Rethinking Deep Learning-Based Rainfall-runoff Modelling for Data-Rich Hydrology

by

Qiutong Yu

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Civil Engineering

Waterloo, Ontario, Canada, 2026

© Qiutong Yu 2026

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Usman T. Khan Associate Professor, Civil Engineering, Lassonde School of Engineering, York University
Supervisor(s)	Bryan A. Tolson Professor, Department of Civil and Environmental Engineering, University of Waterloo
	Jimmy Lin Professor, Cheriton School of Computer Science, University of Waterloo
Internal Member	John Quilty Assistant Professor, Department of Civil and Environmental Engineering, University of Waterloo
Internal Member	Chul Min Yeum Associate Professor, Department of Civil and Environmental Engineering, University of Waterloo
Internal-External Member	Grant Gunn Assistant Professor, Department of Geography and Environmental Management, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Chapter 2 of this thesis consists of a published paper (Yu et al., 2024), which I co-authored with Dr. Bryan A. Tolson, Dr. Hongren Shen, Dr. Ming Han, Dr. Juliane Mai, and Dr. Jimmy Lin. My contribution to this paper included conceptualization, methodology, software implementation, validation, visualization, original draft, review and editing. Dr. Bryan A. Tolson contributed to conceptualization, methodology, review and editing. Dr. Hongren Shen contributed to software implementation and methodology. Dr. Ming Han contributed to data preparation and software implementation. Dr. Juliane Mai contributed to data preparation, methodological guidance, review and editing. Dr. Jimmy Lin contributed to methodological guidance, review and editing.

Chapter 3 of this thesis consists of a published paper (Yu & Tolson, 2026), which I co-authored with Dr. Bryan A. Tolson. My contribution to this paper included conceptualization, methodology, software implementation, validation, visualization, original draft, review and editing. Dr. Bryan A. Tolson contributed to methodology, review and editing.

Yu, Q., Tolson, B. A., Shen, H., Han, M., Mai, J., & Lin, J. (2024). Enhancing long short-term memory (LSTM)-based streamflow prediction with a spatially distributed approach. *Hydrology and Earth System Sciences*, 28(9), 2107–2122. https://doi.org/10.5194/hess-28-2107-2024

Yu, Q., & Tolson, B. A. (2026). How much historical data do we need? The role of data recency and training period length in LSTM-based rainfall-runoff modelling. *Journal of Hydrology*, 668, 135046. https://doi.org/10.1016/j.jhydrol.2026.135046

Abstract

Sequential deep learning (DL) architectures, particularly Long Short-Term Memory (LSTM) networks, have become widely used in large-sample rainfall-runoff modelling. By training a single model across many basins using meteorological forcings and static basin attributes, these models can exploit cross-basin rainfall–runoff patterns and often achieve strong overall predictive performance. This modelling paradigm has developed alongside the increasing availability of large hydrological datasets. While previous DL-based hydrological studies mainly benefited from increasing the number of gauged basins for regional training, hydrological datasets have continued to expand the information available to DL models in other ways. Lake-river routing network products and catchment discretization provide finer spatial detail, long-term hydrometeorological archives extend historical coverage, and high-frequency forcing and streamflow records support sub-daily prediction. These developments create opportunities for DL-based rainfall–runoff modelling, but they also raise methodological questions about how expanded datasets should be represented, selected, and matched with suitable model architectures. This thesis investigates how DL-based rainfall-runoff modelling can be adapted to use hydrological datasets that provide finer spatial detail, longer historical coverage, and higher temporal resolutions, through three complementary studies on subbasin-scale modelling, recency-aware training, and model architecture for data-intensive settings.

The first study examines how subbasin-scale spatial information can be incorporated into LSTM-based streamflow prediction. Regional LSTM models typically use basin-averaged forcings and attributes, treating each basin as a lumped response unit. This representation supports large-sample training, but it also reduces information on within-basin variability and drainage connectivity. To address this limitation, the study proposed the Spatially Recursive (SR) model, a hybrid framework that applies a regionally trained lumped LSTM at the subbasin scale and routes the resulting local streamflow predictions through a lake–river hydrological routing model. The method was evaluated in the Great Lakes region and compared with the original lumped LSTM. The SR model achieved comparable performance at trained locations and improved prediction in larger basins. These results show that lake–river routing network information and subbasin-level response units can help regional LSTM models use spatially detailed hydrological data more effectively.

The second study investigates how long historical records should be used when many training basins are available. Reanalysis products and hydrometric archives provide decades of meteorological forcing and streamflow data, but the relevance of older observations depends on their similarity to the prediction period. Using hydrometeorological records from 1374 North American watersheds spanning 1950–2023, this study evaluates the effects of training-period length and data recency through backward-expanding, forward-expanding, and sliding-window experiments. The results show that recent records have greater value for LSTM training than distant historical records. Extending the training period with older data produced limited gains and sometimes reduced performance. The benefit of increasing the number of training watersheds also depended on data recency, especially for prediction in ungauged basins. Spatial diversity improved generalization most clearly when recent observations were included. Peak-flow analyses further showed that very short recent windows may provide insufficient exposure to rare high-flow events, indicating that recency-aware training still requires adequate temporal coverage. These findings suggest that large-sample training data should be selected by considering both spatial diversity and temporal relevance.

The third study explores whether enhanced recurrent architecture can support large-sample hourly rainfall-runoff modelling. At hourly resolution, models must process longer input sequences and preserve relevant information across many more time steps than in daily modelling. To examine this issue, this study implemented MF-xLSTM, which combines a multi-frequency daily–hourly input structure with an xLSTM model backbone. The model was evaluated using hourly data from 430 Canadian basins and compared with LSTM-based benchmarks under the same multi-frequency input setting. MF-xLSTM achieved performance comparable to the LSTM benchmarks when predicting unseen periods and showed stronger overall accuracy in pseudo-ungauged basins. It also produced low-flow volume bias closer to zero, suggesting that xLSTM's memory mechanisms may help retain information from earlier parts of long input sequences that are relevant to delayed storage release such as groundwater baseflow. These results indicate that xLSTM is a plausible model backbone for large-sample hourly rainfall–runoff modelling.

Overall, this thesis demonstrates that the benefits of expanded hydrological datasets depend on how these data are used in model development. In a data-rich hydrological setting,

simply adding more data does not necessarily lead to better generalization. Finer spatial data need to be represented in ways that preserve within-basin heterogeneity and drainage connectivity; long historical records need to be selected with attention to temporal relevance; and high-frequency observations require architectures capable of retaining information across longer sequences. This thesis provides evidence on how DL-based rainfall-runoff modelling can make more effective use of expanded hydrological datasets.

Acknowledgements

I am deeply grateful to my supervisors, Dr. Bryan A. Tolson and Dr. Jimmy Lin, for their guidance and support throughout my doctoral studies. Their expertise and advice have been invaluable to my academic growth and to the completion of this work.

I would like to sincerely thank my defense committee members, Dr. John Quilty, Dr. Chul Min Yeum, Dr. Grant Gunn, and Dr. Usman T. Khan, for taking the time to review this thesis and providing thoughtful suggestions.

I would like to thank my colleagues in the Hydrology Research Group for the shared experiences, discussions, and companionship that made this doctoral journey more enjoyable.

I would also like to thank all my friends, near and far, for the support they have offered over the years. I honour the memory of my late friend, Zhelong, whose encouragement was especially meaningful to me when I first began my doctoral studies.

I am deeply grateful to my parents. From my earliest school years to my doctoral studies, their love and support have remained unwavering.

I am also thankful for Caviar, my British Shorthair, who kept me company and did not cause any damage to my workspace.

Words cannot fully express my gratitude to my wife, Yingfang, for her love, understanding, and the many sacrifices she has made. This thesis would not have been possible without her by my side.

Finally, to my beloved daughter, Valerie, you are the greatest surprise and the most precious gift I have received during my doctoral journey. You have brought me joy unlike any other and taught me lessons beyond science. Thank you for helping me grow alongside you. I hope that you will find what you love and always have the strength to keep moving forward on your own path, with the same persistence that has carried me through these years.

Dedication

This thesis is dedicated to my wife and daughter,

whose love, like the Yangtze River, always carries me home.

Table of Contents

Examining Committee Membership ..... **ii**

Author's Declaration ..... **iii**

Statement of Contributions ..... **iv**

Abstract ..... **v**

Acknowledgements ..... **viii**

Dedication ..... **ix**

List of Figures ..... **xiii**

List of Tables ..... **xvii**

Chapter 1: Introduction **1**

1.1 Research Background 1

1.1.1 Overview of large-scale hydrological datasets 1

1.1.2 Overview of deep learning for rainfall-runoff modelling 3

1.2 Research Gaps 6

1.3 Thesis Contributions 8

1.4 Thesis Structure 8

Chapter 2: Enhancing LSTM-based streamflow prediction with a spatially distributed approach **9**

2.1 Introduction 10

2.2 Material and Methods 13

2.2.1 Overview of Spatially Recursive (SR) model 13

2.2.2 Selection of additional gauging locations for concept validation 21

2.2.3 Performance metrics 23

2.2.4 Experimental design 24

2.3 Results and Discussion 25

2.3.1 Extrapolating LSTM to small Non-GRIP-GL basins 25
2.3.2 Comparing lumped LSTM and SR model on GRIP-GL basins 26
2.3.3 New testing basins and impact of routing network delineation 30
2.4 Conclusion 32
Chapter 3: The role of data recency and training period length in LSTM-based rainfall-runoff modelling 35
3.1 Introduction 36
3.2 Methodology 39
3.2.1 Model settings and performance metrics 39
3.2.2 Study area and data 40
3.2.3 Experimental design 42
3.2.4 Peak flow analysis 43
3.3 Results and Discussion 44
3.3.1 Impact of training period length and recency 44
3.3.2 Is there a minimum training period length for reliable performance? 47
3.3.3 How do training strategies affect peak flow predictions? 48
3.3.4 Implications and limitations of findings 50
3.4 Conclusion 52
Chapter 4: Exploring xLSTM for Large-Sample Hourly Rainfall-Runoff Modelling 55
4.1 Introduction 56
4.2 Data and Methods 60
4.2.1 Dataset 60
4.2.2 Definition of flashy basins 61
4.2.3 Data partitioning 64
4.2.4 Benchmark models 65

4.2.5 Extended Long Short-term Memory (xLSTM) 66
4.2.6 Hyperparameter tuning 74
4.2.7 Evaluation criteria 75
4.3 Results and Discussion 76
4.3.1 Effects of xLSTM block stacking 76
4.3.2 Performance comparison between xLSTM and benchmark LSTM 77
4.3.3 Impacts of basin flashiness 80
4.3.4 Computational efficiency 82
4.4 Conclusion 82
Chapter 5: Conclusions, Limitations and Future Work 85
5.1 Conclusions 85
5.2 Limitations 86
5.3 Future Work 87
References 89
Appendices 109
A2 Supplement for Chapter 2 109
A2-1 Map of the streamflow gauging stations 109
A2-2 Raven configurations for the routing-only mode 109
A2-3 Lumped LSTM configurations 112
A3 Supplement for Chapter 3 115
A4 Supplement for Chapter 4 118

List of Figures

Figure 2-1. The spatially recursive model workflow for an arbitrary basin of interest. Note the workflow is conceptually model-agnostic (e.g. the lumped LSTM can be replaced by another data-driven model), while each component needs to generate outputs compatible with other components. Components with multiple arrows are utilized multiple times throughout the workflow..... 14

Figure 2-2. An example lake-river routing network for the Ontonagon River Watershed (USGS gauge ID: 04040000), which is one of the GRIP-GL calibration basins. (A) The network from the default GRIP-GL common routing product. (B) The network from the initial NALRP without simplification. (C) The simplified NALRP network delineated by a minimum lake area of 5 km^2 and a minimum subbasin drainage area of 500 km^2..... 19

Figure 2-3. Comparison of observation (grey), lumped LSTM prediction (red) and SR model prediction (blue) for the 4 small Non-GRIP-GL basins. KGE values over the period from 2001 to 2017 are given for each basin. 25

Figure 2-4. Boxplots of KGE validation scores for the SR model (blue) and the lumped LSTM (red). (A) The results of all basins participating in each validation experiment. (B) The results of basins with a size larger than 2000 km^2. 27

Figure 2-5. Comparison of time series of observation (grey), lumped LSTM prediction (red) and SR model prediction (blue) for representative GRIP-GL validation basins, over the selected period from 2009 to 2012. KGE values over the whole study period from 2000 to 2017 are given for each basin. (A) The 2 basins that show the most significant improvement as compared to the lumped LSTM. (B) The 2 basins that show the most significant degradation as compared to the lumped LSTM. (C) The two basins where the SR model achieves the highest KGE scores. (D) The two basins where the SR model achieves the lowest KGE scores. 29

Figure 2-6. Overall model performances as adopting different routing network delineations in the 8 large Non-GRIP-GL basins, during the whole study period 2000-2017. The basins are sorted from left to right in ascending order according to their sizes, from smallest (2343 km^2) to largest (6760 km^2)..... 31

Figure 2-7. Comparison of the time series of observation (grey), lumped LSTM prediction (red) and SR model predictions with different routing networks for the Vermilion River basin

05129115. Note that the lines of NALRP time series all follow extremely similar trends as the GRIP-GL Routing series and thus are not distinguishable. 31

Figure 2-8. Comparison of the routing networks of 3 large Non-GRIP-GL basins to illustrate the variation in different lake densities. (A) Gauged basin 05129115 (5 lakes, 14% of basin is covered by lakes). (B) Gauged basin 02KJ004 (11 lakes, 10% of basin is covered by lakes) and (C) Gauged basin 04260500 (4 lake, 4% of basin is covered by lakes). The three illustrated routing networks were delineated to mimic the GRIP-GL routing delineation strategy. 32

Figure 3-1. Locations of all 1374 gauging stations included in this study. Due to the large geographical extent, an inset map displays stations located in Alaska and Yukon. Stations of training watersheds are indicated by green circles, validation watersheds by yellow squares, and testing watersheds by red triangles……...41

Figure 3-2. Train/test temporal splits adopted for each experiment. The left, middle, and right panels correspond to Experiments 1, 2, and 3, respectively. 43

Figure 3-3. Results of Experiment 1. In the upper row of subplots, the left panel illustrates four temporal splits (1A–1D), each with a fixed end year of 2010 and progressively longer training periods by including older historical data, thereby extending backward in time. The heatmaps show the median KGE performance under temporal testing (220 training watersheds; middle panel) and spatiotemporal testing (137 pseudo-ungauged watersheds; right panel), across increasing numbers of watersheds used for training (x-axis). Beneath the heatmaps, cumulative distribution functions (CDFs) of KGE scores are shown for spatiotemporal testing with 220 (left) and 1100 (right) training watersheds, highlighting differences in the overall distribution of model performance across temporal splits. 45

Figure 3-4. Results of Experiment 2. In the upper row of subplots, the left panel illustrates four temporal splits (2A–2D), each with a fixed start year of 1950 and progressively longer training periods by including more recent data, thereby extending forward in time. The remaining panels are consistent with Figure 3-5 in both layout and evaluation setup. 46

Figure 3-5. Results of Experiment 3. In the upper row of subplots, the left panel illustrates four temporal splits (3A–3D), each using a fixed-length training window that shifts forward in time. The remaining panels are consistent with Figure 3-5 in layout and evaluation setup. 46

Figure 3-6. Results of training on the most recent 11, 6, and 3 years of data, in comparison with configurations 1A and 1B from Experiment 1. Heatmaps are formatted consistently with earlier experiments to facilitate comparison. The left CDF plot presents the performance of models trained on 1100 watersheds under temporal testing, while the right CDF plot shows the results under spatiotemporal testing. 47

Figure 3-7. APE of model predictions for the 1085 peak flow events under the temporal testing scenario. The left subplot shows results from models trained on 220 watersheds, and the right subplot shows results from models trained on 1100 watersheds. Boxplots summarize the distribution of APE values for different training period start years (x-axis), with outliers indicating instances of large prediction errors. 49

Figure 3-8. APE of model predictions for the 681 peak flow events under the spatiotemporal testing scenario..... 49

Figure 4-1. Relationship between the daily-based RBI and the GOUE (NSE) across the 430 basins. Each point represents one gauged basin……63

Figure 4-2. Spatial distribution of the 430 study catchments across Canada. Circles denote gauged catchments used for model training and temporal testing, while triangles denote pseudo-ungauged catchments used for spatial testing. Red markers indicate flashy catchments, and grey markers indicate non-flashy catchments. 64

Figure 4-3. The architecture of the xLSTM network, adapted from Beck et al. (2024). 67

Figure 4-4. Schematic comparison of the original MF-LSTM implementation (Acuña Espinoza et al., 2025) and the MF-LSTM/MF-xLSTM implementations used in this study for constructing the multi-frequency input sequence. 73

Figure 4-5. Cumulative KGE distributions of the different models. The left panel shows KGEs of 344 basins from temporal testing, the right panel shows KGEs of 86 basins from spatial testing. 79

Figure 4-6. Hourly hydrograph comparison for two gauges in the western Greater Toronto Area during July 2023. The left panel shows 02HC009 (East Humber River near Pine Grove; drainage area = 196 km^2; RBI = 0.304), which is characterized by a more attenuated runoff response. The right panel shows 02HC030 (Etobicoke Creek below Queen Elizabeth Highway; drainage area = 208 km^2; RBI = 0.720), which is characterized by a flashier runoff response..... 81

Figure A2-1. Locations of the streamflow gauging stations used in this study. In total, there are 224 streamflow gauging stations (shown as dots on the map). The 212 stations used in the GRIP-GL study are shown in dark blue dots (calibration) and light blue dots (validation). The stations of the 4 small Non-GRIP-GL basins (described in Sec 3.1 of the main article) are shown in light red dots, and the stations of the 8 large Non-GRIP-GL basins (described in Sec 3.2 of the main article) are shown in light green dots…….109

Figure A2-2. The .rvi file structure for implementing the routing-only setup..... 110

Figure A2-3. The .rvh file structure for representing lakes/reservoirs in the routing model..... 111

Figure A3-1. Nash–Sutcliffe Efficiency (NSE) results for Experiment 1……..116

Figure A3-2. NSE results for Experiment 2..... 117

Figure A3-3. NSE results for Experiment 3..... 117

List of Tables

Table 2-1. Summary of the 12 Non-GRIP-GL gauging basins selected for this study..... 22

Table 2-2. Median KGE for prediction performance of the lumped LSTM and the SR models. Top performing model is indicated by bold font. 27

Table 3-1. Performance improvements inferred from Gauch, Mai, et al. (2021) showing the effects of watershed quantity and training period length on NSE…….37

Table 3-2. Hyperparameter settings used for the LSTM network. The first six hyperparameters listed in the table were fine-tuned over the specified ranges, and the selected values are highlighted in bold and underlined. 40

Table 3-3. Dataset partitioning scheme. 41

Table 4-1. Overview of the performance metrics…….75

Table 4-2. Performance comparison of different xLSTM stacking configurations using 3-fold cross-validation within the training set. 76

Table 4-3. Median metrics for sMTS-LSTM (Gauch, Kratzert, et al., 2021), MF-LSTM (Acuña Espinoza et al., 2025), and proposed MF-xLSTM in temporal testing. The number of basins is denoted by '***n***'. 78

Table 4-4. Median metrics for the sMTS-LSTM, MF-LSTM, and MF-xLSTM in spatial testing. 78

Table A2-1. Hyperparameter settings used for the lumped LSTM…….112

Table A2-2. CaSR-v2 forcing variables used for the lumped LSTM. For initial lumped LSTM training, these variables are derived for each of the gauged basins. For the SR-model, these variables are derived for each of subbasins in the routing network. 113

Table A2-3. Static attributes used for the lumped LSTM. For initial lumped LSTM training, these variables are derived for each of the gauged basins. For the SR-model, these variables are derived for each of subbasins in the routing network..... 114

Table A3-1. Temporal distribution of streamflow data availability across all 1100 training watersheds for each training period block (1950–1965, 1965–1980, 1980–1995, 1995–2010)……115

Table A3-2. Summary of Seasonal Mann–Kendall (MK) trend test results (1950–2023) for precipitation (P), mean temperature (T), streamflow (Q) across all 1374 watersheds. .. 115

Table A3-3. Static watershed attributes used as LSTM input variables. 116
Table A4-1. CaSR forcing variables ……..118
Table A4-2. Static attributes..... 118
Table A4-3. General hyperparameters tuning grid. The selected values are highlighted in bold. 119
Table A4-4. xLSTM-specific hyperparameters 119

Chapter 1: Introduction

1.1 Research Background

Hydrology, as a complex and dynamic field, plays a pivotal role in understanding and managing water resources, especially in the face of climate change and increasing anthropogenic pressures. Rainfall-runoff modelling is a central task in hydrology because streamflow prediction underpins flood forecasting, drought assessment, reservoir operation and water resources management. A persistent challenge is that rainfall–runoff relationships vary across catchments and can shift over time under climate change and anthropogenic disturbance (Milly et al., 2008; Deb et al., 2019). As a result, hydrological models must generalize across heterogeneous catchments and changing hydroclimatic conditions (Addor et al., 2020). Large-sample hydrology uses datasets containing many catchments to support statistically robust hydrological inferences that are less dependent on individual study sites (Gupta et al., 2014). This approach supports the identification of similarities and differences among hydrological systems and links these differences to underlying climate, landscape and human controls (Addor et al., 2020). In hydrological modelling, it provides a basis for more systematic model evaluation, catchment classification, regionalization, uncertainty assessment and generalization beyond individual case studies (Gupta et al., 2014; Addor et al., 2020).

1.1.1 Overview of large-scale hydrological datasets

The development of large-sample rainfall–runoff modelling has been closely linked to the availability of open benchmark datasets. One of the most influential benchmark datasets is CAMELS-US (Catchment Attributes and Meteorology for Large-sample Studies), which provides daily meteorological forcings, streamflow observations, and catchment attributes for 671 minimally disturbed basins in the contiguous United States (Newman et al., 2015; Addor et al., 2017). The CAMELS framework was later extended to many other regions, including CAMELS-CL for 516 catchments in Chile and CAMELS-GB for 671 catchments in Great Britain (Alvarez-Garreton et al., 2018; Coxon et al., 2020). These regional datasets support model evaluation across different hydroclimatic and physiographic conditions, while maintaining a relatively standardized catchment-scale data structure.

Large-sample datasets have expanded in terms of spatial coverage and temporal extent. HYSETS provides a multisource database for 14,425 North American watersheds, including

hydrometric records, watershed boundaries, meteorological data, and physiographic attributes (Arsenault et al., 2020). At the global scale, Kratzert et al. (2023) introduced Caravan, a global community dataset that harmonizes meteorological forcings, streamflow observations, and catchment attributes from seven existing large-sample hydrology datasets. The initial release covers 6830 catchments. GRDC-Caravan extends this dataset using open discharge records from the Global Runoff Data Centre, covering 5356 catchments in 25 countries from 1950 to 2023 (Färber et al., 2025).

Other datasets have expanded large-sample hydrology toward higher temporal resolution and finer spatial representation. LamaH-CE provides data for 859 gauged catchments in Central Europe, with more than 60 catchment attributes and both daily and hourly hydrometeorological time series, making it relevant for high-temporal-resolution modelling (Klingler et al., 2021). CAMELSH extends large-sample hydrology at the hourly scale across the contiguous United States by providing hourly meteorological time series, catchment attributes, and watershed boundaries for 9008 catchments. Among these catchments, 3166 include hourly streamflow records (Tran et al., 2025). CAMELS-SPAT further supports spatially distributed analysis by providing streamflow observations, meteorological forcing data, and geospatial data for 1426 basins in the United States and Canada at lumped, sub-basin, and gridded resolutions (Knoben et al., 2025).

In addition to catchment benchmark datasets, global hydrographic products provide spatial information needed to represent drainage networks and within-basin heterogeneity. HydroSHEDS is a widely used global hydrographic dataset that provides hydrologically conditioned elevation, flow direction, flow accumulation, river network, and basin boundary products (Lehner et al., 2008; Lehner & Grill, 2013). Related products further extend this hydrographic framework: HydroBASINS provides nested basin boundaries, HydroRIVERS represents river reaches, HydroLAKES provides global lake polygons and associated lake attributes, and HydroATLAS provides standardized hydro-environmental attributes for sub-basins and river reaches (Lehner & Grill, 2013; Messager et al., 2016; Linke et al., 2019).

Overall, these large-scale datasets indicate that the volume and diversity of hydrological data available for model training have expanded substantially, including more gauged basins, longer records, higher temporal resolution, finer spatial representation, and richer catchment

attributes. This expansion of large-sample datasets has provided an important foundation for the recent development of AI-based hydrological modelling.

1.1.2 Overview of deep learning for rainfall-runoff modelling

Rainfall-runoff models are commonly grouped into physically-based, conceptual and data-driven approaches. Physically-based models represent hydrological processes through governing equations and spatially explicit descriptions of catchment properties, boundary conditions and water movement (Freeze & Harlan, 1969; Beven, 2012; Clark et al., 2015). Conceptual models simplify the rainfall–runoff system into linked storages and flux relationships, with parameters calibrated to reproduce observed hydrological responses (Wagener et al., 2003; Beven, 2012). These models are scientifically valuable because their structures reflect the understanding of hydrological processes. However, these models face persistent difficulties in large-sample modelling. Parameters calibrated for one basin may not transfer well to another, and regional calibration can reduce performance relative to local calibration. Equifinality further complicates regionalization because multiple parameter sets or model structures may produce similarly acceptable simulations, reducing the uniqueness and transferability of calibrated parameters (Wagener et al., 2003; Beven, 2006). Data-driven rainfall–runoff models differ from physically-based and conceptual models by relying primarily on empirical input–output relationships rather than predefined process equations or storage structures. They estimate streamflow as a nonlinear function of historical hydroclimatic inputs, particularly time-varying variables such as precipitation, temperature and potential evapotranspiration (Young & Beven, 1994; Shamseldin, 1997). Machine learning (ML) is a major class of data-driven modelling approaches, and deep learning (DL) is a subset of ML that uses multi-layer neural networks to learn nonlinear relationships from data.

Recurrent neural networks (RNNs) provide a suitable framework for rainfall–runoff modelling because they process sequential inputs while updating a hidden state through time. This structure is relevant because rainfall–runoff response depends on both recent meteorological forcing and antecedent catchment conditions. A basic RNN updates a hidden state through time, allowing previous forcing conditions to influence later predictions. However, standard RNNs are limited by vanishing and exploding gradients, which restrict their ability to preserve long-term information (Bengio et al., 1994). LSTM networks address this limitation through a gated cell

state that controls how information is added, retained and released (Hochreiter & Schmidhuber, 1997).

Kratzert et al. (2018) argue that the LSTM cell state can be interpreted as a learned representation of slowly varying catchment storage, while the input, forget and output gates regulate when precipitation, temperature and antecedent wetness should affect runoff generation. In their pioneering large-sample study, Kratzert et al. (2018) further showed that LSTM models trained on meteorological forcings could achieve performance comparable to or better than a conceptual hydrological model benchmark. Kratzert, Klotz, Shalev, et al. (2019) further demonstrated that a single LSTM trained across hundreds of basins using meteorological forcings and static catchment attributes could outperform both regional and locally calibrated hydrological benchmarks. These results showed that DL models can learn regional hydrological patterns across multiple basins. Static catchment attributes are important in this framework because they provide information on why basins respond differently to similar meteorological forcing. By encoding controls such as catchment size, topography, soil, vegetation, aridity and snow influence, these attributes allow a regionally-trained LSTM model to produce catchment-specific rainfall–runoff responses (Kratzert, Klotz, Herrnegger, et al., 2019; Kratzert, Klotz, Shalev, et al., 2019).

This regional learning framework is relevant to prediction in ungauged basins (PUB), where streamflow records are unavailable for local calibration (Sivapalan et al., 2003). Traditional regionalization methods typically transfer model parameters from gauged to ungauged basins using spatial proximity, physical similarity or regression relationships (Oudin et al., 2008). DL instead learns a direct mapping from meteorological forcings and catchment attributes to streamflow across many gauged basins. When applied to an ungauged basin, the model uses its meteorological forcings and static attributes to estimate streamflow without calibrating model parameters to local discharge observations. Arsenault et al. (2023) showed that LSTMs clearly outperformed several traditional model-dependent regionalization methods for streamflow prediction in ungauged basins. The advantage of this approach lies in learning regional hydrological behavior from data rather than transferring calibrated parameter sets. Despite these strengths, LSTM models remain constrained by a common limitation of data-driven models: their predictions are less reliable when hydrological conditions fall outside the

training data. This limitation is particularly relevant for rare extremes, such as unusually large flood peaks or severe drought events, where relevant examples may be scarce during training (Frame et al., 2022). Moreover, Baste et al. (2025) found that LSTMs exhibit an upper limit in predicted discharge under synthetic extreme-precipitation experiments.

Furthermore, other sequential DL architectures have been explored as alternatives or complements to LSTM for streamflow prediction. Liu et al. (2024) found that a vanilla Transformer was not competitive with LSTM on the CAMELS-US dataset, particularly for high-flow metrics. They attributed this limitation partly to a mismatch between vanilla attention and rainfall–runoff dynamics: streamflow depends on evolving catchment states and water-storage memory, whereas the vanilla Transformer does not explicitly maintain a recurrent hydrological state. Wang et al. (2025) reported that a state space model (SSM) performed similarly to LSTM on the CAMELS-US dataset, with slight improvements in overall accuracy metrics but no consistent advantage across regions or flow-bias metrics. Overall, current evidence suggests that alternative sequence architectures have not yet demonstrated a consistent or decisive advantage over LSTM for standard large-sample rainfall–runoff simulation, and broader hydrological evaluation remains needed.

Sequential DL architectures such as LSTM, Transformer and SSM typically use basin-specific, spatially aggregated time series as inputs. In this formulation, the contributing area upstream of each hydrometric station or point of interest is treated as a lumped hydrological response unit, and spatial variability within the basin is largely removed. Graph neural networks (GNNs) provide a different approach because gauging stations, sub-catchments or river reaches can be represented as nodes, while flow connections can be represented as edges. Previous studies have used GNN-based models to incorporate river-network structure and river routing into streamflow prediction (Sun et al., 2022; Mosaffa et al., 2026). However, such models commonly depend on basin-specific discretization or river-network topology, which are difficult to standardize across diverse basins. This limits their direct use in large-sample rainfall–runoff modelling, where models are typically trained across many basins within a unified framework and evaluated for their ability to generalize to ungauged basins.

Although LSTM models can achieve high predictive accuracy, their predictions are not necessarily constrained by hydrological principles. For example, trained LSTM models may

generate discharge under hypothetical no-precipitation conditions or show weak consistency with water-balance expectations, even when standard performance metrics remain high (Yokoo et al., 2022). This concern has motivated physics-guided and hybrid modelling approaches. Frame et al. (2022) showed that adding mass conservation to an LSTM did not improve extreme-flow prediction; the unconstrained LSTM generally performed better for peak flows and out-of-sample events. This result indicates that physical constraints must be incorporated carefully, because a simplified constraint can reduce model flexibility without adequately representing unobserved fluxes and missing processes.

A more flexible hybrid direction is to make process-based or conceptual hydrological model structures differentiable, allowing model parameters or embedded neural-network components to be learned through gradient-based training (C. Shen et al., 2023). Unlike purely data-driven predictors, this framework retains process-based model structure while using neural networks to improve parameterization or selected process representations. Feng et al. (2022) showed that differentiable process-based models can approach LSTM-level accuracy while also producing physically interpretable internal variables such as snow storage, soil water, groundwater and evapotranspiration. However, differentiable hydrological models are still influenced by the assumptions and simplifications of the selected hydrological backbone (Feng et al., 2023). Applying this framework to a different hydrological backbone requires either implementing the model directly in a differentiable and numerically stable form, or using a differentiable surrogate to approximate the behavior of the original model when full reimplementation is impractical (Tsai et al., 2021).

1.2 Research Gaps

In large-sample hydrology, DL-based rainfall–runoff models are typically built on lumped basin representations. In this setting, each basin is treated as a single hydrological response unit: gridded meteorological forcings are clipped to the basin boundary and converted into basin-averaged or area-weighted time series, while physiographic attributes are summarized at the basin scale. This formulation has enabled sequential architecture (e.g., RNNs, Transformer, SSMs), to be trained across many basins using standardized large-sample datasets. However, it also removes much of the spatial variability within individual basins, such as localized precipitation patterns, elevation gradients, and soil heterogeneity. Architectures such as CNNs or

GNNs can, in principle, make greater use of spatially structured information. However, their applications commonly depend on basin-specific grids, subbasin discretization, or river-network structures, which are difficult to standardize across diverse basins. This limits the feasibility of training such models within a unified large-sample framework and constrains their direct transfer to basins without streamflow observations. This creates a methodological gap between models that can generalize across basins and models that can explicitly represent within-basin spatial heterogeneity.

DL-based rainfall–runoff models are typically trained across many basins, allowing them to learn a wide range of rainfall–runoff patterns and hydroclimatic regimes from spatially diverse datasets. Under this multi-basin training framework, it remains unclear whether model training still requires several decades of historical records. Older data may contain additional hydrological information, but their value may be reduced by hydroclimatic non-stationarity under climate change, as historical rainfall–runoff relationships may become less representative of the conditions encountered when flood forecasting models are deployed for operational prediction in later periods. Additionally, the quality of older data may also be affected by historical limitations in instrument precision, measurement practices, data processing consistency, and missing data. This raises a practical issue for large-sample model development, namely whether training datasets should be expanded both spatially, by including as many basins as possible, and temporally, by extending records as far as possible into the past. While spatial expansion has been demonstrated in previous studies (Kratzert et al., 2024), temporal expansion remains less explored. It is unclear whether shorter but more recent records across many basins can provide sufficient information for robust generalization.

High-frequency data further expand the modelling problem. Moving from daily to hourly rainfall–runoff modelling substantially lengthens the input sequence, increasing both computational demand and the difficulty of retaining hydrologically relevant information from earlier time steps. Moreover, standard LSTMs rely on recurrent sequence processing and sigmoid-bounded gates, which can limit their scalability and the extent to which stored information is revised when hydrological conditions change rapidly. These challenges motivate the assessment of architectures that retain the time-recursive state representation of LSTM while providing greater capacity for high-resolution and large-sample hydrological data.

1.3 Thesis Contributions

This thesis has five core contributions as follows.

