https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

A novel hybrid fine-tuning method for supercharging deep learning model development for hydrological prediction

M.S. Jahangir[1,2], J. Quilty[3], C. Shen[4], A. Scott[5], S. Steinschneider[2], J. Adamowski[1]

[1]Department of Bioresource Engineering, McGill University, Montreal, H9X 3V9, Canada
[2]Department of Biological and Environmental Engineering, Cornell University, Ithaca, 14853, United States
[3]Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, N2L3G1, Canada
[4]Department of Civil and Environmental Engineering, Pennsylvania State University, College town, 16802, United States
[5]Department of Mechanical Engineering, University of Waterloo, Waterloo, N2L3G1, Canada
Correspondence to: Mohammad Sina Jahangir (mohammad.jahangir@mcgill.ca)

Abstract. This study proposes a novel hybrid method that substantially accelerates and improves deep learning (DL) model development for streamflow prediction. The method leverages a combination of a long short-term memory (LSTM) network and random forests. A hybrid encoder-decoder model is designed, where a pre-trained LSTM is utilized as an encoder to extract temporal features from the input data. Subsequently, the random forest decoder processes the encoded information to make streamflow predictions. Our method was tested on 421 catchments in the continental United States and 324 in Germany, both selected from two CAMELS datasets. The hybrid method has several benefits. First, it is much more efficient and robust than training LSTMs on each catchment individually (~14x faster). Second, it is much less computationally expensive than LSTM fine-tuning (i.e., feasible on a CPU-based workstation). Third, it achieves superior accuracy compared to a catchment-wise training strategy (e.g., 9.2% improvement in the median in Nash-Sutcliffe Efficiency (NSE)), shows competitive performance compared to regional LSTM models when trained with fewer data, and through fine-tuning, improves regional LSTM performance in out-of-training samples by 13.13% (median NSE). To our knowledge, this is the first decision-tree model integrated within a DL workflow to enhance fine-tuning efficiency of pre-trained models in new locations. This hybrid approach holds significant promise for future applications in hydrological modeling, particularly considering the imminent rise of geospatial foundation models in hydrology that will rely on transfer learning techniques for effective deployment.

1 Introduction

Accurate hydrological prediction is necessary for reliable water resource management. The efficacy of deep learning (DL) models for hydrological prediction, especially long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), is well established. Previous work has shown that LSTM models outperform conceptual hydrological models in streamflow prediction (e.g., Kratzert et al., 2018). Unlike process-based models, which rely on differential and linear equations to model underlying physical processes (Xie et al., 2021), DL models function as conditional estimators, mapping input variables to target variables without explicitly considering the governing equations of the modeled system (Feng et al., 2022; Jahangir et al., 2023).

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

The success of LSTM models in hydrological prediction is primarily attributed to two key factors. First, LSTM cells effectively represent multi-timescale temporal dynamics through feedback connections (Elman, 1990), making them well-suited for processing hydrological series (Jahangir et al., 2023). Second, the encoder-decoder (ED) structure of the model – in which past hydrometeorological data are encoded using the LSTM cell and decoded using linear (Kratzert et al., 2018) or non-linear (Yin et al., 2021) layers - can effectively process the non-linear relationships between input and output series (Cui et al., 2022; Jahangir and Quilty, 2024).

Previous studies have emphasized the benefits of training regional LSTM models for hydrological prediction (e.g., Feng et al., 2020; Kratzert et al., 2024, 2019). Rather than developing a separate model for each catchment, a single LSTM is trained on all catchments of interest, allowing it to generalize across diverse hydrologic conditions. When adequately trained on data from diverse catchments, these models can accurately predict a broad spectrum of hydrological processes (Kratzert et al., 2024). While diverging from traditional, catchment-wise hydrologic modeling, this top-down approach aligns with DL model development practices in other fields (Brown et al., 2020).

Although regional LSTMs outperform conceptual and catchment-wise strategies regarding prediction accuracy, this increased effectiveness comes with costs. High data requirements, computational demands of model optimization, and the need for specialized expertise create barriers, particularly for practitioners without DL training or access to high-performance computing (HPC). Lacking large-scale GPUs, cloud infrastructure, or technical expertise, many struggle to develop and deploy DL models tailored to local hydrological conditions.

The high computational burden of model training also limits thorough hyperparameter optimization and the development of multiple regional DL models for sensitivity analysis, scenario evaluation, or improved prediction. For example, while ensemble modeling consistently excels in forecasting (Bojer and Meldgaard, 2021), it remains impractical for most practitioners due to resource constraints. Even universities often lack the computing power needed for the most advanced models forwarded by major technology companies, widening the gap in DL adoption.

As a result, despite their demonstrated efficacy, regional LSTM models see limited real-world application, with many practitioners defaulting to legacy models and catchment-wise training strategies (Kratzert et al., 2024), either by choice or by necessity. Water managers typically prioritize localized predictions over broad-scale generalizable, further challenging DL integration into operational hydrology. This disparity underscores a critical challenge: while DL-based hydrologic models continue to advance, their practical use remains constrained by data availability, expertise gaps, and computational limitations, reinforcing the need for more accessible, efficient solutions.

One approach to tackle this problem is transfer learning (TL; Pan and Yang, 2010). The general concept in TL is to speed up the learning procedure by extracting meaningful representations from the relevant source domain(s) and transferring this knowledge to the target domain. By leveraging existing knowledge rather than starting from scratch, TL reduces a DL model's dependency on large training datasets (Zhao et al., 2024). It also can enhance model generalization, mainly when the target domain data is limited (e.g., streamflow in poorly gauged regions, water quality, edge-of-field runoff).

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

Transfer learning can encompass both zero-shot prediction tasks, such as prediction in ungauged basins (PUB; Arsenault et al., 2023; Fang et al., 2024; Nearing et al., 2024), and few-shot learning, where the source model is modified and fine-tuned for a new task or dataset (Hu et al., 2016; Niu et al., 2020; Zhao et al., 2024). The latter is beneficial when applying DL hydrologic models in poorly gauged regions or for individual sites not included in the source training set. This was first demonstrated by Ma et al. (2021), which showed that an LSTM trained regionally on catchments within the continental United States (CONUS) could be effectively transferred to and fine-tuned for regions in China, Chile, and the United Kingdom. Their work showed that performance generally increased with additional data for both initial training and fine-tuning, and that the source data used for regional model training substantially impacted the best fine-tuning strategy. Khoshkalam et al. (2023) showed the effectiveness of DL-based TL for hydrological prediction in snow-dominated catchments by fine-tuning a regionally trained LSTM for CONUS on data from watersheds in southern Quebec, Canada. Their results suggested that applying TL with consistent hydrometeorological data across the source and target datasets resulted in more reliable and satisfactory results. TL efficacy has also been shown for Alpine regions, where a regional LSTM model trained on data from CONUS was fine-tuned for four catchments around the Tibetan Plateau (Yao et al., 2023). Recently, Khoshkalam et al. (2025) demonstrated the benefits of data integration (using past lags of streamflow) and catchment clustering for TL.

Although the current study shares similarities with prior research on DL-based TL for hydrological prediction in gauged catchments, it approaches the problem differently. Here, the focus is on exploring a novel method to reduce TL's computational costs, improve accessibility, and increase regional model performance. This study contributes two key TL advances in DL-based hydrologic modeling:

(a) Proposing a novel hybrid TL approach that integrates a regional LSTM with a random forest (RF) model for efficient TL and fine-tuning, particularly in cases where limited computational resources or insufficient data make training regional models across hundreds of catchments impractical; and

(b) Evaluating the potential of the hybrid approach as a complementary tool to large sample regional hydrological prediction.

This research analyzes the hybrid approach and two other fine-tuning strategies for large sample hydrological prediction and investigates their effectiveness on two benchmark datasets: CAMELS-US (Addor et al., 2017) and CAMELS-DE (Loritz et al., 2024). Our results show that the hybrid approach is an efficient and practical modeling method for hydrological prediction that is accessible to all water resources practitioners. We emphasize that this work does not advocate abandoning regional DL model development for hydrology. Rather, it tries to develop strategies that hydrologists and water resource practitioners can use to leverage emerging DL-based hydrologic modeling, including geo-foundation models (Xie et al., 2023), for single-catchment applications that are both efficient and accessible.

The rest of the paper is organized as follows: Section 2 provides an overview of the dataset used for model development. Section 3 outlines the hybrid TL development strategy and details the evaluation metrics. Section 4 presents and discusses the results and explains research limitations. Section 5 concludes the paper, and directions for future research are presented.

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

2 Dataset and Case Studies

Two case studies were adopted for model development and benchmarking. In the first case study, we used 421 catchments from CAMELS-US (Addor et al. (2017)). In the second case study, we used 324 catchments from CAMELS-DE (Loritz et al. (2024)). The data and selection criteria for each case study are described below in subsections.

2.1 CAMELS-US

The CAMELS-US dataset encompasses 671 catchments in the CONUS with minimal human interference. These catchments are categorized into 18, 2-digit hydrological units (HUCs) based on the U.S. Geological Survey categorization. Daymet (Thornton et al., 2014) meteorological variables, including precipitation (mm/day), maximum and minimum air temperature (°C), shortwave downward radiation (wat/m^2), and vapor pressure (pa), were used as input.

