



A diversity-centric strategy for the selection of spatio-temporal training data for LSTM-based streamflow forecasting

Everett Snieder and Usman T. Khan

Department Civil Engineering, Lassonde School of Engineering, York University, Toronto, Canada

Correspondence: Everett Snieder (esnieder@yorku.ca)

Received: 31 May 2024 – Discussion started: 14 June 2024

Revised: 25 October 2024 – Accepted: 6 December 2024 – Published: 13 February 2025

Abstract. Deep learning models are increasingly being applied to streamflow forecasting problems. Their success is in part attributed to the large and hydrologically diverse datasets on which they are trained. However, common data selection methods fail to explicitly account for hydrological diversity contained within training data. In this research, clustering is used to characterise temporal and spatial diversity, in order to better understand the importance of hydrological diversity within regional training datasets. This study presents a novel, diversity-based resampling approach to creating hydrologically diverse datasets. First, the undersampling procedure is used to undersample temporal data and to show how the amount of temporal data needed to train models can be halved without any loss in performance. Next, the procedure is applied to reduce the number of basins in the training dataset. While basins cannot be omitted from training without some loss in performance, we show how hydrologically dissimilar basins are highly beneficial to model performance. This is shown empirically for Canadian basins; models trained on sets of basins separated by thousands of kilometres outperform models trained on localised clusters. We strongly recommend an approach to training data selection that encourages a broad representation of diverse hydrological processes.

ered amongst the best ways to mitigate flood damage. Many Canadian communities lack any sort of FEWS, making them vulnerable to flood damage. Over the past 3 decades, machine learning (ML) models have been increasingly applied for streamflow prediction, and they currently have significant potential for improving the accuracy and coverage of FEWSs in flood-prone regions. Recently, several large-sample studies have shown that ML models can consistently outperform traditional, physics-based hydrological models (Mai et al., 2022; Arsenault et al., 2023a; Kratzert et al., 2019b), which underscores their proficiency for FEWSs.

ML model development has typically followed the same format as physics-based models, in that a single model is parameterised and calibrated on an individual basin, which is referred to as a locally trained model. The work by Kratzert et al. (2019b) demonstrated that the accuracy of ML models can be improved by training a model on a set of basins rather than an individual basin, which is referred to as a regionally trained model. Regional training relies on deep learning architectures such as long short-term memory networks (LSTMs), which have recently surged in popularity for streamflow forecasting applications and are considered to be state of the art (Fang et al., 2022). Recent advances in regional learning have focused on improvements to model architectures (Nevo et al., 2022; Girihagama et al., 2022) and benchmarking against traditional physics-based models (Lees et al., 2021; Arsenault et al., 2023a).

Broadly speaking, physics-based and ML hydrological models benefit from diverse training data, which improves their performance for unseen future conditions. Locally trained models are only provided with temporally diverse data at a single point in space. In contrast, regionally trained models have been empirically shown to outperform locally

1 Introduction

Floods constitute a major threat to populations and infrastructure and are projected to increase in severity due to climate change and urbanisation. Flood early warning systems (FEWSs), which rely on models that predict streamflow, provide advanced notice of flood risk and are consid-

trained models in several studies (Kratzert et al., 2019b; Zhang et al., 2022), and their success is in part attributable to the spatial and temporal hydrological diversity contained in the multi-basin datasets on which they are trained (Kratzert et al., 2019b).

However, there is currently little guidance on optimum basin (spatial) selection of training data for these models. Often times, models are trained on complete large-sample datasets such as those included in the Caravan dataset (Kratzert et al., 2023). However, with increasingly large and global hydrological datasets, it is not always practical or feasible to train on all available data, especially when conducting computationally expensive tasks such as hyperparameter selection, which is required to achieve optimum model performance, or creating multimodel ensembles. Therefore, there is a need for improved guidance on efficient methods for training data selection to maximise model performance and generalisation. Many of the deep learning advances in hydrology (e.g. Kratzert et al., 2019a, Klotz et al., 2022, Lees et al., 2022, and Gauch et al., 2021a) have utilised well-established large-sample basins (see Caravan (Kratzert et al., 2023) and the datasets therein). Canada is a country with a diverse hydrological landscape, characterised by coastal regions, mountains, urban areas, and exposed rock in the form of the Canadian Shield. The effectiveness of regionally trained models has not yet been well established on highly diverse Canadian basins.

When training a model to predict streamflows in some region of interest, the goal is to select the most relevant spatial and temporal data while avoiding data that have either no impact or a negative impact on model performance. Unsupervised clustering has been used in previous studies as a data-driven approach to identify spatial and temporal diversity within a training dataset (Toth, 2009; Kratzert et al., 2024). The application of clustering as a means to identify spatial and temporal diversity is in itself nothing new. Many studies, which are reviewed below, have applied clustering to spatial and temporal data as a means to quantify hydrological diversity. However, the treatment of hydrological diversity generally follows one of two approaches: it is used to generate either hydrologically diverse datasets or datasets with homogeneous hydrological conditions. The former aims to generalise models to a wide range of conditions, promoting balanced performance and good generalisation, while the latter aims to simplify the learning problem, improving performance in similar conditions. Both approaches have been used successfully for temporal streamflow clustering. Antil and Lauzon (2004) applied a self-organising map (SOM) to streamflow data in a single basin to create a training dataset with a balanced representation of diverse hydrological states. In contrast, Toth (2009) used an SOM to classify streamflow into homogeneous subsets, on which individual models are trained and combined in a modular format. Their approach was found to improve overall prediction accuracy, which can be attributed to the error diversity of the col-

lection of trained models. Snieder et al. (2021) partitioned streamflows into typical streamflows and high streamflows, in order to undersample typical streamflows and oversample high streamflows, which is found to improve performance for the latter, which is desirable for FEWS applications. The same motivation has led to numerous applications of clustering on basins, particularly in regional training schemes. For example, Gauch et al. (2021b) showed that implicitly increasing hydrological diversity of regional training datasets, by iteratively increasing the number of basins, as well as the amount of data in each basin, improves model generalisation. However, the study does not explicitly quantify hydrological diversity, in part due to the absence of a widely agreed upon metric for hydrological similarity (Oudin et al., 2010). However, other studies have used clustering to estimate hydrological diversity, such that basin selection can explicitly account for hydrological diversity. These cases tend to use some form of clustering (either supervised or unsupervised) to quantify the hydrological diversity within training data and the effects it has on model generalisation. Zhang et al. (2022) applied *K*-means clustering to a set of 35 mountainous basins in China based on hydroclimatic attributes, finding that a model trained on all available basins typically outperformed those trained on individual clusters. Hashemi et al. (2022) applied a similar approach by classifying basins into distinct hydrological-regime-based hydrometeorological thresholds. As done in Zhang et al. (2022), their study compared locally and globally trained models, finding only minor differences in the performance between the two. A common problem in comparing global and locally trained models is that these comparisons typically do not control for sample size. As a result, the improved performance of the global model can be impacted by the regularisation effect on the sample size. In other words, deep learning models trained on small datasets may be overfitted and, thus, poorly generalised. Fang et al. (2022) accounted for this potential issue. Their study used an existing “ecoregion” basin classification, where basins were classified by the United States Environmental Protection Agency, and evaluated the effects of additional training basins at three similarity intervals. Their study showed that, counterintuitively, “far” or “dissimilar” basins amongst the training dataset often produced greater improvements in model performance when compared to the inclusion of “close” basins. They speculate that distant basins provide a regularisation effect. Fang et al. (2022) called for further investigation into the effect that hydrological diversity has on model generalisation and underlined the need for a systematic approach. Kratzert et al. (2024) characterised hydrological diversity by applying *K*-means clustering to basin attributes, finding that models trained on basins with similar hydrological characteristics outperform randomly sampled basin sets of the same sizes. However, in every case, they show that LSTMs trained on hundreds of basins outperform those trained on smaller subsets. They also demonstrated how regional learning improves performance for ex-

treme events and thus for FEWSs, as the training datasets contain a higher number of extreme events spread across all basins. Many of these studies assume that similar basins are most useful to one another in the context of regional learning. We challenge this assumption, and we seek to determine to what extent hydrologically similar data are beneficial for training.