1. A spatially distributed modelling approach was developed to incorporate within-basin spatial heterogeneity into LSTM-based rainfall–runoff modelling. This approach provides a practical hybrid modelling strategy that preserves the spatial generalizability of lumped LSTM models while introducing a finer-scale spatial representation. (Chapter 2)
2. The thesis demonstrates the value of finer-scale spatial representation for LSTM-based rainfall–runoff modelling, particularly in large basins where spatial aggregation may lead to the loss of hydrometeorological variability. (Chapter 2)
3. A systematic assessment of training data recency was conducted for large-sample LSTM-based rainfall–runoff modelling. This work provides empirical evidence that the recency of training data is an important consideration, rather than assuming that longer historical records are always more beneficial. (Chapter 3)
4. This thesis clarifies the relative roles of spatial diversity and temporal relevance in constructing training datasets for large-sample rainfall–runoff modelling, showing that shorter but more recent records can support robust model performance under multi-basin training. (Chapter 3)
5. An exploratory assessment of xLSTM-based architectures was conducted for large-sample hourly rainfall–runoff modelling. It provides preliminary evidence for future development of DL-based rainfall–runoff models designed for increasingly large hydrological datasets. (Chapter 4)

1.4 Thesis Structure

The chapters of this thesis are organized as follows. Chapter 2 presents a spatially distributed modelling approach that incorporates within-basin spatial heterogeneity to enhance lumped LSTM-based streamflow prediction. Chapter 3 presents an empirical assessment of how training-period selection affects LSTM-based rainfall-runoff modelling. Chapter 4 assesses the xLSTM architecture for large-sample hourly rainfall-runoff modelling. Chapter 5 concludes the thesis by synthesizing the major findings, discussing limitations, and outlining future research directions.

Chapter 2: Enhancing LSTM-based streamflow prediction with a spatially distributed approach

This chapter is a mirror of the following published article with minor changes to increase its consistency with the body of the thesis. References are unified at the end of the thesis.

Yu, Q., Tolson, B. A., Shen, H., Han, M., Mai, J., & Lin, J. (2024). Enhancing long short-term memory (LSTM)-based streamflow prediction with a spatially distributed approach. *Hydrology and Earth System Sciences*, 28(9), 2107–2122. https://doi.org/10.5194/hess-28-2107-2024

Summary

Deep learning (DL) algorithms have previously demonstrated their effectiveness in streamflow prediction. However, in hydrological time series modelling, the performance of existing DL methods is often bound by limited spatial information, as these data-driven models are typically trained with lumped, spatially aggregated input data. This study proposes a hybrid approach, namely the Spatially Recursive (SR) model, that integrates a lumped long short-term memory (LSTM) network with a physics-based hydrological routing simulation for enhanced streamflow prediction. The lumped LSTM was trained using basin-averaged meteorological and hydrological variables derived from 141 gauged basins located in the Great Lakes region of North America. The SR model applies the trained LSTM at the subbasin scale to generate local streamflow predictions, which are then routed to the basin outlet using the hydrological routing model. The efficacy of the SR model was evaluated for streamflow prediction at 224 gauged stations across the Great Lakes region and compared with that of the standalone lumped LSTM model. The results indicate that the SR model achieved performance levels on par with the lumped LSTM in basins used for LSTM training. Additionally, the SR model predicted streamflow more accurately in large basins, such as those with drainage areas greater than 2000 km², underscoring the substantial information loss associated with basin-wise feature aggregation. Furthermore, the SR model outperformed the lumped LSTM when applied to basins that were not part of the LSTM training, namely pseudo-ungauged basins. These findings suggest that lumped LSTM predictions, especially in large and ungauged basins, can be reliably improved by representing spatial heterogeneity at finer resolution through the SR model.

2.1 Introduction

Reliable streamflow prediction is critical in water resources management. Following recent developments in Artificial Intelligence (AI), an increasing number of hydrological studies have focused on adopting deep learning (DL) techniques, such as long short-term memory (LSTM), to improve basin-scale streamflow prediction compared to traditional physically-based hydrological models and conventional machine-learning (ML) algorithms (Kratzert et al., 2018; Frame et al., 2021; Gauch, Kratzert, et al., 2021). LSTM is a type of Recurrent Neural Network (RNN) that can capture long-term dependencies in hydrological time-series data and has demonstrated promising results in tasks such as streamflow prediction (Kratzert et al., 2018; Gauch, Kratzert, et al., 2021), precipitation forecasting (Tao et al., 2021), and drought monitoring (Wu et al., 2022). A recent large-sample model intercomparison study, namely the Great Lakes Runoff Intercomparison Project in the Great Lakes region (GRIP-GL; Mai et al., 2022), showed that LSTM model exhibited significant superiority in streamflow predictions compared to 12 other physically-based hydrological models, regardless of whether they were lumped or spatially distributed (Mai et al., 2022). The development of LSTM in hydrology has been driven by the need for more accurate and sophisticated models that can handle the complex and non-linear relationships in hydrological processes.

Despite of the recent popularity of data-driven modelling (e.g., ML and DL models) in hydrological modelling studies, process-based hydrological models (physically-based or conceptual model) continue to be used for operational streamflow forecasting. In contrast with data-driven prediction, traditional process-based models often rely more on spatially distributed representation of the region or basin being simulated. They utilize gridded meteorological forcings and, more importantly, break up the basin into various smaller response units such as grid cells (fully-distributed model) or subbasins (semi-distributed model). Compared to lumped models, distributed models take into account spatial variability at a finer resolution and also incorporate the routing process within the simulated basin. With the recognition that data-driven models and process-based models possess distinct advantages, the hybridisation of these two types of models has drawn growing attention in environmental modelling studies. Hybrid models can be categorized into two primary structural types: serial and parallel. In most cases, serial hybrid model involves the sequential coupling of one data-driven model and one process-based model (Hunt et al., 2022). This is typically achieved by feeding (to train) a data-driven model

with the outputs of a process-based model (Frame et al., 2021; W. Liu et al., 2021; Nevo et al., 2022), which implies that the data-driven model usually serves as a post-processor within a hybrid modelling workflow. On the contrary, Bindas et al. (2024) employs a DL network to infer the parameterizations of a river routing model for enhanced streamflow prediction. In a parallel hybrid model, the data-driven model and process-based model are integrated in parallel with each simulating different processes (Slater et al., 2023). In general, these hybrid modelling approaches allow researchers, to a certain extent, to incorporate spatial variability of input variables into the data-driven prediction scheme.

It is widely acknowledged that having ample training data is advantageous for DL models. Kratzert et al. (2024) showed that LSTM models trained across many basins generally achieve better predictive performance than models trained separately for individual basins. Nevertheless, in previous studies regarding LSTM-based streamflow prediction (to the best of our knowledge), most regionally-trained LSTM models only consider the spatial heterogeneity between training basins where LSTM inputs are at the lumped training basin scale. That is, each attribute (i.e., LSTM input variable) is computed for the entire basin and the basin is considered a single response unit such that streamflow is only predicted at the outlet of the basin (Kratzert et al., 2018; Kratzert, Klotz, Shalev, et al., 2019; Feng et al., 2020; Gauch, Kratzert, et al., 2021; Xie et al., 2022; Arsenault et al., 2023; Tang et al., 2023; Pokharel et al., 2023). Typically, this involves cropping the gridded dynamic input variables (e.g., precipitation and temperature) to the basin polygon and calculating a basin-average (or weighted average) to produce the lumped time series of the dynamic input variables (Lees et al., 2022).

The rationale behind employing a lumped model is due to the architectural limitation of LSTM networks which are not compatible with gridded data (i.e., image-like data) with various shapes (i.e., basin outlines), as inputs. Additionally, a lumped model enables effective learning of static basin attributes, such as drainage area and frequency of high precipitation (Kratzert, Klotz, Shalev, et al., 2019). However, the aggregation of climatic forcings results in neglecting the heterogeneous spatial distribution of various rainfall events. A study by Hunt et al. (2022) explains that a lumped LSTM probably failed to appropriately characterize rainfall over a large and arid basin due to averaging rainfall to the basin scale. Wang and Karimi (2022) argue that lumped LSTM rainfall-runoff models are unable to fully utilize the spatial variability of input

features. In their experiments, the spatial variability of rainfall was represented by a 20-element vector feature. For each of the 10 basins where they trained the LSTM, the vector consists of rainfall data at all hydrological response units within the basin. Their results show that LSTM trained on spatial-distributed rainfall data outperformed those driven by basin-averaged rainfall data. Nonetheless, their method does not clarify how to generalize the process of supplying the LSTM with spatially distributed rainfall information in the context of predicting outcomes in ungauged basins (PUB).

Conceptually, the predictive accuracy of lumped hydrological modelling will eventually degrade as basin size increases. This is due to the fact that meteorological forcing inputs are simply not well approximated by assuming they are constant over space. For example, in their paper presenting the CAMELS dataset of attributes for 671 basins in the contiguous United States, Addor et al. (2017) caution against using the largest of these basins for lumped hydrological modelling. They argued that the significance of basin-averaged input attributes diminishes with an increase in basin drainage area, because larger basin tends to necessitate a heightened consideration of spatial heterogeneity, requiring the incorporation of a spatially distributed representation. Nevertheless, the threshold at which basin area leads to poor lumped model performance is not precisely known and will likely vary by watershed location. In a study of benchmarking multiple hydrological models, Newman et al. (2017) excluded basins in the CAMELS-US dataset over 2000 km^2 in drainage area. Additionally, the lumped LSTM modelling study by Kratzert, Klotz, Shalev, et al. (2019) and the original set of basins in the Caravan dataset (Kratzert et al., 2023) both employ 2000 km^2 as an upper threshold. Given the past use of the 2000 km^2 threshold, we apply this criterion to define a ‘large basin’ in our study.

Hydrological routing across a drainage network is a general technique commonly used in distributed hydrological models, and it allows the model to represent the transport of water more accurately throughout large heterogeneous basins. Bindas et al. (2024) presents a novel differentiable river routing method to improve the streamflow prediction in a single large basin. They employed a regionally-trained LSTM to predict discharge at the subbasin-level and then map the LSTM predictions to a river network for routing. The results of the final routing model show promise by simulating predicted subbasin-level discharge from a lumped LSTM. However, their study presents limited empirical evidence demonstrating the superiority of their routing

model over the lumped LSTM. Specifically, the comparison was conducted within one single basin for a short testing period of one year, and the routing model outperforms the lumped LSTM in one of the three untrained gauges with larger than 2000 km^2 of drainage area. As for methodology, an intermediate process (scale-specific and basin-specific) is required to translate the LSTM predictions into lateral flow inputs for each reach in the river network. Furthermore, the routing model requires training of a multilayer perceptron network to update the parameters. As such, their basin-scale approach is challenging to directly generalize to new basins, or to different spatial scales for modelling.

Our study aims to identify an easy-to-implement, generalizable, regional-scale (or larger) approach for applying spatially-distributed inputs to effectively improve upon lumped data-driven streamflow prediction, especially in large, ungauged basins. In pursuit of this goal, we propose the Spatially Recursive (SR) model. The SR model first employs a lumped LSTM model (regionally-trained on a large sample of basins) to predict local streamflow at subbasins discretized from the basin of interest. Then, it utilizes a semi-distributed hydrological routing-only model, capable of explicit lake simulation, to route subbasin streamflow to the basin outlet. The data-driven prediction model is considered spatially recursive because it is trained at the basin scale and further applied at the subbasin scale to incorporate finer-resolution forcing data and subbasin attributes.

This chapter is structured as follows: Section 2.2 provides a description of the SR model, datasets, and experimental design. Section 2.3 presents key results and discussion. Finally, Section 2.4 concludes the findings.

2.2 Material and Methods

2.2.1 Overview of Spatially Recursive (SR) model

The proposed SR model is composed of three workflow components (see Figure 2-1), a regional LSTM for basin outlet streamflow prediction that is trained using a large sample of basins, a vector-based lake-river routing network that breaks up basins into subbasins, and a process-based routing model that only simulates the movement of LSTM predicted subbasin-level streamflow through the routing network. The fundamental concept of the SR model is to firstly employ a regional LSTM to predict local streamflow at each subbasin outlet (delineated in the lake-river routing network) for a basin where streamflow at the basin outlet is of interest.

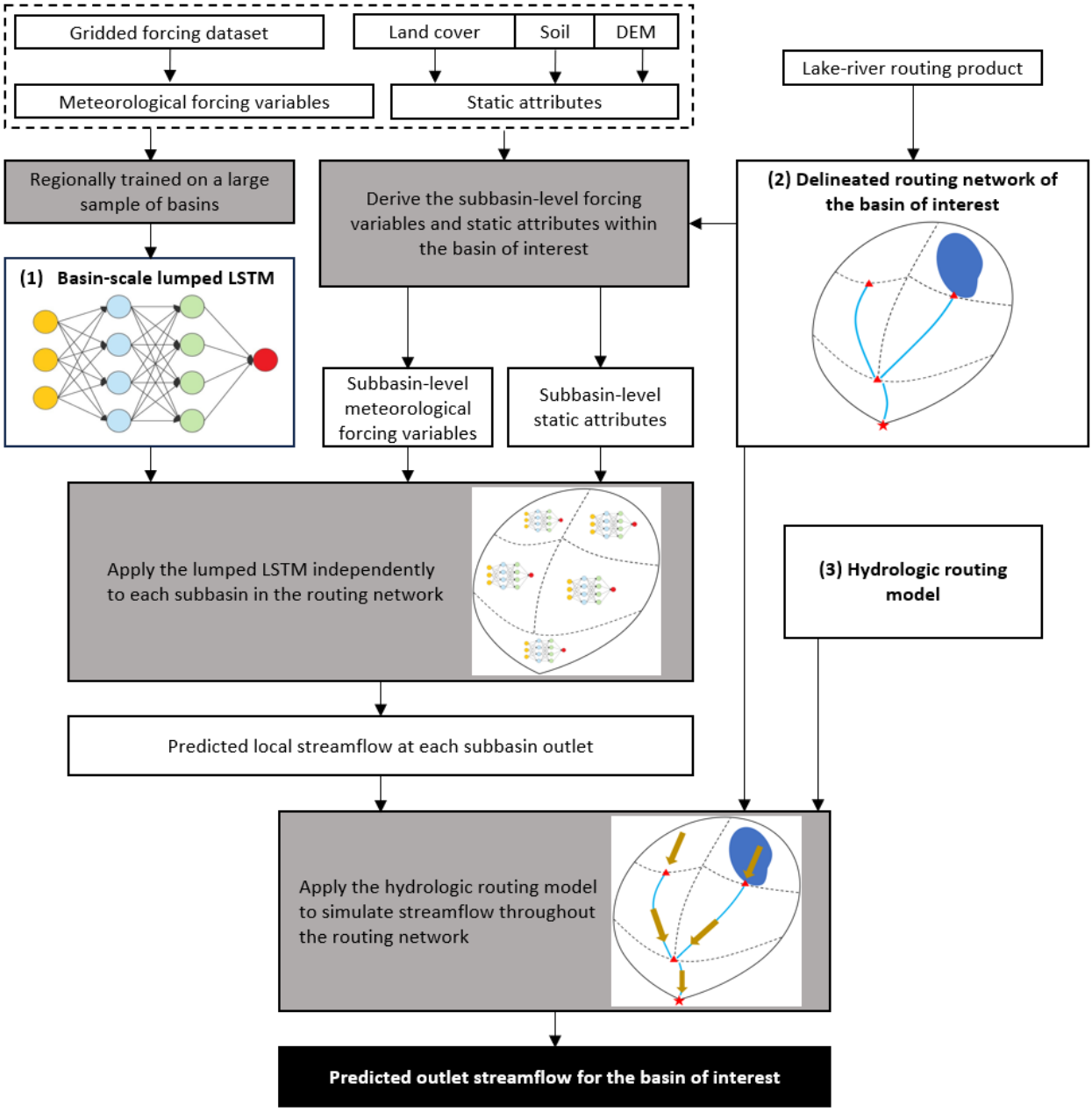


Figure 2-1. The spatially recursive model workflow for an arbitrary basin of interest. Note the workflow is conceptually model-agnostic (e.g. the lumped LSTM can be replaced by another data-driven model), while each component needs to generate outputs compatible with other components. Components with multiple arrows are utilized multiple times throughout the workflow.

Note that local streamflow for a subbasin is defined as the streamflow at the subbasin outlet that would occur if there were no upstream subbasins. Then, the predicted streamflow at each subbasin outlet serves as input to the routing-only model, which simulates how water is transported through the lake-river routing network and ultimately the streamflow at the basin

outlet. The subbasin-scale predictions are therefore used as intermediate inputs to provide spatially distributed local inflows to the routing-only model, so that the final basin-outlet prediction can account for within-basin variability in meteorological forcings and subbasin attributes while explicitly representing lake–river routing.

In general, none of the three workflow components are novel when considered individually. The regionally-trained LSTM is applied to directly produce the local streamflow at the subbasin-scale (i.e., the spatial scale of the routing model). Thus, no need to apply scale transformation when we use the LSTM prediction as the direct input to the routing model. Additionally, the SR model does not require any further training/calibration of either the regional LSTM or the routing model, making it generalizable to any designated gauge or basin outlet within the LSTM training region. The lake-river routing product, such as the MERIT-hydro global database (Yamazaki et al., 2019) and the North American Lake-River Routing Product (NALRP; Han et al., 2020), which features vector-based watershed discretization, should typically have a finer spatial resolution to break up each basin polygon into subbasins, at least for the majority of the streamflow gauging stations used for LSTM training. The following subsections explain the specifics of the components we selected in this study to demonstrate the spatially recursive modelling approach.

2.2.1.1 GRIP-GL Lumped LSTM

In this study, we precisely replicated the lumped LSTM built for streamflow prediction in the Great Lakes region of North America by Mai et al. (2022). That is, we trained our version of the lumped LSTM model using the same hyperparameters and input features as the LSTM model trained in the GRIP-GL project. The decision to employ an existing lumped LSTM for streamflow prediction, rather than attempting to add more basins and retrain a new LSTM, was intentional as we wanted to demonstrate explicitly that the proposed spatially distributed methodology works to improve upon an existing lumped LSTM without the need for LSTM retraining.

A brief summary of the model setup, training, and testing procedures of the lumped LSTM is presented here. Section A2-3 details the hyperparameter settings (Table A2-1) and the full list of all model input features (Table A2-2 and Table A2-3). Following the training procedures outlined by Mai et al. (2022), our trained model is an ensemble of 10 LSTM models

with the same architecture but different random seeds, and the final prediction is the average of the 10 models' outputs. Each LSTM model was simultaneously trained on 141 gauged basins (also referred to as 'calibration basins' in the GRIP-GL project) located in the Great Lakes region, over the period spanned from January 2000 to December 2010 (referred to as 'calibration period' in GRIP-GL). The LSTM model was constructed using the NeuralHydrology Python library (Kratzert et al., 2022). It was implemented to conduct sequence-to-one prediction, that is, the LSTM model uses a 365-day input sequence ending on the target date to predict the daily average streamflow on that date.

The input features were derived for each basin, which include the target variable (observational streamflow at the daily time scale), 9 dynamic variables (i.e., meteorologic forcings, listed in Table A2-2), and 30 static basin attributes describing soils, topography, land cover types, and climate (listed in Table A2-3). The observed discharge data is from either Water Survey Canada (WSC) or United States Geological Survey (USGS). The meteorologic forcings and climatic attributes were taken from the Canadian Surface Reanalysis Version 2 (CaSR-v2; Gasset et al., 2021). CaSR-v2 is a gridded reanalysis product which covers North America with a 10 km by 10 km spatial resolution on an hourly time step and this product was downloaded for the region from the CaSPAr archive (Mai et al., 2020). Soil attributes were derived from the Global Soil Dataset for Earth System Models (GSDE; Shangguan et al., 2014). Topological attributes (e.g., mean elevation, mean slope) were computed from the HydroSHEDS digital elevation model (DEM) product (Lehner et al., 2008). And the North American Land Change Monitoring System (NALCM; CEC, 2024a) product was used to derive the land cover attributes. The target variable and dynamic variables were aggregated from hourly to daily timescale. All dynamic variables and static attributes were spatially averaged for each gauged basin in the study. All input features derived for our replication of the GRIP-GL LSTM model were completely consistent with the derived input features in the original GRIP-GL study. This consistency check against GRIP-GL was important because our LSTM input derivation scripts are reapplied here for numerous and generic smaller subbasin polygons. Most importantly, the resultant trained GRIP-GL LSTM model rebuilt here generated practically identical quality hydrographs as the GRIP-GL LSTM in Mai et al. (2022) with differences in the median KGE performance metric of less than 0.01 (due only to different random seeds used in training).

After the model training, the lumped LSTM model was evaluated in three validation experiments according to the testing procedures outlined by Mai et al. (2022). These three validation experiments are used consistently throughout this study and are listed as follows:

1. **Temporal validation**, conducted for 141 calibration basins which were used to train the LSTM, predicting the daily streamflow over the period from January 2011 to December 2017 (referred as 'validation period' in GRIP-GL).

2. **Spatial validation**, conducted for 71 basins which were not used for model training (referred as 'validation basins' in GRIP-GL), predicting the daily streamflow over the training/calibration period (January 2000 to December 2010).

3. **Spatiotemporal validation**, conducted for the 71 validation basins, predicting the daily streamflow over the validation period (January 2011 to December 2017).

Note that in this study we used the term 'validation' and 'testing' interchangeably in order to be consistent with the experimental design of GRIP-GL, and it is not same as the 'validation' terminology normally used in machine learning applications.

2.2.1.2 Spatially distributed prediction based on a lake-river routing network

The lake-river routing network for an arbitrary basin defines the connectivity between the lakes/reservoirs, river channels and subbasins, as well as initial values for subbasin, lake, and channel characteristics required for running a spatially distributed hydrological simulation. A routing product is defined as a collection of routing networks covering large geographic regions and all included networks in a routing product should be delineated using the same source Geographic Information System (GIS) products (Han et al., 2023).

In this study, we tested our SR model with two routing products, the GRIP-GL common routing product (used in Mai et al. (2022)), and the NALRP v2.1 (Han et al., 2020). Both products were generated by the BasinMaker Python library (Han et al., 2023), which supports delineation of vector-based routing networks from any DEM and user-defined lake polygons. The GRIP-GL common routing product was derived from the HydroSHEDS DEM, with a spatial resolution of 3 arcseconds. The river network and subbasin (i.e., discretization) were defined by a constant flow accumulation threshold of 5000. That is, for a given point of interest, the contributing drainage area would be at least 5000 DEM cells, which corresponds to

approximately 40.5 km^2. On the other hand, the NALRP was produced based on the MERIT DEM (Yamazaki et al., 2019) with the same 3 arcseconds resolution, and the value of flow accumulation threshold is 2000 DEM cells (approximately 16.2 km^2). Additionally, attributes of lakes were taken from the HydroLAKES database (Messager et al., 2016) for both the GRIP-GL routing product and the NALRP. While the NALRP product included all lakes in HydroLAKES, the GRIP-GL routing product does not include small lakes with an area less than 5 km^2.

We use the GRIP-GL routing product directly in order to replicate in our SR model the precise routing networks Mai et al. (2022) used for their semi-distributed hydrological models. The average subbasin size in the routing networks of all GRIP-GL basins is approximately 131 km^2. In contrast, we used the NALRP product here to provide flexibility to run our SR model in non-GRIP-GL basins and to evaluate SR-model sensitivity to alternative routing model configuration decisions. The BasinMaker library includes post-processing functions to further simplify the routing networks in the NALRP. When applying simplification to the routing network, the resolution of the routing network was reduced (Han et al., 2023) by incorporating fewer vector elements (river channels, subbasin polygons etc.) into the routing network. For example, users can specify a minimum lake area threshold to remove the lakes with an area smaller than the threshold, and a minimum subbasin drainage area (MDA) threshold to merge upstream subbasins (with a drainage area smaller than the MDA threshold) to their downstream subbasin. It should be noted that the subbasin will not be merged if there is a lake or gauge located within it. Figure 2-2 shows a single basin discretized into three example lake-river routing networks, including the GRIP-GL routing network in Figure 2-2A, the original high-resolution NALRP network in Figure 2-2B, and a simplified NALRP network in Figure 2-2C (derived from the original NALRP network by applying BasinMaker functions). Note that our lake subbasins include only one lake which is completely contained within the subbasin boundary and our non-lake subbasins only have a single channel reach.

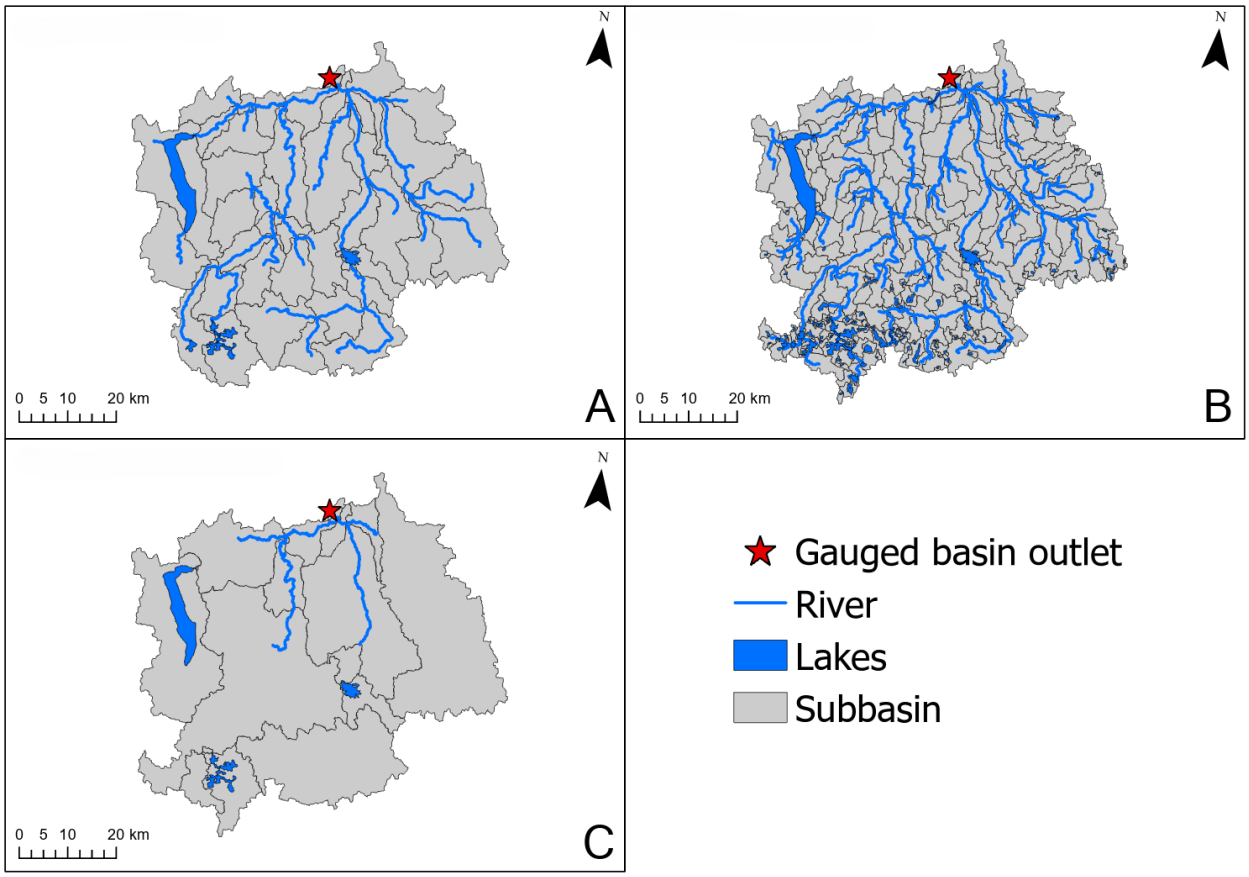


Figure 2-2. An example lake-river routing network for the Ontonagon River Watershed (USGS gauge ID: 04040000), which is one of the GRIP-GL calibration basins. (A) The network from the default GRIP-GL common routing product. (B) The network from the initial NALRP without simplification. (C) The simplified NALRP network delineated by a minimum lake area of 5 km^2 and a minimum subbasin drainage area of 500 km^2.

As shown in Figure 2-1, given the trained LSTM and the corresponding lake-river routing network discretizing a basin into subbasins, LSTM input features at the subbasin level need to be derived. This derivation is based on the original geospatial data and spatiotemporal data (see Section 2.2.1.1) and thus leverages the inherent spatial variability in the original data (as opposed to simply applying the basin-scale average input features in a basin to all the subbasins in that basin). With subbasin-level LSTM input features, the lumped LSTM is then deployed in each subbasin to predict the local subbasin outlet streamflow.

2.2.1.3 Routing-only mode in the Raven hydrological modelling framework

In this study, we constructed a physically-based routing-only model in the Raven hydrological modelling framework (Craig et al., 2020) to move local subbasin outlet streamflow

through the routing network in each basin. Note the lake-river routing network generated by BasinMaker incorporates all the necessary Raven routing model inputs and parameters (e.g., channel roughness, lake outlet characterization).

The intermediate results from our previous step in Section 2.2.1.2 (i.e., the LSTM local streamflow predictions at subbasin-level) can be seen as distributed subbasin-specific streamflow fluxes at various points within a routing network of streams and lakes (if any). The routing-only mode in the Raven framework can simulate the routing of distributed surface runoff (Craig et al., 2020; Han et al., 2020) but here we use it for the first time to route local subbasin outlet streamflow. This is accomplished within Raven by representing the local streamflow input (at each subbasin) as hourly precipitation instantly flushing to the subbasin outlet.

The routing model combines this local streamflow at the hourly timestep with the upstream subbasin streamflow that were routed from the subbasin inlet to the subbasin outlet. This process continues from upstream to downstream subbasins to the basin outlet. The routing model time step is hourly even though input streamflow from the LSTM are daily averages. We transformed the daily LSTM streamflow predictions to hourly routing simulation inputs by assuming the streamflow is constant over the day. That is, for a given date, the LSTM-predicted streamflow is assigned to all 24 hourly time steps. The simulated hourly streamflow at each basin outlet (stream gauge) is aggregated to daily streamflow which is the final prediction of the SR model.

The routing model is initialized with the lakes filled to the crest outlet elevation (point of zero outflow). For lake subbasins, the flushing operation sends all local subbasin streamflow into the lake instantly (rather than the subbasin outlet at the outlet of the lake) and then it becomes subject to the lake routing process in Raven. Since, water area is an input attribute in our LSTM, the LSTM has implicitly been trained to at least partially reflect lake routing impacts. Hence, our approach within Raven means that for lake subbasins, the local subbasin streamflow delivery to the subbasin outlet is only approximate as this typically small fraction of the total streamflow reaching the subbasin outlet has lake routing impacts applied in two ways instead of only once. This approximation is unavoidable within Raven modelling system but to be clear, the impacts are negligible given that for most lakes, local subbasin streamflow is only a small fraction of the total streamflow entering the lake considering all upstream subbasins.

The Raven framework provides the options to manipulate the routing algorithms and to calibrate routing-related parameters. In this study, we utilized the default configurations of the Raven framework and refrained from calibrating the routing-only model. We selected the diffusive wave channel routing option (where an analytical solution to the diffusive wave equation is used to relate inflow and outflow in each reach) and level-pool outflows from lakes are assumed to be governed by the broad-crested weir equation. Since subbasin-level streamflow is predicted by a calibrated (trained) lumped LSTM model, we assume that routing calibrated fluxes through a reasonably configured default routing model will typically yield reasonable quality results. Effectively, this approach provides a lower-bound estimate of the SR model performance given the routing model parameters are uncalibrated. The Raven configurations used in this study are described in Section A2.2.

2.2.2 Selection of additional gauging locations for concept validation

We posit that the effective scale of LSTM prediction (i.e., generalizability on various watershed sizes) might be affected by the range and distribution of the drainage area of the training/calibration basins, due to the variation of streamflow pattern in watersheds with different size. For instance, hydrological responses in small watersheds tend to be raging and flashy (Camera et al., 2020). The lumped LSTM was trained exclusively on basins with a drainage area exceeding 200 km^2 in accordance with the selection criteria of gauging basins in the GRIP-GL project. Considering the watershed delineation schemes (GRIP-GL routing and simplified NALRP) deployed in this study, many of the delineated subbasins would have a drainage area smaller than 100 km^2. Furthermore, the GRIP-GL basins show a skewed distribution in terms of sizes, with 109 out of the 141 calibration basins having a drainage area between 200 to 2000 km^2. Similarly, 56 out of the 71 validation basins fall within this range. To assess the LSTM performance at subbasin-level (smaller than 200 km^2) and larger watersheds (greater than 2000 km^2), we selected 12 additional gauging basins (not used in the GRIP-GL) in the Great Lakes region (within the bounding box defined by the minimum and maximum latitudes and longitudes of the GRIP-GL drainage basin) for spatial and spatiotemporal validation. These basins (summarized in Table 2-1) include four small basins with a drainage area below 100 km^2 and eight large basins ranging in size from 2000 to 7000 km^2.

Table 2-1. Summary of the 12 Non-GRIP-GL gauging basins selected for this study.

WSC or USGS station number	Description	Drainage Area (km^2)
02HC017	Etobicoke creek at Brampton, ON, Canada	69
01415000	Tremper kill near Andes, NY, USA	86
04LA006	Mollie River at Highway No.144, ON, Canada	93
04105700	Augusta Creek near Augusta, MI, USA	95
05129115	Vermilion River near Crane Lake, MN, USA	2343
02KF001	Mississippi River at Fergusons Falls, ON, Canada	2660
02KJ004	Dumoine (Riviere) A La Sortie Du Lac Robinson, QC, Canada	3760
02KB001	Petawawa River near Petawawa, ON, Canada	4120
04260500	Black River at Watertown, NY, USA	4827
01529950	Chemung River at Corning, NY, USA	5195
04LA002	Mattagami River near Timmins, ON, Canada	5570
04LF001	Kapuskasing River at Kapuskasing, ON, Canada	6760

All Non-GRIP-GL gauged basins are selected based on the following additional criteria:

1. The basin is not heavily regulated by dams or reservoirs.
2. The basin has less than 5% of missing data in streamflow observation for the study period.
3. The gauge ID at the basin outlet is included in the NALRP and thus defines a pre-existing routing network.

These criteria, along with the obvious requirement that none of the 212 existing GRIP-GL gauges could be used as additional testing basins, functioned to eliminate more than 1,000 streamflow gauges in the region from consideration and hence resulted in a relatively small sample size of additional test basins.

2.2.2.1 Comparison of different routing structures

This analysis aims to investigate the sensitivity of SR model prediction quality to the chosen delineation method (i.e., the routing network source and spatial resolution). BasinMaker postprocessing functions were applied to simplify the initial NALRP routing network for the 8 large Non-GRIP-GL basins. Firstly, small lakes were removed by using the same minimum lake

area threshold as the GRIP-GL routing product (5 km^2). Secondly, subbasins were merged by specifying the MDA threshold. For each basin, we delineated 7 routing networks which are defined as follows.

1. Mimic GRIP-GL routing, by using the same discretization strategy as the GRIP-GL routing product.
2. *NALRP_10%*, the MDA threshold was calculated as 10 percent of each basin's total drainage area.
3. *NALRP_100*, the MDA threshold is 100 km^2 for all basins.
4. *NALRP_300*, the MDA threshold is 300 km^2 for all basins.
5. *NALRP_500*, the MDA threshold is 500 km^2 for all basins.
6. *NALRP_800*, the MDA threshold is 800 km^2 for all basins.
7. *NALRP_1000*, the MDA threshold is 1000 km^2 for all basins.