Two criteria were used to select catchments for model development and evaluation. The first criterion excluded catchments with consecutive missing values in the streamflow records to avoid the need for imputing large data gaps. The second criterion excluded catchments with fewer than 10,000 records, as the study compares TL with both catchment-wise and regional model development approaches. Applying these criteria resulted in 421 catchments for model development and evaluation (see Figure 1(a)). The list of the selected catchments is provided in supplementary information (SI, Table S1). We used specific discharge (mm/day) values taken from Caravan (Kratzert, 2023). The final selected dataset spanned from June 1984 to December 2012. The dataset was divided into three subsets: training (76.5%), validation (8.5%), and testing (15%). The training and validation subsets, comprising the first 85% of the records, covered the period from June 6, 1984, to November 24, 2007. The testing subset, representing the final 15% of the data, covered the period from November 25, 2007, to January 15, 2012. The validation set was used for hyperparameter selection and early stopping (see Section 3 for details). The test set, unseen by the DL models during training and validation, was employed for performance evaluation.

2.2 CAMELS-DE

CAMELS-DE comprises data from 1,555 streamflow catchments across Germany, offering hydrometeorological time-series spanning up to 70 years from January 1951 to December 2020 (median length of 46 years; minimum of 10 years). The dataset contains catchments with areas ranging from 5 to 15,000 km² and contains several attributes, including soil characteristics, land cover, hydrogeologic properties, and information on human influences (Loritz et al., 2024). The dataset also provides simulation results from a regionally trained LSTM network trained to all 1,555 sites, which we used as a benchmark for evaluating the hybrid method (Loritz et al., 2024). For this analysis, we selected 324 catchments with no missing value (see Table S2 and Figure 1(b)) and utilized the same input features as the regional LSTM model from CAMELS-DE: mean precipitation (mm/day), precipitation standard deviation (mm/day), mean radiation (wat/m^2), and mean minimum and maximum temperature (°C). Specific discharge (mm/day) serves as the target variable. The period from October 1, 1970, to

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

December 31, 1999, is used as training, while October 1, 1965, to September 30, 1970, and January 1, 2001, to December 31, 2020, are used for validation and testing, respectively (please see Loritz et al. (2024) for details).

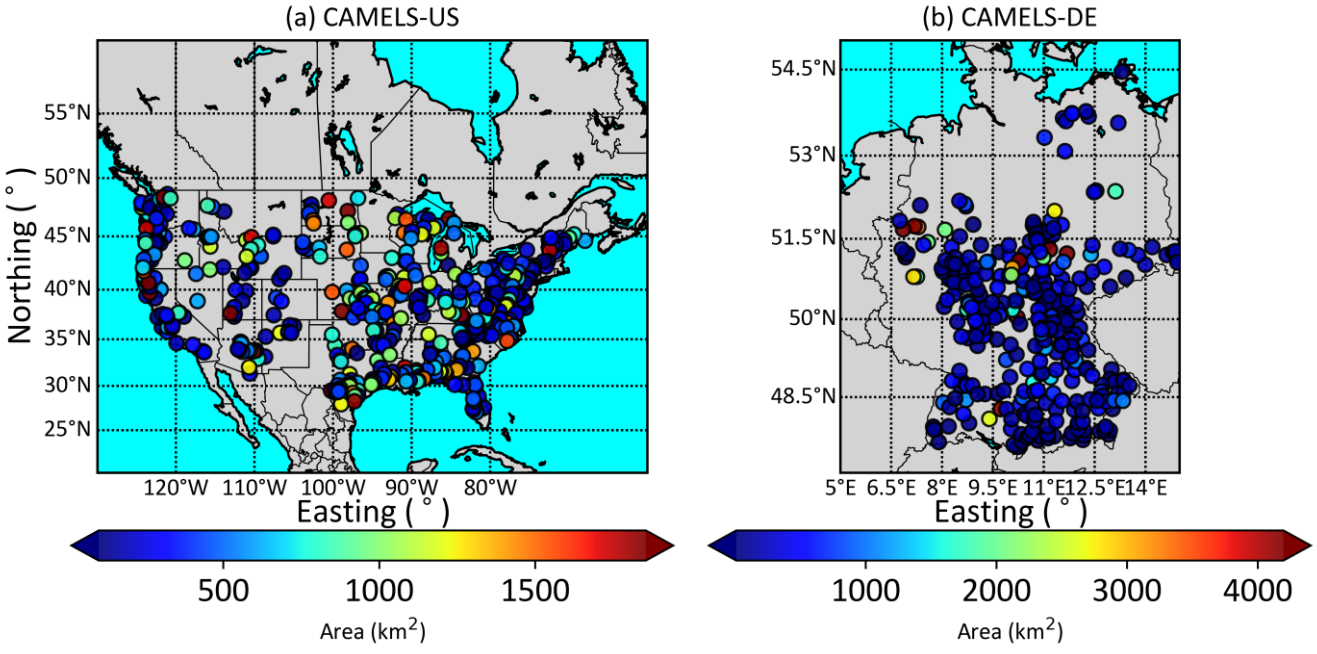


Figure 1: The selected gauges associated with catchments in (a) CAMELS-US (421) and (b) CAMELS-DE (324).

3 Methodology

In this section, we first present the formulation of the LSTM model, followed by the details of implementing the TL methods. Finally, we outline the evaluation metrics used in this study. While LSTM is a well-established approach for hydrological prediction, and its formulation has been extensively detailed in prior literature, we provide it here to establish the foundation for our hybrid TL methodology.

3.1 Long Short-Term Memory Network

LSTM networks, introduced by Hochreiter and Schmidhuber (1997) in the late 1990s, were designed to overcome the challenges of vanishing and exploding gradients faced by recurrent neural networks (RNNs) when processing long data sequences. These gradient issues, which stem from either the rapid decay or exponential growth of gradients during backpropagation, hinder the ability of RNNs to capture long-term dependencies. LSTMs mitigate this problem by incorporating internal memory cells and gating mechanisms, enabling the selective retention or dismissal of information over

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

prolonged timesteps. The formulation of an LSTM cell associated with timestep t (see Figure 2) is as follows (Kratzert et al., 2019):

$$\mathbf{i}^t = \text{sigmoid}(\mathbf{W}_i\mathbf{X}^t + \mathbf{U}_i\mathbf{h}^{t-1} + \mathbf{b}_i) \quad \textbf{Eq 1}$$

$$\mathbf{f}^t = \text{sigmoid}(\mathbf{W}_f\mathbf{X}^t + \mathbf{U}_f\mathbf{h}^{t-1} + \mathbf{b}_f) \quad \textbf{Eq 2}$$

$$\mathbf{g}^t = \tanh(\mathbf{W}_g\mathbf{X}^t + \mathbf{U}_g\mathbf{h}^{t-1} + \mathbf{b}_g) \quad \textbf{Eq 3}$$

$$\mathbf{o}^t = \text{sigmoid}(\mathbf{W}_o\mathbf{X}^t + \mathbf{U}_o\mathbf{h}^{t-1} + \mathbf{b}_o) \quad \textbf{Eq 4}$$

$$\mathbf{c}^t = \mathbf{f}^t \odot \mathbf{c}^{t-1} + \mathbf{i}^t \odot \mathbf{g}^t \quad \textbf{Eq 5}$$

$$\mathbf{h}^t = \mathbf{o}^t \odot \tanh(\mathbf{c}^t) \quad \textbf{Eq 6}$$

where $\text{sigmoid}(x) = 1/e^{-x}$, $\tanh(x) = (e^{2x}-1)/(e^{2x}+1)$, and $\mathbf{g}^t$, $\mathbf{h}^t$, $\mathbf{c}^t$ are the cell input, recurrent state, and cell state at timestep t, respectively. The $\odot$ operator indicates the element-wise multiplication. $\mathbf{X}^t$ is the input covariates matrix (static and dynamic features) associated with timestep t. $\mathbf{W}$ and $\mathbf{b}$ indicate weight and bias, respectively. Data is processed and passed to the adjacent cell through three main subnetworks, or gates: input ($\mathbf{i}^t$), forget ($\mathbf{f}^t$), and output ($\mathbf{o}^t$) gates. A linear head (dense layer) is used at the output of the LSTM cell to map the last encoded input step $\mathbf{h}^t$ (i.e., the recurrent state, hereafter called the context vector) to the target:

$$\hat{y} = \mathbf{W}_l\mathbf{h}^t + \mathbf{b}_l \quad \textbf{Eq 7}$$

where $\hat{y}$ is the specific discharge prediction. The LSTM model is an example of an encoder-decoder structure, where the LSTM cell is the encoder and is used to map the input signal (with dimension $365 \times \text{number of features}$) to a context vector (with dimension $\|\mathbf{h}\| \times 1$), and the linear head is the decoder, mapping the context vector to the target (often a scalar value). While any neural network can be used as the decoder, such as another LSTM model (Jahangir et al., 2023; Kao et al., 2020), the linear head has been shown to be sufficient and efficient to train. Even with a linear head, dropout connections (randomly setting $\mathbf{W}$ entries to zero at a pre-specified rate) are necessary during training to avoid overfitting (see Feng et al., 2020; Kratzert et al., 2018). We explore how a non-linear, non-neural network mapping function (f), namely a RF, can be used as an alternative to the linear head and how this can support more flexible, efficient, and accessible TL in hydrologic contexts.

3.2 Hybrid LSTM-RF Model and Baseline TL Strategies

Using RF as the decoder has several potential benefits, motivating its use in this research. First, RF's ensemble-based structure enables complex, non-linear relationships to be captured by constructing multiple decision-trees on random subsets of data. During training, the algorithm automatically evaluates and ranks feature importance, eliminating the need for manual input variable selection (Cutler et al., 2007). Additionally, by combining predictions from multiple trees and leveraging randomization in both data and feature sampling, random forests exhibit strong resilience to overfitting (Breiman, 2001). For the hybrid LSTM-RF model, the target discharge is estimated by:

$$\hat{y} = \boldsymbol{f}(\mathbf{h}^t) = \boldsymbol{f}(\text{LSTMCell}(\mathbf{X})) \quad \textbf{Eq 8}$$

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

where $\boldsymbol{f}$ is a RF. The schematic of the proposed hybrid model is shown in Figure 2.

