurements for their calibration and validation (van Tiel et al., 2020). Azam et al. (2021)

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

further documented major uncertainties in Himalayan glacio-hydrological modeling arising from equifinality and glacier dynamics due to data scarcity.

Although in-situ hydrologic observations began in the nineteenth century and expanded throughout the twentieth, the number of publicly available in-situ monitoring sites is decreasing due to expense and maintenance issues in high-elevation locations (Shahgedanova et al., 2021). Riggs et al. (2023) compiled a comprehensive assessment of publicly available global streamflow gauges, finding that out of 45,837 gauges, 37% are discontinued and 77% do not contain real-time data. The reluctance of some countries to share key hydrological data of and around transboundary basins, prove a further hindrance (Gleason and Hamdan, 2017; Gleason and Smith, 2014). To counter the limitation due to data unavailability, regionalization attempts have been made to transfer model parameters learnt from gauged basins to ungauged ones (Peel and Blöschl, 2011; Razavi and Coulibaly, 2013). The regionalization of these models have been accomplished by either spatial proximity, physical similarity, scaling relationships, regression methods, or hydrological signature methods (Arsenault and Brissette, 2014; Guo et al., 2021; Razavi and Coulibaly, 2013). However, a lack of consistent success and large uncertainties have limited the regionalization of traditional models to mountainous and glaciered basins (Guo et al., 2021; Nepal et al., 2017; Razavi and Coulibaly, 2013; Vinze and Azam, 2023). We are, therefore, interested in exploring alternatives methods, specifically DL-based methods, for modeling these ungauged regions.

Improvements in primary data availability have enabled exploration of ML and DL models in hydrology (Aryal et al., 2023; Hsu et al., 1995; Maier et al., 2010; Maier and Dandy, 1996, 2000; Saha and Chandra Pal, 2024; Zeng et al., 2026). These include satellite and reanalysis products, such as ERA5-Land (Muñoz-Sabater et al., 2021), together with the collation of standardized large-sample hydrologic datasets (e.g., the CAMELS benchmark catchment archives; Addor et al., 2017; Häge et al., 2023; Koch et al., 2022; Knoben et al., 2025; Loritz et al., 2024; and Caravan dataset; Färber et al., 2025; Kratzert et al., 2023). The flexibility of multi-layered DL models enable them to extract insights from complex data through higher level representations of the underlying data structures (Liu et al., 2025b; Nearing et al., 2021; Painter and Destouni, 2026; Shen, 2018; Sit et al., 2020). These data-driven models have consistently shown to perform as well as, if not better than, conventional physical and conceptual models to predict hydrological variables (Arsenault et al., 2023; Boodoo et al., 2025; Fang et al., 2017; Feng et al., 2020; Hu et al., 2018; Kratzert et al., 2018, 2019; Lees et al., 2022; Liu et al., 2025b; Rahmani et al., 2021). Furthermore, these models have successfully proven themselves capable of predicting variables of interest in ungauged regions (Arsenault et al., 2023; Kratzert et al., 2019; Nearing et al., 2024; Willard et al., 2025; Zhang et al., 2024). Kratzert et al. (2019) evaluated the extrapolating skill of an LSTM on basins that were not included in training and showed a comparable performance as the state-of-the-art physically based hydrological model. Arsenault et al. (2023) subsequently showed that LSTMs clearly outperform traditional regionalization methods in ungauged basins. Ma et al. (2021) demonstrated the power of an LSTM to transfer hydrologic knowledge from regions with dense data to scarce target regions better than locally trained LSTM models. Fang et al. (2022) showed synergistic benefits of combining data from multiple heterogeneous regions over

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

training on homogeneous local data alone. Similar performance is seen in Graph Neural Networks (GNNs) (Mosaffa et al., 2026; Sun et al., 2021).

Building on these advances, recent research has extended the use of deep learning to high mountain and alpine basins, where complex cryospheric processes pose additional modeling challenges (Bolibar et al., 2020; Chiogna et al., 2018; Ji et al., 2021; Li, 2023; Mohammadi et al., 2025; Ougahi and Rowan, 2025; Wang, 2023; Yang et al., 2023). Anderson and Radic (2022) demonstrated how a convolutional long short-term memory network (CNN-LSTM) can learn physically consistent principles of runoff generation using only temperature, precipitation, and streamflow data across three distinct regimes: glacial, nival, and pluvial. Their findings revealed that the model develops regime-specific sensitivity to changes in input fields and that cell states can be linked to basin glacier coverage, serving as indicators of glacier runoff. The model was able to differentiate between melt sources, such as glacier melt and snowmelt-driven flows, despite the absence of information on melt origin or specific glacial inputs. DL models have shown promise in capturing nonlinear glacier responses and improving representation of extreme mass balance rates (Bolibar et al., 2022; van der Meer et al., 2025), as well as identifying key climatic and topographic controls on runoff generation in glaciated watersheds (Aguayo et al., 2025a; Chen et al., 2022; Hao et al., 2024). Collectively, these findings underscore the potential of data-driven approaches to address the complexity of cryospheric processes and enhance predictive skill under data-scarce conditions without bespoke representation of process. However, a common limitation of these studies is that they either test within the same catchment region – what Gleason & Durand (2020) termed ‘semi-gauged’ evaluation, where models are trained on temporally partitioned data from the target region – or evaluate a single architecture without systematic comparison of training data conditions. The question of which DL architecture generalizes most reliably to truly ungauged glacierized catchments, and under what training data conditions, therefore remains unanswered.

In the purely data-driven literature, the strong ungauged generalization demonstrated by Kratzert et al. (2019), Arsenault et al. (2023), and Fang et al., (2021a) is achieved within broadly temperate, continental hydroclimatic pools — diversity within CAMELS is fundamentally different from including a cryospheric regime that is absent from most global training archives. Physics-informed approaches offer a potentially distinct perspective: differentiable models such as δHBV have been shown to exceed LSTM performance specifically in ungauged extrapolation, with structural constraints reducing sensitivity to training distribution mismatch (Feng et al., 2022, 2024; Ji et al., 2025). These advantages have since been extended to global and multi-forcing settings (Feng et al., 2024; Liu et al., 2025a; Song et al., 2026). Yet even these advances do not include glacierized catchments, and none systematically vary training data composition to test whether structural advantages persist when the target regime is absent from training entirely. The architecture comparison literature is similarly incomplete: Liu et al. (2025b) provide the most rigorous DL benchmarking to date — confirming LSTM remains the strongest model for daily streamflow regression — but their only ungauged test holds basins out in space while retaining temporal overlap, and is conducted entirely within CAMELS, and they explicitly identify glacierized high-mountain regions as an unaddressed gap.

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

This convergence of gaps motivates two open questions. First, in a hydroclimatic regime as distinct as glacierized high-mountain basins, how does the composition of training data — specifically the inclusion of glacierized catchments and hydroclimatic representativeness of training data relative to the target region — compare with architecture choice in determining generalization skill? Second, do physics-informed structural constraints provide sufficient resilience to compensate for missing representativeness, or does training data composition dominate regardless of architecture? These questions have direct practical consequences for data collection priorities and model selection in high-mountain water towers, and for whether DL-based daily prediction in ungauged glacierized basins can reliably serve as a foundation for downstream climate impact assessment — a pipeline that Aguayo et al. (2025) demonstrate is feasible for Patagonia but that requires the kind of systematic generalization characterization this study provides.

The aim of this study is to characterize the conditions under which three fundamentally different DL architectures — LSTM, δHBV, and GNN — generalize to ungauged glacierized high-mountain basins, by systematically evaluating the relative roles of hydroclimatic representativeness of training data, training data size, and glacierized catchment inclusion. We deliberately compare three established architectures rather than the newest attention-based or foundation models. Recent benchmarking shows that the advantage of such models over LSTM is task-dependent, emerging in long-horizon autoregressive and zero-shot forecasting rather than the daily streamflow regression addressed here, where no attention-based model meaningfully outperforms LSTM, including under spatial cross-validation (Liu et al., 2025b). Our three were instead chosen to span the dominant paradigms in data-driven hydrology — purely data-driven (LSTM), spatially structured (Graph WaveNet), and physics-informed differentiable (δHBV) modelling. We address two research questions. First: under what training data conditions — defined along axes of dataset size, hydroclimatic representativeness of the training pool, and inclusion of representative glacierized catchments — do these architectures generalize reliably to completely ungauged glacierized basins evaluated strictly out-of-sample in both space and time? Second: how does model architecture modulate sensitivity to these training data conditions, and what are the practical computational tradeoffs among architectures for deployment in regions where training data is inherently scarce? We have chosen glacierized high-mountain catchments as our target region, but the experimental framework is directly transferable to other hydroclimatically unique and data-sparse regions such as the hyperarid tropics or the Arctic. We note that none of the architectures evaluated includes an explicit glacier mass balance module; this study therefore characterizes data-driven generalization skill under a realistic operational workflow — where externally-modeled glacier runoff is available as forcing (Rounce et al., 2023) rather than claiming process representation fidelity. Establishing which architecture and training strategy generalizes reliably under historical conditions is a necessary prerequisite for downstream climate impact assessment: a model that cannot generalize historically cannot be trusted to produce reliable projections under future glacier change scenarios (Aguayo et al., 2025b; Thébault et al., 2026).

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

2 Data and Methods

2.1 Data

2.1.1 Gauge and meteorology data

We used a combination of glacial inputs (see section 2.1.3) and the Caravan dataset (Kratzert et al., 2023) - an aggregate of seven existing open large-sample hydrology datasets. The data are standardized globally ensuring that all the catchments share a common set of meteorological and landscape variables that are derived from the same source datasets using the same procedures. Caravan includes area-normalized daily streamflow observations for 6,830 basins (1981-2020). The meteorological forcing data is obtained from ERA5-Land and has global coverage with a spatial resolution of 9 km (Muñoz-Sabater et al., 2021). Additionally, there are 2 sets of catchment attributes: one derived from HydroATLAS and the other from the daily ERA5-Land time-series. We further added the Caravan extension version of CAMELS-DK (Koch, 2022) and CAMELS-CH (Höge et al., 2023). CAMELS-DK includes 308 catchments from Denmark and CAMELS-CH covers 331 basins from Central European countries including Switzerland, Austria, France, Germany, and Italy. These additions are critical for our study as these regions contain many glacierized catchments. Out of ~7,500 basins, we selected those basins with continuous streamflow record which resulted in a total of 2,845 basins across the globe (Table S1). The maximum continuous streamflow record was identified to be 1988-2008 and we chose this as our training period.

From these data, we derive features. Features are simply the input data we use to make predictions with DL. In this study, the forcings (Table S2a & Table S2b), derived from ERA5-Land, and static attributes (Table S3) of the Caravan basins describe the features. In DL, label is used to describe the variable of interest that we are trying to predict, in this case streamflow normalized by basin area (mm/day). All three models were forced with the same ERA5-Land meteorological fields, so any biases in the reanalysis affect the models uniformly and do not confound comparisons across architectures.

2.1.2 Data partitioning

We partitioned the 2,845 Caravan basins into three groups. The three training partitions represent a deliberate gradient of hydroclimatic representativeness — operationalized through progressive physiographic filtering to increase coverage of the target glacierized regime — from globally diverse (A), through mixed high-mountain (M), to exclusively glacierized high-mountain (G). This design allows the effect of training data representativeness to be evaluated independently from training data volume.

- **All Global Catchments (A) – 2,845 basins:** This is the largest and most generalized dataset in our study with continuous streamflow data within our training period. Basins from glacierized regions, other mountain regions,

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

and non-mountain regions altogether form All Global Catchments (A). This dataset represents the theory that more diverse training data yields a more skilful model, regardless of hydroclimatic representativeness.

- **Mixed High Mountain Catchments (M) – 1,020 basins:** All the high mountain catchments in (A) are grouped as Mixed High Mountain Catchments (M). (M) includes both glacierized and non-glacierized mountain basins, hence, "mixed". These basins are identified by intersecting the high mountain regions of the world defined by Karagulle et al. (2017) with the (A) dataset. To define this intersection, we use the spatial-join function in the geopandas package in python to find the polygons of (A) basins that intersect with the high and scattered high mountain (K3) class polygons (https://rmgsc.cr.usgs.gov/gme/). This dataset tests a theory that hydroclimatic representativeness (high mountains are most representative of glacierized catchments) should yield more skilful models but balances the amount of training data available (more than just glacierized catchments) against the traditional regionalization approach.

- **Glacial High Mountain Catchments (G) - 283 basins:** A refinement of (M) containing all glaciated high mountain basins gives the Glacial High Mountain Catchments (G). Glaciers are defined as per the Randolph Glacier Inventory (RGI). This subset is, again, formed by intersecting the (M) basin polygons with the glacier polygons (https://nsidc.org/data/nsidc-0770/versions/6) using the geopandas spatial-join function. This dataset represents prediction closest to traditional regionalization, where models are trained on regions similar to the target basins.

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

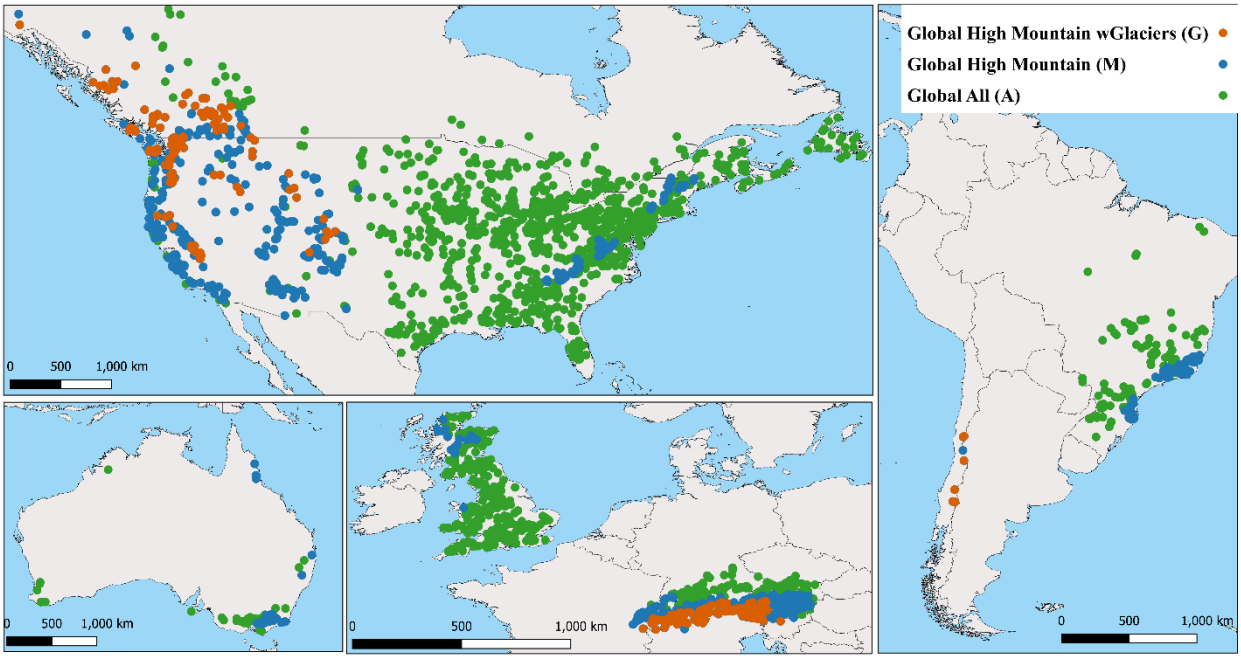


Figure 1: Spatial Distribution of Caravan dataset for each partition across different continents. The global distribution of A, M, and G is shown here, and the lack of spatial representation is apparent as there are no Caravan basins in Asia or Africa. Asia contains the world's highest mountains and thus makes a perfect illustration of the lack of public and standard data needed to train models for glacierized basins. Our G dataset is therefore composed of Alpine, Patagonian, and North American glaciers.

2.1.3 Glacierized inputs

Categorization of G basins based on glacier coverage

For further analysis, we categorized the G basins based on the first-order and second-order regions defined in the RGI. Out of 19 first-order regions, this study included three first-order regions and six second-order region categories as shown in Table S4. We again categorized the G basins into low, moderate, high and very high classes based on the glacier area percentage by binning them into [2, 20, 40] (Table 1). To do this, we identified the total area covered by the RGI glacier polygons in each of these G basins and computed the percentage of area with respect to basin area.

Table 1
Count of basins in each glacier category

Glacier Category	Glacier Percentage	Basin Count
Low	< 2	188
Moderate	2 - 20	89
High	20 - 40	4
Very High	>40	2

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

Glacier Data

We used monthly glacier runoff for the RGI obtained from Rounce et al. (2023). They simulated each RGI glacier independently using a hybrid glacier-evolution framework that couples the PyGEM mass-balance module with the OGGM glacier-dynamics module to produce monthly glacier runoff for each glacier. For each basin in the G dataset, we identified all RGI glaciers whose outlines were within the basin boundary. We aggregated the monthly runoff from all of the glaciers inside the basin to get the basin scale monthly glacier runoff. To derive daily glacial runoff from the monthly values, we applied a temperature-index disaggregation method, assuming melt is proportional to daily positive degree days (PDDs). Daily PDDs were computed as follows:

$$T_i^+ = \max(0, T_i - T_{base})$$

where T_i is the daily mean air temperature and T_{base}= 0 C. For each month, daily weights were calculated as

$$w_i = \frac{T_i^+}{\Sigma_{j\epsilon m} T_j^+}$$

and daily runoff was then estimated as

$$R_i = R_m \times w_i$$

where R_m is the total monthly runoff. In months with no positive degree days, melt was distributed evenly across all days.

