Improving Discharge Predictions in Ungauged Basins: Harnessing the Power of Disaggregated Data Modeling and Machine Learning

Aggrey Muhebwa[1], Colin J. Gleason[2], Dongmei Feng[3] and Jay Taneja[1]

[1]Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA, USA

[2]Department of Civil and Environmental Engineering, University of Massachusetts Amherst, MA, USA

[3]Department of Chemical and Environmental Engineering, University of Cincinnati, Ohio, USA

Abstract

Current machine learning methods for discharge prediction often employ aggregated basin-wide hydrometeorological data (lumped modeling) for parametric and non-parametric training. This approach may overlook the spatial heterogeneity of river systems and their impact on discharge patterns. We hypothesize that integrating temporal-spatial hydrologic knowledge into the data modeling process (distributed/disaggregated modeling) can improve the performance of discharge prediction models. To test this hypothesis, we designed experiments comparing the performance of identical Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) models forced with either lumped or distributed features. We gather meteorological forcing and static attributes for the Mackenzie basin in Canada- a large and unique basin. Importantly, discharge performance is assessed out-of-sample with k-fold replication across gauges. Results reveal a 9.6% improvement in the mean Nash-Sutcliffe Efficiency (NSE) and a 4.6% improvement in mean Kling-Gupta Efficiency (KGE) when LSTMs are trained with distributed information. Notably, the models exhibit consistently unbiased predictions, with a negligible relative bias (RBias ≈ 0.0) across all predictions. These experiments and results demonstrate the importance of integrating topologically guided geomorphologic and hydrologic information (distributed modeling) in data-driven discharge predictions.

Plain Language

Accurate river discharge prediction is critical for sustainable water resource management and effective flood mitigation. Traditional methods often treat the entire river basin as a homogenous unit, neglecting crucial hydrologic and hydrometeorological variations that significantly impact water flow across different locations. This "lumped" approach can lead to inaccurate predictions. We propose a "distributed" modeling approach incorporating detailed information about the river basin's spatial heterogeneity. Applying this method to the Mackenzie River, a vast and complex river system in Canada, resulted in significantly more accurate discharge predictions compared to traditional lumped models. This confirms the critical importance of considering the river basin's spatial variability for better understanding and predicting water flow dynamics. Our work paves the way for enhanced water management strategies and improved flood preparedness by providing more precise and reliable discharge predictions.

Main Points

1. Current Machine Learning models rely on aggregated hydrometeorological data, ignoring the spatial heterogeneity inherent in river systems.
2. Incorporating topological-guided spatiotemporal hydrologic data can improve understanding of discharge dynamics within the river basin.

1. Introduction

The hydrologic cycles that generate river discharge are stochastic, complex, and non-deterministic systems characterized by processes and events whose dynamics depend on various direct (e.g., meteorological and environmental factors) and indirect (e.g., human interactions) inter-connected phenomena (Dimitriadis et al., 2021; Zounemat-Kermani et al., 2021). This complexity ensures that in situ monitoring via gauges is the best way to understand rivers: a direct measurement is best. However, continuous in situ monitoring of global rivers is challenging due to logistical difficulties, expense, and politics (Hannah et al., 2011; Wu et al., 2016; Gleason & Hamdan, 2017). Despite these challenges, the importance of monitoring river discharge cannot be overstated, as it aids in detecting climatic and environmental changes across time and space.

As a result of these challenges, process-based hydrology models are often deployed to estimate river discharge. Process-based models are rapidly scalable to different hydro-meteorological conditions and can explain and interpret underlying model performance. However, they are highly dependent on their calibrated parameters, which can degrade significantly when applied to rivers with different average discharges, seasonal variations, river widths, and geographical characteristics (e.g., Wagener et al., 2011; Arsenault et al., 2014; Pool & Seibert, 2021). This is important for modeling discharge in remote and developing regions where many assumptions must be made to achieve accurate predictions (Marshall et al., 2005; Thyer et al., 2009; Clark et al., 2016; Pilz et al., 2020). The needs and benefits of process-based models are an especially circular problem in ungauged basins between the need for robust models to replace gauges and the need for more gauged data to calibrate them. Watershed regionalization techniques such as spatial calibration, interpolation, and regression of basin and hydro-meteorological characteristics are often used to adopt these models and their parameters to ungauged basins (Hrachowitz et al., 2013; Pagliero et al., 2019; Belvederesi et al., 2022). Finally, models can simulate future projections based on physically realistic processes, i.e., 'what if' scenarios (Montanari & Koutsoyiannis, 2012; Basijokaite & Kelleher, 2021; Mai et al., 2022). This is especially important given the expected increase in the intensity and frequency of hydrological extremes due to climate change (Shrestha et al., 2021; Leng, Tang, and Rayburg, 2015; Tabari, 2020).

Despite their widespread adaptation and credibility in hydrology, process-based models have several limitations that hinder their ability to fully capture the complexities of real-world hydrologic systems. First, the dominant physical processes that govern water movement and transformation within a watershed exhibit significant temporal-spatial heterogeneity, reflecting variations in fluvial, geomorphological, and soil characteristics (Kirchner, 2006; McDonell et al., 2007; Sidle et al., 2017; Royall, 2021). This heterogeneity challenges the development of a single model structure that adequately represents all interacting processes across the diverse landscapes encountered in natural watersheds. Second, equifinality - the ability of multiple parameter settings to produce similar model outputs - obscures a proper process-based understanding of models with many parameters, making it difficult to discern the proper combination of underlying mechanisms responsible for hydrologic responses. Third, the limited spatial and temporal scales at which process-based models are typically developed and calibrated constrain their ability to effectively represent fast-evolving temporal-spatial variability in physical processes across different scales (Yoshida et al., 2022; Clark et al., 2015a, Clark2015b; Clark, 2016). This limitation hinders their applicability in assessing and predicting hydrologic response under changing climate and land-use scenarios. To address these limitations, modelers must incorporate heterogeneity and temporal-spatial variability of physical processes into their models or use remote sensing to gather more primary data (e.g., Oubanas et al., 2018; Xie et al., 2021; Tsai et al., 2021).

Therefore, gauges are the best means of monitoring rivers, but they are impractical to deploy globally. Hydrologic models and remote sensing are excellent tools, whether used separately or in combination, but they have unique challenges, especially in ungauged basins. How, then, can we best combine the richness of primary data with process-based hydrologic knowledge and sparse in situ data? We argue the answer can be found in machine learning (ML). Early ML studies (e.g., Hsu et al.,

1995) demonstrated the ability of feed-forward networks to outperform calibrated process-based models in predicting discharge across flow regimes. Recent studies (e.g., Ouyang et al., 2021; Feng et al., 2020; Feng 2021; Ma et al., 2021; Kratzert et al., 2019a; Kratzert2019b) have shown that Long short-term memory (LSTM) artificial neural networks can outperform process-based models in ungauged basins. Transfer learning (e.g., Zhuang et al., 2020; Tan et al., 2018; Long et al., 2017; Zamir et al., 2018; Ma et.al, 2021), which is analogous to regionalization (Kittel et al., 2020; Wang et al.,2021; Yang et al., 2020; Oudin et al., 2008), also shows promise in tuning ML models to well-measured basins and applying them to ungauged basins. At its core, ML for hydrology involves the automatic discovery of inherent temporal-spatial patterns in historical hydrological data. While current ML approaches have demonstrated improved streamflow predictions, they still have several limitations. First, ML models, especially deep learning models, are still relatively non-interpretable, meaning we can produce accurate streamflow hydrographs without knowing how or why they were produced or which combinations of hydrological processes improved the model's learning process. However, ML is moving toward improved interpretability (e.g., Marcinkevič & Vogt, 2020; Lundberg et al., 2017; Lundberg, 2020; Wanner et al., 2020; Lees et al., 2021), but for now, it remains a powerful predictive tool that often divides opinions in the traditionally process-based discipline of hydrology. Second, ML models are complex and require access to specialized computing, such as GPU clusters. Third, ML models typically require much more training data with stricter consistency requirements than hydrologists are used to working with: the amount of data needed for quality ML training far outstrips the amount needed to calibrate a model or remote sensing technique (Mastorakis, 2018; D'Amour et al., 2020; Seifert & Rasp, 2020).

Current ML for hydrology retrofits ML techniques to hydrological data. However, we argue that aspects of hydrologic modeling and remote sensing for hydrology can be easily implemented in an ML-driven hydrology framework to move toward a more hydrologically aware and purpose-built ML for the discipline. For instance, hydrologists have long known that distributed modeling—where inputs are spatiotemporally heterogeneous - outperforms lumped modeling - where inputs are spatiotemporally homogeneous (Baroni et al., 2019; Ntegeka et al., 2014; Fry & Maxwell, 2018; Tran et al., 2018; Muhammad et al., 2019; Dembele et al., 2020). Yet almost all previous ML in hydrology has been lumped modeling. Moving to distributed ML would allow known correlations between altitude and temporal-spatial variation in isotopic signatures of snowmelt, glacier melt, and rainwater to express themselves in the predictions (Immerzeel et al., 2010; Pokhrel et al., 2018; Scown et al. 2020; Fujita et al., 2008; Nepal et al. 2014; Pant & Semwal, 2021). This shift would require changes to the input structure of ML models, but it should improve them considerably. Further, since ML requires huge quantities of training data, remotely sensed inputs are the best way of obtaining this needed primary data in ungauged basins (Gleason and Durand, 2020) in conjunction with globally available climate model output currently used in ML-driven hydrology modeling (e.g., Larnier & Monnier 2020; Ma et al., 2021; Feng et al., 2020; Asanjan et al., 2018; Kratzert et al., 2019; Kratzert 2019b; Ouyang et al., 2021).

Therefore, we compare the impact of aggregating LSTM training data over the entire upstream basin (lumped modeling) against separating upstream basin information based on the Strahler River order system (distributed modeling) while holding the LSTM architecture and input data constant. This tests the hypothesis that creating a distributed LSTM model based on topologically organized geomorphologic and hydrologic information can improve discharge estimation performance in ungauged basins. We demonstrate this comparison in ungauged basins by training generalizable machine learning models in hydrologically similar basins to validation zones in ungauged basins. We also compare results to previously published LSTMs and a recent remotely sensed data assimilation product (Feng et al., 2021). Ultimately, we aim to show how tenets of hydrologic modeling improve ML in ungauged basins.

2. Data and Methods

2.1. Data

We tested our proposed ML approach on the Mackenzie basin (Figure 1). This basin covers approximately 1.8 million square kilometers and encompasses various climatic conditions, including mountainous, cold temperate, subarctic, and arctic zones. The Mackenzie River drains approximately one-fifth of Canada's total land area, including the Rocky and Mackenzie mountains and the Canadian Shield. It contains over 39,000 river reaches in the MERIT Basin River network (Lin et al., 2019), developed on the MERIT HYDRO topography data (Aziz and Burn, 2006; Yamazaki et al., 2019). We selected a subset (n = 69) of all gauge stations in the Mackenzie basin, limited to those with at least ten years of consistent daily gauge data available from Environment and Climate Change Canada (ECCC). These gauge data formed the basis of training and validation for our work.

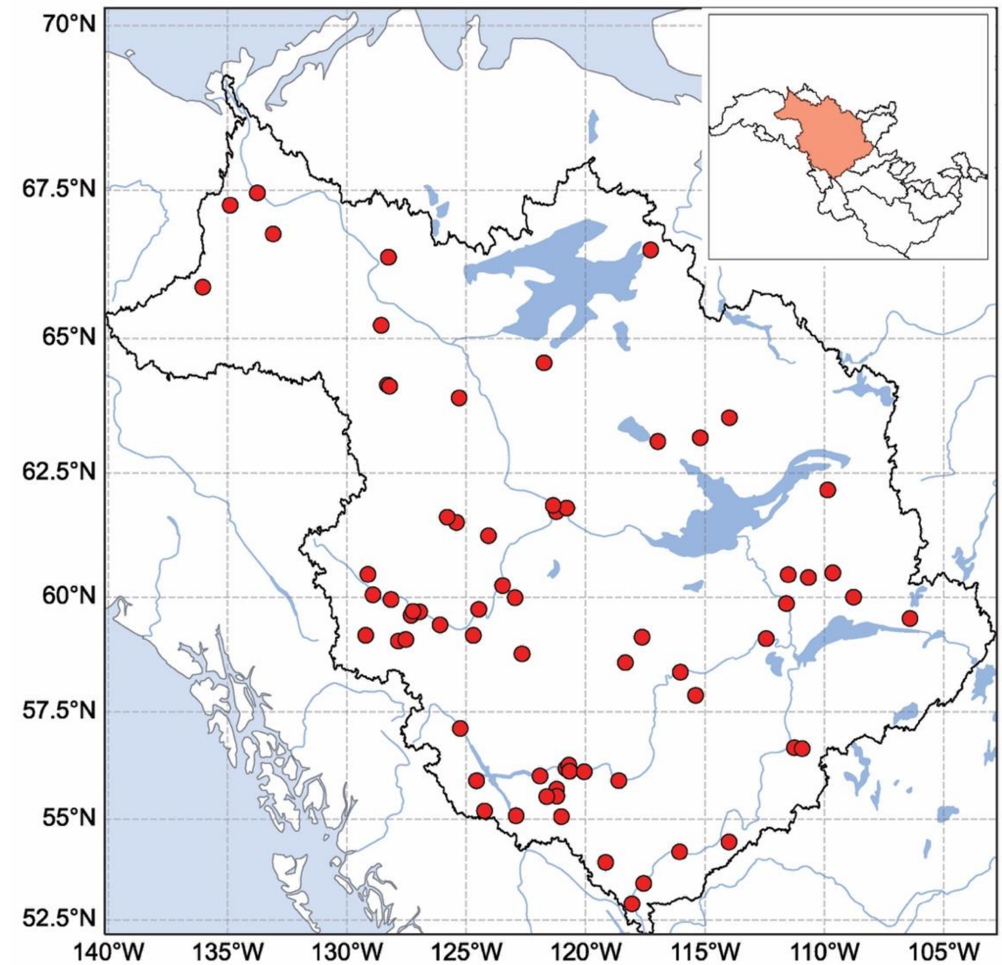

Figure 1: A Map showing the location of gauge stations (red circles) in the Mackenzie basin used in the study. Inset shows a map of the 20 biggest basins in Canada, including the Mackenzie Basin (shaded).

