Hydrology Research

Hydrology Research Vol 57 No 3, 501 doi: 10.2166/nh.2026.119

Is smart sampling worth it? Impact of training data selection on the performance of LSTMs in streamflow prediction

Benedikt Heudorfer [a,*] and Ralf Loritz[b]

[a] Karlsruhe Institute of Technology (KIT), Institute of Meteorology and Climate Research – Atmospheric Trace Gases and Remote Sensing, Karlsruhe, Germany
[b] Karlsruhe Institute of Technology (KIT), Institute for Water and Environment, Karlsruhe, Germany
*Corresponding author. E-mail: benedikt.heudorfer@kit.edu

BH, 0000-0001-7801-9375

ABSTRACT

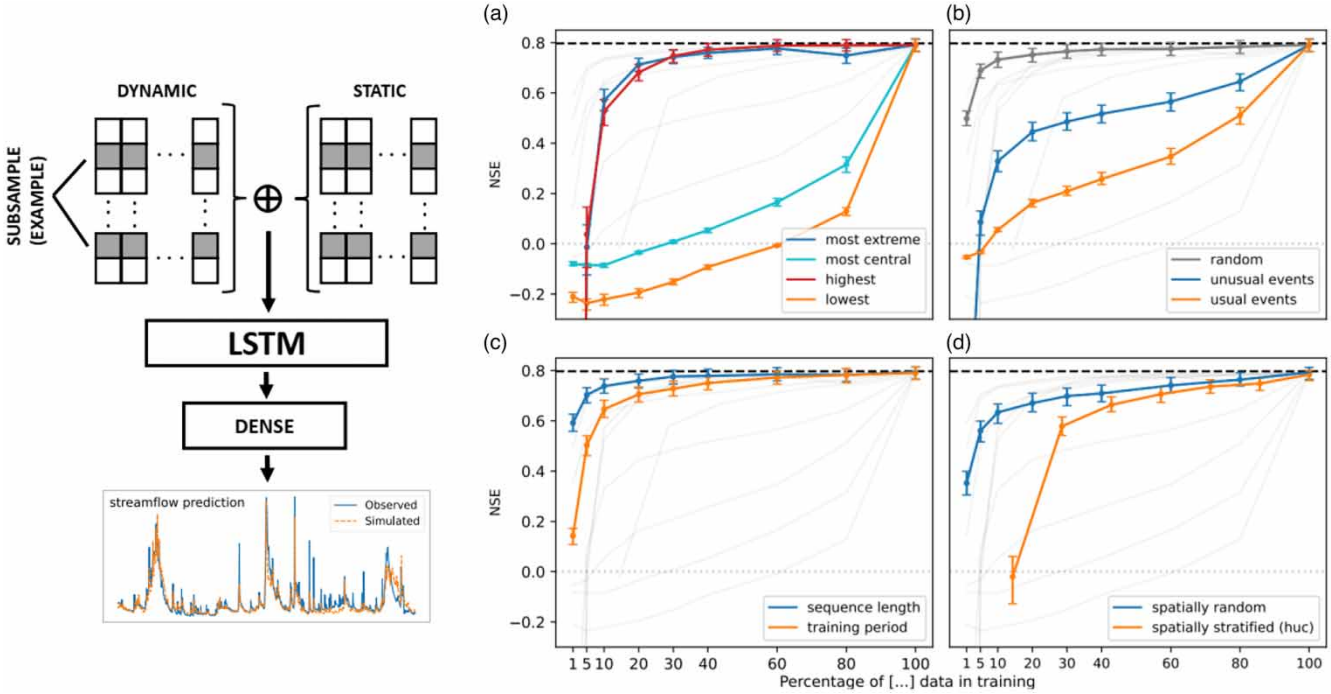
Deep learning models, particularly long short-term memory networks (LSTMs), have set new standards in streamflow prediction but require extensive data and computational resources. This raises a practical question: which parts of the data are truly indispensable, especially when computational budgets are limited or when only sparse observations are available? To address this, we examine which type of data is most essential in model training. We systematically ablate the CAMELS-US dataset using four families of sampling strategies – hydrological extremes, event rarity/statistical representativity, temporal context, and spatial representativity – and use the ablated datasets in training. Among all tested approaches, sampling based on statistical representativity via random sampling consistently outperformed more targeted strategies, achieving strong performance (NSE $>$ 0.7) and good representativity on as little as 10% of the data. Sampling hydrological extremes is the second-most efficient strategy, particularly when high-flow and low-flow extremes are sampled jointly, but with the largest performance gains stemming from high-flow events. Concerning temporal context, surprisingly short sequence lengths ($<$3 weeks) and training periods ($<$2 years) were sufficient for competitive performance (NSE $>$ 0.7). These findings provide practical guidance for efficient data selection in data-driven modeling and provide groundwork for future studies on training strategies.

Key words: CAMELS-US, deep learning, LSTM, sampling strategies, streamflow prediction

HIGHLIGHTS

- Demonstrates that random sampling outperforms targeted data sampling strategies for LSTM-based streamflow prediction.
- Reveals that high-flow data are information-dense but not solely sufficient, emphasizing the need for diverse flow conditions.
- Shows competitive model performance using only ~10% of randomly sampled data, $<$3-week sequence lengths, or $<$2-year records.
- Provides practical guidance for efficient data selection to reduce computational costs.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Recent advances in machine learning have significantly improved the accuracy of streamflow predictions up to the point that deep learning methods now represent the state of the art in terms of predictive capacity. While earlier works using artificial neural networks made good progress in hydrological modeling (Mount *et al.* 2016; Kratzert *et al.* 2018), the breakthrough only came with long short-term memory (LSTM) networks when feeding them a combination of static and dynamic information as input from multiple catchments simultaneously (Kratzert *et al.* 2019a, b). This model setup consistently outperforms conventional hydrological models such as the TOPMODEL, SWAT, SAC-SMA, and GloFAS by up to 10–30% (Feng *et al.* 2020; Kratzert *et al.* 2019a, 2021; Lees *et al.* 2021; Ma *et al.* 2021; Acuña Espinoza *et al.* 2024; Nearing *et al.* 2024). And while more recent attempts with Transformers (Liu *et al.* 2024) or Mamba (Eddin *et al.* 2025) can match the performance of the LSTM, so far they have not significantly exceeded it.

Deep learning models achieve this performance by leveraging large-scale datasets such as the CAMELS-US (Catchment Attributes and MEteorology for Large-sample Studies, Addor *et al.* 2017) or CARAVAN (named after a series of CAMELS, Kratzert *et al.* 2023) to predict streamflow across extensive sets of gauging stations. These increasingly large (and growing) datasets contain multi-decadal streamflow time series for hundreds to thousands of basins, accompanied by extensive sets of catchment attributes. Thus, it is computationally expensive – if not prohibitive – for many researchers to train on these datasets, especially CARAVAN, particularly if a hyperparameter search or model ensembles are implemented. This might explain why applications using e.g. CARAVAN remain relatively rare, and why we still see many single-basin (or few basin) studies despite the evidence that training LSTMs on single catchments is not advisable (Kratzert *et al.* 2024). At the same time, there is long-standing and growing evidence that the volume of data alone is not the sole determinant of predictive performance (Vrugt *et al.* 2002; Singh & Bárdossy 2012; Huo *et al.* 2019; Gauch *et al.* 2021b; Auer *et al.* 2024; Gupta 2024; Snieder & Khan 2025). Deep learning models do require substantial amounts of data to learn effectively, but with increasing training data size, performance gains diminish once a certain saturation is reached (Gauch *et al.* 2021b; Kratzert *et al.* 2024). Roscher *et al.* (2024) discussed that indiscriminately adding data can introduce noise, bias, and spurious correlations, harming generalization if data quality and task relevance are not addressed. Experiments on a harmonized dataset (Gupta 2024) underpin this, showing a saturation point in the number of training catchments, after which performance stagnates or decreases. Thus, Roscher *et al.* (2024) advocate a data-centric perspective that prioritizes dataset curation, evaluation, and quality over scale.

These insights are not limited to data-driven models. Also in process-based hydrological modeling, data quality and informativeness can prove more critical than data quantity. For example, Singh & Bárdossy (2012) showed that a relatively small, carefully selected subset of the available data was sufficient to calibrate a conceptual rainfall–runoff model. This finding suggests that large portions of hydrological time series contribute little additional information for learning the rainfall–runoff relationship within such models. Importantly, however, these results apply only to conceptual models. A well-initialized conceptual model already encodes substantial process knowledge, and therefore gains little from typical, uninformative days with average rainfall and runoff conditions. In contrast, data-driven models behave fundamentally differently: they start from random initialization and effectively operate as random functions before training. For these models, even the first few informative data points can induce large parameter updates and strongly shape the emerging representation. This distinction highlights that the relevance of individual time steps depends not only on hydrological information content but also on the model class and its initialization.