For comparison, we consider two other fine-tuning strategies previously suggested by Ma et al. (2021): 1) LSTM^{A}, where the weights of all LSTM layers were updated, and LSTM^{B}, where only the weights of the linear head were updated and the LSTM cell weights were kept frozen (see Figure 2).

The proposed hybrid LSTM-RF has several potential advantages over these two other fine-tuning strategies. Fine-tuning a DL model is highly sensitive to the choice of learning rate. An inappropriate learning rate can lead to a significant decline in performance. Furthermore, as deep learning models grow in complexity and size, the computational resources required for fine-tuning them increase substantially. Feeding the context vector into a non-neural network model like RF can be a straightforward post-processing step. This is beneficial primarily for hydrologists who may not have extensive experience in DL model development but still want to use a regional LSTM or other foundational geophysical models. Using this approach, researchers can fine-tune a model trained on hundreds or thousands of catchments for their specific sites of interest without requiring deep expertise in DL model architecture or access to HPC resources. This substantially lowers the technical barrier to applying advanced DL models in hydrology.

For this study, we integrated the post-processing step within the broader modeling framework, ensuring seamless application. However, a catchment-wise post-processing approach, where the outputs of the LSTM model are extracted and subsequently fed into an RF or another interpretable model, is equally viable. This flexibility allows practitioners to adapt the methodology based on their computational constraints and expertise while benefiting from the knowledge embedded in large-scale, regional DL models for site-specific applications.

In this work, we utilized Google's TensorFlow 2.16 and TF-DF 1.9.2 (updated to YDF) Python 3.11 libraries for most simulations, enabling seamless integration of neural networks and decision trees (hybrid or composite models). We also repeated many DL development experiments with Pytorch 2.5.1 to ensure reproducibility.

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

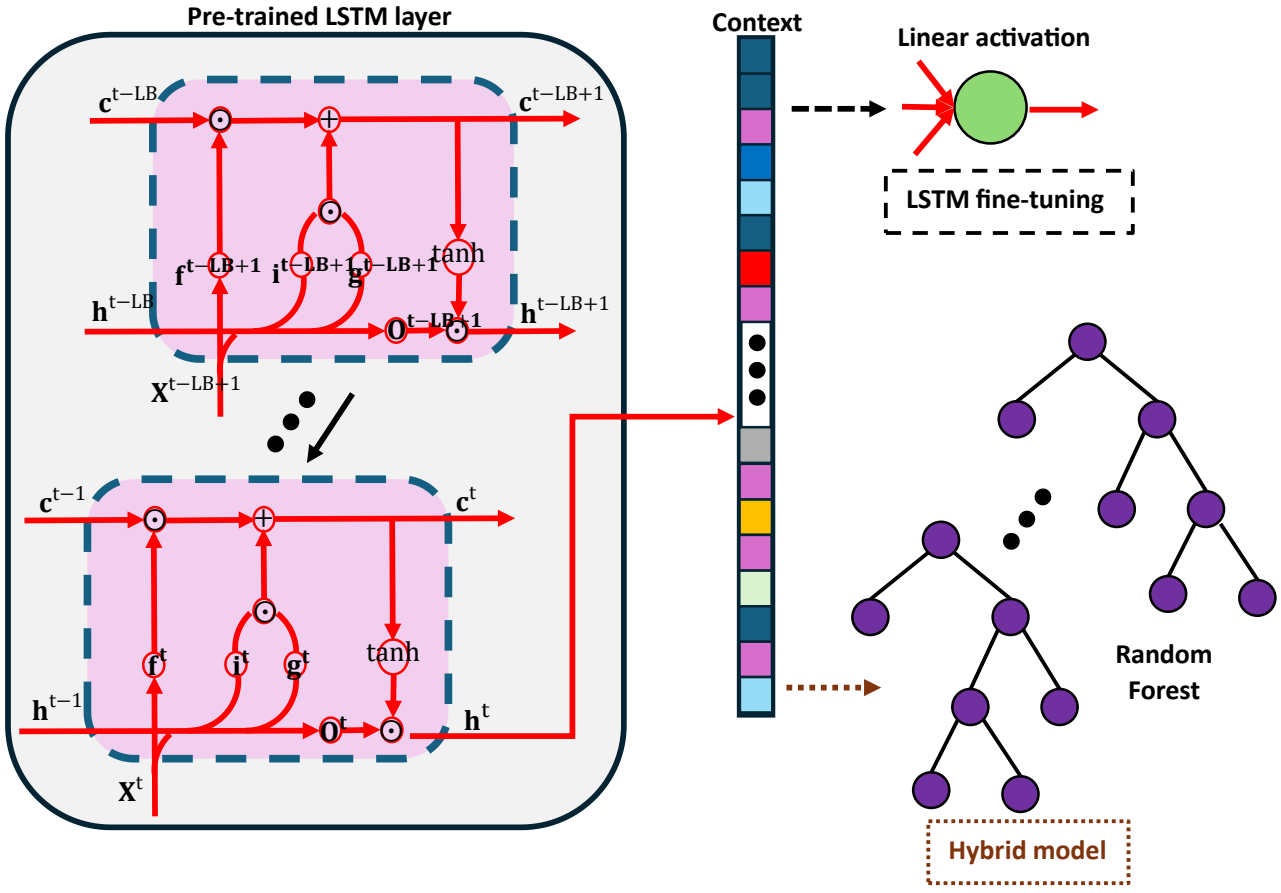


Figure 2: The schematic of the proposed hybrid LSTM-RF model. Instead of the regular linear head, a RF is used to map the context vector to the target. LB is the lookback period. For all experiments, a LB of 365 days was adopted.

3.3 Transfer Learning Experiments

We define two TL scenarios with different numbers of experiments (e.g., model developing type), which are summarized in Table 1. The goal of Scenario 1 is to showcase the effectiveness and efficiency of the hybrid TL method in situations when the available regional LSTM has limited source data (both in terms of the number of catchments and static features). This situation might arise in regions around the world where the number of locations for regional LSTM training is limited or in cases where available landscape features for model inputs are unavailable at the target catchments, complicating TL using a more sophisticated regional LSTM (see Khoshkalam et al. (2023) for details). In addition, Scenario 1 is relevant to situations where analysts are interested in fine-tuning a relatively simpler regional LSTM model that is cheaper and faster to train.

In Scenario 1, a base LSTM was trained on 50 randomly selected catchments (these catchments are listed for each experiment in Table S3). Afterward, the base LSTM model was fine-tuned on the remaining catchments (371 in CAMELS-US and 274 for CAMELS-DE). The base LSTM was only trained using four static features that are likely available for most

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

target sites: average discharge (mm/day), average precipitation (mm/day), and average maximum and minimum temperature (°C). Fine-tuning was conducted catchment-wise using only data from the training period. The fine-tuned models were benchmarked against both a regional LSTM model (trained on all 421 and 324 catchments in CAMELS-US and CAMELS-DE, respectively) and LSTM models trained separately for each catchment. Thus, for Scenario 1, there are five experiments in total: two fine-tuned LSTMs ($LSTM^A$ and $LSTM^B$), the hybrid model, the full regional LSTM, and catchment-wise LSTMs. Note that $LSTM^A$ and $LSTM^B$ were only developed for CAMELS-US and not CAMELS-DE, as our results for CAMELS-US demonstrated the superiority of the hybrid model (see Section 4.1.1).

We also investigated increasing the number of random catchments used to train the base model from 50 to 100; however, this did not significantly improve the performance of the fine-tuned models. This outcome is likely due to the base model being trained with minimal static features in this scenario, thereby limiting improvements in model generalizability with more catchments. To assess the impact of randomness, we also experimented with 10 different random seeds to select the 50 random catchments and evaluated their influence on fine-tuning performance. The results across all seeds were consistent. Therefore, we present the results from one representative random seed in the main text and include another in the SI.

In Scenario 2, the primary goal is to evaluate whether the hybrid approach can improve the performance of a "well-trained" regional LSTM model developed with a larger set of input features and hundreds of catchments, where the risk of negative (adverse) TL outcomes is high (i.e., overfitting and drop in performance; see Wang et al. (2019)). A secondary goal is to assess whether the hybrid approach works better when the base model is sub-optimal (i.e., the weights have not been fully optimized) or when the base model is trained on fewer catchments. Scenario 2 is only applied to CAMELS-US (421 catchments). Unlike Scenario 1, all models in Scenario 2 are trained using 27 static catchment attributes covering topography, climate indices, hydrological signatures, land cover, soil, and geological characteristics. The 27 static catchment attributes are the same as those reported in Kratzert et al. (2019).

Three hybrid models are developed in Scenario 2. The first, called hybrid-371, is our primary hybrid model for this scenario. Here, a base LSTM model is trained on 371 catchments and fine-tuned with the hybrid approach for all 421 catchments. We fine-tune the base LSTM model with its final (optimal) weights, and compare this hybrid model to the base regional model based on performance for the 371 sites used in training and for the 50 sites left out of training. The goal of this comparison is to investigate whether fine-tuning an optimal regional model with the hybrid approach can improve its performance on in- and out-of-sample locations.

