Water Resources Research

RESEARCH ARTICLE
10.1029/2023WR036853

Special Collection:
Advancing Interpretable AI/ML Methods for Deeper Insights and Mechanistic Understanding in Earth Sciences: Beyond Predictive Capabilities

Streamflow Prediction in Human-Regulated Catchments Using Multiscale Deep Learning Modeling With Anthropogenic Similarities

Arken Tursun[1], Xianhong Xie[1], Yibing Wang[1], Dawei Peng[1], Yao Liu[1], Buyun Zheng[1], Xinran Wu[1], and Cong Nie[1]

[1]State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing, China

Key Points:

- Long Short-Term Memory is able to capture irregular streamflow patterns by utilizing a novel collection of multiscale attributes as input
- The performance of the model is extremely sensitive to the training data period across human-regulated catchments
- Differentiable Parameter Learning may be susceptible to the dynamics of human activities, and struggles to provide accurate simulations

Correspondence to:
X. Xie,
xianhong@bnu.edu.cn

Citation:
Tursun, A., Xie, X., Wang, Y., Peng, D., Liu, Y., Zheng, B., et al. (2024). Streamflow prediction in human-regulated catchments using multiscale deep learning modeling with anthropogenic similarities. *Water Resources Research*, *60*, e2023WR036853. https://doi.org/10.1029/2023WR036853

Received 3 DEC 2023
Accepted 28 AUG 2024

Author Contributions:
Conceptualization: Arken Tursun, Xianhong Xie
Data curation: Arken Tursun, Yao Liu
Funding acquisition: Xianhong Xie
Methodology: Arken Tursun
Project administration: Xianhong Xie
Resources: Yibing Wang, Buyun Zheng
Software: Arken Tursun
Supervision: Xianhong Xie
Validation: Arken Tursun
Visualization: Yibing Wang, Dawei Peng
Writing – original draft: Arken Tursun

Abstract Accurate streamflow prediction in human-regulated catchments remains a formidable challenge due to the complex disturbance of hydrological processes. To consider human disturbance in hydrological modeling, this study introduces a novel static attribute collection that combines river-reach attributes with catchment attributes, referred to as multiscale attributes. The attribute collection is assembled into two deep learning (DL) methods, that is, the Long Short-Term Memory (named as Multiscale LSTM) and the Differentiable Parameter Learning (DPL) model, and the performance is evaluated across 95 human-regulated catchments in the United States (USA) and 24 catchments in the Yellow River Basin in China. In the USA, the Multiscale LSTM and the DPL models achieve similar performance with median Kling-Gupta Efficiency (KGE) of 0.78 and 0.71, respectively. However, in the Yellow River Basin, the KGE values are 0.58 for Multiscale LSTM and 0.24 for DPL. These results highlight the DL models' ability to leverage multiscale attributes for improved performance compared to traditional catchment attributes. The performance of Multiscale LSTM and DPL models is predominantly influenced by river-scale attributes, encompassing factors such as connectivity status index (CSI), degree of regulation (DOR), sediment trapping (SED), and number of dams. Additionally, satellite-derived attributes such as mean and maximum river width (Width), slope and mean water surface elevation (WSE) from the Surface Water and Ocean Topography River Database (SWORD) contribute valuable insights into anthropogenic influences. Moreover, our study highlights the significance of selecting the appropriate training data period, which emerges as the most dominant factor affecting model performance across human-regulated catchments. The diversity of data during the training period enables the model to capture a broad spectrum of hydrological signatures within these catchments. Consequently, this study emphasizes the advantages of Multiscale LSTM and underscores the significance of considering both natural and anthropogenic signatures to enhance hydrological predictions within human-regulated environments.

Plain Language Summary Understanding how water flows in rivers is crucial for managing reservoirs, preventing floods, and making smart decisions about water use in human-regulated catchments. Previous research using data-driven models has predominantly focused on long short-term memory (LSTM) and differentiable parameter learning (DPL) models in natural catchments with minimal human influences. Identifying the effectiveness of these models in human-regulated catchments has posed a significant challenge. To address this gap, we intentionally compared the performance of LSTM and DPL models, incorporating multiscale attributes as inputs. The results underscore the significant superiority of the LSTM model over the DPL model in human-regulated catchments, emphasizing the crucial role of river attributes in enhancing model performance. Additionally, we observed that the DPL model exhibited higher sensitivity to the dynamics of human activities, and struggles to provide accurate simulations during periods of elevated human impacts. These findings elucidate the divergent capabilities of these models in representing hydrological processes in human-regulated catchments.

1. Introduction

Streamflow is an essential water resource for human survival, ecosystem balance, and geophysical processes, and its spatiotemporal pattern is generally disrupted by human activities (Marvel et al., 2019; Singh & Basu, 2022). Over the past century, the rising demand for freshwater necessary to accommodate the evolving needs of society has led to the construction of infrastructure such as dams, reservoirs, and canals for water storage (Déry et al., 2021; Ficklin et al., 2018; Grill et al., 2019). Consequently, more than two-thirds of river systems

Writing – review & editing: Xianhong Xie

worldwide have undergone significant alterations of their flows (Liu et al., 2023). Thus, development of highly accurate hydrological models is crucial to understanding future hydroclimate scenarios in the context of global climate change across human-regulated catchments (Tu et al., 2023; Zhang et al., 2023).

Data-driven models based on deep learning (DL) methods have been applied to obtain satisfactory predictions in different scientific areas (Gavahi et al., 2023; Jiang et al., 2023; Kumar & Rao, 2023; Tian et al., 2021). DL is an effective tool for estimating various hydrological variables (Li et al., 2022; Nevo et al., 2022; Wright et al., 2022; Yousefi & Toffolon, 2022; Zhang et al., 2022). As a typical DL model, long short-term memory (LSTM) has become a prominent technique in streamflow prediction (Feng et al., 2020; Kratzert et al., 2018, 2021; Nearing et al., 2022). Even for prediction in ungauged basins (PUB), LSTM has significant advantages compared to classical hydrological models, because it can transfer the knowledge learned from data-rich regions to ungauged ones (Arsenault et al., 2023; Feng et al., 2021; Zhang et al., 2022). LSTM models enhance hydrological modeling by identifying input–output relationships with satisfactory precision and efficiency (Wei et al., 2023). Despite the increasing interest in the application of LSTM for streamflow prediction, it is important to note that very few studies have utilized LSTM to predict streamflow within exceedingly large and human-regulated catchments, and conducted a concurrent evaluation of the reliability of these simulations. Consequently, a substantial research gap is still present in the exploration of LSTM networks for streamflow simulation, particularly in the dynamic context of evolving human impacts.

The fusion of physically based models and DL models, termed hybrid models, is increasingly gaining attention in current research, and represents a promising approach for future investigations (Ng et al., 2023). Different types of hybrid models have been extensively used in the estimation or prediction of streamflow (Wi & Steinschneider, 2022), stream temperature (Du et al., 2022), and evaporation (Bennett & Nijssen, 2021). More recently, Botterill and Mcmillan (2023) introduced a hybrid model featuring an encoder-decoder framework, which is a hydrological process-informed differentiable model that employs the acquired signatures and current meteorological data to predict streamflow. Furthermore, Feng, Beck, et al. (2023), Feng, Tan, et al. (2023) proposed a differentiable, parameter-learning (DPL) model, utilizing neural network (NN) to predict static and dynamic parameters within Hydrologiska Byråns Vattenbalansavdelning (HBV)-like process-based model. These hybrid models exhibit performance closely approaching the state-of-the-art results from LSTM and allow users to simulate and interpret unobserved variables, similar to classical hydrological models (Feng, Beck, et al., 2023; Feng, Tan, et al., 2023). However, the majority of existing studies primarily focused on natural catchments with minimal human influence (Chen et al., 2022; Gholizadeh et al., 2023; Michel et al., 2022; Zhang et al., 2022). Nevertheless, it is still necessary to systematically assess and compare hybrid models (e.g., the DPL) with LSTM specifically within human-regulated catchments.

Hydrologists often tend to choose hydrological signatures that can capture the diverse aspects of flow regimes in natural catchments based on catchment attributes (Ma, Feng, et al., 2021; Ma, Montzka, et al., 2023; Thomas Lees, 2021). However, classical catchment attributes resulting in uncertainty in human-regulated catchments that diminishes model performance (Kratzert et al., 2021; Ouyang et al., 2021). Given the diverse and direct impacts of human activities on streamflow patterns, it is clear that a comprehensive understanding of methods for accurately selecting hydrologic signatures is still lacking. Consequently, there is an opportunity from global river morphological attributes from the Surface Water and Ocean Topography (SWOT) River Database (SWORD), such as elevation, width, and slope, are available in a static format (Durand et al., 2023; Riggs et al., 2022). It is possible to leverage the capabilities of LSTM and hybrid models to identify an optimized set of hydrologic signatures that can provide the most relevant information for effectively addressing the complex challenges associated with streamflow simulation in human-regulated catchments.

