HESS Opinions: Never train an LSTM on a single basin

Frederik Kratzert[1], Martin Gauch[2], Daniel Klotz[3], and Grey Nearing[4]

[1]Google Research, Vienna, Austria
[2]Google Research, Zurich, Switzerland
[3]Helmholtz Centre for Environmental Research - UFZ, Leipzig, Germany
[4]Google Research, Mountain View, California, USA

Correspondence: Frederik Kratzert (kratzert@google.com)

Abstract. Machine learning (ML) has played an increasing role in the hydrological sciences. In particular, Long Short-Term Memory networks (LSTMs) are popular for rainfall–runoff modeling. A large majority of studies that use this type of model do not follow best practices, and there is one mistake in particular that is common: training deep learning models on small, homogeneous data sets, typically data from only a single hydrological basin. In this position paper, we show that LSTM rainfall-runoff models are best when trained with data from a large number of basins.

1 Machine learning requires different intuitions about hydrological modeling

Regionalizing rainfall–runoff models across multiple watersheds is a longstanding problem in the hydrological sciences (Guo et al., 2021). The most accurate streamflow predictions from conceptual and process-based hydrological models generally require calibration to data records in individual watersheds. Hydrology models based on machine learning (ML) are different – ML models work best when trained on data from many watersheds (Nearing et al., 2021). In fact, this is one of the main benefits of ML-based streamflow modeling.

Because ML models are trained with data from multiple watersheds, they are able to learn hydrologically diverse rainfall–runoff responses (Kratzert et al., 2019b) in a way that is useful for prediction in ungauged basins (Kratzert et al., 2019a). However, prediction in ungauged basins is not the only reason to train ML models on data from multiple watersheds. Models trained this way have better skill even in individual, gauged watersheds with long training data records, and they are also better at predicting extreme events (Frame et al., 2022).

The purpose of this paper is to effect a change in intuition. ML requires a top-down modeling approach, in contrast to traditional hydrological modeling that is usually most effective with a bottom-up approach. We do not mean top-down vs. bottom-up in the sense discussed by Hrachowitz and Clark (2017), who use these terms to differentiate between lumped, conceptual (top-down) vs. distributed, process-based (bottom-up) models. Instead, we mean that traditional hydrology models (both lumped conceptual models and process-based models) are typically developed, calibrated, and evaluated at a local scale, ideally using long and comprehensive data records. Then, in this bottom-up approach, after a model is developed, we might work on regionalization strategies to extrapolate parameters and parameterizations to larger areas (e.g., Samaniego et al., 2010; Beck et al., 2016). In contrast, with ML modeling the best approach is to start by training on all available data from as many

watersheds as possible, and then work to fine tune models for individual catchments. The effort then goes into localizing large scale models, instead of regionalizing small scale models.

This paper focuses on rainfall–runoff modeling with Long Short-Term Memory (LSTM) networks because this is currently the most common type of ML model used in surface hydrology. The use of LSTMs for rainfall–runoff modeling is motivated by the fact that LSTMs are state-space models and are therefore structured similarly to how hydrologists conceptualize watersheds (Kratzert et al., 2018). The use of LSTMs in hydrology research has increased exponentially in the last several years (see Fig. 1). We see no reason to suspect that the lessons learned about big data with this type of model are not general, and several reasons to suspect that they are; namely it is important to recognize that, across application domains, machine learning models trained on large training data sets out-perform smaller, more specialized models (Sutton, 2019).

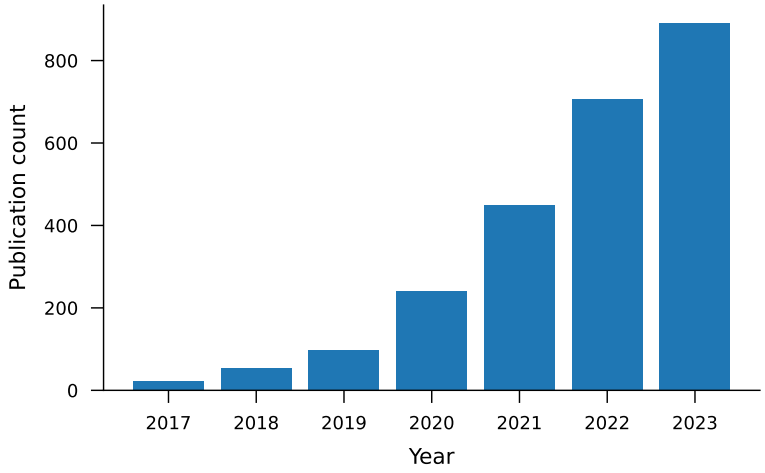

Figure 1. Number of hydrological publications related to rainfall–runoff modeling with LSTMs over time, based on data retrieved from Google Scholar in April 2024.

To understand the current state of practice with LSTM-based rainfall runoff modeling, we collected papers returned by a keyword search on Google Scholar for "*rainfall-runoff modeling LSTM streamflow*" sorted by relevance. From this search, we surveyed the top 50 papers per year for the years 2021, 2022, and 2023 and skipped papers that did not involve training models or developing systems for training models. Of those 150 papers surveyed, 122 trained models on individual catchments and 28 trained models on multiple catchments (of which four were co-authored by one or more authors of this paper) (Fig. 2). We collected these 150 papers for review in April, 2024, more than four years after the original regional LSTM rainfall–runoff modeling papers (Kratzert et al., 2019a, b) were published. The list of 150 papers is included in the data repository released with this paper.

It is important to recognize that there is usually no reason in practice to train LSTM streamflow models using data from only a small number of watersheds. There is enough publicly available streamflow data to train robust ML models. For example,

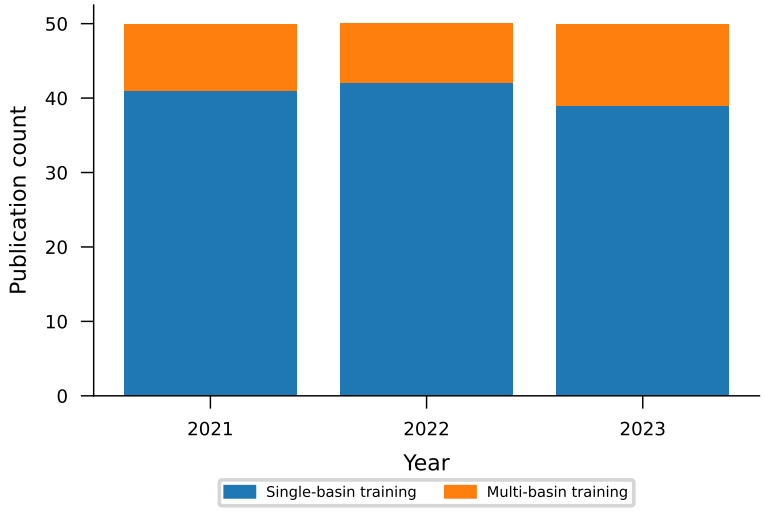

Figure 2. Fractions of 50 peer-reviewed research articles papers per year (starting one year after the original LSTM regional modeling papers were published Kratzert et al., 2019a, b) that train LSTM rainfall–runoff models on single vs. multiple basins, based on articles retrieved from Google Scholar in April 2024.

the various CAMELS datasets such as CAMELS-US (Newman et al., 2015), the Global Runoff Data Center (BAFG), or the Caravan dataset and its extensions (Kratzert et al., 2023). It is possible to fine tune large–sample models to individual locations and/or for specific purposes (e.g., Ma et al., 2021), however fine tuning is outside the scope of this paper. It is sufficient for our purpose to show that training large-scale ML models is better than training small-scale ML models for streamflow prediction and fine tuning would only widen this difference.

In summary, the large majority of LSTM papers published in hydrology journals train models on small datasets from single catchments. This is unfortunate because it does not leverage the primary benefit of machine learning, which is the ability to learn and generalize from large datasets. The rest of this paper illustrates how and why that is a problem when using LSTMs specifically for rainfall-runoff modeling.

2 Skill gaps between local and regional models

Figure 3 shows differences in performance between models trained on single basins vs. multiple basins (regional). Subpanel (a) shows this comparison for two traditional hydrology models and subpanel (b) shows the comparison for LSTM models. Notice that in subpanel (a), single-basin models perform better than regional models, and in subpanel (b) this is reversed.

Subpanel (a) of Fig. 3 shows cumulative density functions (CDFs) over Nash–Sutcliffe Efficiencies (NSEs) for 489 CAMELS basins from a conceptual model (mHM) and a process-based model (VIC). These models were calibrated and run by other research groups without our involvement, and the data from these models were borrowed from the benchmarking study by

Kratzert et al. (2019b). The fact that conceptual and process-based hydrological models perform worse when regionally calibrated is, to our knowledge, a consistent finding across hydrological modeling studies (e.g., Beck et al., 2016; Mizukami et al., 2017).

Subpanel (b) of Fig. 3 shows the same NSE CDFs for LSTM models. We tuned the hyperparameters and trained an ensemble of ten single-basin LSTM models separately for each of 531 basins in the CAMELS data set (Newman et al., 2015; Addor et al., 2017) using a standard train/validation/test data split with approximately ten years of data in each split. We similarly trained an ensemble of ten LSTM regional models with data from all 531 CAMELS catchments simultaneously using hyperparameters taken from Kratzert et al. (2021). Details about how LSTM models were hypertuned, trained, and tested can be found in Appendix A.