This approach preserves the physically-modeled monthly glacier runoff magnitude — which encodes basin-specific glacier hypsometry, area, and calibrated mass balance parameters from PyGEM (Rounce et al., 2023) — while distributing melt proportionally to daily positive degree days within each month, consistent with standard temperature-index disaggregation practice (Hock, 2003). The resulting daily glacier runoff therefore provides information beyond temperature forcing alone, specifically the absolute magnitude of glacier melt contribution to total basin runoff for each month, which the DL models are not otherwise exposed to.

2.2 Methods

We compare the three deep learning models by their model type, model components, their data representation, information flow inside the model, their neighborhood consideration, and processing approach. We also compare them by the data and resources required to train them.

2.2.1 LSTM model and its architecture

An LSTM is a type of recurrent neural network (RNN) in which a single memory cell, consisting of regulatory gate units, is designed to take sequential data over a long period of time (Hochreiter and Schmidhuber, 1997). The regulatory gates enable

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

the LSTM to learn long-term dependencies without experiencing exploding and vanishing gradients (Hochreiter and Schmidhuber, 1997). Each memory cell consists of 3 gates maintaining and updating the cell state (or memory): forget gate (f_t), input gate (i_t) and output gate (o_t). The forget gate compares the relevancy of the previous information in the cell state to the current input. Then, the input gate learns new information from the input and decides which pieces of the new information to store in the current state. Finally, the output gate controls which pieces of information in the current state to output. This architecture of the LSTM makes it useful as a hydrological model to capture the storage effects within hydrological catchments. LSTMs are particularly well-suited to time series analysis as it learns patterns and dependencies through back-propagation and gradient-based optimization and have been extensively used in hydrology (Fang et al., 2021b; Feng et al., 2020; Frame et al., 2022; Hu et al., 2018; Kratzert et al., 2018, 2019; Lees et al., 2022; Shen et al., 2021).

In this paper, we use the NeuralHydrology (Kratzert et al., 2022) python package to train the LSTM model. We calibrate the model with the ERA5-Land variables (meteorological forcing data and model state variable) listed in Table S2a and the static attributes in Table S3 to predict daily streamflow. We followed a similar approach to define model architecture as in (Kratzert et al., 2019). To predict the streamflow (label) of each day, the model received the static and the dynamic forcings of the previous 365 days. This is called the 'lookback period' and it is one of the hyperparameters whose choice affected the model's learning. Additional hyperparameters were set as follows 50 epochs in batches of 256, i.e., for calibration, the model took 256 random samples of the training data in each epoch for 50 epochs.

2.2.2 GNN model and its architecture

A Graph Neural Network (GNN) is a type of deep learning algorithm that learns from unstructured data, i.e., graphs (Bronstein et al., 2017; Roth and Liebig, 2022; Scarselli et al., 2009). A graph is defined by a set of nodes and edges and their relationship is generally represented by a weighted adjacency matrix. The nodes contain features that can be dynamic or static. Graph WaveNet proposed by Wu et al. (2019) is a type of a GNN employed for spatial-temporal graph modeling to learn hidden patterns from graphs. Graph WaveNet uses a Graph Convolution Layer (GCN) in conjunction with a gated Temporal Convolution Layer (TCN) to capture the spatial dependencies and temporal trends of a node respectively (Wu et al., 2019). To construct the adjacency matrix needed for the Graph WaveNet and define a prior graph topology, we followed a similar approach as defined by Sun et al. (2021). To do so, we computed the pairwise Euclidean distance between the nodes based on the static attributes of the nodes. Each node had a set of neighbors defined based on those edge lengths that exceeded a certain percentile cutoff value of the resulting edge length CDF. Although the training and testing gauges together made up the nodes in the adjacency matrix, the model could only see the training nodes during training whereas during testing, all the nodes were visible to the model such that the target basins could have the training basins in their neighborhood pool.

We edited the Graph WaveNet model used by Sun et al. (2021) to train the models for our study. The model took both the static and dynamic forcings as input (similar to LSTM) to predict daily discharge values (Sun et al., 2021). The model

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

architecture is similar to Sun et al. (2021) with 8 TCN/GCN blocks and dilation factor of 2. The models were trained with 365 lookback days for 50 epochs, consistent with the LSTM. As with the LSTM, the Graph WaveNet is sensitive to the lookback period, and we erred in using a longer period which can in theory capture the delayed runoff from snowmelt as the longer period encompasses the dominant annual hydrological processes/responses. The graph size for each of the experiment and partition was the total number of basins that were used for training as well as testing.

2.2.3 Differentiable HBV model and its architecture

Differentiable modeling is a hybrid approach connecting physical modeling and machine learning (Shen et al., 2023; Tsai et al., 2020), leveraging both the learning and adaptation capability of ML and the process clarity of the process-based models (PBMs). Neural networks complement the PBMs by providing parameterization or process representations, whereas the PBM serve by defining governing constraints and structural framework. Feng et al. (2022) described δHBV as a modified form of the HBV model, a bucket type model with 5 state variables that employs a neural network to determine model parameters. They designed an LSTM to take both static attributes and meteorological forcing as input and return the physical parameters that are required by the HBV to compute the state variables. The end-to-end learning of the LSTM ensures that the physical parameters are computed with the daily inputs ensuring hydrological relationships are conserved and no direct target variable is required for the embedded neural network. Furthermore, they allowed daily dynamicity in physical parameters for evapotranspiration (ET) (for finer control on ET efficiency based on landscape properties and vegetation) and runoff (to allow influence from forcing history). Finally, the model computes the streamflow output with a routing model that convolves a unit hydrograph with runoff. In later advances, Song et al. (2023) demonstrated the use of adjoint schemes to enable implicit numerical schemes which address the numerical error introduced by simple sequential calculations of HBV. Differentiable modeling has also been applied to routing (Bindas et al., 2024), stream temperature (Rahmani et al., 2023) and ecosystem modeling (Aboelyazeed et al., 2023).

The inputs to the embedded LSTM model in δHBV are limited to three forcing variables as required by HBV. In addition to the dynamic forcings, the static attributes of the basins are used to compute the physical parameters. Similar to the previous two models, the δHBV model demonstrated sensitivity towards the choice of lookback period, and we again used a lookback period of 365 days. We trained the model for 50 epochs in batches of 100.

Table 2 summarizes the basin input data and key hyperparameters used in each of the models. We fixed other training hyperparameters to be the same as the previous studies. Due to the spatiotemporal learning of the Graph WaveNet model, on a single GPU it required higher VRAM with limited training batches. Likewise, the dynamic inputs for δHBV are limited to three variables due to the constraint of the conceptual HBV model. However, the absence of such constraints in the other two models allowed them to train with other dynamic variables. Table 3 summarizes the models.

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

Table 2
Model inputs, hyperparameters and resources used for each model

	Model Inputs			Hyperparameters			Resources used for training			
	Static Attributes	**Target Variables**	**Dynamic Inputs**	**Sequence Length**	**Epochs**	**Batch Size**	**CPU**	**GPU**	**RAM**	**VRAM**
LSTM	Table S3	Streamflow	Table S2a	365	50	128	8	1	300G	Min. 11G
DHBV	Table S3	Streamflow	SI Table S2b	365	50	100	8	1	120G	Min. 11G
Graph WaveNet	Table S3	Streamflow	Table S2a	365	50	2 (A), 4 (M), 8 (G)	8	1	300G	>40G

Table 3
Differences between models

	LSTM	**δHBV**	**Graph WaveNet**
Model Type	Recurrent Neural Network (RNN)	Hybrid model (combination of traditional physics-based model (HBV) and a data-driven approach (LSTM))	Spatial-temporal Graph Neural Network (GNN)
Model components	Gates, cell state, output state	Modified HBV components, LSTM based neural network	Convolutional layers, pooling layers, fully connected layers
Data Representation	Sequential data (time-series)	Sequential data with physical constraints	Graph-structured data (nodes and edges)
Information Flow	Sequential, influenced by historical information	Concurrent, influenced by spatial-temporal context and hydrological processes	Concurrent, influenced by neighbors' information
Neighborhood consideration	Not explicitly considered	Hydrological relationships	Considers node neighborhoods
Processing Approach	Learns patterns and dependencies through backpropagation and gradient-based optimization	Leverage both physics-based equations and data driven learning for accurate predictions	Learns spatial and temporal dependencies through end-to-end learning

3 Experimental Design

Traditional hydrological regionalization holds that prediction skill is maximised when training catchments are as similar as possible to the target region (Wagener et al., 2007). Machine learning research, however, suggests that skill improves with

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

both data volume and moderate diversity in training data, even when representativeness is reduced (Fang et al., 2022; Ma et al., 2021). Our experimental design directly tests this tension in the specific context of ungauged glacierized basins, where the target regime is simultaneously the most hydrologically distinct and the most data-scarce — making the tradeoff between representativeness and dataset size particularly consequential.

The four experiments are designed to independently isolate three axes of training data influence on DL generalization to ungauged glacierized basins:

- training data volume (the quantity of training data available),
- hydroclimatic representativeness (how closely the training regime matches the target glacierized domain), and
- glacierized catchment inclusion (whether glacierized catchments are explicitly present in training).

Experiments 1 and 2 test the hydroclimatic representativeness and size axes while retaining glacierized basins within all training partitions. Experiments 3 and 4 test the representativeness axis by excluding glacierized catchments from training while independently varying volume. Experiment 5 tests the marginal value of explicit glacier process information as a model input, independent of training data composition. Importantly, in all experiments, skill is assessed completely out-of-sample, consistent with ungauged basins. That is, we are assessing each model only in its ability to correctly estimate streamflow in our target region without providing any training data from the gauges used to judge skill in either space or time. We do provide training data in similar glacierized regions, but never from the catchments we test in. To assess the models along the axis of hydroclimatic representativeness of the training pool, the first experiment trains each model with three data partitions (A, M and G) with increasing levels of representativeness. The 2nd and 3rd experiments evaluate the models along the axis of training data size by limiting the size of the training data in each partition while the 3rd and 4th experiments test the models along the axis of inclusion of the representative basins of the target region.

To eliminate sampling bias and representation bias, for each experiment we perform a Repeated K-Fold cross validation to randomly sample 56 (20%) test basins from G with K=100 times. This results in 100 different test sets and training sets for each experiment ensuring the training sample set is not biased. We also enforce an identical 1:1 mapping of testing across models such that for example, on trial 27, the GNN, LSTM, and δHBV, all have exactly the same held-out test data regardless of its training on A, M, or G. This test split is then randomly redrawn for example, trial 28, but again, for each data partition (A, M, G), the test set is passed to each of the three models identically and training drawn from the remainder of the partition according to the experiment. We do this to obtain robust predictions that are insensitive to the exact test set and to ensure that differences in skill are always obtained on the same test set for each of K trials. Note that we specify train and test sets here to signify the basins used to train the models and to test the skill of the models. However, within the training set we further split the training set into 70/30 split of training and validation basins to optimize each ML model's parameters as it trains.

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

3.1 Experiment 1: Non-Exclusive Full Set Training (Baseline Performance)

For this experiment, models were trained using all non-held-out basins from each respective dataset to establish a performance baseline. For each of our 100 trials, we sample 56 basins (20%) of the G dataset and hold these data out from each of the three sets A, M and G. We then train each model on the remaining data, which results in an unequal number of training data across sets (i.e., 2,789 basins in A, 964 basins in M, and 227 basins of G). This experiment evaluates the skill of the models in manifold ways, whereby all data in each partition is used to train each ML method. ML theory predicts that A will have the highest skill here (Fang et al., 2021a; Kratzert et al., 2019; Wi and Steinschneider, 2022), whereas regionalization of hydrology would expect either M or G to have the highest skill. We note that the GraphWaveNet configuration required excessive runtime to acquire the full replicate of 100 trials, so we truncated Experiment 1A for the GraphWaveNet at 36 repeat trials. All other experiments contain the full suite of replicates.

3.2 Experiment 2: Non-Exclusive Size-Matched Selection

In this experiment, we used the same volume of training data across our partitions, exploring the skill of models artificially limited to the same amount of training data but with varying degrees of hydroclimatic representativeness to our target region. As with all experiments, a common test set is defined by randomly sampling 20% of the Global Glaciers (G) set. We then randomly sample 227 basins for training, which is equivalent to the complement of G (i.e., all G not used in training). Note that A and M both encompass G, so there is a random chance glaciated basins are chosen in the training data. We hypothesize that G will have the best performance here as the A and M datasets, although diverse, no longer have a training data volume advantage and G represents the most similar training data.

3.3 Experiment 3: Exclusive Size-Matched Selection

For this experiment, we repeat Experiment 2 but require that basins selected from the larger datasets must be exclusive and not present in the smaller subsets. That is, we still use 227 basins to train each of the three models, but in this experiment glacierized catchments are not available to train M or A – all the data must be sampled from non-glaciated basins. This experiment therefore combines with Experiment 2 to control the effect of inclusion in training data within a controlled sample size. We again expect G to perform best here.

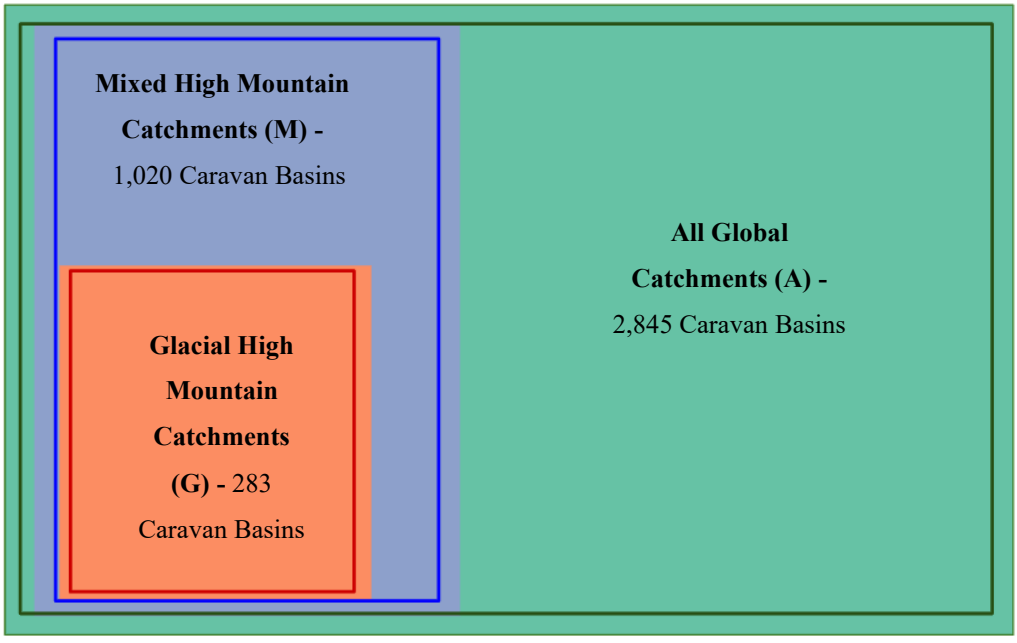
3.4 Experiment 4: Exclusive Full Set Selection

Finally, models were trained using all non-held-out basins exclusive to each respective dataset. This results in the same structure for G as experiments 2 and 3, but M now uses all the basins excluding G (737 basins) basins, and A dataset has all basins not in M or G (1,825 basins). This affects the 100 random sampling of the training set for A and M. Therefore, we fix our trial size as 10 with random initialization. This experiment therefore tests the effect of training sample size within a controlled heterogeneity and directly pits ‘small homogeneous training’ against ‘large dissimilar training’.

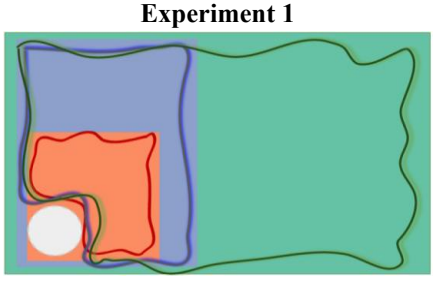
https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

3.5 Experiment 5: Addition of Glacial Runoff Data

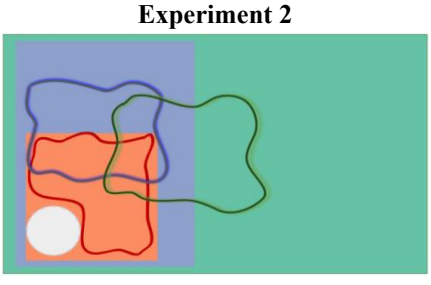
We repeat the first 10 trials from experiment 1 for G set with addition of daily glacial runoff data as dynamic input to the models. This experiment tests the effect of availability of glacial data on the process of model learning. For LSTM and Graph WaveNet, the disaggregated daily glacier runoff is added as an explicit additional input feature for each basin. However, this
395 data is added to the precipitation time series to represent glacier contribution in the dPL-HBV runs.