2.2.3 Performance metrics

In this study, the Kling–Gupta Efficiency (KGE; Gupta et al., 2009) is used to evaluate the performance of the lumped LSTM model and the proposed SR model. KGE measures the degree of correspondence between two time series (e.g., observations versus a model prediction of those observations) and here it is computed for daily average streamflow time series. KGE is defined as follows in Equation 2-1:

$$KGE = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\alpha-1)^2} \qquad (2-1)$$

where r is the Pearson correlation coefficient which measures the linear correlation between the observed time series and predicted time series, β denotes the bias term which indicates whether the model is prone to overestimate or underestimate the streamflow, and α denotes error in flow variability. The range of KGE is (-∞, 1], where KGE = 1 signifies a perfect prediction. It is common to use KGE = 0 as the threshold for determining whether the model exhibits good predictive performance (Knoben et al., 2019). On the other hand, in the GRIP-GL project paper, Mai et al. (2022) carefully argues that in general, a KGE less than 0.48 would be considered a poor model and models with higher KGEs are medium or higher quality.

2.2.4 Experimental design

Three sets of experiments are used to evaluate the quality of the SR model. The first experimental set involved implementing the SR model on the four small Non-GRIP-GL basins. The main objective of this task was to validate the predictive capabilities of the lumped LSTM model in estimating streamflow at a local subbasin-level (LSTM extrapolation to small basins). This was conducted by testing the lumped LSTM on basins that are much smaller than the minimum drainage area in the training dataset. Additionally, we also implemented a single-subbasin routing model in Raven to ensure our approach to push local streamflow into Raven worked as expected.

In the second set of experiments, we utilized the SR model to predict streamflow on the 212 GRIP-GL basins. Of these, 141 basins were used for training the lumped LSTM (calibration basins), and 71 basins were not used for training (i.e., validation basins). This task aims to evaluate the overall performance of the SR model as compared to the lumped LSTM. Following the evaluation scheme of the GRIP-GL project, the KGE was calculated separately for temporal validation (trained locations and untrained period), spatial validation (untrained locations and trained period), and spatiotemporal validation (untrained locations and untrained period). Moreover, for validation basins (untrained locations), we calculated the KGE for the whole study period (2001-2017). As mentioned earlier, only the GRIP-GL common routing product was used as the routing structure in this task.

For the third set of experiments, the SR model was applied to the eight large Non-GRIP-GL basins. These basins are equivalent to the validation basins in the GRIP-GL project, where we had no previous knowledge or experience applying the LSTM. Each basin was tested with the seven routing structures described in Section 2.2.2.1, in order to investigate the impacts of using alternative routing networks. Note that these routing networks derived from MERIT DEM differ from those in the second set of experiments, which were derived from HydroSHEDS DEM, as described in Section 2.2.1.2.

For all experiments, the KGE metric was calculated for the lumped LSTM model predictions and for the semi-distributed SR model predictions, respectively. The lumped LSTM is the benchmark model for comparing with the integrated SR model.

2.3 Results and Discussion

Regarding all hydrographs (line plots) in this section, only the data from January 2009 to December 2012 were plotted (i.e., 2 years in calibration period and 2 years in validation period) for a better visualization, and the displayed KGE for each basin was calculated for the whole study period (2001 - 2017). Note that the lumped LSTM and SR models both make predictions starting from the year of 2001 as their LSTMs takes the year of 2000 as the initial input sequence.

2.3.1 Extrapolating LSTM to small Non-GRIP-GL basins

The lumped LSTM prediction quality on the four small basins (69 - 95 km^2) is quite good with KGE values for these new test locations of 0.785, 0.812, 0.634 and 0.489, respectively. These KGEs compare favourably with the median GRIP-GL validation performance level reported in Mai et al. (2022) of 0.767 for the same LSTM applied to much larger basins. Figure 2-3 shows the daily observed streamflow, the predicted streamflow from the lumped LSTM and SR model at each of these four small basins.

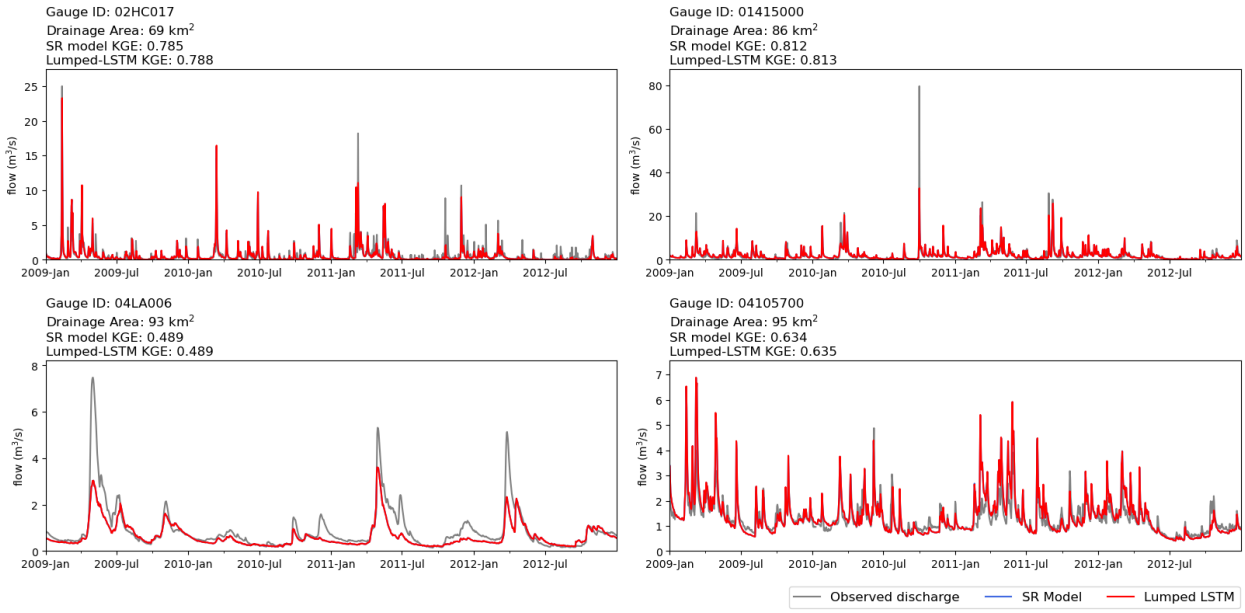


Figure 2-3. Comparison of observation (grey), lumped LSTM prediction (red) and SR model prediction (blue) for the 4 small Non-GRIP-GL basins. KGE values over the period from 2001 to 2017 are given for each basin.

The predictions from the SR model (blue) are not visible because, as expected, they are the virtually the same as the predictions from the lumped LSTM (red). This can be explained by the fact that the delineation is geometrically identical to the basin outline (i.e., no spatial

discretization and each basin is only a single subbasin). In the routing simulation, the LSTM-predicted subbasin streamflow would be directly flushed without delay to the basin outlet, making it equivalent to a lumped prediction. Overall, these results indicate the lumped LSTM adequately extrapolates to much smaller basins than those it was trained for and the translation of LSTM-predicted streamflow into the routing model is correct.

2.3.2 Comparing lumped LSTM and SR model on GRIP-GL basins

The SR modelling approach should be able to outperform the lumped LSTM for basins with large drainage areas, because spatially distributed modelling will mitigate the information loss caused by feeding the lumped LSTM with basin-averaged features, and the SR model will take advantage of maintaining such spatial heterogeneity. Therefore, the results are evaluated in two ways: results first include all basins (corresponding to the validation experiment in the GRIP-GL study), while the second way focuses solely on basins larger than 2000 km^2.

The performance results of the lumped LSTM and the SR model at GRIP-GL basins are summarized in Table 2-2 and visually compared in Figure 2-4. In the temporal validation experiment, the predictions by the SR model shows a comparable level of quality to that of the lumped LSTM. The interquartile range of the lumped LSTM is slightly narrower than that of the SR model, suggesting less variability in the KGE score distribution. The results indicate that the lumped LSTM better captures the temporal trends and seasonal patterns at trained locations, while the SR model, relying on an uncalibrated process-based routing model, shows no improvement (slightly reduced KGEs) relative to the lumped LSTM results for both large (> 2000 km^2) and small (< 2000 km^2) basins. However, the SR model results for all basins are better than the best of 12 physically-based/inspired GRIP-GL hydrological model (see Table 2-2).

In the spatial validation and spatiotemporal validation experiments, both the SR model and the lumped LSTM exhibit similar performance degradation relative to temporal validation performance (considering all basins). These two experiments primarily focus on assessing the models' robustness in predicting streamflow in an ungauged basin scenario (where no local streamflow observations were used to train the model). It is worth mentioning that the LSTM prediction at each subbasin is practically also a prediction of streamflow in an ungauged basin (refer to the first task described in Section 2.3.1).

Table 2-2. Median KGE for prediction performance of the lumped LSTM and the SR models. Top performing model is indicated by bold font.

Validation experiment	Number of basins	Best of other 12 models in GRIP-GL*	Lumped LSTM	SR Model
Temporal Validation	141	0.790	**0.819**	0.804
Temporal Validation (drainage area **over** 2000 km^2)	32	0.824	**0.840**	0.836
Temporal Validation (drainage area **below** 2000 km^2)	109	0.773	**0.810**	0.794
Spatial Validation	71	0.607	0.767	**0.779**
Spatial Validation (drainage area **over** 2000 km^2)	15	0.598	0.708	**0.821**
Spatial Validation (drainage area **below** 2000 km^2)	56	0.627	**0.778**	0.777
Spatiotemporal Validation	71	0.589	**0.744**	0.732
Spatiotemporal Validation (drainage area **over** 2000 km^2)	15	0.614	0.663	**0.747**
Spatiotemporal Validation (drainage area **below** 2000 km^2)	56	0.597	**0.758**	0.729

*The 12 models are physically-based / physically-inspired models.

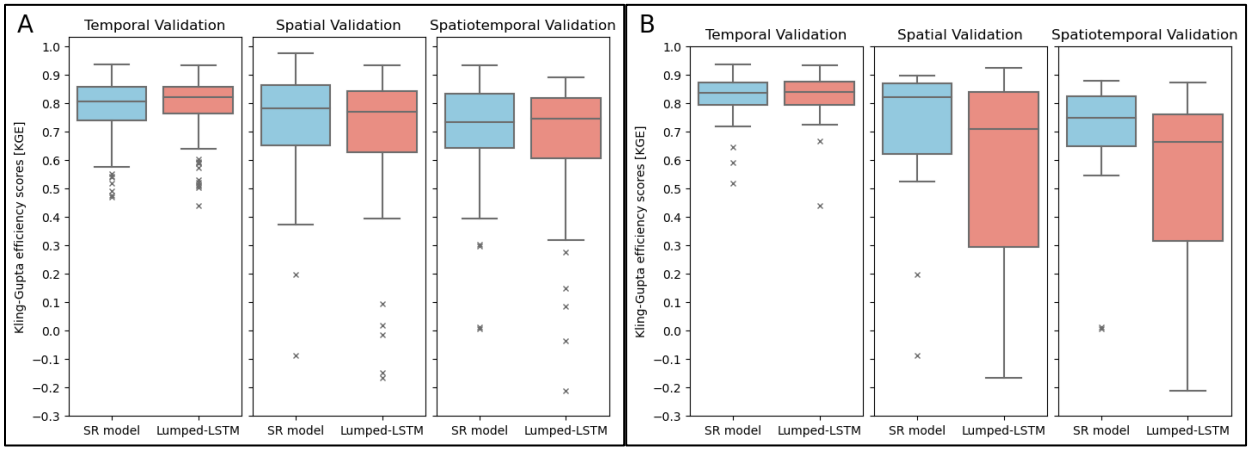


Figure 2-4. Boxplots of KGE validation scores for the SR model (blue) and the lumped LSTM (red). (A) The results of all basins participating in each validation experiment. (B) The results of basins with a size larger than 2000 km^2.

As with the temporal validation, when considering all basins in spatial or spatiotemporal validation, the SR model shows no real difference in median KGEs relative to the lumped LSTM results. However, the advantage of the SR model in larger basins becomes apparent in these untrained locations. As compared to the lumped LSTM, for large basins over 2000 km^2, the median KGE of the SR model is 0.113 KGE units higher in spatial validation, and 0.084 KGE units higher in spatiotemporal validation. These improvements for large basins came with no real performance drops for small basins (< 2000 km^2) as the median KGE values for both the SR

model and the lumped LSTM were within a 0.030 KGE units difference across all three validation modes (0.016 for temporal validation, 0.001 for spatial validation and 0.029 for spatiotemporal validation). Furthermore, the SR model results are substantially better than the best of 12 physically-based/inspired GRIP-GL hydrological models for all basins. This is notable given that seven of these 12 models in GRIP-GL were spatially-distributed (not lumped) and utilized the same routing network discretization as the SR model for each basin.

Figure 2-5 displays the time series of observations, model predictions, and KGE scores, for representative GRIP-GL validation basins. Figure 2-5A shows the hydrographs of the two basins where the SR model demonstrates the largest improvement (better by more than 0.5 KGE units) compared to the lumped LSTM. Among them, 02KF005 (Ottawa River at Britannia) is the largest basin studied in the GRIP-GL project, with almost a 90000 km^2 drainage area. The other basin is approximately 6923 km^2 in size (WSC gauge 02LG005, Gatineau Riviere Aux Rapides Ceizur). The deficiency of the lumped LSTM model is evident as it fails to capture the peaks and seasonal variations at these two large watersheds, and the lumped LSTM predicts constantly low flow throughout the study period. Figure 2-5B shows compares hydrographs at WSC gauges 02HM010 (Salmon River at Tamworth) and 02LB007 (South Nation River at Spencerville), where the prediction accuracy of the SR model shows the worst degradation relative to the lumped model (by 0.17 and 0.12 KGE units). These two gauges are small (588 km^2 broken into 9 subbasins in the SR model and 277 km^2 broken into 3 subbasins in the SR model) and within the range of training basin sizes. It is important to note that the spatial resolution of CaSR-v2 dataset is 10 km by 10 km, thereby each grid covers an area of approximately 100 km^2 and so breaking these small basins up into a handful of smaller subbasins is likely unnecessary in terms of representing the spatial rainfall patterns. In general, the SR model aligns with the overall pattern of the observed streamflow, but it tends to underestimate peak flows, and the lumped LSTM predicts larger peaks on these small basins. This underestimation of peak flow events is not evident in most of the high-quality SR model hydrographs and in fact in the larger two basins shown in Figure 2-5C, the SR model is predicting higher peaks than the lumped LSTM. Figure 2-5C shows the two basins where the SR model achieves the highest KGE scores (both over 0.9 and both large basins) and these are both notably improved over the lumped LSTM. The hydrographs of two basins where the SR model achieves the lowest KGE scores (both around 0) are shown in Figure 2-5D. It is evident that both lumped LSTM and SR model were unable to

simulate flow in these basins which, according to the observed hydrograph, appear to be substantially impacted by regulation. Their poor performance in regulated basins is unsurprising, as the LSTM training data do not indicate the degree of regulation.

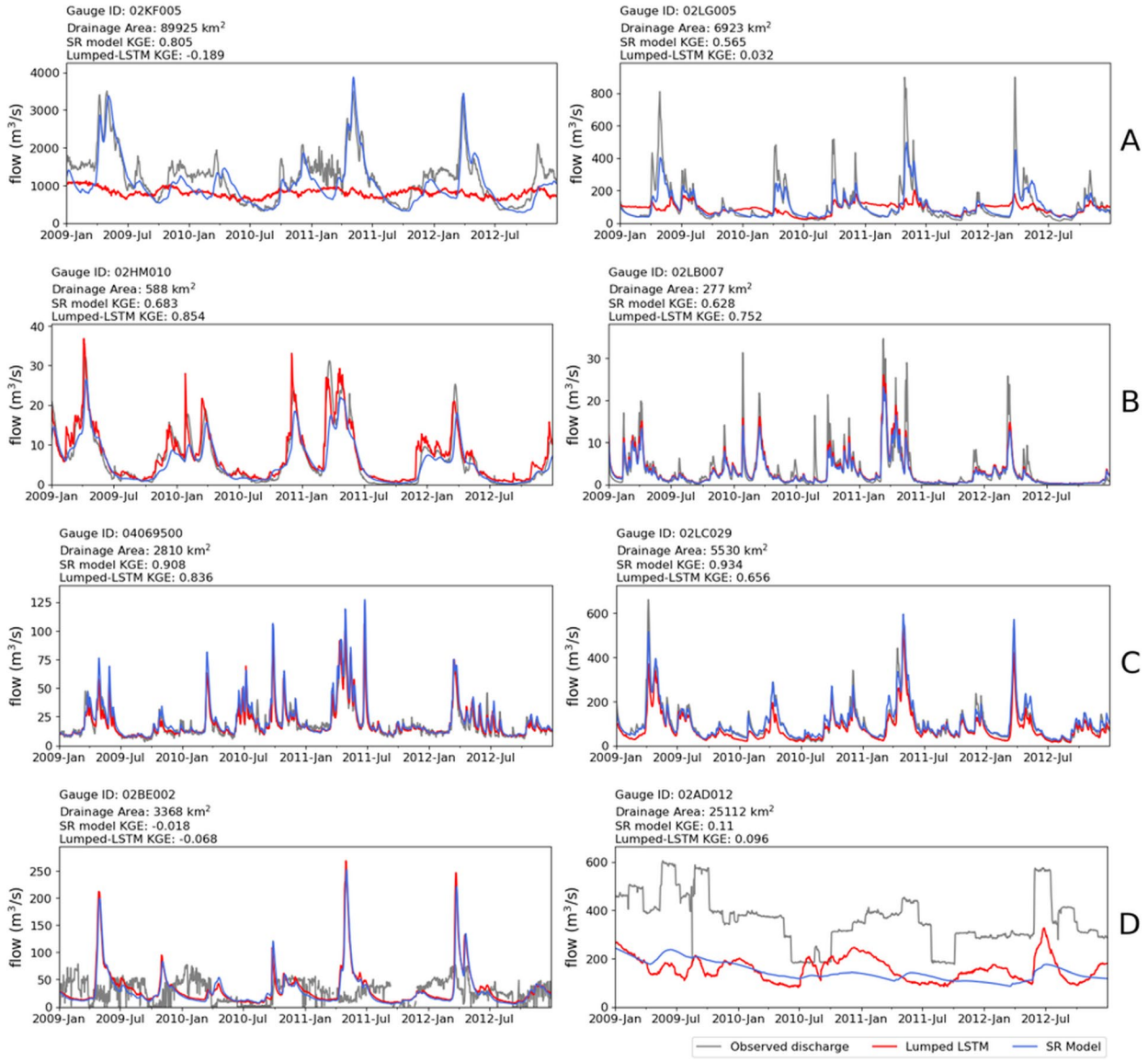


Figure 2-5. Comparison of time series of observation (grey), lumped LSTM prediction (red) and SR model prediction (blue) for representative GRIP-GL validation basins, over the selected period from 2009 to 2012. KGE values over the whole study period from 2000 to 2017 are given for each basin. (A) The 2 basins that show the most significant improvement as compared to the lumped LSTM. (B) The 2 basins that show the most significant degradation as compared to the lumped LSTM. (C) The two basins where the SR model achieves the highest KGE scores. (D) The two basins where the SR model achieves the lowest KGE scores.

2.3.3 New testing basins and impact of routing network delineation

Eight large basins (not used in the GRIP-GL study) were identified as suitable additional independent testing basins according to the criteria in Section 2.2.2 in order to conduct further comparisons between the lumped LSTM and the SR model built with varying routing networks. Like the GRIP-GL validation basins, neither model was trained on these eight new basins.

Figure 2-6 summarizes the comparative results and shows the KGE of each model in each of the eight basins. From Figure 2-6, it can be seen that the choice of the delineation method (routing network resolution) has a minor impact on prediction performance. In general, the differences in overall KGE scores among the seven different resolution routing networks are small at each basin. This could be attributed to the consistent representation of lakes across all the routing networks (all resolutions retain lakes more than 5 km^2 in area in the network) combined with the crucial role these lakes play in modelling the transport of water.

In terms of relative model performances, the SR model outperforms the lumped LSTM model in seven out of the eight tested basins (improved by an average of 0.160 KGE units, using the GRIP-GL routing network resolution). This result strongly reinforces the findings in Section 2.3.1, showing that the SR model tends to exhibit better relative performance in basins with larger drainage areas. The SR model achieved the worst KGE score in the basin identified by the USGS gauge 05129115 (Vermilion River near Crane Lake), and this is the only basin where the lumped LSTM (KGE score of 0.716) outperformed the SR model (KGE scores from 0.534 to 0.600). As depicted in Figure 2-6, the KGE scores at this basin exhibit a gradual decrease as the resolution of the routing network becomes coarser, such as with NALRP_800 and NALRP_1000. Figure 2-7 shows the hydrographs for USGS gauge 05129115 and while it is evident that while the SR model successfully captures the timing of the peaks, it consistently underestimates their magnitude in comparison to the observation and the lumped LSTM.

The degradation in performance could be attributed to the heavy presence of lakes within this basin (as shown in Figure 2-8A). Among the 8 tested basins, 05129115 stands out with the largest fraction of its area covered by lakes, accounting for approximately 14% of its total area. In contrast, the SR model shows significant improvement in basin 02KJ004, which has the second highest proportion of lake coverage, approximately 10% (as depicted in Figure 2-8B). However, the lake areas are vastly different in these two basins. Lakes in 05129115 are

concentrated mainly as one very large lake in the middle of the basin, whereas the lakes in 02KJ004 are more numerous, smaller, and generally long and narrow. On the other hand, the SR model obtains a substantial improvement over the lumped LSTM (by over 0.347 KGE units) in a lake-sparse basin gauged by USGS station 04260500 (see Figure 2-8C). Notably, this is also the additional testing basin where the SR model achieved the highest KGE score of 0.894.

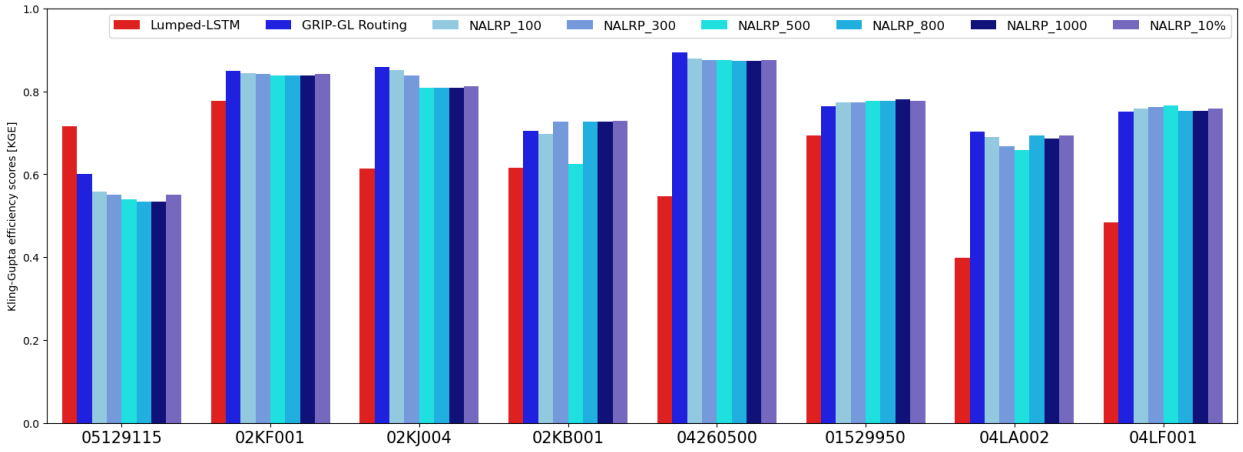


Figure 2-6. Overall model performances as adopting different routing network delineations in the 8 large Non-GRIP-GL basins, during the whole study period 2000-2017. The basins are sorted from left to right in ascending order according to their sizes, from smallest (2343 km^2) to largest (6760 km^2).

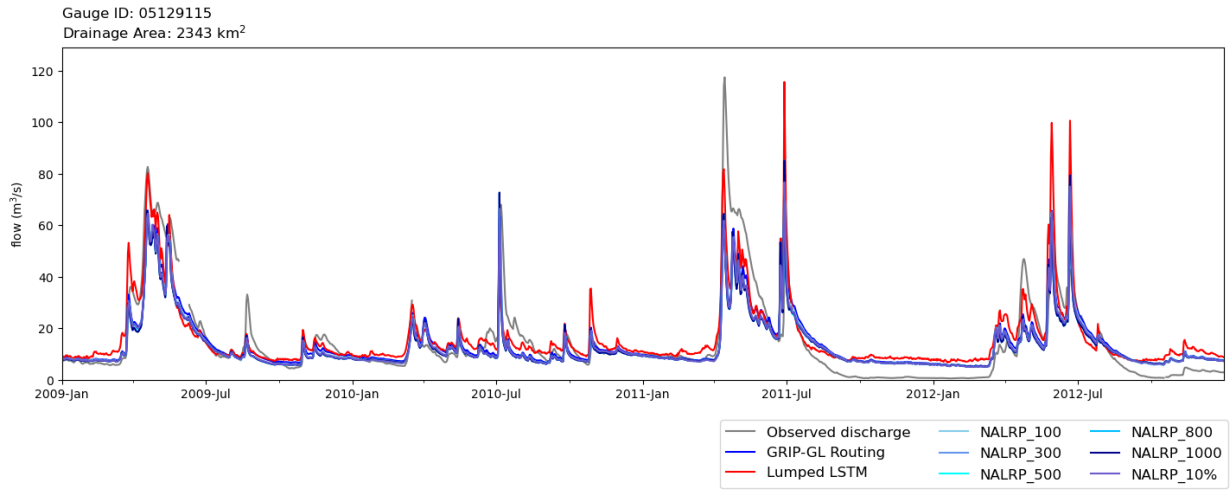


Figure 2-7. Comparison of the time series of observation (grey), lumped LSTM prediction (red) and SR model predictions with different routing networks for the Vermilion River basin 05129115. Note that the lines of NALRP time series all follow extremely similar trends as the GRIP-GL Routing series and thus are not distinguishable.

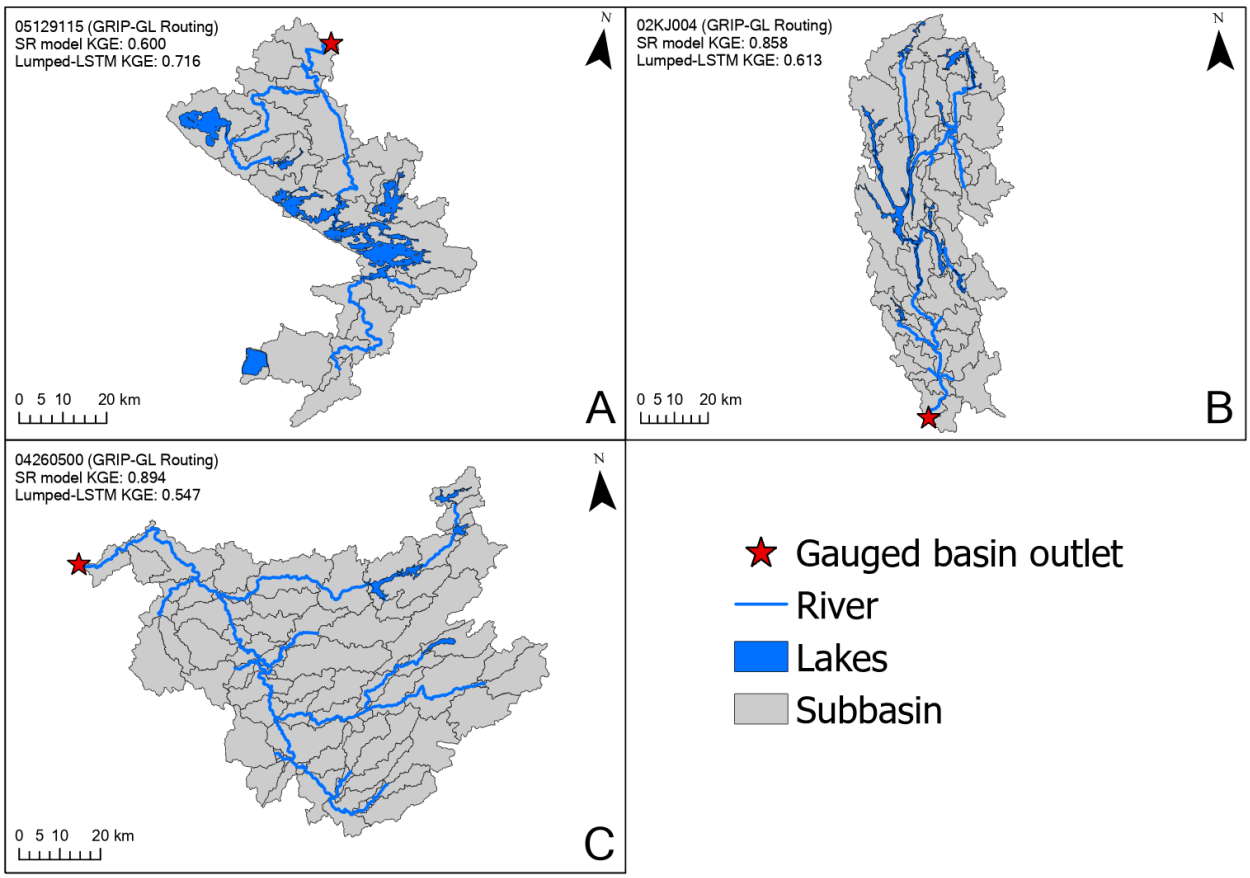


Figure 2-8. Comparison of the routing networks of 3 large Non-GRIP-GL basins to illustrate the variation in different lake densities. (A) Gauged basin 05129115 (5 lakes, 14% of basin is covered by lakes). (B) Gauged basin 02KJ004 (11 lakes, 10% of basin is covered by lakes) and (C) Gauged basin 04260500 (4 lake, 4% of basin is covered by lakes). The three illustrated routing networks were delineated to mimic the GRIP-GL routing delineation strategy.

2.4 Conclusion

In this study, we proposed a hybrid modelling approach named the Spatially Recursive (SR) model that aims to enhance the accuracy of streamflow predictions made by lumped LSTM model. For a basin of interest, a regionally-trained lumped LSTM is used to predict the local streamflow at the subbasin-level (as delineated in the basin's lake-river routing network), and then a process-based hydrological routing-only model simulates the transport of local streamflow from the subbasin outlet to the basin outlet. The novelty of the SR model is threefold: (1) It considers the spatial variability of input variables at finer spatial resolution by having smaller response units than the training dataset (i.e., in our case, this is from basin-scale to subbasin-scale); (2) It integrates physically-based lake-river hydrological routing with the data-driven learning to form a generalizable modelling approach for enhanced streamflow prediction in large

ungauged basins; (3) It operates without the need for further fine-tuning, parameter transfer, or training/calibration, given the trained LSTM is available.

Three sets of experiments were conducted to examine the applicability and performance of SR model. First, we validated the concept of predicting streamflow at local subbasin-level with an LSTM trained using much larger basins. This was done by predicting streamflow at four small testing basins (< 100 km^2) which were used as mimic local subbasins. The results revealed that the lumped LSTM can indeed be applied for predicting streamflow in basins below the minimum drainage area threshold of the training dataset.

Subsequently, the SR model was evaluated on 212 basins from the GRIP-GL project. The results showed that the SR model is comparable to lumped LSTM in terms of overall performance across basins of all drainage areas. The SR model exhibits a noticeable advantage in predicting streamflow in large basins (>2000 km^2), which demonstrates that incorporating spatially distributed inputs can be beneficial to the hydrological modelling in large basins, due to the fact that the spatial heterogeneity is naturally more significant in larger regions. The empirical performance improvement over the lumped LSTM is most significant in a PUB context. For the 15 large GRIP-GL validation basins, the median KGE levels over the 10-year training period and the 7-year testing period, are 0.11 and 0.08 KGE units higher, respectively, than those of the lumped LSTM. Importantly, for smaller basins (< 2000 km^2), the performance gains for large basins do not result in significant performance drops, as the median KGE difference of SR model and the lumped LSTM were within 0.03 KGE units in all three validation modes.

Lastly, we investigated the impacts of the routing network delineation by testing the SR model in 8 additional large basins (2343 km^2 to 6760 km^2). This out-of-sample testing showed that the SR model was substantially better than the lumped LSTM for seven of eight basins by an average of 0.16 KGE units, over the 17-year period. It further corroborated the superiority of our method for modelling large heterogeneous basins. Results also clearly show that the substantial performance gains of the SR model over the lumped LSTM are not sensitive to a range of routing network configurations, and these performance gains occur in basins with both large and small fractions of the basin covered by lakes.

Moreover, these improvements of the SR model relative to the lumped LSTM model did not require calibration or additional training after the original lumped LSTM was trained. In fact, the reported results of the SR model reflect a conservative estimate (i.e., lower bound) of performance, considering the following factors: (1) The regional LSTM within the SR model could be improved with additional training data (Kratzert et al., 2024); (2) The SR model routing parameters could be calibrated and regionalized to further improve validation results; and (3) The regional LSTM could be purpose-built to train on a sufficient number of small basins better matching the subbasin-level spatial scale at which the lumped LSTM would be applied within the SR model (e.g., 131 km^2 as the average subbasin size of the GRIP-GL routing product).

The findings of this study highlight the importance of considering spatially distributed inputs to streamflow prediction and demonstrate a new way that data-driven models can benefit from such information. This research opens up new avenues for future research regarding hybrid modelling in hydrology, by improving an existing data-driven model with uncalibrated hydrological routing approach. The simplicity of our approach, combined with the growth of regionally-trained LSTMs for streamflow prediction, means that our approach should be accessible to all hydrological modellers. Future refinement of the proposed SR modelling approach could focus on two key aspects: the training strategy of the lumped DL-based predictor (e.g., larger dataset, targeted training basin sizes, different neural networks etc.), and the calibration of hydrological routing-related parameters.

Chapter 3: The role of data recency and training period length in LSTM-based rainfall-runoff modelling

This chapter is a mirror of the following published article with minor changes to increase its consistency with the body of the thesis. References are unified at the end of the thesis.

Yu, Q., & Tolson, B. A. (2026). How much historical data do we need? The role of data recency and training period length in LSTM-based rainfall-runoff modelling. *Journal of Hydrology*, 668, 135046. https://doi.org/10.1016/j.jhydrol.2026.135046

Summary

Rainfall-runoff models based on Long Short-Term Memory (LSTM) networks typically require extensive datasets for training. While increasing the number of watersheds generally improves model performance, the appropriate temporal scope of training data—especially under potential non-stationarity from climate change—remains unclear, and the importance of data recency remains underexplored. This study investigates whether decades of historical data are necessary when training with large numbers of watersheds, and whether shorter, more recent training periods can match the performance of extended records. Using hydrometeorological data from 1374 North American watersheds (spanning 1950–2023), we systematically evaluate the effect of data recency on LSTM performance through three complementary experimental designs: (1) backward-expanding training periods (fixed recent data + progressively older blocks), (2) forward-expanding periods (fixed old data + progressively newer blocks), and (3) sliding-window periods (fixed-length windows moving forward). Results reveal that newer data universally improves predictions, while older data contributes marginally or negatively to model performance. Notably, the benefit of increasing the number of watersheds is conditional on data recency, as Prediction in Ungauged Basins (PUB) performance improves most significantly with recent data. These findings provide the empirical evidence for the importance of data recency, demonstrating that data recency—not just data volume—is critical for LSTM-based streamflow prediction, underscoring the need for recency-aware training strategies.