The choice to use 531 CAMELS basins for training and testing LSTMs comes from the suggestion by Newman et al. (2017), who selected these basins from the full CAMELS dataset for model benchmarking. We use this set of CAMELS benchmark gauges for the remainder of this study. Figure B1 in Appendix B shows the same comparison as Fig. 3, but subpanel (b) in that figure shows NSE CDFs for only the 489 CAMELS basins with mHM and VIC runs.

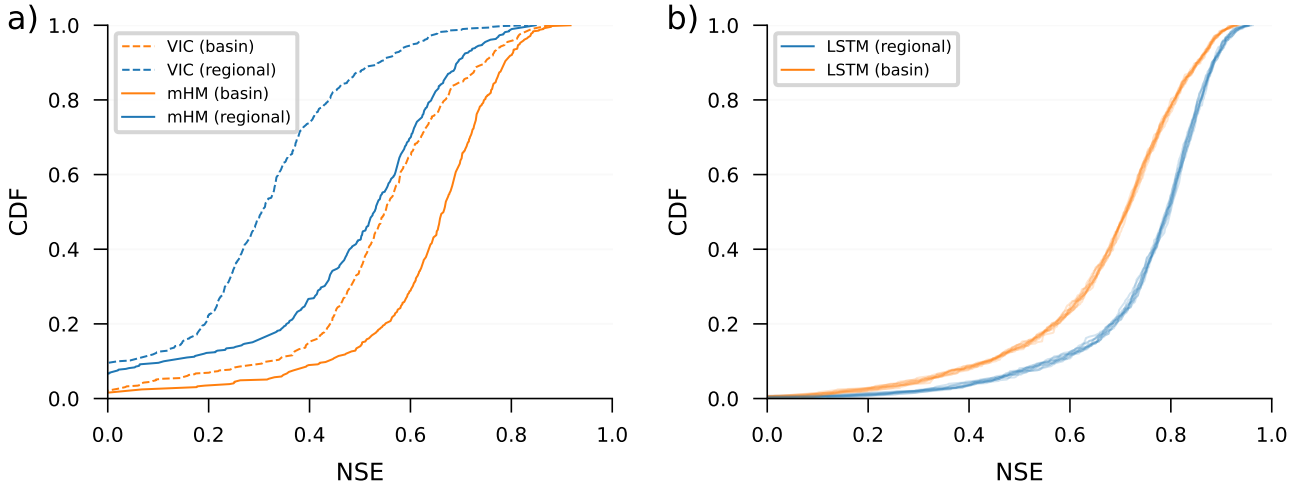

Figure 3. Cumulative Density Functions (CDF) of Nash–Sutcliffe Efficiencies (NSE) of simulated volumetric discharge over watersheds in the CAMELS data set for models trained on individual basins (basin; orange lines) vs. on multiple basins (regional; blue lines). Subplot (a) shows a conceptual model (mHM) and a process-based model (VIC), both calibrated using data from 489 watersheds by (different) research groups that are familiar with each model – VIC single basins: Newman et al. (2017), VIC regional: Mizukami et al. (2017), mHM single basins: Mizukami et al. (2019), and mHM regional: Rakovec et al. (2019). Subplot (b) shows the same NSE CDFs from LSTMs trained in two ways: regional LSTMs (blue) were trained using data over the training period from all 531 CAMELS basins, and single-basin LSTMs (orange) were trained using data over the training period from each CAMELS basin individually. An ensemble of ten LSTM models were trained in all cases, where randomness in the LSTM repetitions is due to randomness in the initial weights prior to training. We recommend averaging hydrographs from this kind of repetition, as was done by Kratzert et al. (2019a) and Kratzert et al. (2019b).

The takeaway from Fig. 3 is that whereas traditional hydrological models are more accurate when calibrated to a single watershed, LSTMs are more accurate when trained on data from many watersheds. Other hydrological metrics are reported in Appendix E, however the main point of our argument holds regardless of which metric is used for evaluation.

3 Why this matters for extreme events

Training on large-sample data sets with hydrologic diversity means that the training envelope is larger, making it less likely that any new prediction will be an extrapolation. Intuitively, the training envelope refers to the ranges of data where model performance is well-supported by the training process. If the training set includes a very humid basin, then the model is more likely to have seen large precipitation events, so that a new extreme precipitation event seen during inference is less likely to be outside of the training envelope. As an example of this, Nearing et al. (2019) discussed how watersheds can move within the training envelope as (e.g., climate) conditions within a catchment change, and how this causes changes in the modeled rainfall–runoff response in individual watersheds.

We can look at the target data to see an example of how this diversity in training data helps. The LSTM model used by Kratzert et al. (2019a) and Kratzert et al. (2019b) have a linear "head" layer that produces a scalar estimate of streamflow at each timestep by taking a weighted sum of the values of the LSTM hidden state. The weights for this weighted sum in the head layer are parameters that are tuned during training. The LSTM hidden state is a real-valued vector in (-1, 1), and has a size equal to the number of cell states or memory states in the LSTM (for a hydrologically-centered overview of the structure of an LSTM, please see Kratzert et al., 2018), which means that the maximum (limiting) value of the scalar streamflow estimate from the model is defined by the sum of the absolute values of weights in the head layer (see also Appendix C). More diversity in training data (here, training targets) causes the model to expand the range of weights in the head layer to accommodate higher flow values.

This effect can be seen in Fig. 4, which shows the theoretical maximum prediction from each of 531 single-basin LSTM models and from a single LSTM model trained on all 531 CAMELS basins. During inference (test period) there are a total of 10 streamflow observations across all 531 catchments that are above the regional model's theoretical maximum when trained on data from all 531 watersheds. However when separate models are trained per catchment, there are more than 6,000 streamflow observations that are above the theoretical maximums for each model in its respective catchment. Figure 5 shows how this effect manifests in an example hydrograph from one particular basin (not chosen at random). More examples are given in Appendix D. Notice that no model captures all of the extreme events, even in the training data set (which is common for physically-based models as well; Frame et al., 2022).

In summary, training on a larger dataset means the model is able to adapt internal weights and biases to account for more extreme events.

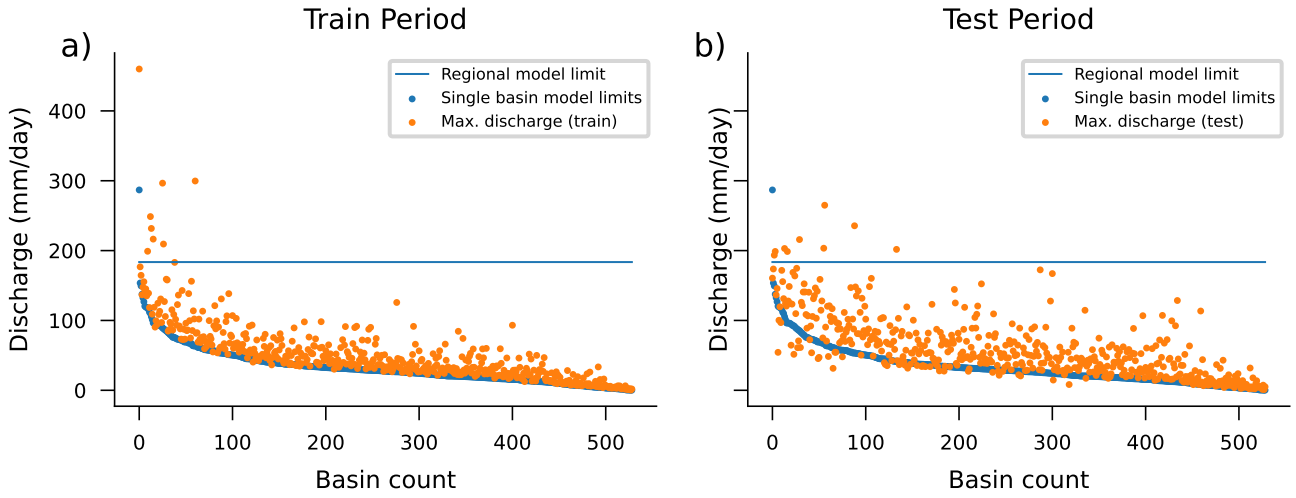

Figure 4. Theoretical maximum streamflow prediction of an LSTM model with a linear head layer when trained per basin (blue dots) and for all 531 CAMELS basins together (blue line). Orange dots represent the maximum streamflow values per basin in the train period (subpanel a) and test period (subpanel b). In the test period, there are 10 flow values above the maximum prediction of the regional model, while there are more than 6,346 flow values above the theoretical maximums of their respective single-basin models across all basins.

4 How many basins are necessary?

Figure 6 shows how test period performance increases as more basins are added to the training set. The blue line in this figure was created by grouping 531 CAMELS basins into different sized training and test groups randomly without replacement. This was done using k-fold cross-validation so that all models were tested on the same basin(s) where they were trained (during different time periods) and every basin was used as a test basin exactly once in each grouping size. For example, the 531 basins were grouped into, e.g., 5 disjoint groups (each with around 107 basins), then data from the training period (1999 - 2000) of one group was used to train an ensemble of ten LSTM models, and data from the test period (1980 - 1989) of the same group was used to evaluate that ensemble of trained LSTM models. This procedure was repeated for each of the remaining four groups, and a similar procedure was used for each different size of basin grouping shown along the x-axis of Fig. 6. Figure 6 plots the average (over ten ensemble members) of the median (over 531 CAMELS basins) test-period NSE for various basin groupings. More details about basin groupings can be found in Appendix A3.