The objective of this study is to examine the effect that the formation of hydrologically diverse training datasets have on model performance and generalisation. Hydrological diversity is quantified using clustering, which is applied separately to streamflow (temporal) and basins (spatial). This topic is analysed throughout two experiments. In the first experiment, we evaluate the effects of removing non-diverse streamflow data from regional training datasets. The latter test is repeated but by undersampling non-diverse basins instead of streamflow values from a larger subset. In the second experiment, we compare the effects of adding similar and dissimilar basins to a training dataset for some region of interest. The purpose of this second experiment is to compare the contribution of additional basins to model generalisation with respect to their hydrological similarity to the evaluation set.

While numerous studies have applied clustering to streamflow and basins, to the extent of the knowledge of the authors, the use of clustering to explicitly create spatially and temporally diverse training datasets is a novel approach. The outcome of these experiments has the potential to improve methods for the creation of training datasets for regionally trained models. This topic is investigated on sequence-to-sequence (Seq2Seq) LSTMs for daily streamflow forecasts of 1–3 d in Canadian basins. Basins are sampled from across Canada, and models are trained using historic hydrometeorological data from the past 36 years.

2 Methods and data

2.1 Input and target variables

This study uses data retrieved from the HYSETS dataset, which contains hydrometeorological data for over 14 000 basins across North America (Arsenault et al., 2020). The target variable, the future state of streamflow, is predicted using dynamic and static input features. The models in this study are autoregressive (AR), meaning past streamflow at the target gauge is used as one of the input features (Nearing et al., 2022). Additional dynamic features from the HYSETS database include daily basin-averaged minimum temperature, maximum temperature, precipitation, and snow water equivalent (SWE), which are listed in Table 2. AR LSTMs are used since they are more accurate than non-AR LSTMs (Nearing et al., 2022) and since the Canadian hydrometric network is largely available in real time. The static basin attributes, which are summarised in Table S2, allow the model

Table 1. Streamflow resampling configurations in experiment 1a. Labels denote the resampling type: number of clusters (K for CUS, cluster-based undersampling) and resampling range (ϕ).

Label (K, ϕ)	Unique basins	Samples (per basin)	Total samples (thousands)
Baseline	64	4380	280
CUS _{Q} ($K = 6, \phi = 0.25$)	64	1095	70
CUS _{Q} ($K = 6, \phi = 0.50$)	64	2190	140
CUS _{Q} ($K = 12, \phi = 0.50$)	64	2190	140
RUS _{Q} ($\phi = 0.50$)	64	2190	140

to transfer learnt information between basins. Some of the static attributes are included in the HYSETS database, while additional attributes are calculated based on the dynamic time series, which generally follow those included in the CAMELS dataset (Addor et al., 2017). The additional attributes are calculated based on the time series data already contained within HYSETS and do not rely on any external databases. Unfortunately, there are no attributes that reveal whether a catchment is regulated by built infrastructure. However, the model may still be able to learn the effects of built infrastructure implicitly through the rainfall–runoff relationship. Lastly, the input feature set also contains a one-hot-encoded basin label (Lees et al., 2022), which enables the model to distinguish between streamflows in different basins. Static basin attributes are used as input features for the streamflow forecasting model and in the basin clustering method, which are described in Sect. 2.1 and 2.5, respectively.

2.2 Basin selection

This study only considers Canadian basins from the HYSETS database. Basins are removed if they have less than 80 % data availability within any of the training, validation, or testing periods, which span a total of 36 years from October to September of 1982–1994, 1994–2006, and 2006–2018, respectively, following the split-sample method (KLEMEŠ, 1986). The training partition is used to train the models, validation is used to fine-tune LSTM hyperparameters, and the test partition is used to calculate model performance. While records in some basins exist prior to 1982, it is imperative that the data used to train and evaluate basins are from the same time period and are of a similar size. Including records from before 1982 results in fewer basins to choose from, which tends to reduce the hydrological variability of the basin set. Next, some basins are removed due to missing static attributes. These criteria produce a set of approximately 2000 basins, with highly variable attributes, according to Table 1.

2.3 Sequence-to-sequence LSTM models

LSTM models, with a Seq2Seq architecture (Cho et al., 2014), are used to generate forecasts at a daily resolution, at multiple lead times. Seq2Seq models are composed of an encoder and decoder; the encoder transforms an input sequence into a fixed length context vector, which is provided to the decoder, which outputs predictions. Recently, several studies have demonstrated the aptness of Seq2Seq models for predicting runoff at multiple lead times (Xiang et al., 2020; Girihagama et al., 2022; Zhang et al., 2022).

The Seq2Seq models in this study use hyperparameter values that are common for LSTM rainfall–runoff models. The models in this study use a hidden layer size of 128 cells, a dropout rate of 0.2, a batch size of 32, and Adam optimisation with a decaying learning rate of 0.001 to 0.0001 across a total of 80 epochs. Input and output sequences of 7 and 3 d were used, respectively. While an input sequence of 365 d is commonly used for streamflow prediction (Kratzert et al., 2019b; Arsenault et al., 2023b), Gauch et al. (2021b) noted that short sequences are better suited to small basin sets and have been used in AR models (Nevo et al., 2022).

Typically, for regional training, the error terms of individual basins are normalised based on streamflow variance of that basin. Using the typical variance-based regularisation applied to the cost function produces a relative increase in weight applied to low-variance basins. While this works well for non-AR models, AR models have a tendency to develop an overreliance on recent streamflow observations, which can manifest in a positive timing error (Snieder, 2019). The resulting models may be barely distinguishable from the naive model (i.e. the most recent streamflow observation). Highly seasonal, naturalised basins are most prone to this problem, as streamflow tends to change gradually with time; thus, a model that outputs recent streamflow observations might be mistakenly seen as accurate. Specialised performance metrics such as the persistence index (PI) are often used to identify this problem in real-time forecasting applications (Nevo et al., 2022). This is simply because the PI normalises the error relative to the naive model, with a PI of less than 0 corresponding to a non-informative forecast relative to real-time observations. The same is not the case for the widely used Nash–Sutcliffe efficiency (NSE), which can be misleadingly high in the same cases (Knoben et al., 2019). For this reason, we propose that basin persistence (i.e. mean-squared deviation between observations at the current time and forecast time) be used to regularise the cost function for regionally trained models. Instead of placing more weight on low-variance basins, persistence-based regularisation places more weight on basins that have low error between recent and future streamflow values. Failure to do so results in models that are not adequately trained in those basins and produces non-informative forecasts (relative to the naive model).

In this study, basins are normalised using the formulation proposed in Kratzert et al. (2019b) for NSE* but substituting

the basin variance for the persistence corresponding to the forecast lead time. The persistence-based cost function PI* is given by

$$PI^* = \frac{1}{B} \sum_{b=1}^B \sum_{t=1}^T \frac{\sum (q_t - \hat{q}_t)^2}{\sum (p_b + \epsilon)^2}, \quad (1)$$

in which q_t is the observed streamflow, \hat{q}_t is the predicted streamflow, ϵ is a constant (0.1) that prevents the function from exploding to negative infinity (Kratzert et al., 2019b), and p_b corresponds to the persistence of an individual basin b in a set of B basins, given by

$$p_b = \sqrt{\sum_{t=1}^T (q_t - q_{t-L})^2}, \quad (2)$$

in which L is the forecast lead time of \hat{q}_t .