(a)



(b)



(c)

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

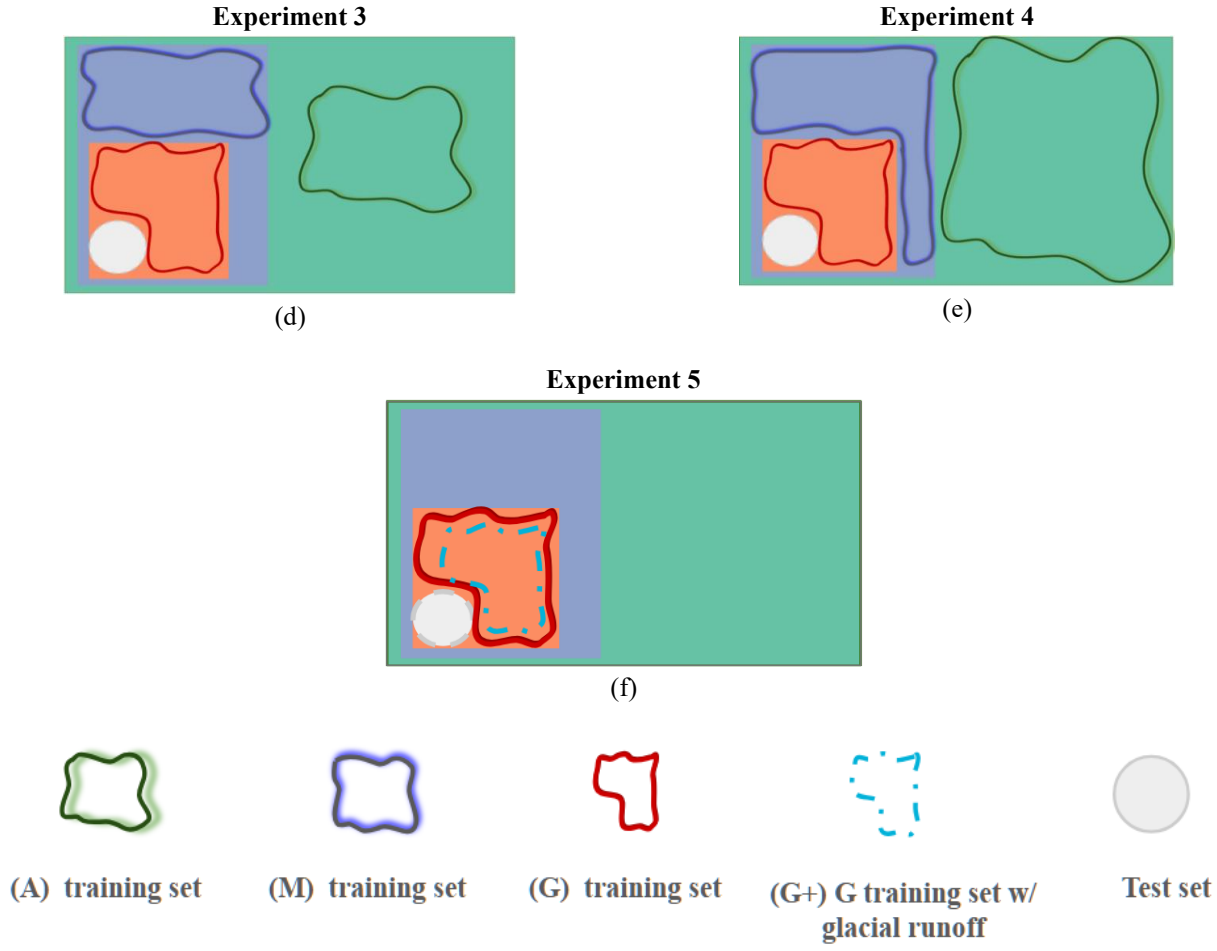


Figure 2: Partitioning of Caravan dataset and experimental design: (a) Partitioning of the Caravan dataset into A, M and G partition; (b-e) The selection of Kth training and test set for each partition for experiment 1-4. (f) G training set with and without glacial runoff data. For each of the 100 trials of Repeated K-Fold cross validation, the white circle takes 100 different positions within the G partition, even though it is drawn in the same position for this illustration of the Kth trial.

3.6 Evaluation metrics for model performance

Our experimental design ensures out-of-context training in both space and time. The training period for all the experiments is 1988-2005 and the test period is 2008-2014. Previous methods that train on the past and test on the present (for example, Kratzert et al. (2018)) simulate a gauge that has gone offline, where we test for gauges that never existed. Therefore, our held-out test data are held out in space and time- the sternest test possible for each method. To evaluate the models, we calculate NSE, KGE and RMSE between the observations and predictions at the daily timestep and summarize these as CDFs across our 100 random trials for each of our 56 held-out basins in each trial. The 5,600 NSE values represent 100 trials of 56 basins

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

each; trials share substantial overlap in training data and are not statistically independent. We use the distribution to characterize sampling variability rather than for hypothesis testing.

4 Results

This section summarizes model performance across the four experimental setups described in Section 3. We first present basin-scale NSE for the full test period in Table 4 and Figs. 3-4, which together summarize both median skill and the full distribution of performance across basins, models (LSTM, δHBV, Graph WaveNet), and data partitions (A, M, G). We then focus on the glacier-relevant summer months in Table 5 and Figs. 5, which report median summer NSE and the corresponding CDFs. In each of the subsection (Sections 4.1-4.2), we discuss each experiment in turn, explicitly linking the narrative to the relevant rows of Figs. 3-5 and to the summary statistics in Tables 4 and 5. Section 4.3 further stratifies performance by glacier-coverage category (low, moderate, high, very high) for the G basins, enabling us to assess how skill changes along a continuum of glacier influence. Finally, Section 4.4 examines Experiment 5, in which explicit glacier runoff forcing is added as an additional input, and evaluates how this glacier data alters basin-scale NSE and pBias across models and glacier-coverage classes (Figs. 8-10).

4.1 Metrics for the entire test period

For experiment 1 (maximum data per partition), each model demonstrates strong out-of-context predicting skill as NSE scores for 70% or more of test basins are equal or greater than 0.5 regardless of training partition (A, M, G) as shown in Figs. 3a, 3b, and 3c. The median NSE values of the three models range between 0.53 and 0.8 for every partition (Table 4). Regardless of the data partition, LSTM slightly exceeds the performance of the other two models (Table 4); however, this performance improvement is marginal (Figs. 4a, 4b, and 4c). The LSTM and δHBV demonstrate more-or-less data partition agnostic behavior in performance, i.e., irrespective of the data representativeness of the training pool relative to target basins, these models perform equally well (Figs. 3a and 3b). On the contrary, the performance of Graph WaveNet improves as the level of representativeness increases (median NSE: 0.53 (A) <0.58 (M) <0.68 (G)) (Table 4).

Table 4
Median NSE Values for entire test period

	All Global (A)			**Mixed High Mountain (M)**			**High Mountain with Glacier (G)**		
	LSTM	**δHBV**	**Graph WaveNet**	**LSTM**	**δHBV**	**Graph WaveNet**	**LSTM**	**δHBV**	**Graph WaveNet**
Exp 1	***0.71***	0.67	0.53	***0.70***	0.67	0.58	***0.68***	0.66	***0.68***
Exp 2	0.49	***0.54***	0.24	0.57	***0.59***	0.55	***0.68***	0.66	***0.68***
Exp 3	-1.43	***0.01***	-0.03	0.41	***0.53***	0.46	***0.69***	0.66	***0.69***
Exp 4	-0.96	-0.09	***0.17***	0.48	***0.56***	0.50	***0.67***	0.65	0.65

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

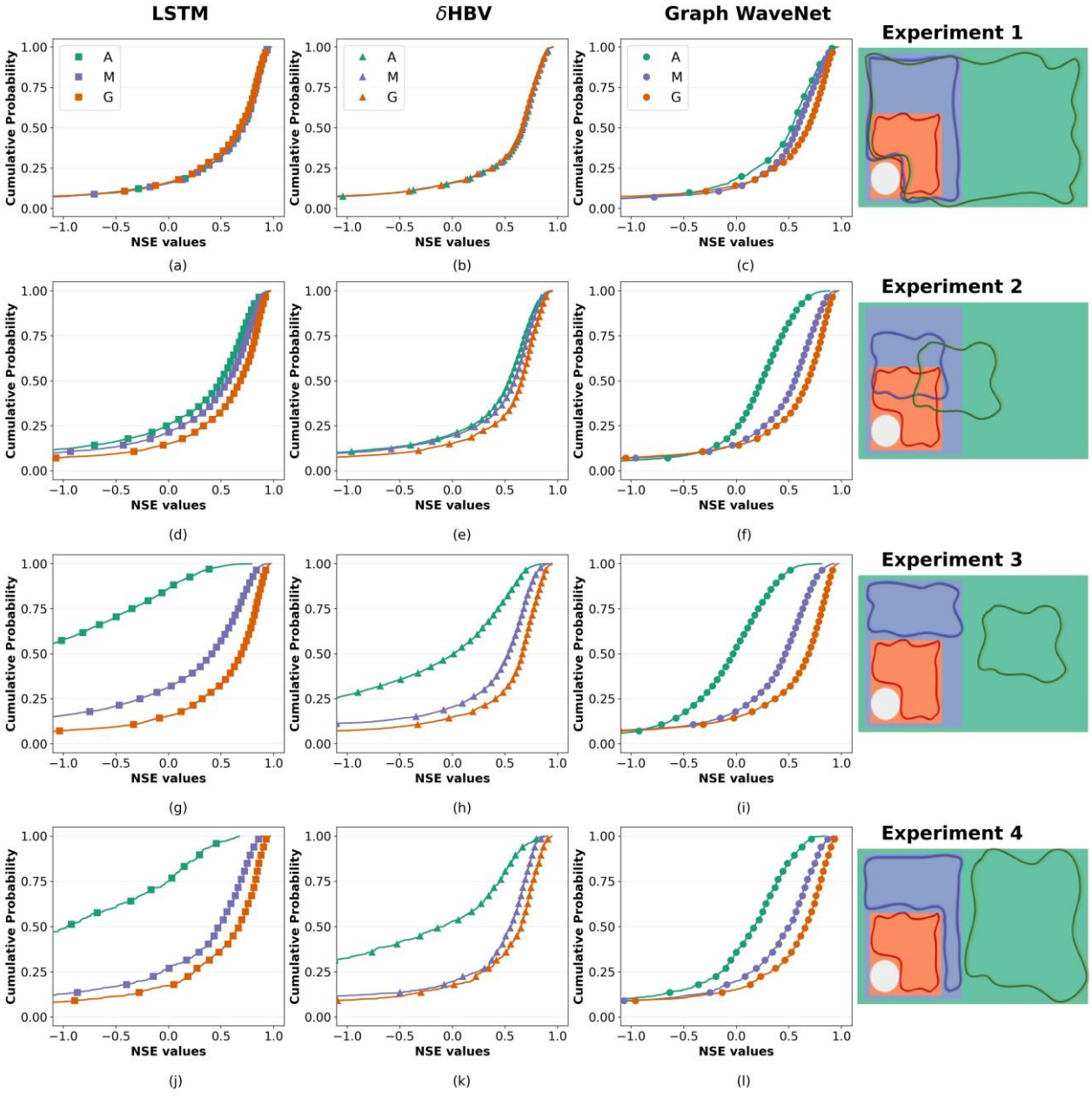


Figure 3: CDF comparison of NSE values of LSTM, δHBV and Graph WaveNet when trained with 3 different data partitions. Rows 1-4 showing performance for experiments 1-4 respectively. Each line of the CDF plot comprises 5600 sample points from each of the 100 trials with 56 test basins for experiments 1-3, while the lines in experiment 4 consist of 560 sample points from 10 trials with 56 test points each. * *Graph WaveNet in experiment 1 for A partition has only ~2300 samples due to excessive resource needs required to train the model.*

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

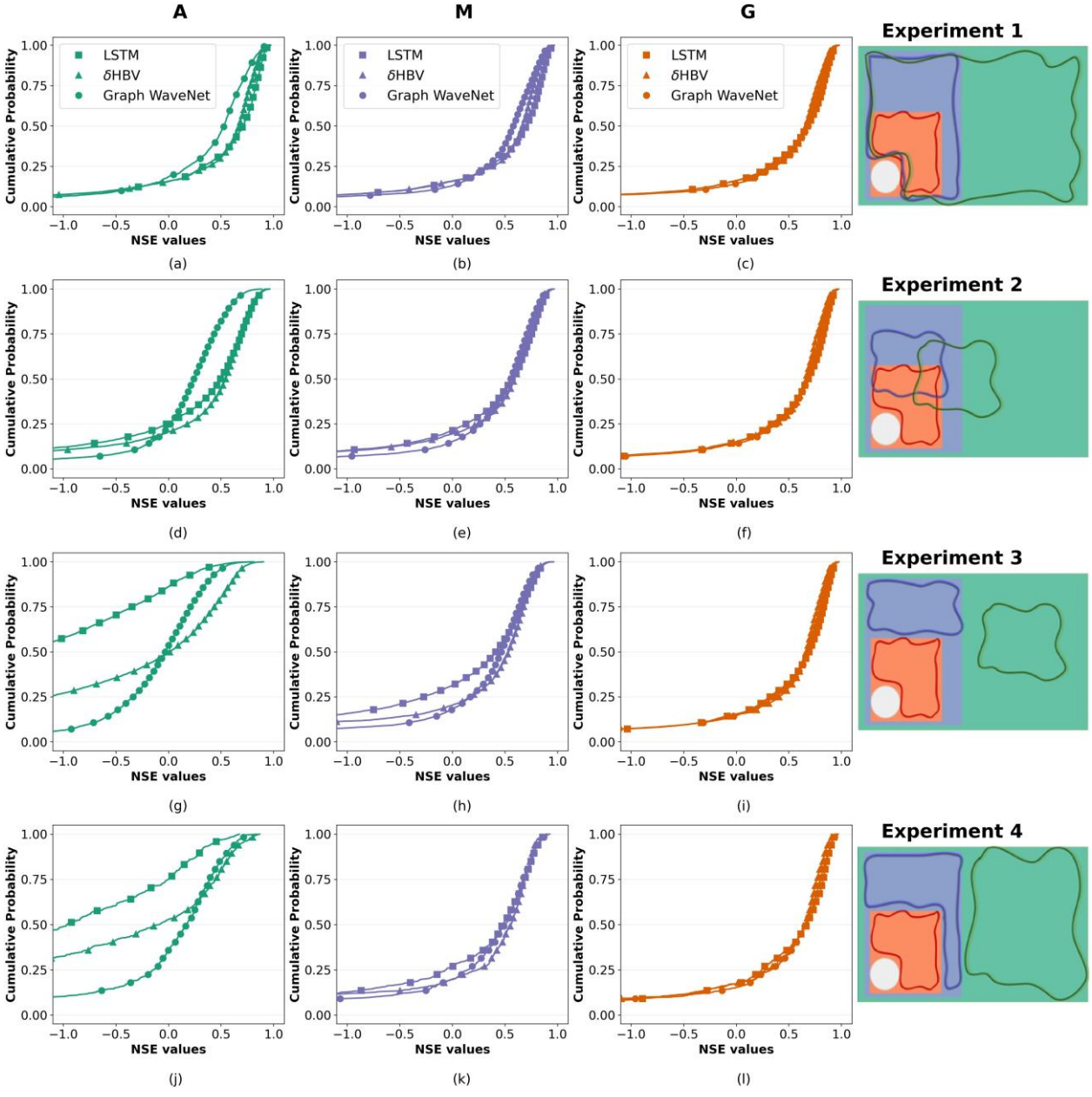


Figure 4: CDF comparison of NSE values of 3 DL model architectures trained with different partitions. Each column represents 3 different data partitions – A, M, and G respectively. Rows 1-4 showing performance for experiments 1-4 respectively. Each line of the CDF plot comprises 5600 sample points from each of the 100 trials with 56 test basins for experiments 1-3, while the lines in experiment 4 consists of 560 sample points from 10 trials with 56 test points each. **Graph WaveNet in experiment 1 for A partition has only ~2300 samples due to excessive resource needs required to train the model.*

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

In experiment 2- Non-Exclusive Size-Matched Selection, as expected, demonstrate a distinct difference in performance across the three partitions where the homogeneous partition trains better than other two (Figs. 3d-3f). Experiment 2 uses the same volume of training data in each partition; however, it does not limit sampling between partitions (Fig. 2c). As hypothesized, each model shows the best performance when trained with the homogeneous G dataset and the performance degrades with increasing heterogeneity (G > M > A) (Figs. 3d-3f). The range of median NSE values for the three models are [0.55, 0.59] and [0.24, 0.55] for M and A respectively (Table 4). The performance of δHBV is better than the other two in both heterogeneous partitions (A and M) (Figs. 4d-4e). The Graph WaveNet model trained with the size limited heterogeneous A partition demonstrates weakest performance with median NSE being 0.24 (Table 4). Nevertheless, each of the three models (except for Graph WaveNet trained with A partition) continue to perform well: NSE greater than 0.5 is true for 50% of test basins across all models (Figs. 4d-4f).