The blue line in Fig. 6 shows performance (median NSE) increasing as the size of the training dataset increases. This effect continues up to the maximum size of the CAMELS data set (531 basins). In other words, it is better to have more basins in the training set, and even these 531 basins are most likely not enough to train optimal LSTM models for streamflow.

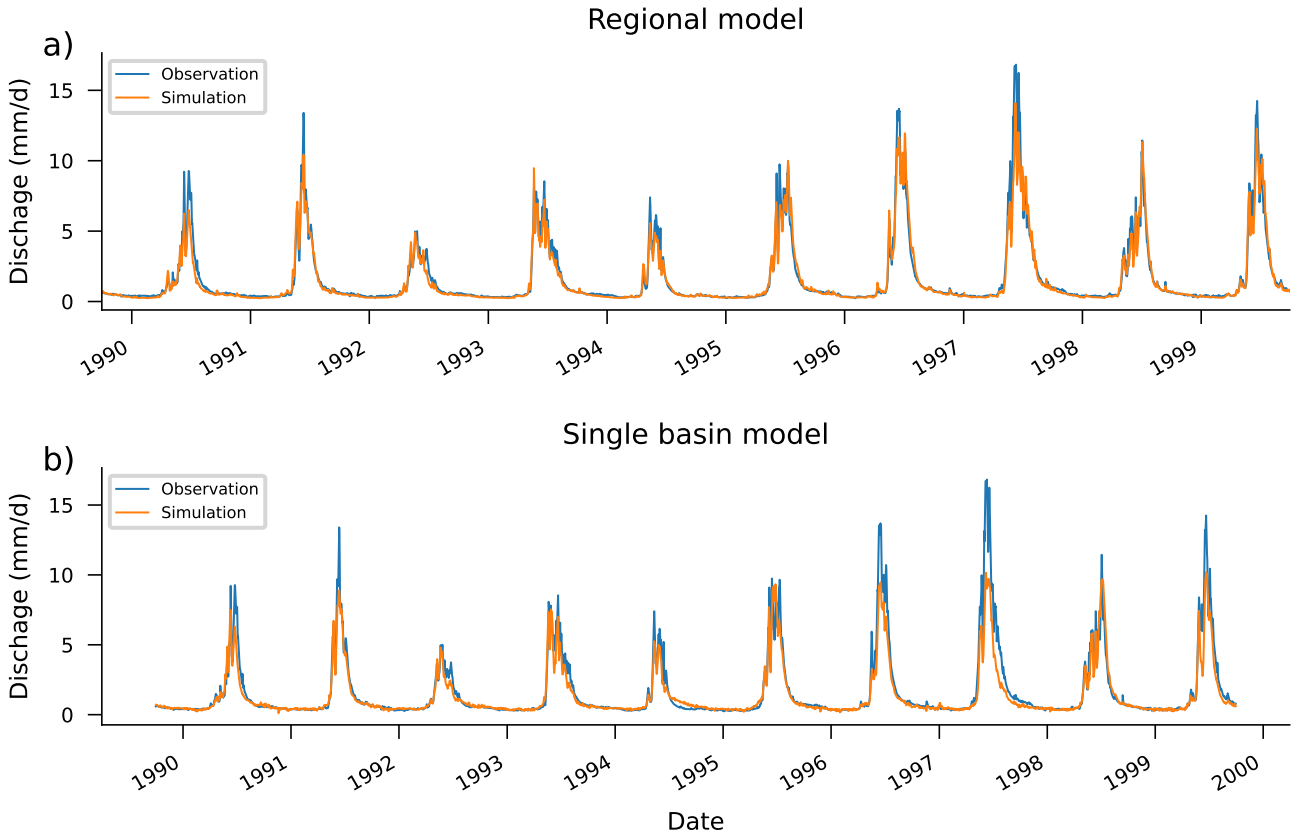

Figure 5. Observed and simulated hydrographs from the test period (1989 - 1999) in a particular basin (13011900). Notice how the high-flow effect outlined in Fig. 4 manifests in the differences between hydrographs predicted by a regional LSTM (subpanel a) vs. a single-basin LSTM (subpanel b). This example was chosen to highlight this effect (not chosen randomly), however the effect is similar in most basins. This gauge is on Lava Creek in Wyoming with a drainage area of 837 square kilometers and exhibits a strong seasonal flow pattern.

5 Is hydrological diversity always an asset?

There are at least two factors to consider when choosing training data: volume and variety. Volume refers to the total amount of data used for training (more is always better, as far as we have seen), and variety refers to the (hydrologic) diversity of data. Diversity might be in the form of different geophysical catchment attributes, different types and magnitudes of events, or different hydrological behaviors.

Figure 6 provides examples of training on less hydrologically diverse basin groups. The orange line in Fig. 6 shows the mean (over ten ensemble members) of the median NSE (over 531 CAMELS basins) from training and testing models on basins grouped by USGS hydrological unit codes (HUCs). There are 18 HUCs represented in the CAMELS data set, with between 2 and 79 basins per HUC. The green line in Fig. 6 shows the mean (over ten ensemble members) of the median

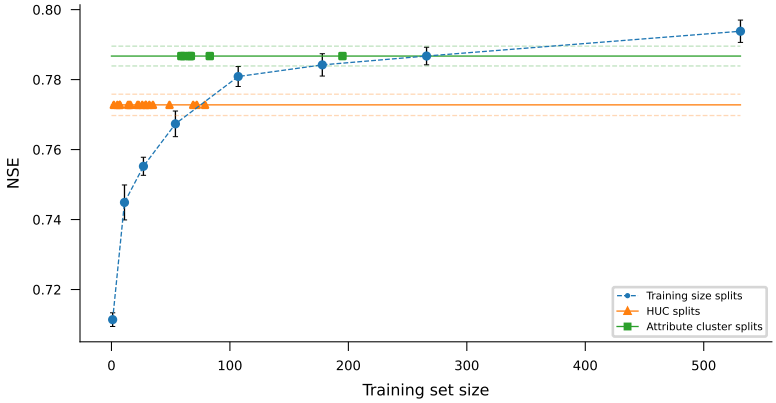

Figure 6. Median NSE scores of simulated volumetric discharge over 531 CAMELS basins for LSTM models trained and tested on splitting the 531 CAMELS basins into different groupings for training and testing. All models were tested on the same basins where they were trained (during different time periods). The blue line represents basin groupings of different sizes chosen randomly without replacement so that all 531 basins were modeled exactly once for each grouping size. Orange and green lines show results from grouping the 531 basins into sets based on USGS hydrological unit codes (orange), and k-means clustering of basin attributes (green). Dots on the orange and green lines indicate the sizes of the (18 and 6) basin groups used in those splits, since there are a different number of CAMELS basins in each HUC and a different number of basins in each attribute cluster. Blue dots and solid orange and green lines represent the median model performance over 531 basins, while blue error bars and dashed orange and green lines represent the standard deviation of this median over a ten-member ensemble.

NSE (over 531 CAMELS basins) from training and testing models on basin groups derived from k-means clustering on static catchment attributes. The CAMELS data set includes catchment attributes related to climate, vegetation, pedology, geology, and topography, and we clustered using 25 catchment attributes described in Tab. A1 in Appendix A5. We selected a k-means cluster model based on a maximin criterion on silhouette scores, which resulted in a model with 6 clusters ranging from 59 to 195 basins per cluster. Details about how models were trained and tested on HUCs and attribute clusters can be found in Appendices A4 and A5.

Figure 6 provides evidence that there *might* be ways to construct training sets that could potentially result in better models than simply training on all available streamflow data. This conclusion is hypothetical because in all examples shown in Fig. 6, models trained on any subset of the 531 CAMELS basins performed worse, on average, than models trained on all 531 CAMELS basins. *However*, separating the training set into hydrologically similar groups of basins results in models that perform better than models trained on random basin groups of similar size. **It is an open question as to whether a larger data set than CAMELS (e.g., Kratzert et al., 2023) might be divisible into hydrologically similar groups that individually perform better than a model trained on all available data.** This could happen if, for example, the curve in Fig. 6 becomes asymptotic at some point beyond the size of the CAMELS data set, and if the performance of models trained on

hydrologically-informed basin groups continues to increase with sample size. Note that this analysis does not account for the value of hydrologic diversity for prediction in ungauged basins.

The takeaway is that even if enough basins exist to divide your training data into hydrologically-informed training sets, one is likely better off simply training a single model with all available data. At least, one should perform an analysis like what is shown in Fig. 6 to understand whether splitting the training set helps or hurts. We are interested to see (through future work) what these tradeoffs look like with larger training sets.

6 Are bigger models better everywhere?

Even though the best model on average is the model trained on all 531 CAMELS basins, it is not the case that the model trained on all 531 CAMELS basins is better in every basin. Figure 7 shows the number of basins for which models trained on each grouping (size, HUC, attributes cluster) perform statistically better than (green), not statistically different than (orange), or statistically worse than (blue) the regional model trained on all 531 basins. These statistical tests were done using two-sided Wilcoxon signed-rank test over ten repetitions of each model with a significance level of $\alpha = 0.05$. All models perform worse than the full regional model in more basins than they perform better.