Nevertheless, the findings of Singh & Bárdossy (2012) are consistent with broader evidence that the information content of hydrological time series is highly heterogeneous, with only certain periods contributing meaningfully to learning or model refinement (Pool *et al.* 2019). Furthermore, recent works (e.g. Staudinger *et al.* 2025) explored the influence of data quantity on predictive performance by comparing process-based and statistical models, using smaller single-basin datasets. Their findings indicated that data-driven models, particularly LSTMs, outperform process-based approaches if enough data are available. However, the generalizability of their conclusions is constrained by the scale and scope of the datasets used. In addition, while Staudinger *et al.* (2025) focused on the intercomparison of different models, a rigorous analysis of sampling strategies would be helpful.

Together, these findings underscore that without increasing the informativeness of training data, simply expanding the training dataset does not necessarily enhance predictive skill. This highlights the central role of data selection in both conceptual and data-driven modeling approaches. It is especially relevant if one considers that increased dataset sizes lead to increased computational cost for large-scale training. As a result, we identify a growing interest in 'smart' sampling strategies that select sections of training data that exhibit higher information density, potentially enabling more efficient training. Identifying such 'smart' sampling strategies would be beneficial for modeling tasks such as model distillation, pre-training, and transfer learning. But to leverage these strategies effectively, we first need to better understand if there are characteristics of hydrological data that more significantly contribute to successful model training than others. The key questions are: how much data is truly necessary for effective training, and is there a lower limit below which performance notably degrades? Moreover, do we genuinely require extensive historical records spanning decades and multiple catchments, or could strategically selected subsets suffice? Identifying what makes certain data subsets more informative or valuable could enable targeted model initialization and cost-effective pre-training strategies.

In this study, we systematically examine data requirements of LSTMs by using various sampled subsets of the CAMELS-US dataset – characterized by hydrological extremity (high/low flows), event rarity and statistical representativity (usualness/unusualness/random), temporal context (sequence length/training period length), and spatial representativity (stratified/random) – and evaluate how subsampling impacts predictive performance and generalization across catchments. Ultimately, our objective is to offer practical guidance for more efficient training strategies and to contribute to ongoing discussions regarding data sufficiency in deep learning-driven hydrologic modeling.

2. METHODS

2.1. Data

In this study, we rely on the previously published dataset CAMELS-US (Newman *et al.* 2015; Addor *et al.* 2017), which provides long-term, catchment-scale hydroclimatic time series (dynamic features) as well as descriptive attributes (static features) for a total of 671 basins across the contiguous United States. From this dataset, we use the commonly adopted subset of 531 catchments defined by Newman *et al.* (2017) to ensure comparability with prior studies using this subset (e.g. Kratzert *et al.* 2021; Sun *et al.* 2021; Frame *et al.* 2022; Klotz *et al.* 2022; Liu *et al.* 2024; Heudorfer *et al.* 2025). As variables/features, we use streamflow as the target feature, the meteorological forcing variables listed in Table 1 as dynamic input features (i.e. time variant), and the catchment attributes listed in Table 1 as static input features (i.e. time invariant). Thereby, the static features were fed into the model alongside the dynamic features (for more details, see Section 2.2). To allow state-of-the-art model performance, we further adopt the training data makeup from Kratzert *et al.* (2021), who achieved benchmark performance

Table 1 | List of all dynamic and static input features used in this study

	Description	Unit
Dynamic input feature (meteorological forcing)		
t_max	Daily air temperature (maximum) at 2 m above ground	°C
t_min	Daily air temperature (minimum) at 2 m above ground	°C
prcp	Daily precipitation sum	mm
srad	Shortwave downward solar radiation	W/m^2
vp	Daily average near-surface vapor pressure	Pa
Static input feature (attribute)		
elev_mean	Catchment mean elevation	masl
slope_mean	Catchment mean slope	m/km
area_gages2	Catchment gages area (GAGESII estimate)	km^2
p_mean	Mean daily precipitation	mm/day
pet_mean	Mean daily PET (Priestley–Taylor estimate)	mm/day
aridity	Aridity Index: Ratio of mean PET to mean precipitation (PET/P)	–
frac_snow	Fraction of precipitation that presumably falls as snow (precipitation on days colder than 0 °C)	%
high_prec_freq	Frequency of high precipitation days (days with five times the mean daily precipitation or more)	days/year
high_prec_dur	Average duration of high precipitation events (average number of consecutive days with five times the mean daily precipitation or more)	days
low_prec_freq	Frequency of low precipitation days (days with less than 1 mm of precipitation)	days/year
low_prec_dur	Average duration of low precipitation days (average number of consecutive days with less than 1 mm of precipitation)	days
frac_forest	Fraction of the catchment area that is forested	%
lai_max	Maximum monthly mean of the leaf area index (LAI)	–
lai_diff	Difference between the maximum and minimum monthly mean of the LAI	–
gvf_max	Maximum monthly mean of the green vegetation fraction (GVF)	–
gvf_diff	Difference between the maximum and minimum monthly mean of the GVF	–
soil_depth_pelletier	Depth of soil from surface to bedrock according to Pelletier *et al.* (2016)	m
soil_depth_statsgo	Depth of soil according to the STATSGO model (Miller & White 1998)	m
soil_porosity	Volumentric porosity of the soil	–
soil_conductivity	Saturated hydraulic conductivity of the soil	cm/h
max_water_content	Maximum water content of the soil (estimated based on soil_porosity and soil_depth_statsgo)	m
sand_frac	Sand fraction of the soil (excluding soil material >2 mm)	%
silt_frac	Silt fraction of the soil (excluding soil material >2 mm)	%
clay_frac	Clay fraction of the soil (excluding soil material >2 mm)	%
carbonate_rocks_frac	Fraction of the catchment area characterized as 'carbonate sedimentary rocks'	%
geol_permeability	Subsurface permeability (log_{10})	m^2

Note: All data were taken from CAMELS-US (Newman *et al.* 2015; Addor *et al.* 2017). Note that for the dynamic input features, every three variants were used (Daymet, Maurer, and NLDAS, details see Section 2.1).

by including the variables precipitation, solar radiation, minimum temperature, maximum temperature, and vapor pressure from all three daily meteorological forcing datasets linked with the CAMELS-US dataset, namely the 'Daymet', 'NLDAS', and 'Maurer' forcing datasets (for details on these datasets, see Newman *et al.* 2015; Addor *et al.* 2017). This makes a total of 15

meteorological dynamic input features. As static features, we use 27 catchment attributes (Table 1) as a subset of the original CAMELS-US selection. We use this subset to allow strict comparability with previous studies, where the same subset was used (e.g. Kratzert *et al.* 2019a, b; Gauch *et al.* 2021a; Heudorfer *et al.* 2025). In one experiment, we split the territory of the contiguous US into seven sensible hydrologic macro-regions. These seven regions are a grouping of the original 18 hydrologic unit code (HUC) regional divisions (USGS & NRCS 2025). We choose this region grouping based on guidance given by Feng *et al.* (2023), who originally defined these regions based on hydrological similarity.

2.2. The LSTM model

The core of the model employed in this study is a LSTM network (Hochreiter & Schmidhuber 1997). LSTMs belong to the class of recurrent neural networks and maintain an internal state (called 'cell') that evolves with each new input. Additions and subtractions from the cell state are governed by a series of gating functions that regulate how information is incorporated, retained, or released. An input gate manages and regulates how much new information is added to the cell state, a forget gate determines how much of the previous state is preserved, and an output gate controls how the cell state contributes to the network output. Thereby, the cell state acts as memory that accumulates relevant information across time, functioning similarly to state variables in conceptual hydrological models.

Here, we adopt the LSTM in the same way as in Heudorfer *et al.* (2025), which employs it as a simple two-layer neural network. Following Kratzert *et al.* (2019a, b), static input features are repeated at every time step and simply concatenated to the dynamic features. All features are then standardized globally across all basins and then jointly fed into one LSTM layer. The LSTM layer, in turn, feeds into a fully connected output layer, which predicts discharge. This model setup is called entity-aware (Heudorfer *et al.* 2024, 2025), because it combines static and dynamic information in order to create spatially (entity-/catchment-specific) differentiated awareness of the dynamic input–output relation. Codewise, it is based on the implementation by Acuña Espinoza *et al.* (2025). Because our study deals exclusively with testing different data ablation procedures and data sampling strategies, no further model-specific development or hyperparameter tuning was implemented.