Two other hybrid models are also developed and compared to assess the utility of the hybrid approach in situations with regional models fit with sub-optimal weights or fewer catchments. The second hybrid model (hybrid-421) is developed by fine-tuning a base model trained to all 421 catchments but using pre-final weights (checkpoints, e.g., epoch 3). This second model is meant to showcase whether fine-tuning early optimization weights (i.e., a sub-optimal model) can be utilized for rapid DL model development, especially when the goal is scenario analysis and impact assessment of specific catchments rather than training models for numerous catchments (Khan and Coulibaly, 2010). This approach is practical, as early checkpoints of advanced DL models are often publicly available, and deploying these models for inference is far less computationally

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

demanding than training even smaller DL models. We used a model checkpoint from the third epoch of the regional LSTM model but also achieved promising results with the hybrid approach utilizing the first epoch (discussed further in Section 4.2). The optimal checkpoint selection for fine-tuning remains an open research question that we do not explore in detail here.

The third hybrid model (hybrid-50) fine-tunes a base model trained to only 50 catchments but using the final weights. The purpose of comparing hybrid-421 and hybrid-50 is to determine whether it is more advantageous to fine-tune a sub-optimal regional LSTM model trained on a larger number of catchments (hybrid-421), or to fine-tune an optimal regional model trained on fewer catchments (hybrid-50). We also compare the hybrid-421 and hybrid-50 models to a regional model without fine-tuning fit to all 421 catchments, using sub-optimal weights.

Thus, in Scenario 2, there are a total of five experiments: hybrid-371 (base: regional, optimal, trained on 371 catchments), hybrid-421 (base: regional sub-optimal, trained on 421 catchments), hybrid-50 (base: regional optimal, trained on 50 catchments), and two regional models without fine-tuning fit to 371 and 421 catchments.

Hyperparameters for the LSTM models, including context size and dropout rate, were selected based on prior studies and preliminary experiments. The final selected hyperparameters for the LSTM models are shown in Table 1. Extensive experiments were conducted to optimize the learning rate and number of epochs to prevent overfitting during fine-tuning of $LSTM^A$ and $LSTM^B$ (highlighting the sensitivity of these approaches to hyperparameter selection). The results confirmed that for $LSTM^A$, it is most beneficial to adopt a small learning rate (5e-5) and fine-tune for a maximum of 15 epochs, while for $LSTM^B$, a larger initial learning rate of 1e-3 and a maximum of 50 epochs were most effective. Early stopping and a dynamic learning rate adjustment were implemented to mitigate overfitting further. Mean squared error was utilized as the loss function, and the Adam optimizer (Kingma and Ba, 2014) was employed to optimize the weights. Inputs and targets were z-normalized before being fed into the models for better convergence. All models were trained to predict specific discharge (mm/day). For the RF hyperparameters (e.g., tree depth) in the hybrid model, the model performance was highly robust across a wide range of settings. Among the factors evaluated, the context size (hybrid model input cardinality) emerged as having the most significant impact on performance. As a result, default hyperparameter values were adopted for the RF models across all experiments, including a maximum tree depth of 16 and an ensemble size of 300. The use of default hyperparameters for the RF hybrid model, in contrast to the careful hyperparameter selection needed for $LSTM^A$ and $LSTM^B$, highlights a key difference in ease of use between these methods.

Table 1- The defined scenarios and their associated experiments. *- indicates the context size and the dropout rate. X indicates not presented. $LSTM^A$ indicates fine-tuning all the weights, while $LSTM^B$ indicates fine-tuning the linear head weights.**

Scenario 1: TL for improving regional LSTM model with limited source data						
Experiment	Regional LSTM model	Catchment-wise LSTM Model	Base LSTM	Fine-tuning ($LSTM^A$)	Fine-tuning ($LSTM^B$)	Hybrid approach
CAMELS-US	256*-0.4**	128-0.2	128-0.2	Update the weights of the LSTM cell and linear head	Update the linear head weights	Base LSTM-RF
CAMELS-DE	128-0.4	X	64-0.2	X	X	Base LSTM-RF

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

Scenario 2: TL for improving regional LSTM model with ample source data						
Experiment	Regional LSTM model	Catchment-wise LSTM Model	Base LSTM	Fine-tuning ($LSTM^A$)	Fine-tuning ($LSTM^B$)	Hybrid approach
CAMELS-US	256-0.4	X	256-0.4	X	X	Base LSTM-RF

3.4 Performance Evaluation

The two most common deterministic, scale-independent performance measures in hydrology were adopted for evaluating the performance: the modified Kling–Gupta efficiency (Kling et al., 2012) and NSE (McCuen et al., 2006) (see SI for mathematical formulations). A NSE and KGE of 1 is ideal, and NSE and KGE values greater than 0.75 are considered good (Crochemore et al., 2015; Moriasi et al., 2007; Palash et al., 2024). The percent bias of the top 2% peak flow range (FHV) and the percent bias of the bottom 30% (baseflow) range (FLV) was also assessed (see Feng et al. (2020) and Yilmaz et al. (2008)). To identify whether the hybrid approach significantly improved prediction performance compared to the regional LSTM model, the Wilcoxon signed-rank test was adopted (Wilcoxon ,1945). The Wilcoxon signed-rank test was applied to the absolute error values from both the hybrid approach and the regional LSTM model, separately for each site. This was used to determine if there was a statistically significant difference in prediction performance, indicating whether the hybrid approach offered a meaningful improvement over the regional LSTM model on a site-by-site basis. A significance level of 95% was used to evaluate the significance of the Wilcoxon test. All reported metrics are associated with the test set.

4 Results and Discussion

This section presents and analyses each scenario's results and corresponding experiments. First, we examine the outcomes of utilizing TL for effective hydrological prediction when source data is limited (Scenario 1). Next, we discuss the results related to enhancing the performance of a regional model built on more expansive source data (Scenario 2). The role of the LSTM in encoding the metrological forcing into informative clusters is briefly discussed, and finally, the limitations of this study are addressed at the end of this section.

4.1 Scenario 1

4.1.1 CAMELS-US

Figure 3 shows the distribution of NSE and KGE values for the 371 catchments excluded from base model training in Scenario 1 for the hybrid, $LSTM^A$ and $LSTM^B$ models. For comparison, the solid line represents the median of the regional model trained on all 421 catchments, while the dashed line indicates the median of the catchment-wise LSTM models (see SI for details). The median KGE and NSE values for the base model trained to 50 catchments are also shown in the figure (dashed-dotted line). The results demonstrate that the hybrid approach outperforms the other two TL methods. Specifically, the hybrid model achieves a 3.99% (1.11%) and 4.84% (6.04%) improvement in KGE and NSE medians compared to $LSTM^A$ ($LSTM^B$),

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

respectively. The hybrid model also improved performance over the two other TL methods for the worst-performing catchments, showing better performance in the lower portion of the distribution. Additionally, the hybrid model showed a 4.14% and 9.19% improvement in KGE and NSE medians, respectively, compared to the catchment-wise training method. Results suggest that $LSTM^A$ and $LSTM^B$ resulted in relatively similar performance.

The regional model did not substantially outperform the catchment-wise approach due to the limited use of static features during training. This underscores the role of incorporating sufficient static features in regional DL development for hydrology and highlights the challenges in transferring regional models to locations with differing static feature availability. Interestingly, the results indicate that using the hybrid approach to fine-tune the base model, even when the base model exhibits poor out-of-sample performance, proves to be more advantageous than developing a regional deep learning model when sufficient static features are not utilized. Overall, the hybrid approach showed promising performance considering that a relatively simple base model (low number of catchments, few variables) was adopted.

Similar results to those above were obtained for another random set of catchments used to train the base model (Table S3 in SI), where the hybrid model generally outperformed the other two methods. The hybrid model showed 4.84%, and 7.50% improvement for KGE and NSE medians (Fig S1) compared to the catchment-wise training method. However, unlike the first adopted pool, it was observed that $LSTM^A$ generally performed better than $LSTM^B$.

In addition to the performance differences above, the experiments indicated that the hybrid method is more efficient than the other two TL methods. The optimization times for the models, shown in Table 2, were recorded on a single node of Alliance Canada's Graham cluster, equipped with a V100 GPU, two CPU cores, and 48 GB of RAM. These hardware settings were chosen based on the availability of Alliance Canada resources and to ensure reproducibility on personal workstations. The hybrid approach was approximately 35% and 32% faster than $LSTM^A$ and $LSTM^B$, respectively. For comparison, and for a specified random seed, optimizing a regional model for all 421 catchments required 3 hours and 48 minutes, the base model (trained on 50 catchments) took approximately 36 minutes, and the catchment-wise training strategy required 8 hours and 42 minutes. It should be noted that recorded times are impacted by many factors, such as model weight initialization and software configurations (e.g., the Python and CUDA versions).