We have not found a way to (reliably) predict which model will perform best in any particular basin. It is not possible to use train period or validation period metrics to (reliably) choose the best model in the test period. Additionally, we have tried extensively to construct a separate predictor model that uses catchment attributes and/or hydrological signatures to predict whether one model will perform better or worse than other models in specific basins. We have not been able to construct a model that performs well at this task. Details of these predictability experiments are out-of-scope for this paper, but a relevant example was given by Nearing et al. (2024).

7 Conclusion

The main point that we would like for readers to take from this opinion paper is that training LSTMs for rainfall-runoff modeling requires using data from many basins. We have seen a number of papers that train large ML models (LSTMs or similar) on very small data sets, and many of these papers then go on to test some type of adaptations that seems to offer improvement. Of course, it is trivial (but most likely uninteresting) to beat improperly trained models. It *would* be interesting to show that adding physics to a well-trained ML model adds information – so far, to our knowledge, all attempts to add physics to (properly trained) streamflow LSTMs in hydrology have produced lower-performing models.

Whatever goal a researcher might have for training an ML-based rainfall-runoff model, there is no reason not to train the model with a large-sample data set. There is enough publicly available streamflow data that there should never be an excuse not to use at least hundreds of basins for training. This is true even if the focus of a particular study is on one or a small number of watersheds.

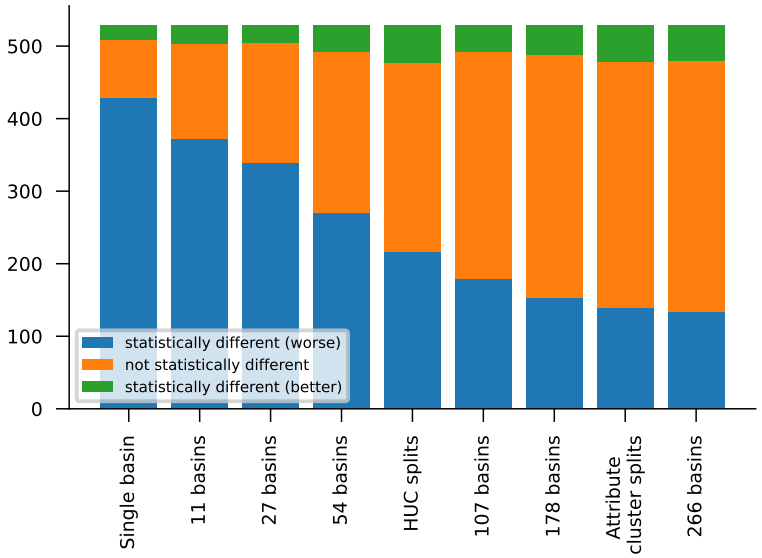

Figure 7. Counts of basins for which models trained on each grouping (sizes, HUC, attributes cluster) perform statistically better than (green), not statistically different than (orange), or statistically worse than (blue) the regional model trained on all 531 basins. Significance was assessed using a two-sided Wilcoxon signed-rank test over ten repetitions of each model with a significance level of $\alpha = 0.05$.

Code and data availability. The run directories of all experiments, including model weights, simulations and pre-computed metrics are available at https://zenodo.org/records/11247607. The code that was used for analysing all experiments and to create all figures, based on the run directories, can be found at https://github.com/kratzert/never-paper. We used the open source Python package NeuralHydrology (Kratzert et al., 2022) to run all experiments. The forcing and streamflow data as well as the catchment attributes used in this manuscript are from the publicly available CAMELS dataset by Newman et al. (2015) and Addor et al. (2017). The simulations from the two hydrological models that were used to create Fig. 3 are available at https://doi.org/10.4211/hs.474ecc37e7db45baa425cdb4fc1b61e1.

Appendix A: Hyperparameter tuning, training, and testing

All models in this paper were trained using data from the CAMELS data set (Newman et al., 2015; Addor et al., 2017). Building on the community benchmarking experiment proposed by Newman et al. (2017), and used by many LSTM modeling studies (e.g., Kratzert et al., 2019a, b, 2021; Frame et al., 2021, 2022; Klotz et al., 2022; Nearing et al., 2022), we trained and tested models on 531 CAMELS basins using time periods for training (1 October 1999 through 30 September 2008), validation (1 October 1980 through 30 September 1989), and testing (1 October 1989 through 30 September 1999). All models were trained and evaluated using NeuralHydrology v.1.3.0 (Kratzert et al., 2022) with an NSE loss function. All LSTMs consist of a single layer LSTM with a linear head layer.

A1 Regional LSTM

The regional LSTM uses hyperparameters from Kratzert et al. (2021). The most important hyperparameters are:

Hidden size 256.

Dropout 40% dropout in the linear output layer.

Optimizer Adam.

Number of epochs 30.

Learning rate Initial learning rate $1e-3$, reduced to $5e-4$ at epoch 20, further reduced to $1e-4$ at epoch 25.

Sequence length 365.

Loss function Adapted NSE loss, see Kratzert et al. (2019b).

After training, we picked the weights from the epoch with the highest validation metric (median NSE across all basins) and evaluated the model with these weights on test period data from all 531 CAMELS catchments. Validation curves for all ten regional models over 30 training epochs are shown in Figure A1

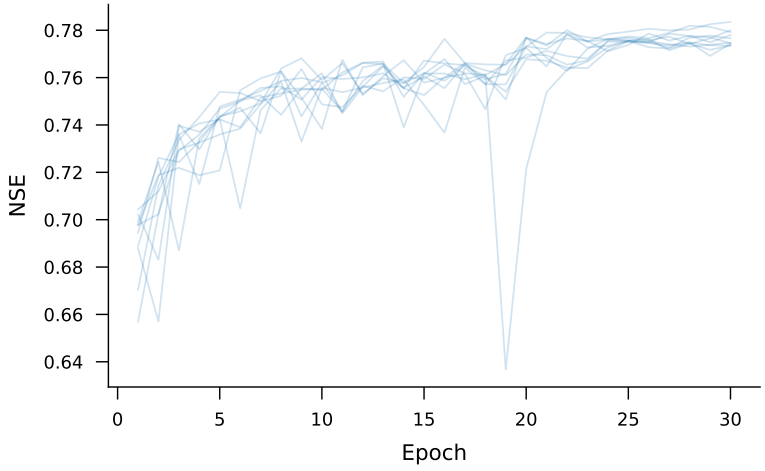

Figure A1. Validation scores of all ten repetitions of the regional model over training epochs.

A2 Single-basin LSTMs

Single-basin LSTMs were trained for each basin individually, and use the same basic architecture as described in Section A1: a single layer LSTM followed by a linear head layer. Hyperparameters were tuned specifically for ech basin for the experiments in this paper using the following two-step procedure. Both steps were done with a grid search:

First step: We used 3 repetitions of each hyperparameter setting for each basin ($n = 531$) with different random seeds for initializing the weights. All models were run for 100 epochs using a the Adam optimizer with a learning rate of $5e-3$ and a batch size of 256. During training, the model was validated after every 4 epochs on validation period data. Hyperparameters were chosen using the model settings with the highest median NSE scores over the 3 repetitions in any validation epoch.

Hidden size (8, 16, 32)

Dropout rate on the head layer (0.0, 0.2, 0.4, 0.5)

For all models, we used the same sequence length ($n = 365$) as for the regional model.

Second step: Using the hyperparameters chosen from the first step, we tuned the learning rate and batch size in a similar way, maximizing over the median NSE over 3 model repetitions:

Learning rate $(5e-3, 1e-3, 5e-4, 1e-4)$

Batch size $(128, 256, 512)$

For each basin separately, we picked model weights from the best validation epoch of the model with the highest NSE score over all validation epochs from all models in each basin.

Final training and evaluation: Given the set of per-basin optimized parameters, we trained ten models per basin, each with a distinct random seed. All statistics reported in this paper for all models are from test period data, except where otherwise noted.

A3 Random size basin splits

Models reported in Fig. 6 were tuned (hyperparameters chosen), trained, and tested in a way that is similar to the single-basin LSTMs described in Appendix A2. For random basins splits we divided the 531 CAMELS basins into random sets without replacement using 6 different sizes of splits that were chosen by (approximately) dividing the full 531 basin group into [50, 20, 10, 5, 3, and 2] basin groups. An example of one of these random splits with 5 groups (approximately 107 basins per split) is shown in Fig. A2.