Our training data include both static and dynamic variables. Static variables, such as bed slope, sinuosity, and stream length, do not change over a few decades. Dynamic features, on the other hand, reflect changing hydrologic processes. We gathered daily data from 1981 to 2010, including simulated discharge and runoff from the GRADES database (Lin et al., 2019), reach averaged widths obtained from the Global Long-term river Width (GLOW) database (Feng et al., 2022), and climate model data. Climate data were from the Global Land Data Assimilation System (GLDAS)-2.1 model (Rodell et al., 2014; Beaudoing and Rodell, 2019) and included three hourly climate data gridded at 0.25 x 0.25 degrees resolution, which were downsampled to daily data. These data were downloaded from the Google Earth Engine platform (Gorelick et al., 2017). Previous studies have shown that stationary data are relatively easy to model with ML (e.g., Hosking, 1984; Dickey and Pantula, 1987). Appendix A lists all variables used in this study.

We include river width as one of the input features for all models used in this study. Previous studies have shown that river width has a strong correlation with river discharge (Gleason and Smith, 2014; Gleason et al., 2014; Hagemann et al., 2017; Brinkerhoff et al., 2019; Feng et al., 2019; Feng et al., 2021). However, Landsat-derived river widths are only available at best every 16 days, considering cloud cover and seasonality. This is not a problem for hydrological approaches, but long short-term memory (LSTM) models require training data without gaps (e.g., Bengio and Gingras, 1995; Che et al., 2018; Lim et al., 2021). Therefore, we impute a complete width record from the Landsat observations in the GLOW dataset (Feng et al., 2022). Imputation is a statistical process of determining and assigning replacement values for missing or invalid data points in a multivariate dataset by leveraging possible correlations between covariates (Buck, 1960; Jamshidian and Mata, 2007). Thus, we estimated missing width values using a regression model fitted with the remaining covariates in the dataset. We chose this imputation approach to retain river widths as a strong predictor of discharge.

To compare lumped and distributed ML approaches, we trained and tested our models only at gauges with at least five upstream reaches. This ensured that we had sufficient data to quantify the impact of upstream climatology factors on daily discharge at a given gauge station. We also limited our selection to gauges with at least ten years of daily discharge data. Preliminary tests indicated that this was the scale of data needed to train an LSTM model accurately without overfitting (Ying, 2009). Finally, we selected Strahler River orders with at least four gauge stations: order 4 (25 gauge stations), order 5 (23 gauge stations), order 6 (13 gauge stations), order 7 (4 gauge stations), and order 8 (4 gauge stations). This gave us a total of 69 gauge stations.

2.2. Sequential Learning Via LSTMs

Our ML models are based on the LSTM model architecture. This artificial neural network, introduced by Hochreiter and Schmidhuber in 1997, excels at processing sequential data, a hallmark of hydrometeorological and hydrologic time series. LSTMs have demonstrated remarkable success in diverse applications, including language modeling, video understanding, music transcription, and, crucially, discharge prediction for hydrology (e.g., Eck and Schmidhuber, 2002; Srivastava et al., 2015; Ghosh et al., 2016; Ouyang et al., 2021; Feng et al., 2020; Kratzert et al., 2019). Unlike standard neural networks that solely capture the spatial context of data, LSTMs are uniquely equipped to extract temporal and spatial information embedded within the training data (e.g., Yin et al., 2017; Wu and Prasad, 2017).

This ability to grasp the intricate interplay of spatial and temporal dynamics is paramount for accurately modeling hydrological processes. Structurally, an LSTM network comprises a series of identical recurrent neural networks, each building upon the information passed from its predecessor. This cascading architecture enables LSTMs to handle the sequential context inherent in historical data, particularly in hydrologic time series. Unlike traditional RNNs, LSTMs possess an inherent memory mechanism that allows them to retain information over extended periods, effectively overcoming the vanishing gradient problem (Chung et al., 2014; Hu et al., 2018). This memory capability empowers LSTMs to capture long-term temporal dependencies, where desired outputs depend on inputs presented far in the past (lookback window). This capability is critical for modeling physical processes unfolding at varying spatial resolutions, a characteristic of hydrological phenomena. Consequently, the lookback window size dictates how much information a model can learn about a particular physical process at any given time.

The LSTM network architecture can be implemented in either a unidirectional or bidirectional fashion (Graves & Schmidhuber, 2005; Siami-Namini et al., 2019; Fraiwan & Alkhodari, 2020). Unidirectional LSTMs process and encode features in a forward manner, sequentially learning information from each feature at each timestep $t = \{t[0], t[1], t[2], \ldots, t[n]\}$. However, they only utilize information from preceding timesteps (t_i-1) to enhance prediction at the current timestep (t_i). This unidirectional approach limits the model's ability to capture dependencies between features and information encoded in subsequent timesteps (t+1).

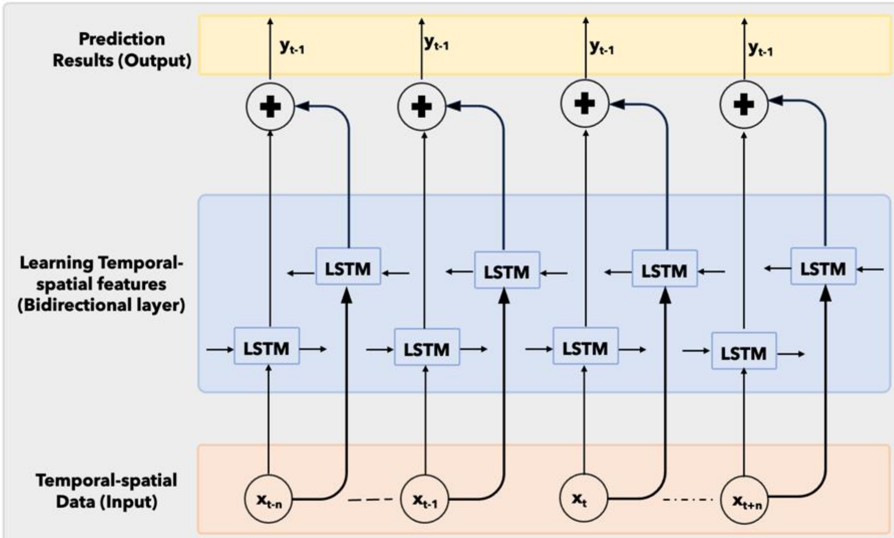

Figure 2: An architectural overview of a Bidirectional Long Short-Term Memory (Bi-LSTM) Network for time series prediction, showcasing the flow of temporal-spatial data through LSTM units in both forward and backward directions to enhance feature learning and improve prediction accuracy.

In contrast, bidirectional LSTMs combine two unidirectional LSTMs operating in opposite directions, as shown in Figure 2. The core LSM equations, shared by both forward and backward passes, are summarized as follows:

Forget gate: $f_t = \sigma\left(W_f \cdot \left[h_{prev}, x_t\right] + b_f\right)$ (1)

Input gate: $i_t = \sigma\left(W_i \cdot \left[h_{prev}, x_t\right] + b_i\right)$ (2)

Candidate state: $\tilde{C}_t = tanh\left(W_c \cdot \left[h_{prev}, x_t\right] + b_C\right)$ (3)

Cell state update: $C_t = f_t * C_{prev} + i_i * \tilde{C}_t$ (4)

Output gate: $o_t = \sigma\left(W_\circ \cdot \left[h_{prev}, x_t\right] + b_0\right)$ (5)

Hidden state update: $h_t = o_t * tanh(C_t)$ (6)

Where x_t is the input at timestep t , h_{prev} and C_{prev} are the previous hidden and cell states, respectively. Furthermore, the final output at each timestep t, represented as y_t , in a Bidirectional LSTM is the concatenation of the forward and hidden states represented as $y_t = \left[h_t{}^{(f)}; t_t{}^{(b)}\right]$. Finally, for the forward pass, $h_{prev} = h_{t-1}{}^{(f)}$ and $C_{prev} = C_{t-1}{}^{(f)}$ while for the backward pass, $h_{prev} = h_{t+1}{}^{(b)}$ and $C_{prev} = C_{t+1}{}^{(b)}$.

This architecture enables the model to learn encoded features forward and backward, simultaneously processing information from past and future timesteps. This bidirectional approach is particularly advantageous in hydrological modeling, where river discharge at the next timestep (t_i+1) can provide valuable context for improving prediction at the current timestep (t_i). For instance, knowledge of future rainfall patterns can inform the model about potential changes in river discharge.

Additionally, Bi-directional LSTMs have demonstrated superior prediction accuracy, efficiency, and stability in various applications (e.g., Ma et al., 2021; Atef and Eltawil, 2020; Siami-Namini, et al., 2019; Althelaya et. al., 2018), underscoring their versatility and effectiveness in handling diverse time-series data, robustness to noise, and long-term trends than uni-directional LSTMs. Finally, the structure of Bi-LSTMs offers more opportunities to improve performance through epochs and hyperparameter tuning. Recognizing the importance of this bidirectional relationship, we employed the bidirectional LSTM network architecture for our experiments.

To mitigate overfitting and enhance model generalizability, we employed several strategies. Regularization techniques (Bickel et al., 2006; Ghojogh & Crowley, 2019) impose constraints on the model's coefficient estimates (learned parameters), effectively preventing it from overfitting the training data and improving its generalizability to new data. This is achieved by adding a penalty term to the loss function - the measure of how well the model fits the training data. The penalty term typically increases with the complexity of the model, thus incentivizing simpler models that generalize better to unseen data. Additionally, we utilized dropout layers (Hinton et al., 2012; Wager et al., 2013) between each LSTM layer. These dropout layers randomly drop a certain percentage of connections during training, effectively preventing individual neurons from becoming overly reliant on specific features in the training data. This stochasticity enhances model generalizability by encouraging it to learn more robust and transferable data representations.

We opted for a bidirectional LSTM network with four layers. This architecture was chosen based on its ability to capture both temporal and spatial dependencies in the data, which is crucial for accurate hydrological modeling. Increasing the number of layers beyond four yielded minimal performance improvements, suggesting that the four-layer architecture was sufficient for capturing the relevant patterns in the data. Finally, we selected the Swish activation function (Ramachandran et al., 2017) for the output layer. This activation function has a smoother and more non-linear nature compared to ReLU - the most common activation function in ML, which enhances the flow of gradients through the network, contributing to improved performance. In addition to its computational efficiency, Swish also mitigates the dying ReLU problem, a phenomenon where ReLU neurons become inactive during training. By maintaining active neurons throughout the training process, Swish ensures that the network continues to learn and adapt. Furthermore, Swish offers efficiency advantages over ReLU, particularly when training deep neural networks with numerous layers, further reducing computational burdens. Overall, our hyperparameter tuning strategy and network architecture choices resulted in a robust and generalizable bidirectional LSTM model capable of accurately predicting hydrological time series.

2.3. Experiment Design

We hypothesize that an LSTM model trained with topologically organized distributed geomorphologic and hydrologic information should outperform the same LSTM that lumps the same training data. To this end, we estimate discharge in five ways: three experiments with identical ML models per Section 2.2 but with different organizations of the training data, and comparisons with two state-of-the-art approaches: an assimilation product (RADR- Feng et al., 2021) and a recently published LSTM model (PUB-LSTM- Kratzert et al., 2019). By organizing the training data consistently with topology, we aim to capture these spatial relationships and allow the ML model to learn more intricate patterns in the data. This approach differs from traditional methods that aggregate data into a single-point representation, which may lead to the loss of critical spatial information.

2.3.1. Experiments and literature comparisons

I. **At-station experiment**: We used dynamic and geomorphological static variables and climate data in a 25 km buffer around a given gauge station as input features to an ML model. These are the fewest possible data we can use to train any ML model that leverages temporal and spatial information encoded in historical data around a gauge station.

II. **Lumped experiment:** In addition to leveraging local information around the river outlet (the at-station experiment), we included integrated aggregated climate data from the largest possible upstream basin. Therefore, this experiment has static and dynamic variables from the prediction reach and averaged upstream climatology. This represents the approach taken by Ouyang (2021), Feng (2020; 2021), Ma (2021), and Kratzert (2019a; 2019b), among others.

III. **Distributed experiment:** Here, we expanded on the methodology used in experiments (I) and (II) by segmenting the upstream climate data according to the Strahler River order system. Although traditional clustering methods such as DBSCAN are better at clustering data (e.g., Brinkerhoff et al., 2020; Muhebwa et al., 2021), we chose the Strahler River ordering because it is an objective, consistent, and physically meaningful method for hierarchical clustering of hydrometeorological information, making it useful for various hydrological and geomorphological studies. This stratification was applied to dynamic variables in the entirety of the upstream basin. Thus, for a river system encompassing 'n' orders of upstream sub-basins, we introduced a more nuanced set of input features. Specifically, for each river order, we generated a distinct set of input features corresponding to each of the modeled hydrometeorological processes. The total number of additional input features was thus calculated as (n*x), where 'x' represents the total number of these processes. By averaging the data across all sub-basins per order (Figure 3), we were able to effectively capture the spatial variability of hydrological processes, resulting in more accurate river discharge predictions. The distributed approach aligns with those of Baroni et al. (2019) and Moore et. al. (1991), who emphasize the effectiveness of integrating data from various sources and considering spatial variability in hydrological processes, respectively. This method adheres to the principles of distributed data modeling, as it enhances river discharge prediction by incorporating the spatial distribution of hydrological processes, such as snowmelt, soil moisture, and evapotranspiration, across the watershed.