This study proposes to employ river-reach attributes that comprehensively encompass various river characteristics. These attributes are combined with catchment attributes to create a novel attribute collection known as multiscale attributes. The objective of this study is to investigate whether LSTM with multiscale attributes (Multiscale LSTM) can extract valuable information for streamflow prediction in human-regulated catchments. We selected a state-of-the-art DPL model that not only clearly describes physical processes but also adheres to physical laws, such as mass conservation, while producing interpretable outputs. Deliberate comparisons between the LSTM model and the DPL model were designed to determine their ability to learn from these multiscale attributes. We examined whether these attributes can outperform traditional catchment attributes or scenarios where no attributes are used. Furthermore, we aim to explore how the diversity, quantity, and temporal

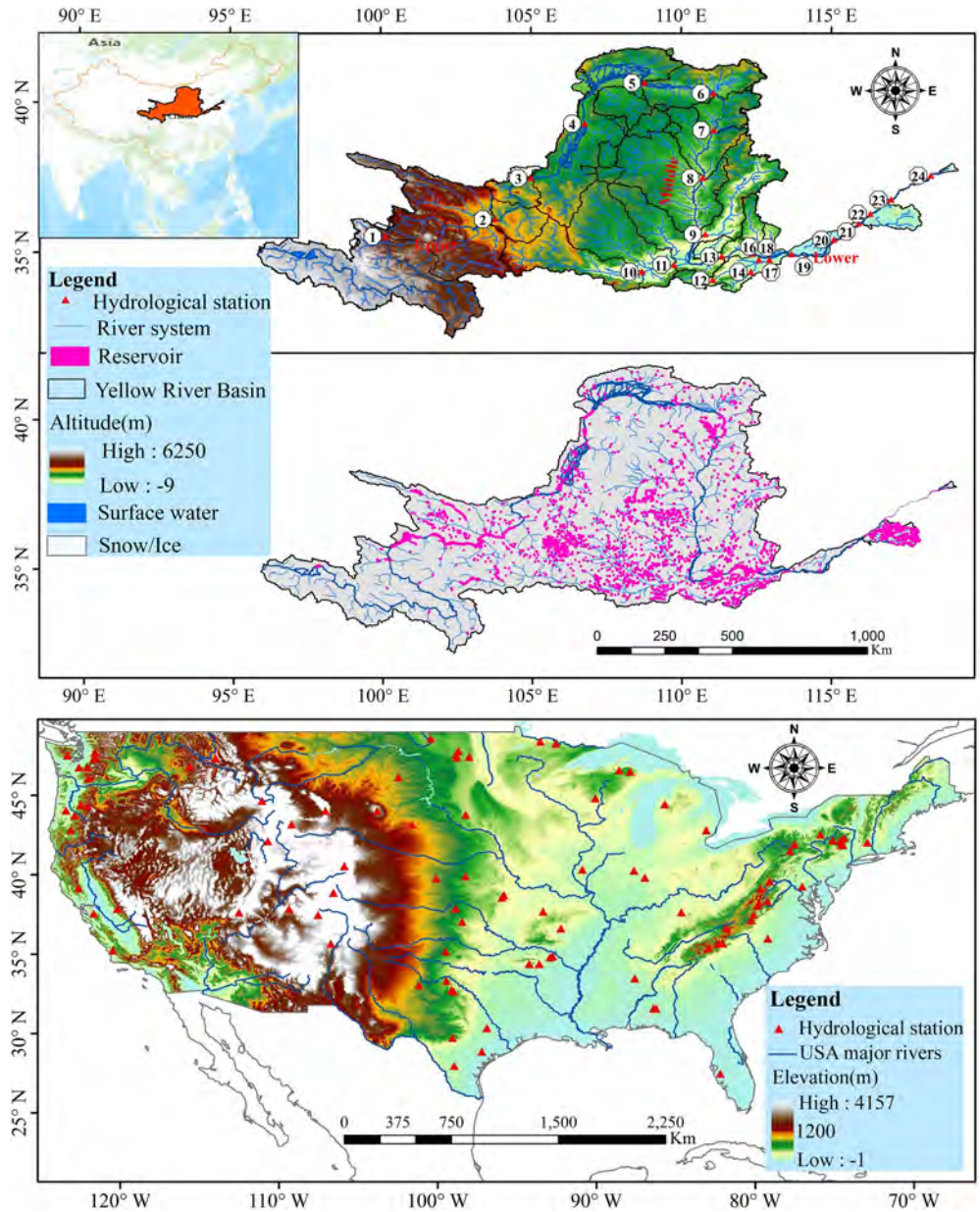

Figure 1. Locations of the stream gauges and spatial distribution of the reservoirs across the Yellow River Basin (above) the 95 catchments across the entire United States included in the CAMELS data set (below).

distribution of the training data affect the performance of the LSTM and DPL models within the context of evolving human influence dynamics. This study is expected to offer insights into the influence of the training data characteristics and provide a comprehensive understanding of the model performance in different human influence scenarios.

2. Materials and Methods

2.1. Study Area

The Yellow River Basin, located in northern China (Figure 1), which is the world's sixth-longest river system. Covering an immense expanse of approximately 750,000 km^2, the Yellow River Basin accounts for 8% of China's land area. Notably, a significant portion of this region is characterized by elevations surpassing 3,000 m above sea level, predominantly situated within the arid and semiarid zones of northwestern China. Additionally, we

expanded our data set to include 95 catchments from Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) data set (Addor et al., 2017) across the United States (USA). These catchments were selected based on a criterion where the connectivity status index (CSI) of the river is less than 95% (Figure 1).

Since the implementation of the ecological restoration project in 1999, the Yellow River Basin has witnessed a remarkable resurgence of its vegetative cover. Over this period, the vegetation coverage surged from 31.6% to an impressive 65% by 2017 (Jiang et al., 2022; Wang et al., 2021). In addition to these ecological restoration efforts, significant human-made alterations have been made to the landscape, particularly in the early 2000s when several key reservoirs and water infrastructure projects were initiated (Liu et al., 2023). The Yellow River Basin contains more than 3,000 reservoirs that have significantly modified streamflow by altering its timing and magnitude (Jiang et al., 2022; Ni et al., 2022; Xu, Ma, et al., 2022; Xu, Wang, et al., 2022). Thus, the variations in human activities before and after 2000 had a discernible impact on the flow of the Yellow River Basin, leading to noticeable changes in the water discharge (Figure A1).

2.2. Data Sets

The meteorological forcing data used in this study include precipitation, temperature, potential evaporation, and wind speed. These forcing sets were derived from ERA5-Land, a data set that generates primary land surface components through the ERA5 atmospheric model (Muñoz-Sabater et al., 2021; Xu, Ma, et al., 2022; Xu, Wang, et al., 2022). ERA5-Land offers a consistent portrayal of the historical evolution of land-related variables spanning from 1950 to the present day, providing data at a spatial resolution of 11 km with hourly temporal frequency. Using the hourly data sets, we computed the area-weighted spatial average for each variable in each basin from gridded meteorological forcing data. Daily streamflow data from a total of 24 watersheds situated within the Yellow River Basin were used for this study. The size of these watersheds ranges from 2,450 to 747,000 km^2, with a median area of approximately 330,000 km^2, as depicted in Figure 1. The 24 stations were mainly distributed along the main stream of the river. The daily streamflow data covered the period from 1980 to 2014 and were sourced from the Loess Plateau Subcenter, National Earth System Science Data Center, National Science & Technology Infrastructure of China (http://loess.geodata.cn). Please note that there are data gaps from 1998 to 2002 across the Yellow River basin. To maintain data integrity and avoid introducing uncertainties, we did not reconstruct the missing data for this period.

For our analysis of the impacts of human regulation and multiscale attributes on discharge prediction, we utilized the original CAMELS data set (Addor et al., 2017). This data set encompasses 671 catchments located across the United States. Within the CAMELS data set, there are 58 static attributes involving various aspects of basin topography, climatic indices, hydrological signatures, as well as information on land cover, soil composition, and geology. The data set also includes daily climate inputs (such as temperature, precipitation, radiation, and humidity), streamflow records spanning from 1980 to 2014.

The multiscale attributes are defined by combining river-reach attributes with catchment attributes (Tables 1 and 2), yielding in a unique collection of attributes (Figure B1). In this study, we calculated two sets of catchment attributes. The first was based on HydroATLAS data, and the second was derived from daily climate variables extracted from ERA5_Land (Kratzert et al., 2023). Notably, a majority of the basins within HydroATLAS were smaller than the actual catchment boundaries within the Yellow River basin. To address this issue, we performed spatial joins between the HydroATLAS polygons and carried out calculations to aggregate catchment attributes. These attributes encompassed diverse categories, namely hydrology, land cover characteristics, soils and geology, and climatology.