Choosing hyperparameters was done as described in Appendix A2, except that for these splits we did not use 3 random repetitions and only trained up to 30 epochs to reduce computational expense. We also expanded the hyperparameter search slightly due to our experience training larger models – the hyperparameter ranges for the two grid search stages were:

First step:

Hidden size (8, 16, 64, 128, 256)

Dropout rate on the head layer (0.0, 0.2, 0.4, 0.5)

Second step: In the second stage, we used a batch size of 256 always (as with the regional model), and tuned over the following multi-stage learning rates, where the index is the epoch on which the learning rate switched to the listed value:

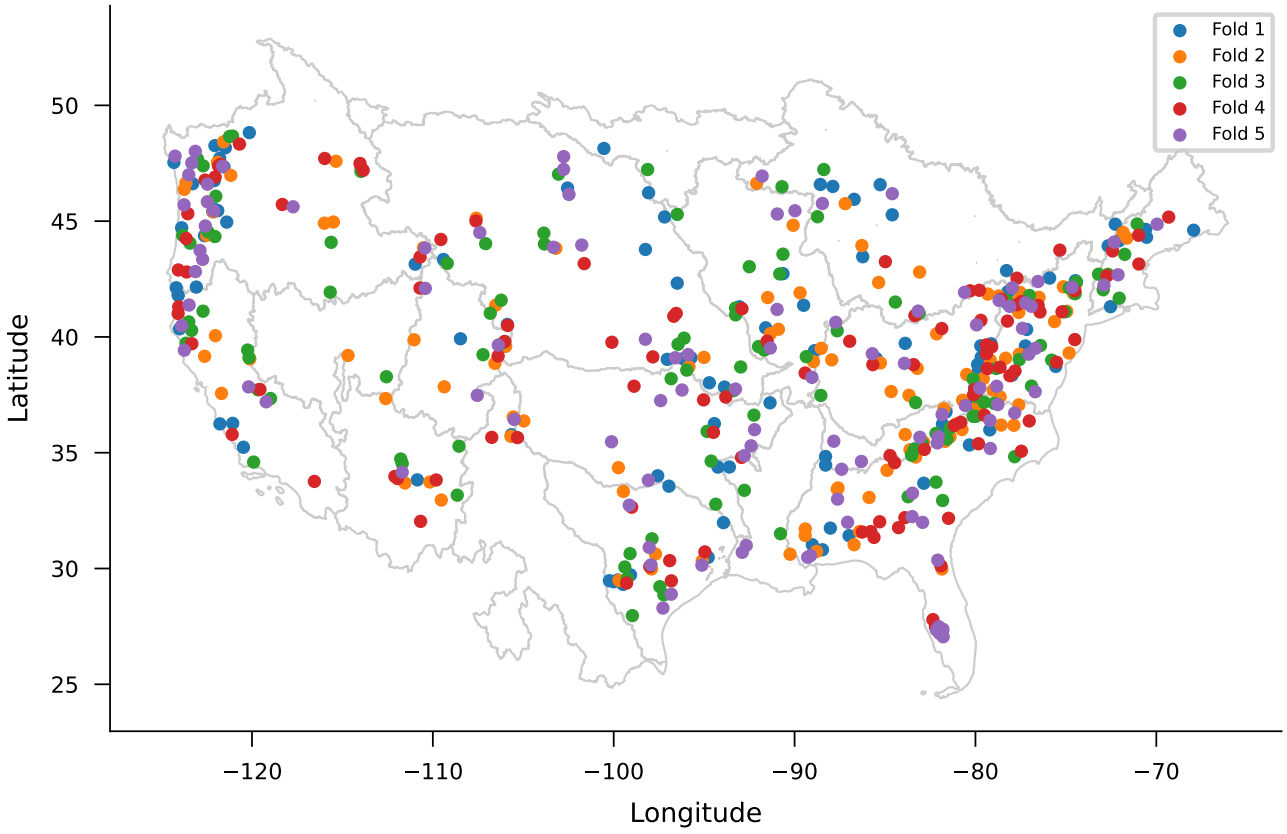

Figure A2. Basin location of random split with 5 groups of approximately 107 basins per group.

- 0: 5e-3, 10: 1e-3, 25: 5e-4
- 0: 1e-3, 10: 5e-4, 25: 1e-4
- 0: 5e-4, 10: 1e-4, 25: 5e-5
- 0: 1e-4, 10: 5e-5, 25: 1e-5

A4 Hydrological unit code splits

The orange curve in Fig. 6 shows median NSE scores over CAMELS basins that result from LSTM models trained on basin groups defined by USGS HUCs. Figure A3 shows the locations of the 531 CAMELS basins by HUC region. The set of 531 basins was divided according to these geographical regions and a separate model was trained on all basins from each. Hyperparameter tuning was done as described in Appendix A3. Testing was done as described in Appendix A2.

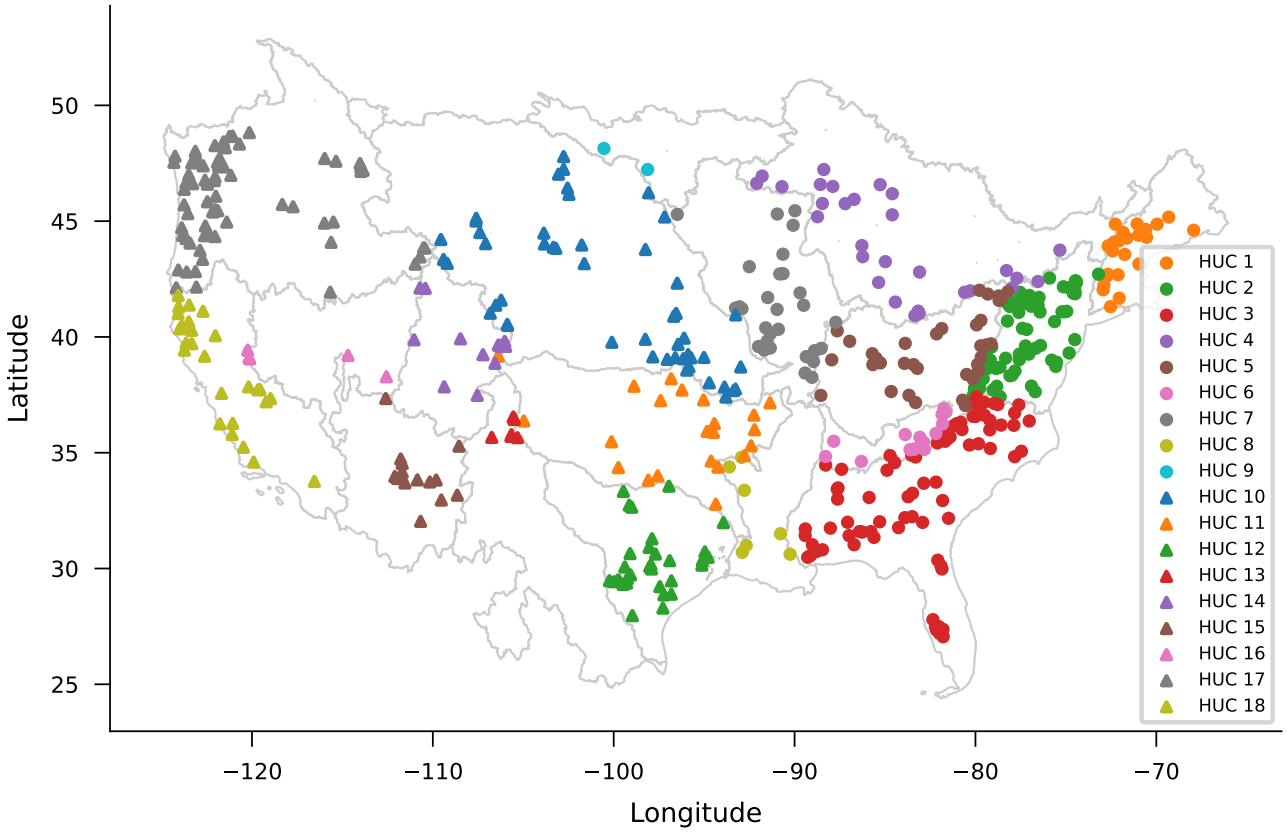

Figure A3. Spatial location of basins split by USGS hydrological unit code 02.

A5 Attribute clusters splits

The green curve in Fig, 6 shows median NSE scores over CAMELS basins that result from LSTM models trained on basin groups defined by k-means clustering based on static catchment attributes. The catchment attributes used for clustering are described in Table A1. These are almost the same attributes that were used by Kratzert et al. (2019b) but without carbonate rocks fraction and the seasonality of precipitation (the former is often zero and the latter is categorical, both of which make clustering slightly more difficult).

We performed k-means clustering on these 25 basin attributes (all attributes were normalized), using 300 iterations and 10 random initializations. Using a maximin criterion on silhouette scores for between 3 and 100 clusters, we chose to divide basins into six groups with sizes of 83, 195, 67, 61, 66, and 59 basins. These clusters are mapped in Fig. A4.