IV. **Comparison datasets:** We compare our approach against off-the-shelf results from the RADR model and a re-implementation of the PUB-LSTM model. The RADR (Feng et al., 2021) model was calibrated on data from 1984 to 1998 and assimilated with remotely sensed discharge data from 1984 to 2018 for the entire Arctic region (including the Mackenzie basin). Data assimilation in process-based modeling provides time-dependent distributed estimates that are updated whenever new data become available, i.e., the model's states are updated in response to how it performs at a given time (McLaughlin, 1995; Clark et al.,2008). We also implemented the PUB-LSTM model defined in Kratzert (2019) – a state-of-the-art unidirectional LSTM model. We trained this model with data defined in the lumped experiment but consolidated the data from all gauge stations into a single set (irrespective of the river order) before performing k-fold cross-validation. This means that each subset of stations in training/validation can contain data across any of the orders 4 to 8.

Our approach requires us to develop order-specific ML models given the rigid requirements for LSTM training. That is, each of our three ML experiments has five different LSTMs - one for each order from 4 to 8, as these orders contain sufficient training data. In order to apply our model to an ungauged basin, we would need first to identify the order of the river reach of interest and then select the appropriate order model to deploy. This means that our methods cannot predict flows in orders other than 4-8, but in return for this compromise, we can estimate flows quickly, efficiently, and accurately in ungauged basins, as shown below. Further, global datasets like those used to build our models already identify the order of all global rivers, so there is no additional computational burden on future users of these methods.

2.3.2. Validation design and applicability to ungauged basins

Our objective is to develop ML models that can accurately forecast daily river discharge in ungauged basins: watersheds lacking discharge monitoring stations (gauge stations). A standard approach in machine learning is to split the model's input data into training and validation sets by a particular ratio (Wu et al., 2013; Rácz et al., 2021; Shen et al., 2022). This implies that training and validation occur on data from the same distribution, known as independent and identically distributed (IID) data, where each random variable follows the same probability distribution, and all variables are independent. Consequently, it is simple to train models that perform well on training and validation data but struggle to generalize effectively to unseen data, a phenomenon known as overfitting. However, our goal is to transfer hydrological knowledge to ungauged basins. Therefore, we employ cross-validation to assess the performance of our ML models. Cross-validation (Stone, 1987; Rao et al., 2008; Refaeilzadeh et al., 2009; Berrar, 2019) is a technique where multiple ML models are trained on subsets of the available input data and evaluated on complementary subsets of the same data. This introduces heterogeneity in the training data by repeated resampling, thereby improving the ability of models to generalize to previously unseen data.

Since we use stream order as a unifying concept for our distributed modeling, we must build, train, and validate models that function per order. Previous studies (e.g., Feng et al., 2021; Kratzert el at., 2019; Sun et al., 2021) have either treated training data as a single entity, thereby making it easier to implement out-of-sample testing using k-fold validation (dividing data into groups of approximately equal sizes) or splitting training data by a given percentage (e.g., 70/30 split) for models trained and tested on IID data. Conversely, different Strahler River orders in our training data have unequal gauge stations (Table 1), making it difficult to implement an identical k-fold validation strategy. The imbalance in data across different orders can result in model uncertainties. We mitigate this by combinatorial training data selection for individual models in each order and by maintaining an equal number of stations (x) in each training and validation subset. This strategy of organizing training data maintains a relatively consistent volume of training data across the entire data strata. Consider a stream order with n stations; we can create sets of all possible combinations of stations in that order where each set contains x stations where x is any arbitrary number less than n. We chose x=3 for our experiment as a tradeoff between the minimum number of stations in each order (orders 7 and 8 each have 4 stations) and the computation time to train models for all subsets in each order. We then train a model on each subset and evaluate it on the complementary subsets of the same order. Therefore, in a basin with n=25 gauge stations, we try all combinations of x=3 training and (n-x)=22 validation stations. For stations with many subsets, i.e., orders 4 to 6 (Table 1), we randomly select 24 sets from all possible ${}^{n}C_{x}$ combinations to balance model compute time with statistical representativeness. Preliminary experiments to increase the size of the sets from 24 to 50 and 100 had no substantial improvement/degradation in model performance. Our results are presented as distributions of predictions across the complementary (validation) sets instead of reporting the results of individual or selected ML models that may perform particularly well or poorly at a gauge station. Therefore, the width of these distributions corresponds to the sensitivity of our three experiments to a particular combination of training/validation data.

Note that orders 7 and 8 have sufficient data to train and test but insufficient data to cross-validate. Also, remember that we build per-order ML models; thus, the performances here reflect only rivers of that order. Finally, given the available gauge data in the Mackenzie, we cannot predict in orders below 4 and above 8.

Table 1: Table showing the number of generated and contributed sets used for training in each Strahler River order.

Strahler order	Number of gauge stations (n)	Number of training stations per set (x)	Number of ungauged validation stations per set (n-x)	Possible training/validation combination sets (${}^{n}C_{x}$)	Number of selected sets used to report results
4	25	3	22	2300	24
5	23	3	21	1771	24

6	13	3	10	286	24
7	4	3	1	4	4
8	4	3	1	4	4

Ultimately and importantly, all results represent an ungauged case where validation is only done on the n-x stations not used in training and then tested in combinations per Table 1. This represents a common hydrologic situation where some gauge data are in a basin but not in areas where desired. Our methods would use the gauge data in hand, per order, to make estimates at all ungauged reaches of the basin of the same order. Here, we withhold gauge data to make that test, and each validation set, is completely independent of the others for a true ungauged case.

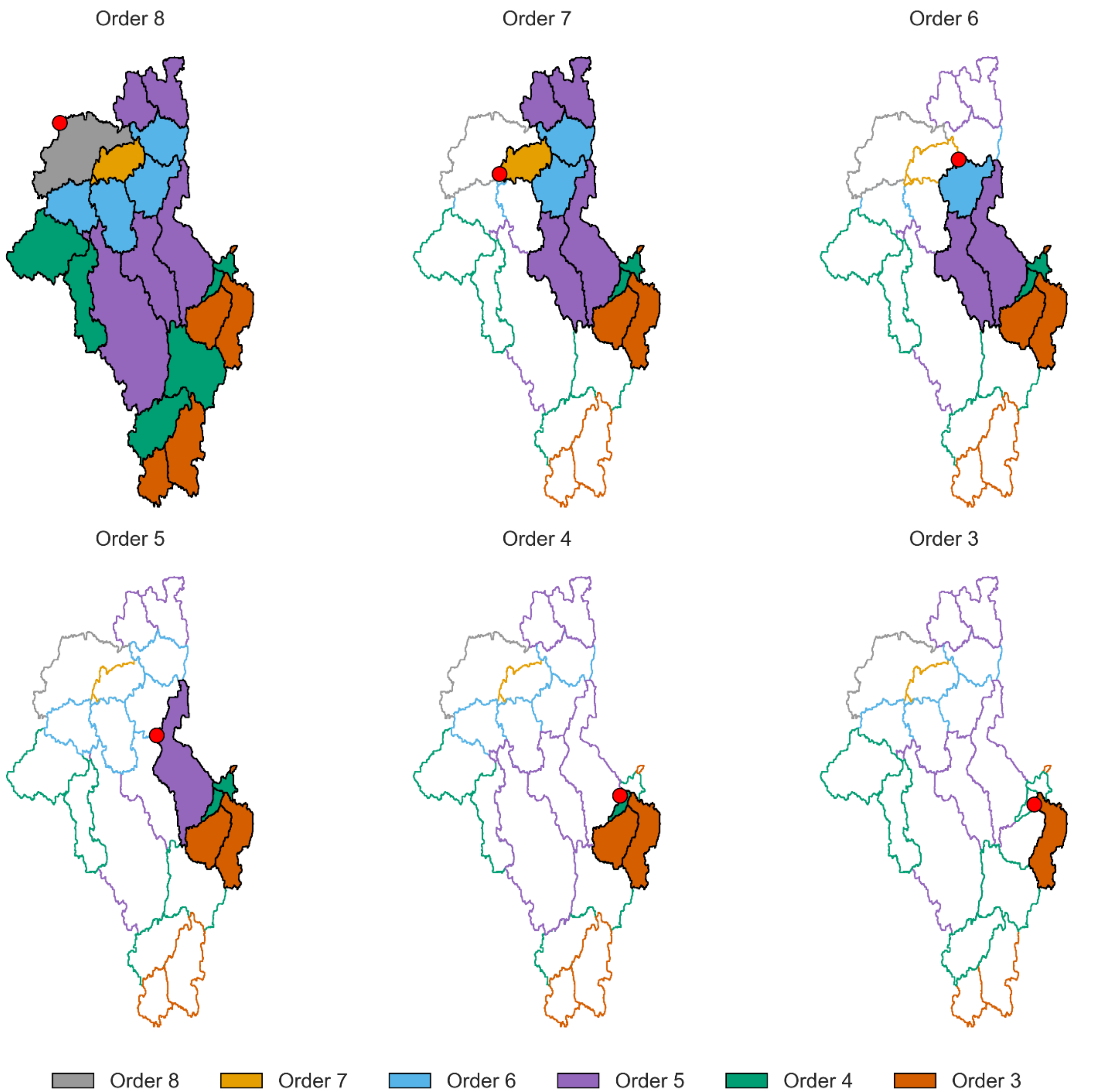

Figure 3: Schematic representation of an order eight basin network. The red circle represents the location of a gauge station on the delineated basin's outlet. At each hierarchical level, a single-order basin and its lower-order basins are selected (filled), while the remaining basins on the same level or not upstream of the selected basin within that level are ignored (hatched). This topological representation integrates the temporal-spatial variation of physical processes at different stages of a river network.

2.4. Evaluation Metrics

We report our results based on four major metrics used to evaluate the performance of discharge prediction models: Kling-Gupta Efficiency (KGE) (Gupta et al., 2009), Nash-Sutcliffe Efficiency (NSE) (Nash & Sutcliffe, 1970), Relative Bias, and Normalized Root Mean Squared Error (NRMSE).

$$\text{KGE} = 1 - \sqrt{(\gamma - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \quad (7)$$

where γ is the Pearson correlation between observed and actual discharge, α is the ratio of the standard deviation of actual vs. observed discharge, and β is the ratio of the mean of observed vs. actual discharge.

$$NSE = 1 - \frac{\sum_{i=1}^{N}\left(Q_i - Q_i^I\right)^2}{\sum_{i=1}^{N}(Q_i - \bar{Q})^2} \quad (8)$$

Where Q_i is the observed discharge at timestep ι and Q_i^I is the simulated discharge at timestep ι. These standard hydrology metrics assess different aspects of the hydrograph and errors in both timing and volume of water (e.g., Lin et al., 2019; Hagemann et al., 2017).

3. Results

Our experiments show that a distributed data modeling approach outperforms at-station and lumped approaches in training ML models for predicting discharge in ungauged basins. Figure 4 illustrates this outcome by presenting cumulative distribution functions (CDFs) for KGE and NSE across the experiments defined in Section 2.3.1. Note that all results pertain to ungauged cases where validation is performed exclusively on the n-x stations not used for training and then tested in combinations as per Table 1.

3.1. Predictions in Ungauged basins

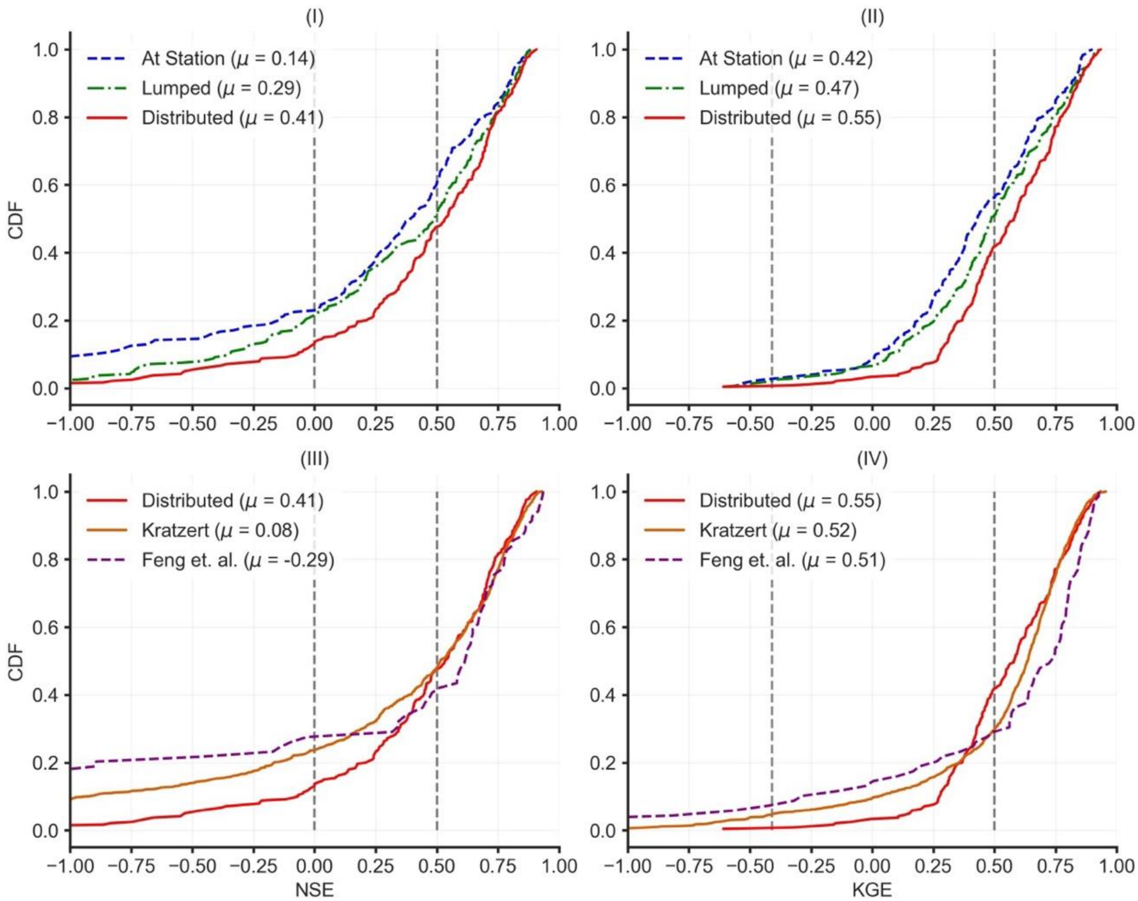

Figure 4: Cumulative distribution functions (CDFs) of NSE and KGE for defined experiments and selected benchmarks calculated from distributions across all Strahler River orders. Figures (I) and (II) compare the performance of models in the at-station and lumped experiments against the models trained with data from the distributed experiment. Figures (III) and (IV) compare the performance of models in the distributed experiment against two literature models: Feng et al. (2021) and Kratzert et al. (2019). A shift to the right indicates an improvement in model performance. Baseline models from the literature show lower skill than the ML here when all models perform poorly ($-\infty <$ NSE&KGE ≤ 0.0) but better performance when all models have good predictions ($0.5 <$ NSE &KGE ≤ 1.0). The distributed model outperforms the at-station and lumped models across the entirety of the results. CDFs are preferred because they represent the overall model performance across the entire test dataset.