The river-reach attributes utilized in this study originated from the Global Free Flowing River (FFR) data set, as described in detail by Grill et al. (2019). The FFR data set defines free-flowing rivers based on an extensive literature review and identifies five primary pressure factors that influence river connectivity. These pressure indicators include the degree of river fragmentation (DOF), degree of regulation (DOR), sediment trapping index (SED), consumption water use (USE), road density (RDD), and urban areas (URB). These proxy indicators were then integrated using a weighting model to derive the river CSI for every river reach. In our analysis, river reaches with a CSI of ≥95% were considered to have good connectivity status, while those below 95% were classified as impacted (Figure B1). Please note that the way for calculating river-reach attributes significantly differs from that of catchment attributes. Rather than computing an average value based on all river-reaches within a catchment, our approach specifically focused on the attributes at the outlet of each catchment. This deliberate choice allowed

Table 1
Multiscale Attributes Used to Train the Long Short-Term Memory and Differentiable Parameter Learning Models Across the Yellow River Basin

Static variables	Static variable description	Median	Range
Catchment attributes			
Area	Catchment area (km^2)	330,000	[2,450, 747,000]
ele_mt	Elevation (m)	1,675	[985, 4,800]
p_mean	Mean daily precipitation (mm d^{-1})	1.9	[0.3, 2.7]
ari_ix	Global aridity index (%)	58.6	[37, 75]
pnv_pc	Potential natural vegetation extent (%)	0.54	[0, 81]
pet_mean	Mean daily potential evaporation (mm d^{-1})	3.7	[2.3, 4.3]
aet_mm	Monthly mean actual evapotranspiration (mm)	45	[34, 65]
high_prec_dur	Average duration of high precipitation events (-)	1.3	[1.2, 1.4]
high_prec_freq	Frequency of high precipitation days (-)	0.05	[0, 0.8]
low_prec_freq	Frequency of low precipitation days (-)	0.57	[0.50, 0.77]
low_prec_dur	Average duration of low precipitation events (-)	4.9	[3.5, 8.3]
swc_pc_	Soil water content (%)	58	[30, 73]
glc_pc	Forest cover extent (%)	1.45	[0.01, 12.6]
cly_pc	Clay fraction in soil (%)	18.2	[10, 22]
frac_snow	Fraction of precipitation falling as snow (mm d^{-1})	0.06	[0, 0.35]
moisture_index	Mean annual moisture index (-)	−0.5	[0.3, −0.75]
pst_pc	Pasture cover extent (%)	36	[9, 58]
ire_pc	Irrigated area extent (%)	10	[0, 31]
River attributes			
CSI	River connectivity status index (%)	68	[21, 85]
DOR	Degree of regulation (%)	46	[80, 100]
DOF	Degree of fragmentation (%)	38	[0, 100]
URB	Night light intensity in urban areas (%)	8	[0, 45]
USE	Water use for irrigation, industry, municipal (%)	13	[0, 70]
Num_dams	Number of dams on the river (-)	18	[0, 133]
WSE	Reach average water surface elevation (m)	102	[6, 1,903]
Width_var	Width variance along the reach (m)	1.01	[0, 90.9]
Width	Reach average width (m)	87	[15, 596]
Max_width	Maximum width value across the channel (m)	296	[20, 1,209]
Slope	Reach average slope (m/km)	0.49	[0, 30]
Dist_out	Distance from the river outlet (m)	1,145,247	[1,564, 1,875,638]
Facc	Maximum flow accumulation value (km^2)	3,158	[234, 1,473,597]

us to capture and represent the anthropogenic signatures at the finer scale of individual river-reaches. The geometric data set of the global river attribute information for every river reach, the values of all pressure indicators (DOF, DOR, SED, USE, RDD, and URB) and the values for the CSI are available at https://doi.org/10.6084/m9.figshare.7688801 (Grill et al., 2019).

For the extraction of remote sensing-based river-reach scale characteristics, our primary data source was SWOT River Database (SWORD v16), a comprehensive database regarding global river networks and morphological attributes. SWORD provides a cohesive data set comprising river reaches, each approximately 10 km in length, and nodes spaced approximately 200 m apart. These data sets include crucial hydrological and morphological features

Table 2
Multiscale Attributes Used to Train the Long Short-Term Memory and Differentiable Parameter Learning Models Across the 95 Catchments in USA

Static variables	Static variable description	Median	Range
Catchment attributes			
elev_mean	Elevation (m)	560	[33, 3,370]
slope_mean	Catchment mean slope (m/km)	31	[1, 160]
area_gages2	Catchment area (km^2)	357	[8.9, 7,582]
frac_forest	Forest fraction (-)	0.79	[0.0, 1.0]
lai_max	Maximum monthly mean of the leaf area index (-)	3.26	[0.73, 5.47]
lai_diff	Difference between the maximum and minimum of the leaf area index (-)	2.27	[0.47, 4.64]
dom_land_cover_frac	Fraction of the catchment area associated with the dominant land cover (-)	0.92	[0.31, 1.0]
root_depth_50	Root depth at 50th percentiles (m)	0.18	[0.12, 0.25]
soil_depth_statsgo	Soil depth (m)	1.35	[0.61, 1.50]
soil_porosity	Volumetric soil porosity (-)	0.44	[0.37, 0.52]
soil_conductivity	Saturated hydraulic conductivity (cm/hr)	1.39	[0.45, 8.79]
p_mean	Mean daily precipitation (mm/day)	3.0	[1.09, 8.79]
aridity	PET/P	0.85	[0.22, 2.60]
high_prec_freq	Frequency of high precipitation days (days/year)	22	[9, 29]
high_prec_dur	Average duration of high precipitation events (days)	1.32	[1.09, 1.97]
low_prec_freq	Frequency of dry days (days/year)	255	[178, 322]
low_prec_dur	Average duration of dry periods (days)	5.22	[2.83, 16.46]
frac_snow	Fraction of precipitation falling as snow (-)	0.13	[0, 0.72]
p_seasonality	Seasonality and timing of precipitation (-)	0.11	[−1.37, 0.92]
pet_mean	Mean daily PET (mm/day)	2.60	[1.90, 3.88]
max_water_content	Maximum water content (m)	0.53	[0.17, 0.76]
geol_porostiy	Subsurface porosity (-)	0.14	[0.01, 0.24]
geol_permeability	Subsurface permeability (m^2)	−14	[−16, −11]
River attributes			
CSI	River connectivity status index (%)	81	[24, 96]
DOR	Degree of regulation (%)	0	[0, 100]
DOF	Degree of fragmentation (%)	57	[0, 100]
URB	Night light intensity in urban areas (%)	0	[0, 16]
SED	Sediment trapping index (%)	0	[0, 97]
USE	Water use for irrigation, industry, municipal (%)	0.20	[0, 32]
Width	Reach average width (m)	84	[30, 23,128]
Slope	Reach average slope (m/km)	0.29	[0, 25.37]
WSE	Reach average water surface elevation (m)	197	[0, 1,347]
Width Variance	Width variance along the reach (m)	1,134	[21, 1,492]
Max Width	Maximum width value across the channel (m)	45	[36, 23,128]
WSE Variance	WSE variance along the reach (m)	0.46	[0, 7,930]

such as width, slope, meander length, maximum width, and width trend. SWORD integrates data from multiple global hydrography databases into a unified product, ensuring consistency and accuracy across various attributes.

We collected dam and reservoir data from georeferenced global dam and reservoir data sets (GeoDAR) (Wang et al., 2022). We calculated the numbers of reservoirs and river dams in every catchment and added these data into static attributes to improve the streamflow modeling.

2.3. Methods

2.3.1. LSTM

LSTM introduces several crucial components, including memory cells and gates, as proposed by Hochreiter and Schmidhuber. (1997). These gates include the input gate, forget gate, and output gate, each serving a distinct function (Kratzert et al., 2019; Tiggeloven et al., 2021). The input gate governs the information that should be incorporated, the forget gate decides what should be discarded from memory, and the output gate determines what information to produce as output. Importantly, these gates undergo automatic and simultaneous training based on the input data to predict the target variable (Ha et al., 2021; Ma, Feng, et al., 2021; Ma, Montzka, et al., 2023; Shen et al., 2021). The incorporation of a memory cell endows LSTM with a unique ability to capture and comprehend long-term dependencies, such as those associated with snow accumulation and subsurface water storage that are crucial for accurate streamflow predictions. Detailed information regarding the LSTM model parameters is provided in Table 3. Moreover, Figure 2 visually describes the intricate mechanisms operating within an LSTM cell, showcasing its capacity to effectively model sequential data and extract valuable insights for hydrological applications.