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

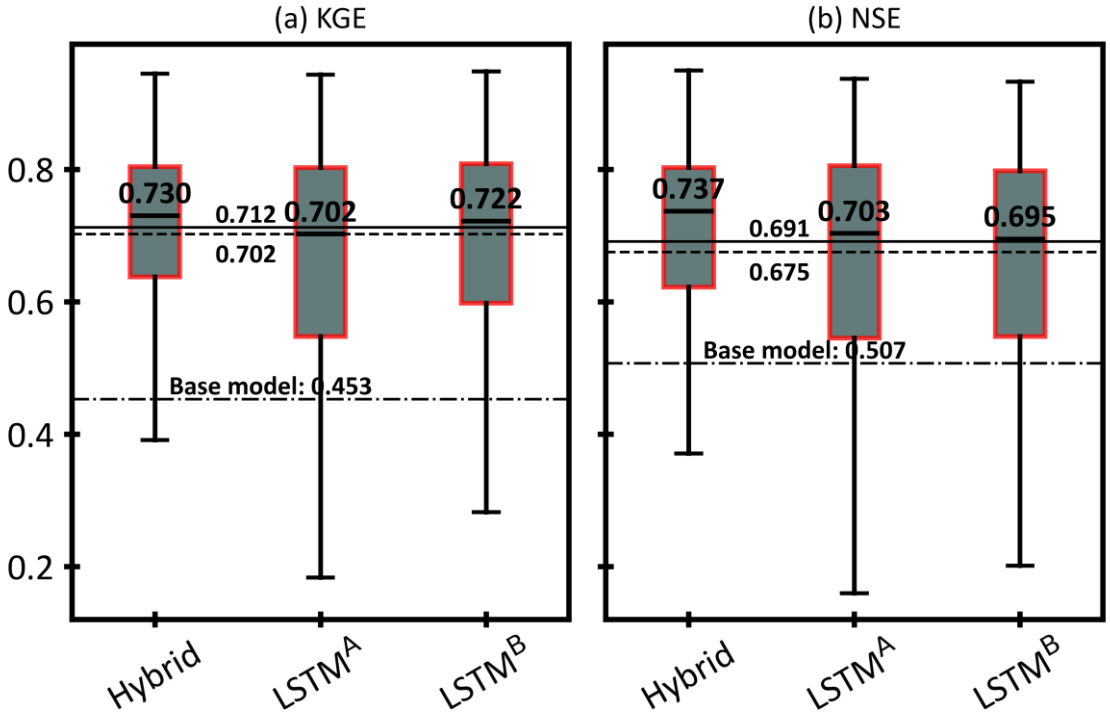


Figure 3: (a) NSE and (b) KGE distribution over the 371 catchments for the different approaches to improve the base model. For comparison, the solid line indicates the regional model median trained on all 421 catchments (using minimum static features), while the dashed line shows the median of the catchment-wise LSTM models.

Table 2- Wall time comparison of TL methods. All the recorded times are in seconds. The reported statistics are based on wall times averaged across all 421 catchments.

	Method		
Stat	Hybrid	$LSTM^A$	$LSTM^B$
Average	47.22	72.36	69.23
Median	42.87	69.46	62.47
St.D.	13.31	5.96	29.95
Min.	37.37	67.92	34.56
Max.	88.61	84.87	159.89

4.1.2 CAMELS-DE

Figure 4 shows the empirical cumulative distribution function (CDF) of KGE and NSE values for the 274 catchments not used for base model training. Three CDFs are shown: one for the hybrid approach, which fine-tunes the base model trained on 50 catchments, one for the regional benchmark trained on all 1555 catchments in CAMELS-DE, and one for the base model.

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

Recall that we did not develop TL models (LSTM^{A} and LSTM^{B}) for CAMELS-DE. The findings indicate that the hybrid approach performs competitively with the regional model, even though the base model was trained with fewer static features and only 50 catchments and shows inferior performance on the out-of-training catchments. Although the median NSE of the hybrid model was 2.54% lower than that of the regional model, the hybrid model achieved a minimum NSE value of 0.369 compared to 0.094 for the regional model (catchment ID: DE810470). This disparity is evident in the CDF plot, where the regional model displays a heavier lower tail. When comparing the medians of the differences, the regional model demonstrated marginal improvements of 0.57% in KGE and 1.48% in NSE over the hybrid model.

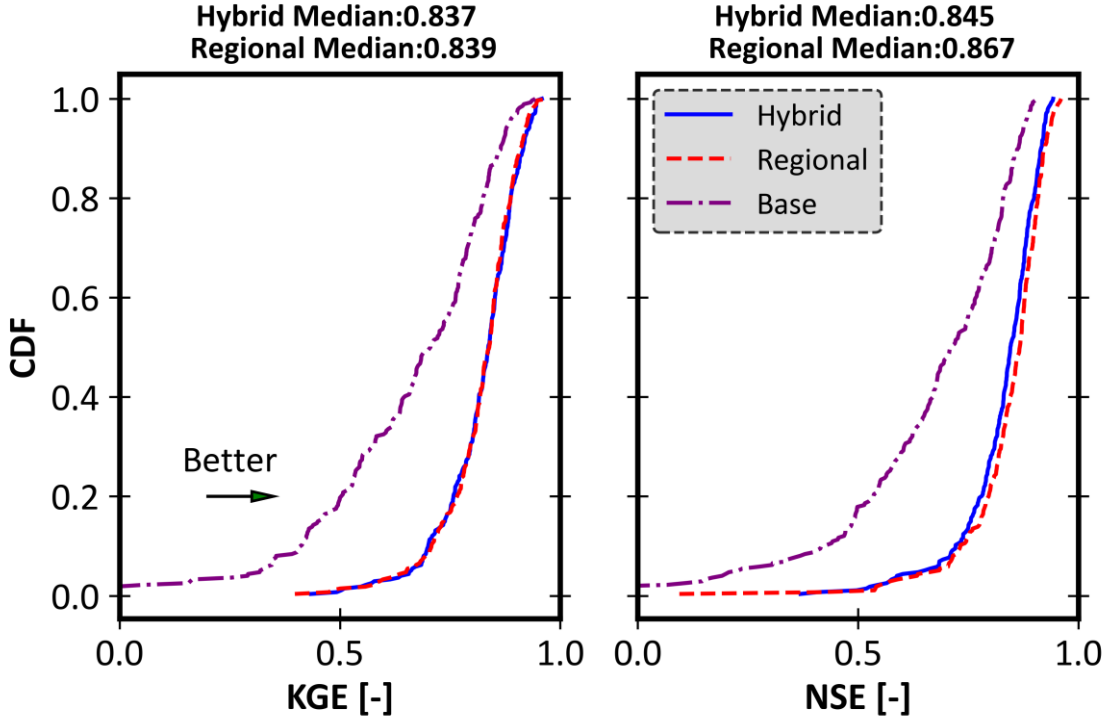


Figure 4: NSE and KGE empirical CDF for the selected 274 catchments in Germany.

The performance of the models under low-flow and high-flow conditions was also evaluated. The empirical CDFs for FLV and FHV, corresponding to the regional and hybrid models, are presented in the SI (Figure S2). Results indicate that the hybrid model performed comparably to the regional model, with a slight advantage in high-flow regimes. Specifically, the hybrid model achieved median FLV and FHV values of 4.34% and -16.95%, respectively, compared to 4.37% and -17.77% for the regional model.

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

4.2 Scenario 2

The results of fine-tuning a regional model using the proposed hybrid method are presented. In this analysis, both optimal and sub-optimal regional models were evaluated. Given that the hybrid approach consistently outperformed the alternative methods ($LSTM^A$ and $LSTM^B$), only the hybrid model was selected for further experimentation.

4.2.1 CAMELS-US

Figure 5 presents the NSE and KGE values for both the hybrid model (hybrid-371) and the corresponding regional model fit to 371 sites and using optimal weights. Each model is represented by two groups. The dashed boxplot indicates catchments (50) that were not included in the training of the regional (base) model. The results confirm that fine-tuning the regional model leads to significant performance improvements for catchments that were that were not used in training (50), showing a 14.11% and 13.13% improvement in median KGE, and NSE, respectively. For the catchments already adopted in the training of the regional model, the hybrid method resulted in 0.66% and 1.19% improvement in median KGE and NSE, respectively. Repeated analyses using various random seeds consistently demonstrated that the hybrid method enhances the performance of a regional model for out-of-sample catchments but does not meaningfully improve performance for catchments already included in model training for a regional model fit to many sites and with many static features.

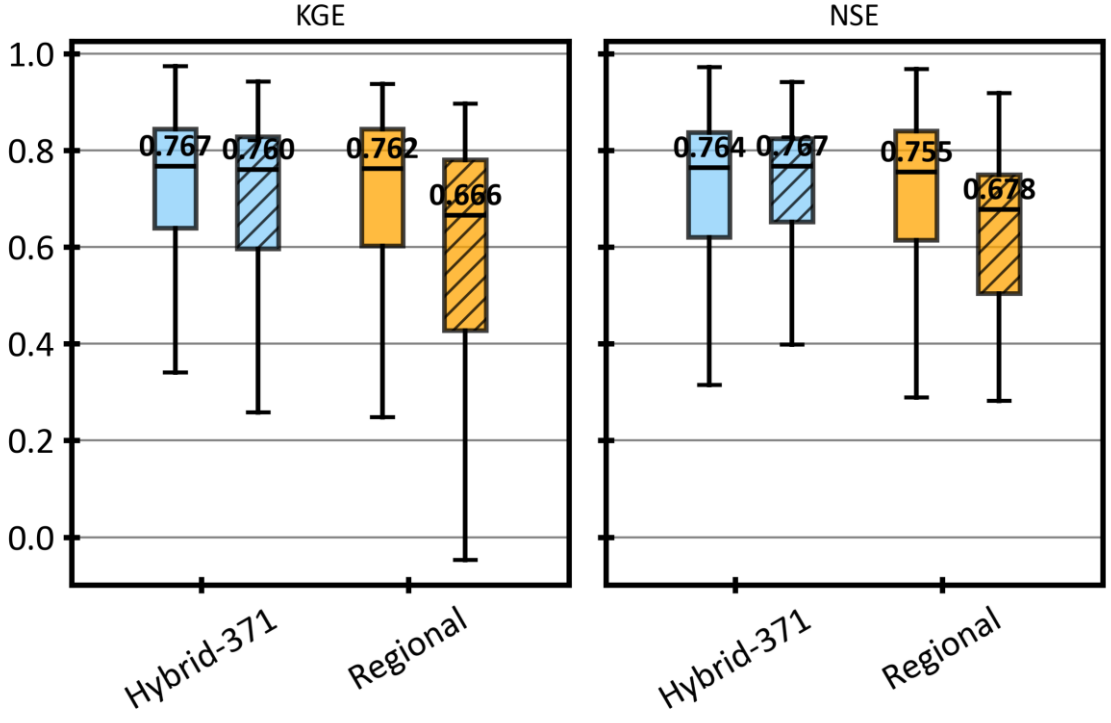


Figure 5: NSE and KGE for hybrid (hybrid-371, light blue) and regional (base, orange) models. The dashed box plots are associated with the catchments that were not used in training the regional model.