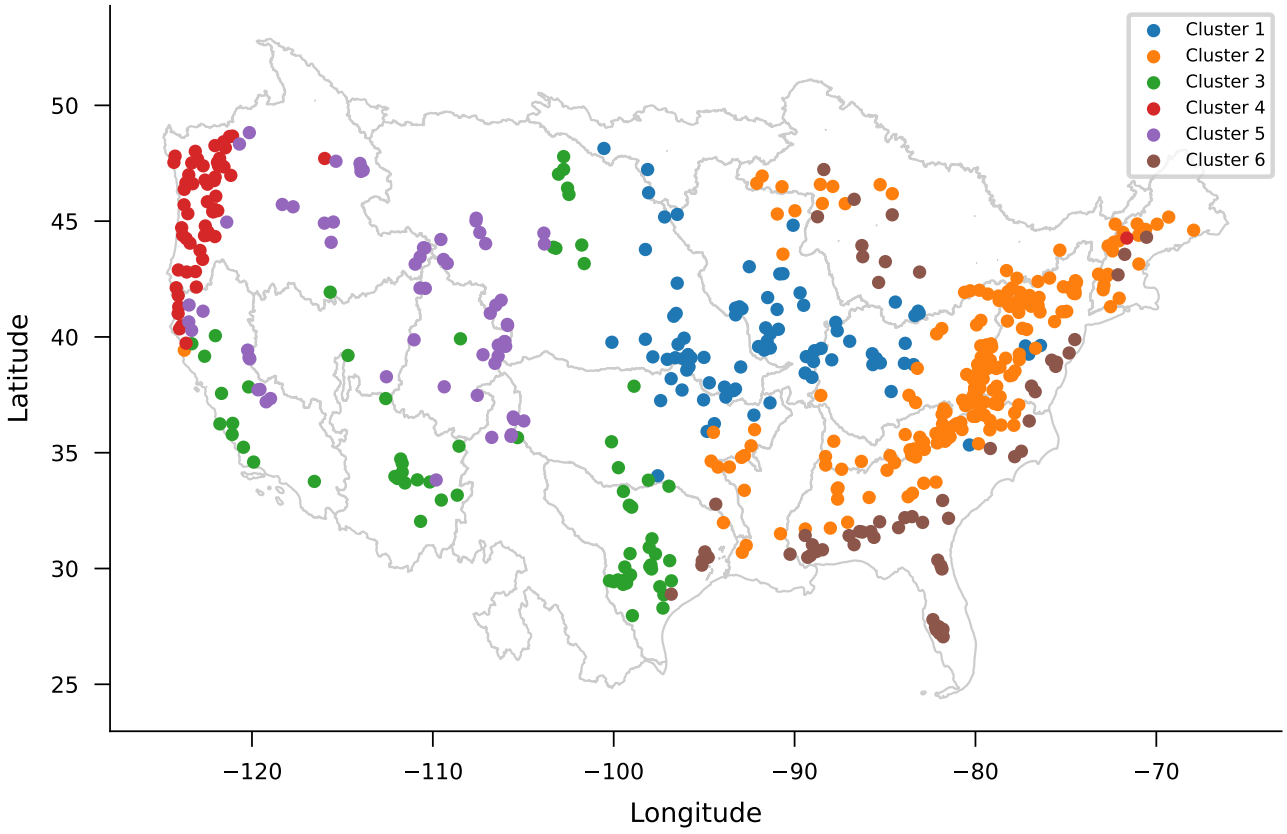

Figure A4. Spatial location of basins split by k-means clustering on the basin attributes.

Appendix B: Regional vs. Single-Basin Model Comparison

Figure 3 illustrates a difference between how LSTM models and conceptual models behave when trained regionally vs. locally (on single basins). Specifically, the LSTM performs best when trained regionally while traditional models perform best when trained locally.

Figure B1 illustrates the same comparison, but where the LSTM NSE CDFs in subpanel (b) only consider the same 489 CAMELS basins that are used in the mHM and VIC NSE CDFs in subpanel (a). This is a more direct comparison than what is shown in Fig. 3, however the results of the comparison are qualitatively identical.

Table A1. Table of catchment attributes used in this experiments. Description taken from the data set Addor et al. (2017).

p_mean	Mean daily precipitation.
pet_mean	Mean daily potential evapotranspiration.
aridity	Ratio of mean PET to mean precipitation.
frac_snow_daily	Fraction of precipitation falling on days with temperatures below $0^\circ C$.
high_prec_freq	Frequency of high precipitation days (>= 5 times mean daily precipitation).
high_prec_dur	Average duration of high precipitation events (number of consecutive days with >= 5 times mean daily precipitation).
low_prec_freq	Frequency of dry days (< 1 mm/day).
low_prec_dur	Average duration of dry periods (number of consecutive days with precipitation < 1 mm/day).
elev_mean	Catchment mean elevation.
slope_mean	Catchment mean slope.
area_gages2	Catchment area.
forest_frac	Forest fraction.
lai_max	Maximum monthly mean of leaf area index.
lai_diff	Difference between the max. and min. mean of the leaf area index.
gvf_max	Maximum monthly mean of green vegetation fraction.
gvf_diff	Difference between the maximum and minimum monthly mean of the green vegetation fraction.
soil_depth_pelletier	Depth to bedrock (maximum 50m).
soil_depth_statsgo	Soil depth (maximum 1.5m).
soil_porosity	Volumetric porosity.
soil_conductivity	Saturated hydraulic conductivity.
max_water_content	Maximum water content of the soil.
sand_frac	Fraction of sand in the soil.
silt_frac	Fraction of silt in the soil.
clay_frac	Fraction of clay in the soil.
geol_permeability	Surface permeability (log10).

Appendix C: Theoretical prediction limit

Figure 4 shows the theoretical maximum prediction limits for regional and single-basin LSTMs. To understand how those limits were derived, it is important to understand how the output of the LSTM layer is computed and how this output translates into the model prediction.

The output of the LSTM layer, $\mathbf{h}_t$, is computed according to the following equation:

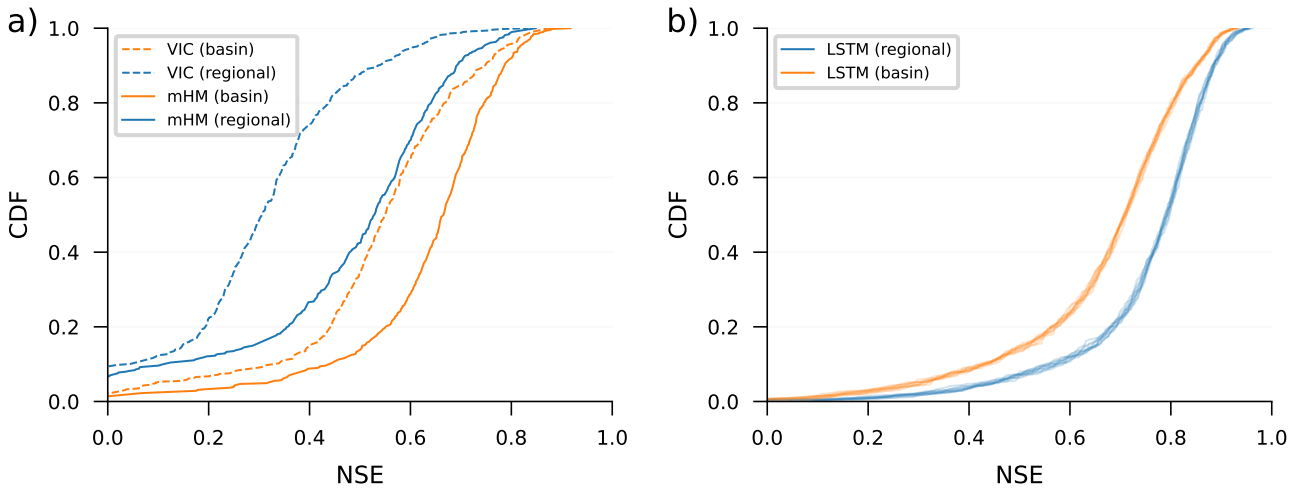

Figure B1. This figure is identical to Fig. 3 except that subpanel (b) uses all 531 CAMELS basins that Newman et al. (2017) recommended using for model benchmarking, and which were used in all other experiments reported in this study.

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t), \tag{C1}$$

where $\mathbf{o}_t$ is the output gate at time step t, $\tanh()$ is the hyperbolic tangent function and $\mathbf{c}_t$ the LSTM cell state of time step t. The output gate is computed according to the following equation:

$$\mathbf{o}_t = \text{sigmoid}(\mathbf{W}\mathbf{x}_t + \mathbf{V}\mathbf{h}_{t-1} + \mathbf{b}), \tag{C2}$$

where $\text{sigmoid}()$ is the logistic function, $\mathbf{x}_t$ are the input features of time step t, $\mathbf{h}_{t-1}$ the hidden state (or LSTM output) from the previous time step $t-1$, and $\mathbf{W}$, $\mathbf{V}$, and $\mathbf{b}$ are learnable model parameters.

Finally, in the case of our model architecture, the output of the LSTM is passed through a linear layer that maps from the hidden size of the LSTM to 1, the models prediction. More formally, the model prediction $\hat{y}_t$ at time step t is computed according to the following equation:

$$\hat{y}_t = \mathbf{W}\mathbf{h}_t + b, \tag{C3}$$

where $\mathbf{W}$ and $\mathbf{b}$ are another set of learnable model parameters, specific to this linear layer. Since our model maps to a single output value, $\mathbf{W}$ is of shape [hidden size, 1]. With $\mathbf{h}_t$ of shape [hidden size], we can write Eq. C3 as:

$$\hat{y}_t = b + \sum_{i=0}^{n} \mathbf{W}_i * \mathbf{h}_{i,t} \tag{C4}$$

Knowing that each element of $\mathbf{h}_t$ is in $(-1, 1)$ and that $\mathbf{W}$ and b are fixed after training, the maximum possible value a trained LSTM (of our architecture) can predict can be computed by

$$upperlimit = b + \sum_{i=0}^{n} \mathrm{abs}(\mathbf{W}_i), \tag{C5}$$

where $\mathrm{abs}()$ is the function that returns the absolute value. Note that this value is in the space of training labels and if the labels were normalized for training, the $upperlimit$ needs to be re-transformed into discharge space to get the $upperlimit$ in e.g. mm per day.