The LSTM model used in this study is designed to predict 1 day of streamflow at a single time step, using the data for the different dynamic variables recorded over the previous 365 days. Moreover, we formulated distinctive features that comprehensively described topographic, land-cover, and climatic attributes for each catchment. These features served as static inputs. In addition to static catchment attributes, we introduced another special static attribute at the scale of river reaches. These attributes provide insights into the influence of human activities on the respective river reaches. This attribute integration enables the model to capture the effects exerted on river reaches due to human-induced factors. The selection of multiscale (catchment, river) attributes was guided by the aim to offer hydrological insights that would enable the LSTM model to effectively differentiate among the various catchment behaviors related to rainfall-runoff processes.

2.3.2. Differentiable, Learnable Process-Based Model

We trained the Differentiable, Parameter learning (DPL) model as introduced by Feng, Beck, et al. (2023), Feng, Tan, et al. (2023). The DPL model develops a process-based model as an evolvable backbone on a differentiable computing platform and uses intermingled NNs for parameterization. We used dynamic parameters for the HBV model, which means that the model obtains a new parameter every day from the parameter estimation NN. The DPL model is designed to support differentiable programming because it is implemented within a DL platform. Detailed information regarding the DPL model parameters is provided in Table 4. Within the DPL model, certain components can be replaced with NNs, allowing for the corresponding updates to the model structure. A distinctive feature of the model lies in the training of the NNs in the "end-to-end" manner, without any supervisory data guiding the outputs of the NN units. The principal loss function of the DPL model is computed based on the observed data and the outputs of the process-based model.

2.4. Experiment Setup

In this study, we propose a hypothesis that the performance of both the LSTM and DPL models will be noticeably influenced by the diversity and quantity of the training data set. To examine this hypothesis, we conducted a series of experiments involving variations in the amount of the source training data (Figure 2). We trained the LSTM and DLP models with different training strategies. Specifically, we trained the models with five different random seeds for every single experiment, and the performance was selected based on the maximum of the five model predictions. Due to the different random seeds and irregular behavior of streamflow, the models achieve very different performance characteristics in every experiment for single catchments, and choosing the max value of Kling-Gupta Efficiency (KGE) among them improves the model performance.

2.4.1. Setup With Multiscale Attributes

In this study, our hypothesis posits that the both LSTM and DPL models will exhibit a tendency to cluster similar basins together within a high-dimensional space. This clustering facilitates the utilization of similar memory cells for human-regulated catchments, contingent upon the information contained within the multiscale attributes employed as inputs. Consequently, the models should learn to recognize anthropogenic similarities in a manner

Table 3
List of Long Short-Term Memory Model Hyperparameters

No	LSTM parameters	Value
1	Number of LSTM layers	2
2	Hidden states	128
3	Initial forget bias	3
4	Dropout rate	0.4
5	Learning rate	0.005,0.001
6	Batch size	256
7	Optimizer	Adam
8	Number of training epochs	50
9	Sequence length	365
10	Loss function	NSE

conducive to rainfall-runoff prediction. To validate this hypothesis and enhance the robustness of our analysis, experiments were conducted in both the Yellow River basin and the United States. From the CAMELS data set, we specifically selected 95 catchments out of the total 671, employing a criterion where the CSI of the river is less than 95%.

In this experiment, for across the Yellow River Basin, we deliberately selected the training period from 1 January 1986 to 30 December 2009. And then we used 5 years' worth of data from 1 January 2010 to 30 December 2014, to evaluate the model's performance. Before the training, we used 6 years' worth of data from 1 January 1980 to 30 December 1985, as validation data to select the hyperparameters. For the 95 catchments across the CAMELS basins, we trained the LSTM and the DPL models on 15 years' worth of data from 1 October 1980 to 30 September 1995, and evaluated its performance on another 15 years' worth of data from 1 October 1995 to 30 September 2010. The 5 years' worth of data from 1 October 1980 to 30

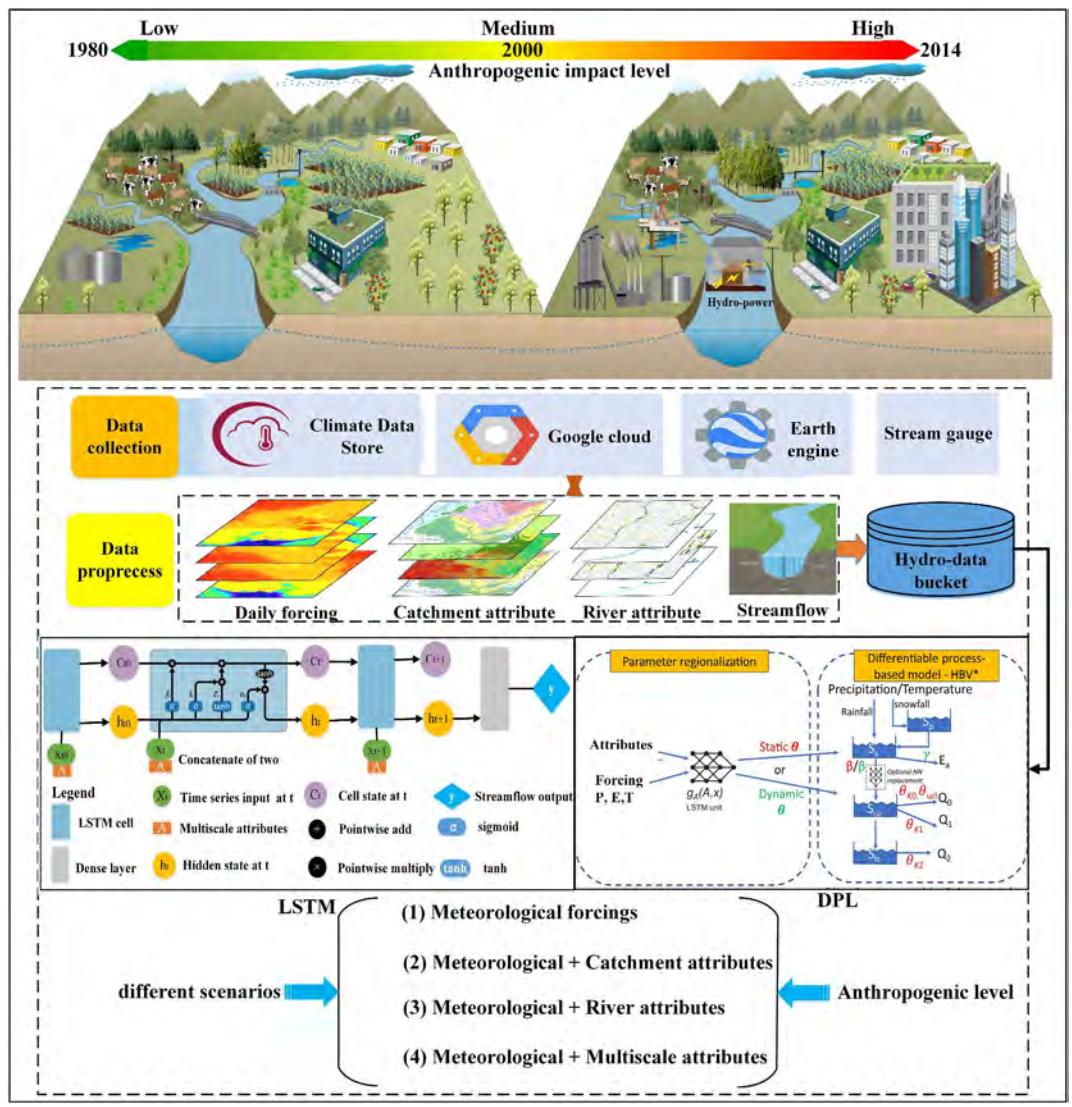

Figure 2. Sketch of the long short-term memory and Differentiable Parameter Learning models. The two models were trained with different strategies to consider anthropogenic impact levels.

Table 4
List of Differentiable Process-Based Model ($\delta_n(\gamma_r, \beta_t)$) Hyperparameters

No	Model parameters	Value
1	Warm-up option	1 year
2	Parameters	Dynamic parameters
3	Hidden state	128
4	Number of multi-component model	16
5	Number of HBV parameters	13
6	Learning rate	0.005
7	Batch size	512
8	Optimizer	Adam
9	Number of training epochs	100
10	Sequence length	365
11	Loss function	RMSE

September 1985, were used as validation data to select the hyperparameters. Furthermore, for the LSTM model, we trained the models from 1 October 1980 to 30 September 1995, and from 1 October 1995 to 30 September 2010. It is essential to emphasize that due to the data gaps from 1998 to 2002 across the Yellow River Basin, the training and validation periods differ significantly for the Yellow River Basin from the CAMELS basins.