2.4 Performance metrics

Models are evaluated using two performance metrics: NSE and PI. The NSE, given in Eq. (3), is amongst the most widely used metrics for hydrological models and effectively normalises the mean-squared model error based on streamflow variance. The PI, which is given in Eq. (4) and used in the basin regularisation function described above, is a similar metric, but instead of normalising squared residuals using the mean, it normalises forecasts based on the squared error between the streamflow at the current and forecast time steps (Kitanidis and Bras, 1980).

$$NSE = 1 - \frac{\sum_{t=1}^T (q_t - \hat{q}_t)^2}{\sum_{t=1}^T (q_t - \bar{q})^2}, \quad (3)$$

$$PI = 1 - \frac{\sum_{t=1}^T (q_t - \hat{q}_t)^2}{\sum_{t=1}^T (q_t - q_{t-L})^2}, \quad (4)$$

where \hat{q}_t is the predicted streamflow, and \bar{q} is the mean observed streamflow. q_{t-L} is the observed streamflow, shifted by the lead time L of the forecast such that it represents the real-time observable streamflow in an operational context. Both metrics range between $-\infty$ and 1, with 1 being perfect and values less than 0 indicating performance worse than each respective baseline.

2.5 Clustering

Clustering is a simple yet effective way to identify hydrologically diverse data for training streamflow forecasting models. This study uses clustering to identify two forms of hydrological diversity. First, it is applied to streamflow records of individual basins to identify diverse streamflow conditions. Second, it is applied to static basin attributes to identify basins with diverse hydrometeorological attributes. Note that both methods are independent of one another. This study uses the

constrained K -means clustering algorithm (Bennett et al., 2000), which allows for the specification of a minimum cluster size. This avoids a problem that occurs with clustering streamflow, which is that infrequent flood streamflows typically produce a very small cluster, constraining the number of samples that are available when drawing an even number of samples from each cluster (Toth, 2009).

The first application of clustering is to identify hydrologically diverse streamflows. Previous studies have applied clustering to the input vectors of ML models (Anctil and Lauzon, 2004; Abrahart and See, 2000). However, such approaches do not guarantee that streamflow is the main variable by which clusters are discriminated. For that reason, we engineer a feature set for clustering streamflow based solely on the target streamflow data, which encourages diverse streamflow conditions between clusters. The engineered feature set includes streamflow (q_t), two streamflow gradient features (given as $(q_{t-3} - q_t)/3$ and $q_{t-1} - q_t$), and two day-of-year features (given as $\sin(2\pi t/365)$ and $\cos(2\pi t/365)$, where t is the day of year, i.e. 1 to 365). While the day of year is discontinuous between 365 and 1, the sine and cosine decomposition offers a continuously changing pairing across each year, which increases the likelihood of clusters spanning from December to January. The streamflow gradient encourages the representation of rising and falling limbs within the clusters, which are not distinguishable using only the streamflow state.

The second application is on basins which are clustered based on their static attributes. Due to the large number of features (39) and collinearity between features, principal component analysis (PCA) is used to reduce the feature set to eight principal components. Names and statistics of the static attribute set are provided in Table S1 in the Supplement.

Both clustering applications are used to inform a simple resampling procedure that aims to maximise hydrological diversity. The cluster-based undersampling (CUS) procedure is as follows. Given N training examples (either streamflow samples or basins),

1. select an undersampling rate ϕ as a fraction of N and a number of clusters K ,
2. cluster records into K clusters with a minimum cluster size of $\phi N/K$,
3. sort samples based on distance to the cluster centroid,
4. select samples 1 to $\phi N/K$ from each cluster to form the training dataset.

Each undersampling strategy is illustrated in Fig. 1. Panels (a), (c), and (e) illustrate the raw streamflow time series (a), clustered streamflow (c), and undersampled streamflow (e). Note how the time series in (e) contains fewer typical streamflows and proportionally more high streamflows compared to the continuous records in (a) and (c). Similarly,

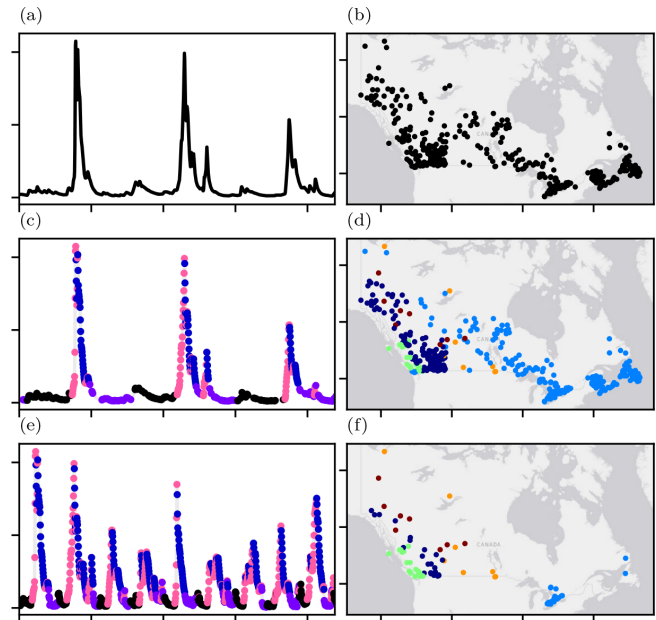


Figure 1. Left column, from top to bottom: unclustered (a), clustered (c), and undersampled streamflow (e) for basin 01AD003 (HYDAT ID (Canadian National Water Data Archive); located along St Francis River in New Brunswick). Right column, from top to bottom: unclustered (b), clustered (d), and undersampled basins (f). Cluster colours are arbitrary, and there is no connection between temporal (a, c, e) and spatial (b, d, f) cluster colours. The World Gray Canvas base map in (b), (d), and (f) was provided by Esri et al. (2022a).

panels (b), (d), and (f) show Canadian basins (b), clustered basins (d), and diverse undersampled basins (f).

CUS applied to streamflow and CUS applied to basins are denoted as CUS_Q and CUS_B , respectively. Selecting an equal number of examples (streamflow or basins) from each cluster results in a balanced variety of hydrological conditions within the training set. There are several reasons why such a training dataset is desirable. First, balanced hydrological conditions encourage balanced performance across different streamflows or basins. Models trained on imbalanced datasets, such as streamflow records in which low streamflows drastically outnumber high streamflows, may be biased towards low streamflow conditions (Snieder et al., 2021). The same reasoning applies to basin selection. A regionally trained model may be biased towards areas with dense spatial coverage. By selecting an equal number of each type of basin, we encourage a balanced spread of hydrological characteristics in the training basin dataset, which translates to good generalisation across a broader range of basins.

Due to the large and diverse feature set, feature importance is calculated to interpret the dominant basin attributes that distinguish clusters. Since K -means clustering does not inherently quantify feature importance, a random forest (RF) classifier is used as a surrogate for approximating feature im-

portance. RFs contain an intrinsic importance metric that is commonly used in hydrology (Tyrallis et al., 2019). An RF with 256 estimators and a max depth value of 6 is used. Since the RF is simply fitting the outcome of the constrained K -means clustering, the RF is expected to achieve near-perfect accuracy, without the need for hyperparameter tuning.

Note that the two clustering applications described above are standalone and can be combined by using them in series. While a unified spatio-temporal clustering method could be used to cluster samples in a single step, separating spatial and temporal clustering allows for each method to be assessed independently.

2.6 Experiments

2.6.1 Experiment 1: evaluating streamflow and basin redundancy in training datasets

In this experiment, CUS is applied to streamflow (experiment 1a) and separately applied to basins (experiment 1b). These experiments are designed to determine the extent to which non-diverse (i.e. redundant) data can be removed from training datasets, without any loss in model performance. The CUS-generated datasets are compared with random undersampled (RUS) datasets, which consist of ϕN samples (sampled without replacement).

To evaluate CUS_Q (experiment 1a), a set of 64 randomly sampled basins is established. CUS_Q is applied to each basin individually, then merged to form the training dataset. Several resampling configurations are considered, including ϕ values of 0.25 to 0.50 and K of values 6 and 12. These configurations are compared against two baseline models, which are trained on (1) the entire dataset and (2) an RUS, which are also undersampled at 0.25 and 0.50 to match the sample size of the CUS_Q datasets. The parameters for each training configuration are listed in Table 1.

The framework outlined above is replicated to evaluate the effects of spatial undersampling (experiment 1b). Beginning this time with a set of 128 randomly sampled basins, we establish subsets of basins sampled for varying numbers of clusters. Basin subset comparisons are made against a baseline model that is trained on the entire set of basins and RUS subsets. The purpose of this experiment is to determine whether clustering can effectively be used to identify a subset of basins that is sufficiently hydrologically diverse such that it can be used to train a model capable of generalisability on the complete set. In this experiment, the basins are trained on the entire streamflow records (i.e. no temporal resampling). The parameters for each training configuration are listed in Table 2.