As a loss function, the basin-averaged NSE (Kratzert *et al.* 2019b) was used to train the model for 30 epochs, with the final epoch weights being used for testing. Model performance is primarily evaluated by the regular Nash–Sutcliffe Efficiency (NSE, Nash & Sutcliffe 1970). Additionally, the Kling–Gupta Efficiency (KGE, Gupta *et al.* 2009), high-flow volume bias of the top 2% flow values (FHV, Yilmaz *et al.* 2008), and low-flow volume bias of the bottom 30% values (FLV, Yilmaz *et al.* 2008) are calculated for supporting analysis. The training period is from 01 October 1999 to 30 September 2008, and the testing period is from 01 October 1989 to 30 September 1999. Excluding gaps, the training period contained approximately 1,652,300 samples, where each sample represents the combination of dynamical meteorological forcing (i.e. 365 days of 15 meteorological variables) and static features (i.e. 27 catchment characteristics) associated with each daily streamflow value.

All experiments utilize the same model setup to maintain a consistent architecture across all experiments. Following the bootstrapping procedure employed already in Heudorfer *et al.* (2025), each experiment (combination of sampling strategy and increment, see Section 2.3) was replicated with five different seeds for model weight initialization. Test period predictions from all seeds were pooled per catchment, from which 100 bootstrap samples (each comprising 80% of the pooled set) were drawn. We chose $B = 100$ bootstrap samples because the Monte-Carlo uncertainty of bootstrapped quantiles decreases $1/\sqrt{B}$ (Davison & Hinkley 1997), and since our pooled set of predictions is already large, we deemed 100 draws sufficient for stable estimates. In the end, for each bootstrap sample, performance metrics were calculated, resulting in distributions of metrics per catchment. Median, 5 and 95% quantiles were derived from these distributions and serve as performance estimates with uncertainty bounds. Additionally, throughout the results section, we use NSE = 0.7 as a threshold indicating good performance, following Moriasi *et al.* (2007). This is used as rough guidance, and not as a sharp pass/fail criterion, which is not recommended (Clark *et al.* 2021).

2.3. Experimental setup and sampling strategy

To better understand how the composition of training data influences predictive performance and generalization in LSTM-based streamflow models, we conducted systematic subsampling experiments. In each experiment, subsets of data are sampled at incremental percentages of the full dataset (1, 5, 10, 20, 30, 40, 60, and 80%). We chose progressively coarser increments to save computational time, since there is a diminishing return in performance gains with increasing dataset size.

We investigated four families of sampling strategies (Table 2). The first family aims to evaluate the effect of flow magnitude. This is done by looking at how hydrological extremes and average conditions affect predictive skill. For this, we designed four sampling strategies based on the distribution of streamflow (Table 2). In the second family, we explore how training on statistically representative hydrological events, i.e. typical (usual), versus atypical (unusual) events influences model prediction and generalization. To quantify event representativity, we used the identification of critical events (ICE) algorithm (Singh & Bárdossy 2012) based on the Tukey half-space depth function (Tukey 1975) derived from the multivariate distribution of streamflow and the antecedent precipitation index (see details in Supplementary Appendix Text A1). In the third family, we investigate the impact that the historical context and memory encoded in sequence length have on model training. We did this by systematically sampling temporal aspects of the time series, namely sequence length and training period. And

Table 2 | Overview of the sampling strategies used in the individual experiments with associated descriptions

Experiment	Experiment	Description
Hydrological extremes	Most extreme	Training subsets include an equal proportion of the lowest and highest streamflow values (e.g. the 10% subset comprises the lowest 5% and highest 5% of streamflow). This strategy prioritizes the inclusion of floods and droughts equally.
	Most central	Training subsets are selected around the mean streamflow. For instance, the 10% subset includes values within 5% above and below the mean, aiming to test predictive accuracy with exclusively normal conditions in training.
	Highest/lowest	Training subsets include only the highest (e.g. the 10% subset comprises only the 10% highest flows) or lowest (vice versa) streamflow values. These strategies exclusively emphasize either extreme, helping us understand how focused training on either end of the streamflow distribution impacts predictions.
Event rarity and statistical representativity	Random	Training samples are randomly drawn from the full dataset of all catchment, i.e. random draw independent of catchment affiliation. Serves as a baseline with a balanced representation of all flow conditions.
	Unusual events	Training samples are those with the lowest Tukey half-space data depth, i.e. on the outer rim (near the convex hull) of the point cloud spanned by streamflow and API, i.e. the most unusual samples (events). Prioritizes samples representing atypical, rare conditions (e.g. unusual rainfall–runoff relationships). See Supplementary Appendix Text A1 for details.
	Usual events	Training samples are those with the highest Tukey half-space data depth, i.e. at the center of the point cloud spanned by streamflow and API, i.e. the most usual samples (events). Prioritizes samples representing the most common or central hydrometeorological conditions. See Supplementary Appendix Text A1 for details.
Temporal context	Sequence length	Samples the length of the input sequence (or lookback window) of the LSTM, relative to the full, standard sequence length of 365 days. For example, using a 10% sequence length equals 37 days (rounded).
	Training period	Sample the overall length of the full historical training period, retaining only the most recent subset. For example, using 10% of the training period implies using 06 November 2007 to 30 September 2008 (329 days). The full length is 01 October 1999 to 30 September 2008 (3,287 days).
Spatial representativity	Spatially random	Training samples consist of catchments randomly drawn across space from the full dataset (531 catchments). For example, using 10% implies using the training data of 53 random catchments out of the full dataset of 531 catchments. This sampling strategy is different from normal random sampling because a full-time series of individual catchments is included when sampled.
	Spatially stratified (HUC regions)	Catchments were selected based on a macro-region classification derived from HUC, as specified in Supplementary Appendix Table A1. Individual macro-regions were selected for training, and testing was done over all 531 catchments in the dataset. This aimed at testing the importance of region-specific climatic and physiographic conditions during training. See Supplementary Appendix Text A2 for details about this stepwise sampling routine.

in the fourth and final family, we tested how spatial sampling influences the generalization capability of LSTM models across different geographic and hydrological settings.

2.4. Explicit analysis of sample representativity

In this study, we examine the predictive capacity of information contained in different samples of the CAMELS-US in an entity-aware LSTM for predicting streamflow across the whole dataset. From an information-theoretic perspective, this is the examination of how representative different samples are for the population in terms of information usable by the LSTM for streamflow prediction. In information theory, the Kullback–Leibler divergence (D_{KL}) is an explicit measure that allows quantification of sample representativity. D_{KL} is a measure of how one (sample) distribution Q diverges from a second, reference (population) distribution P (Weijs & Van De Giesen 2013), where higher D_{KL} values indicate higher divergence, and lower values indicate higher similarity, and a value of zero denotes identical distributions:

$$D_{KL}(Q|P) = \sum_x Q(x) \log\left(\frac{Q(x)}{P(x)}\right)$$

We use D_{KL} to explicitly calculate the divergence of the streamflow training samples derived by the different sampling strategies outlined in Section 2.3 from the full streamflow dataset. This is to add an additional layer of analysis, aimed at strengthening conclusions drawn from results in Section 3. Specifically, we use the freely available Python implementation in the UNITE toolbox published alongside the paper by Álvarez Chaves *et al.* (2024). Since the CAMELS-US dataset has frequent zero values for streamflow in the dataset (due to a couple of individual arid, ephemeral basins), k-nearest neighbor (kNN), as the most accurate distribution estimation method for D_{KL} (Álvarez Chaves *et al.* 2024), does not produce sensible D_{KL} quantities. Thus, we resort to binning as an estimation method, and determine bin width with the Freedman–Diaconis rule (for more information, see Álvarez Chaves *et al.* 2024). This choice is further justified by the comparative nature of our analysis, which only requires relative methodological consistency and not strictly optimal absolute accuracy in estimating D_{KL}.

3. RESULTS AND DISCUSSION

In this section, we show how the different sampling strategies affect model performance, providing results for each sampling strategy family in separate sections.

3.1. Effect of sampling hydrological extremes

Figure 1(a) shows the sampling strategies associated with the extremity of flows. When the highest flows are exclusively passed to the model, performance quickly approaches its maximum. While up to a sample size of 5%, the model still performs poorly (NSE < 0), only 20% of the total training dataset is needed to surpass the good performance threshold of NSE $= 0.7$. Since sampling both high and low flows together ('most extreme' experiment) closely follows the performance curve associated with high flow-only sampling, we assumed that the performance of the 'most extreme' sampling is mostly due to high flows, and less due to low flows. By comparison, low-flow sampling underperforms, since even up to a sample size of 80% of the total training dataset, model performance does not elevate above NSE $= 0.2$ (Table 3), albeit with significantly lower uncertainty. This means that low flows do not provide the model with useful information. Of course, this is surely in part an effect of using NSE as a loss function, since the NSE is more sensitive to high flows. We can contrast this by looking at the widely used (e.g. Kratzert *et al.* 2021; Frame *et al.* 2022) FLV metric (Supplementary Appendix Table A4), designed specifically for evaluating the fit of low flows (Yilmaz *et al.* 2008). However, even if only low flows are sampled, the FLV does not show a particularly good fit in low flows either. The FLV generally shows extremely high volatility and uncertainty, rendering its results not conclusive. Frame *et al.* (2022) already made the same point and deemed FLV too volatile for robust analysis. The FHV equally exhibits high uncertainty (Supplementary Appendix Table A3), but shows the best fit to high flows when high flow sampling is used. Strikingly, it is really only the absolute extreme values that prove useful in that case, as Supplementary Appendix Table A3 shows that the 1% high flow sample produces an outstanding FHV performance of -0.409, but sharply drops to the FHV $= 20$ mark in the next increment (5%), and stays there for all other sampling increments. The importance of really only looking at the absolute extreme values was also discussed recently in Baste *et al.* (2025).