We conducted multiple experiments for both DPL and LSTM, optimizing their respective hyperparameters to achieve optimal performance. The multiscale attributes for the catchment and river reach scale attributes may have different impacts on the performance of DLP and LSTM. Thus, we designed four different scenarios: (a) Base, in which the LSTM model did not use any catchment or river attributes as input; (b) Catchment attribute, in which both the LSTM and DPL models were fed with catchment attributes; (c) River attribute, in which both models were fed with river attributes; and (d) Multiscale attribute, in which all of the catchment and river attributes were used as static attributes in the LSTM and DPL models. The four scenarios can differentiate the contributions of each type of attribute to the LSTM and DPL model performance.

2.4.2. Setup With Varying Data Length Under Shifting Anthropogenic Influence

Our study considered the temporal evolution of human impacts within the Yellow River Basin. This basin underwent a shift from relatively modest anthropogenic activities before 2000 to a substantial intensification of human influences post-2000. In light of this temporal division, we carefully divided our data set into two distinct periods: a low anthropogenic impact period, representing the conditions before 2000, and a high anthropogenic impact period, characterizing the years following 2000. For each of these periods, we trained the model using different lengths of observed data, ranging from 1 to 10 years, and validated model performance using two years of data and evaluated the performance on another 2 years' worth of data. The KGE results from this experiment are presented for the test period.

Our main goal was to understand the influence of the shifting human dynamics on the performance of the LSTM and DPL models. To evaluate this influence, we executed a series of experiments with systematic variations in the amount of the source training data. Subsequently, we applied both LSTM and DPL models to each of these discrete temporal segments, allowing us to examine their performance under different anthropogenic conditions. Furthermore, we systematically explored the impact of varying data lengths on the model proficiency by manipulating the duration of the training data. These comprehensive experiments were designed to provide insights into how the diversity, quantity, and temporal distribution of training data influence the effectiveness of these models.

2.4.3. Evaluate Model Performance Across Changing Anthropogenic Impact Periods

In this part, we conducted experiments meticulously tailored to thoroughly assess the performance of the two models across the two distinct temporal periods as outlined previously. Initially, we trained both LSTM and DPL using varying lengths of the training data, specifically 3, 5, 7, and 10 years of data, during the period of low anthropogenic impact. Subsequently, we tested the performance of the models from 2001 to 2014, representing the high anthropogenic impact period. Additionally, we conducted an inverse experiment wherein the models were trained using data from the high anthropogenic period, spanning 3, 5, 7, and 10 years, and then evaluated their performance on the low anthropogenic period using data from 1990 to 1996. The outcomes of these experiments not only facilitated a thorough evaluation of model performance but also highlighted the adaptability and versatility of the models, demonstrating their potential utility in scenarios characterized by evolving human impacts.

2.5. Evaluation Metrics

In this study, the simulation accuracy of each model is evaluated by statistical error measurements and discharge process error, that is, the KGE criterion (Gupta et al., 2009). The KGE is commonly used to verify the goodness of the hydrological model prediction results. It is calculated as follows:

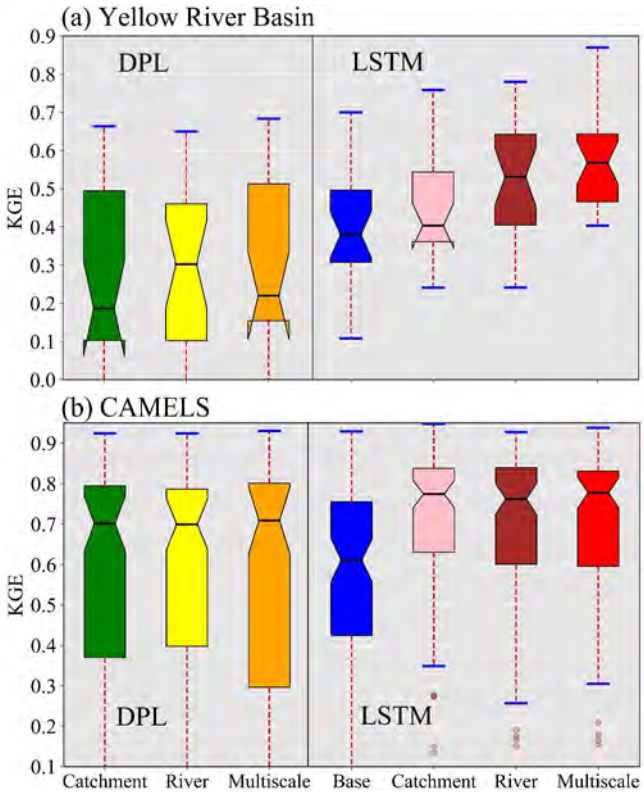

Figure 3. Performance of Differentiable Parameter Learning (left) and long short-term memory (right) with different static attributes in the testing period (2010–2014): (a) for the Yellow River Basin and (b) for CAMELS basins in USA.

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (a-1)^2 + (b-1)^2} \quad (1)$$

where r is the correlation coefficient between the observed and model-simulated streamflow, a is the ratio of the standard deviation of the model-simulated flows to the standard deviation of the observed flows, and b is the ratio of the mean of the simulated flows to the mean of the observed flows.

3. Results

3.1. Impacts of Multiscale Attributes on Streamflow Prediction

Figure 3 presents the overall results for the LSTM and DPL models fed with different selections of static attributes. For the DPL model, the median KGE values regarding the scenarios with the catchment, river, and multiscale attributes are 0.19, 0.30, and 0.22, respectively across the Yellow River Basin. The performance is substantially improved in the CAMELS basins as the median KGE over 0.70. For the LSTM model, the median KGE values of LSTM regarding the scenarios are larger than 0.4 across the Yellow River Basin, and it also shows better performance in the CAMELS basins as the KGEs over 0.78. It is important to emphasize that even without any static attributes, the base LSTM already obtained similar results with the DPL model across the Yellow River Basin, although it is not comparable to the DPL model that are fed with different selection of attributes for the CAMELS basins. Relative to the base scenarios, for the CAMELS basins, the LSTM presents the KGE values increase nearly by 0.15 by adding different attributes selections. Moreover, both LSTM and DPL models with river attributes could nearly reaches the performance of those with classical catchment attributes and multiscale attributes for the CAMELS basins, which means both DPL and LSTM models learn the anthropogenic signatures by using river attributes. Considering the impacts of river attributes, both the LSTM and DPL models fed with river attributes perform better than those fed with catchment attributes across the Yellow River Basin. In contrast, there is no significant improvements for both the LSTM and DPL models with multiscale attributes relative to the catchment attributes for the CAMELS basins.

In Figure 4, we present the spatial distribution of the differences (ΔKGE) between the LSTM and DPL models across the Yellow River Basin and the CAMELS basins. The overall improvement with the Multiscale LSTM is more significant, and distinct spatial patterns can be observed. We calculated Δ_{mean}KGE which is the mean difference between the comparison model (i.e., Multiscale LSTM) and the reference model (i.e., Multiscale DPL). The differences in KGE for the models are as follows: $\Delta_{\mathrm{mean}}\mathrm{KGE} = 0.29$, $\Delta_{\mathrm{median}}\mathrm{KGE} = 0.30$ across the Yellow River Basin, and $\Delta_{\mathrm{mean}}\mathrm{KGE} = 0.27$, $\Delta_{\mathrm{median}}\mathrm{KGE} = 0.08$ for the CAMELS basins. Both the mean and median differences indicate that the LSTM model provides more reliable streamflow simulations than the DPL model. Spatially, the DPL model struggled to provide acceptable predictions in the middle and lower reaches of the Yellow River Basin. It has poor performance in most of the lower reaches, with ΔKGE ranging from 0.5 to 0.9, indicating that their KGE values were less than zero. There are eight catchments in the lower reaches with ΔKGE larger than 0.25. For the CAMELS basins, the LSTM model performs better in 63 out of the 95 catchments than the DPL models. The DPL model has the advantage in the Great Plains catchments, while the LSTM model may degrade in the arid catchments.