Comparing results from at-station, lumped, and distributed experiments reveals that incorporating increasing amounts of upstream basin data universally enhances discharge estimation. In Figures 4(I) and (II), the rightward shift of the distributed experiment's cumulative distribution function (CDF) curve relative to those of the at-station and lumped experiments indicates performance improvement. Order level-specific models trained with minimal data (at-station experiment) achieve 77% positive NSE predictions and 92% positive KGE predictions. KGE and NSE values range between ($-\infty$, 1]; positive values are generally desirable, while negative NSE values indicate that the mean of observed values is a better predictor than the predicted value. KGE is a more 'forgiving' metric that takes a value of -0.41 when the mean hydrograph is predicted (NSE scores 0 in this case), as shown by Knoben, Freer, & Woods (2019). Incorporating aggregated upstream basin information (lumped experiment) into model training yielded no significant performance improvement (P-value > 0.05). However, training the same models with topologically organized data (distributed modeling) led to a 6.4-point increase in mean NSE and a 9.8-point increase in mean KGE.

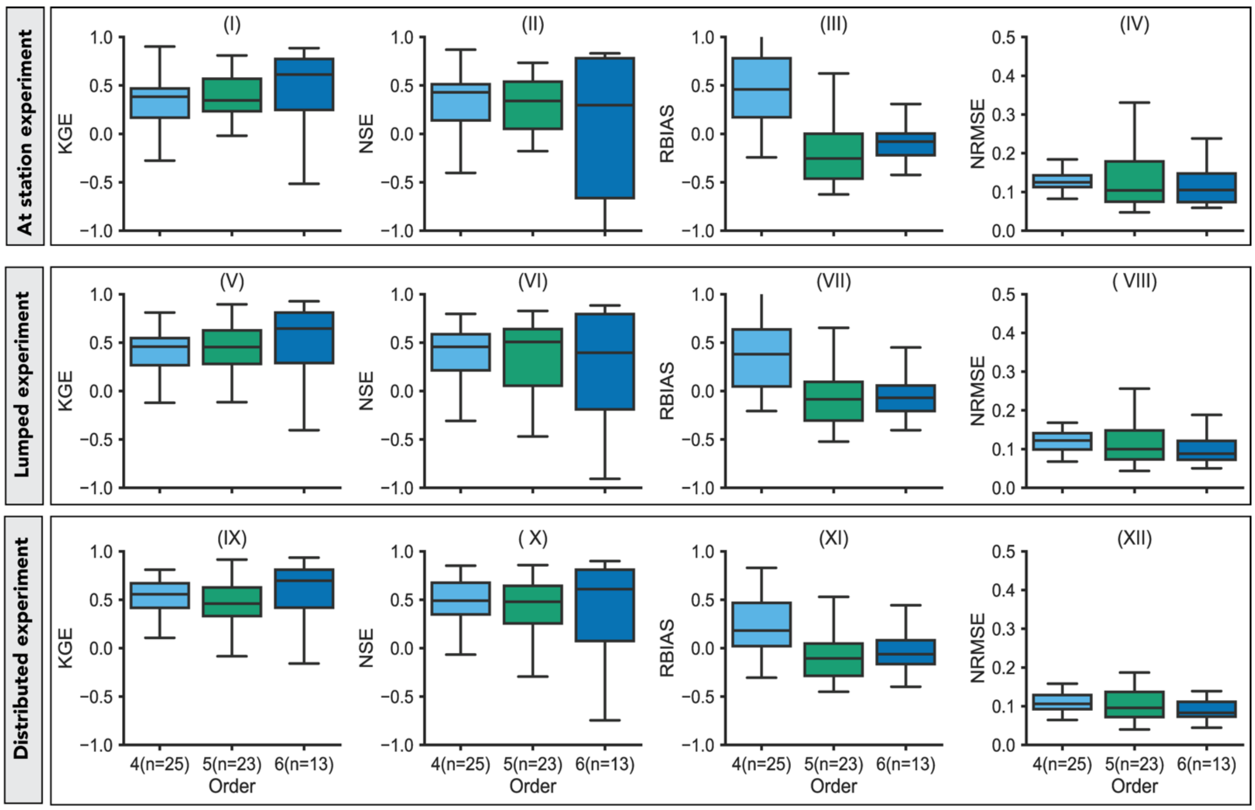

Figure 5: Top to Bottom: Distribution comparisons of selected metrics on held-out predictions for at-station (I-IV), lumped (V-VII), and distributed (IX-XII) experiments. Note that distributions for seventh and eighth orders are not included due to the limited number of gauge stations in the training set. Figure S1 shows a distribution comparison across all experiments and literature models.

When ML models were trained with the least possible data (at-station experiment), i.e., Figure 5(I)-(IV), we observed a significant ($p \leq 0.05$) improvement in median KGE from 0.38 to 0.61 as basin size increased from order 4 to order 6, which is observed across all experiments. NSE, however, was relatively constant across orders, with a noticeable increase in the interquartile range (IQR) for the largest order with ten stations. When we compared similar spatial orders across the three experiments (columns in Figure 5) - at-station, lumped, and distributed experiments - we observe an improvement in both NSE and KGE scores as orders increase and more information is added to the data modeling process. Consider Figures 5(I), (V), and (IX), KGE improved from 0.38 to 0.56 in the fourth order, 0.34 to 0.46 in the fifth order, and 0.61 to 0.69 in the sixth order, from at station to distributed experiments respectively. Likewise, we observe an equivalent improvement in NSE, i.e., Figures 5(II), (VI), and (X) from 0.42 to 0.48 in the fourth order, 0.34 to 0.47 in the fifth order, and 0.29 to 0.60 in the sixth order. Additionally, these skill gains are accompanied by consistently unbiased predictions with negligible relative bias (RBias ≈ 0.0) across all models and orders.

When we compare the performance of literature models on an order level basis (Figure S1), we observe a much more substantial improvement in performance as the number of sub-basins increases. The RADR model (Feng et al., 2021) had the most noticeable improvement in skill scores, with median KGE improving from 0.63 in the fourth order to 0.77 in the sixth order, while median NSE improved from 0.47 to 0.58 in the corresponding orders. On the other hand, Kratzert et al. (2019) demonstrated an improvement in KGE from 0.68 in the fourth order to 0.72 in the sixth order but a decline in NSE scores from 0.72 in the fourth order to 0.56 in the sixth order.

We compare the results of the distributed experiment against model predictions of both a reimplementation of an ML model proposed by Kratzert et al. (2019) with minor modification and off-the-shelf results of a remote sensing data assimilation over the same basin and time from Feng et al. (2021), i.e., Figure 4(III)-(IV). Performance across all three methods was largely similar but with noticeable differences in 'good' and 'bad' regions of skill, which is more pronounced with the KGE metric (that rewards correlation per Eq. 1). The distributed modeling approach has 13% of all NSE values and 3% of all KGE values as negative predictions across the entire experiment, the Kratzert et al. model has 22% of all NSE values and 7% of KGE values as negative predictions across all orders, and the Feng et al. model has 28% of all NSE values and 13% of all KGE values as negative predictions across all Strahler river orders. Thus, the distributed LSTM we propose here produces fewer 'bad' hydrographs that are worse than the mean compared to the other two methods.

However, when all models perform well, the two literature models outperform our LSTM, although performance is quite similar ($p > 0.05$).

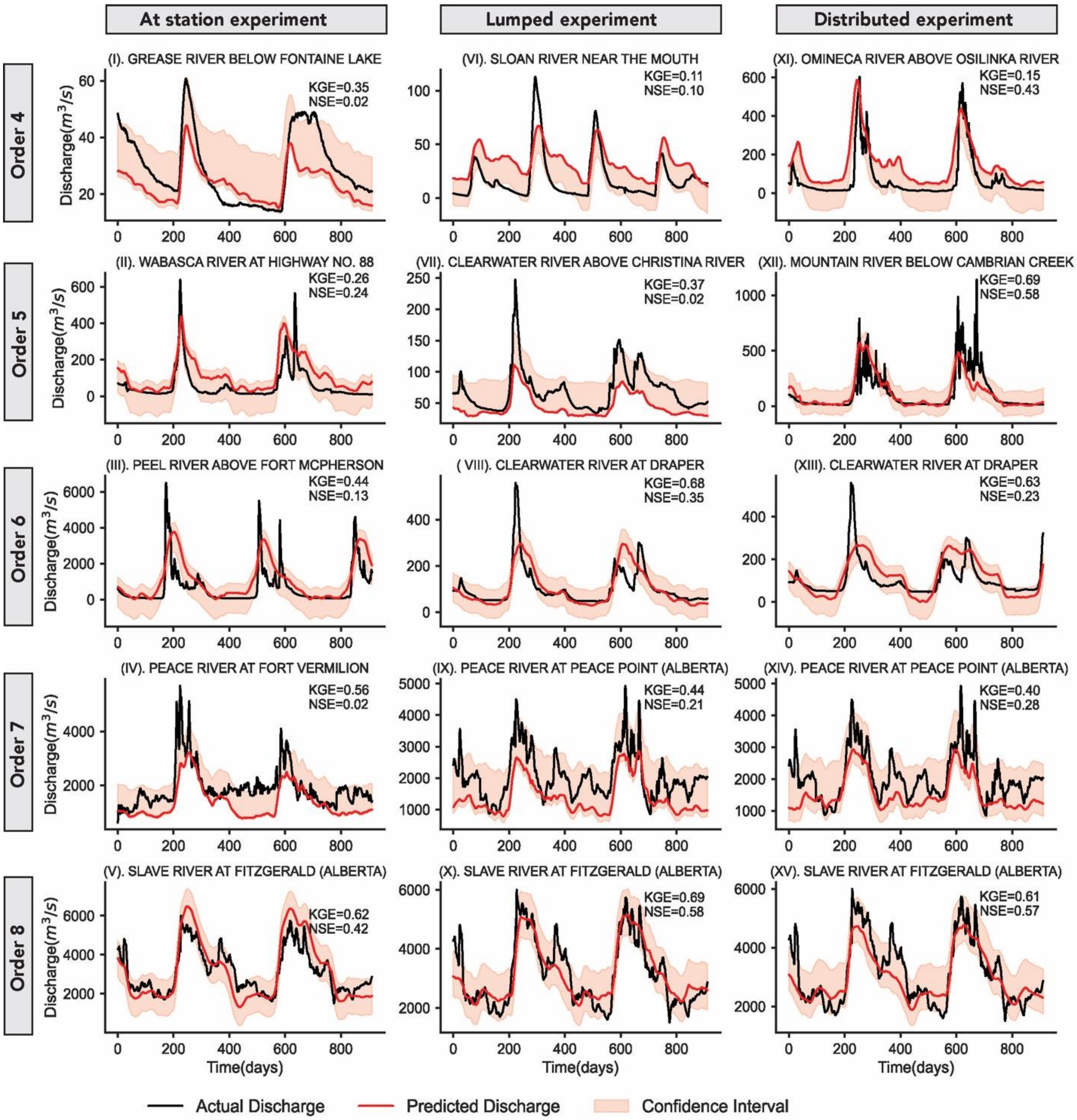

Figure 6: Representative hydrographs showing randomly selected models with $0.0 < \text{NSE} \leq 0.6$ in each of the experiments: At-station (left), lumped (middle), and distributed (right) experiments across the defined orders, i.e., from order 4 (top) to order 8 (bottom). Here, we plot hydrographs for the first 2.5 years.

Figure 6 shows hydrographs of randomly selected ML models in orders 4 to 8 whose NSE scores lie between 0.0 and 0.6. Here, we use $0.0 < \text{NSE} \leq 0.6$ as a representative average performance range across the prediction distribution. Across individual experiments, the models' confidence to re-create discharge increases as sub-basins increase. For example, absolute relative bias (|RBIAS|) improves from 0.24 to 0.007 in the station experiment, 0.80 to 0.002 in the lumped experiment, and

0.82 to 0.06 in the distributed experiments, as the number of sub-basins increases (i.e., from fourth to eight order). Note that as relative bias approaches zero, model predictions become increasingly unbiased and reliable, thereby enhancing the confidence and reliability with which they can inform impactful water management decisions. Nevertheless, notable differences in hydrographs remain across the defined experiments. Consider the fourth order across the three experiments, normalized root mean squared error (NRMSE) reduces from 0.17 in the at-station experiment to 0.09 in the distributed experiment, indicating an improvement in model performance in response to additional hydrological information in the training data.