In experiment 3, when further limitation is imposed on size-limited heterogeneous partitions by withholding glaciated basins from the training set (A and M) (Fig. 2d), the performance of the models further declines (Figs. 3g-3i). The median NSE for the three models declines to a range of [-1.43, 0.01] for partition A and to [0.41, 0.53] for partition M (Table 4). This lack of representation and size limitation in A dataset particularly affects the LSTM model as more than 80% of the test basins have NSE<0 (Fig. 3g). Although the other two models are not as severely impacted, their performance also degrades considerably as almost 50% of the test basins have NSE scores less than zero for A dataset (Fig. 4g). However, this drastic decline in performance of models is not seen when trained with M dataset as ~50% or more of test basins have NSE greater than 0.5 for all three models (Fig. 4h).

Lastly, in experiment 4 the models are trained with all available basins, excluding the test basins, exclusive to each of the partitions resulting in different training sizes across partitions (Fig. 2e). There is a slight improvement in the performance of the models by increasing the training data size from experiment 3 even when representation basins are not included. The median NSE increases to a range of [-0.96, 0.17] for partition A and [0.48, 0.56] for partition M (Table 4). The LSTM benefits most out of the three with the increased training size. However, the lack of representative basins still clearly limits the models' ability as the NSE values are still much lower than in all-inclusive training dataset of experiments 1 and 2.

These results reveal a consistent pattern across all four experiments: the three axes of training data influence operate hierarchically. When training data includes large volume including glacierized catchments — Experiment 1 — all three architectures perform well regardless of partition. When volume is equalized across partitions — Experiment 2 — hydroclimatic representativeness of the training pool relative to the target regime emerges as the dominant factor, with G outperforming M outperforming A across all models. When representativeness is explicitly removed by excluding glacierized catchments — Experiments 3 and 4 — skill collapses in the A partition regardless of architecture, and even with the additional data in Experiment 4 only partially restores skill, confirming that data volume cannot substitute for representativeness. Non-

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

glacierized mountain catchments (M partition) partially preserve skill across Experiments 3 and 4, demonstrating that hydroclimatic partial hydroclimatic representativeness — through non-glacierized mountain basin inclusion —provides benefit. Across all conditions, LSTM achieves the highest ceiling but is most sensitive to training data composition; δHBV maintains the most consistent performance due to its structural constraints; and GNN improves systematically with training data representativeness of the target regime, performing best with homogeneous glacierized data and degrading most under decreasing training data representativeness.

4.2 Metrics for the summer months (Jun-Aug) of the test period

Table 5
Median NSE Values for the summer months of test period

	All Global (A)			Mixed High Mountain (M)			High Mountain with Glacier (G)		
	LSTM	**δHBV**	**Graph WaveNet**	**LSTM**	**δHBV**	**Graph WaveNet**	**LSTM**	**δHBV**	**Graph WaveNet**
Exp 1	***0.5***	0.43	0.23	***0.50***	0.44	0.36	***0.48***	0.40	*0.47*
Exp 2	0.16	***0.19***	-0.15	0.29	*0.29*	**0.30**	***0.50***	0.40	*0.47*
Exp 3	-2.83	*-0.59*	-0.65	0.00	***0.14***	0.11	***0.50***	0.40	*0.48*
Exp 4	-2.37	-0.80	*-0.35*	0.09	***0.25***	0.14	***0.44***	0.37	0.42

For Experiment 1, the summer-period NSE distribution mirrors the full evaluation period but with markedly reduced predictive skill. LSTM and δHBV remain largely partition-independent (Figs. 5a–5b), with LSTM retaining a slight advantage (Figs. 6a–6c). Graph WaveNet continues to exhibit regional dependence, improving with greater training data representativeness of the target glacierized regimeto (Table 5). However, performance declines substantially in summer: fewer than 50% of basins achieve NSE ≥ 0.5 across partitions (except Graph WaveNet with A and M; Figs. 5a–5c). Median NSE values range from 0.23–0.50 (Table 5). For experiment 2, the distinction in performance across the three partitions is much more evident for the summer months (Figs. 5d-5f). With equal dataset size and cross-partition sampling, models trained on G basins outperform those trained on M and A. The range of median NSE values degrades to ~0.30 from [0.55, 0.59] for M and to [-0.15, 0.19] from [0.24, 0.55] for A in summer compared to entire test period (Table 4 & Table 5). Both LSTM and δHBV show similar performance when trained with M and A basins with ≤40% of the test basins had NSE higher than 0.5 (Figs. 6d-6e), however, Graph WaveNet degrades further in performance with decreasing training data representativeness, as test basins with NSE ≤0.5 went from ~50% to ~70% for M partition (Fig. 5f).

In experiment 3, further restricting size-limited partitions by excluding glaciated basins (A and M; Fig. 2d) leads to additional performance degradation in summer. For LSTM and Graph WaveNet trained with A partition, the proportion of test basins

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

that have NSE ≥0.5 goes from ~5% (Figs. 3g and 3i) to nearly zero (Figs. 5h and 5i). While the M partition shows less decline than A partition over the full period (Figs. 3g-3i), the effect is more visible in summer months (Figs. 5g-5i). Even models trained on G basins deteriorate, with median NSE falling from [0.66, 0.69] to [0.4, 0.5] (Table 5).

Lastly in experiment 4, removing hydroclimatic representativeness of the training pool relative to the target regime while increasing training size to include all available basins in each partition (Fig. 2d) slightly improves summer performance compared to experiment 3 (Figs. 6g-l). However, the CDF still shifts left for all models and partitions: for example, ~65% of the test basins trained on G achieve NSE ≥0.5 over the full period (Fig. 4l), but this decreases to ~50% in summer (Fig. 6l). M and A partition follow similar degradation trends (Figs. 6j and 6k).

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

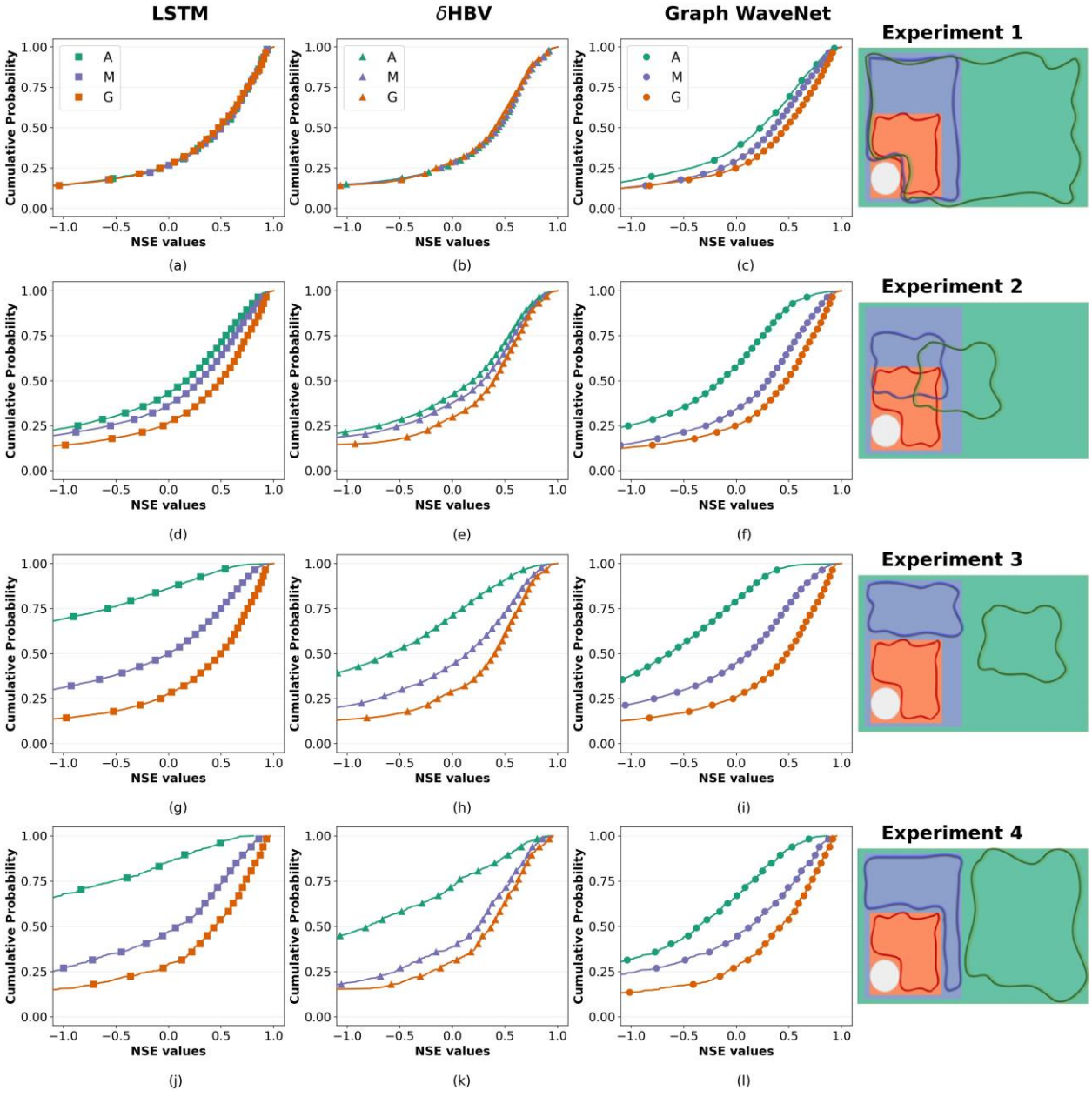


Figure 5: CDF comparison of NSE values of LSTM, δHBV and Graph WaveNet when trained with 3 different data partitions over the summer months of the test period. Rows 1-4 showing performance for experiments 1-4 respectively. Each line of the CDF plot comprises 5600 sample points from each of the 100 trials with 56 test basins for experiments 1-3, while the lines in experiment 4 consists of 560 sample points from 10 trials with 56 test points each. * *Graph WaveNet in experiment 1 for partition A has only ~2300 samples due to excessive resource needs required to train the model.*

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

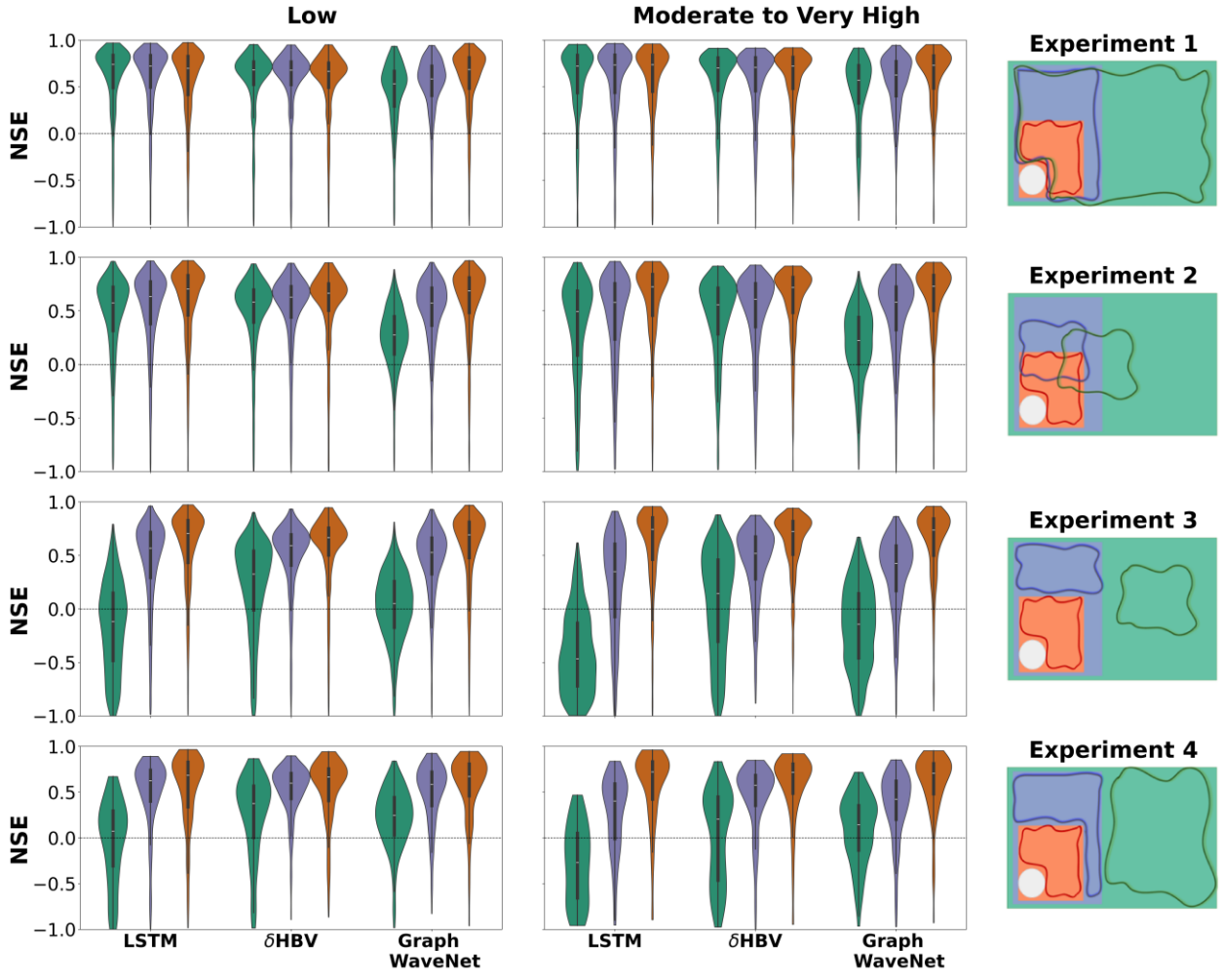
4.3 Performance evaluation by Glacier Coverage Category

To understand how the training design and model architecture interact with glacier influence, we first analyze the distribution of model performance for each of the glacier coverage classes – Low (188 basins with <2%), Moderate (89 basins with 2-20%), High (4 basins with 20-40%) and Very-high (2 basins with >40%). We then analyze the seasonal hydrographs of one representative basin from each of the classes. For each category, we compare the performance of the three models (LSTM, δHBV, and Graph WaveNet) and examine how their simulations differed across the four experimental setups and training data partitions.

4.3.1 Metric distribution by glacier category

Across all experiments, the violin plots of NSE for daily flows show consistent patterns across glacier categories and models (Fig. 6). For low-glacier basins, all three models, trained with M or G partition, achieve relatively high skill with most NSE values ~0.6 in every experiment and only a modest widening of the distribution as the experiments become more restrictive from 1 to 4. However, the A partition fails to maintain the tight distribution as restrictions are imposed in the dataset. This sensitivity to the experimental setup is more prominent in moderate to very high glaciated basins across data partitions. The spread of NSE increases, and the lower tail extends into negative values, especially for LSTM in Experiments 3 and 4. For the summer season, NSE distributions are generally higher and tighter in low-glacier basins across all four experiments (Fig. 7) whereas moderate to very high glacierized basins, without representation basin in the training set (Experiment 3 and 4), show broad distributions and frequent negative NSEs.

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026



530 **Figure 6:** Distribution of Nash–Sutcliffe efficiency (NSE) for the three models (LSTM, δHBV, and Graph WaveNet; x-axis) across glacier categories and experimental setups. Columns correspond to glacier classes (Low, Moderate to Very High), rows to the four experimental training configurations (Experiments 1–4).