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

The potential of fine-tuning for improving sub-optimal regional models was also assessed. The distributions of NSE and KGE for a regional LSTM model and two hybrid model variants are presented in Figure 6. For clarification, the "hybrid-50" method corresponds to the case where the base model was trained on only 50 randomly selected catchments (see SI), like Scenario 1, but utilizing all available (27) static features. In contrast, the "hybrid-421" method involves fine-tuning the sub-optimal regional LSTM model (using the epoch-3 checkpoint). The results demonstrate that the hybrid approach is effective and delivers competitive performance compared to the regional LSTM model, even when the base model was trained on a limited subset of catchments (hybrid-50). Overall, fine-tuning the sub-optimal regional model led to marginal performance gains, with median KGE and NSE values improving by 0.92% and 1.19%, respectively. Comparing the results of hybrid-421 and hybrid-50 indicates that fine-tuning a sub-optimal regional model trained to more sites is more beneficial compared to fine-tuning an optimal regional model trained with fewer catchments.

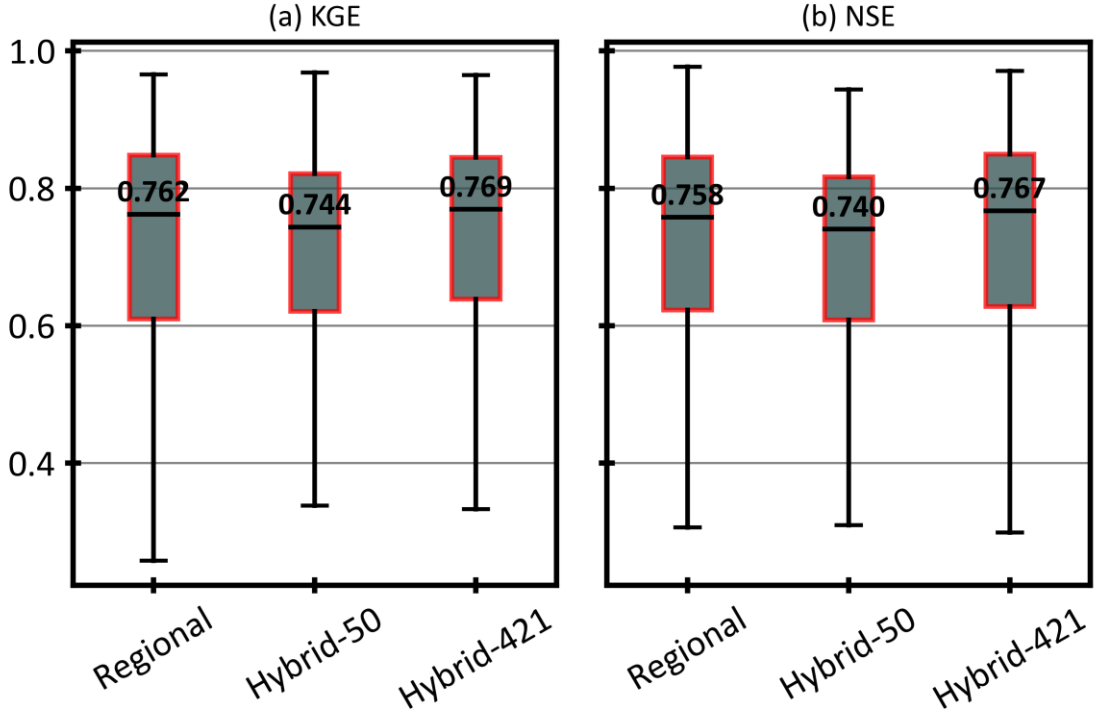


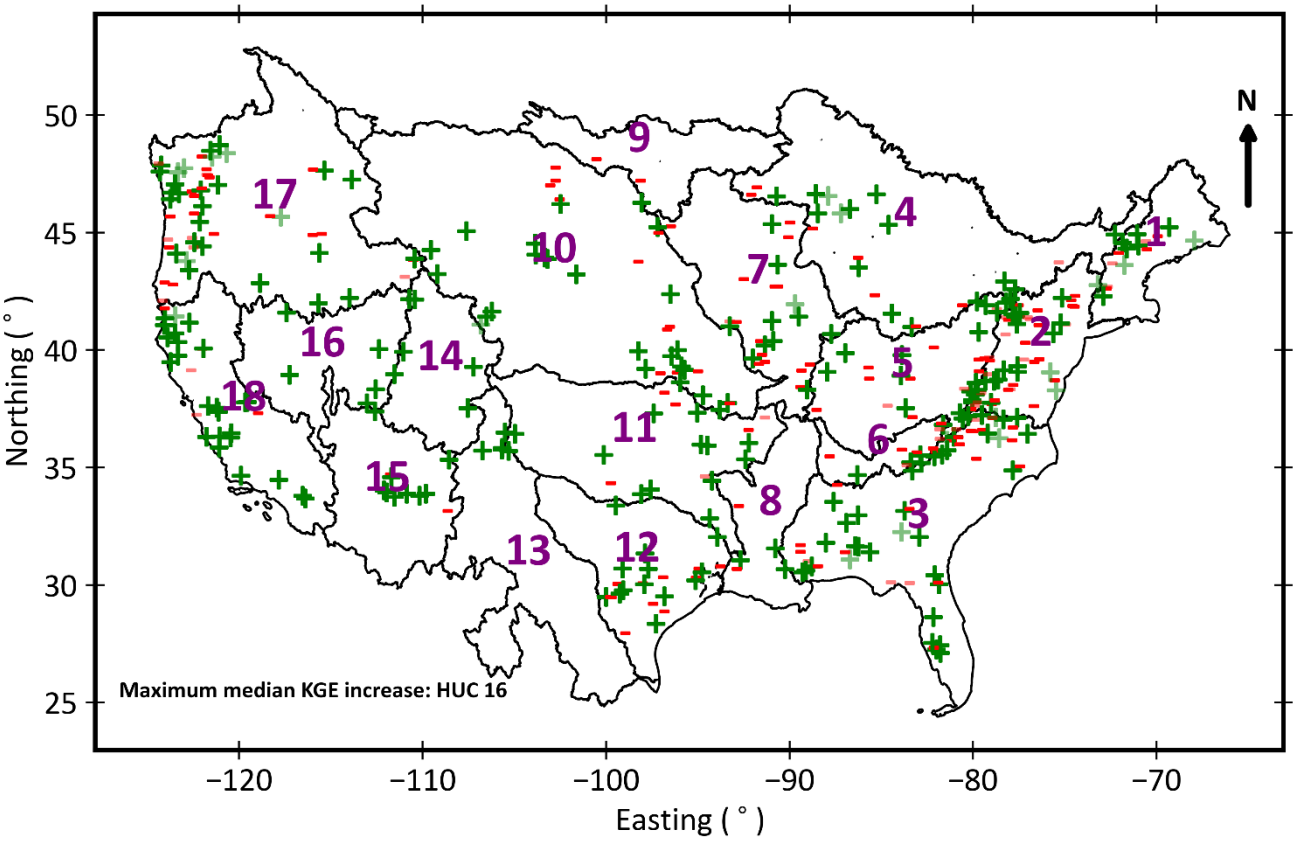
Figure 6: NSE and KGE distribution over the 421 catchments associated with the regional, hybrid-50, and hybrid-421 models.

The results presented for hybrid-421 correspond to the weights retrieved at the third epoch; however, we also tested earlier and later checkpoints and different weight initializations. Fine-tuning the weights from the first epoch of a pre-trained regional model also demonstrated competitive performance, yielding KGE and NSE values of 0.763 and 0.748, respectively, compared to 0.766 and 0.752 achieved by the regional model (see Table S5). It is important to note that the results provided were not selectively presented. Our experiments showed that the epoch yielding the best performance from the fine-tuned model varied

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

and depended on the initialization of the model's weights. Optimally selecting the checkpoint for fine-tuning is challenging, relies on many factors (e.g., the designed pipeline), and falls outside the scope of this work.

Even though the median performance metrics in Scenario 2 suggest incremental improvements for in-sample locations included in the regional (base) models, it is useful to assess the specific catchments where performance gains occurred with the hybris approach and whether these gains are statistically significant. This analysis was performed using the Wilcoxon test, with the results illustrated in Figure 7. Catchments with significant KGE improvements are marked by green positive markers, whereas those with significant decreases are indicated by red negative markers. Transparent markers represent catchments where the differences were not statistically significant. The analysis reveals that KGE significantly increased (decreased) in 53.51% (31.72%) of the catchments. A similar analysis was conducted for NSE (see Figure S3 in SI), showing that in 43.60% (41.38%) of the catchments, NSE was significantly increased (decreased). For KGE, the greatest improvement was observed in the Great Basin Region (HUC 16), with a median improvement of 0.161 and a mean improvement of 0.220. Conversely, the largest performance decline occurred in the Souris-Red-Rainy Region (HUC 9), with a median and mean decrease of -0.220. For NSE, the greatest improvement was observed in the Lower Colorado Region (HUC 15), with a median increase of 0.102 and a mean increase of 0.212. Conversely, the largest performance decline occurred in the Souris-Red-Rainy Region (HUC 9), with a median and mean decrease of -0.111. However, HUC 9 only includes two catchments.



https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

Figure 7: KGE increase (green +) or decrease (red -) across the selected 421 catchments, based on comparing hybrid-421 to the regional LSTM model fit to 421 catchments and optimal weights. Only catchments with KGE>-1 were considered for comparison. Non-significant changes are shown in transparent color. The numbers indicate the HUC.