Table 2. Basin resampling configurations in experiment 1b. Labels denote the resampling type: number of clusters (K for CUS_B) and resampling range (ϕ).

Label (K, ϕ)	Unique basins	Samples (per basin)	Total samples (thousands)
Baseline	128	4380	560
CUS _B ($K = 2, \phi = 0.50$)	64	4380	280
CUS _B ($K = 8, \phi = 0.50$)	64	4380	280
CUS _B ($K = 32, \phi = 0.50$)	96	4380	420
RUS _B ($\phi = 0.50$)	64	4380	280
RUS _B ($\phi = 0.75$)	96	4380	420

2.6.2 Experiment 2: cross-comparison of two clusters of basins

The next experiment is designed to determine to what extent hydrologically dissimilar basins are useful to one another for model training. In experiment 2a, basins are divided into two clusters (which are referred to as C0 and C1) using the K -means method described in Sect. 2.5. The reasoning behind two clusters is to maximise the hydrological dissimilarity between basins in each cluster (based on the static basin attributes). In the first experiment, for each cluster, a baseline model is trained on 32 basins that belong to that cluster. Next, we compare the effects of adding 32 similar basins (labelled as + similar), by adding 32 dissimilar basins to the training set (i.e. from the other cluster, labelled as + dissimilar). In all cases, only the original 32 basins are evaluated (those used to train the baseline model); the performance of the additional training basins is not reported. This produces five unique training sets: 32 basins in cluster 0, 64 basins in cluster 0, 32 basins in cluster 1, 64 basins in cluster 1, and 32 basins in each cluster 0 and 1.

Next in experiment 2b, experiment 2a is repeated but with cluster-based streamflow undersampling applied to the training dataset. The resulting models are trained on half the amount of training data as those in experiment 2a. This experiment provides a comparison point between models trained in 32 basins without CUS_Q and models trained on 64 basins with CUS_Q, as both configurations have the same number of samples. The 32-basin configurations have greater temporal representation within the evaluation basins, while the 64-basin CUS_B configurations have greater spatial diversity, at the expense of temporal data. These comparisons reveal which is more useful to model generalisation: temporal data from within the subject basin or data from outside the basin. One distinction in this experimental configuration is that it does not use one-hot-encoded basin labels, as many of the test basins are not included in the training dataset. The labels are removed to ensure that the model does not develop any dependencies on them, since they are not available for the test basins. The models trained for experiments 2a and 2b are summarized in Table 3.

Table 3. LSTM size and training datasets used in experiment 2. Configurations are grouped by experiment.

Experiment	Code	Cells	n basins ($k = 0$)	n basins ($k = 1$)	Training data (years)	n samples (thousands)
Exp. 2a	cluster(0)- n (32)	128	32	0	12	140
Exp. 2a	cluster(0)- n (64)	128	64	0	12	280
Exp. 2a	cluster(1)- n (32)	128	0	32	12	140
Exp. 2a	cluster(1)- n (64)	128	0	64	12	280
Exp. 2a	cluster(0,1)- n (64)	128	32	32	12	280
Exp. 2b	cluster(0)- n (32) cus	128	32	0	6	70
Exp. 2b	cluster(0)- n (64) cus	128	64	0	6	140
Exp. 2b	cluster(1)- n (32) cus	128	0	32	6	70
Exp. 2b	cluster(1)- n (64) cus	128	0	64	6	140
Exp. 2b	cluster(0,1)- n (64) cus	128	32	32	6	140

3 Results and discussion

3.1 Experiment 1a: cluster-based temporal undersampling

Examples of temporal clustering results are provided in Fig. 2 for basin 01AD003 and Fig. 3 for basins 01AD003 and 07AF002 (located along the McLeod River in Alberta), where each belong to different basin clusters from Sect. 3.3. These results are for six clusters and minimum cluster sizes of 365. Although associations between clusters and specific hydrological characteristics can be expected to vary between individual basins, the results from basins 01AD003 and 07AF002 characterise four seasonal periods, as well as rising and receding limbs. Distinguishing between rising and falling limbs is consistent with previous studies that used streamflow clustering (Toth, 2009). Ensuring that distinct seasons are represented in the clustering results is important, as streamflow drivers are known to change throughout the year.

The performance of models trained on a set of 64 randomly sampled basins is shown in Fig. 4 in terms of NSE (a–c) and PI (d–f) for three cases: without resampling, with cluster-based temporal undersampling (CUS_Q), and random temporal undersampling. The cumulative density functions (CDFs), which have an optimum shape “J”, represent the proportion of basins (in the evaluation set) that fall below the performance along the x axis. The baseline model (no resampling) performs reasonably well across basins, with 100 % and 75 % of basins achieving an NSE greater than 0.5 at the 1 and 3 d lead times, respectively. Roughly 95 % of basins achieve a positive PI, indicating lower error than the naive model.

In comparison, the basin sets with CUS_Q at a rate of 0.5 (meaning that they use 6 out of 12 years of available training data) achieve the same level of performance as the baseline for both the 6-cluster and 12-cluster cases. The RUS model trained on 6 years of randomly sampled data performs very poorly. Similarly, the CUS_Q model trained on 3 years of

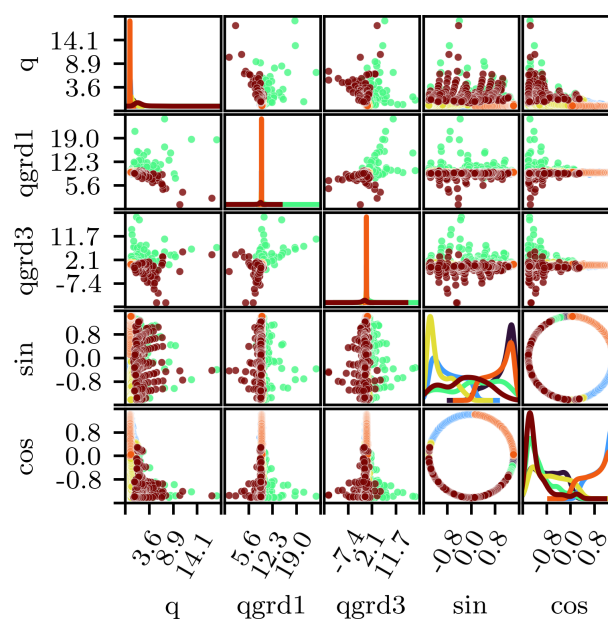


Figure 2. Scatterplot matrix illustrating the temporal clustering results for basin 01AD003 streamflows ($K = 6$). Markers are coloured by cluster and kernel smoothed histograms are shown along the diagonal. Axis labels q , $qgrd1$, $qgrd3$, \sin , and \cos are the short forms for q_t , $q_{t-1} - q_t$, $(q_{t-3} - q_t)/3$, $\sin^{-1}(t/365)$, and $\cos^{-1}(t/365)$, respectively.

cluster-based data performs poorly, indicating that key hydrological processes are no longer sufficiently represented in the reduced training data. This also indicates that a lower limit of the extent to which CUS_Q can be used is somewhere between an undersampling rate of 0.25 and 0.5. Finally, in no cases do any undersampled configurations outperform the model trained on all data.

These results highlight how a simple clustering method can be used to efficiently identify subsets of data that are sufficiently representative of the hydrological processes, which are identified by the clustering method, contained in each

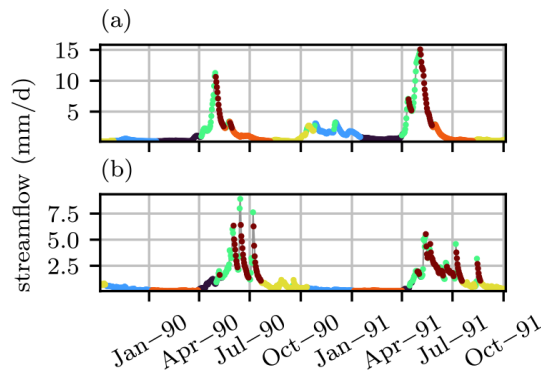


Figure 3. Hydrographs for basin 01AD003 (a) and basin 07AF002 (b) with observations from October 1989–October 1991, coloured by cluster. Cluster colours are arbitrarily assigned.