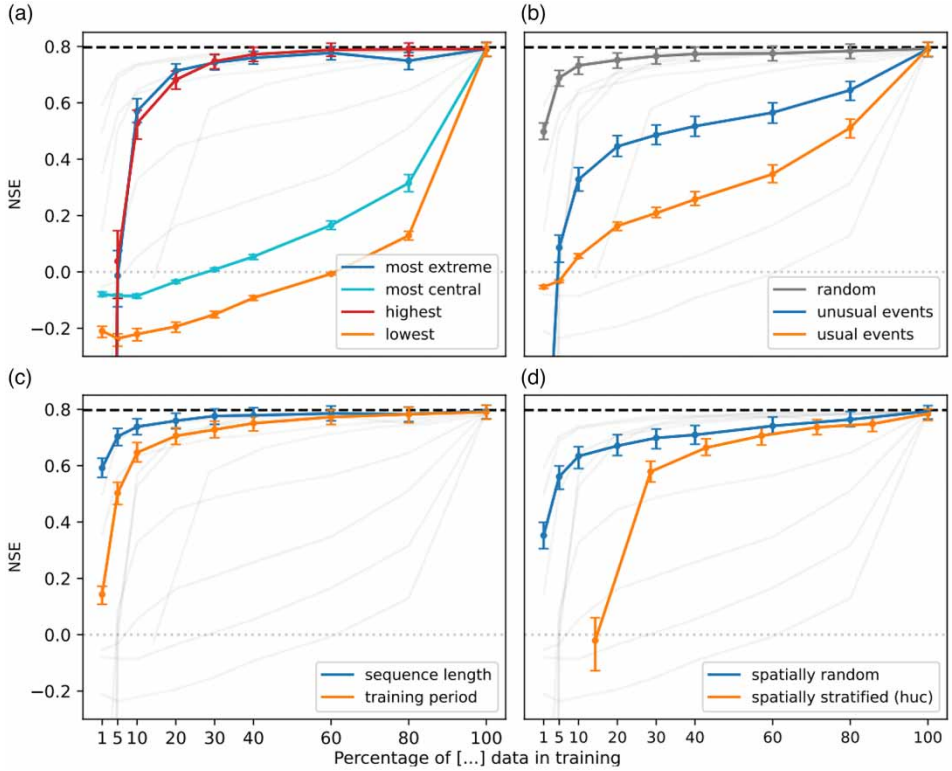


Figure 1 | NSE scores of all sampling strategy experiments (described in Table 2 and Section 2.3) across the sampled percentages of total train data. The dotted horizontal line is NSE = 0, the dashed horizontal line is the maximum possible performance for the total train data. NSE scores represent the median bootstrapped score value (see Section 2.2), with the uncertainty brackets representing the associated 5 and 95% bootstrapped quantiles. The *x*-axis locations of the spatially stratified experiment are the average sample size for the first, second through seventh region aggregation level (see Supplementary Appendix Text A2). Gray lines indicate the location of the performance in the respective other experiments for better orientation. The same plot but with KGE values is shown in Supplementary Appendix Table A2, with FHL values in Supplementary Appendix Table A3 and FLV values in Supplementary Appendix Table A4.

Furthermore, what's interesting in Figure 1(a) is that data grouped around the mean seems to have some information content for the model as well, since sampling of streamflow values grouped around the mean ('most central' experiment) performs slightly better compared with the low flow sampling experiment. This indicates strikingly that normal flow conditions are more relevant for model performance than low flows. It means that, even though information density is dissimilarly distributed across high, normal, and low flows, Figure 1(a) clearly shows that low and normal flows are not useless, but still contribute to performance, albeit on a lower magnitude. This is further underpinned by the fact that random sampling outperforms high flow sampling (Figure 1(b)). We can read this as strong support for the previously made suggestions to abandon the catchment-averaged NSE loss (Baste *et al.* 2025) in favor of more balanced approaches like quantile regression, pinball loss, or related strategies (e.g. Papacharalampous *et al.* 2019; Lamontagne *et al.* 2020; Jahangir *et al.* 2023) or probabilistic forecasting strategies (Gneiting & Katzfuss 2014), and contra the use of a high-flow-only loss function like frequently proposed in the flood forecasting community. This is true at least if balanced metrics of overall model fit are desired. If the focus is on a good high flow fit, metrics geared toward high flow fit (e.g. FHV, Supplementary Appendix Table A3) might prove to be better, as already noted by Mizukami *et al.* (2019).

3.2. Effect of sampling event rarity and statistical representativity

Figure 1(b) shows that the model performance of samples based on the Tukey half-space-based classification into usualness and unusualness of flow values underperforms most other sampling strategies (Figure 1(a)–1(d)): Both Tukey half-space-based sampling strategies only reach a threshold of NSE = 0.7 with a sample size of 80% (Table 3). The fact that sampling for unusualness does not produce performance competitive with e.g. high flow sampling is surprising. This is because extremes, particularly floods, are theoretically characterized by low data depth and should therefore make up a significant portion

Table 3 | Median NSE scores (and uncertainty in brackets) as shown in Figure 1

	1%	5%	10%	20%	30%	40%	60%	80%	100%
most extreme	– 5.599 (± 0.835)	– 0.011 (± 0.097)	0.571 (± 0.044)	**0.714** (± 0.025)	**0.742** (± 0.026)	**0.760** (± 0.025)	**0.778** (± 0.025)	**0.749** (± 0.029)	**0.793** (± 0.023)
most central	– 0.079 (± 0.007)	– 0.086 (± 0.008)	– 0.086 (± 0.007)	– 0.034 (± 0.006)	0.008 (± 0.005)	0.055 (± 0.009)	0.163 (± 0.015)	0.314 (± 0.029)	**0.790** (± 0.024)
highest	– 5.520 (± 0.765)	0.042 (± 0.104)	0.530 (± 0.052)	0.681 (± 0.035)	**0.748** (± 0.024)	**0.773** (± 0.025)	**0.788** (± 0.023)	**0.790** (± 0.023)	**0.792** (± 0.024)
lowest	– 0.212 (± 0.020)	– 0.237 (± 0.024)	– 0.222 (± 0.022)	– 0.194 (± 0.019)	– 0.151 (± 0.012)	– 0.092 (± 0.009)	– 0.007 (± 0.003)	0.127 (± 0.015)	**0.792** (± 0.025)
random	0.497 (± 0.031)	0.689 (± 0.026)	**0.731** (± 0.031)	**0.751** (± 0.028)	**0.764** (± 0.028)	**0.772** (± 0.026)	**0.775** (± 0.025)	**0.783** (± 0.025)	**0.793** (± 0.025)
unusual events	– 0.967 (± 0.137)	0.086 (± 0.049)	0.329 (± 0.042)	0.442 (± 0.037)	0.485 (± 0.036)	0.516 (± 0.037)	0.567 (± 0.038)	0.645 (± 0.033)	**0.791** (± 0.025)
usual events	– 0.053 (± 0.006)	– 0.032 (± 0.005)	0.055 (± 0.008)	0.164 (± 0.014)	0.209 (± 0.017)	0.257 (± 0.025)	0.346 (± 0.032)	0.507 (± 0.036)	**0.792** (± 0.024)
sequence length	0.593 (± 0.036)	**0.703** (± 0.030)	**0.738** (± 0.029)	**0.759** (± 0.027)	**0.776** (± 0.027)	**0.778** (± 0.027)	**0.786** (± 0.027)	**0.783** (± 0.025)	**0.792** (± 0.025)
training period	0.145 (± 0.034)	0.505 (± 0.039)	0.645 (± 0.035)	**0.706** (± 0.028)	**0.728** (± 0.029)	**0.752** (± 0.027)	**0.773** (± 0.025)	**0.783** (± 0.026)	**0.792** (± 0.025)
spatially random	0.352 (± 0.050)	0.561 (± 0.043)	0.631 (± 0.039)	0.672 (± 0.039)	0.698 (± 0.037)	**0.709** (± 0.034)	**0.743** (± 0.030)	**0.762** (± 0.026)	**0.792** (± 0.024)
			14%	**28%**	**42%**	**57%**	**71%**	**86%**	**100%**
spatially stratified (huc)			– 0.027 (± 0.092)	0.578 (± 0.038)	0.663 (± 0.029)	**0.706** (± 0.030)	**0.739** (± 0.026)	**0.746** (± 0.027)	**0.783** (± 0.021)