Figure 5 represents the explicit performance of the Multiscale LSTM and Multiscale DPL models across the 24 catchments in the Yellow River Basin. The Multiscale LSTM shows KGE values above zero at all 24 gauges, while the Multiscale DPL presents KGE values above zero at 22 gauges. A further examination of the number of cases of good performance (KGE > 0.5) reveals that 19 out of 24 catchments have KGE > 0.5 for the Multiscale LSTM, while 6 catchments show KGE > 0.5 for the DPL model. Additionally, we observe that in the source region catchment (Catchment ID 1), the DPL model nearly reaches the performance of LSTM. However, in some catchments with drought and low annual streamflow (Catchment ID 15), the DPL model outperforms LSTM. This

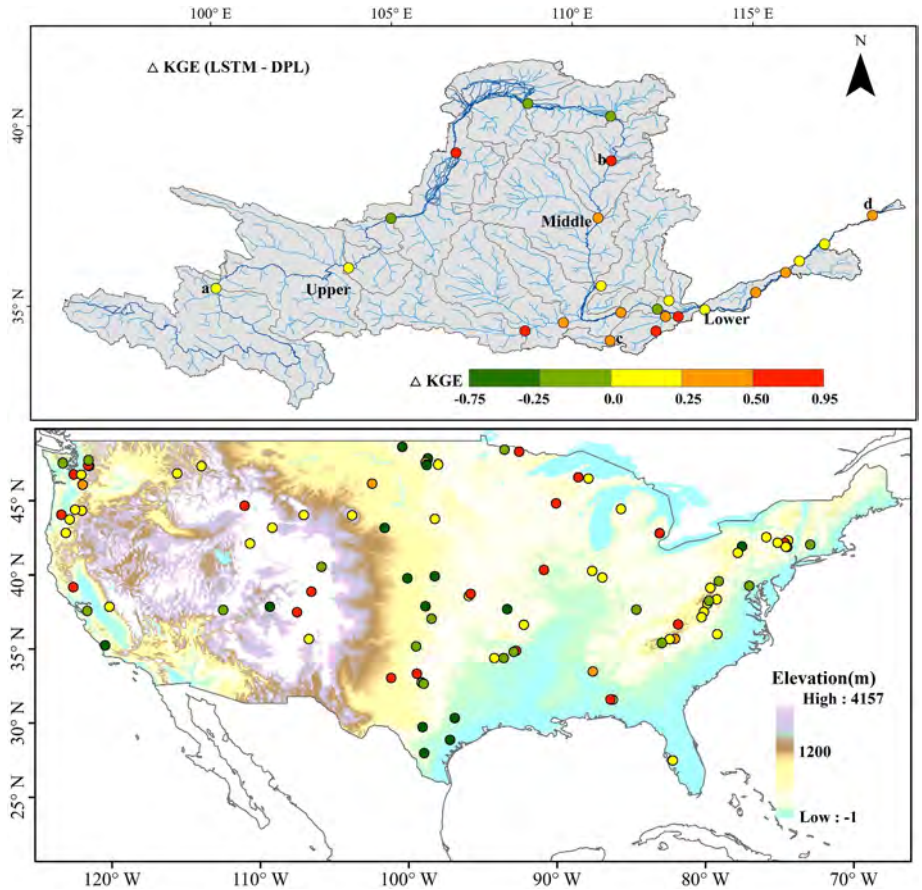

Figure 4. Spatial distribution of the difference in Kling-Gupta Efficiency (KGE) between the long short-term memory and Differentiable Parameter Learning (DPL) models across Yellow River Basin (above) and USA (below). The KGE difference is computed as $\Delta KGE = KG_{ELSTM} - KG_{EDPL}$.

is expected, because catchments with relatively dry and low average flows may result in worse performance for the LSTM model. Overall, the results clearly indicate that the LSTM model demonstrates noticeably better performance than the DPL for most of the catchments. This provides further proof that the addition of catchments with different anthropogenic and hydrological signatures endows the model with the ability to establish relationships that the DPL cannot describe across large and data-sparse conditions.

We present four hydrographs for the Yellow River Basin to illustrate the performance of the Multiscale LSTM and the Multiscale DPL models (Figure 6). The LSTM model accurately simulated both low and high flows for catchment ID 1 in the source region (Figure 6a). However, the DPL model underestimated peak flow, even though it was able to represent flow duration curves in the source region, which has low anthropogenic impacts (number of dams = 2). For the large catchment ID 7 in the middle reach, the LSTM and DPL models achieved KGE values of 0.42 and −0.32, respectively (Figure 6b). Notably, DPL struggled to provide good simulations with an increase in human interventions (number of dams = 46).

Additionally, DPL could not reach the observed magnitudes of peak flow and baseflow, while the LSTM model moderately mitigated this problem. For catchment ID 11, which experiences low annual flow and drought conditions, both the LSTM and DPL models nearly represent high flow, but the DPL model significantly underestimated the baseflow (Figure 6c). For Catchment ID 24, the largest in the Yellow River Basin (area = 766,138 km^2, number of dams = 152), is challenging for the DPL because it was unable to accurately represent the streamflow process. By contrast, the LSTM model improved both peak and flow durations in this extremely large catchment (Figure 6d).

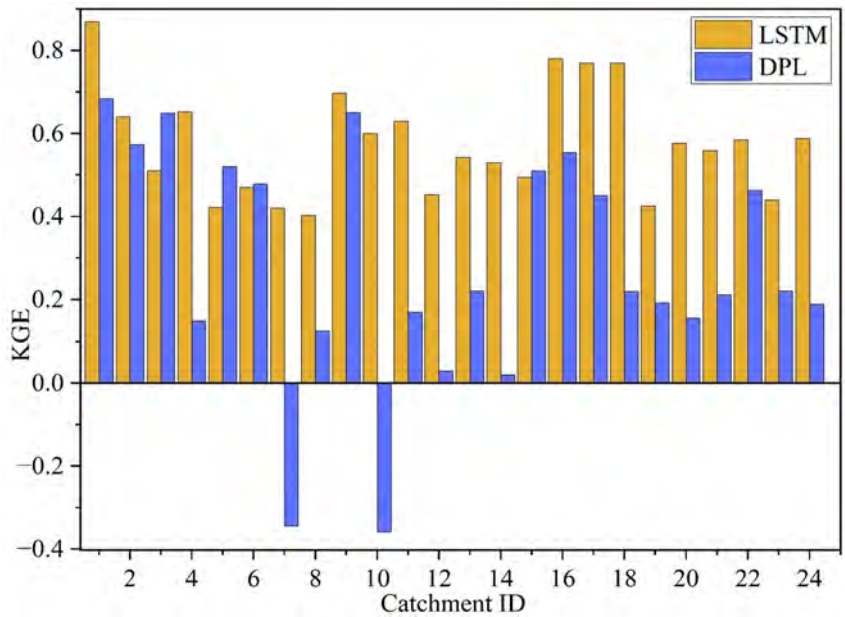

Figure 5. Performance of the Multiscale long short-term memory and the Multiscale Differentiable Parameter Learning models in the estimation period for the Yellow River Basin.

3.2. Impact of Data Length Under Different Human Interventions

The results depicted in Figure 3 illustrate promising performance of both LSTM and DPL models in the CAMELS basins; however, their effectiveness appears to diminish in the Yellow River Basin. This challenge can be attributed to the substantial portion (>80%) of the Yellow River Basin being situated in dryland areas with significant human activities. Consequently, in this subsection and subsequent sections, we shift our focus to the Yellow River Basin rather than the CAMELS data set to examine the influence of human intervention on the performance of the two models.

Figure 7 shows the annual flow trends during different human-impact periods (before and after 2000). The results of this trend analysis unveiled distinctive spatio-temporal variations in streamflow patterns during the two different periods. Notably, discernible decrease in streamflow was observed from 1980 to 2000. These trends were predominantly witnessed in the middle and lower reaches of the Yellow River, except two gauges. However, a contrasting pattern with consistent increases was observed in streamflow from 2001 to 2014. Therefore, the following analysis describes the results for the second experiment, where we explore how the time-series length (1–10 years of observed data) in the training data set affects the simulation quality of LSTM and DPL during different human impact periods (before and after 2000).