3.1 Introduction

Deep learning (DL) models typically require large volumes of data to achieve robust generalization. In the context of large-scale rainfall-runoff modelling, dataset size can refer to either the number of watersheds (spatial diversity) or the length of the training period (temporal coverage). Increasing the number of watersheds has been consistently shown to improve the performance of long short-term memory (LSTM) networks (Kratzert et al., 2024). Although several studies have reported that extending the training period can also benefit LSTM performance (Kratzert et al., 2018; Boulmaiz et al., 2020; Gauch, Mai, et al., 2021; Ayzel & Heistermann, 2021; Arsenault et al., 2023), some evidence from these studies also suggests that performance improvements may plateau beyond a certain training length. For instance, Boulmaiz et al. (2020) examined how LSTM performance depends on the length of training data at 20 gauged watersheds. Their experimental design involved five scenarios with progressively extended training periods of 3, 6, 9, 12, and 15 years, all ending in the year of 1999, thereby increasing temporal coverage by moving backward in time while keeping the validation (2000-2004) and testing periods (2005-2014) fixed. Results showed that LSTM performance improvements plateaued after nine years of training data, suggesting diminishing returns from including older observations. However, this finding is constrained by the small sample size of only 20 watersheds, limiting its applicability to large-scale LSTM modelling applications. Moreover, Gauch, Mai, et al. (2021) evaluated how LSTM model performance responds to variations in training period length and spatial coverage using a much larger dataset that includes 531 watersheds. They trained three sets of LSTM models, each with training periods starting from October 1999 and lasting for 3, 6, and 9 years, respectively. The testing period spans from October 1989 to September 1999. While the results confirm that longer training periods are beneficial, they also suggest that increasing the number of watersheds contributes more significantly to performance improvement—measured in Nash–Sutcliffe Efficiency (NSE)—than merely extending the training period, as summarized in Table 3-1 below.

Although these findings do not imply that a longer training period is unnecessary, they clearly highlight two points: (1) the marginal gain from extending the training period is small when the number of watersheds is large, and (2) the marginal gain from increasing the number of watersheds is substantial even when the training period is short. While longer training periods may expose the model to a broader range of hydrological conditions—thereby aiding

generalization and convergence—they may also incorporate outdated patterns that are less representative of present or emerging hydrological regimes.

Table 3-1. Performance improvements inferred from Gauch, Mai, et al. (2021) showing the effects of watershed quantity and training period length on NSE.

Change	Training Setup	Median NSE Before	Median NSE After	Improvement (ΔNSE)
Increasing number of Watersheds	3-year training period	0.43	0.68	+0.25
(13 → 531)	9-year training period	0.63	0.74	+0.11
Extending Training Period	13 watersheds	0.43	0.63	+0.20
(3 → 9 years)	531 watersheds	0.68	0.74	+0.06

One of the fundamental principles in hydrological modelling is the assumption of stationarity (Yang et al., 2021), which posits that the key statistical properties (e.g., mean, variance) of hydrological processes remain temporally invariant when represented by models, allowing for the use of historical data to predict future events (Deb et al., 2019). Once trained, models such as LSTM networks operate as fixed functions, mapping input sequences to output sequences using a static set of parameters. In this sense, a trained LSTM can be considered a model that assumes stationarity in the input-output relationship—not because the architecture itself imposes stationarity, but because the training/operational process assumes that the learned relationships hold consistently across time. However, as anthropogenic interventions and climate change significantly alter hydrological systems (Arnell & Gosling, 2013; M. Chen et al., 2021), this assumption may no longer be valid (Ajami et al., 2017; Z. Wang & Yang, 2024). Shifts in rainfall patterns, temperature regimes, and land surface conditions can introduce non-stationarity (Milly et al., 2015; Villarini & Wasko, 2021), rendering historical patterns potentially unrepresentative of current or future conditions.

Nevertheless, in large-sample, multi-basin training, the presence of non-stationarity in a particular watershed does not necessarily imply that the model faces unfamiliar conditions (i.e., conditions requiring extrapolation). As highlighted by Martel et al. (2025), analogous hydroclimatic patterns may exist elsewhere in the training dataset (e.g., recent conditions in a watershed located in Ontario may resemble historical conditions in another watershed located in Ohio), meaning the model may have effectively "seen" these hydroclimatic patterns despite local

temporal shifts. Their study demonstrated that continental LSTM models are well suited for climate-change impact studies because they were trained on watersheds from diverse climate zones, suggesting that broad spatial diversity helps mitigate some of the challenges posed by non-stationary local patterns at individual watersheds.

This leads to a core question: in the context of training large-scale rainfall–runoff models, is it necessary to use long training periods that cover as much historical data as possible, or is the temporal relevance of training data more important than their sheer quantity? In other words, while spatial diversity appears to benefit consistently from increasing dataset size, it is unclear whether the same "more is better" principle applies to temporal coverage. This perspective motivates an explicit examination of data recency (or data recentness), that is, when the training data were acquired, in addition to how many years of data or how many watersheds are included in the training. Following Shen et al. (2022), we define data recency as the temporal proximity of training data to the prediction period (i.e., the testing period in this study).

Our underlying assumption is that the value of recency lies in the similarity between the training period sample and the anticipated system characteristics in the prediction period, where similarity involves climatic and anthropogenic similarity including, for example, data quality and changing land use. This formulation reflects the notion that recent data is likely more relevant to the prediction target because it better captures the evolving hydrological system dynamics. Shen et al. (2022) found that calibrating physically-based hydrological models to the latest data is a superior approach compared to using old data, highlighting the value of data recency. Building on this insight, our study investigates whether the same principle holds for data-driven models such as LSTM. Specifically, we evaluate whether prioritizing recent observations during training can enhance model performance, without requiring structural changes or continuous retraining. To this end, our study examines whether training on shorter, more recent subsets of data—rather than decades-long historical records—can yield comparable performance, and whether the necessity of long training periods diminishes as the number of watersheds increases. To validate these hypotheses, we conduct extensive experiments using a comprehensive dataset comprising hydrometeorological records from over 1,300 watersheds across North America, spanning 1950 to 2023 and representing diverse climatic and hydrological regimes. Experiments are designed to isolate and assess the effects of data recency, and to provide evidence-based recommendations

for selecting training data while balancing computational costs, data availability, and prediction accuracy.

The remainder of the chapter is organized as follows: Section 3-2 describes the methodology, including the model configurations, datasets, and experimental design. Section 3-3 reports the results and discusses their implications. Finally, Section 3-4 concludes the paper.

3.2 Methodology

3.2.1 Model settings and performance metrics

A standard PyTorch-based LSTM neural network was employed to simulate daily streamflow based on meteorological and geophysical inputs. The model architecture comprises an input embedding layer, a single-layer LSTM module, a dropout layer, and an output head. During the forward pass, input features are passed through the embedding network, producing a concatenated sequence of input representations. This embedded sequence is then processed by the LSTM layer, which returns the full sequence output as well as the final hidden and cell states. The sequence output is passed through a dropout layer and subsequently through the output head, which generates the final streamflow predictions.

The model architecture and training routines were implemented using the open-source NeuralHydrology library (Kratzert et al., 2022). All models were trained for 30 epochs using four NVIDIA RTX A6000 GPUs. To reduce variance from random initialization, the final streamflow predictions were obtained by averaging an ensemble of 10 LSTM models, each trained with a different random seed.

Hyperparameters were tuned through five-fold cross-validation on a representative subset of 220 watersheds from the training dataset, following a procedure similar to that of Mai et al. (2022). The selected hyperparameters are summarized in Table 3-2 and were kept fixed for all subsequent experiments to ensure comparability. Model performance of the trained LSTM models are then evaluated using the KGE metric (Gupta et al., 2009), as defined in Equation 2-1. In this study, we report the median KGE across all watersheds in the testing set, under two evaluation settings:

- **Temporal Testing:** Evaluates model predictions on unseen years at training watersheds, capturing the model's capability to forecast beyond its training period.

- **Spatiotemporal Testing:** Evaluates model predictions on unseen years at pseudo-ungauged (i.e., unseen/untrained) watersheds, assessing the model's capability to generalize in the prediction in ungauged basins (PUB) context.

Table 3-2. Hyperparameter settings used for the LSTM network. The first six hyperparameters listed in the table were fine-tuned over the specified ranges, and the selected values are highlighted in bold and underlined.

Hyperparameter	Tested Range/Selected Value
Hidden size	128, **256**, 512
Batch size	64, 128, **256**
Epochs	15, 20, 25, **30**, 35, 40
Learning rate[1]	(0: 0.005, 10: 0.001, 20: 0.0005, 30: 0.0001), (0: 0.001, 10: 0.0005, 20: 0.0001, 30: 5e-05), **(0: 0.0005, 10: 0.0001, 20: 5e-05, 30: 1e-05)**, (0: 0.0001, 10: 5e-05, 20: 1e-05, 30: 5e-06)
Sequence length	30, 180, **365**
Dropout	0.2, 0.3, **0.4**
Loss function[2]	NSE*
Optimizer	Adam

[1](**0**: x_1, **10**: x_2, **20**: x_3) denotes learning rate x_1 for epoch 0 to 9, learning rate x_2 for epoch 10 to 19, and learning rate x_3 after epoch 20.

[2]Basin-averaged NSE loss, as defined in Equation A2-2.

3.2.2 Study area and data

This study utilizes the HYSETS dataset (Arsenault et al., 2020), which offers harmonized daily hydrometeorological records across North America from 1950 to 2023. The meteorological data rely exclusively on the Quality-Controlled-Station version of HYSETS, ensuring a consistent data source and uniformly applied quality-control procedures across the study period. Moreover, as documented in Arsenault et al. (2020), all hydrometric records underwent quality control before inclusion, and regulated gauges were excluded, leaving only natural-flow gauges. These steps minimize potential biases due to regulation or inconsistent data quality.

To ensure data reliability, only watersheds with ≤10% missing discharge values were retained, resulting in 1374 gauged watersheds across the U.S. and Canada (location shown in Figure 3-1). These watersheds were partitioned into subsets for training, validation, and testing, as listed in Table 3-3. Of the 1100 training watersheds, 68% have complete discharge records (i.e., with no missing values) throughout the 1950–2010 training period.

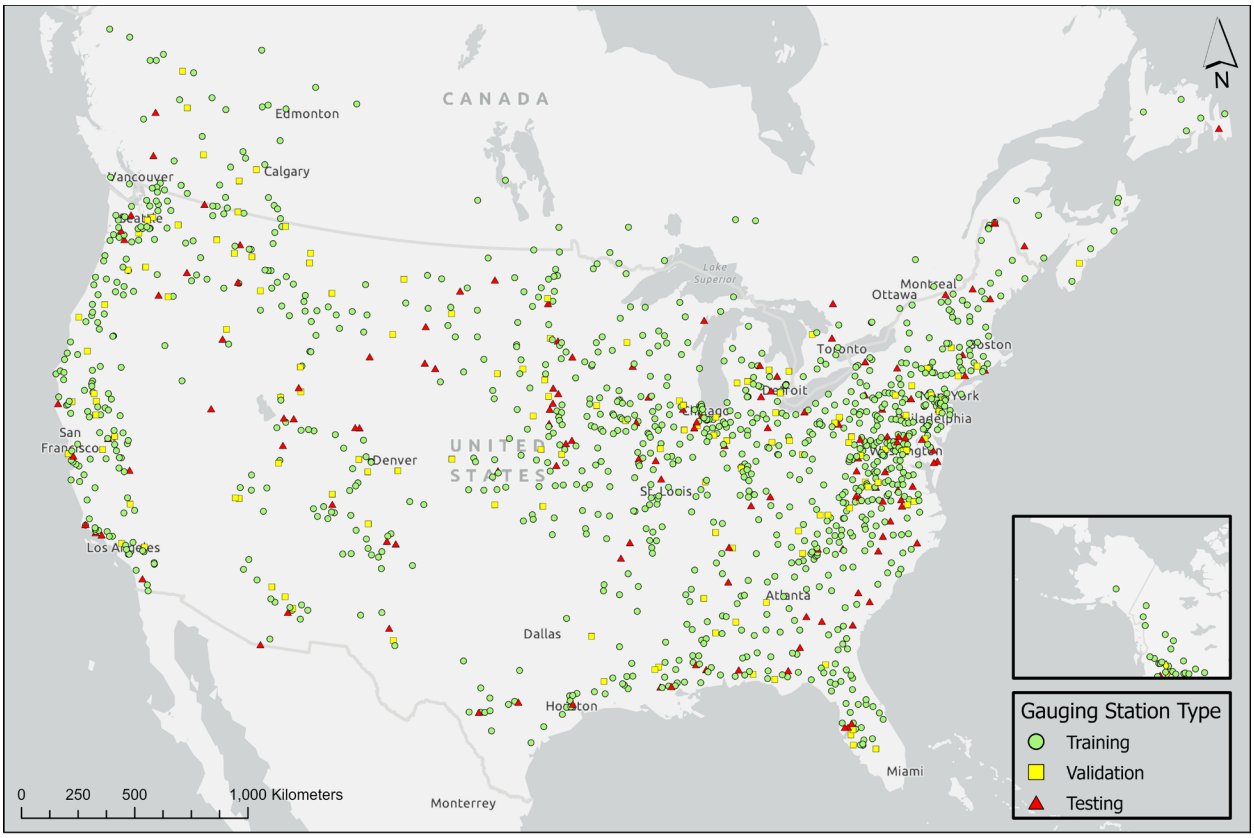


Figure 3-1. Locations of all 1374 gauging stations included in this study. Due to the large geographical extent, an inset map displays stations located in Alaska and Yukon. Stations of training watersheds are indicated by green circles, validation watersheds by yellow squares, and testing watersheds by red triangles.

Table 3-3. Dataset partitioning scheme.

Dataset Partition	Number of Watersheds	Temporal Coverage	Purpose
Training Set	1100	1950/01/01 – 2010/12/31	Used to train LSTM model
Validation set	137	2000/01/01 – 2010/12/31	Used to monitor training convergence (early stopping); excluded from final testing
Temporal testing set	220	2011/01/01 – 2023/12/31	Randomly sampled from training watersheds to test on years not seen during training
Spatiotemporal testing set	137	2011/01/01 – 2023/12/31	Fully excluded from training/validation to evaluate predictions at ungauged locations

To ensure computational feasibility for the extensive iterative model training process, the number of input variables was deliberately constrained. Each watershed's input comprised three dynamic meteorological forcings - daily precipitation, maximum temperature, and minimum temperature - along with 16 static watershed attributes, as specified in Table A3-3 in the appendix. We focused on these three dynamic forcings because they are widely recognized as the most critical forcing variables in rainfall-runoff modelling. For static attributes, we used those provided in the HYSETS dataset, selecting only the ones without missing values for the watersheds included in this study.

3.2.3 Experimental design

Three complementary experiments were designed to evaluate model performance over a common testing period from 2011 to 2023, using varying combinations of training period lengths (ranging from 16 to 61 years) and training sample sizes (ranging from 220 to 1100 watersheds). Each experiment was repeated across different training sample sizes to assess the interaction between temporal coverage and spatial diversity on model performance. Collectively, these three experiments enable a targeted assessment of how training data recency affects model performance. The training/testing split strategies are illustrated in Figure 3-2.

In Experiment 1, backward-expanding training periods were used. The training dataset always included the most recent 16-year period (1995–2010) and was then expanded backward in time by adding up to three 16-year blocks: 1980–1995, 1965–1980, and 1950–1965. This assessed whether extending the training period with older data improves the model's ability to predict future conditions. In Experiment 2, forward-expanding training periods were used. The training dataset always included the earliest data period (1950–1965) and progressively added more recent information in 16-year blocks: 1965–1980, 1980–1995, and 1995–2010. This assessed whether improvements in performance are primarily driven by the inclusion of more recent training data, rather than simply by extending the training period length. In Experiment 3, sliding window training periods were used. To isolate the effect of training period timing independent of length, this experiment used a fixed training window of 16 years and slid it forward through time: 1950–1965, 1965–1980, 1980–1995, and 1995–2010. This assessed whether more recent or more distant historical periods offer greater predictive value for unseen future years, under identical training lengths.

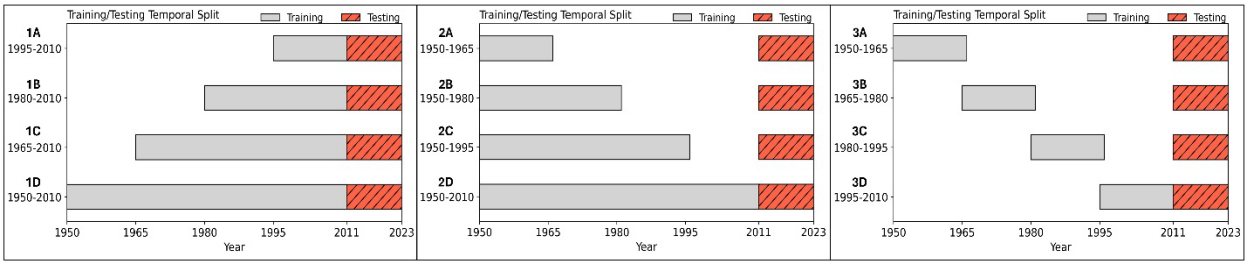


Figure 3-2. Train/test temporal splits adopted for each experiment. The left, middle, and right panels correspond to Experiments 1, 2, and 3, respectively.

3.2.4 Peak flow analysis

Arsenault et al. (2023) noted that LSTM models, like other DL methods, have potential difficulties in extrapolating to extreme conditions that fall outside the range of the training data. This raises concerns about training strategies that rely heavily on recent observations, as shorter training periods may contain fewer peak flow events and thus limit the model's exposure to such extremes. In contrast, longer training period with historical data may encompass a broader range of hydrological variability, including rare high flow events, potentially enhancing the model's ability to simulate them. To examine this issue, we evaluate how the length of the training period affects model performance specifically for peak flow events, in order to assess the necessity of including long historical records. This helps determine whether LSTM can develop an understanding of hydrological processes using only short but recent data, or if longer historical records are required to move beyond simply extracting information from limited datasets. In this study, peak flow events are defined as those with magnitudes exceeding the 99.9th percentile of observed discharge in the testing watersheds during the testing period (2011-2023). For each identified event, we compute the absolute percentage error (APE) between the predicted and observed discharge at the corresponding time step, using the formula:

$$APE = \frac{|predicted - observed|}{observed} \times 100\% \qquad (3-1)$$

Additionally, we quantified the number of peak flow events in the training watersheds during each training period using a consistent 99.9th percentile threshold, to ensure that there is no potential bias if different periods contain substantially different numbers of events. The results are reported in Table A3-1.

3.3 Results and Discussion

3.3.1 Impact of training period length and recency

Experiment 1 examines the impact of training period length by extending the dataset backward in time from a fixed end year (i.e., 2010), progressively incorporating older historical records. As shown in the upper row of Figure 3-3, the heatmaps demonstrate that temporal testing produces consistently stable KGE scores (0.79-0.81) regardless of training period length, while spatiotemporal testing shows only marginal improvements when the training period is extended into the past. Model trained on the only most recent data (1A) perform comparably to or even better than those trained on longer historical records (1C or 1D), regardless of the number of watersheds used for training. The CDFs provide further insight into this pattern. First, increasing the number of training watersheds from 220 to 1100 results in a notable rightward shift in the KGE distributions, confirming the benefit of larger spatial coverage. However, within each panel, the CDF curves for all four temporal splits exhibit a similar shape, indicating that the temporal extent of training data has minimal influence on the overall distribution of model performance. This convergence across temporal splits reinforces the finding that including older historical data does not significantly enhance generalization in either spatial or temporal extrapolation contexts.

Experiment 2 investigates the effect of data recency by progressively shifting the end year of the training period forward in time, while keeping the start year fixed (i.e., 1950). As shown in the upper row of Figure 3-4, both the heatmaps and CDFs reveal that incorporating more recent data into the training set substantially improves model performance, regardless of whether the testing scheme is temporal or spatiotemporal. Conversely, the model trained with old-only data (2A and 2B) shows markedly inferior performance. More importantly, the results show that the benefit of additional spatial data is much more pronounced when recent data are included, as the median KGE of 2A rises only slightly from 0.51 to 0.55 when increasing the watershed count from 220 to 1100, whereas the same increase in watershed count under 2D leads to a jump in median KGE from 0.55 to 0.68. Taken together, the findings from Experiments 1 and 2 suggest that while extending the training period does provide more training data to the LSTM, the temporal recency of that data is more crucial than the duration alone.

Experiment 3 investigates the impact of data recency while controlling for training period length by fixing it to 16-years window and progressively shifting the window forward in time. As shown in the upper row of Figure 3-5, both the heatmaps and the CDF plots consistently indicate that models trained with more recent data (3C and 3D) significantly outperform those trained with older data (3A and 3B), despite all configurations using the same amount of training data. This trend holds across both PUB (spatiotemporal testing) and non-PUB (temporal testing) settings. Like the pattern observed in Experiment 2, the improvement gained from increasing the number of watersheds is conditional on data recency. Specifically, for the older training periods (3A and 3B), increasing the watershed count from 220 to 1100 yields little to no performance gain. In contrast, under recent training periods (3C and 3D), the same increase in spatial coverage leads to substantial improvement.

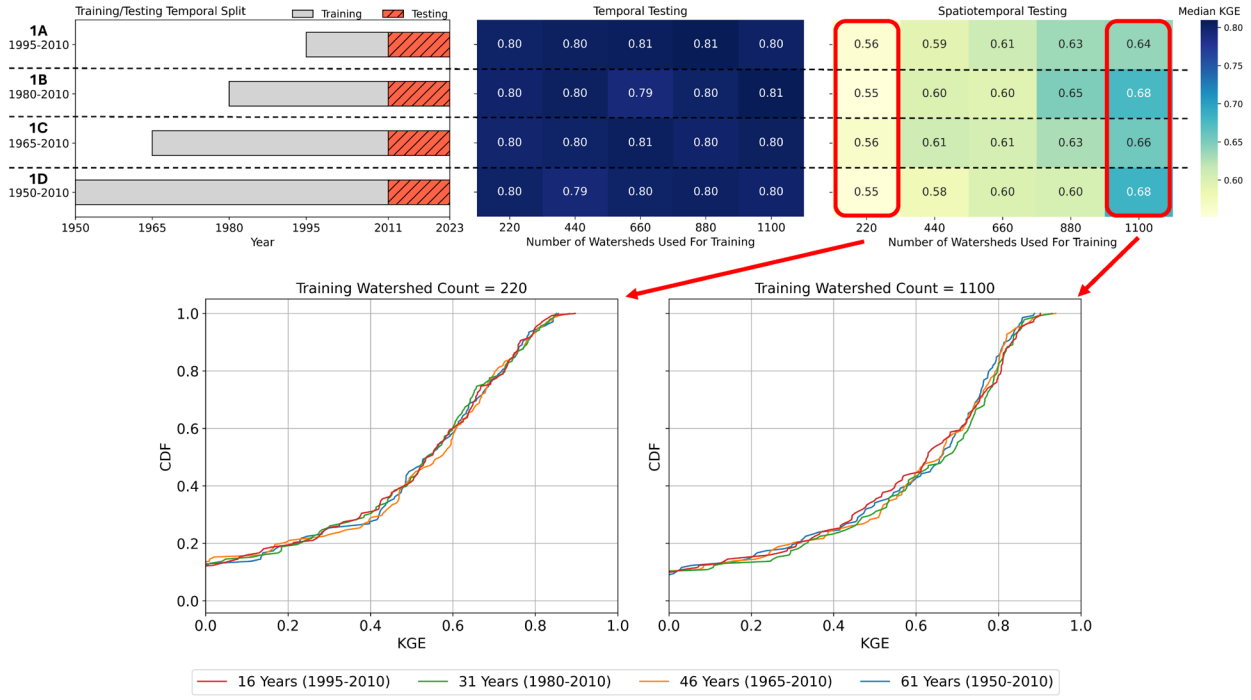


Figure 3-3. Results of Experiment 1. In the upper row of subplots, the left panel illustrates four temporal splits (1A–1D), each with a fixed end year of 2010 and progressively longer training periods by including older historical data, thereby extending backward in time. The heatmaps show the median KGE performance under temporal testing (220 training watersheds; middle panel) and spatiotemporal testing (137 pseudo-ungauged watersheds; right panel), across increasing numbers of watersheds used for training (x-axis). Beneath the heatmaps, cumulative distribution functions (CDFs) of KGE scores are shown for spatiotemporal testing with 220 (left) and 1100 (right) training watersheds, highlighting differences in the overall distribution of model performance across temporal splits.

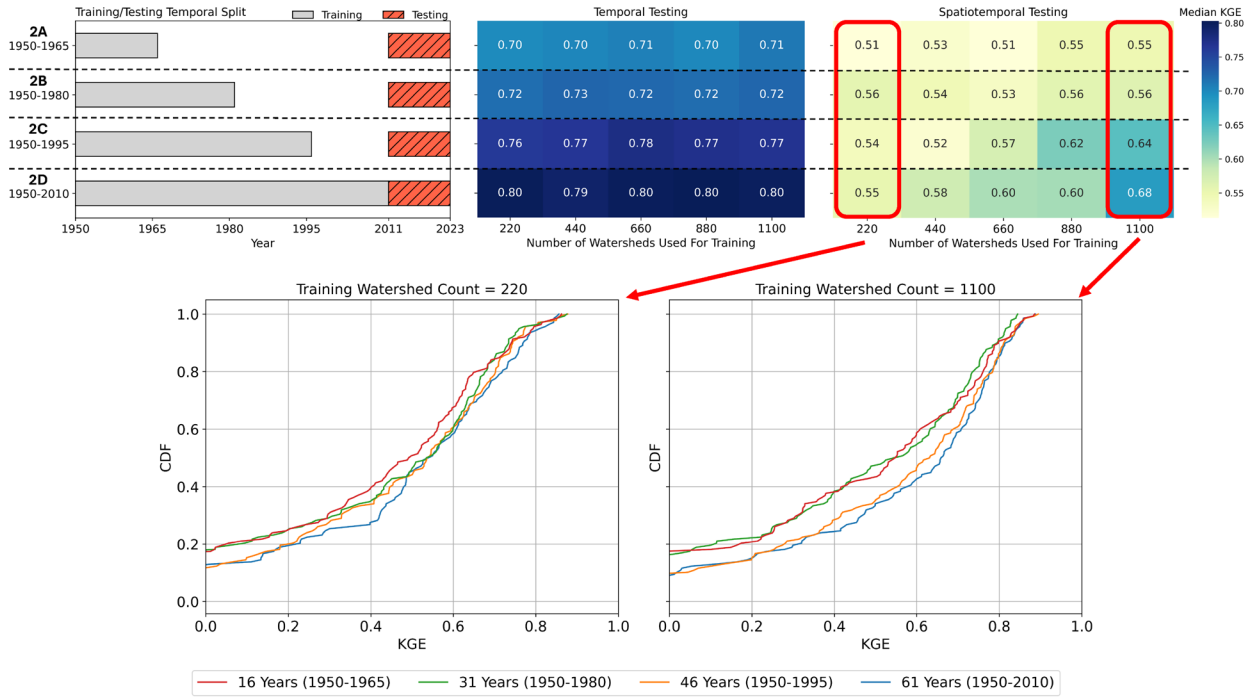


Figure 3-4. Results of Experiment 2. In the upper row of subplots, the left panel illustrates four temporal splits (2A–2D), each with a fixed start year of 1950 and progressively longer training periods by including more recent data, thereby extending forward in time. The remaining panels are consistent with Figure 3-5 in both layout and evaluation setup.

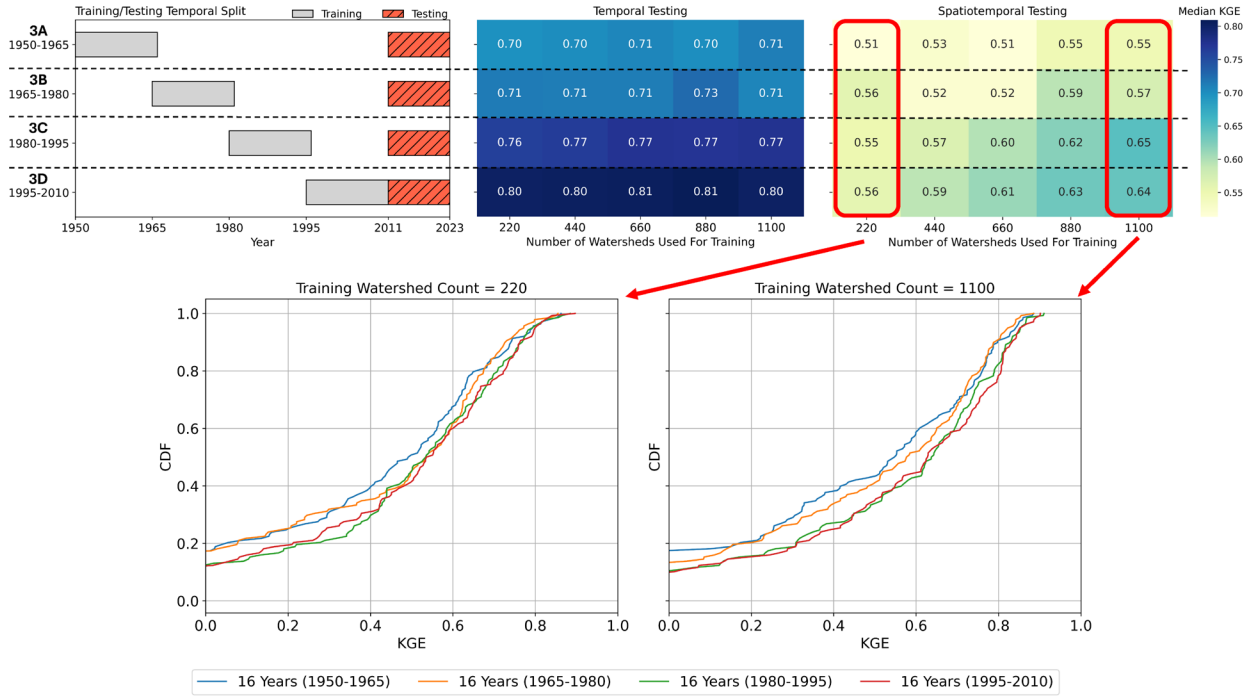


Figure 3-5. Results of Experiment 3. In the upper row of subplots, the left panel illustrates four temporal splits (3A–3D), each using a fixed-length training window that shifts forward in time. The remaining panels are consistent with Figure 3-5 in layout and evaluation setup.

The corresponding results of the three experiments based on NSE are provided in Figure A3-1, Figure A3-2, and Figure A3-3, respectively, and are highly consistent with those shown in Figures 3-3 to 3-5. Therefore, using the NSE instead of the KGE as the evaluation metric does not alter the findings.

3.3.2 Is there a minimum training period length for reliable performance?

To explore whether there exists a threshold in training period length below which model performance clearly degrades, we build upon Experiment 1 by further shortening the training period. Specifically, we examine training periods of 11, 6, and 3 years, with start years of 2000, 2005, and 2008, respectively. The results are shown in Figure 3-6.

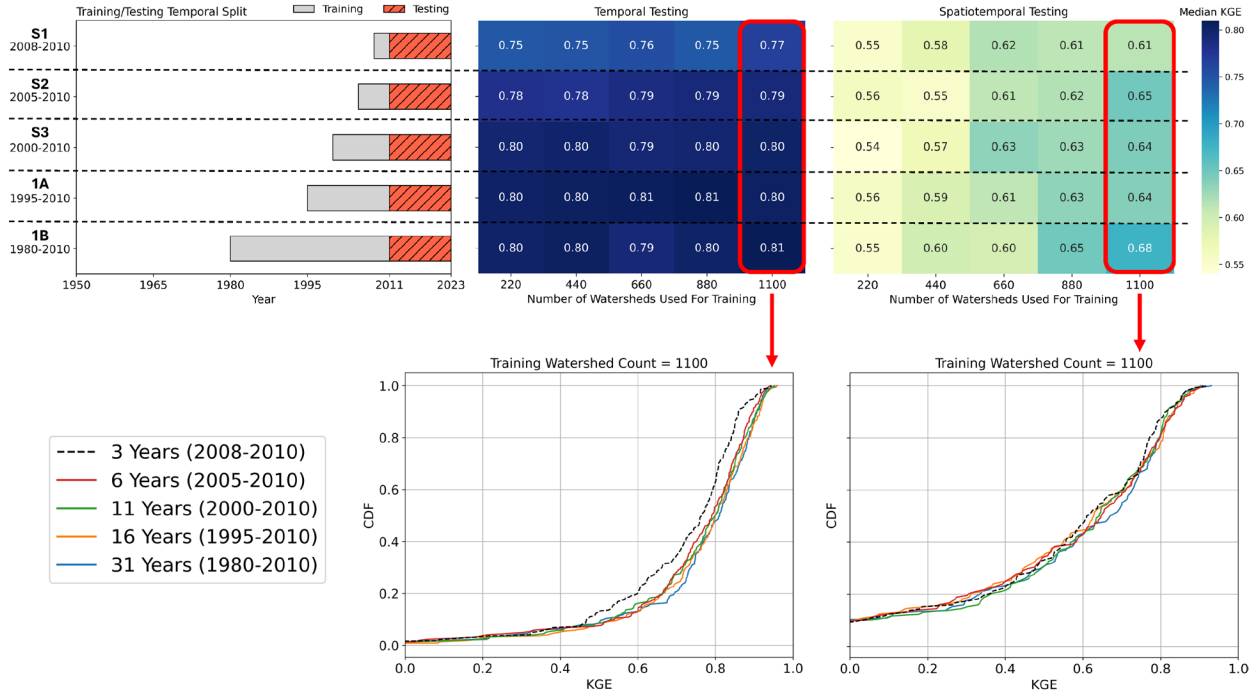


Figure 3-6. Results of training on the most recent 11, 6, and 3 years of data, in comparison with configurations 1A and 1B from Experiment 1. Heatmaps are formatted consistently with earlier experiments to facilitate comparison. The left CDF plot presents the performance of models trained on 1100 watersheds under temporal testing, while the right CDF plot shows the results under spatiotemporal testing.

In the temporal setting, model performance remains relatively stable when the training period begins in 1980, 1995 and 2000. However, performance begins to decline when the start year shifts to 2005 and drops significantly for the 2008-start configuration (3-year training period), as evidenced by a pronounced leftward shift in the CDF and lower median KGE. In the

spatiotemporal setting, the impact of reduced training length is less severe but still apparent: the model trained on just three years of data exhibits the lowest median KGE among all tested configurations.