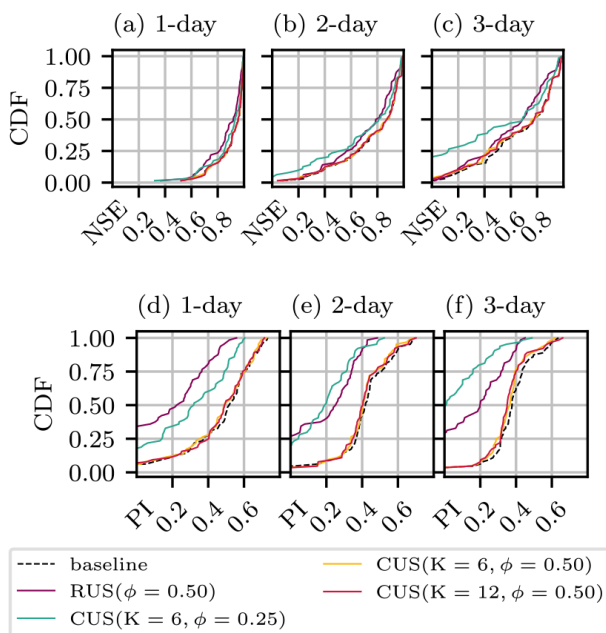


Figure 4. CDFs showing model performance according to NSE (a–c) and PI (d–f) for RUS (red), two CUS_Q configurations (blue and green), and no resampling baseline (black). Panels (a), (b), and (c), as well as (d), (e), and (f), correspond to forecasts of 1, 2, and 3 d, respectively.

basin such that there is no loss in temporal generalisation. In addition, these results show that a significant proportion of hydrological data within a continuous series are redundant and needlessly add to the computational burden of training, which is especially relevant to computationally expensive tasks such as hyperparameter optimisation (HPO). Reducing the computational requirement of HPO speeds up model development or allows for more extensive HPO, potentially improving model accuracy and thus FEWS reliability. Improvements in runtimes, which are reported in Tables S3 and S4 in

the Supplement, were typically found to be proportional to undersampling rates.

A limitation of this experiment, as well as the subsequent experiments, is that they were conducted using AR LSTMs. If AR inputs are not available, the model hyperparameters would need to be reconfigured, and model performance would be expected to decrease. While the experimental results are expected to transfer to non-AR models, this would need to be empirically confirmed. They would need to be repeated to confirm transferability.

3.2 Experiment 1b: cluster-based spatial undersampling

In experiment 1b, CUS_B is used to select training basins. As with temporal undersampling, reducing the number of training basins required to train models has the potential to drastically reduce the computational demand of training, especially across large domains such as the thousands of gauged basins spread across Canada.

First, a baseline model is trained on a set of 128 randomly sampled basins. The basins are then grouped into K clusters, sampled at rates of 0.5 and 0.75. As with the previous experiment, RUS configurations are included at the same resampling rates, which in this case consist of models trained on randomly sampled basin sets. The CDFs for each training configuration are shown in Fig. 5. Unlike the temporal streamflow undersampling in experiment 1a, the basin subsets are unable to match the performance of the baseline model that is trained on the complete set of basins, which is visibly shown by the CDFs for PI. As expected, the configurations with a greater number of training basins perform closest to the baseline. The CUS_B configurations narrowly outperform the RUS configurations with the same undersampling rates, most of all at a rate of 0.50.

3.3 Experiment 2: cross-comparison of two clusters of basins

To better understand the extent to which including hydrologically similar basins in the training dataset can benefit model performance, we consider an extreme case in which basins are grouped into two clusters, which are referred to as C0 and C1. The choice of two clusters is based on the maximum silhouette score, which is a commonly used measure of cluster cohesion (Rousseeuw, 1987). An important note on the result of the maximum silhouette score is that the result depends on the use of constrained K -means clustering, which uses a minimum cluster size of 64. A lower minimum cluster size tends to increase the optimum number of clusters; details are provided in the Supplement. Another reason for using two clusters is that it simplifies the analysis. For example, we compare the effects of similar and dissimilar clusters; in contrast, a greater number of clusters would require

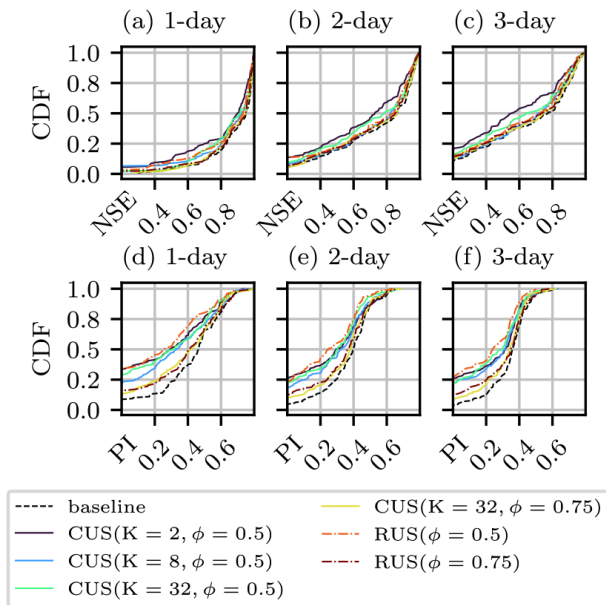


Figure 5. CDFs for NSE (a–c) and PI (d–f) for models trained on various basin subsets, which include a baseline (black line, includes all basins), cluster-based undersampling (coloured solid lines), and random undersampling (coloured dashed lines). The number of basins sampled from each cluster in the CUS_B configurations is equal to B/K. Panels (a), (b), and (c), as well as (d), (e), and (f), correspond to forecasts of 1, 2, and 3 d, respectively.

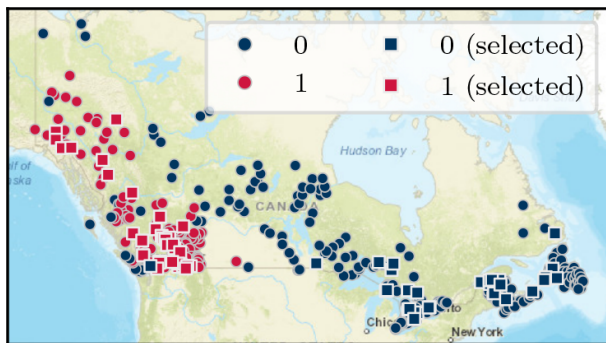


Figure 6. Clustering ($K = 2$) result for basins across Canada. For each cluster, 32 square markers indicate basins selected for the baseline and evaluation sets in the two-cluster experiments. World Street Map base map provided by Esri et al. (2022b).

a larger number of training configurations at varying degrees of hydrological similarity

Figure 6 shows the spatial distribution of the cluster labels across Canada. The C0 basins tend to be located at low elevations, along coastlines and east of the Rocky Mountains. In contrast, the C1 basins are mainly confined to higher elevations in the Rocky Mountains. Mean basin attributes for each cluster are provided in Table S2.