Figure 8 shows the results obtained for the effects of varying training period lengths in different human-regulated periods. During the low-human-impact period, we observed improved model performance with increasing amount of training data. Initially, the model's performance demonstrated a stable upward trend, with KGE reaching its optimal value of 0.63 when 9 years of data were used for training. As the training period extended to 3 years, the LSTM achieved a good level of accuracy (KGE > 0.3). Remarkably, as the training length expanded from 4 to 10 years, all models consistently exhibited KGE values above 0.4. These simulation results exhibited a positive correlation with the time-series length of the input data. However, even for an extended training period, certain catchments still exhibited poor performance (Figure 8a). Transitioning to the high-level human-impacts period, the performance of the LSTM model also tended to improve with the increasing training period, albeit with some differences from the low-human-impacts period (Figure 8b). For instance, when the training period length was 4 years, the model displayed superior performance compared to the models trained for 5, 6, and 7 years. Most of the models demonstrated KGE values within the range of 0.35–0.5 and did not exhibit a smooth, consistent improvement with longer training periods. Additionally, the model performance did not surpass a KGE

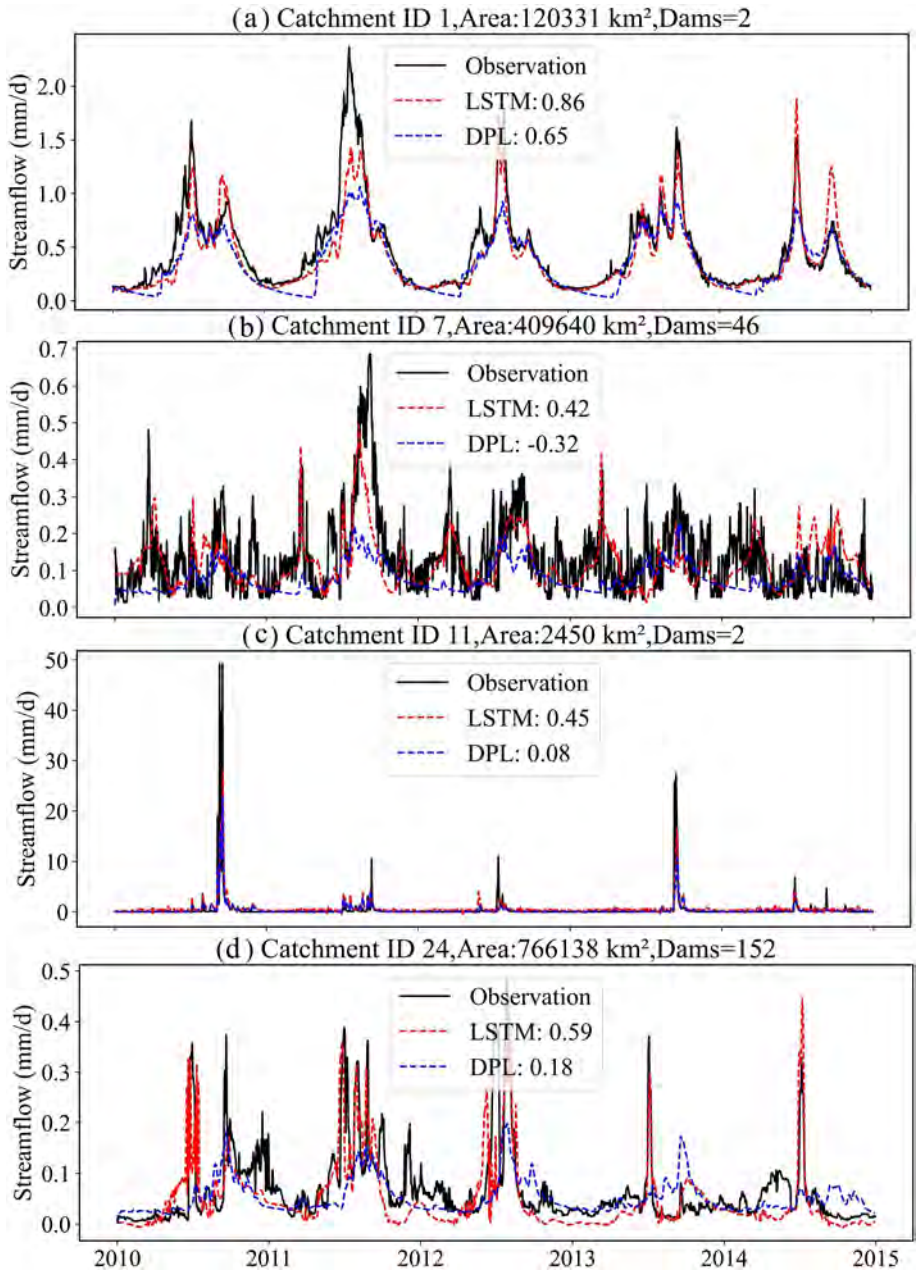

Figure 6. Time series comparisons for four catchments from the Yellow River Basin with Kling-Gupta Efficiency (KGE) values—Multiscale long short-term memory versus Multiscale Differentiable Parameter Learning. The numbers in the legends show the KGE metric for the whole testing period (2010–2014), and Dams denotes the number of water dams in the specific catchments.

value of 0.6 across various training period lengths. In summary, the model performance for the low-human-impact period is not as good as for the high-human-impact period.

During the low-human-impact period, the DPL model exhibited a trend of increasing performance with longer training period. However, this upward trend was not as pronounced, and the KGE values remained within the range of 0.2–0.4 (Figure 8c). Furthermore, most of the cases during this period showed KGE values lower than 0.4, with the exception of the model trained for 10 years, which managed to achieve a KGE value higher than 0.4. In the high-human impact period, the behavior of the DPL model displayed significant variation. There was no discernible trend of significant increase in performance, and the model did not exhibit improvement with increasing

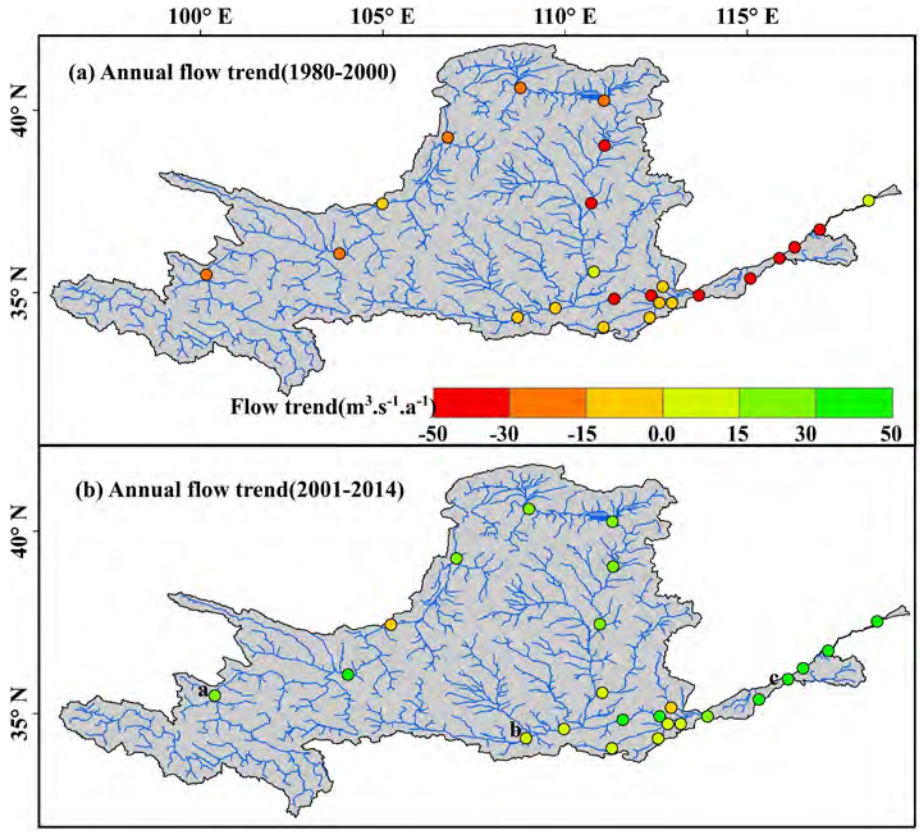

Figure 7. Annual streamflow trends between two different periods across the Yellow River Basin.

training data length (Figure 8d). Across all models trained with varying data lengths, the KGE values remained below 0.4. These observations strongly indicate that the DPL model struggled to achieve reliable performance during the high-human-impact period, in stark contrast to its behavior in the low-human-impact period.

We further investigated the model performance differences (ΔKGE = KGE (LSTM) − KGE (DPL)) with varying data lengths for both the LSTM and DPL models with the results depicted in Figure 9. For the low-human-impact period, when the training data length was less than 5 years, there was no substantial distinction between the LSTM and DPL models, with both achieving a ΔKGE of approximately 0.12. However, when the input data length ranged from six to 10 years, the LSTM model displayed significant improvement attributable to the augmented training data length. Consequently, a noticeable disparity emerged between the two models, with ΔKGE hovering at approximately 0.21 (Figure 9a). In the high-human-impact period, a substantial divergence was observed between the LSTM and DPL models. As the human impact increased, it had a pronounced impact on the performance of the DPL model, leading to five cases with ΔKGE values exceeding 0.2 (Figure 9b). This finding underscores the sensitivity of DPL to human impacts, and LSTM model demonstrating greater resilience and performance improvement in response to longer training data even during high-human-impact periods.