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

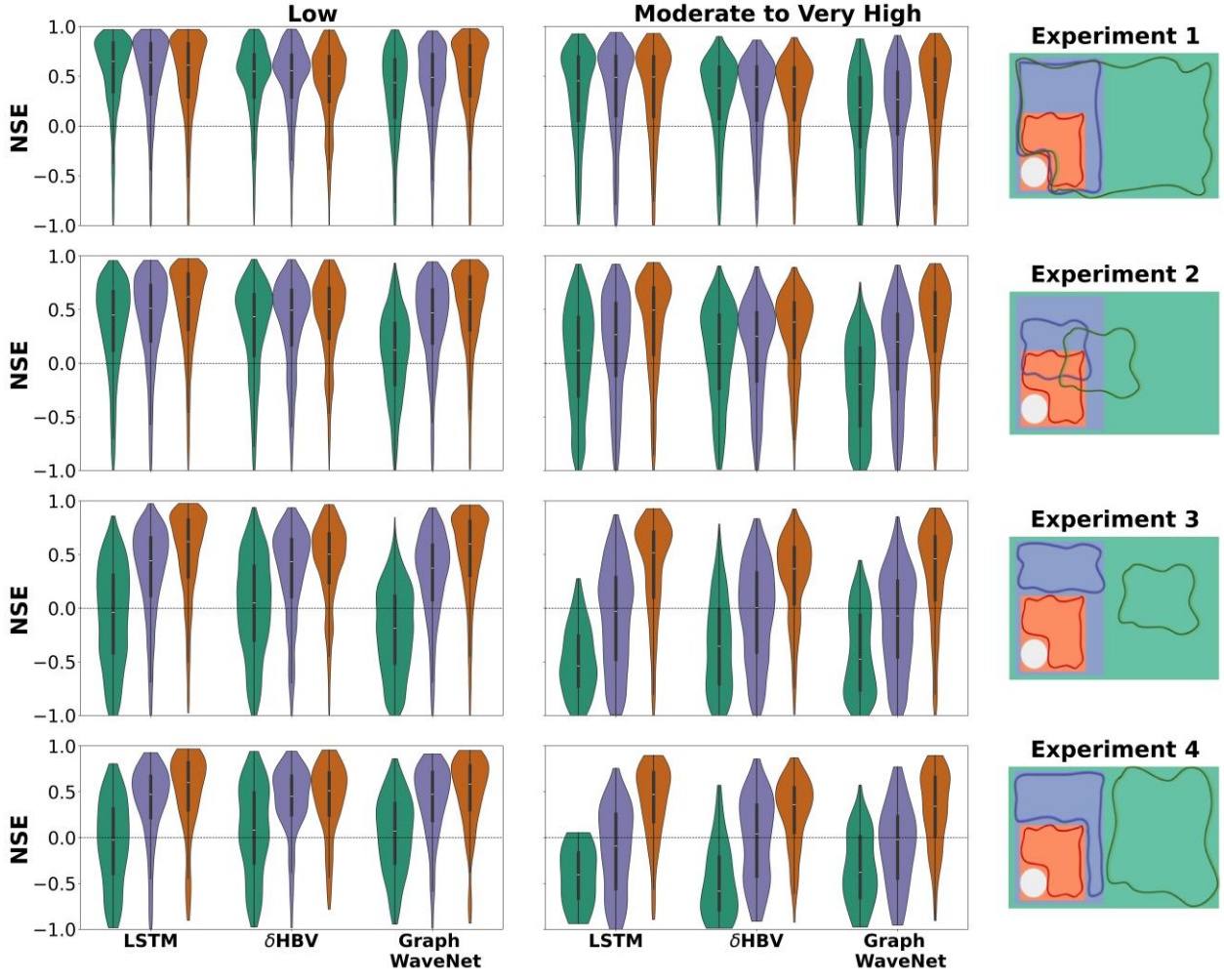


Figure 7: Distribution of Nash–Sutcliffe efficiency (NSE) for the three models (LSTM, δHBV, and Graph WaveNet; x-axis) across glacier categories and experimental setups for the summer months of the test period. Columns correspond to glacier classes (Low, Moderate to Very High), rows to the four experimental training configurations (Experiments 1–4).

4.3.2. Hydrograph characteristics by Glacier Category

Low Glacier Coverage (<2%)

Most of the basins with low glacier coverage behave similarly to non-glacierized mountain catchments. The typical hydrograph peaks in late spring or early summer that are driven by snowmelt and rainfall, with little contribution from glacier ice, and recedes after the snowmelt pulse (Fig. S1). The simulated hydrographs for a representative low glacierized basin (hysets_08NF001) are shown in Fig. S7. In Experiment 1, each of the models is successful in capturing the timing of early summer peak of this representative basin: the simulated hydrographs mirror the observed with a single distinct early-summer

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

peak followed by recession (except LSTM A3). Peak magnitude biases are relatively small, though some differences emerge across models and experiments. The Graph WaveNet and δHBV tend to underestimate the peak flow followed by LSTM when trained on A and M partitions across experiments (Fig. S7). Particularly, Graph WaveNet under-predicts the peak (on the order of ~20-50% low) when trained with A dataset. Each of the models show a modest low bias in peak magnitude when trained with G dataset.

Moderate Glacier Coverage (2-20%)

Moderately glacierized basins exhibit a mix of snowmelt- and glacier melt-driven flow, resulting in a more prolonged summer high-flow period than low-glacier basins (Fig. S8). Typically, these catchments still experience a snowmelt-driven rise in discharge from late spring to early summer, but unlike the low-glacier case, the presence of glacier ice sustains and augments the flow into mid- and late-summer. In Experiment 1and 2, both δHBV and LSTM simulate the mid-summer peak and sustain flows with reasonable accuracy across data partitions. However, Graph WaveNet, trained with A or M dataset, underestimates the peak producing a much lower summer flow than the observed across experiments (Experiment 1-4). In Experiment 3 and 4, performance of δHBV and LSTM slightly deteriorates with M dataset, however, they overestimate the peak flow when trained with A dataset, particularly, LSTM trained with A dataset but without representativeness or size increment over predicts by almost 50%.

High Glacier Coverage (20–40%)

Highly glacierized basins are strongly dominated by glacier melt contributions in summer. The result is a pronounced late-summer peak in flow – with the melting of seasonal snowpack followed by melting of glacier ice reaching its maximum rate (Fig. S9). Under Experiments 1 and 2, δHBV and LSTM both capture the hydrologic cycle well; performance deteriorates in Experiments 3 and 4, especially for the LSTM in terms of timing. GraphWaveNet was the worst performing model in all experiments and shows poor performance outside Experiment 1.

Very High Glacier Coverage (>40%)

Basins with very high glacier coverage exhibit strong glacier-driven hydrology. The summer streamflow is dominated by glacier melt and peak flow occurs in late summer. These basins present the greatest challenge for the models due to the limited training data. The only cases where the models exhibit reasonable performance is when the LSTM are trained on glacierized basins. The δHBV and Graph WaveNet are unable to reproduce the seasonal cycle in terms of magnitude in particular.

4.4. Model Performance with additional glacier data

Stratifying results by basin glacier coverage (Figs. 8-9) and by seasonal hydrographs (Fig. 10) reveal that the impact of glacier forcings is strongly conditioned on glacier fraction and model architecture. Including glacier information changes total volume bias only slightly but can have a more noticeable effect on NSE (Figs. 8-9). For pBias, more than 80–93% of basins for every

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

model and glacier class change by less than 5% when glacier data are added. In low-glacier basins, only 6–8% of basins per model fall above the band and 3–9% fall below (LSTM: 10 above, 15 below out of 162; δHBV: 7 above, 5 below out of 166; Graph WaveNet: 12 above, 12 below out of 160). In moderate–to–very-high glacier basins, 84–90% of basins remain within 5% (LSTM: 5 above, 4 below out of 79; Graph WaveNet: 4 above, 4 below out of 78), with δHBV showing the strongest asymmetry: 12 of 79 basins (~15%) move above the band while only one basin falls below, indicating a subset where adding glacier information increases already positive biases and leads to larger overestimation of total flow. Overall, however, the dominance of points within the band shows that glacier data do not systematically alter total volume bias for most basins.

The NSE comparison reveals a more mixed response. For low-glacier basins, 62–72% of basins per model lie within the ±0.05 band, implying little change in efficiency. Among the remaining basins, LSTM has 28 (17.3%) with NSE increases and 30 (18.5%) with decreases; δHBV has 18 (10.8%) improved and 28 (16.9%) degraded; and Graph WaveNet has 35 (21.9%) improved and 26 (16.2%) degraded. Points below the band for LSTM and δHBV show a wider spread, including several basins where NSE becomes strongly negative with glacier data, whereas points above the band are generally modest for upward shifts from already moderate–high NSE. Thus, in low-glacier basins, adding glacier information is largely neutral and can worsen NSE for a minority of catchments, especially for LSTM and δHBV, while Graph WaveNet shows slightly more improvements than degradations. In moderate–to–very-high glacier basins, 59–62% of basins remain within the ±0.05 band (47–48 basins per model), again indicating limited change for most catchments. For LSTM, 17 of 79 basins (21.5%) improve by more than 0.05 NSE and 15 (19.0%) degrade; for δHBV, 11 basins (13.9%) improve and 21 (26.6%) degrade; and for Graph WaveNet, 13 of 78 basins (16.7%) improve while 17 (21.8%) degrade. As in the low-glacier class, the spread of points below the band is generally larger than above, with several basins exhibiting a transition from moderate or positive NSE without glacier data to strongly negative NSE with glacier data. Improvements above the band are typically smaller in magnitude and concentrated among basins that were already well simulated.

Taken together, these results show that explicit glacier information has little impact on total volume bias for most basins and produces a heterogeneous response in NSE: it yields modest improvements for a substantial subset of basins, particularly in glacier-influenced regions, but also introduces notable degradations in a smaller yet non-negligible group of catchments. It's unclear whether the quality of the glacier information is a factor.

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

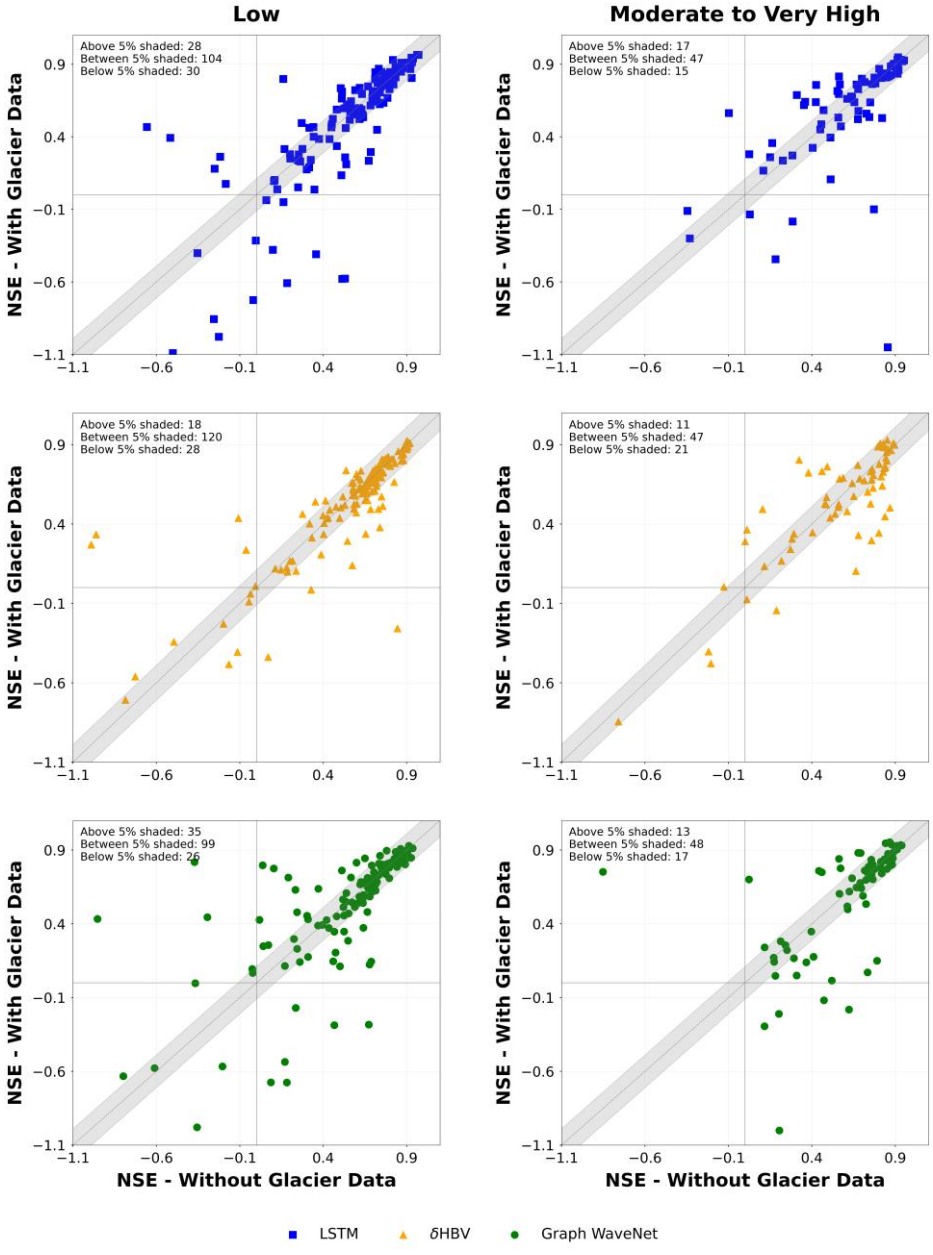


Figure 8: Change in basin-scale NSE when glacier data are included for three models (rows: LSTM in blue, δHBV in orange, Graph WaveNet in green) and glacier-coverage classes (columns: Low & Moderate to Very High). Each point represents one basin.

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

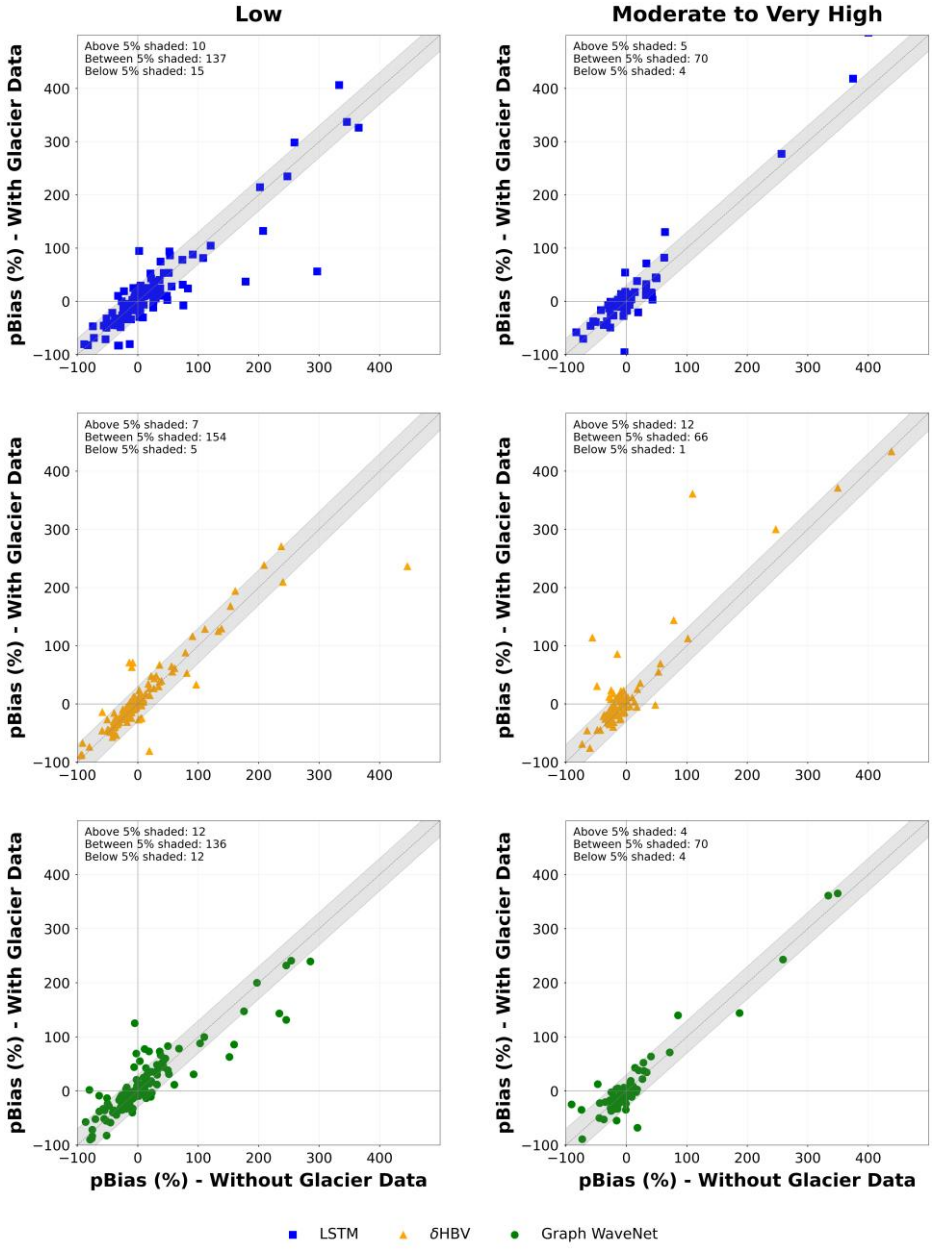


Figure 9: Change in basin-scale pBias when glacier data are included for three models (rows: LSTM in blue, δHBV in orange, Graph WaveNet in green) and glacier-coverage classes (columns: Low & Moderate to Very High). Each point represents one basin.