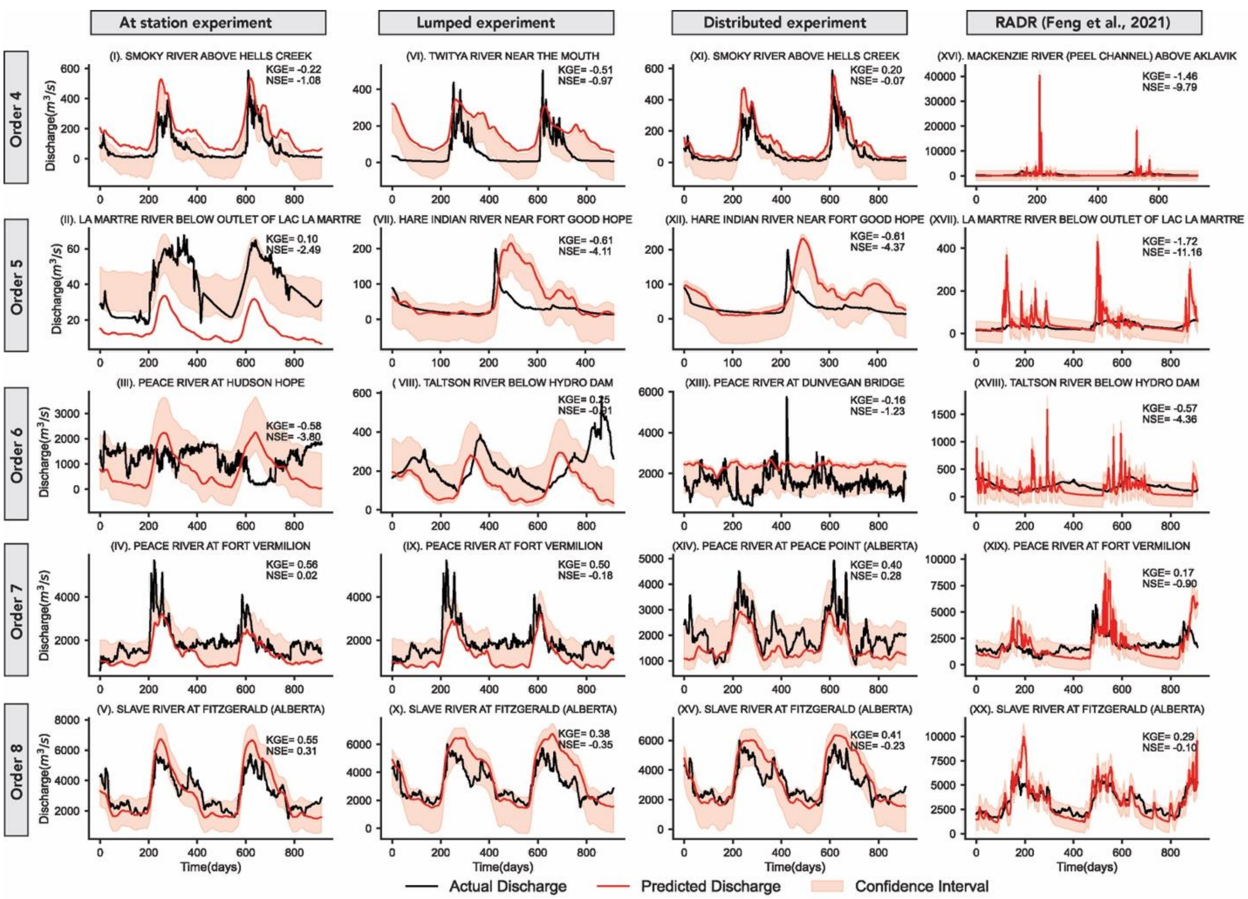

Figure 7: Left to right: Representative hydrographs showing the worst performing ML models in each of the experiments and the non-ML literature model; At station experiment, lumped experiment, distributed experiment, and RADR model (Feng et al., 2021) across the defined orders, i.e., from order 4 (top) to order 8 (bottom). The RADR model overestimates peak flows and underestimates base flows in lower orders. Here, we plot hydrographs for the first 2.5 years.

Figure 7 represents hydrographs of worst-performing models with NSE scores below 0.0 ($-\infty <$NSE <0.0) across orders 4 to 8. This NSE range encompasses the entirety of potentially bad model predictions within the predicted discharge distribution, providing a comprehensive view of model shortcomings across the defined experiments. We observed that across all experiments, models within orders 4 and 5 demonstrated a significant difficulty in precisely reproducing discharge hydrographs, indicated by a high uncertainty in model predictions. Interestingly, higher-order models, specifically those of orders 7 and 8, exhibited a consistent ability to capture the underlying trend of the actual discharge despite persisting uncertainty in the finer details. The underwhelming performance of RADR models reinforces the observation in Figure 4: process-based models, while valuable for capturing established physical and hydrologic laws, often struggle to adapt to real-world scenarios marked by significant, unpredictable fluctuations; that is, their inflexible structure hinders their ability to adapt to these deviations, leading to less accurate discharge predictions.

Different geographical and climatic regions have dominant physical processes at different temporal-spatial scales. Results in section 3.1 showed that integrating this knowledge of temporal-spatial variations (distributed modeling) improved the discharge prediction of ML models.

Earlier studies (e.g., Kahraman et al., 2021; Dey & Fuentes, 2020) showed that longer lookback windows with a longer ‘memory’ of past hydrologic conditions improve model performance. However, this performance improvement comes with increased computational power and time. To further evaluate the impact of the lookback window on model performance, we repeat experiments defined in section 2.3.1 with varying lookback window sizes of 30, 90, 180, and 270 days. Pairwise comparisons of distributions for both at-station and lumped experiments indicate that the size of the lookback window has no impact on model performance (P-value $>$ .05). However, there is a significant difference between distributions of results for lookback pairs (30, 90), (30, 270) days of the distributed experiment (P-value $\leq$ .05).

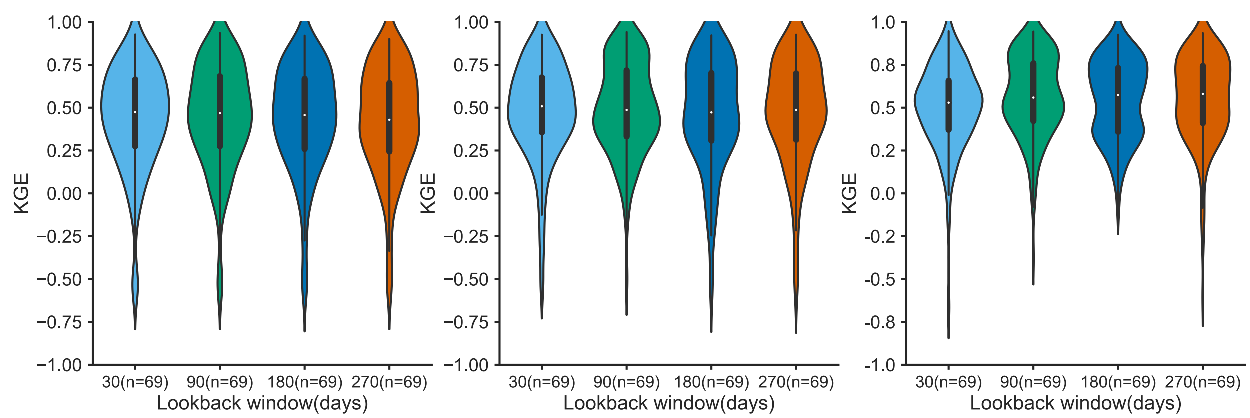

Figure 8: Left to right: Pairwise comparison of KGE distributions with varying lookback window sizes and corresponding statistical significance tests across the three experiments. Inter-experiment comparisons show that distributions of lookback for at-station and lumped experiments are similar. At the same time, there is an observable difference in the distributions of lookback windows of the distributed experiment.

4. Discussion

We confirm our hypothesis that distributed modeling outperforms lumped modeling for our architecture, but that by Kratzert et al's (2019) lumped LSTM has superior performance to our distributed model when all models predict well. Our model outperformed the literature models when all models produced poor hydrographs (Figure 4, 7), and our skill scores have a much higher 'floor' than the literature models. However, we have a lower 'ceiling' as well - the literature models' performance exceeds ours when all models perform well, although the difference between this study and the literature is much more pronounced at lower skill (where our results improve skill). We attribute the superior performance of the Feng et al. RADR product at the high skill areas to three factors. First, RADR was calibrated on remotely sensed data drawn from the same distribution (independent and identically distributed data). Second, the model was assimilated on heterogeneous data from the entire Arctic region (as compared to our models trained on data from only the Mackenzie basin). Finally, the superior performance of process-based models can be attributed to their deep-rooted understanding of hydrologic, geomorphologic, and hydrometeorological processes. This comprehensive knowledge enables process-based models to effectively simulate the complex and interconnected interactions between various processes within a river basin. This theoretical foundation grounded in the principles of hydrology and river system dynamics not only enhances their predictive accuracy but also ensures the physical consistency and interpretability of the results. We attribute the Kratzert et al. model's better performance to a different training strategy as compared to the distributed experiment. Whereas models in the distributed experiment were trained and validated on order-specific training data, the Kratzert et al. model used a k-fold validation strategy and trained on the entire spectrum of data (all 69 gauge stations), following the original model implementation proposed by the authors. This strategy ensured the model was trained on more diverse data, enhancing its generalization to previously unseen data. This also offers the advantage of enabling flow prediction for all rivers within the basin. However, in our study, we didn't follow Kratzert et al. because our distributed experiment exhibits two notable advantages: first, when all models performed poorly (Figure 6), models in the distributed experiment still performed better than literature models. In general, we attribute poor performance (poor generalization) to limited training data, a reality for much of the world where training data are rare, nonexistent, or proprietary (Gleason & Smith, 2014). Second, acknowledging the influence of physical processes on the hydrologic cycle, the existence of these processes at different spatial resolutions, and their varying dominance across different geographical regions, order-specific models in the distributed experiment firmly integrate this hydrological knowledge in the data modeling process as compared to the literature models. One possible explanation of why models in the distributed experiment perform better when all models have low skill scores is that despite limited training data, these models are better than literature models at leveraging the high correlation between temporal-spatial variability and physical processes to extract meaningful patterns in the training data. This capability is particularly relevant when considering discharge estimation on a global scale, where well-hydrologically mapped regions are scarce.

We also observe that while RADR has the highest skill score when all models perform well, it also has the lowest skill scores when all models generally perform poorly (Figures 4(III) and (IV)). One possible explanation is that process-based models, which rely heavily on established physical and hydrological principles, often struggle to adapt to poorly understood scenarios or environments with significant uncertainties. This limitation is further compounded by their potential inability to capture emergent phenomena and human impacts - complex interactions or patterns that arise spontaneously and are not yet fully understood or integrated into existing hydrologic theories. Thus, the scientific robustness of process-based models, while grounded in established principles, can inadvertently narrow their score, hindering their ability to dynamically adapt to and accurately model these evolving and multifaceted riverine environments. Thus, while each model possesses unique advantages, a distributed data modeling approach offers a more applicable and scalable solution for global-scale discharge estimation.

Further, we observed that even the best-performing models in the at-station experiment (Figure S2) fail to recreate medium to high peak discharges by a considerable margin in the lower orders. This is not surprising, given that peak discharges are a function of events in the upstream basin, e.g., after maximum rain intensity or melting of accumulated snow (Volpi et al., 2018; Jones, 2000; Furey & Gupta, 2005; Kabeja et al., 2020), information that is not included in the training data. Indeed, the impact of the knowledge of events in the upstream basin becomes more prevalent as more information is added to the training data. This is visible in the hydrographs of both the lumped and distributed experiments in Figure 6 (average-performing) and Figure S2 (best-performing), in which models recreate most of the peak discharges (or miss them by a small margin). To verify this, we aggregated the top 10 peak flows of each station. We observed that the mean error of the best-performing models across each experiment (defined as the average of the top 10 peaks in each order) reduced from 2901.58 m^3s^{-1} in the lumped experiment to 2518.74 m^3s^{-1} in the distributed experiment and observed a similar pattern between the same orders across the two experiments.

We attribute the high correlation between pairs of lookback windows for both the at-station and lumped experiments to the fact that both experiments ignore spatial variations of events in the upstream basin (physical processes). On the other hand, we attributed the differences across the lookback window pairs of the distributed experiment to the integration of knowledge of both temporal and spatial variations of physical processes in the data modeling process, indicating that the impact of dominant physical processes on model performance is prevalent at different temporal-spatial scales. We found that at various temporal scales (with similar spatial scales), a lookback of as little as 90 days was enough to capture temporal information encoded in the training data. As such, we saw no additional value in longer lookback windows, although this could be different for different geographical regions and data.

We do not report individual skill scores of the seventh and eighth orders (Figure 5) due to the limited number of gauge stations (Table 1). Further, data availability limits the minimum number of gauge stations (x) to include in each subset, which reduces data heterogeneity for each order-specific model. For instance, on order 8, x=3 represents 75% of the data as training, while on order 4, x=3 represents only 12% (Table 1). We chose to keep x constant instead of a constant train/test ratio because this allows sharing model hyper-parameters (and structure) and makes it easier to compare the results of models trained on the same number of gauge stations (x) across different orders of the same experiment. Finally, randomly selecting 24 subsets from all possible combinations for spatial resolutions with many gauge stations (Table 1) is not the best representation of complete data heterogeneity. However, we experimented with up to 100 validation sets and observed no substantial change in model performance. Future work could explore all possible combinations of training and testing and/or vary x to learn the effect of increasing the training sample.