These results suggest that the performance drop observed for very short training periods is due to limited data volume. When the training period is sufficiently long (e.g., 10-15 years), recent data contribute substantially to model performance (see the results of Experiment 2). Nonetheless, there exists a minimum temporal window—between 3 and 6 years—below which even recent data fail to provide sufficient variability (e.g., interannual hydrological variation or full seasonal cycles) needed for robust performance. Note that this threshold is specific to our large-scale setup with 1100 training watersheds and should not be interpreted as a universal guideline for other modelling scenarios.

3.3.3 How do training strategies affect peak flow predictions?

The evaluation on peak flow event prediction was conducted for all model configurations from Experiment 1 and Section 3.3.2 that were trained on either 220 or 1100 watersheds, with the training period ending in 2010. The distribution of APEs across all identified peak flow events is summarized using boxplots in Figure 3-7 (temporal testing) and Figure 3-8 (spatiotemporal testing). The temporal testing set contains 1085 identified peak flow events, whereas the spatiotemporal testing set includes 681 events.

Under the temporal testing setting, models trained with longer historical data periods (start years of 1950, 1965, and 1980) generally exhibit smaller errors when predicting peak flow events. In contrast, models trained using only short recent periods (e.g., starting from 1995, 2000, 2005, or 2008) tend to produce higher APEs, with errors increasing progressively as the training period becomes shorter. The results indicate that including data prior to 1980 has minimal effect on peak flow prediction accuracy. When the training watershed count is 220, the median APE increases by only 0.9% when the start year shifts from 1950 to 1980 but rises by 11.9% when the start year shifts from 1950 to 2008. A similar pattern is also seen with 1100 training watersheds: the median APE remains unchanged between 1950 and 1980 but increases by 6.9% from 1950 to 2008. Under the spatiotemporal testing scenario, the increase in median APE from 1950 to 1980 remains insignificant (e.g., 1.0% for 220 training watersheds and 1.8% for 1100 training watersheds), while more substantial increases appear only when the start year is later than 2000

(e.g., from 1950 to 2008, the median APE increased by 3.4% for 220 training watersheds and 7.8% for 1100 training watersheds).



Figure 3-7. APE of model predictions for the 1085 peak flow events under the temporal testing scenario. The left subplot shows results from models trained on 220 watersheds, and the right subplot shows results from models trained on 1100 watersheds. Boxplots summarize the distribution of APE values for different training period start years (x-axis), with outliers indicating instances of large prediction errors.



Figure 3-8. APE of model predictions for the 681 peak flow events under the spatiotemporal testing scenario.

A comparison between Figure 3-7 and Figure 3-8 further reveals that increasing spatial diversity of training data (e.g., from 220 to 1100 training watersheds) leads to a notably greater reduction in peak flow prediction error for spatiotemporal testing than for temporal testing. This

pattern is consistent with the KGE heatmaps in Figures 3-3 to 3-6, which likewise demonstrate that spatial diversity has a substantially greater impact in the PUB context.

3.3.4 Implications and limitations of findings

Several factors may explain why LSTM models trained on more recent data consistently achieve better predictive performance. While it is beyond the scope of this work to attribute the finding to specific factors, we conducted preliminary analyses to inform future research. First, hydroclimatic conditions could be non-stationary. To assess whether non-stationarity is present, the Seasonal Mann–Kendall test (Hirsch et al., 1982) was applied to evaluate long-term monotonic trends in basin-scale precipitation, temperature, and discharge over 1950–2023. As summarized in Table A3-2, a large proportion of basins exhibit statistically significant ($p < 0.05$) monotonic trends in precipitation (68% of 1374 watersheds), temperature (71%), and discharge (72%), providing clear empirical evidence of hydroclimatic non-stationarity in the study domain. Such long-term trends, along with short-term climate variability and gradual shifts, may cause recent data to differ substantially from older records in terms of hydrological relevance. Nonetheless, Martel et al. (2025) argued that regional LSTM can remain effective for simulating streamflow under climate change impacts.

Second, older meteorological and streamflow data may be of lower quality due to limitations in historical measurement technologies or inconsistencies in data processing or missing data rates. Early streamflow records (e.g., from the 1950s–1970s) are often less complete than more recent observations, raising the possibility that uneven temporal distribution of missing data could bias the experimental results. For instance, if 50% of the observations in the older period were missing, this would be an obvious cause of older data being found to be not beneficial when included in training. To ensure this is not a potential confounding factor, we examined the distribution of missing values in training watersheds across training periods and summarized the results in Table A3-1. Overall, the vast majority of watersheds retain high data completeness across different periods, making it unlikely that the findings can be attributed to unevenly distributed missing data. For example, our findings indicated that the 1965–1980 period was not beneficial for model training, despite its data characteristics being broadly consistent with those of the two most recent training periods (1980–1995 and 1995–2010).

Third, watershed characteristics can be impacted by anthropogenic influences—such as land use change, urbanization, or reservoir operations—which alter runoff generation processes over time. Lastly, streamflow series exhibit strong temporal autocorrelation, meaning that data from periods immediately preceding the prediction/testing window carry more predictive information. Future research would be required to try and determine which of these factors are most important.

The implications of the results differ by evaluation setting. In PUB task, multi-basin training allows the model to learn cross-basin commonalities—hydrological knowledge that can be transferred to ungauged watersheds. Our experiments show that both spatial diversity and data recency are important for PUB: models trained on many watersheds with only old data (or on fewer watersheds with only recent data) achieve poor performance, whereas strong PUB performance requires both dimensions together. This aligns with the view of Kratzert et al. (2024) that LSTM-based rainfall–runoff models benefit from large-sample, hydrologically diverse training in order to capture transferable information. By contrast, temporal testing exhibits a different pattern: in our large-scale training setting (≥220 training watersheds), performance is strongly influenced by temporal variability within individual basins (i.e., changes in hydrometeorological conditions over the years), while further increasing the number of training watersheds does not improve the model's ability to predict temporal changes. Results from Experiments 2 and 3 support this interpretation, showing that extending the temporal coverage of the training data to include more recent years is particularly effective at improving performance in temporal testing, whereas adding more training watersheds yields no comparable gains. Extending this perspective, more challenging prediction scenarios such as prediction in ungauged regions (PUR; Feng et al., 2021)—where entire contiguous regions are without gauges—may further amplify the value of data recency. Investigating how data recency interacts in PUR settings therefore represents a potential direction for future research.

The heatmap results in Figure 3-3 show the inclusion of some or all data before 1980 either doesn't change prediction quality or slightly degrades prediction quality. There is a substantial increase in training computational cost associated with including this apparently unhelpful data. For instance, in this study, training a model using data from 1950 to 2010 across

220 watersheds required approximately 25 minutes per epoch, while training on data from 1980 to 2010 took only five minutes per epoch.

While the findings are robust across tested watersheds, they are subject to two key limitations. All experiments were conducted using North American datasets, which limit the generalizability of the results beyond this region. Other continents may exhibit different patterns of anthropogenic influence, and their historical meteorological and hydrometric data are often collected by multiple independent agencies, each potentially using its own—and possibly evolving—data collection protocols. Furthermore, the analysis was based on a relatively simple LSTM architecture, without incorporating more sophisticated designs such as encoder–decoder structure or physics-informed constraints. Substantially more complex LSTM models may respond differently to our tests and show different observed effects of data recency.

Given these findings, models deployed in operational contexts should be updated regularly to reflect evolving hydrological conditions. As demonstrated by the poor performance of configurations 2A, 2B, 3A, 3B, and 3C, relying continually on a fixed model trained on aging historical data can lead to declining accuracy over time. Avoiding such outdated setups is critical to maintaining reliable forecasts in dynamic environments.

3.4 Conclusion

This study systematically evaluated how the choice of training data—in terms of both temporal length and recency—affects LSTM performance for streamflow prediction. Results from the three core experimental designs demonstrate that using distant-past data in North America—especially records more than about 30 years before the testing/application period—is not beneficial.

Across all our analyses, we find that in the non-PUB context, expanding the training dataset by adding more watersheds does not enhance model performance. In contrast, under the PUB setting, increasing the number of training watersheds leads to substantial performance gains. This benefit is particularly evident when models are trained on short, recent periods, indicating that greater spatial diversity can compensate for limited temporal coverage. As a result, the model's performance becomes less reliant on long historical records, suggesting that the need for extended temporal coverage is diminished when extrapolating to data-scarce watersheds.

Furthermore, training exclusively on old data results in a performance plateau that cannot be overcome merely by increasing the number of training watersheds. By comparing model performance across different training period lengths and varying numbers of watersheds, our results suggest that the benefit of expanding spatial coverage is conditional on the recency of the training data. Substantial improvement with increasing number of watersheds is observed when the model is trained on recent data (e.g., configuration 1A), whereas little to no improvement is seen when the training relies solely on old data (e.g., configurations 2A, 2B, 3A, 3B). This indicates that data recency—in addition to spatial diversity—is a critical factor for effective prediction. Based on this finding, we recommend that in developing operational LSTM-based forecasting models, practitioners avoid, at least initially, relying on overly old historical data. Instead, it is more effective to train models on recent periods across a larger number of watersheds. Moreover, if the training dataset poses a substantial computational burden, it is preferable to discard older records rather than exclude watersheds, in order to retain spatial diversity while preserving data recency.

As for the minimum training period length required for reasonable model performance, this study does not aim to establish a general recommendation. We acknowledge that the optimal length can vary depending on modelling context, such as the number of training watersheds, the study region, or the start and end years of the testing period. Following the guidance of Kratzert et al. (2018), we adopted the 15-year training window as a lower bound for daily streamflow prediction, which served as the baseline for our experimental design.

To evaluate whether using a short and recent training period might impair the model's ability to predict extreme events due to the lack of long historical records, we conducted targeted analyses in Section 3.3.3. The results show that, in both PUB and non-PUB contexts, peak flow prediction errors associated with short training periods using recent data (e.g., less than 15 years) can be mitigated by including longer historical records extending back to 1980. However, further extending the training period to include data from 1950—effectively doubling its length—leads to only minimal improvement in peak flow prediction accuracy.

To conclude, climate and hydrological systems may exhibit non-stationary behavior over time, and older records may also be of lower quality due to measurement limitations, both of which make it problematic to rely on models trained under the presumption of stable and

accurate historical datasets. In this context, the way in which training data is selected becomes critically important. Using only historical data that no longer reflects current system behavior can degrade model performance, whereas augmenting a training dataset of more recent data with such older data generally yields no substantial improvement or degradation. In situations where training on the entire record becomes prohibitively costly, it is preferable to prioritize retaining a larger number of watersheds to maximize spatial diversity, even if this requires shortening the temporal span and focusing only on more recent data. Conversely, when computational resources are sufficient or more advanced, large-capacity model architectures are employed, the most robust strategy is to leverage all available data, including older data. In practice, however, emphasizing more recent data allows the model to learn relationships that are more relevant to present-day conditions, effectively adapting to non-stationarity. This strategy does not change the model's structure but rather aligns its fixed parameters with a more relevant subset of the historical record. Future work should investigate whether similar patterns regarding the importance of data recency hold when applying substantially more complex LSTM architectures or when using datasets from different continents.

Chapter 4: Exploring xLSTM for Large-Sample Hourly Rainfall-Runoff Modelling

This chapter is a replicate of the following manuscript that is currently in preparation. All references are unified at the end of the thesis.

Yu, Q., Tolson, B. A., Lin, J. (2026). Exploring xLSTM for Large-Sample Hourly Rainfall-Runoff Modelling (manuscript in preparation)

Summary

Hourly rainfall–runoff modelling using DL models commonly involves long input sequences and steeper short-term streamflow variations. Under this setting, conventional LSTMs may encounter architectural bottlenecks: information from distant time steps must be carried through a compact recurrent state, and sigmoid-bounded gates can limit how quickly stored information is revised when runoff conditions change abruptly. This study examines whether extended long short-term memory (xLSTM), which introduces exponential gating and expanded memory structures, can improve large-sample hourly streamflow prediction. To isolate the effect of the model backbone, xLSTM and LSTM-based benchmarks were evaluated under the same input setting and hyperparameter configuration. Experiments were conducted using hourly streamflow and meteorological records from 430 Canadian basins over 2011–2023. Model performance was assessed through extrapolation to unseen periods and pseudo-ungauged basins. Compared with LSTM-based benchmarks, the xLSTM-based model achieved comparable overall accuracy for temporal extrapolation and improved overall accuracy for spatial extrapolation. It also showed generally better low-flow simulation. These results suggest that xLSTM is a plausible alternative to the standard LSTM cell for developing large-sample rainfall–runoff models. Further work is needed to refine xLSTM configurations for hydrological modelling and to evaluate their robustness across larger and more diverse datasets.

4.1 Introduction

Recent years have seen the emergence of a deep learning (DL)-based paradigm in large-sample hydrology, in which rainfall–runoff relationships are learned directly from data using recurrent neural networks, particularly long short-term memory (LSTM) models. Early LSTM-based hydrological modelling studies showed that a single regional LSTM model trained across many basins can outperform hydrological benchmark models for daily streamflow simulation (Kratzert et al., 2018; Kratzert, Klotz, Shalev, et al., 2019). This regional learning strategy has also been shown to be promising for prediction in ungauged basins (PUB), because it can learn transferable hydrological patterns jointly from many gauged basins within a shared model, rather than requiring post-calibration transfer or regionalization of model parameters (Arsenault et al., 2023).

Despite these advances, most large-sample studies have focused on simulating daily streamflow (e.g., Kratzert, Klotz, Shalev, et al., 2019; Lees et al., 2021; Frame et al., 2022; Jin et al., 2022; Li et al., 2022; Mai et al., 2022; Arsenault et al., 2023; Baste et al., 2025; J. Liu et al., 2025; Wang et al., 2025), even though many applications, such as flood early warning and reservoir operation, depend on hydrological dynamics at sub-daily scales. Flood peaks often rise and recede within hours, and operational usefulness frequently depends on whether peak magnitude and timing are accurately predicted. Hourly rainfall–runoff modelling therefore provides information on rapid runoff responses that may be smoothed or hidden at daily resolution. Recent high-resolution datasets, such as the hourly CAMELS-US dataset (Gauch, Kratzert, et al., 2021), CAMELSH dataset (Tran et al., 2025), and CAMELS-SPAT (Knoben et al., 2025), have made large-sample hourly modelling feasible.

Hourly rainfall–runoff modelling is especially important for basins with high flashiness, where streamflow can rise and recede within short periods following precipitation events (Baker et al., 2004). Tudaji et al. (2025) argue that the need for sub-daily information is especially pronounced in small and rapidly responding basins. However, flashiness is not controlled by drainage area alone. Continental-scale analyses show that its drivers vary across hydroclimatic regions and reflect interacting effects of climate, topography, soils, subsurface storage, and human modification (Gannon et al., 2022). In large-sample hourly modelling, separately

evaluating flashy basins therefore provides a targeted test of whether a model has learned the sub-daily hydrological dynamics.

Previous daily-scale LSTM studies commonly used a one-year look-back period, in which a 365-day input sequence was provided to the model to predict streamflow at the last time step of the sequence (Kratzert et al., 2018; Mai et al., 2022; Arsenault et al., 2023). If the same look-back period is retained at hourly resolution, the input sequence becomes 24 times longer than its daily counterpart. This increase in sequence length is particularly challenging for conventional LSTM architectures, because their hidden states are updated sequentially through time. As a result, processing a full one-year look-back window at hourly resolution would not only be computationally expensive but could also make training prone to instability, as excessively long recurrent sequences can make model optimization numerically unstable, potentially due to problematic gradient propagation through many time steps. In addition to this computational burden, very long input sequences also make it more difficult for conventional LSTMs to preserve hydrologically relevant long-term dependencies. As modern hydrology moves toward high-frequency prediction and continental- to global-scale modelling, these limitations create a need for model architectures that retain the strengths of LSTM while improving their ability to handle long input sequences and large datasets efficiently.

To address these difficulties associated with long hourly input sequences, Gauch et al. (2021) proposed the multi-timescale LSTM (MTS-LSTM). Its core idea is that catchments are damped systems in which the history of total mass and energy inputs remains important, but the impact of high-frequency variation diminishes with increasing look-back time. MTS-LSTM therefore processes older inputs at coarser temporal resolution (e.g., daily) and more recent inputs at finer resolution (e.g., hourly), allowing the model to retain a long hydrological look-back window without processing the entire sequence at the finest temporal resolution. In the general MTS-LSTM, different timescales are handled by separate LSTM branches. The hidden and cell states from the coarser-resolution branch are passed through transfer layers and used to initialize the finer-resolution branch. Because each branch has its own recurrent cell, MTS-LSTM can use different numbers of input variables (i.e., different input dimensionalities) at different timescales. Gauch et al. (2021) also introduced a shared variant, sMTS-LSTM, in which the same LSTM cell is reused across temporal resolutions: after the coarse-resolution sequence is

processed, the hidden and cell states corresponding to the start of the finer-resolution sequence are retrieved and used to initialize a second forward pass over the recent finer-resolution inputs. This shared formulation requires the same number of input variables (i.e., same input dimensionality) across different timescales, whereas the general MTS-LSTM is more flexible when different per-timescale inputs are available. When the two architectures were evaluated using the same meteorological forcing inputs, Gauch et al. (2021) found that sMTS-LSTM achieved slightly better daily and hourly predictive performance than MTS-LSTM, although the absolute differences were small. This result supports the use of sMTS-LSTM as a strong benchmark when input dimensionality is consistent across timescales.

Acuña Espinoza et al. (2025) further developed this idea in the multi-frequency LSTM (MF-LSTM), which handles multiple timescales within a single LSTM cell while allowing different input dimensionalities. Instead of using two forward passes as in sMTS-LSTM, or separate recurrent branches as in MTS-LSTM, MF-LSTM constructs a single multi-frequency sequence by concatenating the older coarse-resolution segment with the recent high-resolution segment. This sequence is then processed by a single LSTM cell. Frequency-specific embedding networks are used to map inputs from different temporal frequencies into a common feature dimension before they are passed to the recurrent backbone. This gives MF-LSTM the architectural simplicity of a shared recurrent model while retaining the flexibility to accommodate different numbers of input variables across different timescales. Acuña Espinoza et al. (2025) showed that MF-LSTM achieved predictive performance comparable to the MTS-LSTM and sMTS-LSTM architectures of Gauch et al. (2021).

Although hourly large-sample datasets are becoming increasingly available, existing studies on large-sample hourly rainfall–runoff modelling remain limited. Gauch et al. (2021) provided a key benchmark for this setting by introducing MTS-LSTM to address the long-sequence problem in hourly modelling. Subsequent large-sample hourly studies have largely followed this multi-timescale strategy. Wilbrand et al. (2023) used MTS-LSTM to examine the effect of global versus local datasets, whereas Acuña Espinoza et al. (2025) built on the MTS-LSTM framework by proposing MF-LSTM. Therefore, under the single-source forcing setting considered in this study, sMTS-LSTM and MF-LSTM provide two established and directly

comparable LSTM-based benchmarks for evaluating model performance in large-sample hourly rainfall–runoff modelling.

Beyond LSTM-based architectures, attention-based models and state-space models (SSMs) have recently been explored as alternative sequence backbones for rainfall–runoff modelling. J. Liu et al. (2024) found that, for daily flow simulation, the vanilla Transformer was less competitive than LSTM on large-sample rainfall–runoff benchmarks, especially for high-flow metrics, whereas a modified Transformer with convolutional embeddings designed to encode local temporal relations achieved performance closer to LSTM. Across a broader set of hydrological prediction tasks, J. Liu et al. (2025) further showed that model rankings depend on task formulation: LSTM remained highly competitive for standard rainfall-runoff regression, while attention-based models became more competitive in autoregressive forecasting based only on historical target observations and in zero-shot forecasting without task-specific hydrological training. Moreover, Wang et al. (2025) reported that SSMs can outperform LSTM in several hydrological settings, particularly in snowmelt-driven or intermittent-flow watersheds, although their performance was weaker in flashier, high-magnitude flow regimes.

Attention-based and SSM architectures have become prominent in sequence modelling because they target two key challenges: representing long-context dependencies and scaling computation to long sequences and large datasets (Gu et al., 2021). Both challenges are directly relevant to high-resolution rainfall–runoff modelling. Conventional LSTM remains a strong baseline because its recurrent state can emulate the catchment storage-release dynamics (Kratzert et al., 2018). However, as summarized by Beck et al. (2024), conventional LSTM is limited by the difficulty of revising previous storage decisions, restricted storage capacity, and the lack of parallelizability caused by hidden-to-hidden temporal dependencies. These limitations motivated the development of extended LSTM (xLSTM), which retains the recurrent structure of LSTM while updating it with new gating and memory mechanisms (Beck et al., 2024). Recent hydrological studies have explored xLSTM-based variants for streamflow prediction (R. Liu et al., 2025; Yang et al., 2026; Zhang et al., 2026). However, these applications remain primarily local case studies, and dedicated evaluations of xLSTM as the model backbone for large-sample hydrological modelling remain limited.

This study provides an exploratory assessment of xLSTM for large-sample hourly rainfall–runoff modelling across 430 Canadian basins. We implement a multi-frequency xLSTM architecture, termed MF-xLSTM, by replacing the LSTM backbone in the MF-LSTM architecture with an xLSTM backbone, while retaining the same input representation. This replacement-based design is intended to isolate the effect of the model backbone. The MF-xLSTM is compared against two LSTM-based benchmarks, namely sMTS-LSTM and MF-LSTM. Model performance is assessed in terms of temporal extrapolation to unseen periods and spatial generalization to pseudo-ungauged basins. We further evaluate the models specifically for flashy basins to examine whether high streamflow flashiness poses distinct challenges for hourly prediction. Results show that, even without dedicated tuning of xLSTM-specific hyperparameters, MF-xLSTM achieved overall predictive performance comparable to the LSTM-based benchmarks, while improving spatial extrapolation to ungauged basins and reducing low-flow volume bias. Taken together, this study provides preliminary evidence on the potential of replacing LSTM backbone with xLSTM backbone in established hydrological models for future modelling applications.

4.2 Data and Methods

4.2.1 Dataset

A total of 430 hydrometric gauges across Canada were selected from the Water Survey of Canada (WSC) network. All selected gauges are classified as natural (not regulated), thereby reducing the direct influence of major hydraulic infrastructure on the observed discharge records. Instantaneous streamflow observations were obtained from the WSC hydrometric server (Water Survey of Canada (WSC), n.d.). The WSC sub-daily streamflow records used in this study are available from 2011 onward. Instantaneous observations were aggregated to hourly timescale by averaging all available measurements within each hourly interval. All streamflow and meteorological forcing time series were standardized to Coordinated Universal Time (UTC), and the common data period used in this study extends from 2011-01-01 00:00 to 2023-12-31 23:00 UTC. Within this period, all selected gauges had less than 5% missing values in the aggregated hourly streamflow series.

Meteorological forcings were derived from the Canadian Surface Reanalysis version 3.1 (CaSR v3.1; Gasset et al., 2026), produced by ECCC. Most CaSR variables were available at an

hourly resolution; air temperature, originally provided at a 3-hourly resolution, was linearly interpolated to the hourly time step to ensure temporal consistency with the discharge observations and other forcing variables. Gauge-associated catchment boundaries were obtained from the Canadian Lake and River Hydrofabric (CLRH; Tolson et al., 2023) and used as the common spatial units for basin-level aggregation. For each gauge, meteorological forcing time series were derived from the gridded CaSR products by overlaying the corresponding CLRH basin polygon with the CaSR grid and computing area-weighted averages over the intersecting grid cells. The CLRH dataset also provides hydrological and physiographic attributes. The same CLRH basin polygons were used to derive land-cover class fractions from the North American Land Change Monitoring System (NALCM; CEC, 2024) and soil attributes from the Global Soil Dataset for Earth System Models (GSDE; Shangguan et al., 2014). The meteorological forcing variables and static basin attributes used as model inputs are summarized in Table A4-1 and Table A4-2, respectively.

4.2.2 Definition of flashy basins

The selected basins were classified by their degree of flashiness, and model performance was examined separately for the high-flashiness subset. Flashiness was quantified using the Richards–Baker Flashiness Index (RBI), a dimensionless metric that measures the magnitude of short-term streamflow oscillations relative to total discharge (Baker et al., 2004). For a streamflow series, RBI is defined as

$$RBI = \frac{\sum_{i=1}^{n} |q_i - q_{i-1}|}{\sum_{i=1}^{n} q_i} \qquad (4-1)$$

where i indexes consecutive time steps, q_i denotes streamflow at time step i, q_{i-1} denotes streamflow at the preceding time step, and n is the number of time steps included in the calculation period. In this formulation, the numerator represents the cumulative absolute change in streamflow between consecutive time steps, whereas the denominator represents the sum of streamflow values over the same period. Larger RBI values therefore indicate stronger short-term flow oscillations relative to total flow, corresponding to more rapid hydrological responses.

In this study, flashiness was represented by the mean annual RBI calculated over water years from 2011 to 2023. Although RBI was originally introduced using mean daily streamflow, it can also be calculated from sub-daily observations when intra-day flow fluctuations are

relevant (Gannon et al., 2022). Thus, we computed RBI using both daily and hourly streamflow records to assess whether the choice of temporal resolution affected the relative ranking of basin flashiness. The daily streamflow record was derived by aggregating the original hourly observations to daily mean discharge. The absolute magnitude of RBI differed substantially between the two temporal resolutions for the 430 selected basins: hourly-based RBI ranged from 0.0005 to 0.20, whereas daily-based RBI ranged from 0.01 to 0.93. This difference indicates that the RBI is sensitive to the temporal resolution of the input discharge series. When the same formulation is applied to hourly discharge values, the denominator accumulates discharge over a much larger number of time steps than in the daily calculation, while the numerator reflects smaller but more frequent consecutive changes.

The Spearman rank correlation between daily-based RBI and hourly-based RBI was 0.936, indicating that basin ordering was highly consistent across the two temporal resolutions. To further assess whether the use of daily RBI could overlook intraday flow variability, we used the goodness of uniform estimation (GOUE) to quantify the difference between streamflow represented at hourly and daily temporal resolutions. If daily RBI is a meaningful proxy for sub-daily hydrological responsiveness, basins with higher daily RBI should therefore tend to exhibit lower GOUE values.

Following Tudaji et al. (2025), GOUE is defined as the Nash–Sutcliffe efficiency (NSE) between the observed hourly streamflow and an hourly series reconstructed from daily streamflow under the assumption of uniform intraday flow. The daily discharge values were assigned back to all hourly time steps within each corresponding day, producing a daily-derived uniform hourly streamflow series. Therefore, GOUE quantifies how much information on intraday flow variability is lost when the original hourly observations are represented at the daily scale. GOUE was calculated as

$$\mathrm{GOUE} = 1 - \frac{\sum_{i=1}^{n}\left(Q_i^{D\to h} - Q_i^{h}\right)^2}{\sum_{i=1}^{n}\left(Q_i^{h} - \overline{Q^{h}}\right)^2} \qquad (4-2)$$

where Q_i^h is the observed hourly streamflow at time step i, $Q_i^{D\to h}$ is the hourly streamflow reconstructed from daily streamflow by assuming uniform flow within each day, $\bar{Q^h}$ is the mean of the observed hourly streamflow series, and n is the total number of hourly time steps. A larger

GOUE indicates that the daily-derived uniform hourly series closely matches the observed hourly hydrograph, implying weaker intraday streamflow variability. Conversely, a smaller GOUE indicates that the daily-derived uniform hourly series fails to represent intraday flow fluctuations, implying stronger intraday variability. As shown in Figure 4-1, RBI and GOUE exhibited a strong negative monotonic relationship across the 430 basins (Spearman's $\rho=-0.920$, $p<0.001$). This result indicates that basins assigned higher flashiness based on daily RBI also tend to exhibit stronger deviations between hourly streamflow observations and their daily-scale representation. Therefore, daily RBI remains closely associated with intraday flow variability, supporting its use for distinguishing flashy and stable basins in this hourly prediction study.

In this study, we therefore used RBI calculated from daily streamflow to define flashy basins. This choice maintains consistency with the original RBI formulation and is supported by the high agreement between daily-based and hourly-based RBI rankings. Specifically, all 430 selected basins were ranked according to their daily RBI values. The 108 basins with the highest RBI values, representing the upper 25% of the RBI distribution, were classified as flashy basins, whereas the remaining 322 basins were classified as non-flashy basins.

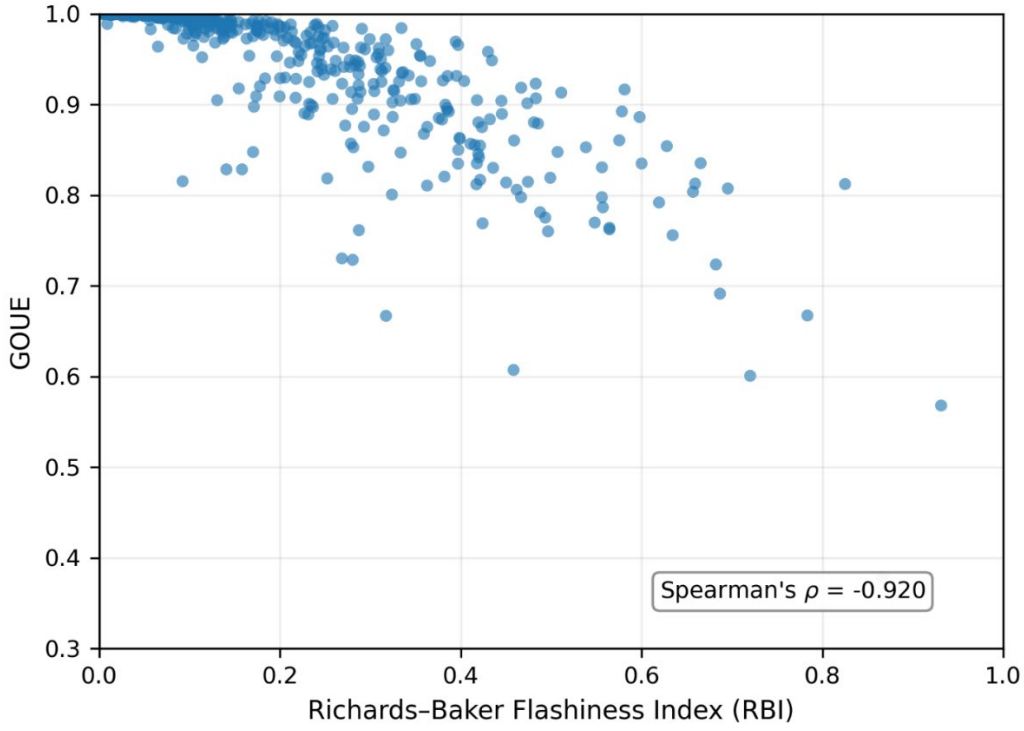


Figure 4-1. Relationship between the daily-based RBI and the GOUE (NSE) across the 430 basins. Each point represents one gauged basin.

4.2.3 Data partitioning

The full set of 430 catchments was first randomly divided into two mutually exclusive basin groups: 344 catchments were assigned to the gauged group, and the remaining 86 catchments were held out as pseudo-ungauged group for spatial testing. Among the 108 catchments previously classified as flashy, 83 were included in the gauged group and 25 were included in the pseudo-ungauged group. Figure 4-2 shows the spatial distribution of the gauged and pseudo-ungauged catchments, separated by flashy and non-flashy catchment types.

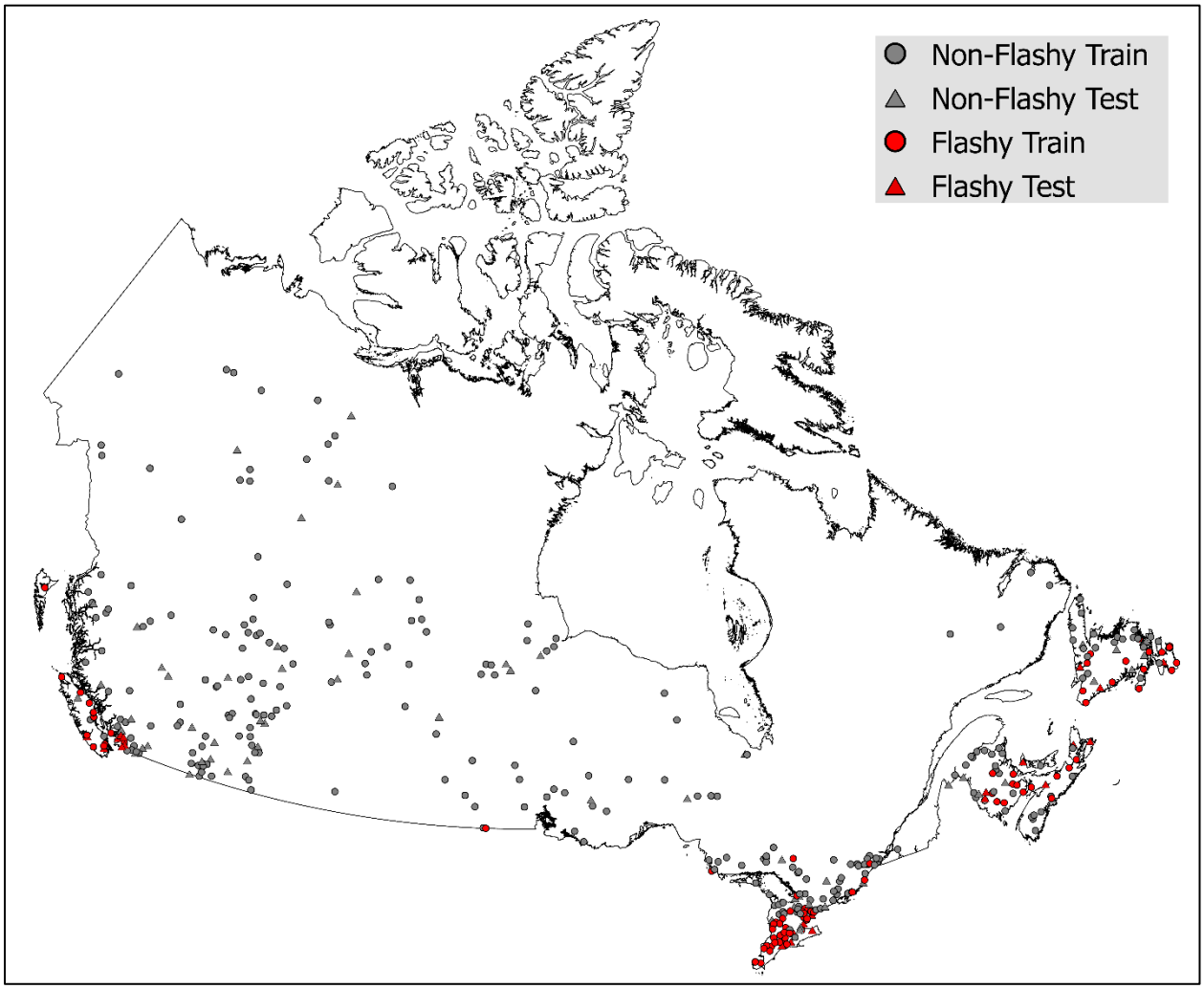


Figure 4-2. Spatial distribution of the 430 study catchments across Canada. Circles denote gauged catchments used for model training and temporal testing, while triangles denote pseudo-ungauged catchments used for spatial testing. Red markers indicate flashy catchments, and grey markers indicate non-flashy catchments.