Note: NSE scores are presented with uncertainty bounds in brackets. All NSE > 0.7 are highlighted in bold font. In the stratified, HUC region-based sampling experiment, the subsampling percentages differ (for explanation, see Supplementary Appendix A2).

of the low-depth/ unusual samples (Singh & Bárdossy 2012). Potentially, data points that rather reflect data errors are caught with the data depth method as well, inhibiting performance. For example, events with high flows but no rainfall due to a failure of the meteorological measurement system would also be highly unusual. However, this is purely speculative. In any case, the assumption that low-depth events are informative for model training – an assumption that held in conceptual modeling (Singh & Bárdossy 2012) – does not appear to hold here. At least with LSTMs, and when characterizing unusualness with the Tukey half-space method as done in Singh & Bárdossy (2012), as implemented in this study.

Interestingly, we can see in Figure 1(b) that random sampling outperforms almost all other sampling strategies. It approaches maximum possible performance much faster than any other strategy (except for sequence length sampling, Figure 1(c)), reaching a threshold of NSE = 0.7 already with a 10% random sample of the total training dataset (Table 3). Random sampling also outperformed two other methods (the Douglas–Peucker algorithm and a scheme called 'consecutive random') in a recent study (Staudinger *et al.* 2025). This does not come as a surprise if we consider that, given the overall dataset size of 1,652,300 samples from 531 catchments (see Section 2.1), a 10% random sample signifies 165,230 samples, or about 311 samples per catchment on average. This is almost equivalent to a full year, meaning that, at a 10% random sample, the chance that the model sees the whole range of flow conditions from all catchments is already high, even though it might not have seen the outermost extreme high flow values that are so important for flood predictions (Baste *et al.* 2025). Theoretically, any non-random sampling strategy can only be better than random sampling if random sampling does not adequately represent the population, or if we train a model for a specific problem (e.g. low flow, high flow). The results show that random sampling reaches population representativity fast, at least within CAMELS-US, but likely also in other similarly large datasets. The efficiency of random sampling can serve as a posterior explanation as to why Nearing *et al.* (2024) only needed to train on 25% of CARAVAN and still produced benchmark global streamflow prediction performance in their worldwide streamflow prediction study.

The representativity of random sampling is further underpinned by the Kullback–Leibler divergence (D_{KL}) depicted in Figure 2. It shows that random sampling consistently produces the highest representativity of the population. Already at low sampling percentages, random sampling has D_{KL} values below 0.01, indicating extremely low divergence i.e. almost

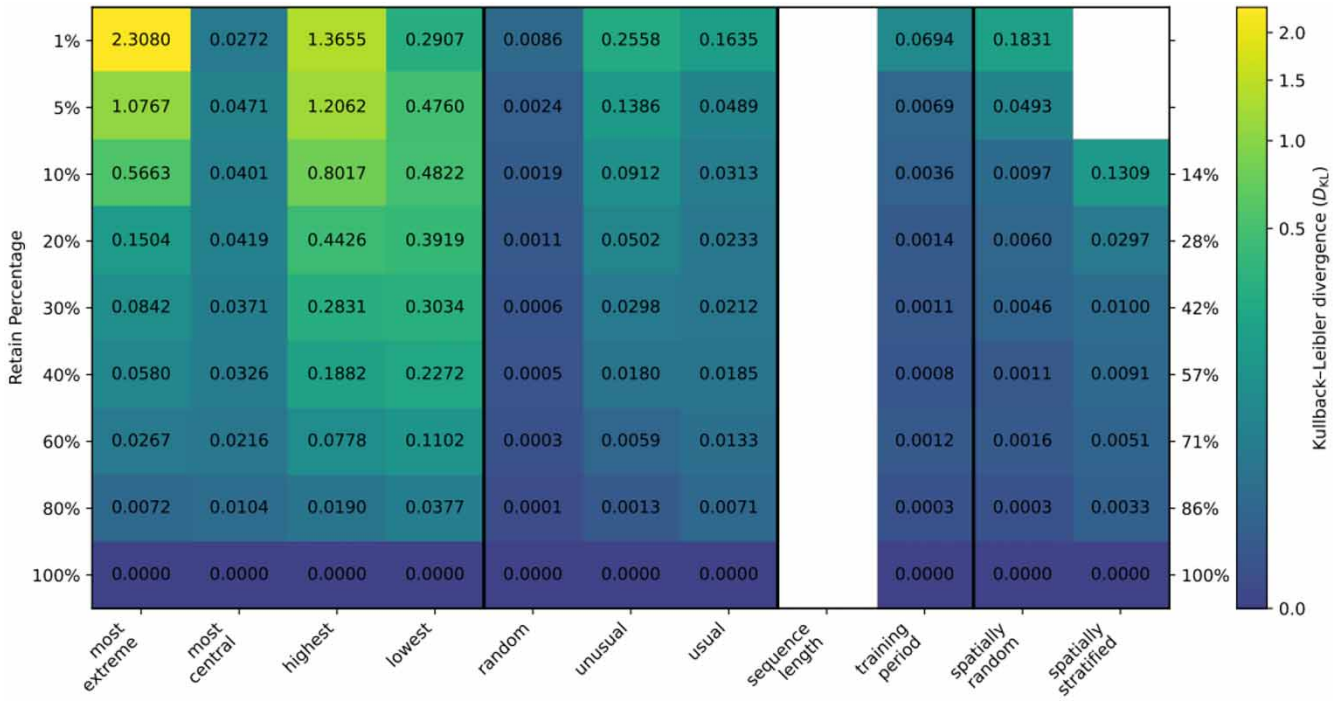


Figure 2 | Kullback–Leibler divergence (D_{KL}) for each sampling strategy and increment, illustrating the divergence of the subsamples from the full dataset. Thereby, higher values equal higher divergence, and zero is identity. D_{KL} is not shown for the 'Sequence length' experiment because its samples cannot be expressed as percentages of the population, preventing the construction of the required sample distribution Q needed for D_{KL} (see Section 2.4).

identity in terms of information theory. However, 'identity' in terms of D_{KL} means overall distribution convergence, whereas we deal with time series prediction here, where timing and not just location of streamflow values are important. This explains how 'most extreme' and 'highest' have high D_{KL} values (i.e. worst agreement with population) despite performing well in terms of NSE performance: Their moments diverge from the population the most, and vice versa for the 'most central' and 'lowest' experiment. However, in general, apart from the special case of the flow magnitude sampling experiments, a relation between low DKL values (Figure 2) and high NSE performance (Figure 1, Table 3) can be attested.

3.3. Effect of sampling temporal context

Figure 1(c) shows the effect of sequence-based sampling routines on performance. It demonstrates that sequence-based sampling is the most efficient sampling strategy tested in this paper, approaching the maximum possible performance faster than any other strategy. Even if the sequence length is reduced to 5%, the resulting model will still maintain a performance of NSE > 0.7 (Table 3). Since 100% represents a standard full year of meteorological data, a sequence length of 5% represents only 18 days or about 2.5 weeks. This is a surprisingly short lookback window, but goes in hand with recent research showing that prior to 14 days before the streamflow value, daily lookback is redundant and can be replaced with longer aggregation periods, e.g. weeks or months (Acuña Espinoza *et al.* 2025). We might explain this behavior by the fact that the model still sees every single day and therefore also meteorological data points in the dataset, albeit only between samples and not within the sequence of individual samples with reduced meteorological context for every streamflow value (sequence length).

In this context, it is important to emphasize that we only look at median scores over all 531 catchments here. This aggregation may obscure different things going on at the local level, for example, higher importance of long-term or short-term memory in specific catchments or events. If we consider catchments with pronounced seasonality, like heavily snow-dominated catchments with large snowpack accumulation in November and December, significantly impacting spring runoff in e.g. the Colorado mountains, we are inclined to think that a long meteorological context is needed to make sense of the individual streamflow value. One might expect that short sequence lengths lead to severely impacted model performance. However, in an auxiliary analysis, we evaluated median NSE performance in snow-dominated catchments (with >50% of precipitation falling as snow) versus rain-dominated (<50% as snow) catchments, and found consistently higher performance in snow-dominated catchments for all sequence lengths, including shortened ones (Supplementary Appendix Table A5). This physically implausible behavior is consistent with previous findings (e.g. Heudorfer *et al.* 2025, especially Supporting information S4 and S5 therein), that proper replication of hydrological behavior in neural networks like LSTMs might not be as robust as we might desire. Ultimately, there are underexplored expectations and assumptions about what these models can and should be able to do. What is missing is a detailed analysis of the quality of generalization capabilities in these models. Future work should focus on catchment-specific or event-specific analyses to further elucidate the physical plausibility of deep learning models.