ML has demonstrated encouraging results in global river discharge predictions and holds the potential to address many existing challenges in hydrology (Shen, 2018; Nearing et al., 2021). However, these advancements have primarily relied on lumped data modeling techniques, which overlook the temporal-spatial variations of physical processes that govern the hydrologic cycle. We have demonstrated that incorporating this knowledge into training data modeling (via our distributed experiments) can further improve the performance of ML models, particularly for predictions in ungauged basins. Further, we have shown that even with limited data, a distributed modeling strategy could provide improved predictions (especially in ungauged basins) than any of the existing benchmarked models. We acknowledge that literature models from ML and hydrologic modeling represented by Kratzert et al. (2019) and Feng et al. (2021) offer unique advantages that can deepen our understanding of global discharge as a proxy for assessing the cascading impacts of climate change on water resources. Therefore, leveraging distributed modeling could further improve the performance of other ML approaches.

5. Conclusion

In this work, we have demonstrated the importance of distributed data modeling in improving the performance of ML models for discharge prediction in ungauged basins. Further, we leverage topologically guided river hierarchies as a proxy for understanding the impact of temporal resolution (lag window) on model performance, specifically examining how much historical context is necessary to improve model performance. We showed that as spatial resolution increases, model performance improves in response to granular hydrological information. This makes our proposed method more applicable for predicting discharge for most global river basins with limited to no data. Our experiments and results demonstrate the importance of integrating hydrological and geographical differences in the data modeling process, a notion that has, until now, been largely ignored when building data-driven hydrology models. With the recent launch of the SWOT mission that will provide more consistent and granular hydrological information on global rivers, our proposed approach has the potential to improve methods for predicting river discharge on a global scale and, as a result, explore the complex, cascading, and often hidden ways that climate change alters global water systems. However, while we did not specifically identify which physical processes are dominant at varying spatial scales, this opens up questions in future work on quantifying the temporal-spatial contribution of distinct features towards model performance and overall interpretability and explainability of ML models in hydrology and physical sciences in general.

Acknowledgment

This work was supported by NASA SWOT Science Team Grant 80NSSC20K1141 that supported C.J Gleason, D. Feng, J. Taneja, and A. Muhebwa. A. Muhebwa was also supported by The Prince Albert II of Monaco Foundation through the Intergovernmental Panel on Climate Change (IPCC) scholarship. D. Feng was also funded by NASA's Terrestrial Hydrology Program 80NSSC21K0977. The main author thanks Craig B. Brinkerhoff for guidance in the initial stages of the project.

Data and Code Availability Statement

Code related to this study can be found online at https://github.com/amuhebwa/rivers_ML . Data used in this study is available at https://zenodo.org/record/6604724 . Data for the RADR model is available at (https://zenodo.org/record/5604980)

References

Agostinelli, F., Hoffman, M., Sadowski, P., & Baldi, P. (2014). Learning activation functions to improve deep neural networks. arXiv preprint arXiv:1412.6830.

Akbari Asanjan, A., Yang, T., Hsu, K., Sorooshian, S., Lin, J., & Peng, Q. (2018). Short-term precipitation forecast based on the PERSIANN system and LSTM recurrent neural networks. Journal of Geophysical Research: Atmospheres, 123(22), 12-543.

Andreadis, K. M., Brinkerhoff, C. B., & Gleason, C. J. (2020). Constraining the assimilation of SWOT observations with hydraulic geometry relations. Water Resources Research, 56(5), e2019WR026611.

Arsenault, R., & Brissette, F. P. (2014). Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches. Water Resources Research, 50(7), 6135-6153.

Aziz, O. I. A., & Burn, D. H. (2006). Trends and variability in the hydrological regime of the Mackenzie River Basin. Journal of hydrology, 319(1-4), 282-294.

Baroni, G., Schalge, B., Rakovec, O., Kumar, R., Schüler, L., Samaniego, L., ... & Attinger, S. (2019). A comprehensive distributed hydrological modelling intercomparison to support process representation and data collection strategies. Water Resources Research, 55(2), 990-1010.

Bengio, Y., & Gingras, F. (1995). Recurrent neural networks for missing or asynchronous data. Advances in neural information processing systems, 8.

Berrar, D. (2019). Cross-Validation.

Basijokaite, R., & Kelleher, C. (2021). Time-Varying Sensitivity Analysis Reveals Relationships Between Watershed Climate and Variations in Annual Parameter Importance in Regions With Strong Interannual Variability. Water Resources Research, 57(1), e2020WR028544.

Beaudoing, H. and M. Rodell, NASA/GSFC/HSL (2019), GLDAS Noah Land Surface Model L4 3 hourly 0.25 x 0.25 degree V2.0, Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC) Accessed: Aug, 2021

Belvederesi, C., Zaghloul, M. S., Achari, G., Gupta, A., & Hassan, Q. K. (2022). Modelling river flow in cold and ungauged regions: a review of the purposes, methods, and challenges. Environmental Reviews, 99(999), 1-15.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of machine learning research, 13(2).

Bickel, P. J., Li, B., Tsybakov, A. B., van de Geer, S. A., Yu, B., Valdés, T., ... & van der Vaart, A. (2006). Regularization in statistics. Test, 15(2), 271-344.

Birhanu, D., Kim, H., & Jang, C. (2019). Effectiveness of introducing crop coefficient and leaf area index to enhance evapotranspiration simulations in hydrologic models. Hydrological Processes, 33(16), 2206-2226.

Brinkerhoff, C. B., Gleason, C. J., & Ostendorf, D. W. (2019). Reconciling at-a-station and at-many-stations hydraulic geometry through river-wide geomorphology. Geophysical Research Letters, 46(16), 9637-9647.

Brinkerhoff, C., Gleason, C., Feng, D., & Lin, P. (2020). Constraining remote river discharge estimation using reach-scale geomorphology. Water Resources Research, 56(11), e2020WR027949.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Journal of the Royal Statistical Society: Series B (Methodological), 22(2), 302-306.

Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. Scientific reports, 8(1), 1-12.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modelling. arXiv preprint arXiv:1412.3555.

Claesen, M., & De Moor, B. (2015). Hyperparameter search in machine learning. arXiv preprint arXiv:1502.02127.

Clark, M. P., Rupp, D. E., Woods, R. A., Zheng, X., Ibbitt, R. P., Slater, A. G., ... & Uddstrom, M. J. (2008). Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. Advances in water resources, 31(10), 1309-1324.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., ... & Rasmussen, R. M. (2015). A unified approach for process-based hydrologic modelling: 1. Modelling concept. Water Resources Research, 51(4), 2498-2514.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., ... & Marks, D. G. (2015). A unified approach for process-based hydrologic modelling: 2. Model implementation and case studies. Water Resources Research, 51(4), 2515-2542.

Clark, M. P., Schaefli, B., Schymanski, S. J., Samaniego, L., Luce, C. H., Jackson, B. M., ... & Ceola, S. (2016). Improving the theoretical underpinnings of process-based hydrologic models. Water Resources Research, 52(3), 2350-2365.

Cramer, W., Guiot, J., Fader, M., Garrabou, J., Gattuso, J. P., Iglesias, A., ... & Xoplaki, E. (2018). Climate change and interconnected risks to sustainable development in the Mediterranean. Nature Climate Change, 8(11), 972-980.

Cui, X., Guo, X., Wang, Y., Wang, X., Zhu, W., Shi, J., ... & Gao, X. (2019). Application of remote sensing to water environmental processes under a changing climate. Journal of Hydrology, 574, 892-902.

Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., & Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3712-3722).

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., ... & Sculley, D. (2020). Underspecification presents challenges for credibility in modern machine learning. arXiv preprint arXiv:2011.03395.

Dey, S., & Fuentes, O. (2020). Predicting solar x-ray flux using deep learning techniques. In 2020 international joint conference on neural networks (ijcnn) (pp. 1–7).

Dembélé, M., Hrachowitz, M., Savenije, H. H., Mariéthoz, G., & Schaefli, B. (2020). Improving the predictive skill of a distributed hydrological model by calibration on spatial patterns with multiple satellite data sets. Water resources research, 56(1), e2019WR026085.

Dickey, D. A., & Pantula, S. G. (1987). Determining the order of differencing in autoregressive processes. Journal of Business & Economic Statistics, 5(4), 455-461.

Dimitriadis, P., Koutsoyiannis, D., Iliopoulou, T., & Papanicolaou, P. (2021). A global-scale investigation of stochastic similarities in marginal distribution and dependence structure of key hydrological-cycle processes. Hydrology, 8(2), 59.

Durand, M., Gleason, C. J., Garambois, P. A., Bjerklie, D., Smith, L. C., Roux, H., ... & Vilmin, L. (2016). An intercomparison of remote sensing river discharge estimation algorithms from measurements of river height, width, and slope. Water Resources Research, 52(6), 4527-4549.

Durand, M., Chen, C., de Moraes Frasson, R. P., Pavelsky, T. M., Williams, B., Yang, X., & Fore, A. (2020). How will radar layover impact SWOT measurements of water surface elevation and slope, and estimates of river discharge?. Remote Sensing of Environment, 247, 111883.

Eck, D., & Schmidhuber, J. (2002). A first look at music composition using lstm recurrent neural networks. Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale, 103, 48.

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

Feng, D., Lawson, K., & Shen, C. (2021). Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modelling and soft data. Geophysical Research Letters, 48(14), e2021GL092999.

Feng, D., Gleason, C. J., Yang, X., Allen, G. H., & Pavelsky, T. M. (2022). How have global river widths changed over time?. Water Resources Research, 58(8), e2021WR031712.

Feng, D., Gleason, C. J., Lin, P., Yang, X., Pan, M., & Ishitsuka, Y. (2021). Recent changes to arctic river discharge. Nature communications, 12(1), 1–9.

Feng, D., Gleason, C. J., Yang, X., & Pavelsky, T. M. (2019). Comparing discharge estimates made via the bam algorithm in high-order arctic rivers derived solely from optical cubesat, landsat, and sentinel-2 data. Water Resources Research, 55(9), 7753–7771.

Fraiwan, L., & Alkhodari, M. (2020). Investigating the use of uni-directional and bi-directional long short-term memory models for automatic sleep stage scoring. Informatics in medicine unlocked, 20, 100370.

Frasson, R. P. D. M., Pavelsky, T. M., Fonstad, M. A., Durand, M. T., Allen, G. H., Schumann, G., ... & Yang, X. (2019). Global relationships between river width, slope, catchment area, meander wavelength, sinuosity, and discharge. Geophysical Research Letters, 46(6), 3252-3262.

Fry, T. J., & Maxwell, R. M. (2018). Using a distributed hydrologic model to improve the green infrastructure parameterization used in a lumped model. Water, 10(12), 1756.

Fu, M., Fan, T., Ding, Z. A., Salih, S. Q., Al-Ansari, N., & Yaseen, Z. M. (2020). Deep learning data-intelligence model based on adjusted forecasting window scale: application in daily streamflow simulation. IEEE Access, 8, 32632-32651.

Fujita, Koji. "Effect of precipitation seasonality on climatic sensitivity of glacier mass balance." Earth and Planetary Science Letters 276.1-2 (2008): 14-19.

Furey, P. R., & Gupta, V. K. (2005). Effects of excess rainfall on the temporal variability of observed peak-discharge power laws. Advances in Water Resources, 28(11), 1240-1253.

Ghojogh, B., & Crowley, M. (2019). The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. arXiv preprint arXiv:1905.12787.

Gleason, C. J., & Hamdan, A. N. (2017). Crossing the (watershed) divide: satellite data and the changing politics of international river basins. The Geographical Journal, 183, 2-15.

Gleason, C. J., & Smith, L. C. (2014). Toward global mapping of river discharge using satellite images and at-many-stations hydraulic geometry. Proceedings of the National Academy of Sciences, 111(13), 4788–4791.

Gleason, C. J., & Durand, M. T. (2020). Remote sensing of river discharge: a review and a framing for the discipline. Remote Sensing, 12(7), 1107.

Ghosh, S., Vinyals, O., Strope, B., Roy, S., Dean, T., & Heck, L. (2016). Contextual lstm (clstm) models for large scale nlp tasks. arXiv preprint arXiv:1602.06291.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press. (pp. 373-416)

Goldstein, R. (1995). Atmospheric limitations to repeat-track radar interferometry. Geophysical research letters, 22(18), 2517-2520.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural networks, 18(5-6), 602-610.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote sensing of Environment, 202, 18-27.

Gupta, H. V., & Kling, H. (2011). On typical range, sensitivity, and normalization of Mean Squared Error and Nash-Sutcliffe Efficiency type metrics. Water Resources Research, 47(10).

Hagemann, M. W., Gleason, C. J., & Durand, M. T. (2017). BAM: Bayesian AMHG-Manning inference of discharge using remotely sensed stream width, slope, and height. Water Resources Research, 53(11), 9692-9707.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580.

Hirpa, F. A., Gebremichael, M., Hopson, T. M., Wojick, R., & Lee, H. (2014). Assimilation of satellite soil moisture retrievals into a hydrologic model for improving river discharge. Remote sensing of the terrestrial water cycle, 206, 319.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735–1780.

Hosking, J. R. (1984). Modelling persistence in hydrological time series using fractional differencing. Water resources research, 20(12), 1898-1908.

Hu, C., Wu, Q., Li, H., Jian, S., Li, N., & Lou, Z. (2018). Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. Water, 10(11), 1543.

Huang, Q., Long, D., Du, M., Han, Z., & Han, P. (2020). Daily continuous river discharge estimation for ungauged basins using a hydrologic model calibrated by satellite altimetry: Implications for the SWOT mission. Water Resources Research, 56(7), e2020WR027309.

Hsu, K.-l., Gupta, H. V., & Sorooshian, S. (1995). Artificial neural network modelling of the rainfall-runoff process. Water resources research, 31(10), 2517– 2530.