For the 344 gauged catchments, the time series were divided into a training period from 1 January 2011 to 31 December 2021 and a testing period from 1 January 2022 to 31 December 2023. The data from the gauged catchments during the training period were used as the training set. The data from the same gauged catchments during the testing period were used as the temporal testing set, which evaluates temporal extrapolation to an out-of-sample period. In addition, the full-period records of the 86 pseudo-ungauged catchments were used as the spatial testing set, which evaluates model generalizability under spatial extrapolation to basins excluded from training.

4.2.4 Benchmark models

In this study, two LSTM-based benchmark models were used. The first benchmark is the shared multi-timescale LSTM (sMTS-LSTM) of Gauch et al. (2021). We used the sMTS-LSTM implementation available in the NeuralHydrology framework (Kratzert et al., 2022). The shared variant was selected because the daily and hourly inputs in this study were derived from the same forcing products and had the same input dimensionality.

The second benchmark is MF-LSTM, which we implemented within the NeuralHydrology framework following the main data-handling principle of Acuña Espinoza et al. (2025): inputs represented at different temporal resolutions are assembled into a single multi-frequency sequence and processed by a single LSTM backbone. To make MF-LSTM compatible with NeuralHydrology data-processing workflow, our implementation differs from the original formulation in two aspects, which are schematically compared in Figure 4-4.

The first difference lies in how the daily and hourly input sequences are derived. In LSTM-based rainfall–runoff modelling, a 365-day look-back window is commonly used to provide the model with long-term hydrometeorological information; at hourly resolution, this corresponds to 8760 time steps. Following this setting, the original MF-LSTM implementation constructs two non-overlapping sequences: the first 351 days are aggregated into a daily sequence, while the final 14 days are retained at the hourly resolution. These two sequences are then passed through separate fully connected (FC) layers, which linearly project the daily and hourly inputs into frequency-specific embedded sequences. The resulting embedded sequences are then concatenated to form a single multi-frequency sequence. In our MF-LSTM implementation, by contrast, the full 365-day window is first aggregated into a daily sequence,

while the final 14 days are also retained as a separate hourly sequence. As a result, the daily and hourly sequences overlap over the final 14 days. The two sequences are then passed through the FC layers. After projection, the daily embedded sequence over the overlapping 14-day period is replaced with the corresponding hourly embedded sequence, producing the final multi-frequency input sequence for the LSTM backbone.

The second difference lies in how static catchment attributes are incorporated into the multi-frequency sequence. In the original MF-LSTM implementation, the linear projection is applied only to dynamic forcings. Static attributes are concatenated to the embedded multi-frequency sequence afterward. In our MF-LSTM implementation, by contrast, static attributes are included before the frequency-specific projection. For both the daily and hourly input sequences, static attributes are first concatenated with the dynamic forcings, and the resulting augmented sequences are then passed through the FC layers to produce frequency-specific embedded sequences.

In this study, all benchmark models were trained for hourly streamflow prediction only; therefore, the training loss was computed exclusively from hourly outputs, and there is no regularization term applied to the loss function to ensure the cross-timescale consistency.

4.2.5 Extended Long Short-term Memory (xLSTM)

This section describes the main mechanisms of xLSTM in detail. Since xLSTM builds on the classical LSTM architecture, we first briefly recap the standard LSTM formulation. In a standard LSTM, at time step t, the cell state is updated additively as

$$c_t = f_t c_{t-1} + i_t z_t \,, \tag{4-3}$$

and the hidden state is obtained from the transformed cell state through an output gate,

$$h_t = o_t \widetilde{h_t} \,, \qquad \widetilde{h_t} = \psi(c_t) \,. \tag{4-4}$$

Here, c_t is the cell state, h_t is the hidden state, z_t is the candidate memory input, and i_t, f_t, and o_t are the input, forget, and output gates, respectively. The function ψ is a nonlinear activation function, typically chosen as the hyperbolic tangent function $\tanh$, which maps the cell state into a bounded hidden representation. In the standard LSTM formulation, all gates are sigmoid-activated,

$$i_t = \sigma(\tilde{i}_t)\,, \quad f_t = \sigma(\tilde{f}_t)\,, \quad o_t = \sigma(\tilde{o}_t)\,, \qquad (4-5)$$

with their pre-activations depending on both the current input x_t and the previous hidden state h_{t-1}. This additive memory path is the basis of the constant error carousel, which allows information to be carried across time steps. The gates regulate how much new information is written, how much previous memory is retained, and how much of the memory is exposed to the next layer or prediction head.

However, because the LSTM relies on sigmoid-bounded gates, scalar memory cells and recurrent memory mixing, it faces inherent limitations in revising stored values, expanding storage capacity, and parallelizing sequence processing when modelling highly complex temporal relationships. Such constraints are relevant in high-resolution time-series modelling tasks, where long input sequences with noisy and intermittent signals can make it difficult to preserve useful historical information, distinguish salient temporal changes from background variability, and update the model state in response to rapidly evolving events while maintaining computational efficiency.

The xLSTM architecture consists of two complementary recurrent units, namely the scalar LSTM (sLSTM) and the matrix LSTM (mLSTM). The former retains scalar memory cells and recurrent memory mixing, whereas the latter replaces scalar memory with a matrix-valued associative memory updated by a covariance rule. These units are further embedded into residual blocks and stacked to construct the full xLSTM architecture.

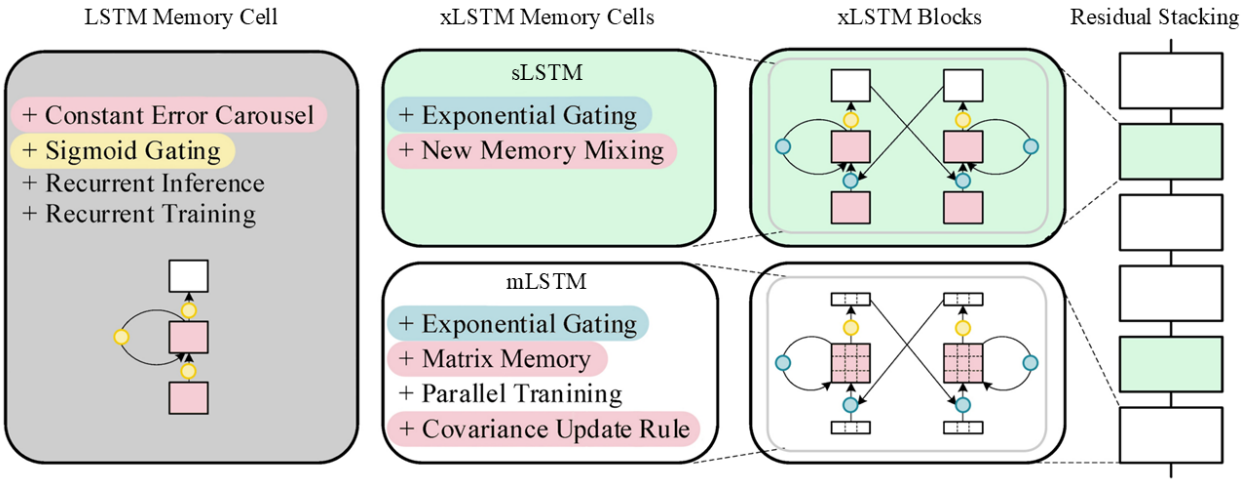


Figure 4-3. The architecture of the xLSTM network, adapted from Beck et al. (2024).

4.2.5.1 Scalar LSTM (sLSTM)

The sLSTM keeps the scalar memory update of LSTM (Equation 4-3) but modifies the gating mechanism so that the model can make stronger and more flexible memory update decisions. At each time step t, given the current input x_t and the previous hidden state h_{t-1}, the candidate update is computed in the same way as in LSTM,

$$z_t = \phi(\tilde{z}_t)\,, \qquad \tilde{z}_t = w_z^\top x_t + r_z h_{t-1} + b_z\,. \tag{4-6}$$

The major difference lies in the input and forget gates. The standard LSTM uses a sigmoid input gate, while sLSTM uses an exponential input gate,

$$i_t = \exp(\tilde{\imath}_t)\,, \qquad \tilde{\imath}_t = w_i^\top x_t + r_i h_{t-1} + b_i\,, \tag{4-7}$$

while the forget gate can be either sigmoid-activated or exponential,

$$f_t = \sigma(\tilde{f}_t) \text{ or } \exp(\tilde{f}_t)\,, \qquad \tilde{f}_t = w_f^\top x_t + r_f h_{t-1} + b_f\,. \tag{4-8}$$

This is the key distinction from the standard LSTM. A sigmoid gate is bounded between 0 and 1, so it can only attenuate information. An exponential gate is not upper-bounded, with an output range of (0, +∞), so the model can assign much larger relative weight to selected inputs or memory states. In sequence modelling terms, this gives sLSTM a stronger ability to revise which information should dominate the memory, rather than only making gradual bounded updates.

The output gate remains sigmoid-activated,

$$o_t = \sigma(\tilde{o}_t)\,, \qquad \tilde{o}_t = w_o^\top x_t + r_o h_{t-1} + b_o\,. \tag{4-9}$$

In Equations 4-6 through 4-9, ϕ is a bounded nonlinearity such as $\tanh$; w_z, w_i, w_f, and w_o represent the input weight vectors between x_t and the pre-activations of the candidate update, input gate, forget gate and output gate, respectively; r_z, r_i, r_f, and r_o are the corresponding recurrent weights from h_{t-1}; and b_z, b_i, b_f, and b_o are the corresponding bias terms.

Because exponential gates can change the scale of the memory (i.e., accumulated cell state), sLSTM introduces a normalizer state,

$$n_t = f_t n_{t-1} + i_t\,, \tag{4-10}$$

the hidden state is then computed by normalizing the memory before applying the output gate,

$$h_t = o_t \widetilde{h_t}\,, \qquad \widetilde{h_t} = \frac{c_t}{n_t}\,. \tag{4-11}$$

This normalization is important because the input and possibly forget gates are no longer constrained to (0,1). The raw cell state c_t can become difficult to interpret or numerically unstable. Dividing by n_t makes the hidden representation behave more like a weighted average of stored information, rather than an uncontrolled accumulation.

To avoid numerical overflow from exponential gates, sLSTM further introduces a stabilizer state,

$$m_t = \max(\log(f_t) + m_{t-1}, \log(i_t))\,. \tag{4-12}$$

The input and forget gates are then rescaled as

$$i'_t = \exp(\log(i_t) - m_t) = \exp(\widetilde{i_t} - m_t)\,, \tag{4-13}$$

$$f'_t = \exp(\log(f_t) + m_{t-1} - m_t)\,. \tag{4-14}$$

In the forward computation, i'_t and f'_t replace i_t and f_t in the memory and normalizer updates. This stabilization changes the numerical scale of the intermediate gates but preserves the normalized output of the cell.

In a standard LSTM used in practice, the recurrent state consists of multiple scalar memory cells whose gated outputs form the hidden representation. The previous hidden state h_{t-1} is mapped into the candidate update and gate computations through hidden-to-hidden recurrent weight matrices. These matrices allow information from different memory cells at the previous time step to jointly influence the current memory update. This interaction among memory cells is referred to as memory mixing. sLSTM preserves this mechanism but modifies how the mixed information is gated by introducing exponential input gates and, optionally, exponential forget gates. In the multi-head sLSTM formulation, the scalar memory cells are further divided into separate heads, where memory mixing is allowed within each head but not across different heads.

To sum up, sLSTM remains a genuinely recurrent network: information at time t depends on h_{t-1}, and this recurrent dependency retains the state tracking behavior of LSTM across a sequence, while making the memory update more flexible through exponential gating and normalization.

4.2.5.2 Matrix LSTM (mLSTM)

The mLSTM addresses a different limitation of the original LSTM: the restricted capacity of scalar memory. Instead of representing memory as a scalar or vector of independent scalar cells, mLSTM uses a matrix memory update. At each time step t, the input x_t is projected into a query, key, and value vectors:

$$q_t = W_q x_t + b_q \,, \tag{4-15}$$

$$k_t = \frac{1}{\sqrt{d}} W_k x_t + b_k \,, \tag{4-16}$$

$$v_t = W_v x_t + b_v \,, \tag{4-17}$$

where W_q, W_k, and W_v are the projection matrices for the query, key, and value vectors, respectively; b_q, b_k, and b_v are the corresponding bias terms; and $1/\sqrt{d}$ is the scaling factor applied to the key projection. The key vector k_t determines how the current input is indexed in memory, the value vector v_t represents the information to be stored, and the query vector q_t is later used to retrieve information from the matrix memory. Although mLSTM uses query, key, and value terminology, it should not be interpreted as standard Transformer self-attention. It does not compute pairwise attention weights over all previous time steps. Instead, it compresses previous key–value information into a recurrent matrix memory and retrieves information from that memory using the current query.

The matrix-valued cell state is then updated as:

$$C_t = f_t C_{t-1} + i_t v_t k_t^\top, \tag{4-18}$$

which is analogous to the additive memory update of LSTM and sLSTM (Equation 4-3), but the update is now applied to a matrix memory. The outer product $v_t k_t^\top$ implements a covariance-style update for storing the current key–value pair in the matrix memory. The forget gate f_t controls the retention of the previous matrix memory, while the input gate i_t controls the strength of the current memory write.

The input gate, forget gate and output gate in mLSTM are:

$$i_t = \exp(\tilde{i_t}) \,, \qquad \tilde{i_t} = w_i^\top x_t + b_i \,, \tag{4-19}$$

$$f_t = \sigma(\tilde{f}_t) \text{ or } \exp(\tilde{f}_t), \qquad \tilde{f}_t = w_f^\top x_t + b_f , \tag{4-20}$$

$$o_t = \sigma(\tilde{o}_t), \qquad \tilde{o}_t = W_o x_t + b_o , \tag{4-21}$$

where w_i and w_f are the input weight vectors for the input and forget gates, respectively; W_o is the weight matrix for the output gate; and b_i, b_f, and b_o are the corresponding bias terms. In contrast to the sLSTM gate equations (Equation 4-7, Equation 4-8, Equation 4-9), which depend on both the current input x_t and the previous hidden state h_{t-1}, the mLSTM gates and key-value projections are computed only from x_t. As a result, the mLSTM memory recurrence can be reformulated for parallel computation, while historical information is still accumulated through the matrix memory. Whereas sLSTM uses a scalar normalizer state, the mLSTM normalizer state is vector-valued:

$$n_t = f_t n_{t-1} + i_t k_t, \tag{4-22}$$

The hidden state is obtained by querying the matrix memory:

$$\tilde{h}_t = \frac{C_t q_t}{\max\{|n_t^\top q_t|, 1\}} , \tag{4-23}$$

followed by an element-wise output gate:

$$h_t = o_t \odot \tilde{h}_t, \tag{4-24}$$

the product $C_t q_t$ retrieves information from the accumulated matrix memory according to the current query. The denominator normalizes this retrieved vector using the alignment between the query q_t and the normalizer state n_t, with the lower bound of 1 preventing division by values close to zero. Finally, the output gate o_t controls which components of the retrieved representation are exposed as the hidden state.

Compared with the standard LSTM and sLSTM, mLSTM changes the memory mechanism in two ways. First, it increases storage capacity by replacing scalar memory with matrix memory. Second, within the additive memory update, mLSTM replaces the cell-input update z_t with the outer-product update $v_t k_t^\top$, which writes a key-value association into the matrix memory. This structure provides a mechanism for retaining and retrieving multiple distinct sequence features within a compressive recurrent memory, which may be beneficial for

long-sequence modelling tasks. However, the mLSTM memory update involves a $d \times d$ matrix, making the per-step computation more expensive than the scalar update in LSTM or sLSTM.

4.2.5.3 Residual stacking of xLSTM blocks

The full xLSTM architecture is constructed by embedding sLSTM and/or mLSTM cells into residual blocks and stacking these blocks along the depth dimension. In this context, a block refers to a complete computational module, including normalization, projection layers, and either an sLSTM or mLSTM cell. The sLSTM and mLSTM cell equations define how information is updated across time steps within each block, whereas residual stacking defines how sequence representations are transformed across depth. Each xLSTM block incrementally transforms the sequence representation produced by preceding blocks, while the residual skip connection helps preserve previously extracted information and facilitates information flow through deeper networks. By stacking and combining both block types, the architecture can combine the recurrent state-tracking ability of sLSTM and the higher-capacity associative memory of mLSTM.

The sLSTM block follows a post-up-projection structure, in which the recurrent memory operation is first applied in the original embedding space and is followed by a gated feed-forward module. In contrast, the mLSTM block follows a pre-up-projection structure, where the input is first projected to a higher-dimensional space before the matrix-memory operation and then projected back to the original embedding dimension. This design is motivated by the fact that the mLSTM matrix memory benefits from operating in a higher-dimensional representation space, whereas the sLSTM relies on scalar memory with recurrent memory mixing and therefore uses the feed-forward module after the memory update.

The effect of different stacking approaches is likely to be task-dependent and has not been examined for rainfall–runoff modelling. Beck et al. (2024) examined several stacked configurations, including sLSTM-only, mLSTM-only, and mixed sLSTM–mLSTM configurations, across language modelling and synthetic sequence tasks. Recent xLSTM-based time-series modelling studies have tended to adopt predefined uses of sLSTM or mLSTM components rather than systematically examining different block orders and network depths (Alharthi & Mahmood, 2024; Wan et al., 2025; Q. Chen et al., 2026). Although these previous studies indicate that the choice of stacking approach can affect sequence modelling performance,

they do not directly establish which stacking approach is appropriate for rainfall–runoff modelling.

Therefore, in this study we evaluated how the composition of stacked blocks would influence the application of xLSTM in rainfall–runoff modelling. This stacking experiment was included as part of the hyperparameter tuning procedure described in Section 4.2.6.

4.2.5.4 MF-xLSTM

Based on the multi-frequency input sequence construction and embedding procedure described in Section 4.2.4, we implemented a multi-frequency xLSTM model, referred to as MF-xLSTM, by replacing the LSTM backbone in the MF-LSTM implementation with an xLSTM block-stack module from the original xLSTM implementation. The purpose of this implementation was to isolate the effect of replacing the backbone. Before the prediction head, dropout was applied to the xLSTM hidden representations as a regularization step, after which the head mapped the resulting representations to discharge predictions.

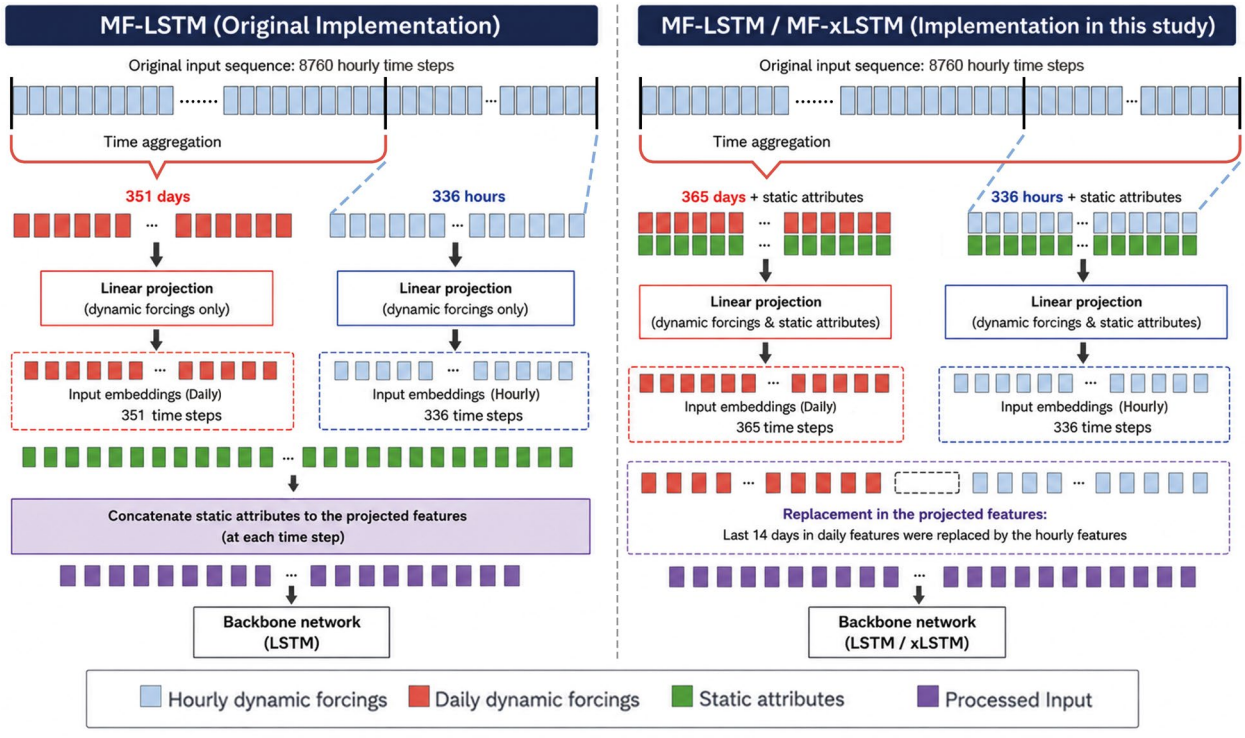


Figure 4-4. Schematic comparison of the original MF-LSTM implementation (Acuña Espinoza et al., 2025) and the MF-LSTM/MF-xLSTM implementations used in this study for constructing the multi-frequency input sequence.

4.2.6 Hyperparameter tuning

Previous large-sample rainfall–runoff studies have commonly compared LSTM variants within a shared benchmark setting, using the same dataset and consistent data splits, while keeping most hyperparameters identical or adopting them from an established LSTM baseline rather than tuning each variant independently (Frame et al., 2022). In this study, we fixed several hyperparameters that have been commonly used in previous LSTM-based rainfall–runoff benchmarks. Specifically, the daily and hourly input sequence lengths were set to $T_D = 365$ and $T_H = 336$, following Gauch et al. (2021) and Acuña Espinoza et al. (2025). This corresponds to using a one-year daily look-back window and representing the most recent 14 days at hourly resolution. Additionally, all models are trained using the Adam optimizer and Basin-averaged NSE loss (Kratzert, Klotz, Shalev, et al., 2019), with the dropout rate fixed at 0.4.

Because our experiments were conducted using a different dataset, the remaining hyperparameters were tuned on the training set through a 3-fold cross-validation. This tuning was carried out for sMTS-LSTM, and the resulting hyperparameter configuration was then used for MF-LSTM and MF-xLSTM without additional tuning. Specifically, the 344 gauged catchments were randomly split into three partitions. For each candidate hyperparameter configuration, the models were trained on two partitions and evaluated on the remaining partition, so that each partition served once as the validation set. This procedure was repeated with five different random seeds. For each validation basin, predictions from the five random seeds were ensembled by averaging the predicted discharge at each time step. The KGE was then computed for each basin using the resulting ensemble prediction, and the final cross-validation score was reported as the median KGE across all 344 basins. The hyperparameter configuration with the highest cross-validation score was retained for the final training and is reported in Table A4-3.

In addition to the general hyperparameters, MF-xLSTM involves several xLSTM-specific hyperparameters. These hyperparameters were not systematically tuned, instead, most of their values were directly adopted from the original xLSTM implementation for language modelling and are reported in Table A4-4. This choice was made for practical reasons, as a full search over the xLSTM-specific hyperparameters would be computationally expensive and time-consuming. Therefore, the MF-xLSTM results should be interpreted as an evaluation of the suitability of the xLSTM architecture under a reasonable but not optimized configuration.

Although a complete hyperparameter search was not performed for MF-xLSTM, we explicitly evaluated the effect of the xLSTM block-stacking approach using the same 3-fold cross-validation procedure described above, while keeping all other hyperparameters fixed to the values listed in Table A4-3 and Table A4-4. We use S to denote one sLSTM-based block, and M to denote one mLSTM-based block. We first assessed single-block configurations, S and M, and then extended the comparison to all two-block configurations, including S-S, S-M, M-S, and M-M. Several three-block configurations were further evaluated to examine the effect of increased depth. The selected stacking configuration was then used for the final training of MF-xLSTM.

4.2.7 Evaluation criteria

The two benchmark models and MF-xLSTM were evaluated using the metrics listed in Table 4-1. KGE and NSE quantified overall hydrograph accuracy, while FHV, FLV, and FMS characterized errors across the flow-duration curve. Peak-Timing and Peak-MAPE were used to assess the timing and magnitude of peak-flow simulations. To account for variability caused by random initialization, all reported results for each architecture were calculated from ensemble hydrographs obtained by averaging predictions from 10 independently trained models with identical hyperparameters but different random seeds.

Table 4-1. Overview of the performance metrics

Metric	Description	Reference	Range and optimum
KGE	Kling–Gupta efficiency	Eq. (9) in Gupta et al. (2009)	(−∞, 1], best: 1
NSE	Nash–Sutcliffe efficiency	Eq. (3) in Nash & Sutcliffe (1970)	(−∞, 1], best: 1
FHV*	High flow bias (top 2 %)	Eq. (A3) in Yilmaz et al. (2008)	(−∞, ∞), best: 0
FLV*	Low flow bias (bottom 30 %)	Eq. (A4) in Yilmaz et al. (2008)	(−∞, ∞), best: 0
FMS*	Bias of the slope of the flow duration curve between the 20th and 70th percentile	Eq. (A2) in Yilmaz et al. (2008)	(−∞, ∞), best: 0
Peak-Timing	Mean absolute time lag (in hours) between observed and simulated peak flows	NeuralHydrology implementation (Kratzert et al., 2022)	[0, ∞), best: 0
Peak-MAPE	Mean absolute percentage error of peak flows	NeuralHydrology implementation (Kratzert et al., 2022)	[0, ∞), best: 0

*Note: FHV, FLV, and FMS are expressed as percentages.

4.3 Results and Discussion

4.3.1 Effects of xLSTM block stacking

Table 4-2 compares the performance of different xLSTM block-stacking configurations. We first evaluated one- and two-block stacks, which showed that mixed configurations (S-M and M-S) clearly outperformed single-type configurations (S, M, S-S, and M-M). For rainfall–runoff modelling, such complementarity is plausible because streamflow prediction requires both the tracking of evolving catchment states and the modelling of context-dependent associations between meteorological forcings and discharge response; these two aspects are respectively aligned with the recurrent memory mixing of sLSTM and the matrix-memory mechanism of mLSTM.

Table 4-2. Performance comparison of different xLSTM stacking configurations using 3-fold cross-validation within the training set.

No. of blocks	Stacking configuration	KGE	NSE
1	S	0.642	0.674
	M	0.635	0.656
2	S-S	0.623	0.677
	S-M	0.681	0.724
	M-S	0.676	0.712
	M-M	0.640	0.701
3	S-S-M	0.698	0.736
	S-M-S	0.696	**0.743**
	S-M-M	**0.701**	0.738
	M-S-M	0.671	0.719
	M-S-S	0.682	0.735
	M-M-S	0.656	0.733

Consequently, the three-block stack evaluation was confined to mixed configurations, so that S-S-S and M-M-M were not included. Compared with the shallower stacks, the three-block configurations generally yielded slight improvements, with S-M-M achieving the highest KGE and S-M-S achieving the highest NSE. Moreover, S-first stacks (S–S–M, S–M–S, and S–M–M)

showed more stable and generally better performance than M-first stacks (M–S–M, M–S–S, and M–M–S). This suggests that model performance was affected not only by stack depth but also by block ordering. However, the three-block configurations were less robust during training. In some runs, the training loss became NaN and remained NaN in subsequent iterations, after which the run had to be terminated and restarted with a different random seed. Considering the limited performance gain and reduced training robustness of three-block stacks, the two-block S-M configuration was selected as the stacking approach for the final MF-xLSTM training and benchmark comparison. It is worth noting that these findings should be regarded as empirical evidence specific to the experimental setting of this study, rather than as a general architectural rule.

4.3.2 Performance comparison between xLSTM and benchmark LSTM

Figure 4-5 illustrates the cumulative distributions (CDF) of per-basin KGE scores of sMTS-LSTM, MF-LSTM, and MF-xLSTM. Table 4-3 reports the median performance metrics in temporal testing, in which the three architectures achieved very similar overall predictive accuracy across the same 344 basins used for model training. The median KGE ranged from 0.821 to 0.828, and the median NSE ranged from 0.835 to 0.842. MF-LSTM produced the highest KGE and NSE in this setting, whereas MF-xLSTM remained close to both benchmark LSTM architectures. Thus, under temporal extrapolation, MF-xLSTM performs comparably to the benchmark LSTM architectures, indicating that the recurrent-cell modifications introduced by xLSTM preserved the model's capacity to learn rainfall-runoff relationships and to generalize to unseen periods.

Table 4-4 reports the results for spatial testing. In contrast to temporal testing, MF-xLSTM showed a more evident advantage in the overall efficiency metrics. Across the 86 pseudo-ungauged basins, MF-xLSTM achieved the highest median KGE and NSE, with values of 0.750 and 0.758, respectively. In comparison, sMTS-LSTM achieved a median KGE of 0.715 and NSE of 0.736, while MF-LSTM achieved a median KGE of 0.722 and NSE of 0.743. The results suggest that the advantage of the xLSTM-based architecture is more evident in spatial extrapolation than in temporal extrapolation. However, the CDF curves were close and partially overlapping, indicating that the overall differences were modest. One possible explanation is that the expanded memory of xLSTM, particularly the matrix memory used in mLSTM block, may

help the model store and retrieve more detailed information about how runoff responds under different basin characteristics and antecedent forcing conditions. This mechanism could be relevant for spatial extrapolation, where the model must transfer learned hydrological patterns to pseudo-ungauged basins with unseen hydrological and physiographic attributes, rather than only to unseen time periods within the basins used for model training.

Table 4-3. Median metrics for sMTS-LSTM (Gauch, Kratzert, et al., 2021), MF-LSTM (Acuña Espinoza et al., 2025), and proposed MF-xLSTM in temporal testing. The number of basins is denoted by '***n***'.

Metric	Flashy basins (***n***=83)			All basins (***n***=344)		
	sMTS-LSTM	MF-LSTM	MF-xLSTM	sMTS-LSTM	MF-LSTM	MF-xLSTM
KGE	**0.818**	0.816	0.813	0.824	**0.828**	0.821
NSE	**0.801**	0.787	0.790	0.839	**0.842**	0.835
FHV	-5.848	-5.185	**-4.798**	-4.360	**-4.351**	-5.843
FLV	**17.033**	24.116	-31.998	22.487	25.336	**-2.829**
FMS	-6.991	-8.140	**-4.548**	-7.079	-7.075	**-4.423**
Peak-Timing	3.693	3.732	**3.059**	4.772	4.889	**4.400**
Peak-MAPE	38.526	38.289	**37.387**	30.324	**29.892**	30.203

Table 4-4. Median metrics for the sMTS-LSTM, MF-LSTM, and MF-xLSTM in spatial testing.

Metric	Flashy basins (***n***=25)			All basins (***n***=86)		
	sMTS-LSTM	MF-LSTM	MF-xLSTM	sMTS-LSTM	MF-LSTM	MF-xLSTM
KGE	0.730	0.735	**0.782**	0.715	0.722	**0.750**
NSE	0.749	0.756	**0.764**	0.736	0.743	**0.758**
FHV	**-7.621**	-9.008	-11.989	**-5.252**	-9.532	-9.719
FLV	**-11.005**	29.894	-25.413	30.806	40.736	**-6.559**
FMS	**-8.310**	-11.769	-8.725	-9.390	-11.259	**-7.850**
Peak-Timing	3.956	**3.866**	4.093	5.508	**5.447**	5.551
Peak-MAPE	38.843	41.829	**38.817**	35.498	36.084	**34.386**

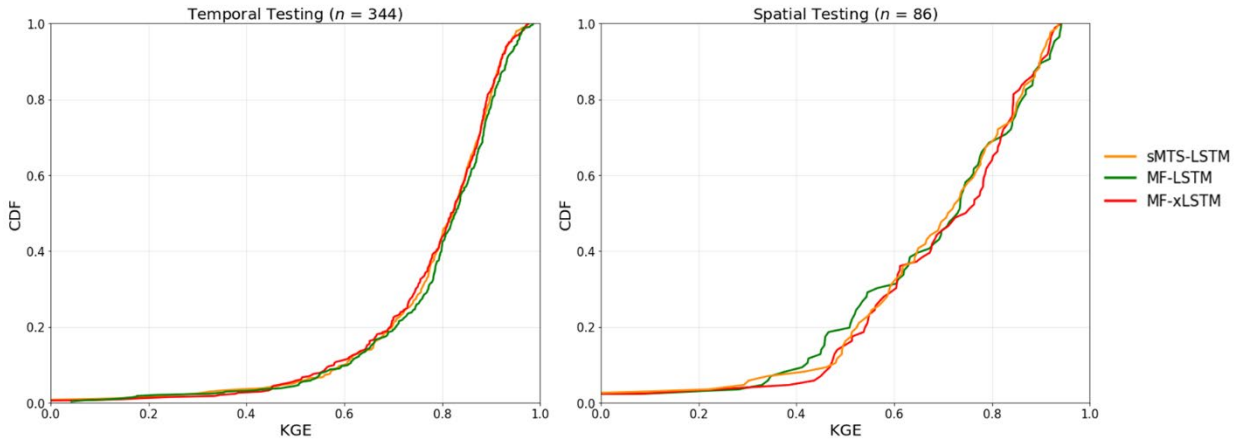


Figure 4-5. Cumulative KGE distributions of the different models. The left panel shows KGEs of 344 basins from temporal testing, the right panel shows KGEs of 86 basins from spatial testing.

The diagnostic flow-bias and peak-flow metrics provide additional information about how the three architectures differ beyond the overall efficiency scores. In both temporal testing and spatial testing, FHV, FMS, Peak-Timing, and Peak-MAPE were broadly comparable across the three architectures. Overall, no single architecture showed a consistent advantage across these four metrics. By contrast, the most notable difference among the diagnostic metrics was observed for FLV. In temporal testing, MF-xLSTM produced the smallest deviation from zero, with an FLV of -2.829, compared with 22.487 for sMTS-LSTM and 25.336 for MF-LSTM. A similar pattern was observed in spatial testing: MF-xLSTM had an FLV of -6.559, compared with 30.806 for sMTS-LSTM and 40.736 for MF-LSTM.

The relatively small FLV deviation achieved by MF-xLSTM is worth noting, as low-flow simulation has been identified as a persistent challenge for LSTM-based rainfall–runoff models, especially in baseflow-dominated watersheds (Gauch, Kratzert, et al., 2021; Kratzert et al., 2021; Jin et al., 2022). Unlike high-flow simulation, which is often strongly linked to precipitation events, low flows are largely sustained by delayed releases from catchment storage, particularly groundwater and subsurface stores (Smakhtin, 2001; Chagas et al., 2024). From a modelling perspective, this interpretation is also consistent with the limitations of conventional recurrent architectures on preserving long-term dependencies. For low-flow simulation, the relevant signal may originate far back in the input sequence. As the effective dependency length increases, information about these slow processes can be diluted, especially when the training objective is

dominated by higher-flow periods. Lees et al. (2021) noted that standard objective functions such as NSE may insufficiently weight low-flow regimes.