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

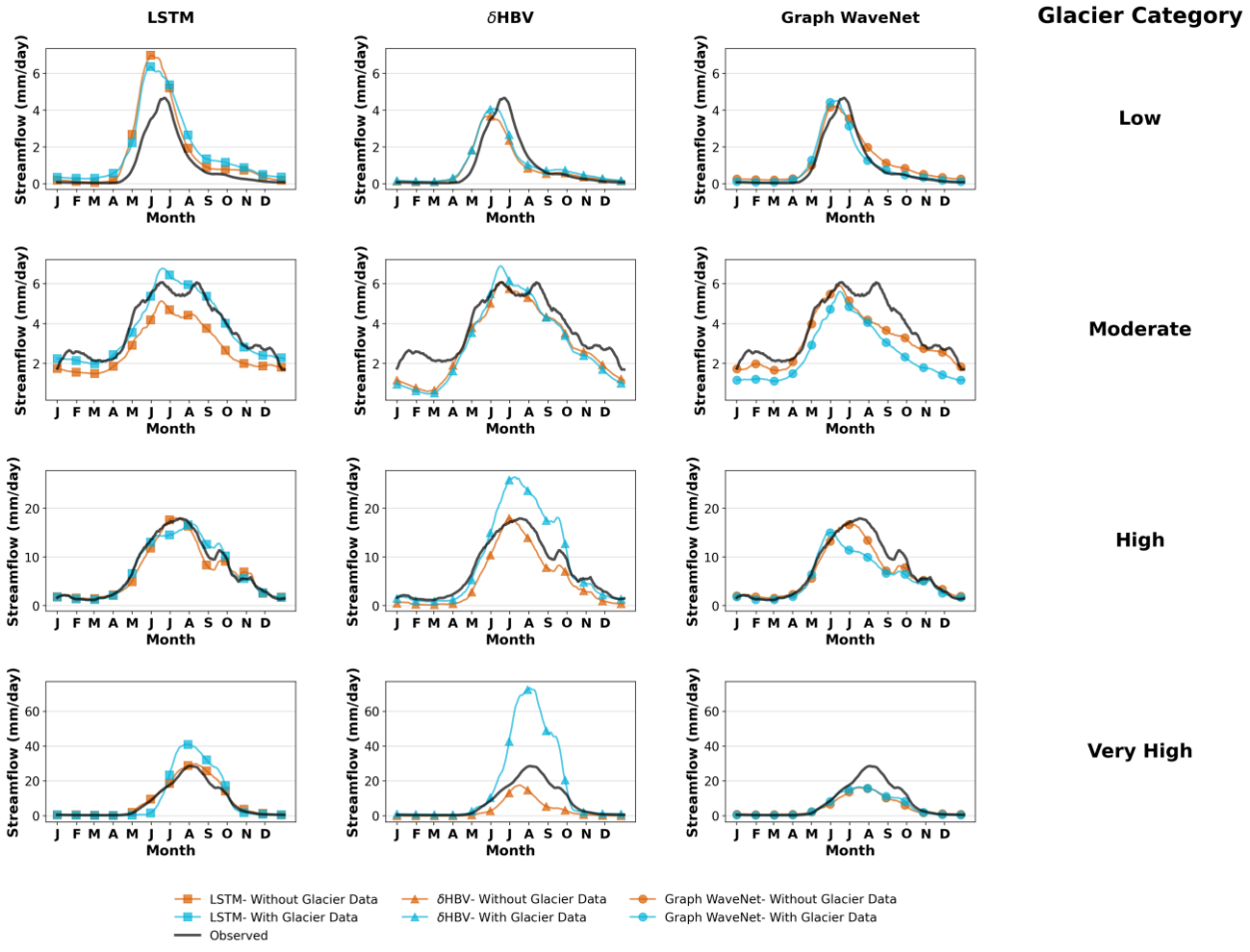


Figure 10: Comparison of monthly mean streamflow with/without glacial inputs across three models (LSTM, δHBV and Graph WaveNet) for representative basins of each glacier category. (Low: hysets_08NF001, Moderate: lamah_201780, High: hysets_08GA071, Very High: hysets_08ME023).

5. Discussion

Across all four experiments, hydroclimatic representativeness of training data to the target glacierized regime emerges as the primary determinant of DL generalization skill — consistently dominating both training data volume and architecture choice — a finding that carries direct practical implications for how DL should be deployed in high-mountain water towers All the model architectures predicted discharge of ungauged glacierized regions well in out-of-context test basins in the entire test period in Experiment 1. The results demonstrated how high the out-of-context skill is when there are large volumes of training data available, regardless of training data representativeness of the target region. This finding is consistent with other studies (K. Fang et al., 2021; Ma et al., 2021). K. Fang et al. (2021) demonstrated the beneficial data synergy effect of combining data from different regions against learning from a homogeneous data of a single region. Similarly, Ma et al. (2021) showed

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

transferring learning from data rich diverse regions to data sparse regions through DL models is a viable and beneficial approach. Our results therefore conform to the ML expectation of 'more data = more skill' in the presence of large volumes of training data (with experiment 4 results).

Experiments 2 and 3 show the effect of limiting the training data size on model learning from the three data partitions. In this data-limited context, each model shows improved performance as the training data representativeness of the target regime increases. This behavior is consistent with hydrologic expectations and compares well with K. Fang et al. (2021), who found higher NSE values when models were trained with a moderately diverse dataset even when the training sample size was held constant between heterogeneous and homogeneous cases. LSTM and δHBV lose more predictive skill than Graph WaveNet when the training dataset is limited in size, with models trained on M performing better than those trained on A. Furthermore, we find that model learning is considerably affected when the limited, diverse dataset is also constrained by excluding representative basins from the regions of interest. In this case, partition A becomes the least informative group, especially for LSTM: its ability to predict out-of-context is drastically reduced when it does not see any representative basins in training, as demonstrated by the drop in skill from Experiment 2 to Experiment 3.

The glacier-stratified analysis highlights a strong dependence of model performance on both glacier coverage and training configuration. Across experiments, all three models perform robustly in low-glacier basins, where hydrographs resemble snowmelt-dominated mountain catchments with a single early-summer peak. Even in weakly glacierized systems, training data design matters: random training sets are less able to sustain performance under extrapolation than partitions that preserve regional or glacier-related structure (M, G). As glacier influence increases, the models become far more sensitive to training design. This behavior indicates that neither physically inspired lumped models nor deep learning architectures can reliably generalize strongly glacierized regimes without explicit exposure to such basins during training. Together, these findings underscore that the training data must also contain basins spanning the relevant glacier coverage spectrum. In low-glacier basins, generic regional training seems adequate, but in glacier-dominated catchments the presence or absence of glacierized donors in the training set largely governs predictive skill. These patterns parallel the findings of Anderson and Radić (2022), who demonstrated for highly glacier intensive basins the internal states of a CNN-LSTM captured glacier runoff without explicit glacier information by associating temperature to streamflow during summer only in glacierized regions.

This also supports our finding that supplying glacier melt did not yield the expected improvements in streamflow skill, even in strongly glacier-influenced basins. If the model already captures these relationships implicitly through its learned states, additional glacier melt information may provide limited incremental benefit, even in strongly glacier-influenced basins. That is, it is more important to have climate-discharge observations in glacierized catchments than it is to have explicit representation of glacial process for ML. This perhaps counterintuitive outcome stems from several interacting issues: the temporal disaggregation method introduces noise relative to observed melt-season hydrographs, and systematic magnitude

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

biases in the glacier product propagates directly into forcings. When models lack representations of delay, storage, or routing for glacier-derived water, the raw melt signal can therefore amplify seasonal bias and produce more negative NSE outcomes rather than correcting them. Moreover, the direct addition of glacier runoff to precipitation could have inflated effective inputs and altered the model water balance.

Although the predictive skill is not as high as LSTM, both the Graph WaveNet and the δHBV show more consistent skill due to their structural constraints; the LSTM loses skill rapidly when G basins are removed from training. This maintains a consistent performance for the two models when the training dataset is limited in size and hydroclimatic representativeness to the target region. Among the three models, LSTM in experiment 1 achieved the highest predictive skill across all experiments. LSTM does not have any structural constraint, so it learns complex relationships among the variables from the data. This is apparent in the A and M partitions with larger diversity and larger size as there is variation in data and larger data to generate the multi-dimensional relationship. As δHBV is an evolved HBV model with LSTM parameterizing the physical model, the HBV model structure constrains the model learning. Likewise, Graph WaveNet defines a prior neighborhood of the basins from the static attributes and therefore integrates information from only the relevant basins. Therefore, the choice of the model depends on the objective of a task as well as availability of training data, i.e. if the focus is on attaining highest predictive skill and a large training data is available, then, LSTM has the capacity to learn best from large heterogeneous data, whereas, if we have limited data in the source dataset, δHBV or Graph WaveNet has the better capability.

Table 6
Time taken (in mins) to train a single epoch of each model for each training configuration

	All Global (A)			Mixed High Mountain (M)			High Mountain with Glacier (G)		
	LSTM	**δHBV**	**Graph WaveNet**	**LSTM**	**δHBV**	**Graph WaveNet**	**LSTM**	**δHBV**	**Graph WaveNet**
Exp 1	75	102	420	30	35	102	7	8	7
Exp 2	10	9	14	4	10	8	7	8	7
Exp 3	5	8	14	4	8	8	7	8	7
Exp 4	60	67	102	15	30	72	7	8	7

We are also interested in practical aspects of these methods, principally the amount of computing resources, time, and expert knowledge needed to deploy them. The δHBV proved the lightest weight architecture requiring the least amount of computation resources and LSTM required the least effort to get to run for our study (Table 6). LSTM, in most experiments, took less time to train a single epoch, however, δHBV could train the models around the same time as an LSTM with higher resources (Table 6), so the LSTM is a very time-efficient model. Moreover, neither the LSTM nor δHBV required GPUs with higher VRAM (using as low as 11GB for training). In contrast, Graph WaveNet training was only possible with 40G or more

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

VRAM per GPU, and even with these more powerful GPUs, the Graph WaveNet usually took twice or more the computation time of either of the other two models, except for training with 'G' partition, as it processes substantial amounts of spatial and temporal data at once (Table 6). In particular, training a Graph WaveNet model with maximum basins as in experiment 1A took around 7 hours for 1 epoch to complete (Table 6).

It is important to note that we have not compared the performance of our DL models against any traditional hydrological models, nor have we attempted to tune the model architectures or hyperparameters to glacial catchments. A lack of model comparison is practical- it would not be efficient to calibrate standard physical or conceptual models on each basin for each training set. Indeed, while the training times of the models described above are considerable, they are orders of magnitude less than calibrating for example, the VIC (Liang et al., 1994) model 5,600 times per experiment (for our 56 basin, k=100 fold cross validation). The use of archived global hydrological models as a foil is another possibility. Global models that produce daily discharge are rare, but Global Reach-Level Flood Reanalysis (GRFR) (Yang et al., 2021), Global Flood Awareness System (GloFAS) (Harrigan et al., 2020), Global River Discharge Reanalysis (GRDR) (Feng and Gleason, 2024), exist and provide daily discharges that cover our study basins and study period. However, these models are produced on either custom grids or river networks that use different forcing than our models. Qualitatively then, we can compare to published skill of global studies, recognizing our inability to control for differences. Feng et al., (2021) published daily skill scores at thousands of pan-Arctic gauges and report median NSE of about 0.27, with similarly around 80% positive NSE. The Arctic is another unique hydrological region with limited training data, and in this context, our DL results suggest similar if not better performance, but quantitative comparisons are difficult as the model in Feng et al. (2021) includes assimilated remotely sensed data and has different forcing. Feng & Gleason (2024) repeated Feng et al. (2021)'s method globally and validated on more than 8,000 gauges, reporting a median NSE of approximately 0.36, showing improved performance globally vs the unique Arctic, consistent with our results. Similarly, Yang et al., (2025) show how training an LSTM with a set of small basins and predicting discharge at grid scale produces out-of-sample streamflow predictions with median KGE of 0.59 globally. Thus, we can conclude that our DL performance is similar to Feng et al. (2021) and Yang et al. (2025).

We also could have tuned our hyperparameters for optimal performance in glacierized catchments. This would have improved skill, but skill alone was not our goal; we sought to compare the off-the-shelf models to question their ability to predict discharge for ungauged glacierized regions with the given the training data setup and answer practical questions about their use as is. Future work could further these comparisons to cross compare whether modeling or DL provides higher skill and seek to find the highest skill possible. Finally, LSTM and Graph WaveNet were trained on a set of five dynamic meteorological variables, while δHBV, in contrast, was trained on only three such variables. LSTMs and Graph WaveNet can easily change inputs, but changing the δHBV here would require rewriting the underlying HBV model- a substantial task. Despite this discrepancy, δHBV exhibited commendable performance even with a more constrained set of inputs, and our implementation this way is consistent with our aim to assess out-of-the-box performance. Considering δHBV's robust performance with a

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

limited set of meteorological variables, there arises a compelling opportunity for further exploration. Future research could delve into, perhaps, incorporating a dedicated glacier module to uncover its full potential in capturing the complexities of glacierized high mountain regions. Finally, our use of ERA5-Land forcings—including its treatment of glacier grid points and snow water equivalent—introduces reanalysis biases that we do not correct. Because all three models are driven by the same ERA5-Land fields, these biases are shared across architectures and primarily affect the absolute interpretation of performance rather than the comparative conclusions we draw.

A further geographic limitation is that the Caravan dataset contains no basins from High Mountain Asia, the Himalayas, or the Karakoram — our G partition therefore comprises exclusively Alpine, Patagonian, and North American glaciers, and results may not generalize to the Hindu Kush-Himalayan system, which represents the largest glacierized area and the most densely affected downstream population globally. Finally, none of the three architectures evaluated includes an explicit glacier mass balance or melt module; this study therefore characterizes data-driven generalization skill under a realistic operational workflow — where externally-modeled glacier runoff is available as forcing — rather than process representation fidelity, and non-stationary glacier dynamics under future climate change are not captured by any model.

While this study evaluates daily regression skill rather than long-term climate projections, our findings have direct implications for DL model selection in climate impact workflows for glacierized basins. Reliable long-term projection of glacier runoff change under climate change presupposes a model that first generalizes dependably to ungauged glacierized catchments under historical conditions — a prerequisite that our results show is non-trivial. Models trained without glacierized basin representation fail substantially even for historical daily simulation, and sensitivity to training data composition intensifies with glacier coverage fraction, precisely the basins most critical for future water resource assessment. Any DL-based climate impact workflow applied to ungauged glacierized catchments must therefore account for the training data conditions we characterize here before projections can be trusted. Aguayo et al. (2025) demonstrate that a full pipeline — from LSTM ungauged prediction to GCM-forced glacio-hydrological projection — is feasible for Patagonian glacierized basins, but their approach depends critically on the kind of historical generalization characterization this study provides. We acknowledge that the connection between daily model fidelity and long-term climate robustness under non-stationary glacier dynamics remains an open challenge across the field (Thébault et al., 2026), and resolving it will require explicit coupling with glacier evolution models and targeted evaluation under synthetic climate warming scenarios — a high-priority direction for future work. Beyond this immediate next step, three directions merit priority. First, extending this experimental framework to Asian glacier basins — through either new open data partnerships or remote sensing-derived discharge proxies for High Mountain Asia — would test whether our training data representativeness findings generalize to the Hindu Kush-Himalayan system, which our dataset does not include but which is hydrologically the most consequential glacierized region globally. Second, evaluating temporal robustness of all three architectures under synthetic climate warming scenarios would directly connect our daily generalization findings to climate impact assessment, addressing whether models that generalize historically remain reliable under the non-

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

stationary glacier dynamics projected for coming decades (Aguayo et al., 2025; Thébault et al., 2026). Third, the architecture comparison could be extended to attention-based models; because their demonstrated advantages are confined to long-horizon autoregressive and zero-shot forecasting rather than the daily regression task studied here (Liu et al., 2025b), this extension would chiefly establish whether that task-dependence also holds in the glacierized ungauged setting, where training-data representativeness rather than architecture governs skill.

Large diverse datasets provide a greater opportunity for model architectures to learn general hydrological behavior, particularly for LSTM. However, when the data is limited by size, the representativeness in the training data relative to target region provides the strongest training setup for the models to learn from. This representativeness in the training data, when limited in size, is especially important for LSTM. Although both δHBV and Graph WaveNet did not attain the same skill as LSTM, both models' structures provide a limited degree of freedom preventing the models from overfitting the data. This leads to a recommendation for the LSTM as the tool of choice for the out-of-context hydrologist, followed by δHBV, unless there is truly a limited amount of training data available (which suggests Graph WaveNet). However, in practice the Caravan data exist and are always available, so unless bespoke inputs are needed it is likely that the LSTM remains the tool of choice for current ML in discharge prediction.

6. Conclusions

Our study sheds light on the comparative performance of three deep learning models—Long Short-Term Memory networks (LSTM), Graph Neural Networks (GNN) - Graph WaveNet, and differentiable parameter learning - differentiable HBV (δHBV)—in simulating discharge in ungauged glacierized high mountain regions under varying data volumes and representativeness in training data structures. The findings underscore the nuanced strengths of each model, with Graph WaveNet being effective in homogenous data settings, whereas LSTM displaying robust performance with larger datasets, and δHBV demonstrating promising outcomes even with a more restricted set of inputs. We aimed for a direct, off-the-shelf comparison, and thus all models were placed in direct service of out-of-context prediction without modification. The better performance of δHBV with fewer variables suggests the potential for enhancing its capabilities, through the addition of a glacier module, encouraging further exploration in harnessing the power of deep learning for precise hydrological predictions in glacierized regions. Moreover, all three of the models demonstrate strong potential for predicting streamflow for out-of-sample in both space and time for glacierized high mountain basins, but their success depends critically on the structure and composition of the training data. Models trained without exposure to glacierized basins struggle to generalize, particularly in strongly glacierized basins, highlighting the importance of including a wide range of basin types across glacier spectrum in training datasets to capture key hydrological processes. This training representation was more important than glacial inputs to generate skill. This study serves as a guide for future workers seeking to harness machine learning in hydrology, especially in contexts of limited data and where hydrological processes are intricate. We suggest that using the Caravan to train an LSTM

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

is the best place to start, as our results suggest this will yield the highest skill. If the region of interest is not represented in Caravan, our results suggest a Graph WaveNet or δHBV will yield better results, but the Graph WaveNet will require substantially more resources than either the LSTM or δHBV.