Hu, Y., Huber, A., Anumula, J., & Liu, S. C. (2018). Overcoming the vanishing gradient problem in plain recurrent networks. arXiv preprint arXiv:1801.06105.

Hunink, Eekhout, J., Vente, J., Contreras, S., Droogers, P., & Baille, A. (2017). Hydrological Modelling using Satellite-Based Crop Coefficients: A Comparison of Methods at the Basin Scale. Remote Sensing, 9(2), 174. https://doi.org/10.3390/rs9020174

Immerzeel, W. W., Van Beek, L. P., & Bierkens, M. F. (2010). Climate change will affect the Asian water towers. science, 328(5984), 1382-1385.

Ishitsuka, Y., Gleason, C. J., Hagemann, M. W., Beighley, E., Allen, G. H., Feng, D., ... & Pavelsky, T. M. (2021). Combining optical remote sensing, McFLI discharge estimation, global hydrologic modelling, and data assimilation to improve daily discharge estimates across an entire large watershed. Water Resources Research, 57(3), e2020WR027794.

Jamshidian, M., & Mata, M. (2007). Advances in analysis of mean and covariance structure when data are incomplete. In Handbook of latent variable and related models (pp. 21-44). North-Holland.

Jiang, D., & Wang, K. (2019). The role of satellite-based remote sensing in improving simulated streamflow: A review. Water, 11(8), 1615.

Jones, J. A. (2000). Hydrologic processes and peak discharge response to forest removal, regrowth, and roads in 10 small experimental basins, western Cascades, Oregon. Water Resources Research, 36(9), 2621-2642.

Kahraman, A., Hou, P., Yang, G., & Yang, Z. (2021). Comparison of the effect of regularization techniques and lookback window length on deep learning models in short term load forecasting. In 2021 international top-level forum on engineering science and technology development strategy: 6th purple mountain forum on smart grid protection and control.

Kabeja, C., Li, R., Guo, J., Rwatangabo, D. E. R., Manyifika, M., Gao, Z., ... & Zhang, Y. (2020). The impact of reforestation induced land cover change (1990–2017) on flood peak discharge using HEC-HMS hydrological model and satellite observations: a study in two mountain basins, China. Water, 12(5), 1347.

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. Water Resources Research, 42(3).

Kittel, C. M., Arildsen, A. L., Dybkjær, S., Hansen, E. R., Linde, I., Slott, E., ... & Bauer-Gottwein, P. (2020). Informing hydrological models of poorly gauged river catchments–a parameter regionalization and calibration approach. Journal of Hydrology, 587, 124999.

Kompas, T., Pham, V. H., & Che, T. N. (2018). The effects of climate change on gdp by country and the global economic gains from complying with the paris climate accord. Earth's Future, 6(8), 1153–1173.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. Water Resources Research, 55(12), 11344–11354.

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. Hydrol Earth Syst Sc 23: 4323–4331.

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(12), 5089–5110.

Larnier, K., & Monnier, J. (2020). Hybrid Neural Network–Variational Data Assimilation algorithm to infer river discharges from SWOT-like data. Nonlinear Processes in Geophysics Discussions, 1-30.

Lehner, F., Wood, A. W., Llewellyn, D., Blatchford, D. B., Goodbody, A. G., & Pappenberger, F. (2017). Mitigating the impacts of climate nonstationarity on seasonal streamflow predictability in the US Southwest. Geophysical Research Letters, 44(24), 12-208.

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., ... & Dadson, S. (2021). Hydrological concept formation inside long short-term memory (LSTM) networks. Hydrology and Earth System Sciences Discussions, 2021, 1-37.

Lin, P., Pan, M., Beck, H. E., Yang, Y., Yamazaki, D., Frasson, R., ... others (2019). Global reconstruction of naturalized river flows at 2.94 million reaches. Water resources research, 55(8), 6499–6516.

Lim, S., Kim, S. J., Park, Y., & Kwon, N. (2021). A deep learning-based time series model with missing value handling techniques to predict various types of liquid cargo traffic. Expert Systems with Applications, 184, 115532.

Long, M., Zhu, H., Wang, J., & Jordan, M. I. (2017, July). Deep transfer learning with joint adaptation networks. In International conference on machine learning (pp. 2208-2217). PMLR.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature machine intelligence, 2(1), 56-67.

M. Hrachowitz, H.H.G. Savenije, G. Blöschl, J.J. McDonnell, M. Sivapalan, J.W. Pomeroy, B. Arheimer, T. Blume, M.P. Clark, U. Ehret, F. Fenicia, J.E. Freer, A. Gelfan, H.V. Gupta, D.A. Hughes, R.W. Hut, A. Montanari, S. Pande, D. Tetzlaff, P.A. Troch, S. Uhlenbrook, T. Wagener, H.C. Winsemius, R.A. Woods, E. Zehe & C. Cudennec (2013) A decade of Predictions in Ungauged Basins (PUB)—a review, Hydrological Sciences Journal, 58:6, 1198-1255, DOI: 10.1080/02626667.2013.803183

Ma, K., Feng, D., Lawson, K., Tsai, W. P., Liang, C., Huang, X., ... & Shen, C. (2021). Transferring hydrologic data across continents–leveraging data‐rich regions to improve hydrologic prediction in data‐sparse regions. Water Resources Research, 57(5), e2020WR028600.

Mai, J., Craig, J. R., Tolson, B. A., & Arsenault, R. (2022). The sensitivity of simulated streamflow to individual hydrologic processes across North America. Nature Communications, 13(1), 1-11.

Marcinkevičs, R., & Vogt, J. E. (2020). Interpretability and explainability: A machine learning zoo mini-tour. arXiv preprint arXiv:2012.01805.

Marshall, L., Nott, D., & Sharma, A. (2005). Hydrological model selection: A Bayesian alternative. Water resources research, 41(10).

Mastorakis, G. (2018). Human-like machine learning: limitations and suggestions. arXiv preprint arXiv:1811.06052.

McLaughlin, D. (1995). Recent developments in hydrologic data assimilation. Reviews of Geophysics, 33(S2), 977-984.

Mhaskar, H. N., & Micchelli, C. A. (1993). How to choose an activation function. Advances in neural information processing systems, 6.

Moore, I. D., Grayson, R. B., & Ladson, A. R. (1991). Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrological processes, 5(1), 3-30.}

Montanari, A., & Koutsoyiannis, D. (2012). A blueprint for process-based modelling of uncertain hydrological systems. Water Resources Research, 48(9).

Muhammad, A., Evenson, G. R., Stadnyk, T. A., Boluwade, A., Jha, S. K., & Coulibaly, P. (2019). Impact of model structure on the accuracy of hydrological modelling of a Canadian Prairie watershed. Journal of Hydrology: Regional Studies, 21, 40-56.

Muhebwa, A., Wi, S., Gleason, C. J., & Taneja, J. (2021). Towards improved global river discharge prediction in ungauged basins using machine learning and satellite observations.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part i—a discussion of principles. Journal of hydrology, 10(3), 282– 290.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., ... & Gupta, H. V. (2021). What role does hydrological science play in the age of machine learning?. Water Resources Research, 57(3), e2020WR028091.

Nepal, S., Flügel, W. A., & Shrestha, A. B. (2014). Upstream-downstream linkages of hydrological processes in the Himalayan region. Ecological Processes, 3(1), 1-16.

Ntegeka, V., Baguis, P., Roulin, E., & Willems, P. (2014). Developing tailored climate change scenarios for hydrological impact assessments. Journal of Hydrology, 508, 307-321.

Ortiz-Bobea, A., Ault, T. R., Carrillo, C. M., Chambers, R. G., & Lobell, D. B. (2021). Anthropogenic climate change has slowed global agricultural productivity growth. Nature Climate Change, 11(4), 306-312.

Oubanas, H., Gejadze, I., Malaterre, P. O., Durand, M., Wei, R., Frasson, R. P., & Domeneghetti, A. (2018). Discharge estimation in ungauged basins through variational data assimilation: The potential of the SWOT mission. Water Resources Research, 54(3), 2405-2423.

Oubanas, H., Gejadze, I., Malaterre, P. O., & Mercier, F. (2018). River discharge estimation from synthetic SWOT-type observations using variational data assimilation and the full Saint-Venant hydraulic model. Journal of Hydrology, 559, 638-647.

Oudin, L., Andréassian, V., Perrin, C., Michel, C., & Le Moine, N. (2008). Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments. Water Resources Research, 44(3).

Ouyang, W., Lawson, K., Feng, D., Ye, L., Zhang, C., & Shen, C. (2021). Continental-scale streamflow modelling of basins with reservoirs: Towards a coherent deep-learning-based strategy. Journal of Hydrology, 599, 126455.

Pagliero, L., Bouraoui, F., Diels, J., Willems, P., & McIntyre, N. (2019). Investigating regionalization techniques for large-scale hydrological modelling. Journal of hydrology, 570, 220-235.

Pant, N., Semwal, P., Khobragade, S. D., Rai, S. P., Kumar, S., Dubey, R. K., ... & Ahluwalia, R. S. (2021). Tracing the isotopic signatures of cryospheric water and establishing the altitude effect in Central Himalayas: A tool for cryospheric water partitioning. Journal of Hydrology, 595, 125983.

Pokhrel, Y., Shin, S., Lin, Z., Yamazaki, D., & Qi, J. (2018). potential disruption of flood dynamics in the Lower Mekong River basin due to upstream flow regulation. Scientific reports, 8(1), 1-13.

Pool, S., & Seibert, J. (2021). Gauging ungauged catchments–Active learning for the timing of point discharge observations in combination with continuous water level measurements. Journal of Hydrology, 598, 126448.

Patz, J. A., Campbell-Lendrum, D., Holloway, T., & Foley, J. A. (2005). Impact of regional climate change on human health. Nature, 438(7066), 310-317.

Pilz, T., Francke, T., Baroni, G., & Bronstert, A. (2020). How to Tailor My Process‐Based Hydrological Model? Dynamic Identifiability Analysis of Flexible Model Structures. Water resources research, 56(8), e2020WR028042.

Rácz, A., Bajusz, D., & Héberger, K. (2021). Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification. Molecules, 26(4), 1111.

Rao, R. B., Fung, G., & Rosales, R. (2008, April). On the dangers of cross-validation. An experimental evaluation. In Proceedings of the 2008 SIAM international conference on data mining (pp. 588-596). Society for Industrial and Applied Mathematics.

Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. arXiv preprint arXiv:1710.05941.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. Encyclopedia of database systems, 5, 532-538.

Rodell, M., P.R. Houser, U. Jambor, J. Gottschalck, K. Mitchell, C. Meng, K. Arsenault, B. Cosgrove, J. Radakovich, M. Bosilovich, J.K. Entin, J.P. Walker, D. Lohmann, and D. Toll, 2004: The Global Land Data Assimilation System, Bull. Amer. Meteor. Soc., 85, 381-394

Royall, D. (2021). Land-use impacts on the hydrogeomorphology of small watersheds.

Scown, M. W. (2020). The sustainable development goals need geoscience. Nature Geoscience, 13(11), 714-715.

Seifert, A., & Rasp, S. (2020). Potential and limitations of machine learning for modelling warm‐rain cloud microphysical processes. Journal of Advances in Modelling Earth Systems, 12(12), e2020MS002301.

Sheffield, J., Wood, E. F., Pan, M., Beck, H., Coccia, G., Serrat-Capdevila, A., & Verbist, K. (2018). Satellite remote sensing for water resources management: Potential for supporting sustainable development in data-poor regions. Water Resources Research, 54(12), 9724-9758.

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. Water Resources Research, 54(11), 8558-8593.

Shen, H., Tolson, B. A., & Mai, J. (2022). Time to Update the Split-Sample Approach in Hydrological Model Calibration. Water Resources Research, e2021WR031523.

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019, December). The performance of LSTM and BiLSTM in forecasting time series. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 3285-3292). IEEE.

Sidle, R. C., Gomi, T., Usuga, J. C. L., & Jarihani, B. (2017). Hydrogeomorphic processes and scaling issues in the continuum from soil pedons to catchments. Earth-Science Reviews, 175, 75-96.

Stone, M. (1978). Cross-validation: A review. Statistics: A Journal of Theoretical and Applied Statistics, 9(1), 127-139.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

Srivastava, N., Mansimov, E., & Salakhudinov, R. (2015, June). Unsupervised learning of video representations using lstms. In International conference on machine learning (pp. 843-852). PMLR.

Sun, A. Y., Jiang, P., Mudunuru, M. K., & Chen, X. (2021). Explore Spatio-Temporal Learning of Large Sample Hydrology Using Graph Neural Networks. Water Resources Research, 57(12), e2021WR030394.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018, October). A survey on deep transfer learning. In International conference on artificial neural networks (pp. 270-279). Springer, Cham.

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., & Srikanthan, S. (2009). Critical evaluation of parameter consistency and predictive uncertainty in hydrological modelling: A case study using Bayesian total error analysis. Water Resources Research, 45(12).

Tian, S., Tregoning, P., Renzullo, L. J., van Dijk, A. I., Walker, J. P., Pauwels, V. R., & Allgeyer, S. (2017). Improved water balance component estimates through joint assimilation of GRACE water storage and SMOS soil moisture retrievals. Water Resources Research, 53(3), 1820-1840.

Tran, Q. Q., De Niel, J., & Willems, P. (2018). Spatially distributed conceptual hydrological model building: A generic top-down approach starting from lumped models. Water Resources Research, 54(10), 8064-8085.

Tsai, W. P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., ... & Shen, C. (2021). From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modelling. Nature communications, 12(1), 1-13.

Volpi, E., Di Lazzaro, M., Bertola, M., Viglione, A., & Fiori, A. (2018). Reservoir effects on flood peak discharge at the catchment scale. Water Resources Research, 54(11), 9623-9636.

Wager, S., Wang, S., & Liang, P. S. (2013). Dropout training as adaptive regularization. Advances in neural information processing systems, 26.