The five most important features, identified by applying an RF to unsupervised clustering outcome, include elevation,

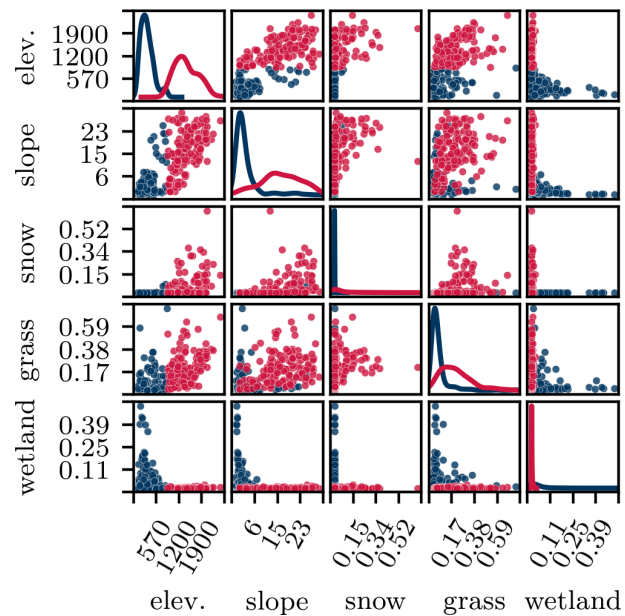


Figure 7. Scatterplot matrix illustrating the five most important features for clustering, as determined using the RF surrogate model. Markers are coloured by cluster, and kernel smoothed histograms are shown along the diagonal.

slope, and three land covers; they are shown in Fig. 7. The relevant features identified here by unsupervised clustering are consistent with the relevant descriptors deemed significant for determining hydrological similarity in the model-based method referenced in Oudin et al. (2010).

First, for each cluster, a model is trained on a set of 32 basins from that cluster. To measure the value of adding hydrologically similar basins, 32 additional basins are added to the baseline training dataset. Finally, to measure the effects of dissimilar basins, 32 dissimilar basins are instead added to the baseline training set. Performance metrics are only calculated for the baseline set of 32 basins. The above process is repeated for each cluster.

Figures 8 and 9 show the performance of the models evaluated on C0 and C1 basins, respectively. In C0 basins, more similar basins produce a notable improvement in performance across all lead times. Adding dissimilar basins instead produces even greater improvements, most notably according to the PI. The same trends are seen with the C1 forecasts, with an even greater difference between the scores of the “+ similar” and “+ dissimilar” training sets.

In C0 basins, adding more similar or dissimilar basins both improve the performance, with dissimilar basins producing larger improvements across all lead times and both metrics. A similar result is observed with the C1 basins, with the addition of dissimilar basins causing comparatively greater improvements in performance.

In experiment 2b, the configurations from experiment 2a are repeated but also incorporating CUS_Q, which is intro-

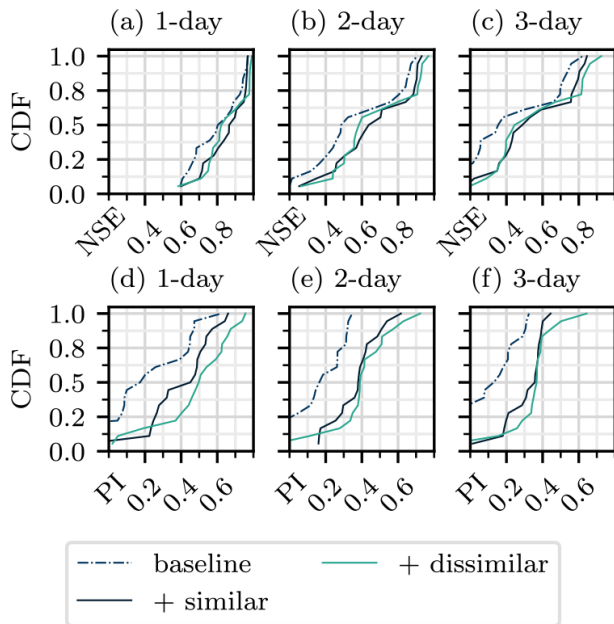


Figure 8. CDFs for models evaluated on C0 basins according to NSE (a–c) and PI (d–f). Each row compares models trained on three different training datasets: a baseline that includes 32 basins in the respective cluster, the baseline plus 32 basins in the same cluster, and the baseline plus 32 basins in the other cluster. Panels (a), (b), and (c), as well as (d), (e), and (f), correspond to forecasts of 1, 2, and 3 d, respectively.

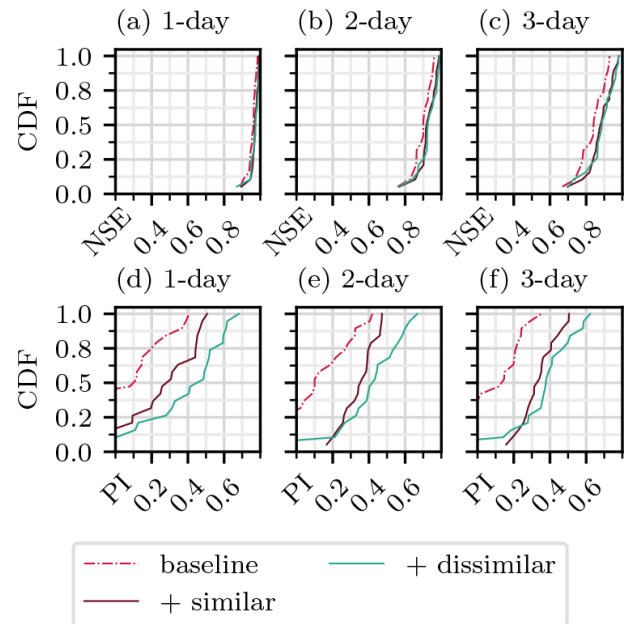


Figure 9. CDFs for models evaluated on C1 basins according to NSE (a–c) and PI (d–f). Each row compares models trained on three different training datasets: a baseline that includes 32 basins in the respective cluster, the baseline plus 32 basins in the same cluster, and the baseline plus 32 basins in the other cluster. Panels (a), (b), and (c), as well as (d), (e), and (f), correspond to forecasts of 1, 2, and 3 d, respectively.

duced in Sect. 3.1. Models from experiments 2a and 2b are compared in Fig. 10 for C0 and C1 evaluation sets. Consistent with the results from experiment 1a in Sect. 3.1, CUS_Q is found to have very little impact on model test performance, which is illustrated by the similar results shown when comparing results without and with CUS_Q undersampling, in panel pairs (a)–(c), (b)–(d), (e)–(g), and (f)–(h). This result indicates that for fixed training dataset size, data from outside the region of interest are much more useful to the training procedure than redundant data within the region of interest, identifiable using the clustering procedure.

3.4 Discussion

Collectively, these results reveal that the addition of training basins with distinctive hydrological characteristics is more useful in terms of improving model performance when compared to the addition of basins with similar characteristics. This outcome might be counterintuitive, since one could expect that training on additional basins that are most similar to the test set would be the most useful. This result also highlights the danger of training models on hydrologically similar basin sets, which is a common approach in the literature (Kratzert et al., 2024; Hashemi et al., 2022).

One explanation for this result is that the input feature set is missing key explanatory variables. Two basins could have

similar input vectors but different corresponding streamflow values. This difference may be explained by processes that are not captured within the input features. For example, two basins may appear to be hydrologically similar based on the available basin attributes and may have a different rainfall–runoff relationship, due to factors not included in the basin attributes, such as surficial geology or the presence of hydraulic structures such as dams. While the LSTM model should be able to distinguish between these two basins using the one-hot-encoded basin labels, incomplete explanatory variables may inhibit the ability of models to transfer learnt behaviour between basins.

Another explanation is that hydrologically similar basins contain a high degree of overlapping input–output patterns, reducing the amount of new information that can benefit predictions in the region of interest, in contrast to dissimilar basins. Information from dissimilar basins could be useful from a hydrological perspective or could simply provide a regularisation effect to the LSTM. Adding basins with distinct hydrological properties from some region of interest might occupy more neural pathways during model training compared to basins that have similar properties, which could mitigate overfitting to the region of interest. However, this explanation is not supported by the fact that constraining the number of cells in the network did not produce comparable regularisation.