Code and data availability

The original Caravan dataset can be downloaded from (https://doi.org/10.5281/zenodo.7540792). The Caravan extension version of CAMELS-CH can be found at (https://doi.org/10.5281/zenodo.7784632). Glacier runoff estimates were derived from the global glacier model output of Rounce et al. (2023; https://doi.org/10.1126/science.abo1324) and obtained directly from the authors.

Author contributions

MM: data curation, software, methodology, formal analysis, investigation, visualization, writing – original draft. CJG: conceptualization, project administration, supervision, writing – review & editing. CB: conceptualization, writing – review & editing. All authors approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Financial support

This research has been supported by funding from the NASA HiMAT grant.

References

Aboelyazeed, D., Xu, C., Hoffman, F. M., Liu, J., Jones, A. W., Rackauckas, C., Lawson, K., and Shen, C.: A differentiable, physics-informed ecosystem modeling and learning framework for large-scale inverse problems: demonstration with photosynthesis simulations, Biogeosciences, 20, 2671–2692, https://doi.org/10.5194/bg-20-2671-2023, 2023.

Ackroyd, C., Skiles, S. M., Rittger, K., and Meyer, J.: Trends in Snow Cover Duration Across River Basins in High Mountain Asia From Daily Gap-Filled MODIS Fractional Snow Covered Area, Frontiers in Earth Science, 9, 2021.

Aguayo, R., Zekollari, H., Hanus, S., Baez-Villanueva, O. M., Mendoza, P. A., and Maussion, F.: Hybrid Glacio-Hydrological Modeling Reveals Contrasting Runoff Changes in Western Patagonia Over the 21st Century, Earth's Future, 13, e2025EF006442, https://doi.org/10.1029/2025EF006442, 2025a.

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

Aguayo, R., Zekollari, H., Hanus, S., Baez-Villanueva, O. M., Mendoza, P. A., and Maussion, F.: Hybrid Glacio-Hydrological Modeling Reveals Contrasting Runoff Changes in Western Patagonia Over the 21st Century, Earth's Future, 13, e2025EF006442, https://doi.org/10.1029/2025EF006442, 2025b.

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

Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F., and Mai, J.: Continuous streamflow prediction in ungauged basins: long short-term memory neural networks clearly outperform traditional hydrological models, Hydrology and Earth System Sciences, 27, 139–157, https://doi.org/10.5194/hess-27-139-2023, 2023.

Aryal, A., Bhusal, A., and Kalra, A.: Evaluating Dry and Wet Season Precipitation from Remotely Sensed Data Using Artificial Neural Networks for Floodplain Mapping in an Ungauged Watershed, Environmental Protection Research, 150–165, https://doi.org/10.37256/epr.3120232255, 2023.

Azam, M., Wagnon, P., Vincent, C., Ramanathan, A., Kumar, N., Srivastava, S., Pottakkal, J. G., and Chevallier, P.: Snow and ice melt contributions in a highly glacierized catchment of Chhota Shigri Glacier (India) over the last five decades, Journal of Hydrology, https://doi.org/10.1016/J.JHYDROL.2019.04.075, 2019.

Azam, M. F., Wagnon, P., Vincent, C., Ramanathan, A. L., Favier, V., Mandal, A., and Pottakkal, J. G.: Processes governing the mass balance of Chhota Shigri Glacier (western Himalaya, India) assessed by point-scale surface energy balance measurements, The Cryosphere, 8, 2195–2217, https://doi.org/10.5194/tc-8-2195-2014, 2014.

Azam, Mohd. F., Kargel, J. S., Shea, J. M., Nepal, S., Haritashya, U. K., Srivastava, S., Maussion, F., Qazi, N., Chevallier, P., Dimri, A. P., Kulkarni, A. V., Cogley, J. G., and Bahuguna, I.: Glaciohydrology of the Himalaya-Karakoram, Science, 373, eabf3668, https://doi.org/10.1126/science.abf3668, 2021.

Bindas, T., Tsai, W.-P., Liu, J., Rahmani, F., Feng, D., Bian, Y., Lawson, K., and Shen, C.: Improving River Routing Using a Differentiable Muskingum-Cunge Model and Physics-Informed Machine Learning, Water Resources Research, 60, e2023WR035337, https://doi.org/10.1029/2023WR035337, 2024.

Bocchiola, D., Mihalcea, C., Diolaiuti, G., Mosconi, B., Smiraglia, C., and Rosso, R.: Flow prediction in high altitude ungauged catchments: A case study in the Italian Alps (Pantano Basin, Adamello Group), Advances in Water Resources, 33, 1224–1234, https://doi.org/10.1016/j.advwatres.2010.06.009, 2010.

Bolch, T., Kulkarni, A., Kääb, A., Huggel, C., Paul, F., Cogley, J. G., Frey, H., Kargel, J. S., Fujita, K., Scheel, M., Bajracharya, S., and Stoffel, M.: The State and Fate of Himalayan Glaciers, Science, 336, 310–314, https://doi.org/10.1126/science.1215828, 2012.

Bolibar, J., Rabatel, A., Gouttevin, I., Galiez, C., Condom, T., and Sauquet, E.: Deep learning applied to glacier evolution modelling, Cryosphere, 14, 565–584, https://doi.org/10.5194/tc-14-565-2020, 2020.

Bolibar, J., Rabatel, A., Gouttevin, I., Zekollari, H., and Galiez, C.: Nonlinear sensitivity of glacier mass balance to future climate change unveiled by deep learning, Nature Communications, 13, https://doi.org/10.1038/s41467-022-28033-0, 2022.

Boodoo, F., Hostache, R., Skifa, N., Guerin, J., and Delenne, C.: Are LSTM and conceptual rainfall-runoff models able to cope with limited training datasets under diverse hydrometeorological conditions?, Model. Earth Syst. Environ., 11, 128, https://doi.org/10.1007/s40808-025-02316-z, 2025.

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P.: Geometric Deep Learning: Going beyond Euclidean data, IEEE Signal Processing Magazine, 34, 18–42, https://doi.org/10.1109/MSP.2017.2693418, 2017.

Brown, R. D. and Mote, P. W.: The Response of Northern Hemisphere Snow Cover to a Changing Climate, Journal of Climate, 22, 2124–2145, https://doi.org/10.1175/2008JCLI2665.1, 2009.

Casassa, G., López, P., Pouyaud, B., and Escobar, F.: Detection of changes in glacial run-off in alpine basins: examples from North America, the Alps, central Asia and the Andes, Hydrological Processes, 23, 31–41, https://doi.org/10.1002/hyp.7194, 2009.

Chen, X., Wang, S., Gao, H., Huang, J., Shen, C., Li, Q., Qi, H., Zheng, L., and Liu, M.: Comparison of deep learning models and a typical process-based model in glacio-hydrology simulation, Journal of Hydrology, 615, 128562, https://doi.org/10.1016/j.jhydrol.2022.128562, 2022.

Chiogna, G., Marcolini, G., Liu, W., Pérez Ciria, T., and Tuo, Y.: Coupling hydrological modeling and support vector regression to model hydropeaking in alpine catchments, Science of The Total Environment, 633, 220–229, https://doi.org/10.1016/j.scitotenv.2018.03.162, 2018.

Fang, G., Yang, J., Chen, Y., Li, Z., Ji, H., and De Maeyer, P.: How Hydrologic Processes Differ Spatially in a Large Basin: Multisite and Multiobjective Modeling in the Tarim River Basin, Journal of Geophysical Research: Atmospheres, 123, 7098–7113, https://doi.org/10.1029/2018JD028423, 2018.

Fang, K., Shen, C., Kifer, D., and Yang, X.: Prolongation of SMAP to Spatiotemporally Seamless Coverage of Continental U.S. Using a Deep Learning Neural Network, Geophysical Research Letters, 44, 11,030-11,039, https://doi.org/10.1002/2017GL075619, 2017.

Fang, K., Kifer, D., Lawson, K., Feng, D., and Shen, C.: The data synergy effects of time-series deep learning models in hydrology, 2021a.

Fang, Z., Wang, Y., Peng, L., and Hong, H.: Predicting flood susceptibility using LSTM neural networks, Journal of Hydrology, 594, 125734, https://doi.org/10.1016/J.JHYDROL.2020.125734, 2021b.

Feng, D. and Gleason, C. J.: More flow upstream and less flow downstream: The changing form and function of global rivers, Science, 386, 1305–1311, https://doi.org/10.1126/science.adl5728, 2024.

Feng, D., Fang, K., and Shen, C.: Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks With Data Integration at Continental Scales, Water Resources Research, 56, https://doi.org/10.1029/2019WR026793, 2020.

Feng, D., Lawson, K., and Shen, C.: Mitigating Prediction Error of Deep Learning Streamflow Models in Large Data-Sparse Regions With Ensemble Modeling and Soft Data, Geophysical Research Letters, 48, https://doi.org/10.1029/2021GL092999, 2021.

Feng, D., Beck, H., Lawson, K., and Shen, C.: The suitability of differentiable, learnable hydrologic models for ungauged regions and climate change impact assessment, https://doi.org/10.5194/hess-2022-245, 2022.

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

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

Fountain, A. G. and Tangborn, W. V.: The Effect of Glaciers on Streamflow Variations, Water Resources Research, 21, 579–586, https://doi.org/10.1029/WR021i004p00579, 1985.

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

Frenierre, J. L. and Mark, B. G.: A review of methods for estimating the contribution of glacial meltwater to total watershed discharge, Progress in Physical Geography: Earth and Environment, 38, 173–200, https://doi.org/10.1177/0309133313516161, 2014.

Gleason, C. J. and Hamdan, A. N.: Crossing the (watershed) divide: satellite data and the changing politics of international river basins, The Geographical Journal, 183, 2–15, https://doi.org/10.1111/geoj.12155, 2017.

Gleason, C. J. and Smith, L. C.: Toward global mapping of river discharge using satellite images and at-many-stations hydraulic geometry, Proc. Natl. Acad. Sci. U.S.A., 111, 4788–4791, https://doi.org/10.1073/pnas.1317606111, 2014.

Guo, Y., Zhang, Y., Zhang, L., and Wang, Z.: Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review, WIREs Water, 8, e1487, https://doi.org/10.1002/wat2.1487, 2021.

Gurtz, J., Zappa, M., Jasper, K., Lang, H., Verbunt, M., Badoux, A., and Vitvar, T.: A comparative study in modelling runoff and its components in two mountainous catchments, Hydrological Processes, 17, 297–311, https://doi.org/10.1002/hyp.1125, 2003.

Hao, H., Hao, Y., Li, Z., Qi, C., Wang, Q., Zhang, M., Liu, Y., Liu, Q., and Jim Yeh, T.-C.: Insight into glacio-hydrologicalprocesses using explainable machine-learning (XAI) models, Journal of Hydrology, 634, 131047, https://doi.org/10.1016/j.jhydrol.2024.131047, 2024.

Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., Barnard, C., Cloke, H., and Pappenberger, F.: GloFAS-ERA5 operational global river discharge reanalysis 1979–present, Earth System Science Data, 12, 2043–2060, https://doi.org/10.5194/essd-12-2043-2020, 2020.

Hochreiter, S. and Schmidhuber, J.: Long Short-term Memory, Neural computation, 9, 1735–80, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Höge, M., Kauzlaric, M., Siber, R., Schönenberger, U., Horton, P., Schwanbeck, J., Floriancic, M. G., Viviroli, D., Wilhelm, S., Sikorska-Senoner, A. E., Addor, N., Brunner, M., Pool, S., Zappa, M., and Fenicia, F.: CAMELS-CH: hydro-meteorological time series and landscape attributes for 331 catchments in hydrologic Switzerland, Earth System Science Data Discussions, 1–46, https://doi.org/10.5194/essd-2023-127, 2023.

Hsu, K., Gupta, H. V., and Sorooshian, S.: Artificial Neural Network Modeling of the Rainfall-Runoff Process, Water Resources Research, 31, 2517–2530, https://doi.org/10.1029/95WR01955, 1995.

Hu, C., Wu, Q., Li, H., Jian, S., Li, N., and Lou, Z.: Deep Learning with a Long Short-Term Memory Networks Approach for Rainfall-Runoff Simulation, Water 2018, Vol. 10, Page 1543, 10, 1543, https://doi.org/10.3390/W10111543, 2018.

Huss, M., Farinotti, D., Bauder, A., and Funk, M.: Modelling runoff from highly glacierized alpine drainage basins in a changing climate, Hydrological Processes, 22, 3888–3902, https://doi.org/10.1002/hyp.7055, 2008.

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

Huss, M., Bookhagen, B., Huggel, C., Jacobsen, D., Bradley, R. s., Clague, J. j., Vuille, M., Buytaert, W., Cayan, D. r., Greenwood, G., Mark, B. g., Milner, A. m., Weingartner, R., and Winder, M.: Toward mountains without permanent snow and ice, Earth's Future, 5, 418–435, https://doi.org/10.1002/2016EF000514, 2017.

Immerzeel, W. W., Lutz, A. F., Andrade, M., Bahl, A., Biemans, H., Bolch, T., Hyde, S., Brumby, S., Davies, B. J., Elmore, A. C., Emmer, A., Feng, M., Fernández, A., Haritashya, U., Kargel, J. S., Koppes, M., Kraaijenbrink, P. D. A., Kulkarni, A. V., Mayewski, P. A., Nepal, S., Pacheco, P., Painter, T. H., Pellicciotti, F., Rajaram, H., Rupper, S., Sinisalo, A., Shrestha, A. B., Viviroli, D., Wada, Y., Xiao, C., Yao, T., and Baillie, J. E. M.: Importance and vulnerability of the world's water towers, Nature, 577, 364–369, https://doi.org/10.1038/s41586-019-1822-y, 2019.

Jansson, P., Hock, R., and Schneider, T.: The concept of glacier storage: a review, Journal of Hydrology, 282, 116–129, https://doi.org/10.1016/S0022-1694(03)00258-0, 2003.

Ji, H., Chen, Y., Fang, G., Li, Z., Duan, W., and Zhang, Q.: Adaptability of machine learning methods and hydrological models to discharge simulations in data-sparse glaciated watersheds, Journal of Arid Land, 13, 549–567, 2021.

Ji, H., Song, Y., Bindas, T., Shen, C., Yang, Y., Pan, M., Liu, J., Rahmani, F., Abbas, A., Beck, H., Lawson, K., and Wada, Y.: Distinct hydrologic response patterns and trends worldwide revealed by physics-embedded learning, Nat Commun, 16, 9169, https://doi.org/10.1038/s41467-025-64367-1, 2025.

Kaser, G., Grosshauser, M., and Marzeion, B.: Contribution potential of glaciers to water availability in different climate regimes, Proc Natl Acad Sci U S A, 107, 20223–20227, https://doi.org/10.1073/pnas.1008162107, 2010.

Koboltschnig, G. R. and Schöner, W.: The relevance of glacier melt in the water cycle of the Alps: the example of Austria, Hydrology and Earth System Sciences, 15, 2039–2048, https://doi.org/10.5194/hess-15-2039-2011, 2011.

Koch, J.: Caravan extension Denmark - Danish dataset for large-sample hydrology (v_03), https://doi.org/10.5281/zenodo.6762361, 2022.

Kratzert, F., Klotz, D., Herrnegger, M., and Hochreiter, S.: A glimpse into the Unobserved: Runoff simulation for ungauged catchments with LSTMs, 2018.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, Water Resources Research, 55, 11344–11354, https://doi.org/10.1029/2019WR026065, 2019.

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

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., and Matias, Y.: Caravan - A global community dataset for large-sample hydrology, Sci Data, 10, 61, https://doi.org/10.1038/s41597-023-01975-w, 2023.

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Kumar Sahu, R., Greve, P., Slater, L., and Dadson, S. J.: Hydrological concept formation inside long short-term memory (LSTM) networks, Hydrol. Earth Syst. Sci., 26, 3079–3101, https://doi.org/10.5194/hess-26-3079-2022, 2022.

Li, X.: Data Driven Discoveries in Streamflow, Vadose Zone, and Baseflow, Ph.D., University of Minnesota, United States -- Minnesota, 266 pp., 2023.

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

Liang, X., Lettenmaier, D. P., Wood, E. F., and Burges, S. J.: A simple hydrologically based model of land surface water and energy fluxes for general circulation models, Journal of Geophysical Research: Atmospheres, 99, 14415–14428, https://doi.org/10.1029/94JD00483, 1994.