Wagener, T., & Montanari, A. (2011). Convergence of approaches toward reducing uncertainty in predictions in ungauged basins. Water Resources Research, 47(6). Wu, X., Jeuland, M., & Whittington, D. (2016). Does political uncertainty affect water resources development? the case of the eastern nile. Policy and Society, 35(2), 151–163.

Wang, H., Cao, L., & Feng, R. (2021). Hydrological Similarity-Based Parameter Regionalization under Different Climate and Underlying Surfaces in Ungauged Basins. Water, 13(18), 2508.

Wang, H., Cheng, Q., Wang, T., Zhang, G., Wang, Y., Li, X., & Jiang, B. (2021). Layover Compensation Method for Regional Spaceborne SAR Imagery Without GCPs. IEEE Transactions on Geoscience and Remote Sensing, 59(10), 8367-8381.

Wanner, J., Herm, L. V., & Janiesch, C. (2020). How much is the black box? The value of explainability in machine learning models.

Wu, H., & Prasad, S. (2017). Convolutional recurrent neural networks forhyperspectral data classification. Remote Sensing, 9(3), 298.

Wu, W., May, R. J., Maier, H. R., & Dandy, G. C. (2013). A benchmarking approach for comparing data splitting methods for modelling water resources parameters using artificial neural networks. Water Resources Research, 49(11), 7598-7614.

Xie, K., Liu, P., Zhang, J., Han, D., Wang, G., & Shen, C. (2021). Physics-guided deep learning for rainfall-runoff modelling by considering extreme events and monotonic relationships. Journal of Hydrology, 603, 127043.

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., & Pavelsky, T. M. (2019). MERIT Hydro: A high-resolution global hydrography map based on latest topography dataset. Water Resources Research, 55(6), 5053-5073.

Yang, X., Magnusson, J., Huang, S., Beldring, S., & Xu, C. Y. (2020). Dependence of regionalization methods on the complexity of hydrological models in multiple climatic regions. Journal of Hydrology, 582, 124357.

Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. arXiv preprint arXiv:1702.01923.

Ying, X. (2019, February). An overview of overfitting and its solutions. In Journal of physics: Conference series (Vol. 1168, p. 022022). IOP Publishing.

Yoshida, T., Hanasaki, N., Nishina, K., Boulange, J., Okada, M., & Troch, P. A. Inference of parameters for a global hydrological model: Identifiability and predictive uncertainties of climate-based parameters. Water Resources Research, e2021WR030660.

Yu, T., & Zhu, H. (2020). Hyper-parameter optimization: A review of algorithms and applications. arXiv preprint arXiv:2003.05689.

Yu, C., Li, Z., Penna, N. T., & Crippa, P. (2018). Generic atmospheric correction model for interferometric synthetic aperture radar observations. Journal of Geophysical Research: Solid Earth, 123(10), 9202-9222.

Zhou, Z., Ren, J., He, X., & Liu, S. (2021). A comparative study of extensive machine learning models for predicting long-term monthly rainfall with an ensemble of climatic and meteorological predictors. Hydrological Processes, 35(11), e14424.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., ... & He, Q. (2020). A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1), 43-76.

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., & Hinkelmann, R. (2021). Ensemble machine learning paradigms in hydrology: A review. Journal of Hydrology, 598, 126266.

Ma, C., Dai, G., & Zhou, J. (2021). Short-term traffic flow prediction for urban road sections based on time series analysis and LSTM_BILSTM method. IEEE Transactions on Intelligent Transportation Systems, 23(6), 5615-5624.

Atef, S., & Eltawil, A. B. (2020). Assessment of stacked unidirectional and bidirectional long short-term memory networks for electricity load forecasting. Electric Power Systems Research, 187, 106489.

Althelaya, K. A., El-Alfy, E. S. M., & Mohammed, S. (2018, April). Evaluation of bidirectional LSTM for short-and long-term stock market prediction. In 2018 9th international conference on information and communication systems (ICICS) (pp. 151-156). IEEE.

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019). A comparative analysis of forecasting financial time series using arima, lstm, and bilstm. arXiv preprint arXiv:1911.09512.

Shrestha, S., Bae, D. H., Hok, P., Ghimire, S., & Pokhrel, Y. (2021). Future hydrology and hydrological extremes under climate change in Asian river basins. Scientific reports, 11(1), 17089.

Leng, G., Tang, Q., & Rayburg, S. (2015). Climate change impacts on meteorological, agricultural and hydrological droughts in China. Global and Planetary Change, 126, 23-34.

Tabari, H. (2020). Climate change impact on flood and extreme precipitation increases with water availability. Scientific reports, 10(1), 13768.

Appendix

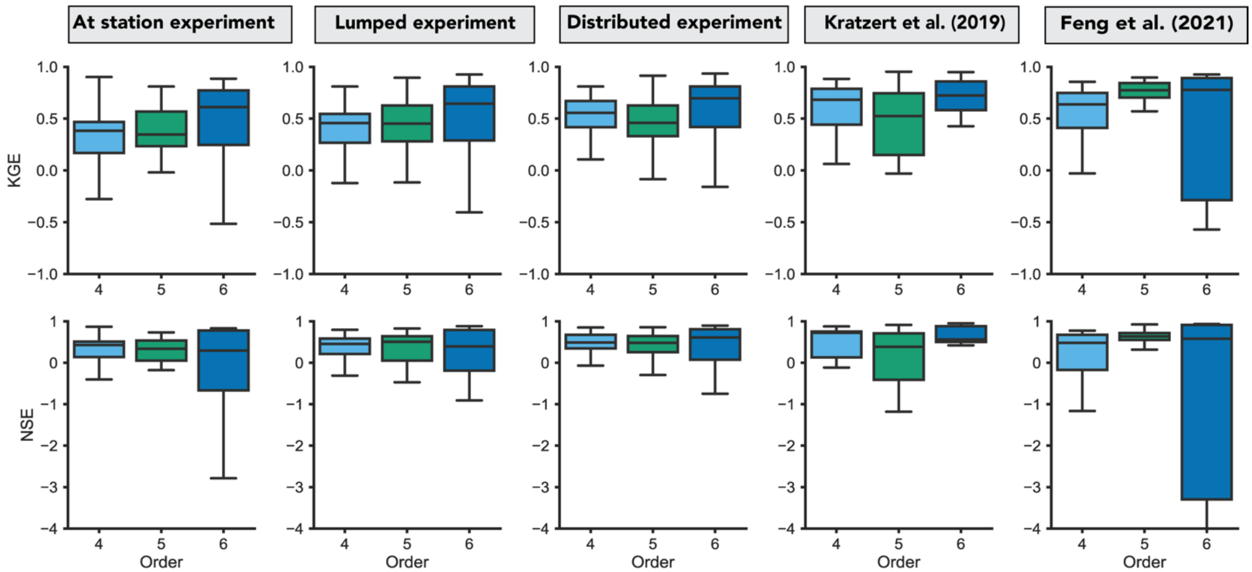

Figure S1: Left to right: Distribution comparisons of selected metrics on held-out predictions for all experiments (i.e., at-station, lumped, and distributed experiments) and literature models: Kratzert et al. (2019) and Feng et al. (2021). Note that distributions for seventh and eighth orders are not included due to the limited gauge stations in the training set.

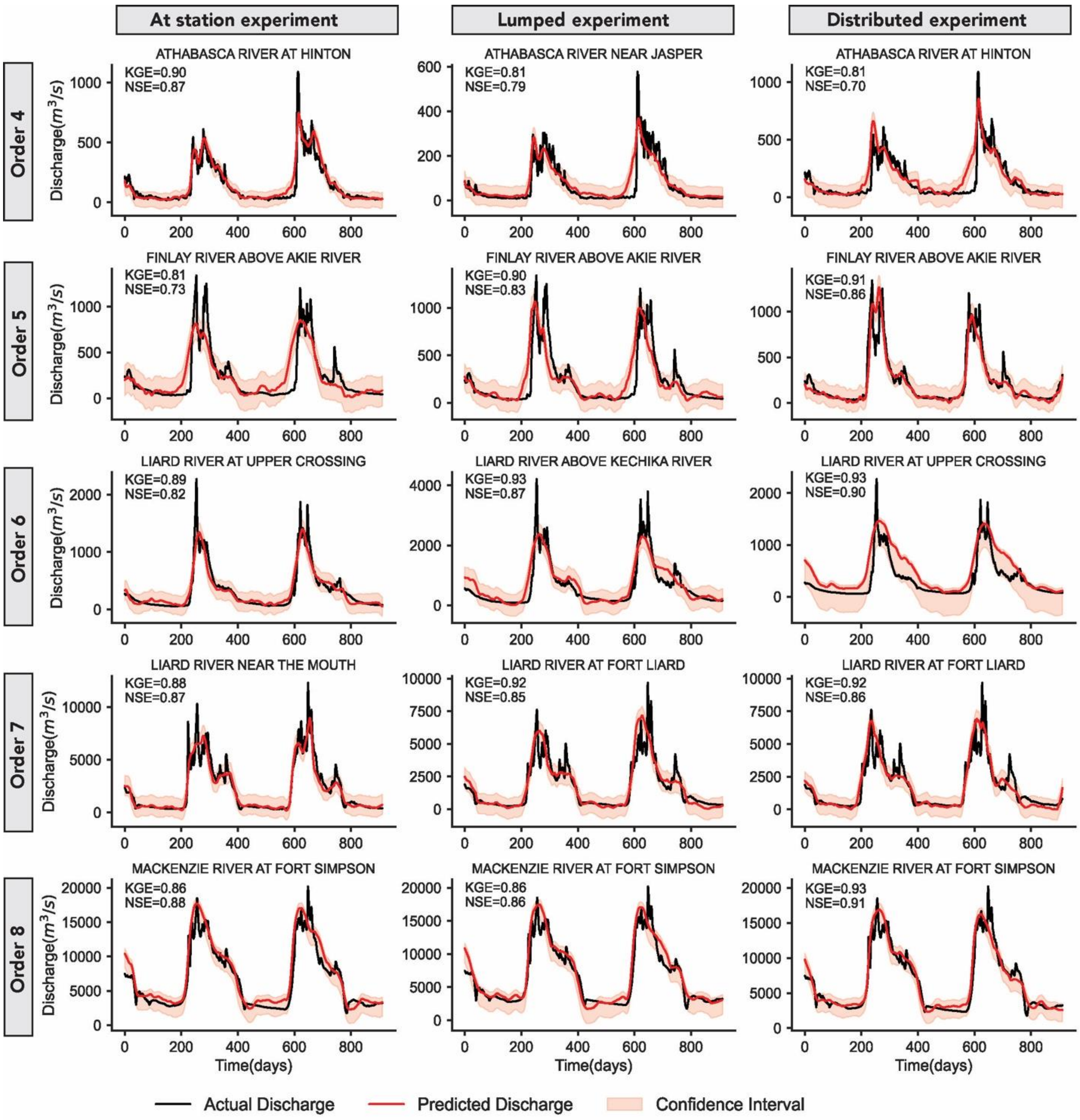

Figure S2: Left to right: Representative hydrographs showing the best performing models in each of the experiments; At-station (left), lumped (middle), and distributed (right) experiments across the defined orders, i.e., from order 4 (top) to order 8 (bottom). Here, we plot hydrographs for the first 2.5 years.

Table 2: Summary of variables used as input features to the LSTM model.

Variable name	Description	Unit
Discharge	In-situ daily river discharge at a gauge station	m^3s^{-1}
Albedo	Albedo	%
Avg_Skin_Temp	Average surface skin temperature	K

PlantCanopyWater	Plant canopy surface water	Kg/m^2
CanopyWaterEvpn	Canopy water evaporation	W/m^2
DirectEvonBareSoil	Direct evaporation free bare soil	W/m^2
Evapotranspn	Evapotranspiration	$Kg/m^2//s$
LngWaveRadFlux	Downward long-wave radiation flux	W/m^2
NetRadFlux	Net long-wave radiation flux	W/m^2
PotEvpnRate	Potential Evaporation rate	W/m^2
Pressure	Pressure	Pa
SpecHmd	Specific humidity	kg/kg
HeatFlux	Heat flux	W/m^2
Sen.HtFlux	Sensible heat net flux	W/m^2
LtHeat	Latent heat net flux	W/m^2
StmSurfRunoff	Storm surface runoff	kg/m^2
BsGndWtrRunoff	Baseflow-groundwater runoff	kg/m^2
SnowMelt	Snow melt	kg/m^2
TotalPcpRate	Total precipitation rate	$kg/m^2/s$
RainPcpRate	Rain precipitation rate	$kg/m^2/s$
RootZoneSoilMstr	Root zone soil moisture	kg/m^2
SnowDepthWtrEq	Snow depth water Equivalent	W/m^2
DwdShtWvRadFlux	Downward short-wave radiation flux	m
SnowDepth	Snow depth	$kg/m^2/s$
SnowPcpRate	Snow precipitation rate	kg/m^2
SoilMst10	Soil moisture (0-10) cm	kg/m^2
SoilMst40	Soil moisture (10-40) cm	kg/m^2
SoilMst100	Soil moisture (40-100) cm	kg/m^2
SoilMst200	Soil moisture (100-200) cm	kg/m^2
SoilTmp10	Soil temperature (0-10) cm	K
SoilTmp40	Soil temperature (10-40) cm	K
SoilTmp100	Soil temperature (40-100) cm	K
SoilTmp200	Soil temperature (100-200) cm	K
NetShtWvRadFlux	Net short wave radiation flux	W/m^2
AirTemp	Air temperature	K
Tspn	Transpiration	W/m^2
WindSpd	Windspeed	m/s
	Reach width	
	Stream length	
	Bed slope	
	Sinuosity	
	Upstream Area	
	Length Dir	
	Stream Drop	

	Mean width	
	Max Width	