Compared with sequence length subsampling, training period sampling underperforms, but it is still better than most other sampling strategies. Around 20% of the training period is necessary to produce a performance of above NSE $= 0.7$ (Table 3). This is about as efficient as high-flow sampling. Given that the total training period is 9 years (see Section 2.1), 20% translates to only about 1.8 years or about 660 days.

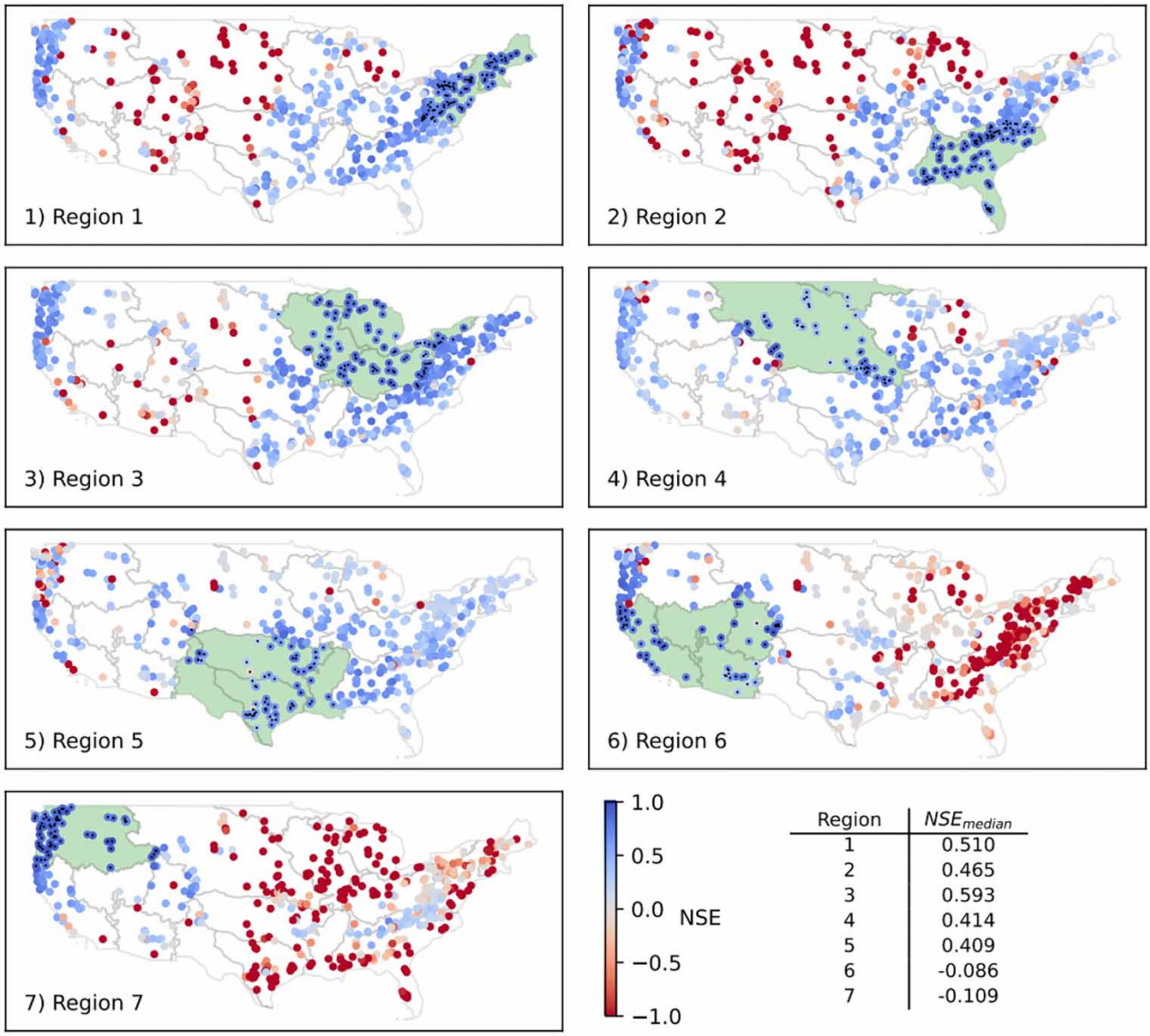
3.4. Spatial representativity and regional transferability

Figure 1(d) shows the performance of spatial sampling strategies. Initial ascendance of the curve associated with random spatial sampling is relatively fast, but further ascendance remains slow compared with most other experiments. In the end, random spatial sampling reaches an NSE of 0.7 at a sample size of 40% of the total dataset (Table 3). This pattern of diminishing return with increasing sample size was also observed by Ma *et al.* (2021) when they increased the number of source basins for transfer learning: once additional basins stopped adding new climate–physiographic combinations, skill gains leveled.

Spatially stratified sampling with respect to HUC regions produces more heavily reduced performance; at least two regions are needed for the model to gain somewhat competitive performance. This experiment especially, but also the relatively slow ascendance of the random spatial sampling curve, makes a good case for the value of region-specific hydrological information in streamflow prediction. This is in line with well-established concepts in hydrologic similarity, where it is understood that different combinations of climate and physiographic catchment settings produce different

characteristic streamflow outputs, which translates into the fact that models need to be trained on data from catchments from different regions. The importance of sampling diverse regions mirrors the conclusions of Ma *et al.* (2021), that 'where' the data come from matters as much as 'how much': their LSTM pre-trained on humid-temperate U.S. basins transferred well to Great Britain but not to arid Chile or snow-dominated China. Ultimately, this further corroborates the limitation of worldwide transfer learning, since we can infer that at least some data from 'new' regions seems necessary for quality predictions.

This is further elucidated in Figure 3, which represents the first step in the stepwise region sampling strategy, where only one single macro-region (Supplementary Appendix Table A1) is used for training, and testing is performed across the full dataset. The figure highlights the ability (or lack thereof) of models trained on region-specific data to extrapolate hydrologic behavior across the contiguous US. Expectedly, models trained on single regions exhibit varying levels of performance when applied to other regions, depending on the degree of hydrologic similarity, a finding that is supported by earlier studies



Region	NSE_{median}
1	0.510
2	0.465
3	0.593
4	0.414
5	0.409
6	-0.086
7	-0.109

Figure 3 | Performance of the model when trained exclusively on individual huc macro-regions, and tested on the full dataset. The figure basically breaks down the first, smallest sampling level in Figure 1. Region designation is detailed in Supplementary Appendix Table A1.

(Ma *et al.* 2021). For example, eastern regions (1–3), which include humid temperate zones with dense vegetation and relatively stable flow regimes, tend to achieve the highest overall median performance ($NSE_{median} = 0.465–0.593$). But they fail to generalize well to the arid and snow-dominated hydrology of the central western and Rocky Mountain regions. Region 5 (Central South) includes mixed forest-agriculture zones and experiences variable precipitation. It shows poor transfer performance to the Northern Pacific area. This likely results from differences in seasonal runoff patterns and orographic effects prominent in the Pacific Northwest. Regions 6 and 7 cover the arid basins and Mediterranean climates of the Pacific South as well as the temperate rainy Pacific North. These regions demonstrate the poorest performance when tested beyond their own boundaries in general ($NSE_{median} = -0.109$ to -0.086), and especially limited transferability to the humid areas in the entire eastern part of the contiguous US.

Interestingly, higher overall median NSE does not always equate to consistent generalization across the whole space, i.e. to high NSE performance everywhere. For instance, Regions 1 and 3, despite having the highest overall median NSE (0.51 and 0.593), show large areas with bad performance. Conversely, Region 4 (Central North) exhibits no major areas of catastrophic failure, even though showing only mediocre overall performance ($NSE_{median} = 0.414$). Further studies about the nature of the hydrological characteristics that provide high or low transferability would be very valuable.

A final note on the limitations of these results, this spatial generalization experiment, like all other experiments in this study, is univariate examination. Regional regionalization patterns could interact with other sampling strategies. For example, it is possible that certain sampling strategies yield better cross-regional transferability than others. This limitation must be considered when digesting these results. Likewise, recent studies using data-driven models (Ouyang *et al.* 2021) have demonstrated that model transferability is also affected by anthropogenic impacts like reservoir operations. The general results of this study are therefore limited to the specific selection of basins, especially since the 531 catchments in the CAMELS-US dataset represent a selection of near-natural catchments.