Litt, M., Shea, J., Wagnon, P., Steiner, J., Koch, I., Stigter, E., and Immerzeel, W.: Glacier ablation and temperature indexed melt models in the Nepalese Himalaya, Sci Rep, 9, 5264, https://doi.org/10.1038/s41598-019-41657-5, 2019.

Liu, B., Yun, X., Pan, B., Xu, X., Gaffney, P. P. J., Lu, H., Luo, L., Sun, G., and Tang, Q.: Assess the impacts of climatic change and human activities on streamflow and floods by using a hybrid-physics-data (HPD) model: A case study in the Lancang-Mekong River Basin, Journal of Hydrology: Regional Studies, 61, 102763, https://doi.org/10.1016/j.ejrh.2025.102763, 2025a.

Liu, J., Shen, C., O'Donncha, F., Song, Y., Zhi, W., Beck, H. E., Bindas, T., Kraabel, N., and Lawson, K.: From RNNs to Transformers: benchmarking deep learning architectures for hydrologic prediction, Hydrology and Earth System Sciences, 29, 6811–6828, https://doi.org/10.5194/hess-29-6811-2025, 2025b.

Maier, H., Jain, A., Dandy, G., and Sudheer, K.: Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions, Environ. Model. Softw., 25, 891–909, https://doi.org/10.1016/j.envsoft.2010.02.003, 2010.

Maier, H. R. and Dandy, G. C.: The Use of Artificial Neural Networks for the Prediction of Water Quality Parameters, Water Resources Research, 32, 1013–1022, https://doi.org/10.1029/96WR03529, 1996.

Maier, H. R. and Dandy, G. C.: Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications, Environmental Modelling & Software, 15, 101–124, https://doi.org/10.1016/S1364-8152(99)00007-9, 2000.

Marta, S., Azzoni, R. S., Fugazza, D., Tielidze, L., Chand, P., Sieron, K., Almond, P., Ambrosini, R., Anthelme, F., Alviz Gazitúa, P., Bhambri, R., Bonin, A., Caccianiga, M., Cauvy-Fraunié, S., Ceballos Lievano, J. L., Clague, J., Cochachín Rapre, J. A., Dangles, O., Deline, P., Eger, A., Cruz Encarnación, R., Erokhin, S., Franzetti, A., Gielly, L., Gili, F., Gobbi, M., Guerrieri, A., Hågvar, S., Khedim, N., Kinyanjui, R., Messager, E., Morales-Martínez, M. A., Peyre, G., Pittino, F., Poulenard, J., Seppi, R., Chand Sharma, M., Urseitova, N., Weissling, B., Yang, Y., Zaginaev, V., Zimmer, A., Diolaiuti, G. A., Rabatel, A., and Ficetola, G. F.: The Retreat of Mountain Glaciers since the Little Ice Age: A Spatially Explicit Database, Data, 6, 107, https://doi.org/10.3390/data6100107, 2021.

van der Meer, M., Zekollari, H., Huss, M., Bolibar, J., Sjursen, K. H., and Farinotti, D.: A minimal machine-learning glacier mass balance model, The Cryosphere, 19, 805–826, https://doi.org/10.5194/tc-19-805-2025, 2025.

Mejía-Veintimilla, D., Ochoa-Cueva, P., Samaniego-Rojas, N., Félix, R., Arteaga, J., Crespo, P., Oñate-Valdivieso, F., and Fries, A.: River Discharge Simulation in the High Andes of Southern Ecuador Using High-Resolution Radar Observations and Meteorological Station Data, Remote Sensing, 11, 2804, https://doi.org/10.3390/rs11232804, 2019.

Mohammadi, B., Gao, H., Pilesjö, P., Tuo, Y., Guo, R., and Duan, Z.: Integrating machine learning with process-based glacio-hydrological model for improving the performance of runoff simulation in cold regions, Journal of Hydrology, 656, 132963, https://doi.org/10.1016/j.jhydrol.2025.132963, 2025.

Moore, R. D., Fleming, S. W., Menounos, B., Wheate, R., Fountain, A., Stahl, K., Holm, K., and Jakob, M.: Glacier change in western North America: influences on hydrology, geomorphic hazards and water quality, Hydrological Processes, 23, 42–61, https://doi.org/10.1002/hyp.7162, 2009.

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

Mosaffa, H., Pappenberger, F., Prudhomme, C., Chantry, M., Rüdiger, C., and Cloke, H.: A GNN routing module is all you need for LSTM Rainfall–Runoff models, Hydrology and Earth System Sciences, 30, 2079–2092, https://doi.org/10.5194/hess-30-2079-2026, 2026.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth System Science Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T. Y., Weitzner, D., and Matias, Y.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, https://doi.org/10.1038/s41586-024-07145-1, 2024.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What Role Does Hydrological Science Play in the Age of Machine Learning?, Water Resources Research, 57, e2020WR028091, https://doi.org/10.1029/2020WR028091, 2021.

Nepal, S., Flügel, W.-A., Krause, P., Fink, M., and Fischer, C.: Assessment of spatial transferability of process-based hydrological model parameters in two neighbouring catchments in the Himalayan Region, Hydrological Processes, 31, 2812–2826, https://doi.org/10.1002/hyp.11199, 2017.

Ougahi, J. H. and Rowan, J. S.: Enhanced streamflow forecasting using hybrid modelling integrating glacio-hydrological outputs, deep learning and wavelet transformation, Sci Rep, 15, 2762, https://doi.org/10.1038/s41598-025-87187-1, 2025.

Painter, S. L. and Destouni, G.: Hydrology in the Age of Artificial Intelligence: From Fragmentation to Coherent Terrestrial Hydrosphere Science, Water Resources Research, 62, e2026WR043509, https://doi.org/10.1029/2026WR043509, 2026.

Peel, M. C. and Blöschl, G.: Hydrological modelling in a changing world, Progress in Physical Geography: Earth and Environment, 35, 249–261, https://doi.org/10.1177/0309133311402550, 2011.

Rabatel, A., Francou, B., Soruco, A., Gomez, J., Cáceres, B., Ceballos, J. L., Basantes, R., Vuille, M., Sicart, J.-E., Huggel, C., Scheel, M., Lejeune, Y., Arnaud, Y., Collet, M., Condom, T., Consoli, G., Favier, V., Jomelli, V., Galarraga, R., Ginot, P., Maisincho, L., Mendoza, J., Ménégoz, M., Ramirez, E., Ribstein, P., Suarez, W., Villacis, M., and Wagnon, P.: Current state of glaciers in the tropical Andes: a multi-century perspective on glacier evolution and climate change, The Cryosphere, 7, 81–102, https://doi.org/10.5194/tc-7-81-2013, 2013.

Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., and Shen, C.: Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data, Environ. Res. Lett., 16, 024025, https://doi.org/10.1088/1748-9326/abd501, 2021.

Razavi, T. and Coulibaly, P.: Streamflow Prediction in Ungauged Basins: Review of Regionalization Methods, Journal of Hydrologic Engineering, 18, 958–975, https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690, 2013.

Roth, A. and Liebig, T.: Forecasting Unobserved Node States with spatio-temporal Graph Neural Networks, in: 2022 IEEE International Conference on Data Mining Workshops (ICDMW), 2022 IEEE International Conference on Data Mining Workshops (ICDMW), 740–747, https://doi.org/10.1109/ICDMW58026.2022.00101, 2022.

Rounce, D. R., Hock, R., Maussion, F., Hugonnet, R., Kochtitzky, W., Huss, M., Berthier, E., Brinkerhoff, D., Compagno, L., Copland, L., Farinotti, D., Menounos, B., and McNabb, R. W.: Global glacier change in the 21st century: Every increase in temperature matters, Science, https://doi.org/10.1126/science.abo1324, 2023.

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

Saha, A. and Chandra Pal, S.: Application of machine learning and emerging remote sensing techniques in hydrology: A state-of-the-art review and current research trends, Journal of Hydrology, 632, 130907, https://doi.org/10.1016/j.jhydrol.2024.130907, 2024.

Scarselli, F., Gori, M., Ah Chung Tsoi, Hagenbuchner, M., and Monfardini, G.: The Graph Neural Network Model, IEEE Trans. Neural Netw., 20, 61–80, https://doi.org/10.1109/TNN.2008.2005605, 2009.

Schaner, N., Voisin, N., Nijssen, B., and Lettenmaier, D. P.: The contribution of glacier melt to streamflow, Environ. Res. Lett., 7, 034029, https://doi.org/10.1088/1748-9326/7/3/034029, 2012.

Shafeeque, M., Luo, Y., Wang, X., and Sun, L.: Altitudinal Distribution of Meltwater and Its Effects on Glacio-Hydrology in Glacierized Catchments, Central Asia, JAWRA Journal of the American Water Resources Association, 56, 30–52, https://doi.org/10.1111/1752-1688.12805, 2020.

Shahgedanova, M., Adler, C., Gebrekirstos, A., Grau, H. R., Huggel, C., Marchant, R., Pepin, N., Vanacker, V., Viviroli, D., and Vuille, M.: Mountain Observatories: Status and Prospects for Enhancing and Connecting a Global Community, mred, 41, A1, https://doi.org/10.1659/MRD-JOURNAL-D-20-00054.1, 2021.

Shen, C.: A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists, Water Resources Research, 54, 8558–8593, https://doi.org/10.1029/2018WR022643, 2018.

Shen, C., Chen, X., and Laloy, E.: Editorial: Broadening the Use of Machine Learning in Hydrology, Frontiers in Water, 3, 38, https://doi.org/10.3389/FRWA.2021.681023/BIBTEX, 2021.

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, F., Kifer, D., Li, L., Liu, X., Ren, W., Zheng, Y., Harman, C. J., Clark, M., Farthing, M., Feng, D., Kumar, P., Aboelyazeed, D., Rahmani, F., Song, Y., Beck, H. E., Bindas, T., Dwivedi, D., Fang, K., Höge, M., Rackauckas, C., Mohanty, B., Roy, T., Xu, C., and Lawson, K.: Differentiable modelling to unify machine learning and physical models for geosciences, Nat Rev Earth Environ, 4, 552–567, https://doi.org/10.1038/s43017-023-00450-9, 2023.

Singh, D., Juyal, V., and Sharma, V.: Consistent seasonal snow cover depth and duration variability over the Western Himalayas (WH), J Earth Syst Sci, 125, 1451–1461, https://doi.org/10.1007/s12040-016-0737-3, 2016.

Singh, P.: Glacier Hydrology, in: Encyclopedia of Snow, Ice and Glaciers, edited by: Singh, V. P., Singh, P., and Haritashya, U. K., Springer Netherlands, Dordrecht, 379–381, https://doi.org/10.1007/978-90-481-2642-2_193, 2011.

Sit, M., Demiray, B. Z., Xiang, Z., Ewing, G. J., Sermet, Y., and Demir, I.: A comprehensive review of deep learning applications in hydrology and water resources, Water Science and Technology, 82, 2635–2670, https://doi.org/10.2166/WST.2020.369, 2020.

Song, Y., Sawadekar, K., Frame, J. M., Pan, M., Clark, M. P., Knoben, W. J. M., Wood, A. W., Lawson, K. E., Patel, T., and Shen, C.: Physics-Informed, Differentiable Hydrologic Models for Capturing Unseen Extreme Events, Water Resources Research, 62, e2025WR040414, https://doi.org/10.1029/2025WR040414, 2026.

Stahl, K. and Moore, R. D.: Influence of watershed glacier coverage on summer streamflow in British Columbia, Canada, Water Resources Research, 42, https://doi.org/10.1029/2006WR005022, 2006.

Sun, A. Y., Jiang, P., Mudunuru, M. K., and Chen, X.: Explore Spatio-Temporal Learning of Large Sample Hydrology Using Graph Neural Networks, Water Resources Research, 57, https://doi.org/10.1029/2021WR030394, 2021.

Thébault, C., Knoben, W. J. M., Addor, N., Newman, A. J., and Clark, M. P.: Varying the Combination of Hydrological Models in Time and Space: Toward a More Accurate Representation of Streamflow in Large-Sample Hydrology, Water Resources Research, 62, e2025WR042272, https://doi.org/10.1029/2025WR042272, 2026.

van Tiel, M., Stahl, K., Freudiger, D., and Seibert, J.: Glacio-hydrological model calibration and evaluation, WIREs Water, 7, e1483, https://doi.org/10.1002/wat2.1483, 2020.

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., and Shen, C.: From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling, Nature Communications, 12, https://doi.org/10.1038/s41467-021-26107-z, 2020.

Vargo, L. J., Anderson, B. M., Dadić, R., Horgan, H. J., Mackintosh, A. N., King, A. D., and Lorrey, A. M.: Anthropogenic warming forces extreme annual glacier mass loss, Nat. Clim. Chang., 10, 856–861, https://doi.org/10.1038/s41558-020-0849-2, 2020.

Vinze, P. and Azam, Mohd. F.: On the transferability of snowmelt runoff model parameters: Discharge modeling in the Chandra-Bhaga Basin, western Himalaya, Frontiers in Water, 4, 2023.

Viviroli, D. and Weingartner, R.: The hydrological significance of mountains: from regional to global scale, Hydrol. Earth Syst. Sci., 8, 1017–1030, https://doi.org/10.5194/hess-8-1017-2004, 2004.

Viviroli, D., Weingartner, R., and Messerli, B.: Assessing the Hydrological Significance of the World's Mountains, mred, 23, 32–40, https://doi.org/10.1659/0276-4741(2003)023%5B0032:ATHSOT%5D2.0.CO;2, 2003.

Viviroli, D., Zappa, M., Gurtz, J., and Weingartner, R.: An introduction to the hydrological modelling system PREVAH and its pre- and post-processing-tools, Environmental Modelling & Software, 24, 1209–1222, https://doi.org/10.1016/j.envsoft.2009.04.001, 2009.

Viviroli, D., Kummu, M., Meybeck, M., Kallio, M., and Wada, Y.: Increasing dependence of lowland populations on mountain water resources, Nat Sustain, 3, 917–928, https://doi.org/10.1038/s41893-020-0559-9, 2020.

Wang, Y.-H.: Bridging the Gap Between the Physical-Conceptual Approach and Machine Learning for Modeling Hydrological Systems, Ph.D., The University of Arizona, United States -- Arizona, 184 pp., 2023.

Wi, S. and Steinschneider, S.: Assessing the Physical Realism of Deep Learning Hydrologic Model Projections Under Climate Change, Water Resources Research, 58, e2022WR032123, https://doi.org/10.1029/2022WR032123, 2022.

Willard, J. D., Varadharajan, C., Jia, X., and Kumar, V.: Time series predictions in unmonitored sites: a survey of machine learning techniques in water resources, Environmental Data Science, 4, e7, https://doi.org/10.1017/eds.2024.14, 2025.

Wu, Z., Pan, S., Long, G., Jiang, J., and Zhang, C.: Graph WaveNet for Deep Spatial-Temporal Graph Modeling, https://doi.org/10.48550/arXiv.1906.00121, 31 May 2019.

Yang, C., Xu, M., Kang, S., Fu, C., and Hu, D.: Improvement of streamflow simulation by combining physically hydrological model with deep learning methods in data-scarce glacial river basin, Journal of Hydrology, 625, 129990, https://doi.org/10.1016/j.jhydrol.2023.129990, 2023.

Yang, Y., Pan, M., Lin, P., Beck, H. E., Zeng, Z., Yamazaki, D., David, C. H., Lu, H., Yang, K., Hong, Y., and Wood, E. F.: Global Reach-Level 3-Hourly River Flood Reanalysis (1980–2019), Bulletin of the American Meteorological Society, 102, E2086–E2105, https://doi.org/10.1175/BAMS-D-20-0057.1, 2021.

https://doi.org/10.5194/egusphere-2026-3798
Preprint. Discussion started: 8 July 2026

Yang, Z., Bai, P., Tian, Y., and Liu, X.: Glacier Coverage Dominates the Response of Runoff and Its Components to Climate Change in the Tianshan Mountains, Water Resources Research, 61, e2024WR037947, https://doi.org/10.1029/2024WR037947, 2025.

Zappa, M., Badoux, A., and Gurtz, J.: The application of a complex distributed hydrological model in a highly glaciated alpine river catchment, 2000.

Zeng, J., Long, D., Zhang, Y., Ryu, D., Wigneron, J.-P., and Huang, Q.: Emerging remote sensing techniques for hydrological applications, Remote Sensing of Environment, 332, 115060, https://doi.org/10.1016/j.rse.2025.115060, 2026.

Zhang, B., Ouyang, C., Cui, P., Xu, Q., Wang, D., Zhang, F., Li, Z., Fan, L., Lovati, M., Liu, Y., and Zhang, Q.: Deep learning for cross-region streamflow and flood forecasting at a global scale, innovation, 5, https://doi.org/10.1016/j.xinn.2024.100617, 2024.

Zhang, L., Su, F., Yang, D., Hao, Z., and Tong, K.: Discharge regime and simulation for the upstream of major rivers over Tibetan Plateau, Journal of Geophysical Research: Atmospheres, 118, 8500–8518, https://doi.org/10.1002/jgrd.50665, 2013.