A relevant comparison can be made with attention-based architectures, which can assign weights to different parts of the input sequence and directly relate distant time steps without relying solely on sequential state propagation (Vaswani et al., 2017). Recent studies have shown that attention-based architectures, such as Transformer, achieved more favourable FLV values than LSTM models (J. Liu et al., 2024, 2025; Wang et al., 2025). These findings suggest that low-flow simulation benefits from architectures that can better preserve and utilize long-term information. In this context, the improved FLV obtained by MF-xLSTM may be attributed to the extended memory mechanisms, which may help retain and transform long-range antecedent information more effectively than conventional LSTM.

Nevertheless, this result should be interpreted with caution. FLV is calculated from the lower 30% of the flow-duration curve using logarithmically transformed simulated and observed flows (Yilmaz et al., 2008), which makes the metric sensitive when flows approach zero. As a result, small absolute differences during very low-flow conditions can produce disproportionately large changes in the metric. Frame et al. (2022) similarly cautioned that FLV can be fragile under dry or frozen conditions because the logarithmic terms can introduce large numerical instability and make the metric sensitive to outlier basins. Thus, the reported median FLV values do not indicate that MF-xLSTM consistently improves low-flow simulation across all basins.

4.3.3 Impacts of basin flashiness

Table 4-3 and Table 4-4 show that the flashy-basin subsets did not produce a substantially different pattern from the full testing sets for KGE, NSE, FHV, and FMS across either temporal or spatial testing. The main exception was FLV. In the full testing sets, MF-xLSTM produced the most favourable FLV among the three architectures, with values closest to zero. However, its median FLV bias became markedly more negative within the flashy-basin subsets, changing from -2.829 to -31.998 in temporal testing and from -6.559 to -25.413 in spatial testing, indicating a shift toward stronger low-flow volume underestimation within the flashy-basin subsets. As discussed above, the more negative median FLV values in the flashy-basin subsets may also partly reflect the sensitivity of the metric to near-zero flows, rather than

providing standalone evidence that MF-xLSTM systematically degrades low-flow simulation in flashy basins.

For the two peak-flow metrics, all three architectures showed a consistent contrast between the flashy-basin subsets and the full testing sets. In both testing setups, flashy basins had lower median Peak-Timing errors but higher median Peak-MAPE. This suggests that the predicted runoff pulses in flashy basins had relatively smaller timing errors, but their peak magnitudes remained more difficult to simulate and were more likely to be underestimated or smoothed. Figure 4-6 illustrates a case example of this pattern. For the less flashy basin 02HC009, the observed hydrograph shows relatively attenuated runoff responses, with lower peak amplitudes and more gradual recessions. Because the runoff events are relatively broad, the simulated and observed peaks can occur at different hours within the same event, leading to noticeable timing offsets. For the flashier basin 02HC030, the observed hydrograph contains sharper runoff peaks, steeper rising limbs, and faster recessions. The models reproduce the timing of several major runoff responses, but some secondary peaks are clearly underestimated or smoothed.

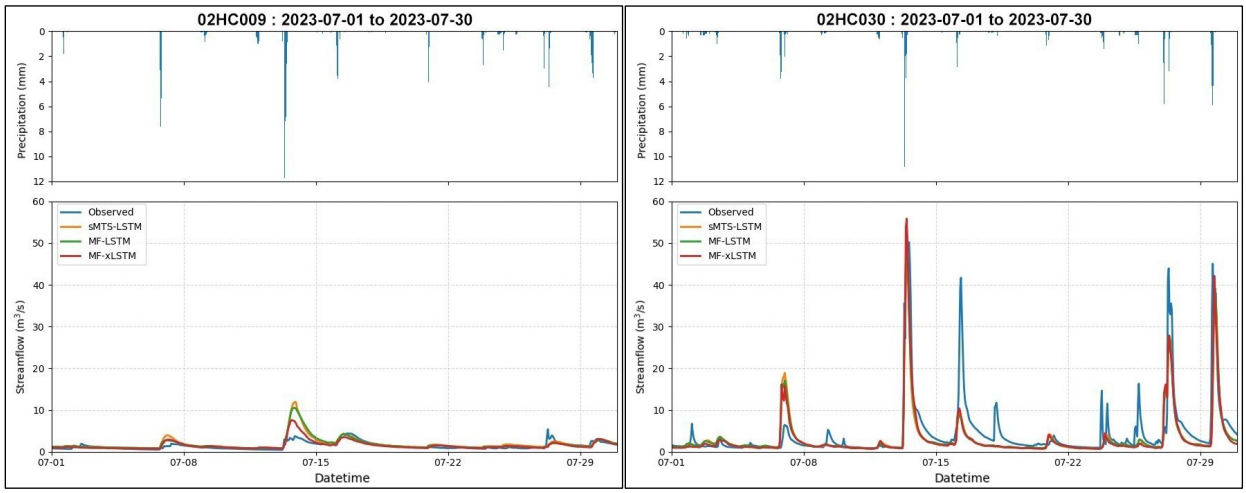


Figure 4-6. Hourly hydrograph comparison for two gauges in the western Greater Toronto Area during July 2023. The left panel shows 02HC009 (East Humber River near Pine Grove; drainage area = 196 km^2; RBI = 0.304), which is characterized by a more attenuated runoff response. The right panel shows 02HC030 (Etobicoke Creek below Queen Elizabeth Highway; drainage area = 208 km^2; RBI = 0.720), which is characterized by a flashier runoff response.

4.3.4 Computational efficiency

Under the same training settings described above, including the look-back window of 365 days, 30 training epochs, and a batch size of 128, sMTS-LSTM and MF-LSTM training took approximately 3.5 hours on the NVIDIA H100 GPU. In contrast, the MF-xLSTM (one sLSTM block and one mLSTM block) required nearly 16 hours. As for inference efficiency, predicting 12 years of hourly streamflow for 86 basins required approximately five minutes for the LSTM-based models and approximately 11 minutes for MF-xLSTM.

The block-stacking experiment in Section 4.3.1 showed that adding mLSTM blocks substantially increased the computational cost and training runtime. This is consistent with Beck et al. (2024), who state that the matrix-memory structure improves memory capacity but introduces additional computational complexity. Nevertheless, the reported runtime should not be interpreted as the fully optimized computational efficiency of xLSTM. The MF-xLSTM experiments were conducted without kernel libraries specialized for the mLSTM and sLSTM computations. Specifically, the mLSTM computations were not run with the Tiled Flash Linear Attention (TFLA)-based mLSTM kernel library, which provides optimized Triton kernels for mLSTM training and inference (Beck et al., 2025). Similarly, the sLSTM computations were not run with FlashRNN, a CUDA/Triton kernel library for accelerating recurrent architectures (Pöppel et al., 2024). Integrating these optimized kernels may reduce the training cost of MF-xLSTM, although this was not evaluated in this study.

4.4 Conclusion

This study explored whether xLSTM can serve as the model backbone for large-sample hourly rainfall-runoff modelling. We implemented MF-xLSTM by replacing the LSTM backbone in the MF-LSTM architecture with an xLSTM backbone, while retaining the same multi-frequency input settings. The model was evaluated against sMTS-LSTM and MF-LSTM across 430 Canadian basins, using temporal testing on unseen periods, spatial testing on pseudo-ungauged basins, and additional evaluation over high-flashiness basins. The results indicate that MF-xLSTM appears to be a viable extension of established LSTM-based hydrological architectures. In temporal testing, MF-xLSTM achieved overall accuracy comparable to the two LSTM benchmarks. In spatial testing, MF-xLSTM achieved the highest median KGE and NSE among the three architectures, suggesting that the expanded memory capacity of xLSTM may

provide additional benefits for transferring learned rainfall–runoff relationships to pseudo-ungauged basins.

The block-stacking experiment also showed that mixed sLSTM-mLSTM configurations outperform single-type configurations, suggesting that recurrent state tracking and the key-value matrix memory may provide complementary information in rainfall-runoff modelling. One possible explanation is that, although LSTM-type recurrent states can be interpreted as representing storage-like catchment dynamics, runoff response is also strongly condition-dependent. Similar precipitation inputs can produce different runoff responses depending on, for example, antecedent soil moisture, snowmelt or frozen-ground conditions, and land-cover settings. From this perspective, the key-value matrix memory of mLSTM may help store and retrieve latent representations of such condition-dependent event-response associations.

The diagnostic metrics provide additional insight into model performance across different flow regimes. Across the full temporal and spatial testing sets, MF-xLSTM achieved FLV values closest to zero, indicating smaller median bias in the low-flow segment of the flow-duration curve. By contrast, differences in FHV, FMS, Peak-Timing, and Peak-MAPE were comparatively small across the three architectures. This pattern suggests that the memory mechanism of xLSTM may help preserve long-range antecedent information relevant to low-flow simulation. However, within the flashy-basin subsets, MF-xLSTM exhibited pronounced negative low-flow volume bias, indicating stronger underestimation of low-flow volumes in rapidly responding catchments.

For KGE, NSE, FHV, and FMS, the flashy-basin subsets showed broadly similar patterns to the full testing sets. Across all three architectures, flashy basins also showed lower peak-timing errors but higher peak-magnitude errors, indicating that the timing of rapid runoff pulses was often captured more closely, while their magnitudes remained more difficult to reproduce. For MF-xLSTM, this pattern is consistent with a limitation commonly reported for LSTM-based hydrological models, namely a tendency to underestimate extreme peak-flow events when runoff responses approach or exceed the range represented in the training data (Baste et al., 2025). In flashy basins, shorter runoff-response durations may reduce peak-timing errors once the triggering rainfall event is captured, while peak magnitudes may more often involve rare or out-of-sample runoff values.

Several limitations of this study should be noted. The xLSTM-specific hyperparameters were largely adopted from the original language-modelling implementation, and only the block-stacking configuration was explicitly examined. Therefore, the reported MF-xLSTM performance may not fully reflect the performance attainable with more targeted hyperparameter tuning. The experiments were also limited to 430 Canadian basins and a fixed multi-frequency input setting. In addition, the training objective was not specifically designed for low-flow, peak-flow, or flashiness-sensitive prediction. Moreover, the apparent advantages of xLSTM in capturing long-term dependencies and condition-dependent associations are interpretations based on the experimental results rather than mechanisms directly validated in this study. These interpretations should therefore be further examined through targeted analyses and validation across larger spatial domains, different hydroclimatic regions, alternative input windows, and finer spatial units. More broadly, this study provides preliminary evidence for further developing xLSTM-based architectures for large-sample rainfall–runoff modelling. As hydrological datasets with greater hydroclimatic diversity become available, architectures with expanded capacity may better represent the nonlinear relationships among input variables and streamflow responses across basins, thereby improving regionalization to ungauged basins. Future work should examine hydrology-specific xLSTM configurations, including more complete hyperparameter tuning, and explore ways to improve computational efficiency through the use of kernel optimization libraries and optimized model implementation.

Chapter 5: Conclusions, Limitations and Future Work

5.1 Conclusions

In summary, the three studies presented in this thesis suggest that the value of large hydrological datasets lies not simply in their size, but in whether the hydrological information they contain can be learned as generalizable representations by deep learning (DL) models. Inappropriate ways of using expanded datasets, such as excessive spatial aggregation or reliance on outdated historical records, can limit model performance even under large-sample training. Conversely, when hydrological information is represented at a more suitable spatial scale, selected with greater relevance to the prediction period, or processed with architectures better suited to large-scale training, expanded datasets can more effectively support rainfall–runoff prediction.

Chapter 2 develops a spatially distributed hybrid modelling framework to retain within-basin spatial heterogeneity in long short-term memory (LSTM)-based streamflow prediction. The results show that basin-averaged forcing variables and static attributes in a lumped LSTM can become limiting in larger basins, where spatial aggregation may remove important spatially variable information. The proposed Spatially Recursive (SR) model provides a practical alternative by applying a regionally trained LSTM at the subbasin scale and routing the resulting local streamflow predictions through a lake-river network. This approach preserves finer-scale spatial information while retaining the advantages of multi-basin LSTM training, and it can be applied to pseudo-ungauged basins without additional LSTM training or routing calibration.

Chapter 3 provides empirical evidence on the effect of training-data recency through large-sample experiments. The results demonstrate that training-period selection is not only a question of how many years are included. For a given number of training watersheds, models trained with recent observations substantially outperformed models trained only on older records. Increasing the number of training watersheds was beneficial for prediction in ungauged basins, but this benefit was limited when the model was trained on outdated data. Conversely, shorter but recent training periods across many watersheds supported robust performance. These findings indicate that temporal coverage should be evaluated by its relevance to the prediction period, not only by its length. The advantage of recent data may reflect greater consistency with current hydroclimatic conditions, data quality, land-surface characteristics, and operational

prediction targets. For operational use, LSTM-based models should therefore be updated regularly rather than relying indefinitely on fixed models trained on aging historical records.

Chapter 4 explores an alternative recurrent architecture with expanded representational and memory capacity to handle the sequence-modelling challenges of large-sample hourly rainfall–runoff prediction. Without extensive hydrology-specific tuning, MF-xLSTM achieved performance close to the LSTM-based benchmark models and showed a modest improvement in prediction for pseudo-ungauged basins. These results suggest that xLSTM is a plausible candidate architecture for future rainfall–runoff modelling applications. The most distinct improvement was observed in the low-flow bias metric. Since low-flow conditions are commonly associated with long-term baseflow and groundwater contributions, the expanded memory capacity of xLSTM may help capture longer-range dependencies more effectively than the standard LSTM, providing a possible explanation for its advantage in low-flow simulation.

5.2 Limitations

A common limitation across the three studies is their geographic scope. Chapter 2 focused on the Great Lakes region, Chapter 3 used watersheds across North America, and Chapter 4 focused on Canadian watersheds. Although these study domains cover a range of hydrological conditions, they remain concentrated in North America and therefore provide limited evidence for global-scale generalization, particularly in climate regions that are poorly represented in the study datasets. In addition, the gauge-selection procedures generally focused on natural or minimally regulated basins. This choice helped reduce the direct influence of human intervention on streamflow records, but it also limits the applicability of the findings to urbanized or strongly regulated basins, where runoff generation and routing are substantially affected by human activities. Future work should therefore evaluate these findings and modelling approaches across global large-sample datasets and in more human-impacted hydrological systems. For regulated basins, this would require additional regulation-related attributes so that DL models can learn to represent the influence of regulation on streamflow.

For Chapter 2, the SR model was evaluated as a proof-of-concept approach, but the appropriate degree of routing-network discretization was not fully resolved. The optimal subbasin scale may depend on forcing-grid resolution, drainage density, and the spatial distribution of lakes. Deriving forcing variables and static attributes at the subbasin scale is also

computationally intensive. In addition, except for some gauged headwater subbasins, the study could not independently validate subbasin-level LSTM predictions because incremental streamflow observations were generally unavailable for individual subbasins.

For Chapter 3, the analysis demonstrated the importance of training-data recency, but it did not further separate the mechanisms that may explain this effect. In particular, the experiments did not distinguish whether the advantage of recent data was mainly associated with hydroclimatic non-stationarity, changes in data quality, or other time-dependent changes in the watershed records. The dynamic forcing variables were also limited to precipitation and temperature, although these are among the most commonly available and reliable dynamic forcing variables in large-sample hydrological datasets. Including additional meteorological variables could change the information available to the LSTM and may affect both overall model performance and the apparent value of recent versus older training data. Finally, the experiments were based on hindcasting simulation models, rather than operational forecasting models with lead times. Although the findings suggest that data recency may be important for streamflow prediction, this implication was not directly tested in a forecasting setting.

For Chapter 4, most xLSTM-specific hyperparameters were adopted from language-modelling implementations. The study therefore does not establish an optimized xLSTM configuration suitable for rainfall–runoff modelling. It also used a fixed multi-frequency input setting rather than full hourly look-back windows, so xLSTM performance under long pure-hourly input sequences remains uninvestigated. Moreover, the dataset was limited to 430 Canadian basins and therefore does not provide sufficient hydroclimatic diversity to support conclusions at the global scale. Finally, because the implementation did not use optimized xLSTM kernels, this study does not provide a conclusive assessment of the computational efficiency of xLSTM for large-sample studies.

5.3 Future Work

One immediate future research direction is to connect the investigations presented across the preceding chapters. In particular, the SR modelling approach developed in Chapter 2 could be jointly examined with the hourly modelling in Chapter 4. This would involve investigating the benefits and limits of the SR model at hourly time steps. Preliminary experiments in this direction suggest that the SR model may not perform well if the routing-network discretization

becomes too fine. Conceptually, larger and less flashy basins may have greater potential for improvement at the hourly time scale when they are discretized into more representative subbasins. Another important extension would be to evaluate the SR model, xLSTM-based models, and the effect of data recency identified in Chapter 3 under explicit forecasting settings. In addition, further hyperparameter tuning experiments could be designed for Chapter 4 to maximize xLSTM performance.

Broader and more open-ended future research should move toward an integrated framework for global-scale rainfall-runoff modelling with large-sample datasets. The three directions examined in this thesis — finer spatial representation, recency-aware training, and higher-capacity sequence modelling architectures — can be viewed as complementary components of such a framework. Future models will need to preserve spatial heterogeneity, prioritize temporally relevant information, and process long, high-resolution hydrological sequences efficiently. This integration is particularly important as large-sample hydrology moves from regional datasets toward continental and global applications, where efficient training, operational updating, and robust generalization across diverse hydroclimatic regions become central requirements. In natural language processing, and particularly in large language models, empirical scaling laws describe how model loss changes predictably with training data size, model size, and computational cost (Kaplan et al., 2020; Hoffmann et al., 2022). It remains unclear whether similar empirical scaling relationships exist in large-sample hydrology. Future research should examine whether predictive performance can increase systematically with data volume, data quality, and model capacity, or whether such gains are ultimately constrained by the intrinsic predictability limits of hydrological systems.

Finally, future research should reconsider how hydrological DL models are interpreted beyond overall accuracy. Functional diagnosis methods are needed to identify where these models fail and whether such failures arise from poor data quality, missing explanatory variables, or architectural limitations. Rather than viewing hydrological DL as a substitute for traditional hydrological models, it may be better regarded as a tool for extracting hydrological knowledge from large and heterogeneous datasets. Although its internal representations may not align with human-defined conceptualizations of hydrological processes, they may still capture complex and broadly shared hydrological patterns through high-dimensional data-driven representations.

References

Acuña Espinoza, E., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., Loritz, R., & Ehret, U. (2025). Technical note: An approach for handling multiple temporal frequencies with different input dimensions using a single LSTM cell. *Hydrology and Earth System Sciences*, *29*(6), 1749–1758. https://doi.org/10.5194/hess-29-1749-2025

Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K., & Mendoza, P. A. (2020). Large-sample hydrology: Recent progress, guidelines for new datasets and grand challenges. *Hydrological Sciences Journal*, *65*(5), 712–725. https://doi.org/10.1080/02626667.2019.1683182

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Ajami, H., Sharma, A., Band, L. E., Evans, J. P., Tuteja, N. K., Amirthanathan, G. E., & Bari, M. A. (2017). On the non-stationarity of hydrological response in anthropogenically unaffected catchments: An Australian perspective. *Hydrology and Earth System Sciences*, *21*(1), 281–294. https://doi.org/10.5194/hess-21-281-2017

Alharthi, M., & Mahmood, A. (2024). xLSTMTime: Long-Term Time Series Forecasting with xLSTM. *AI*, *5*(3), 1482–1495. https://doi.org/10.3390/ai5030071

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., McPhee, J., & Ayala, A. (2018). The CAMELS-CL dataset: Catchment attributes and meteorology for large sample studies – Chile dataset. *Hydrology and Earth System Sciences*, *22*(11), 5817–5846. https://doi.org/10.5194/hess-22-5817-2018

Arnell, N. W., & Gosling, S. N. (2013). The impacts of climate change on river flow regimes at the global scale. *Journal of Hydrology*, *486*, 351–364. https://doi.org/10.1016/j.jhydrol.2013.02.010

Arsenault, R., Brissette, F., Martel, J.-L., Troin, M., Lévesque, G., Davidson-Chaput, J., Gonzalez, M. C., Ameli, A., & Poulin, A. (2020). A comprehensive, multisource database for hydrometeorological modeling of 14,425 North American watersheds. *Scientific Data*, *7*(1), 243. https://doi.org/10.1038/s41597-020-00583-2

Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F., & Mai, J. (2023). Continuous streamflow prediction in ungauged basins: Long short-term memory neural networks clearly outperform traditional hydrological models. *Hydrology and Earth System Sciences*, *27*(1), 139–157. https://doi.org/10.5194/hess-27-139-2023

Ayzel, G., & Heistermann, M. (2021). The effect of calibration data length on the performance of a conceptual hydrological model versus LSTM and GRU: A case study for six basins from the CAMELS dataset. *Computers & Geosciences*, *149*, 104708. https://doi.org/10.1016/j.cageo.2021.104708

Baker, D. B., Richards, R. P., Loftus, T. T., & Kramer, J. W. (2004). A NEW FLASHINESS INDEX: CHARACTERISTICS AND APPLICATIONS TO MIDWESTERN RIVERS AND STREAMS[1]. *JAWRA Journal of the American Water Resources Association*, *40*(2), 503–522. https://doi.org/10.1111/j.1752-1688.2004.tb01046.x

Baste, S., Klotz, D., Acuña Espinoza, E., Bardossy, A., & Loritz, R. (2025). Unveiling the limits of deep learning models in hydrological extrapolation tasks. *Hydrology and Earth System Sciences*, *29*(21), 5871–5891. https://doi.org/10.5194/hess-29-5871-2025

Beck, M., Pöppel, K., Lippe, P., & Hochreiter, S. (2025). *Tiled Flash Linear Attention: More Efficient Linear RNN and xLSTM Kernels* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2503.14376

Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., & Hochreiter, S. (2024). *xLSTM: Extended Long Short-Term Memory* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2405.04517

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166. https://doi.org/10.1109/72.279181

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1–2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Beven, K. (2012). *Rainfall-Runoff Modelling: The Primer* (1st ed.). Wiley. https://doi.org/10.1002/9781119951001

Bindas, T., Tsai, W., Liu, J., Rahmani, F., Feng, D., Bian, Y., Lawson, K., & Shen, C. (2024). Improving River Routing Using a Differentiable Muskingum-Cunge Model and Physics-Informed Machine Learning. *Water Resources Research*, *60*(1), e2023WR035337. https://doi.org/10.1029/2023WR035337

Boulmaiz, T., Guermoui, M., & Boutaghane, H. (2020). Impact of training data size on the LSTM performances for rainfall–runoff modeling. *Modeling Earth Systems and Environment*, *6*(4), 2153–2164. https://doi.org/10.1007/s40808-020-00830-w

Camera, C., Bruggeman, A., Zittis, G., Sofokleous, I., & Arnault, J. (2020). Simulation of extreme rainfall and streamflow events in small Mediterranean watersheds with a one-

way-coupled atmospheric–hydrologic modelling system. *Natural Hazards and Earth System Sciences*, *20*(10), 2791–2810. https://doi.org/10.5194/nhess-20-2791-2020

Chagas, V. B. P., Chaffe, P. L. B., & Blöschl, G. (2024). Regional Low Flow Hydrology: Model Development and Evaluation. *Water Resources Research*, *60*(2), e2023WR035063. https://doi.org/10.1029/2023WR035063

Chen, M., Papadikis, K., & Jun, C. (2021). An investigation on the non-stationarity of flood frequency across the UK. *Journal of Hydrology*, *597*, 126309. https://doi.org/10.1016/j.jhydrol.2021.126309

Chen, Q., Ding, M., Wang, C., & Tong, Y. (2026). Breaking the time barrier: A resilient framework integrating exponential gating mechanisms and matrix memory cells for long-term subseasonal to seasonal wind speed forecasting. *Energy Conversion and Management: X*, *29*, 101441. https://doi.org/10.1016/j.ecmx.2025.101441

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., Arnold, J. R., Gochis, D. J., & Rasmussen, R. M. (2015). A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resources Research*, *51*(4), 2498–2514. https://doi.org/10.1002/2015WR017198

Commission for Environmental Cooperation (CEC). (2024a). *NALCMS - North American Environmental Atlas—Land Cover 2015 30m* (Version 4.0) [Raster digital data [30-m]]. https://www.cec.org/north-american-environmental-atlas/land-cover-30m-2015-landsat-and-rapideye/

Commission for Environmental Cooperation (CEC). (2024b). *NALCMS - North American Environmental Atlas—Land Cover 2020 30m* (Version 2.0) [Raster digital data [30-m]]. https://www.cec.org/north-american-environmental-atlas/land-cover-30m-2020/

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J. K., Lane, R., Lewis, M., Robinson, E. L., Wagener, T., & Woods, R. (2020). CAMELS-GB: Hydrometeorological time series and landscape attributes for 671 catchments in Great Britain. *Earth System Science Data*, *12*(4), 2459–2483. https://doi.org/10.5194/essd-12-2459-2020

Craig, J. R., Brown, G., Chlumsky, R., Jenkinson, R. W., Jost, G., Lee, K., Mai, J., Serrer, M., Sgro, N., Shafii, M., Snowdon, A. P., & Tolson, B. A. (2020). Flexible watershed simulation with the Raven hydrological modelling framework. *Environmental Modelling & Software*, *129*, 104728. https://doi.org/10.1016/j.envsoft.2020.104728

Deb, P., Kiem, A. S., & Willgoose, G. (2019). Mechanisms influencing non-stationarity in rainfall-runoff relationships in southeast Australia. *Journal of Hydrology*, *571*, 749–764. https://doi.org/10.1016/j.jhydrol.2019.02.025

Färber, C., Plessow, H., Mischel, S. A., Kratzert, F., Addor, N., Shalev, G., & Looser, U. (2025). GRDC-Caravan: Extending Caravan with data from the Global Runoff Data Centre. *Earth System Science Data*, *17*(9), 4613–4625. https://doi.org/10.5194/essd-17-4613-2025

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

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. https://doi.org/10.1029/2019WR026793

Feng, D., Lawson, K., & Shen, C. (2021). Mitigating Prediction Error of Deep Learning Streamflow Models in Large Data-Sparse Regions With Ensemble Modeling and Soft Data. *Geophysical Research Letters*, *48*(14), e2021GL092999. https://doi.org/10.1029/2021GL092999

Feng, D., Liu, J., Lawson, K., & Shen, C. (2022). Differentiable, Learnable, Regionalized Process-Based Models With Multiphysical Outputs can Approach State-Of-The-Art Hydrologic Prediction Accuracy. *Water Resources Research*, *58*(10), e2022WR032404. https://doi.org/10.1029/2022WR032404

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

Frame, J. M., Kratzert, F., Raney, A., Rahman, M., Salas, F. R., & Nearing, G. S. (2021). Post-Processing the National Water Model with Long Short-Term Memory Networks for Streamflow Predictions and Model Diagnostics. *JAWRA Journal of the American Water Resources Association*, *57*(6), 885–905. https://doi.org/10.1111/1752-1688.12964

Freeze, R. A., & Harlan, R. L. (1969). Blueprint for a physically-based, digitally-simulated hydrologic response model. *Journal of Hydrology*, *9*(3), 237–258. https://doi.org/10.1016/0022-1694(69)90020-1

Gannon, J. P., Kelleher, C., & Zimmer, M. (2022). Controls on watershed flashiness across the continental US. *Journal of Hydrology*, *609*, 127713. https://doi.org/10.1016/j.jhydrol.2022.127713

Gasset, N., Fortin, V., Dimitrijevic, M., Carrera, M., Bilodeau, B., Muncaster, R., Gaborit, É., Roy, G., Pentcheva, N., Bulat, M., Wang, X., Pavlovic, R., Lespinas, F., Khedhaouiria, D., & Mai, J. (2021). A 10 km North American precipitation and land-surface reanalysis based on the GEM atmospheric model. *Hydrology and Earth System Sciences*, *25*(9), 4917–4945. https://doi.org/10.5194/hess-25-4917-2021

Gasset, N., Khedhaouiria, D., Fortin, V., Dimitrijevic, M., Lauer, A., Wang, X., Bulat, M., Lemogo, L., & Pentcheva, N. (2026, January 27). *The Canadian Surface and Precipitation Reanalysis (CaSR): Highlights from Version 3.1 and Future Developments*. 106th AMS Annual Meeting. https://ams.confex.com/ams/106ANNUAL/meetingapp.cgi/Paper/467408

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

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

Gu, A., Goel, K., & Ré, C. (2021). *Efficiently Modeling Long Sequences with Structured State Spaces* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2111.00396

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

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., & Andréassian, V. (2014). Large-sample hydrology: A need to balance depth with breadth. *Hydrology and Earth System Sciences*, *18*(2), 463–477. https://doi.org/10.5194/hess-18-463-2014

Han, M., Shen, H., Tolson, B. A., Craig, J. R., Mai, J., Lin, S., Basu, N., & Awol, F. (2020). *North American Lake-River Routing Product v 2.1, derived by BasinMaker GIS Toolbox* (Version v2.1) [Dataset]. Zenodo. https://doi.org/10.5281/ZENODO.4728185

Han, M., Shen, H., Tolson, B. A., Craig, J. R., Mai, J., Lin, S. G. M., Basu, N. B., & Awol, F. S. (2023). BasinMaker 3.0: A GIS toolbox for distributed watershed delineation of complex lake-river routing networks. *Environmental Modelling & Software*, *164*, 105688. https://doi.org/10.1016/j.envsoft.2023.105688

Hirsch, R. M., Slack, J. R., & Smith, R. A. (1982). Techniques of trend analysis for monthly water quality data. *Water Resources Research*, *18*(1), 107–121. https://doi.org/10.1029/WR018i001p00107

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. van den, Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., …

Sifre, L. (2022). *Training Compute-Optimal Large Language Models* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2203.15556

Hunt, K. M. R., Matthews, G. R., Pappenberger, F., & Prudhomme, C. (2022). Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States. *Hydrology and Earth System Sciences*, *26*(21), 5449–5472. https://doi.org/10.5194/hess-26-5449-2022

Jin, J., Zhang, Y., Hao, Z., Xia, R., Yang, W., Yin, H., & Zhang, X. (2022). Benchmarking data-driven rainfall-runoff modeling across 54 catchments in the Yellow River Basin: Overfitting, calibration length, dry frequency. *Journal of Hydrology: Regional Studies*, *42*, 101119. https://doi.org/10.1016/j.ejrh.2022.101119

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling Laws for Neural Language Models* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2001.08361

Klingler, C., Schulz, K., & Herrnegger, M. (2021). LamaH-CE: LArge-SaMple DAta for Hydrology and Environmental Sciences for Central Europe. *Earth System Science Data*, *13*(9), 4529–4565. https://doi.org/10.5194/essd-13-4529-2021

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. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

Knoben, W. J. M., Thébault, C., Keshavarz, K., Torres-Rojas, L., Chaney, N. W., Pietroniro, A., & Clark, M. P. (2025). Catchment Attributes and MEteorology for Large-Sample SPATially distributed analysis (CAMELS-SPAT): Streamflow observations, forcing data

and geospatial data for hydrologic studies across North America. *Hydrology and Earth System Sciences*, *29*(20), 5791–5833. https://doi.org/10.5194/hess-29-5791-2025

Kratzert, F., Gauch, M., Klotz, D., & Nearing, G. (2024). HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin. *Hydrology and Earth System Sciences*, *28*(17), 4187–4201. https://doi.org/10.5194/hess-28-4187-2024

Kratzert, F., Gauch, M., Nearing, G., & Klotz, D. (2022). NeuralHydrology—A Python library for Deep Learningresearch in hydrology. *Journal of Open Source Software*, *7*(71), 4050. https://doi.org/10.21105/joss.04050

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

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. *Water Resources Research*, *55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065

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

Kratzert, F., 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. https://doi.org/10.5194/hess-23-5089-2019

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

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

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Kumar Sahu, R., Greve, P., Slater, L., & Dadson, S. J. (2022). Hydrological concept formation inside long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, *26*(12), 3079–3101. https://doi.org/10.5194/hess-26-3079-2022

Lehner, B., & Grill, G. (2013). Global river hydrography and network routing: Baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, *27*(15), 2171–2186. https://doi.org/10.1002/hyp.9740

Lehner, B., Verdin, K., & Jarvis, A. (2008). New Global Hydrography Derived From Spaceborne Elevation Data. *Eos, Transactions American Geophysical Union*, *89*(10), 93–94. https://doi.org/10.1029/2008EO100001

Li, X., Khandelwal, A., Jia, X., Cutler, K., Ghosh, R., Renganathan, A., Xu, S., Tayal, K., Nieber, J., Duffy, C., Steinbach, M., & Kumar, V. (2022). Regionalization in a Global Hydrologic Deep Learning Model: From Physical Descriptors to Random Vectors. *Water Resources Research*, *58*(8), e2021WR031794. https://doi.org/10.1029/2021WR031794

Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H., Tan, F., & Thieme, M. (2019). Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. *Scientific Data*, *6*(1), 283. https://doi.org/10.1038/s41597-019-0300-6

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

Liu, J., Shen, C., O'Donncha, F., Song, Y., Zhi, W., Beck, H. E., Bindas, T., Kraabel, N., & Lawson, K. (2025). From RNNs to Transformers: Benchmarking deep learning architectures for hydrologic prediction. *Hydrology and Earth System Sciences*, *29*(23), 6811–6828. https://doi.org/10.5194/hess-29-6811-2025

Liu, R., Wang, D., Wang, L., Cheng, C., Xia, X., & Yang, Z. (2025). The TAM-xLSTM Model for Hourly River Flow Forecasting: A Case Study of Qiandongnan, Guizhou Province, China. *Water*, *17*(17), 2644. https://doi.org/10.3390/w17172644

Liu, W., Yang, T., Sun, F., Wang, H., Feng, Y., & Du, M. (2021). Observation-Constrained Projection of Global Flood Magnitudes With Anthropogenic Warming. *Water Resources Research*, *57*(3), e2020WR028830. https://doi.org/10.1029/2020WR028830

Mai, J., Kornelsen, K. C., Tolson, B. A., Fortin, V., Gasset, N., Bouhemhem, D., Schäfer, D., Leahy, M., Anctil, F., & Coulibaly, P. (2020). The Canadian Surface Prediction Archive (CaSPAr): A Platform to Enhance Environmental Modeling in Canada and Globally. *Bulletin of the American Meteorological Society*, *101*(3), E341–E356. https://doi.org/10.1175/BAMS-D-19-0143.1

Mai, J., Shen, H., Tolson, B. A., Gaborit, É., Arsenault, R., Craig, J. R., Fortin, V., Fry, L. M., Gauch, M., Klotz, D., Kratzert, F., O'Brien, N., Princz, D. G., Rasiya Koya, S., Roy, T., Seglenieks, F., Shrestha, N. K., Temgoua, A. G. T., Vionnet, V., & Waddell, J. W. (2022). The Great Lakes Runoff Intercomparison Project Phase 4: The Great Lakes (GRIP-GL). *Hydrology and Earth System Sciences*, *26*(13), 3537–3572. https://doi.org/10.5194/hess-26-3537-2022