An interesting and important direction for advancing hydrological prediction lies in identifying where and when different models outperform one another. Although exploring this research avenue falls outside the scope of the current study, preliminary analysis revealed that the hybrid approach serves as a viable alternative in catchments where the regional model underperforms. For instance, in catchments where the regional model yielded KGE and NSE values below 0.75 ("good" threshold), mostly those with lower baseflow ratios and less snowmelt influence, the hybrid model improved performance. When the hybrid approach was used instead of the regional model, the median KGE and NSE increased from 0.762 to 0.788 and from 0.758 to 0.773, respectively.

4.3 LSTM Cell, a Non-Linear Clustering Model

It was observed that fine-tuning a well-trained regional model on the same catchments it was originally trained on can results in marginal performance gains. Also, as demonstrated in previous sections, fine-tuning, whether through the hybrid method or standard approaches, results in substantial performance gains (compared to regional modeling) when applied to a sub-optimal (fewer sites and static features, or early epoch) regional model. Fine-tuning can be performed using either the proposed hybrid approach, which is computationally efficient but may require post-processing, or through standard neural network fine-tuning, which provides an end-to-end solution but is more sensitive to data availability and learning rate customization and requires expertise. Regardless of the method, this approach should be viewed as a complementary tool to regional modeling, for catchments where the regional model demonstrates suboptimal performance.

The context vector represents the transformation of a high-dimensional input space (lag, number of features) into a smaller informative space (**h**). This transformation can be observed as a non-linear dimension reduction, an alternative to linear methods such as principal component analysis. Fine-tuning was performed effectively in cases where the base models exhibited limited generalization, as seen with the regional model trained with minimal static features (Scenario 1) and early checkpoint weights (Scenario 2). However, results indicated that the sub-optimal models are successful at grouping/clustering catchments based on the meteorological forcing. Cosine similarity (Luo et al., 2018) was used to assess this clustering and measure the similarity between context outputs in the catchments of interest. As an example, the average similarity (all samples in the train set) between the context outputs of one random catchment located in HUC 1 (catchment ID: 1022500), with those of 370 other catchments (all not used for training the base model in scenario 1) is shown in Figure 7. As expected, catchments located in HUC 1 showed the most similarity to the selected catchment. The findings of Ma et al. (2021) showed that fine-tuning the linear head of an LSTM model ($LSTM^B$) is more effective when the source and target catchments share similar properties. This effectiveness is due to the reduced variability in context outputs generated by a well-trained LSTM cell (trained over many catchments). When source and target catchments are similar, the linear head can accurately decode the encoded context to the target. In contrast, when catchment properties differ significantly, the encoded space fails to capture the required distinctions,

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

resulting in suboptimal decoding. In essence, the model struggles to differentiate the new catchment from previously learned ones. Obtaining the best encoder for encoding the input forcing, and, consequently, the best encoded representation is an open research avenue (Liu et al., 2023). Advancements in this area have significant potential to enhance data-driven hydrology and improve model transferability across diverse catchments.

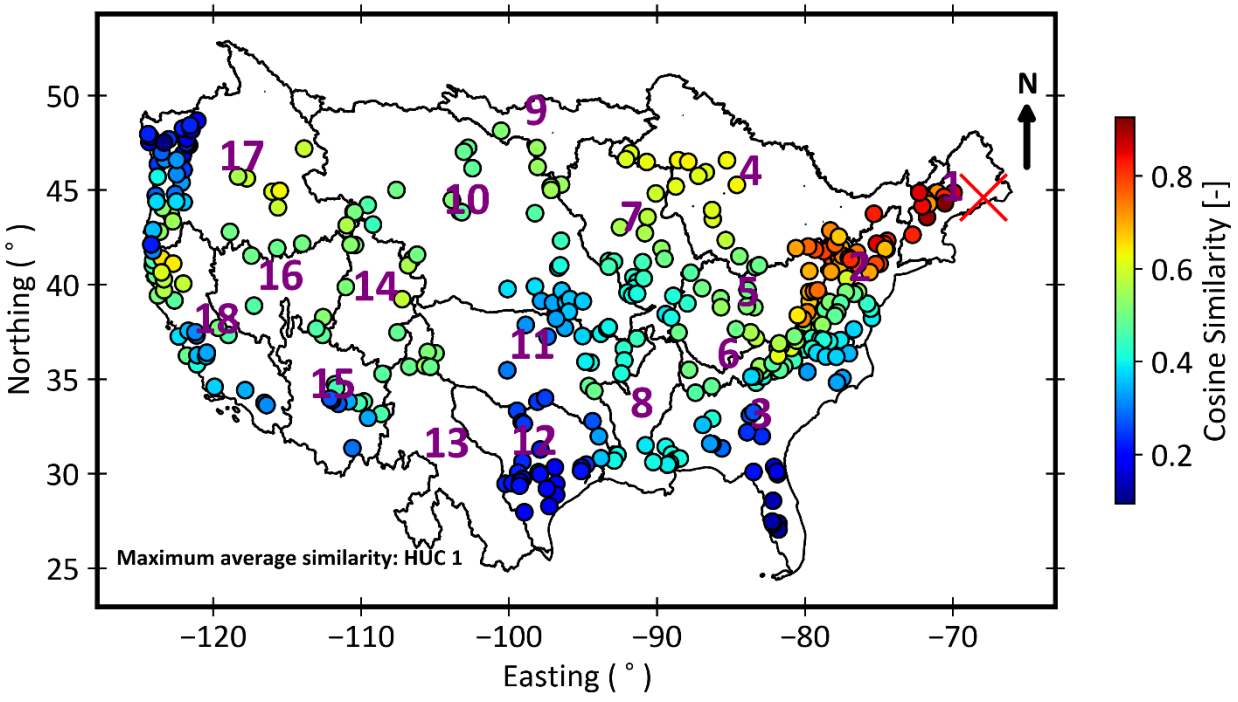


Figure 8: Average Cosine similarity between a random catchment (located in HUC 1, shown by cross) and the other 370 catchments.

5 Conclusion

Accurate hydrological prediction is essential for effective water resource management and the development of reliable early warning systems. Within the hydrology community, data-driven approaches have gained recognition due to their superiority over traditional conceptual modeling in capturing complex hydrological processes. With the growing availability of data, advancements in deep learning (DL) architectures, and increased computational power, the strategic utilization and transfer of complex DL models will play an essential role in the future of data-driven hydrology. It is expected that transfer learning (TL) will be a key component to leveraging these advancements. This study explored a hybrid LSTM–random forest method as an efficient and accurate approach to TL, evaluating its performance across two distinct geographic regions, the continental United States (421 catchments) and Germany (324 catchments), using benchmark datasets (CAEMLS).

Two scenarios were defined to assess the proposed method. In the first scenario, the study evaluated the hybrid model's effectiveness for efficient fine-tuning of a regional LSTM trained with limited data (samples and features). In the second

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

scenario, the hybrid method was assessed for its ability to fine-tune a regional model fit on a large set of sites and features. The results from Scenario 1 demonstrated that the proposed method meaningfully improved performance over other TL strategies ($LSTM^A$ and $LSTM^B$) when the base regional model was fit to a limited number of sites and features, and was also more computationally efficient and less sensitive to hyperparameter selection. The hybrid approach also substantially outperforms catchment-specific DL model development.

In Scenario 2, we found that the hybrid approach resulted in substantial performance improvement when utilized for fine-tuning on catchments excluded from the regional model's training set, but only produced marginal improvements for sites included in regional model training. Furthermore, fine-tuning for sub-optimal regional models fit to more sites led to larger improvements compared to fine-tuning of optimal regional models fit to less sites. The results also demonstrated that the hybrid method is effective in improving the performance of regional DL models in catchments where those regional models underperform, underscoring its potential as a useful tool for hydrological prediction in areas that challenge existing regional models.

Overall, the results of this work highlight the benefits of fine-tuning with the hybrid approach, particularly for regional DL models fit to few sites and features, but also in the context of larger regional models. The hybrid approach can be utilized when developing a model for a new site instead of retraining the model using all the data or developing models catchment-wise.

This work aimed to introduce a novel method that benefits both the scientific community and industry. The findings of this study have caveats because of two main reasons. First, we only tested the approach for catchments with relatively long records (20-30 years) for fine-tuning. This assumption does not hold in many cases, especially in developing countries where monitoring networks are not as extensive and are often newly established compared to developed countries, such as the U.S. and Germany. However, the aim of this research was different and focused on improving efficiency and accuracy, as previous work (e.g., Khoshkalam et al., 2023; Ma et al., 2021) has already demonstrated the benefits of using TL and fine-tuning for data-scarce regions. Second, the checkpoint selection (retrieving the DL model weights) has meaningful impacts on performance. Future research should aim to understand better how checkpoint selection affects fine-tuning outcomes. This knowledge would help practitioners determine when fine-tuning sub-optimal models (which are often released by leading DL model developers) is viable for their specific local applications.

The effectiveness of the hybrid method was tied to the context vector generated by the LSTM cell. A future avenue of research is to investigate how to maximize the information encoded in the context vector, thereby enabling the development of the "best" fine-tuned model. Methods such as self-supervised learning (Liu et al., 2023) or variational modeling (Blei et al., 2017) can be explored to achieve this. Recent advancements in encoder-based large language models have demonstrated their ability to generate context-rich embeddings (Devlin et al., 2019) that enhance downstream predictive tasks across various tasks (Jin et al., 2024). Thus, there is great potential for effectively utilizing such models to produce informative contexts in the hydrologic sciences and for seamless integration into hydrological prediction frameworks accessible to hydrologists and local water resource practitioners worldwide.