3.5. Practical recommendations

Based on our results, we offer a couple of recommendations for practitioners as well as for future modeling studies aiming to develop efficient training strategies, e.g. in the context of pre-training or hyperparameter tuning.

First and foremost, we recommend random sampling across the entire dataset as the most effective general-purpose sampling strategy. Our results indicate that random sampling is most efficient in constructing representative samples (also spatially). We suggest that the lack of representativity is the primary reason why more targeted sampling strategies do not prove to be useful for general-purpose training. If one has more specific prediction aims, e.g. for good high-flow accuracy, more targeted sampling strategies might be useful, i.e. high-flow sampling in this case. However, this really only applies if absolute extreme high flows are used, since in our study, only the top 1% high flow sample showed improved FHV scores. A similar recommendation cannot be made for low-flow prediction, since we find no consistent benefit in adding low-flow samples. However, as noted previously (Frame *et al.* 2022), this inconclusive result might also be due to the instability of the FLV score on which we based this assessment.

Furthermore, very short sequence lengths of about 3 weeks are sufficient to arrive at reasonable performance, allowing significantly reduced train times in LSTMs (Supplementary Appendix Figure A1). An even better cost-performance tradeoff can likely be achieved with the method proposed by Acuña Espinoza *et al.* (2025), where daily values older than 14 days are aggregated into weekly or monthly values. Equally, short training periods of about 2 years can be sufficient for quick-shot training if needed. This opens the possibility to include time series of much smaller length in our datasets. However, performance will likely depend strongly on the representativity of the selected period. In any case, it might be important to preserve the temporal context when providing a training sample, since coherent training period sampling outperformed more targeted sampling strategies.

As a note of caution, however, we point the reader to the limitations shown in auxiliary analyses (Section 3.3, Supplementary Appendix Table A5). Using short sequence lengths did not result in physically plausible performance changes of the LSTM for heavily memory-dependent systems (snow-dominated catchments) when using reduced sequence lengths. These points to the urgent need for a renewed focus on catchment-specific or event-specific analyses of deep learning models in order to evaluate the physical plausibility of these models.

Finally, despite better efficiency of one or the other sampling strategy, our results fundamentally corroborate previous studies (Gauch *et al.* 2021b; Kratzert *et al.* 2024) that more data is always better. Even if performance gains diminish, increasing sample size always leads to higher performance.

4. CONCLUSIONS

In this study, we examined how different data sampling strategies affect the performance of a state-of-the-art LSTM-based streamflow prediction model. The aim was to identify the data needs of neural network models in training, and to evaluate the information density of subsamples of different makeup with respect to streamflow prediction. To do this, we systematically tested data subsets based on flow extremity, representativity, time series length, and spatial coverage.

In general, we found that random sampling consistently outperformed more targeted strategies, reaching strong performance (NSE > 0.7) with as little as 10% of training data. This suggests that random samples are surprisingly representative in large datasets like CAMELS-US. Analysis of the D_{KL} measure underpins this finding, showing that random sampling exhibits the lowest divergence from the full dataset, indicating the best information-theoretic representativity in comparison with all other sampling strategies tested.

In terms of time series length, we found that the length of the training period mattered more than the length of the input sequence (i.e. lookback window). Surprisingly short input sequence lengths (about 2.5 weeks) and training period lengths (about 2 years) were sufficient to ensure good performance (NSE > 0.7). This supports using shorter records to potentially expand datasets to additional data previously unconsidered. With regard to flow conditions, high flows were more information-dense than other parts of the flow spectrum, enabling rapid performance gains with limited data. This is because high flows make up most of the variability in the hydrograph, as opposed to low flows, which contribute least to performance, since even normal flows leverage better performance than low flows. Concerning spatial coverage, region-specific hydrologic information turned out to be essential for generalization. Models trained on geographically restricted regions performed poorly across space. Competitive results were only achieved when broader regional coverage was added to the model, confirming the importance of the inclusion of hydrogeographically similar entities in model training.

Finally, we note that one main limitation of this study is that only univariate performance dependency in sampling was evaluated. This is especially critical since previous studies showed interdependence of sampling strategies (Gauch *et al.* 2021b). Thus, future studies could investigate interactions between sampling strategies. In any case, the present study offers basic guidance for future studies regarding which data to use first, if, for example, researchers want to leverage model distillation, pre-training, or transfer learning techniques in future studies. Due to the proven effectiveness of LSTMs to capture general hydrological system dynamics, we speculate that the results presented in this study might hold for data-driven models in general. However, this needs verification in future studies.

The second main limitation is that, due to the fact that all experiments were conducted in a large-sample setting, conclusions may not directly transfer to small-sample, single-basin, or short-record settings. Moreover, these conclusions are metric-dependent, as performance is assessed primarily using NSE. Even though we did consider alternative metrics (especially KGE, FHV, FLV), conclusions may vary when additional metrics are put to use.

DATA AVAILABILITY STATEMENT

The CAMELS-US dataset is open-source and freely available (Addor *et al.* 2017). All codes underlying the study are available under https://zenodo.org/records/17816198.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Acuña Espinoza, E., Loritz, R., Álvarez Chaves, M., Bäuerle, N. & Ehret, U. (2024) To bucket or not to bucket? Analyzing the performance and interpretability of hybrid hydrological models with dynamic parameterization, *Hydrology and Earth System Sciences*, **28** (12), 2705–2719. https://doi.org/10.5194/hess-28-2705-2024.

Acuña Espinoza, E., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., Loritz, R. & Ehret, U. (2025) Technical note: An approach for handling multiple temporal frequencies with different input dimensions using a single LSTM cel, *Hydrology and Earth System Sciences*, **29**, 1749–1758, https://doi.org/10.5194/hess-29-1749-2025.

Addor, N., Newman, A. J., Mizukami, N. & Clark, M. P. (2017) The CAMELS data set: catchment attributes and meteorology for large-sample studies, *Hydrology and Earth System Sciences*, **21**, 5293–5313.

Álvarez Chaves, M., Gupta, H. V., Ehret, U. & Guthke, A. (2024) On the accurate estimation of information-theoretic quantities from multi-dimensional sample data, *Entropy*, **26** (5), 387. https://doi.org/10.3390/e26050387.

Auer, A., Gauch, M., Kratzert, F., Nearing, G., Hochreiter, S. & Klotz, D. (2024) A data-centric perspective on the information needed for hydrological uncertainty predictions, *Hydrology and Earth System Sciences*, **28** (17), 4099–4126. https://doi.org/10.5194/hess-28-4099-2024.

Baste, S., Klotz, D., Espinoza, E. A., Bardossy, A. & Loritz, R. (2025) Unveiling the limits of deep learning models in hydrological extrapolation tasks, *Hydrology and Earth System Sciences*, **29**, 5871–5891. https://doi.org/10.5194/egusphere-2025-425.

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., Gharari, S., Freer, J. E., Whitfield, P. H., Shook, K. R. & Papalexiou, S. M. (2021) The abuse of popular performance metrics in hydrologic modeling, *Water Resources Research*, **57** (9), e2020WR029001. doi:10.1029/2020WR029001.

Davison, A. C. & Hinkley, D. V. (1997) *Bootstrap Methods and Their Application*. New York, NY: Cambridge University Press.

Eddin, M. H. S., Zhang, Y., Kollet, S. & Gall, J. (2025) *RiverMamba: A State Space Model for Global River Discharge and Flood Forecasting* (No. arXiv:2505.22535). arXiv.

Feng, D., Fang, K. & Shen, C. (2020) Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales, *Water Resources Research*, **56** (9), e2019WR026793. doi:10.1029/2019WR026793.

Feng, D., Beck, H., Lawson, K. & Shen, C. (2023) The suitability of differentiable, physics-informed machine learning hydrologic models for ungauged regions and climate change impact assessment, *Hydrology and Earth System Sciences*, **27** (12), 2357–2373. https://doi.org/10.5194/hess-27-2357-2023.

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

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J. & Hochreiter, S. (2021a) Rainfall–runoff prediction at multiple timescales with a single long short-term memory network, *Hydrology and Earth System Sciences*, **25** (4), 2045–2062. https://doi.org/10.5194/hess-25-2045-2021.

Gauch, M., Mai, J. & Lin, J. (2021b) The proper care and feeding of CAMELS: how limited training data affects streamflow prediction, *Environmental Modelling & Software*, **135**, 104926. https://doi.org/10.1016/j.envsoft.2020.104926.

Gneiting, T. & Katzfuss, M. (2014) Probabilistic forecasting, *Annual Review of Statistics and Its Application*, **1** (1), 125–151. doi:10.1146/annurev-statistics-062713-085831.