Appendix D: Example Hydrographs

Figure 5 shows an example of simulated vs. observed hydrographs for one of the 531 CAMELS basins. Basin 13011900 shown in Fig. 5 has highly is on Lava Creek in Wyoming and has a drainage area of 837 square kilometers. This basin has highly seasonal flow patterns.

Figure D1 and Fig. D2 show similar hydrographs for other basins with different hydrological behaviors. One of these is a small basin on the Salt Creek in Kansas, and has a flashy flow pattern. The other is on the Sauk River in Washington and has non-flashy behavior, but with significant (non-seasonal) peak flows.

Additionally, the code and data repositories released with this paper contain everything necessary to plot simulated and observed hydrographs for any of the 531 CAMELS basins.

Appendix E: Other hydrological metrics

There are a large number of metrics that hydrologists use to assess hydrograph simulations Gupta et al. (2012); Gauch et al. (2023). Several of these metrics are described in Table E1, including bias, correlation, Nash–Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970), Kling–Gupta Efficiency (KGE) (Gupta et al., 2009), and metrics related to hydrograph peaks. Fig. E1 shows differences between regional and single-basin models for these metrics, similar to the NSE comparison shown in Fig. 3.

The takeaway from this figure is that the main message of this paper (do not train a rainfall-runoff LSTM on data from a single-basin) holds regardless of the metric(s) that we focus on. The skill differences are in correlation-based metrics (NSE, KGE, and Pearson-R) as well as variance-based metrics (Alpha-NSE). The latter is an artifact of what we saw in Fig. 4, that training on more, diverse data improves the ability of the model to predict high flows. Figure E1 shows only small improvements in the timing and capture of hydrograph peaks (the Missed-Peaks metric measures whether a peak in the hydrograph was captured at all, not whether the magnitude of the peak was predicted accurately). And we see little or no difference in the two

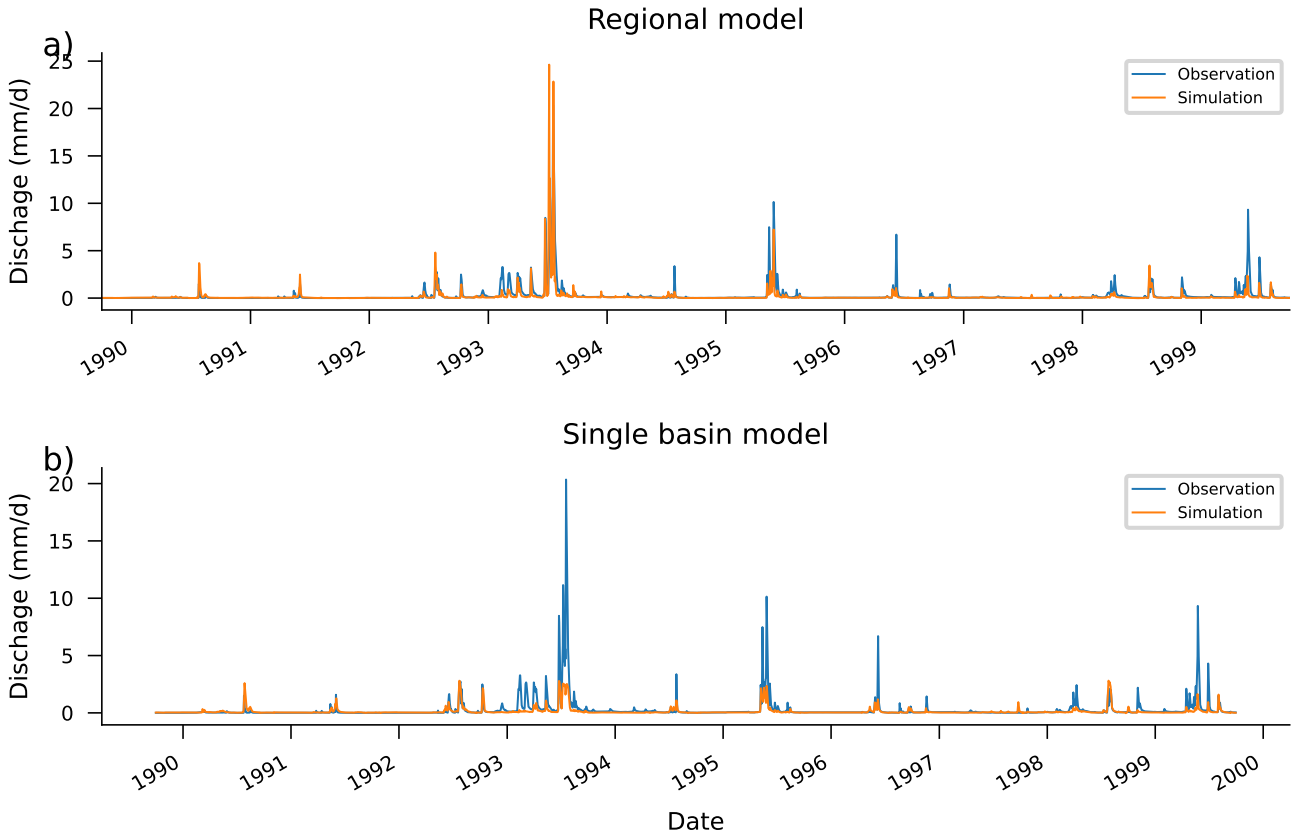

Figure D1. Observed and simulated hydrographs from the test period (1989 - 1999) in a particular basin (06876700). This gauge is on the Salt Creek in Kansas with a drainage area of 1,052 square kilometers, and represents a relatively flashy basin.

bias metrics (Beta-NSE, Beta-KGE), meaning that improvements to catchment-specific mean discharge is not strongly affected by using training data from multiple catchments (i.e., we don't strongly bias one type of catchment by using other types of catchments in training).

Author contributions. FK had the initial idea for this paper. FK and GN set up all experiments. All authors contributed to the analysis of the experiment results. All authors contributed to the writing process with GN doing the majority of the writing.

Competing interests. The authors declare no competing interests.

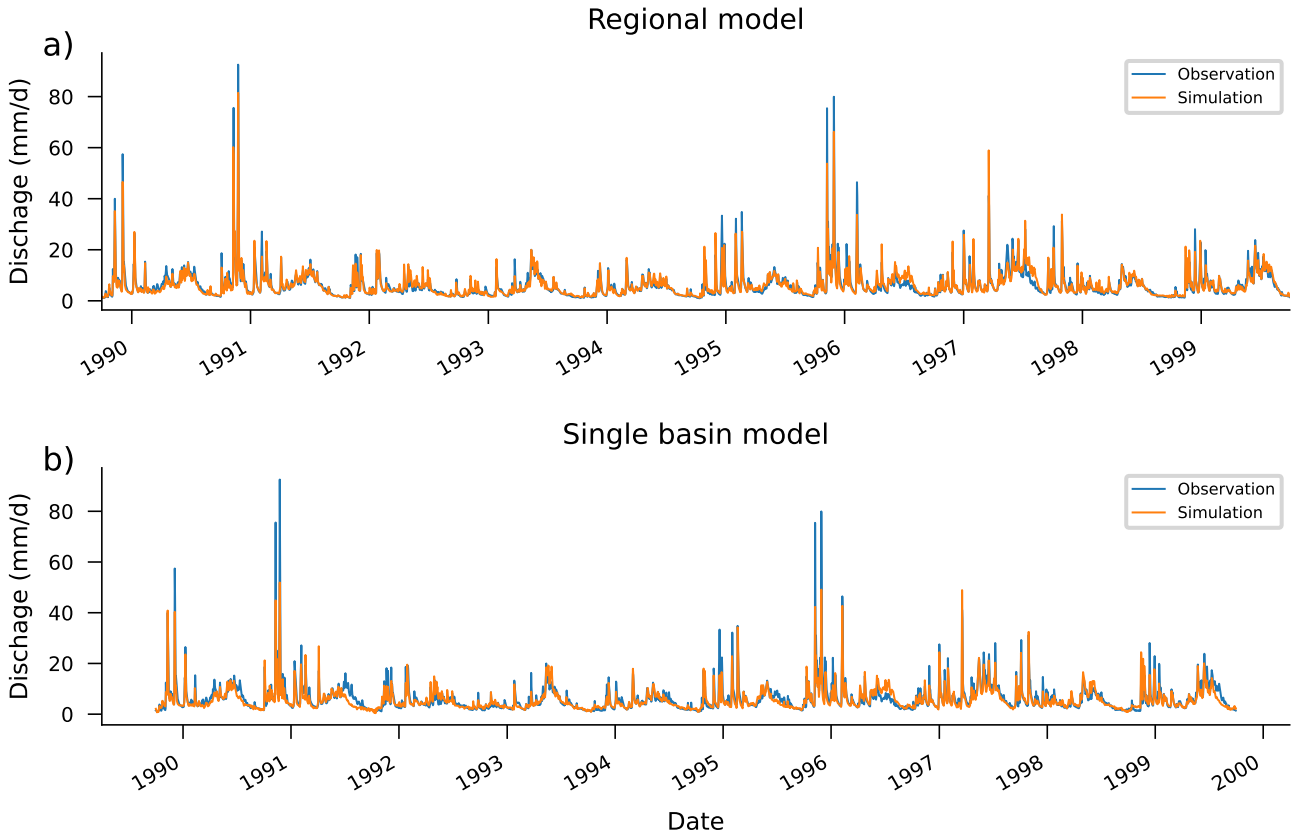

Figure D2. Observed and simulated hydrographs from the test period (1989 - 1999) in a particular basin (12189500). This gauge is on the Sauk River in Washington with a drainage area of 1,849 square kilometers, and represents a basin that with high peaks that are not dominated by a seasonal flow pattern.