Martel, J.-L., Brissette, F., Arsenault, R., Turcotte, R., Castañeda-Gonzalez, M., Armstrong, W., Mailhot, E., Pelletier-Dumont, J., Rondeau-Genesse, G., & Caron, L.-P. (2025). Assessing the adequacy of traditional hydrological models for climate change impact studies: A case for long short-term memory (LSTM) neural networks. *Hydrology and Earth System Sciences*, *29*(13), 2811–2836. https://doi.org/10.5194/hess-29-2811-2025

Messager, M. L., Lehner, B., Grill, G., Nedeva, I., & Schmitt, O. (2016). Estimating the volume and age of water stored in global lakes using a geo-statistical approach. *Nature Communications*, *7*(1), 13603. https://doi.org/10.1038/ncomms13603

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., & Stouffer, R. J. (2008). Stationarity Is Dead: Whither Water Management? *Science*, *319*(5863), 573–574. https://doi.org/10.1126/science.1151915

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., Stouffer, R. J., Dettinger, M. D., & Krysanova, V. (2015). On Critiques of "Stationarity is Dead: Whither Water Management?": On Critiques of "Stationarity is Dead: Whither Water Management?" *Water Resources Research*, *51*(9), 7785–7789. https://doi.org/10.1002/2015WR017408

Mosaffa, H., Pappenberger, F., Prudhomme, C., Chantry, M., Rüdiger, C., & Cloke, H. (2026). A GNN routing module is all you need for LSTM Rainfall–Runoff models. *Hydrology and Earth System Sciences*, *30*(7), 2079–2092. https://doi.org/10.5194/hess-30-2079-2026

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. https://doi.org/10.1016/0022-1694(70)90255-6

Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., Voloshin, D., Kratzert, F., Elidan, G., Dror, G., Begelman, G., Nearing, G., Shalev, G., Noga, H., Shavitt, I., Yuklea, L., Royz, M., Giladi, N., Peled Levi, N., … Matias, Y. (2022). Flood forecasting with machine learning models in an operational framework. *Hydrology and Earth System Sciences*, *26*(15), 4013–4032. https://doi.org/10.5194/hess-26-4013-2022

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

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

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). https://doi.org/10.1029/2007WR006240

Pokharel, S., Roy, T., & Admiraal, D. (2023). Effects of mass balance, energy balance, and storage-discharge constraints on LSTM for streamflow prediction. *Environmental Modelling & Software*, *166*, 105730. https://doi.org/10.1016/j.envsoft.2023.105730

Pöppel, K., Beck, M., & Hochreiter, S. (2024). *FlashRNN: I/O-Aware Optimization of Traditional RNNs on modern hardware* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2412.07752

Raven Development Team. (2023). *Raven: User's and Developer's Manual v3.7*. http://raven.uwaterloo.ca/

Shamseldin, A. Y. (1997). Application of a neural network technique to rainfall-runoff modelling. *Journal of Hydrology*, *199*(3–4), 272–294. https://doi.org/10.1016/S0022-1694(96)03330-6

Shangguan, W., Dai, Y., Duan, Q., Liu, B., & Yuan, H. (2014). A global soil data set for earth system modeling. *Journal of Advances in Modeling Earth Systems*, *6*(1), 249–263. https://doi.org/10.1002/2013MS000293

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, F., Kifer, D., Li, L., Liu, X., Ren, W., Zheng, Y., Harman, C. J., Clark, M., Farthing, M., Feng, D., Kumar, P., Aboelyazeed, D., … Lawson, K. (2023). Differentiable modelling to unify machine learning and physical models for geosciences. *Nature Reviews Earth & Environment*, *4*(8), 552–567. https://doi.org/10.1038/s43017-023-00450-9

Shen, H., Tolson, B. A., & Mai, J. (2022). Time to Update the Split-Sample Approach in Hydrological Model Calibration. *Water Resources Research*, *58*(3), e2021WR031523. https://doi.org/10.1029/2021WR031523

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., Liang, X., McDONNELL, J. J., Mendiondo, E. M., O'Connell, P. E., Oki, T., Pomeroy, J. W., Schertzer, D., Uhlenbrook, S., & Zehe, E. (2003). IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal*, *48*(6), 857–880. https://doi.org/10.1623/hysj.48.6.857.51421

Slater, L. J., Arnal, L., Boucher, M.-A., Chang, A. Y.-Y., Moulds, S., Murphy, C., Nearing, G., Shalev, G., Shen, C., Speight, L., Villarini, G., Wilby, R. L., Wood, A., & Zappa, M. (2023). Hybrid forecasting: Blending climate predictions with AI models. *Hydrology and Earth System Sciences*, *27*(9), 1865–1889. https://doi.org/10.5194/hess-27-1865-2023

Smakhtin, V. U. (2001). Low flow hydrology: A review. *Journal of Hydrology*, *240*(3–4), 147–186. https://doi.org/10.1016/S0022-1694(00)00340-1

Sun, A. Y., Jiang, P., Yang, Z.-L., Xie, Y., & Chen, X. (2022). A graph neural network (GNN) approach to basin-scale river network learning: The role of physics-based connectivity and data fusion. *Hydrology and Earth System Sciences*, *26*(19), 5163–5184. https://doi.org/10.5194/hess-26-5163-2022

Tang, S., Sun, F., Liu, W., Wang, H., Feng, Y., & Li, Z. (2023). Optimal Postprocessing Strategies With LSTM for Global Streamflow Prediction in Ungauged Basins. *Water Resources Research*, *59*(7), e2022WR034352. https://doi.org/10.1029/2022WR034352

Tao, L., He, X., Li, J., & Yang, D. (2021). A multiscale long short-term memory model with attention mechanism for improving monthly precipitation prediction. *Journal of Hydrology*, *602*, 126815. https://doi.org/10.1016/j.jhydrol.2021.126815

Tolson, B. A., Craig, J. R., Lin, S., Aberi, P., Han, M., Wiebe, J., Babra, S., Kolev, S., Stapleton, M., & Shen, H. (2023, June 19). *Canadian Lake and River Hydrofabric (CLRH) version 1*. 2023 Canadian Water Resources Association Annual Conference.

Tran, V. N., Xu, D., Van Nguyen, T., Kim, T., & Ivanov, V. Y. (2025). CAMELSH: A Large-Sample Hourly Hydrometeorological Dataset and Attributes at Watershed-Scale for CONUS. *Scientific Data*, *12*(1), 1307. https://doi.org/10.1038/s41597-025-05612-6

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., & Shen, C. (2021). From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature Communications*, *12*(1), 5988. https://doi.org/10.1038/s41467-021-26107-z

Tudaji, M., Nan, Y., & Tian, F. (2025). Assessing the value of high-resolution rainfall and streamflow data for hydrological modeling: An analysis based on 63 catchments in southeast China. *Hydrology and Earth System Sciences*, *29*(7), 1919–1937. https://doi.org/10.5194/hess-29-1919-2025

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Villarini, G., & Wasko, C. (2021). Humans, climate and streamflow. *Nature Climate Change*, *11*(9), 725–726. https://doi.org/10.1038/s41558-021-01137-z

Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S., & Gupta, H. V. (2003). Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis. *Hydrological Processes*, *17*(2), 455–476. https://doi.org/10.1002/hyp.1135

Wan, H., Wang, J., Gan, Q., Xia, Y., Chang, Y., & Yan, H. (2025). Addressing intermittency in medium-term photovoltaic and wind power forecasting using a hybrid xLSTM-TCCNN model with numerical weather predictions. *Renewable Energy*, *253*, 123618. https://doi.org/10.1016/j.renene.2025.123618

Wang, Y., & Karimi, H. A. (2022). Impact of spatial distribution information of rainfall in runoff simulation using deep learning method. *Hydrology and Earth System Sciences*, *26*(9), 2387–2403. https://doi.org/10.5194/hess-26-2387-2022

Wang, Y., Zhang, L., Yu, A., Erichson, N. B., & Yang, T. (2025). A Deep State Space Model for Rainfall-Runoff Simulations. *Water Resources Research*, *61*(12), e2025WR039888. https://doi.org/10.1029/2025WR039888

Wang, Z., & Yang, Y. (2024). Stationarity of High- and Low-Flows Under Climate Change and Human Interventions Across Global Catchments. *Earth and Space Science*, *11*(1), e2023EA003456. https://doi.org/10.1029/2023EA003456

Water Survey of Canada (WSC). (n.d.). *Sub-daily hydrometrics data download site*. Retrieved February 26, 2026, from https://collaboration.cmc.ec.gc.ca/cmc/hydrometrics/www/UnitValueData/Discharge/corrected/

Wilbrand, K., Taormina, R., Ten Veldhuis, M.-C., Visser, M., Hrachowitz, M., Nuttall, J., & Dahm, R. (2023). Predicting streamflow with LSTM networks using global datasets. *Frontiers in Water*, *5*, 1166124. https://doi.org/10.3389/frwa.2023.1166124

Wu, Z., Yin, H., He, H., & Li, Y. (2022). Dynamic-LSTM hybrid models to improve seasonal drought predictions over China. *Journal of Hydrology*, *615*, 128706. https://doi.org/10.1016/j.jhydrol.2022.128706

Xie, J., Liu, X., Tian, W., Wang, K., Bai, P., & Liu, C. (2022). Estimating Gridded Monthly Baseflow From 1981 to 2020 for the Contiguous US Using Long Short-Term Memory (LSTM) Networks. *Water Resources Research*, *58*(8), e2021WR031663. https://doi.org/10.1029/2021WR031663

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. https://doi.org/10.1029/2019WR024873

Yang, Y., Roderick, M. L., Yang, D., Wang, Z., Ruan, F., McVicar, T. R., Zhang, S., & Beck, H. E. (2021). Streamflow stationarity in a changing world. *Environmental Research Letters*, *16*(6), 064096. https://doi.org/10.1088/1748-9326/ac08c1

Yang, Y., Zhang, W., Li, F., Wang, H., & Zhang, H. (2026). Fusing dynamic physical constraints with PINN-xLSTM to enhance accuracy and physical consistency in runoff prediction under extreme hydrological events. *Journal of Hydrology*, *672*, 135310. https://doi.org/10.1016/j.jhydrol.2026.135310

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

Yokoo, K., Ishida, K., Ercan, A., Tu, T., Nagasato, T., Kiyama, M., & Amagasaki, M. (2022). Capabilities of deep learning models on learning physical relationships: Case of rainfall-runoff modeling with LSTM. *Science of The Total Environment*, *802*, 149876. https://doi.org/10.1016/j.scitotenv.2021.149876

Young, P. C., & Beven, K. J. (1994). Data-based mechanistic modelling and the rainfall-flow non-linearity. *Environmetrics*, *5*(3), 335–363. https://doi.org/10.1002/env.3170050311

Yu, Q., & Tolson, B. (2026). How much historical data do we need? The role of data recency and training period length in LSTM-based rainfall-runoff modeling. *Journal of Hydrology*, *668*, 135046. https://doi.org/10.1016/j.jhydrol.2026.135046

Yu, Q., Tolson, B. A., Shen, H., Han, M., Mai, J., & Lin, J. (2024). Enhancing long short-term memory (LSTM)-based streamflow prediction with a spatially distributed approach. *Hydrology and Earth System Sciences*, *28*(9), 2107–2122. https://doi.org/10.5194/hess-28-2107-2024

Zhang, Q., Wu, J., Wang, T., Tang, C., & Liu, W. (2026). Coupling SWAT and Interpretable Deep Learning to Improve Streamflow Simulation in Arid Regions. *Water Resources Management*, *40*(2), 56. https://doi.org/10.1007/s11269-025-04422-0

Appendices

A2 Supplement for Chapter 2

A2-1 Map of the streamflow gauging stations

Figure A2-1 below shows the modelled gauging station locations (by watershed outlet location). Notably, it emphasizes the positions of non-GRIP-GL stations in comparison to GRIP-GL stations.

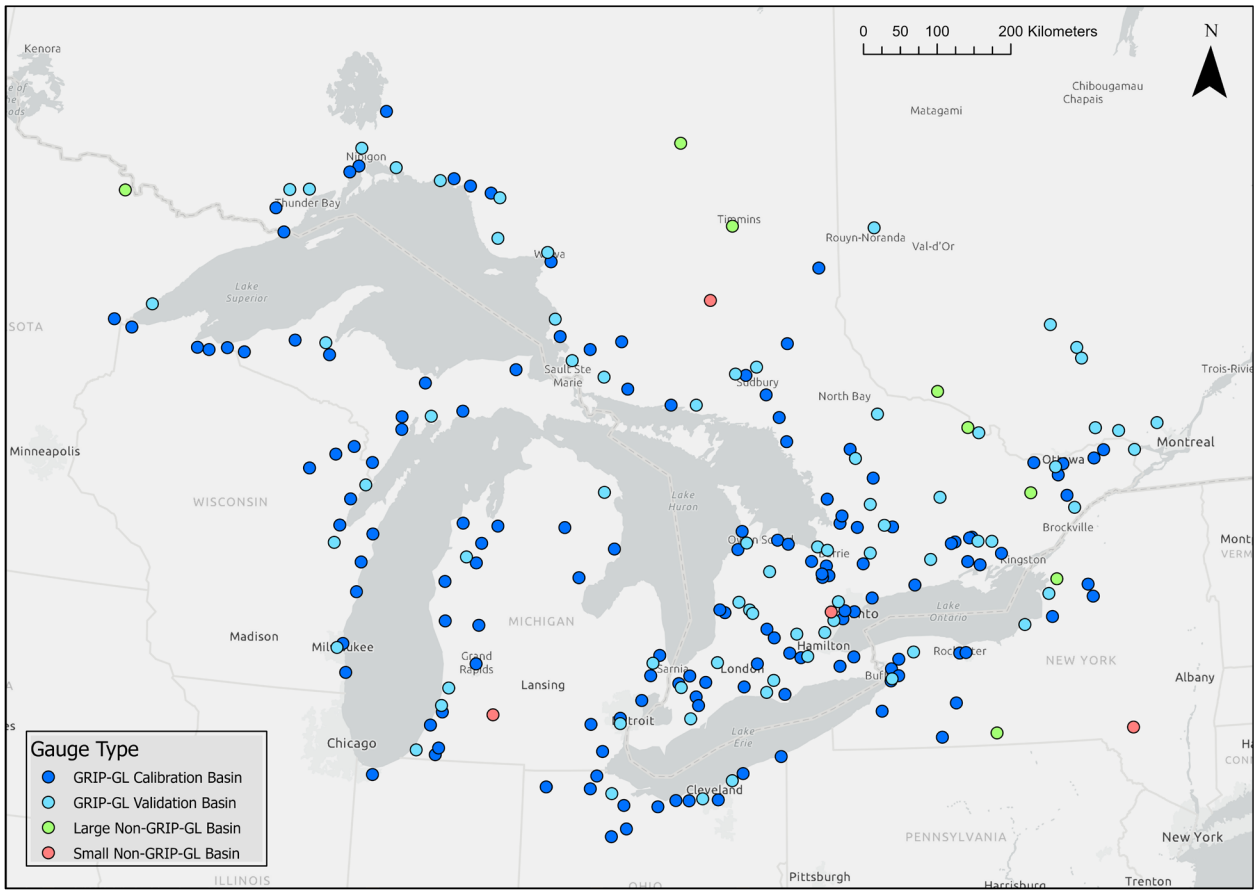


Figure A2-1. Locations of the streamflow gauging stations used in this study. In total, there are 224 streamflow gauging stations (shown as dots on the map). The 212 stations used in the GRIP-GL study are shown in dark blue dots (calibration) and light blue dots (validation). The stations of the 4 small Non-GRIP-GL basins (described in Sec 3.1 of the main article) are shown in light red dots, and the stations of the 8 large Non-GRIP-GL basins (described in Sec 3.2 of the main article) are shown in light green dots.

A2-2 Raven configurations for the routing-only mode

The routing-only model is defined within the Raven hydrologic modelling framework (Craig et al., 2020). It takes the LSTM-predicted subbasin-level streamflow as inputs and then

simulates the water routing through the delineated lake-river network. The mean daily subbasin-level streamflow is taken as precipitation at the hourly time-step (24 hours of constant precipitation intensity rate) and the precipitation in each subbasin is instantly flushed to the subbasin outlet. This approach ensures there is no in-catchment routing delay in the simulation, as would normally occur in a typical routing application that move surface runoff across a subbasin into the channel with a unit hydrograph. The model combines this subbasin-level streamflow at the subbasin outlet with in-channel routed streamflow that entered the subbasin from upstream subbasins. Figure A2-2 below shows the structure of the .rvi file (the primary input file of Raven) that defines the routing-only model.

```
# ------------------------------------------------
# Raven Input file
# ------------------------------------------------
:StartDate             2001-01-01 00:00:00
:EndDate               2017-12-31 00:00:00
:Method                ORDERED_SERIES
:TimeStep              01:00:00
:RunName               02HC017

:CatchmentRoute        ROUTE_DUMP
:Routing               ROUTE_DIFFUSIVE_WAVE
:PrecipIceptFract      PRECIP_ICEPT_NONE
:PotentialMeltMethod   POTMELT_NONE
:SoilModel             SOIL_ONE_LAYER

:HydrologicProcesses
  :Precipitation     PRECIP_RAVEN            ATMOS_PRECIP     PONDED_WATER
  :Flush             RAVEN_DEFAULT           PONDED_WATER     SURFACE_WATER
:EndHydrologicProcesses

:EvaluationMetrics                     NASH_SUTCLIFFE    PCT_BIAS    KLING_GUPTA

:EvaluationPeriod TRAINING 2001-01-01 2010-12-31
:EvaluationPeriod TESTING 2011-01-01 2017-12-31
```

Figure A2-2. The .rvi file structure for implementing the routing-only setup.

The ‘:CatchmentRoute’ command specifies the in-catchment routing method, which is used to represent how water is transported from the subbasin tributaries to the subbasin outlets. Since subbasins are the smallest response unit in this study, we choose the ‘ROUTE_DUMP’ algorithm as the in-catchment routing method, in which all of the water is instantly dumped to the subbasin outlet.

The ‘:Routing’ command specifies the in-channel routing method, which is used to represent how water is transported from upstream to downstream within the primary subbasin channels. We choose the ‘ROUTE_DIFFUSIVE_WAVE’ algorithm as the in-channel routing

method, which applies an analytical solution to the diffusive wave equation through the reach using a constant reference celerity (Raven Development Team, 2023).

The ':PrecipIceptFract' command specifies the method for estimating the fraction of precipitation intercepted by canopy. Since the LSTM-predicted subbasin-level streamflow inputs are taken as precipitation in Raven, we used the 'PRECIP_ICEPT_NONE' algorithm to indicate that there are no canopy interception processes.

The ':PotentialMeltMethod' command specifies the method for estimating the potential snow melt. We adopted the 'POTMELT_NONE' algorithm because snow melt processes are not considered in the routing-only mode.

The ':SoilModel' command specifies the soil structure model used for routing. We used the 'SOIL_ONE_LAYER' mode which defines single soil layer structure. It is worth mentioning that soil water is not simulated in the routing-only mode but Raven requires an input for the ':SoilModel' option, even if it is not used.

The ':HydrologicProcesses' block defines the hydrologic processes to be simulated. The routing-only mode only included two hydrologic processes, ':Precipitation' and ':Flush'. As mentioned previously, the subbasin-level streamflow inputs are taken to mimic precipitation, and then flushed on the land surface to the stream reach.

```
#-----------------------------------------------
# This is a Raven lake rvh file generated
# by BasinMaker v2.0
#-----------------------------------------------
###############################################
# New Lake starts
###############################################
:Reservoir   Lake_108369
  :SubBasinID  3047150
  :HRUID   26
  :Type RESROUTE_STANDARD
  :WeirCoefficient  0.6
  :CrestWidth    10.7983
  :MaxDepth 12.7
  :LakeArea    7515656.781559977
  :SeepageParameters   0   0
:EndReservoir
```

Figure A2-3. The .rvh file structure for representing lakes/reservoirs in the routing model.

Furthermore, the routing-only model is configured to simulate lakes/reservoirs (water levels and outflows) with a surface area greater than five km^2. Figure S3 shows the Raven command block that define a lake/reservoir in the routing network, the command block is automatically generated by the BasinMaker library (Han et al., 2023). All lakes are assumed to be vertically prismatic and are simulated with a lake outlet structure that is a broad-crested weir. Raven applies Equation A2-1 below to simulate the lake/reservoir outflow:

$$Q(s) = \frac{2}{3}\sqrt{2g} * C * L * s^{3/2} \qquad (A2-1)$$

where g is the gravitational constant, C represents the weir coefficient, L is the crest width, and s is the lake stage relative to the weir crest elevation. All lakes are initialized with a water level of zero. The weir coefficient is a fixed value of 0.6. The crest width is assigned with a region-dependent value according to the delineation methodology employed in the BasinMaker library. Each lake is represented by its corresponding 'lake subbasin' in the routing network and there are no channels within the lake subbasin. All lake inlets are defined as upstream subbasin outlets, and the single lake outlet corresponds to the lake subbasin outlet.

A2-3 Lumped LSTM configurations

Recall that the lumped LSTM is a replicate of the LSTM built for the GRIP-GL study (Mai et al., 2022), and the settings/input description below is taken from the supplemental information from that study.

Table A2-1. Hyperparameter settings used for the lumped LSTM.

Hyperparameter	value/range
Hidden size	256
Batch size	64
Epochs	30
Learning rate[1]	(**0**: 0.0005, **20**: 0.0001, **30**: 0.00005)
Sequence length	365 days
Dropout	0.4
Loss function	Nash-Sutcliffe efficiency (NSE)[2]

[1](**0**: x_1, **5**: x_2, **10**: x_3) denotes learning rate x_1 for epoch 0 to 4, learning rate x_2 for epoch 5 to 9, and learning rate x_3 after epoch 10.

[2]Basin-averaged NSE loss – the average of the NSE values across the calibration/training basins (Kratzert et al., 2019).

The loss function is shown in Equation A2-2 below:

$$\mathrm{NSE} = \frac{1}{B} + \sum_{b=1}^{B} \sum_{n=1}^{N} \frac{(\hat{y}_n - y_n)^2}{(s(b) + \epsilon)^2} \qquad (\mathrm{A2} - 2)$$

where B is the number of basins, N is the number of samples (days) per basin B, $\hat{y}_n$ is the prediction of sample n ($1 \leq \mathrm{n} \leq N$), y_n is the observation, and $s(b)$ is the standard deviation of the discharge in basin b ($1 \leq b \leq B$), calculated from the training period (Kratzert et al., 2019). The term ϵ is a small positive constant added to the denominator to ensure numerical stability. Specifically, it addresses the case where the standard deviation of discharge in a basin is very small or even zero.

The input variables required to drive the LSTM are grouped into dynamic attributes (Table A2-2) computed at the daily time step, and static attributes (Table A2-3) which are taken as constant through the LSTM training period.

Table A2-2. CaSR-v2 forcing variables used for the lumped LSTM. For initial lumped LSTM training, these variables are derived for each of the gauged basins. For the SR-model, these variables are derived for each of subbasins in the routing network.

Variable name	Unit
Quantity of precipitation*	mm
Downward solar flux	W/m^2
Minimum air temperature	°C
Maximum air temperature	°C
Specific humidity	kg/kg
U wind component	kts
V wind component	kts
Surface pressure	mb
Potential evapotranspiration (PET)	mm/day

*Note: CaSR precipitation was originally provided in m and was converted to mm.

Table A2-3. Static attributes used for the lumped LSTM. For initial lumped LSTM training, these variables are derived for each of the gauged basins. For the SR-model, these variables are derived for each of subbasins in the routing network.

Attribute Name	Unit
Mean daily precipitation	mm/day
Mean daily PET	mm/day
Aridity (ratio of mean PET to mean precipitation)	-
Mean daily temperature	°C
Fraction of precipitation falling on days with mean daily temperatures below 0°C	-
Frequency of days with high precipitation (≥ 5 times the mean daily precipitation)	days/year
Average duration of high-precipitation events	days
Frequency of dry days (daily precipitation less than 1 mm/day)	days/year
Average duration of dry periods	days
Catchment mean elevation	meter
Standard deviation of catchment elevation	meter
Catchment mean slope	meter/kilometre
Standard deviation of catchment slope	meter/kilometre
Catchment area	km^2
Fraction of land covered by 'Temperate-or-sub-polar-needleleaf-forest'	-
Fraction of land covered by 'Temperate-or-sub-polar-grassland'	-
Fraction of land covered by 'Temperate-or-sub-polar-shrubland'	-
Fraction of land covered by 'Temperate-or-sub-polar-grassland'	-
Fraction of land covered by 'Mixed-Forest'	-
Fraction of land covered by 'Wetland'	-
Fraction of land covered by 'Cropland'	-
Fraction of land covered by 'Barren-Lands'	-
Fraction of land covered by 'Urban-and-Built-up'	-
Fraction of land covered by 'Water'	-
Soil bulk density	g/cm^3
Soil clay content	% of weight
Soil gravel content	% of volume
Soil organic carbon	% of weight
Soil sand content	% of weight
Soil silt content	% of weight

A3 Supplement for Chapter 3

Table A3-1. Temporal distribution of streamflow data availability across all 1100 training watersheds for each training period block (1950–1965, 1965–1980, 1980–1995, 1995–2010).

Period	Mr_lt5	Data_availability (%)	Peakflow_event_count
1950 - 1965	932	96.1	5611
1965 - 1980	1067	99.4	6448
1980 - 1995	1045	99.1	6158
1995 - 2010	1033	98.9	7017
1950 - 2010	**946**	**98.5**	**25234**

Notes:

- Mr_lt5 – Number of watersheds with missing rate below 5% during the given period block. This number is slightly lower during 1950–1965, primarily because a small subset of gauges lack data for the first two years (1950–1951). Additionally, a total of 748 training watersheds contain no missing values during the entire training period (1950–2010).
- Data_availability (%) – Overall data availability across the training set, computed as the ratio of all observed records to the maximum number of records within each period (total number of days in the period × number of watersheds)
- Peakflow_event_count – Total number of peak flow events in each training period across all training watersheds, defined as days exceeding the 99.9th percentile of each watershed's full record (1950–2023). Note that this differs from the definition in Section 3.3 (based on testing watersheds and testing period only) and reflects characteristics of training watersheds.

Table A3-2. Summary of Seasonal Mann–Kendall (MK) trend test results (1950–2023) for precipitation (P), mean temperature (T), streamflow (Q) across all 1374 watersheds.

Variable	Sig_increasing_pct	Sig_decreasing_pct	No_sig_trend_pct
P	64.8	3.4	31.8
T	58.4	13.0	28.6
Q	52.6	19.5	27.9

Notes:

- Sig_increasing_pct – Percentage of watersheds with statistically significant monotonic increasing trends ($p < 0.05$).
- Sig_decreasing_pct – Percentage of watersheds with statistically significant monotonic decreasing trends ($p < 0.05$).
- No_sig_trend_pct – Percentage of watersheds without significant trends ($p \geq 0.05$).

Table A3-3. Static watershed attributes used as LSTM input variables.

Attribute Name	Unit
Drainage Area	km^2
Mean Elevation	meter
Mean Slope	degree
Mean Aspect	degree
Mean Permeability	m^2
Mean Porosity	—
Forest Fraction	—
Grass Fraction	—
Wetland Fraction	—
Water Fraction	—
Urban Fraction	—
Shrubs Fraction	—
Crops Fraction	—
Fraction of Snow & Ice	—
Centroid Latitude	degree North
Centroid Longitude	degree East



Figure A3-1. Nash–Sutcliffe Efficiency (NSE) results for Experiment 1.

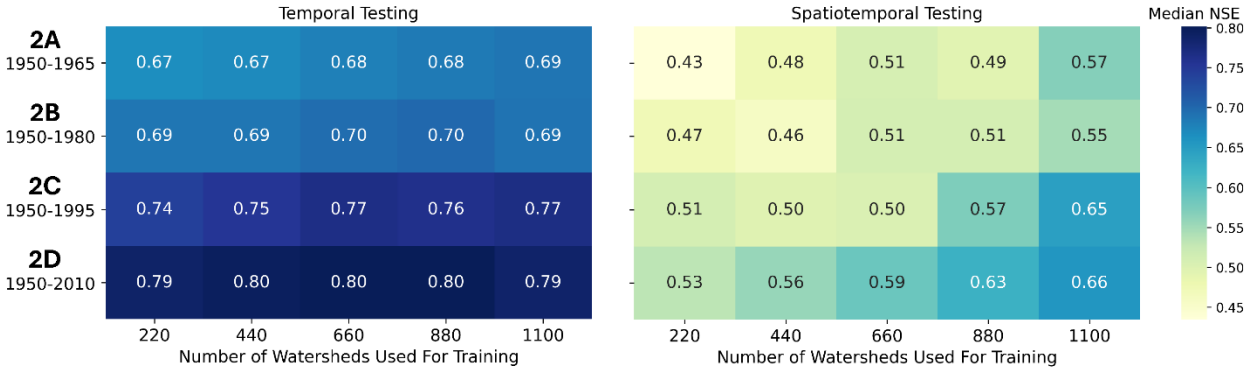


Figure A3-2. NSE results for Experiment 2.

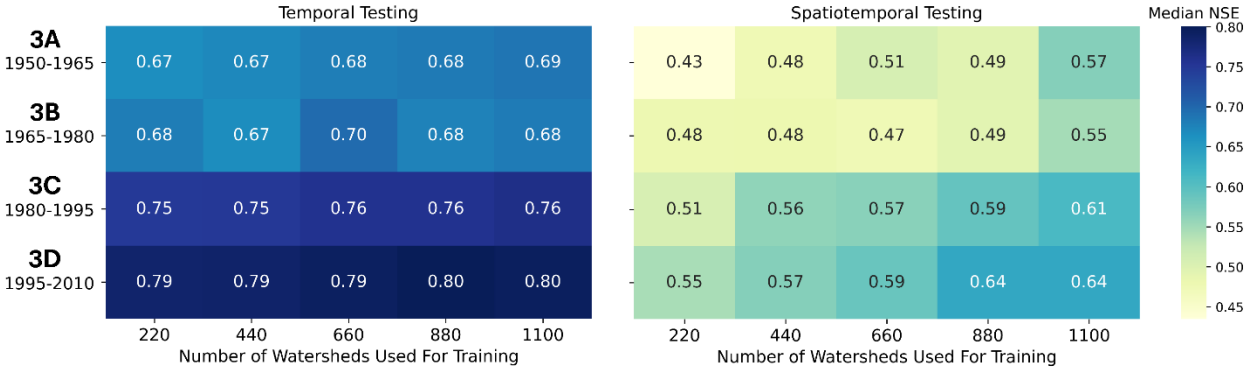


Figure A3-3. NSE results for Experiment 3.

A4 Supplement for Chapter 4

Table A4-1. CaSR forcing variables

Variable name	Level	Unit
Quantity of precipitation*	surface	mm
Air temperature	1.5m	°C
Specific humidity	1.5m	kg/kg
Relative humidity (fraction)	1.5m	-
Downward solar flux	surface	W/m^2
Surface incoming infrared flux	surface	W/m^2
Wind Modulus (derived using UU and VV)	10m	Kts

*Note: CaSR precipitation was originally provided in m and was converted to mm.

Table A4-2. Static attributes

Attribute name	Description	Source
Area-km2	Catchment area	CLRH
Lake-area	Area of lakes	CLRH
Main-channel-length	the longest flow path to the outlet	CLRH
mean-bankfull-width	The average of bankfull width of river channel	CLRH
mean-bankfull-depth	The average of bankfull depth of river channel	CLRH
mean-elev	Catchment mean elevation	CLRH
elev-std*	Standard deviation of catchment elevation	CLRH-derived
mean-slope	Catchment mean slope	CLRH
slope-std*	Standard deviation of catchment slope	CLRH-derived
Wetland	Fraction of land covered by wetland	NALCMS
Cropland	Fraction of land covered by cropland	NALCMS
Barren-Lands	Fraction of land covered by barren grounds	NALCMS
Urban-and-Built-up	Fraction of land covered by man-made structures	NALCMS
Snow-and-Ice	Fraction of land covered by permanent/semi-permanent snow or ice	NALCMS
Forest-total	Fraction of land covered by forest	NALCMS
BD	Soil bulk density	GSDE
CLAY	Soil clay content	GSDE
GRAV	Soil gravel content	GSDE
OC	Soil organic carbon	GSDE

Table A4-2 (continued)

Attribute name	Description	Source
SAND	Soil sand content	GSDE
SILT	Soil silt content	GSDE

*Note: The standard deviations of catchment elevation and catchment slope were derived in this study by re-extracting topographic statistics from the digital elevation model (DEM) used by CLRH within the corresponding CLRH catchment boundaries.

Table A4-3. General hyperparameters tuning grid. The selected values are highlighted in bold.

Hyperparameter	Tested Range/Selected Value
Hidden size	64, 128, **256**, 512
Batch size	64, **128**, 256, 512
Epoch	**30**, 35
Learning rate*	(0: 0.001, 10: 0.0005, 20: 0.0001, 25: 5e-05, 30: 1e-05), **(0: 0.0005, 10: 0.0001, 20: 5e-05, 30: 1e-05)**, (0: 0.0001, 15: 5e-05, 25: 1e-05)

*Note: (0: x1, 10: x2, 20: x3) denotes learning rate x1 for epoch 0 to 9, learning rate x2 for epoch 10 to 19, and learning rate x3 after epoch 20.

Table A4-4. xLSTM-specific hyperparameters

Components	Hyperparameter	Selected Value	Description
sLSTM and mLSTM	embedding_dim	256	Input/output feature dimension of each xLSTM block. It is used as the sLSTM hidden size. Set equal to the hidden size of the benchmark LSTM models.
sLSTM	conv1d_kernel_size	4	Kernel size of the optional causal convolution applied before the sLSTM input and forget gate projections.
	num_heads	4	Number of sLSTM heads.
	bias_init	powerlaw_block dependent	Block-dependent forget-gate bias initialization, used to provide different initial memory scales across the residual stack.
	recurrent_weight_init	zeros	Initialization scheme for the recurrent mixing weights in the sLSTM cell.
	group_norm_weight	True	Enables trainable weights in the multi-head normalization after the sLSTM cell.
	ff_type	ffn_gated	Type of feed-forward module used after sLSTM.
	proj_factor	1.3	Expansion factor of the gated feed-forward module.
	act_fn	gelu	Activation function used in the gated feed-forward module.

Table A4-4 (continued)

Components	Hyperparameter	Selected Value	Description
mLSTM	context_length	687	Maximum sequence length used to construct the causal mask for parallel mLSTM computation; Set to the multi-frequency input length of 687, comprising 351 daily steps followed by 336 hourly steps.
	conv1d_kernel_size	4	Kernel size of the causal depthwise convolution before the mLSTM query and key projections.
	num_heads	4	Number of mLSTM matrix-memory heads.
	bias	False	Bias flag propagated to mLSTM linear projections.
	qkv_proj_blocksize	4	Block size controlling the structured head-wise Q/K/V projections in mLSTM.
	proj_factor	2.0	Internal up-projection factor.