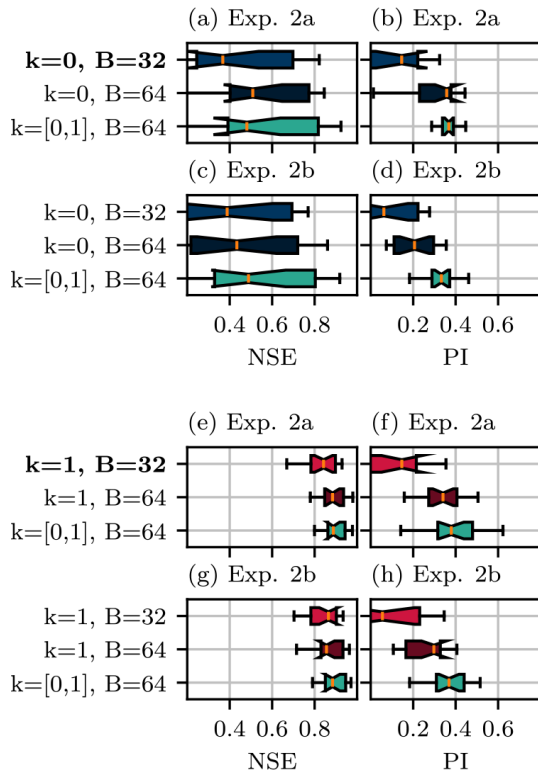


Figure 10. Boxplots showing the NSE and PI of models evaluated on 32 C0 basins for experiments 2a (a, b) and 2b (b, c) and on 32 C1 basins for experiments 2a (e, f) and 2b (g, h) for a lead time of 3 d. The baseline model, which is trained uniquely using the evaluation basins, is indicated in bold. Blue, red, and green colours indicate models trained on C0, C1, and both types of basins, respectively.

A final explanation is that while similar basins provide examples of similar hydrological behaviour, dissimilar basins provide examples of what not to predict. A simple analogy from image classification is that a model trained to classify photos of dogs might benefit more from being trained on some photos of cats than to be trained on more photos of dogs.

Lastly, to better understand the effect that additions of similar and dissimilar basins have on model performance, we consider small, incremental additions. First, we begin with a model trained on four basins in a given cluster. Next, we consider two additions to the training dataset: four C0 basins and four C1 basins. The addition that produces the best performance is retained, and the process is repeated. The models are retrained from scratch for each modification to the training dataset. The outcome is shown in Fig. 11 for models evaluated on four C0 basins (a) and four C1 basins (b). In both cases, the models benefit from the addition of four basins that belong to the same cluster; however, afterwards there are no clear preferences in terms of which basin clusters produce the best improvements. Despite some incremental additions hampering model performance, the performance for

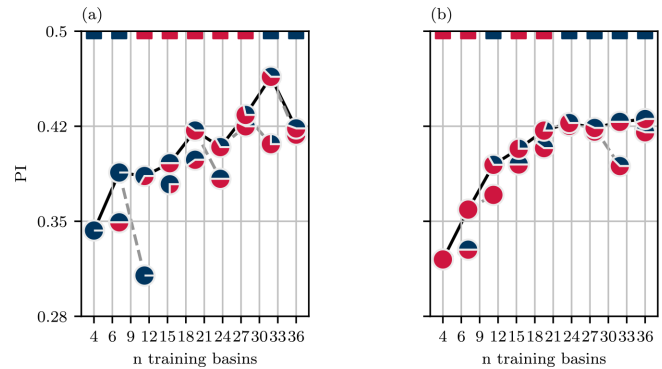


Figure 11. Model performance (PI) across increasing numbers of training basins for models evaluated on four C0 basins (a) and four C1 basins (b). Pie chart markers illustrate the proportion of C0 and C1 basins used in each training dataset. The coloured dashes along the top of each panel indicate which cluster produced the better addition to the training dataset.

each evaluation improves across the larger training dataset, but the improvements decay exponentially, which is consistent with other studies that have looked at model performance across increasing training data (Kratzert et al., 2024; Gauch et al., 2021b). In Fig. 11a, adding four similar basins to the eight-basin training dataset produces a large loss in performance, which highlights the lack of robustness against models trained on very small basin sets. Between C0 and C1 evaluations, C0 basins are more sensitive to new training data, and they are comparatively more likely than C1 basins to exhibit worse performance after the addition of new training basins. Since the size of four basins is a relatively small sample size, the experiment was repeated with four different basins in each cluster, producing similar results, which are shown in Fig. S4.

The experimental results detailed above all assert the importance of hydrologically diverse information-rich training datasets. This is of particular importance in small regions of interest, where far away dissimilar basins may be seen as not relevant to the training task. This study presents many opportunities for future work on curating datasets for hydrological models. While our study uses a simple clustering approach to quantify hydrological diversity, more sophisticated approaches, such as one based on mutual information, may further improve the results. Additionally, relatively little work exists on transferring models between different hydrological regions – which can potentially provide an improved starting point for model training, leading to better performance and more reliable FEWSs.

4 Conclusions

The selection of training data is amongst the most important factors contributing to the performance of streamflow fore-

casting models. Our study showed that the performance of flow forecasting models relies on diverse training data, using a novel use of cluster-based resampling to identify and maximise temporal and spatial hydrological diversity within training datasets. In the first set of experiments, cluster-based undersampling was used to eliminate redundant temporal data from training datasets, drastically reducing the computational demand of model training. The next set of experiments showed how, given some region of interest, data from hydrologically dissimilar basins can be much more useful than data from similar basins. This result is contrary to the intuitive approach of curating training basins for training, which is to train models to a group of hydrologically similar or proximal basins. This outcome also highlights the need for large and hydrologically diverse training datasets. The latter can be combined with cluster-based temporal undersampling to generate diverse training sets that produce more performative models, for a fixed number of training observations, compared to models trained without cluster-based undersampling. Finally, temporal and spatial undersampling routines are combined to demonstrate how, for a fixed number of training samples, spatial hydrological diversity is much more beneficial than temporal diversity. These findings are critical to improving the reliability and accuracy of flood forecasting models and minimising the effects of flooding.

Data availability. The data used in this research are described in the publication by Arsenaault et al. (2020) and available for download through <https://osf.io/rpc3w/> (Center for Open Science). Base maps for several figures were provided by Esri through the Contextily Python module (Esri et al., 2022a, b).

Supplement. The supplement related to this article is available online at <https://doi.org/10.5194/hess-29-785-2025-supplement>.

Author contributions. ES: conceptualisation, code, formal analysis, visualisation, writing (original draft). UTK: conceptualisation, funding acquisition, supervision, writing (editing), revisions.

Competing interests. The contact author has declared that neither of the authors has any competing interests.

Disclaimer. The views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the affiliated organisations.

Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every

effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. The authors would like to thank the HC3 research group at École de technologie supérieure for assembling the HYSETS database, which facilitated this work and other large-sample studies.

Financial support. This research has been supported by the Natural Sciences and Engineering Research Council of Canada (grant nos. RGPIN-2023-05077 and PGS-D 547587) and York University (Graduate funding).

Review statement. This paper was edited by Ralf Loritz and reviewed by Yalan Song and one anonymous referee.

References

- Abrahart, R. J. and See, L.: Comparing Neural Network and Autoregressive Moving Average Techniques for the Provision of Continuous River Flow Forecasts in Two Contrasting Catchments, *Hydrol. Process.*, 14, 2157–2172, [https://doi.org/10.1002/1099-1085\(20000815/30\)14:11/12<2157::AID-HYP57>3.0.CO;2-S](https://doi.org/10.1002/1099-1085(20000815/30)14:11/12<2157::AID-HYP57>3.0.CO;2-S), 2000.
- Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, *Hydrol. Earth Syst. Sci.*, 21, 5293–5313, <https://doi.org/10.5194/hess-21-5293-2017>, 2017.
- Anctil, F. and Lauzon, N.: Generalisation for neural networks through data sampling and training procedures, with applications to streamflow predictions, *Hydrol. Earth Syst. Sci.*, 8, 940–958, <https://doi.org/10.5194/hess-8-940-2004>, 2004.
- Arsenaault, R., Brissette, F., Martel, J.-L., Troin, M., Lévesque, G., Davidson-Chaput, J., Gonzalez, M. C., Ameli, A., and Poulin, A.: A Comprehensive, Multisource Database for Hydrometeorological Modeling of 14,425 North American Watersheds, *Scientific Data*, 7, 243, <https://doi.org/10.1038/s41597-020-00583-2>, 2020 (data available at: <https://osf.io/rpc3w/>, last access: 1 May 2024).
- Arsenaault, 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, *Hydrol. Earth Syst. Sci.*, 27, 139–157, <https://doi.org/10.5194/hess-27-139-2023>, 2023a.
- Arsenaault, 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, *Hydrol. Earth Syst. Sci.*, 27, 139–157, <https://doi.org/10.5194/hess-27-139-2023>, 2023b.
- Bennett, K., Bradley, P., and Demiriz, A.: Constrained K-means Clustering, *Tech. Rep. MSR-TR-2000-65*, Microsoft, <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-2000-65.pdf> (last access: 1 May 2024), 2000.
- Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y.: On the Properties of Neural Machine Trans-