Gupta, A. (2024) Information and disinformation in hydrological data across space: the case of streamflow predictions using machine learning, *Journal of Hydrology: Regional Studies*, **51**, 101607. doi:10.1016/j.ejrh.2023.101607.

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

Heudorfer, B., Liesch, T. & Broda, S. (2024) On the challenges of global entity-aware deep learning models for groundwater level prediction, *Hydrology and Earth System Sciences*, **28** (3), 525–543. https://doi.org/10.5194/hess-28-525-2024.

Heudorfer, B., Gupta, H. V. & Loritz, R. (2025) Are deep learning models in hydrology entity aware? *Geophysical Research Letters*, **52** (6), e2024GL113036. https://doi.org/10.1029/2024GL113036.

Hochreiter, S. & Schmidhuber, J. (1997) Long-short-term memory, *Neural Computation*, **9** (8), 1735–1780. doi:10.1162/neco.1997.9.8.1735.

Huo, X., Gupta, H., Niu, G., Gong, W. & Duan, Q. (2019) Parameter sensitivity analysis for computationally intensive spatially distributed dynamical environmental systems models, *Journal of Advances in Modeling Earth Systems*, **11** (9), 2896–2909. https://doi.org/10.1029/2018ms001573.

Jahangir, M. S., You, J. & Quilty, J. (2023) A quantile-based encoder-decoder framework for multi-step ahead runoff forecasting, *Journal of Hydrology*, **619**, 129269. doi:10.1016/j.jhydrol.2023.129269.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S. & Nearing, G. (2022) Uncertainty estimation with deep learning for rainfall–runoff modeling, *Hydrology and Earth System Sciences*, **26** (6), 1673–1693. https://doi.org/10.5194/hess-26-1673-2022.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K. & Herrnegger, M. (2018) Rainfall–runoff modelling using long short-term memory (LSTM) networks, *Hydrology and Earth System Sciences*, **22** (11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018.

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

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S. & Nearing, G. (2019b) Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, *Hydrology and Earth System Sciences*, **23** (12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019.

Kratzert, F., Klotz, D., Hochreiter, S. & Nearing, G. S. (2021) A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling, *Hydrology and Earth System Sciences*, **25** (5), 2685–2703. https://doi.org/10.5194/hess-25-2685-2021.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G. & Matias, Y. (2023) Caravan – a global community dataset for large-sample hydrology, *Scientific Data*, **10** (1), 61. https://doi.org/10.1038/s41597-023-01975-w.

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

Lamontagne, J. R., Barber, C. A. & Vogel, R. M. (2020) Improved estimators of model performance efficiency for skewed hydrologic data, *Water Resources Research*, **56** (9), e2020WR027101. doi:10.1029/2020WR027101.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G. & Dadson, S. J. (2021) Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models, *Hydrology and Earth System Sciences*, **25** (10), 5517–5534. https://doi.org/10.5194/hess-25-5517-2021.

Liu, J., Bian, Y., Lawson, K. & Shen, C. (2024) Probing the limit of hydrologic predictability with the transformer network, *Journal of Hydrology*, **637**, 131389. https://doi.org/10.1016/j.jhydrol.2024.131389.

Ma, K., Feng, D., Lawson, K., Tsai, W., Liang, C., Huang, X., Sharma, A. & Shen, C. (2021) Transferring hydrologic data across continents – leveraging data-rich regions to improve hydrologic prediction in data-sparse regions, *Water Resources Research*, **57** (5), e2020WR028600. https://doi.org/10.1029/2020WR028600.

Miller, D. A. & White, R. A. (1998) A conterminous United States multilayer soil characteristics dataset for regional climate and hydrology modeling, *Earth Interactions*, **2** (2), 1–26. https://doi.org/10.1175/1087-3562(1998)002<0001:ACUSMS>2.3.CO;2.

Mizukami, N., Rakovec, O., Newman, A. J., Clark, M. P., Wood, A. W., Gupta, H. V. & Kumar, R. (2019) On the choice of calibration metrics for 'high-flow' estimation using hydrologic models, *Hydrology and Earth System Sciences*, **23** (6), 2601–2614. doi:10.5194/hess-23-2601-2019.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D. & Veith, T. L. (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, *Transactions of the ASABE*, **50** (3), 885–900. doi:10.13031/2013.23153.

Mount, N. J., Maier, H. R., Toth, E., Elshorbagy, A., Solomatine, D., Chang, F.-J. & Abrahart, R. (2016) Data-driven modelling approaches for socio-hydrology: opportunities and challenges within the Panta Rhei Science Plan, *Hydrological Sciences Journal*, **61** (7), 1192–1208.

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

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. & Matias, Y. (2024) Global prediction of extreme floods in ungauged watersheds, *Nature*, **627** (8004), 559–563. https://doi.org/10.1038/s41586-024-07145-1.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T. & Duan, Q. (2015) Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, *Hydrology and Earth System Sciences*, **19** (1), 209–223. https://doi.org/10.5194/hess-19-209-2015.

Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B. & Nearing, G. (2017) Benchmarking of a physically based hydrologic model, *Journal of Hydrometeorology*, **18** (8), 2215–2225. doi:10.1175/JHM-D-16-0284.1.

Ouyang, W., Lawson, K., Feng, D., Ye, L., Zhang, C. & Shen, C. (2021) Continental-scale streamflow modeling of basins with reservoirs: towards a coherent deep-learning-based strategy, *Journal of Hydrology*, **599**, 126455. doi:10.1016/j.jhydrol.2021.126455.

Papacharalampous, G., Tyralis, H., Langousis, A., Jayawardena, A. W., Sivakumar, B., Mamassis, N., Montanari, A. & Koutsoyiannis, D. (2019) Probabilistic hydrological post-processing at scale: why and how to apply machine-learning quantile regression algorithms, *Water*, **11** (10), 2126. doi:10.3390/w11102126.

Pelletier, J. D., Patrick D. Broxton, Hazenberg, P., Zeng, X., Troch, P. A., Niu, G.-Y., Williams, Z., Brunke, M. A. & Gochis, D. (2016) A gridded global data set of soil, intact regolith, and sedimentary deposit thicknesses for regional and global land surface modeling, *Journal of Advances in Modeling Earth Systems*, **8**, 41–65. https://doi.org/10.1002/2015MS000526.

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

Roscher, R., Russwurm, M., Gevaert, C., Kampffmeyer, M., Dos Santos, J. A., Vakalopoulou, M., Hänsch, R., Hansen, S., Nogueira, K., Prexl, J. & Tuia, D. (2024) Better, not just more: data-centric machine learning for Earth observation, *IEEE Geoscience and Remote Sensing Magazine*, **12** (4), 335–355. https://doi.org/10.1109/MGRS.2024.3470986.

Singh, S. K. & Bárdossy, A. (2012) Calibration of hydrological models on hydrologically unusual events, *Advances in Water Resources*, **38**, 81–91. doi:10.1016/j.advwatres.2011.12.006.

Snieder, E. & Khan, U. T. (2025) A diversity-centric strategy for the selection of spatio-temporal training data for LSTM-based streamflow forecasting, *Hydrology and Earth System Sciences*, **29** (3), 785–798. https://doi.org/10.5194/hess-29-785-2025.

Staudinger, M., Herzog, A., Loritz, R., Houska, T., Pool, S., Spieler, D., Wagner, P. D., Mai, J., Kiesel, J., Thober, S., Guse, B. & Ehret, U. (2025) How well do hydrological models learn from limited discharge data? A comparison of process- and data-driven models. EGUsphere. Preprint. https://doi.org/10.5194/egusphere-2025-1076.

Sun, A. Y., Jiang, P., Mudunuru, M. K. & Chen, X. (2021) Explore spatio-temporal learning of large sample hydrology using graph neural networks, *Water Resources Research*, **57** (12), e2021WR030394. doi:10.1029/2021WR030394.

Tukey, J. W. (1975) Mathematics and the picturing of data, *Proceedings of the International Congress of Mathematicians*, **2**, 523–531.

USGS & NRCS (2025) *Watershed Boundary Dataset (WBD)* [Dataset]. Available at: https://www.usgs.gov/national-hydrography/watershed-boundary-dataset.

Vrugt, J. A., Bouten, W., Gupta, H. V. & Sorooshian, S. (2002) Toward improved identifiability of hydrologic model parameters: the information content of experimental data, *Water Resources Research*, **38** (12), 1312. https://doi.org/10.1029/2001wr001118.

Weijs, S. V. & Van De Giesen, N. (2013) An information-theoretical perspective on weighted ensemble forecasts, *Journal of Hydrology*, **498**, 177–190. doi:10.1016/j.jhydrol.2013.06.033.

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

First received 18 July 2025; accepted in revised form 13 February 2026. Available online 25 February 2026