References

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

BAFG: https://www.bafg.de/GRDC, "The Global Runoff Data Centre, 56068 Koblenz, Germany".

Beck, H. E., van Dijk, A. I., De Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., and Bruijnzeel, L. A.: Global-scale regionalization of hydrologic model parameters, Water Resources Research, 52, 3599–3622, 2016.

Frame, J. M., Kratzert, F., Raney, A., Rahman, M., Salas, F. R., and Nearing, G. S.: Post-processing the national water model with long short-term memory networks for streamflow predictions and model diagnostics, JAWRA Journal of the American Water Resources Association, 57, 885–905, 2021.

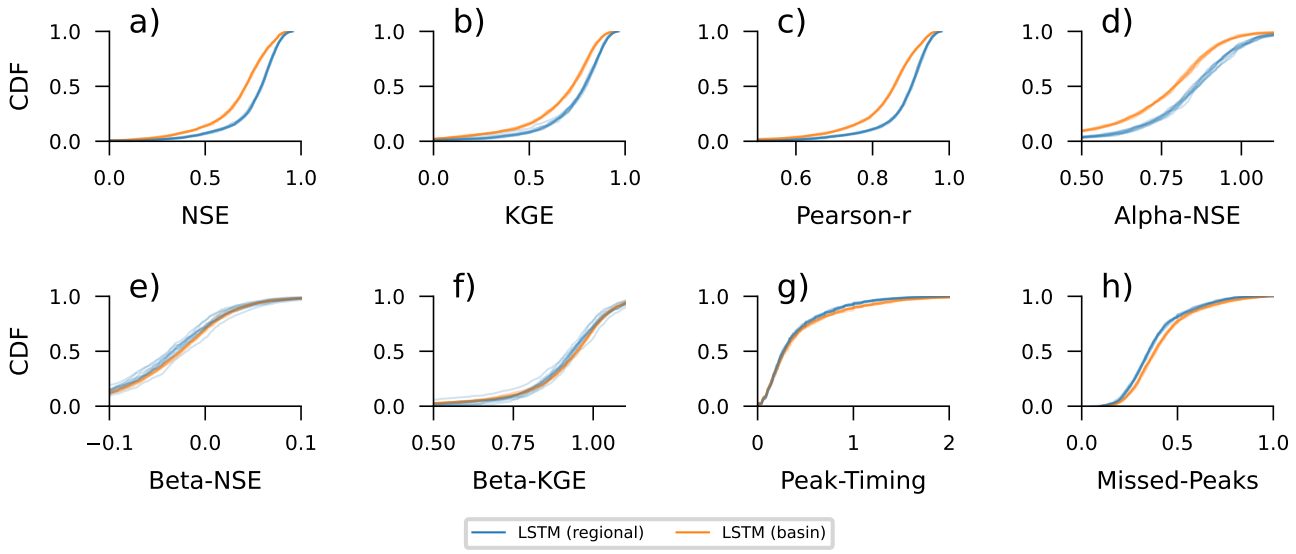

Figure E1. Comparisons between CDFs over 531 CAMELS basins of regional vs. single-basin LSTMs. This is similar to Figure **??**, but for the hydrograph metrics listed in Table E1.

Table E1. A selection of standard hydrograph evaluation metrics.

Name	Description	Reference
NSE^{i}	Nash–Sutcliffe efficiency	Eq. 3 in Nash and Sutcliffe (1970)
KGE^{i}	Kling–Gupta efficiency	Eq. 9 in Gupta et al. (2009)
Pearson-r^{ii}	Pearson correlation	Pearson (1895)
Alpha-NSE^{iii}	Ratio of standard deviations of observed and simulated flow	From Eq. 4 in Gupta et al. (2009)
Beta-NSE^{iv}	Bias scaled by standard deviation of observations	From Eq. 4 in Gupta et al. (2009)
Beta-KGE^{v}	Bias ratio: ratio of mean simulated and mean observed flow	From Eq. 10 in Gupta et al. (2009)
Peak-Timing^{iv}	Mean time lag between observed and simulated peaks	Appendix A in Gauch et al. (2021)
Missed-Peaks^{viii}	Fraction of hydrograph peaks that were missed	Nearing et al. (2022)

^{i}: $(-\infty, 1]$, *values closer to one are desirable.*
^{ii}: $[-1, 1]$, *values closer to one are desirable.*
^{iii}: $(0, \infty)$, *values close to one are desirable.*
^{iv}: $(-\infty, \infty)$, *values close to zero are desirable.*
^{v}: $(-\infty, \infty)$, *values close to one are desirable.*
^{vi}: $(0, 1)$, *values close to zero are desirable.*

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S.: Deep learning rainfall–runoff predictions of extreme events, Hydrology and Earth System Sciences, 26, 3377–3392, 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, Hydrology and Earth System Sciences, 25, 2045–2062, 2021.

Gauch, M., Kratzert, F., Gilon, O., Gupta, H., Mai, J., Nearing, G., Tolson, B., Hochreiter, S., and Klotz, D.: In Defense of Metrics: Metrics Sufficiently Encode Typical Human Preferences Regarding Hydrological Model Performance, Water Resources Research, 59, e2022WR033 918, 2023.

Guo, Y., Zhang, Y., Zhang, L., and Wang, Z.: Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review, Wiley Interdisciplinary Reviews: Water, 8, e1487, 2021.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of hydrology, 377, 80–91, 2009.

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., and Ye, M.: Towards a comprehensive assessment of model structural adequacy, Water Resources Research, 48, 2012.

Hrachowitz, M. and Clark, M. P.: HESS Opinions: The complementary merits of competing modelling philosophies in hydrology, Hydrology and Earth System Sciences, 21, 3953–3973, 2017.

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, Hydrology and Earth System Sciences, 26, 1673–1693, 2022.

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

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward improved predictions in ungauged basins: Exploiting the power of machine learning, Water Resources Research, 55, 11 344–11 354, 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, Hydrology and Earth System Sciences, 23, 5089–5110, 2019b.

Kratzert, F., Klotz, D., Hochreiter, S., and Nearing, G. S.: A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling, Hydrology and Earth System Sciences, 25, 2685–2703, 2021.

Kratzert, F., Gauch, M., Nearing, G., and Klotz, D.: NeuralHydrology—A Python library for Deep Learning research in hydrology, Journal of Open Source Software, 7, 4050, 2022.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., et al.: Caravan-A global community dataset for large-sample hydrology, Scientific Data, 10, 61, 2023.

Ma, K., Feng, D., Lawson, K., Tsai, W.-P., Liang, C., Huang, X., Sharma, A., and Shen, C.: Transferring hydrologic data across continents–leveraging data-rich regions to improve hydrologic prediction in data-sparse regions, Water Resources Research, 57, e2020WR028 600, 2021.

Mizukami, N., Clark, M. P., Newman, A. J., Wood, A. W., Gutmann, E. D., Nijssen, B., Rakovec, O., and Samaniego, L.: Towards seamless large-domain parameter estimation for hydrologic models, Water Resources Research, 53, 8020–8040, 2017.

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

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I—A discussion of principles, Journal of hydrology, 10, 282–290, 1970.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., et al.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, 2024.

Nearing, G. S., Pelissier, C. S., Kratzert, F., Klotz, D., Gupta, H. V., Frame, J. M., and Sampson, A. K.: Physically informed machine learning for hydrological modeling under climate nonstationarity, UMBC Faculty Collection, 2019.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What role does hydrological science play in the age of machine learning?, Water Resources Research, 57, e2020WR028 091, 2021.

Nearing, G. S., Klotz, D., Frame, J. M., Gauch, M., Gilon, O., Kratzert, F., Sampson, A. K., Shalev, G., and Nevo, S.: Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks, Hydrology and Earth System Sciences, 26, 5493–5513, 2022.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J., et al.: 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, 209–223, 2015.

Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., and Nearing, G.: Benchmarking of a physically based hydrologic model, Journal of Hydrometeorology, 18, 2215–2225, 2017.

Pearson, K.: Note on Regression and Inheritance in the Case of Two Parents, Proceedings of the Royal Society of London Series I, 58, 240–242, 1895.

Rakovec, O., Mizukami, N., Kumar, R., Newman, A. J., Thober, S., Wood, A. W., Clark, M. P., and Samaniego, L.: Diagnostic evaluation of large-domain hydrologic models calibrated across the contiguous United States, Journal of Geophysical Research: Atmospheres, 124, 13 991–14 007, 2019.

Samaniego, L., Kumar, R., and Attinger, S.: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, Water Resources Research, 46, 2010.

Sutton, R.: Incomplete Ideas (blog) (Last accessed 19.05.2024), http://www.incompleteideas.net/IncIdeas/BitterLesson.html, 2019.