- lation: Encoder-Decoder Approaches, arXiv [preprint] <https://doi.org/10.48550/arXiv.1409.1259>, 2014.
- Esri, Delorme, HERE: World Light Gray Basemap, Esri, <https://doc.arcgis.com/en/data-appliance/2022/maps/world-light-gray-base.htm> (last access: 1 May 2024), 2022a.
- Esri, Delorme, HERE, USGS, NRCAN: Streets Basemap, Esri, <https://www.arcgis.com/home/webmap/viewer.html?webmap=e833fc01f447451f93ec0dfdfe12eed7> (last access: 1 May 2024), 2022b.
- Fang, K., Kifer, D., Lawson, K., Feng, D., and Shen, C.: The Data Synergy Effects of Time-Series Deep Learning Models in Hydrology, *Water Resour. Res.*, 58, e2021WR029583, <https://doi.org/10.1029/2021WR029583>, 2022.
- Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., and Hochreiter, S.: Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network, *Hydrol. Earth Syst. Sci.*, 25, 2045–2062, <https://doi.org/10.5194/hess-25-2045-2021>, 2021a.
- Gauch, M., Mai, J., and Lin, J.: The Proper Care and Feeding of CAMELS: How Limited Training Data Affects Streamflow Prediction, *Environ. Modell. Softw.*, 135, 104926, <https://doi.org/10.1016/j.envsoft.2020.104926>, 2021b.
- Girihagama, L., Naveed Khaliq, M., Lamontagne, P., Perdikaris, J., Roy, R., Sushama, L., and Elshorbagy, A.: Streamflow Modelling and Forecasting for Canadian Watersheds Using LSTM Networks with Attention Mechanism, *Neural Computing and Applications*, 34, 19995–20015, <https://doi.org/10.1007/s00521-022-07523-8>, 2022.
- Hashemi, R., Brigode, P., Garambois, P.-A., and Javelle, P.: How can we benefit from regime information to make more effective use of long short-term memory (LSTM) runoff models?, *Hydrol. Earth Syst. Sci.*, 26, 5793–5816, <https://doi.org/10.5194/hess-26-5793-2022>, 2022.
- Kitanidis, P. K. and Bras, R. L.: Real-time Forecasting with a Conceptual Hydrologic Model: 2. Applications and Results, *Water Resour. Res.*, 16, 1034–1044, <https://doi.org/10.1029/WR016i006p01034>, 1980.
- KLEMEŠ, V.: Operational Testing of Hydrological Simulation Models, *Hydrolog. Sci. J.*, 31, 13–24, <https://doi.org/10.1080/02626668609491024>, 1986.
- Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G.: Uncertainty estimation with deep learning for rainfall–runoff modeling, *Hydrol. Earth Syst. Sci.*, 26, 1673–1693, <https://doi.org/10.5194/hess-26-1673-2022>, 2022.
- Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, *Hydrol. Earth Syst. Sci.*, 23, 4323–4331, <https://doi.org/10.5194/hess-23-4323-2019>, 2019.
- 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 Resour. Res.*, 55, 11344–11354, <https://doi.org/10.1029/2019WR026065>, 2019a.
- Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, *Hydrol. Earth Syst. Sci.*, 23, 5089–5110, <https://doi.org/10.5194/hess-23-5089-2019>, 2019b.
- 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, *Scientific Data*, 10, 61, <https://doi.org/10.1038/s41597-023-01975-w>, 2023.
- Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin, *Hydrol. Earth Syst. Sci.*, 28, 4187–4201, <https://doi.org/10.5194/hess-28-4187-2024>, 2024.
- Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., and Dadson, S. J.: Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models, *Hydrol. Earth Syst. Sci.*, 25, 5517–5534, <https://doi.org/10.5194/hess-25-5517-2021>, 2021.
- 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.
- Mai, J., Shen, H., Tolson, B. A., Gaborit, É., Arsenaault, R., Craig, J. R., Fortin, V., Fry, L. M., Gauch, M., Klotz, D., Kratzert, F., O’Brien, N., Princz, D. G., Rasiya Koya, S., Roy, T., Seglenieks, F., Shrestha, N. K., Temgoua, A. G. T., Vionnet, V., and Waddell, J. W.: The Great Lakes Runoff Intercomparison Project Phase 4: the Great Lakes (GRIP-GL), *Hydrol. Earth Syst. Sci.*, 26, 3537–3572, <https://doi.org/10.5194/hess-26-3537-2022>, 2022.
- Nearing, G. S., Klotz, D., Frame, J. M., Gauch, M., Gilon, O., Kratzert, F., Sampson, A. K., Shalev, G., and Nevo, S.: Technical note: Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks, *Hydrol. Earth Syst. Sci.*, 26, 5493–5513, <https://doi.org/10.5194/hess-26-5493-2022>, 2022.
- Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., Voloshin, D., Kratzert, F., Elidan, G., Dror, G., Begelman, G., Nearing, G., Shalev, G., Noga, H., Shavitt, I., Yuklea, L., Royz, M., Giladi, N., Peled Levi, N., Reich, O., Gilon, O., Maor, R., Timnat, S., Shechter, T., Anisimov, V., Gigi, Y., Levin, Y., Moshe, Z., Ben-Haim, Z., Hassidim, A., and Matias, Y.: Flood forecasting with machine learning models in an operational framework, *Hydrol. Earth Syst. Sci.*, 26, 4013–4032, <https://doi.org/10.5194/hess-26-4013-2022>, 2022.
- Oudin, L., Kay, A., Andréassian, V., and Perrin, C.: Are Seemingly Physically Similar Catchments Truly Hydrologically Similar?, *Water Resour. Res.*, 46, W11558, <https://doi.org/10.1029/2009WR008887>, 2010.
- Rousseeuw, P. J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis, *J. Comput. Appl. Math.*, 20, 53–65, [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7), 1987.
- Snieder, E., Abogadil, K., and Khan, U. T.: Resampling and ensemble techniques for improving ANN-based high-flow forecast accuracy, *Hydrol. Earth Syst. Sci.*, 25, 2543–2566, <https://doi.org/10.5194/hess-25-2543-2021>, 2021.
- Snieder, E. J.: Artificial Neural Network-Based Flood Forecasting: Input Variable Selection and Peak Flow Prediction Accuracy, PhD thesis, York University, <http://hdl.handle.net/10315/36792> (last access: 1 May 2024), 2019.
- Toth, E.: Classification of hydro-meteorological conditions and multiple artificial neural networks for streamflow

- forecasting, *Hydrol. Earth Syst. Sci.*, 13, 1555–1566, <https://doi.org/10.5194/hess-13-1555-2009>, 2009.
- Tyralis, H., Papacharalampous, G., and Langousis, A.: A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources, *Water*, 11, 910, <https://doi.org/10.3390/w11050910>, 2019.
- Xiang, Z., Yan, J., and Demir, I.: A Rainfall-Runoff Model With LSTM-Based Sequence-to-Sequence Learning, *Water Resour. Res.*, 56, e2019WR025326, <https://doi.org/10.1029/2019WR025326>, 2020.
- Zhang, Y., Ragettli, S., Molnar, P., Fink, O., and Peleg, N.: Generalization of an Encoder-Decoder LSTM Model for Flood Prediction in Ungauged Catchments, *J. Hydrol.*, 614, 128577, <https://doi.org/10.1016/j.jhydrol.2022.128577>, 2022.