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

Data and Code Availability

All data used in this study are publicly available. For CAMELS-US, meteorological data were sourced from Addor et al. (2017), while discharge data were obtained from Kratzert et al. (2023). The CAMELS-DE data were acquired from Loritz et al. (2024). Sample code for reproducing the results is available at: https://github.com/sinajahangir/Hybrid_LSTM_RF.

Author Contribution

MSJ was responsible for designing the experiments, developing the models, drafting the initial manuscript, and conducting the analysis. JQ, AS, CS, SS, and JA assisted with conceptualization, reviewed the manuscript, and provided technical feedback. Additionally, SS assisted with the experiment design and revised the manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgment

This research was supported by Natural Resources Canada (PHIMP23-27P3). We confirm that the funding sources did not influence the results presented in this study.

References

Addor, N., Newman, A.J., Mizukami, N., Clark, M.P., 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrology and Earth System Sciences 21, 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

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

Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational Inference: A Review for Statisticians. Journal of the American Statistical Association 112, 859–877. https://doi.org/10.1080/01621459.2017.1285773

Bojer, C.S., Meldgaard, J.P., 2021. Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting 37, 587–603. https://doi.org/10.1016/j.ijforecast.2020.07.007

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20. Curran Associates Inc., Red Hook, NY, USA, pp. 1877–1901.

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

Crochemore, L., Perrin, C., Andréassian, V., Ehret, U., Seibert, S.P., Grimaldi, S., Gupta, H., Paturel, J.-E., 2015. Comparing expert judgement and numerical criteria for hydrograph evaluation. Hydrological Sciences Journal 60, 402–423. https://doi.org/10.1080/02626667.2014.903331

Cui, Z., Zhou, Y., Guo, S., Wang, J., Xu, C.-Y., 2022. Effective improvement of multi-step-ahead flood forecasting accuracy through encoder-decoder with an exogenous input structure. Journal of Hydrology 609, 127764. https://doi.org/10.1016/j.jhydrol.2022.127764

Cutler, D.R., Edwards Jr., T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random Forests for Classification in Ecology. Ecology 88, 2783–2792. https://doi.org/10.1890/07-0539.1

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805

Elman, J.L., 1990. Finding Structure in Time. Cognitive Science 14, 179–211. https://doi.org/10.1207/s15516709cog1402_1

Fang, S., Johnson, J.M., Yeghiazarian, L., Sankarasubramanian, A., 2024. Improved National-Scale Above-Normal Flow Prediction for Gauged and Ungauged Basins Using a Spatio-Temporal Hierarchical Model. Water Resources Research 60, e2023WR034557. https://doi.org/10.1029/2023WR034557

Feng, D., Fang, K., Shen, C., 2020. Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks With Data Integration at Continental Scales. Water Resources Research 56, e2019WR026793. https://doi.org/10.1029/2019WR026793

Feng, D., Liu, J., Lawson, K., Shen, C., 2022. Differentiable, Learnable, Regionalized Process-Based Models With Multiphysical Outputs can Approach State-Of-The-Art Hydrologic Prediction Accuracy. Water Resources Research 58, e2022WR032404. https://doi.org/10.1029/2022WR032404

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Computation 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu, Q., Zhang, R., Zhou, Y., 2016. Transfer learning for short-term wind speed prediction with deep neural networks. Renewable Energy 85, 83–95. https://doi.org/10.1016/j.renene.2015.06.034

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

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

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J.Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., Wen, Q., 2024. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. https://doi.org/10.48550/arXiv.2310.01728

Kao, I.F., Zhou, Y., Chang, L.C., Chang, F.J., 2020. Exploring a Long Short-Term Memory based Encoder-Decoder framework for multi-step-ahead flood forecasting. J. Hydrol 583, 124631.

Khan, M.S., Coulibaly, P., 2010. Assessing Hydrologic Impact of Climate Change with Uncertainty Estimates: Bayesian Neural Network Approach. Journal of Hydrometeorology 11, 482–495. https://doi.org/10.1175/2009JHM1160.1

Khoshkalam, Y., Rousseau, A.N., Rahmani, F., Shen, C., Abbasnezhadi, K., 2025. Does grouping watersheds by hydrographic regions offer any advantages in fine-tuning transfer learning model for temporal and spatial streamflow predictions? Journal of Hydrology 650, 132540. https://doi.org/10.1016/j.jhydrol.2024.132540

Khoshkalam, Y., Rousseau, A.N., Rahmani, F., Shen, C., Abbasnezhadi, K., 2023. Applying transfer learning techniques to enhance the accuracy of streamflow prediction produced by long Short-term memory networks with data integration. Journal of Hydrology 622, 129682. https://doi.org/10.1016/j.jhydrol.2023.129682

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization.

Kling, H., Fuchs, M., Paulin, M., 2012. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. Journal of Hydrology 424–425, 264–277. https://doi.org/10.1016/j.jhydrol.2012.01.011

Kratzert, F., Gauch, M., Klotz, D., Nearing, G., 2024. HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin. Hydrology and Earth System Sciences 28, 4187–4201. https://doi.org/10.5194/hess-28-4187-2024

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall–runoff modelling using long short-term memory (LSTM) networks. Hydrol. Earth Syst. Sci 22, 6005–6022.

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

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

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

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J., 2023. Self-Supervised Learning: Generative or Contrastive. IEEE Transactions on Knowledge and Data Engineering 35, 857–876. https://doi.org/10.1109/TKDE.2021.3090866

Loritz, R., Dolich, A., Acuña Espinoza, E., Ebeling, P., Guse, B., Götte, J., Hassler, S.K., Hauffe, C., Heidbüchel, I., Kiesel, J., Mälicke, M., Müller-Thomy, H., Stölzle, M., Tarasova, L., 2024. CAMELS-DE: hydrometeorological time series and attributes for 1555 catchments in Germany. Earth System Science Data Discussions 1–30. https://doi.org/10.5194/essd-2024-318

Luo, C., Zhan, J., Xue, X., Wang, L., Ren, R., Yang, Q., 2018. Cosine Normalization: Using Cosine Similarity Instead of Dot Product in Neural Networks, in: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (Eds.), Artificial Neural Networks and Machine Learning – ICANN 2018. Springer International Publishing, Cham, pp. 382–391. https://doi.org/10.1007/978-3-030-01418-6_38

Ma, K., Feng, D., Lawson, K., Tsai, W.-P., Liang, C., Huang, X., Sharma, A., Shen, C., 2021. Transferring Hydrologic Data Across Continents – Leveraging Data-Rich Regions to Improve Hydrologic Prediction in Data-Sparse Regions. Water Resources Research 57, e2020WR028600. https://doi.org/10.1029/2020WR028600

McCuen, R.H., Knight, Z., Cutter, A.G., 2006. Evaluation of the Nash–Sutcliffe Efficiency Index. Journal of Hydrologic Engineering 11, 597–602. https://doi.org/10.1061/(ASCE)1084-0699(2006)11:6(597)

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

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

Niu, S., Liu, Y., Wang, J., Song, H., 2020. A Decade Survey of Transfer Learning (2010–2020). IEEE Transactions on Artificial Intelligence 1, 151–166. https://doi.org/10.1109/TAI.2021.3054609

Palash, W., Akanda, A.S., Islam, S., 2024. A data-driven global flood forecasting system for medium to large rivers. Sci Rep 14, 8979. https://doi.org/10.1038/s41598-024-59145-w

Pan, S.J., Yang, Q., 2010. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering 22, 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Thornton, P.E., Thornton, M.M., Mayer, B.W., Wilhelmi, N., Wei, Y., Devarakonda, R.C., R.B., States). Tyralis, H., Papacharalampous, G., Burnetas, A., Langousis, A., 2014. Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 2. J. Hydrol 577, 123957.

Wang, Z., Dai, Z., Poczos, B., Carbonell, J., 2019. Characterizing and Avoiding Negative Transfer, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, pp. 11285–11294. https://doi.org/10.1109/CVPR.2019.01155

Wilcoxon, F., 1945. Individual Comparisons by Ranking Methods. Biometrics Bulletin 1, 80–83. https://doi.org/10.2307/3001968

Xie, K., Liu, P., Zhang, J., Han, D., Wang, G., Shen, C., 2021. Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships. Journal of Hydrology 603, 127043. https://doi.org/10.1016/j.jhydrol.2021.127043

Xie, Y., Wang, Z., Mai, G., Li, Y., Jia, X., Gao, S., Wang, S., 2023. Geo-Foundation Models: Reality, Gaps and Opportunities, in: Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems, SIGSPATIAL '23. Association for Computing Machinery, New York, NY, USA, pp. 1–4. https://doi.org/10.1145/3589132.3625616

https://doi.org/10.5194/egusphere-2025-846
Preprint. Discussion started: 13 March 2025

Yao, Y., Zhao, Y., Li, X., Feng, D., Shen, C., Liu, C., Kuang, X., Zheng, C., 2023. Can transfer learning improve hydrological predictions in the alpine regions? Journal of Hydrology 625, 130038. https://doi.org/10.1016/j.jhydrol.2023.130038

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

Yin, H., Zhang, X., Wang, F., Zhang, Y., Xia, R., Jin, J., 2021. Rainfall-runoff modeling using LSTM-based multi-state-vector sequence-to-sequence model. Journal of Hydrology 598, 126378. https://doi.org/10.1016/j.jhydrol.2021.126378

Zhao, Z., Alzubaidi, L., Zhang, J., Duan, Y., Gu, Y., 2024. A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. Expert Systems with Applications 242, 122807. https://doi.org/10.1016/j.eswa.2023.122807