3.3. Understanding the Anthropogenic Similarities

Figure 10 shows how varying degrees of human activity impact on model performance. One scenario involved training the LSTM and DPL models in a low impact period (1980–1996) and then assessing their performance in a period with high human impacts (2001–2014). Notably, when the training data were limited to 7 years, the model struggled to produce accurate simulations during the high-human-impact period. In contrast, when the LSTM model was trained in the high-impact period, it exhibited superior performance when tested in the low-impact period (Figure 10a).

The results obtained by the DPL model showed a different pattern. The DPL model trained during the low-impact period did not perform well when tested in the high-impact period and its performance showed no significant

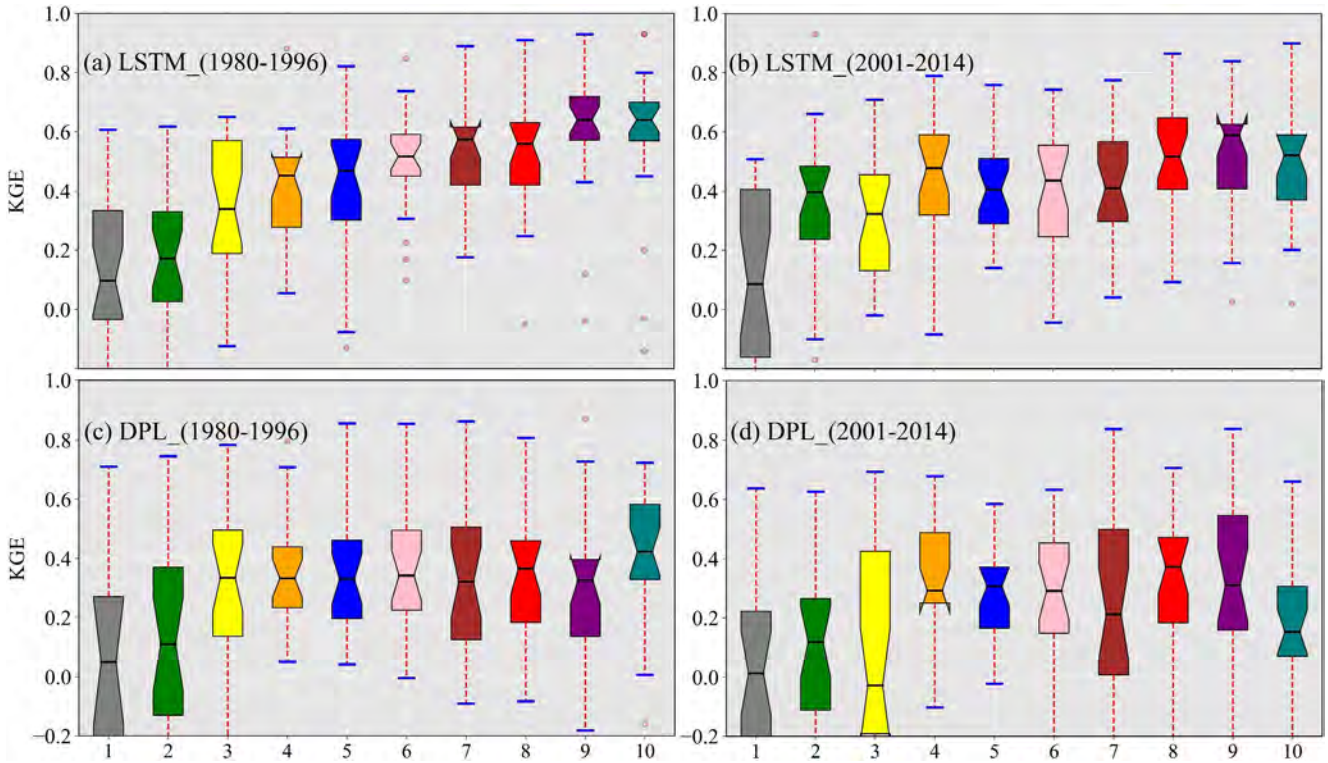

Figure 8. Effect of time-series length (1–10 years of observed data) on Kling-Gupta Efficiency. (a) Low human impact period (1980–1996) for the long short-term memory (LSTM) model; (b) high-human-impact period (2001–2014) for the LSTM model; (c) low-human-impact period (1980–1996) for the Differentiable Parameter Learning (DPL) model; and (d) high-human-impact period (2001–2014) for the DPL model.

correlation with the length of the training data. Furthermore, the DPL model trained during the high-human-impact period still did not provide better simulation results when tested in the low-human-impact period (Figure 10b). Generally, as the volume of the training data increased, the LSTM model consistently demonstrated better simulation accuracy than the DPL model. This trend remained stable even during periods of high human impacts. These findings suggest that when sufficient training data are available, LSTM has the potential to outperform the DPL model, particularly in human-regulated catchments, while the performance of the DPL model is significantly degraded by complex human activities.

We further identify the contribution of the river and catchment attributes by evaluating their impacts on the model's performance across the Yellow River Basin, and we calculated the Spearman's rank correlation of KGE (from the experiments with multiscale attributes described in Section 3.1) and each static attribute. As shown in Figure 11, for the LSTM model, static attributes such as sediment trapping (SED, −0.60), catchment area (area_gauge, 0.50), degree of regulation (DOR, −0.49), degree of water use (USE, −0.44), number of dams in catchments (Num_dam, −0.43), and all anthropogenic attributes exhibit negative correlations with the model's KGE scores. Conversely, the river connectivity status index (CSI, 0.57), mean river width from the SWORD (Width, 0.41) displays a positive correlation with the model's KGE scores (Figure 11a). In contrast, the correlation pattern for the DPL model differs notably from that of the LSTM model. The DPL model's performance (KGE) shows a significant correlation with classical hydrological signatures, including the base flow index (0.53), fraction of snow cover (0.41), low flow frequency (−0.48), slope of the flow duration curve (−0.47), mean water surface elevation (WSE, 0.27), and mean river width (Width, 0.22) from the SWORD data (Figure 11b). In summary, the Multiscale LSTM model highlights the capability of DL models to flexibly incorporate meteorological, hydrological, and anthropogenic features.

Figure 12 presents the KGE value of the LSTM and DPL models associated with the relationship between the number of dams within dammed catchments. A discernible trend emerges from this analysis: in catchments with a small number of dams, the LSTM model shows satisfactory performance with the KGE values higher than 0.7, and the DPL models also perform relatively well. However, as the number of dams increases, the river's natural

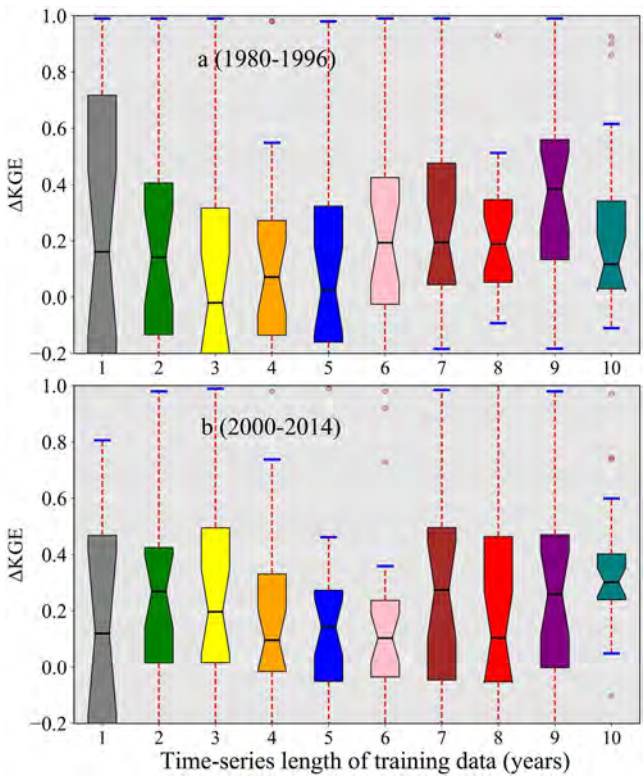

Figure 9. Kling-Gupta Efficiency (KGE) difference between the long short-term memory (LSTM) and Differentiable Parameter Learning (DPL) models, △KGE = KGE (LSTM) − KGE (DPL). (a) For the low-human-impact period and (b) for the high-human-impact period.

connectivity diminishes, and the pressure on the river segments intensifies. This leads to a decrease in the KGE scores for the DPL model, reflecting the growing impact of river damming on model performance. Interestingly, the Multiscale LSTM model demonstrates unique robustness with respect to increasing dam numbers. Even as the number of dams approaches approximately 120, it manages to sustain a relatively consistent level of simulation accuracy, with KGE values exceeding 0.58. This notable performance could be attributed to the inclusion of anthropogenic similarities at river-reach scales. The improvements observed in the Multiscale LSTM model's performance highlight its potential in addressing this challenge. Conversely, as the number of dams exceeds 40, the DPL model's performance tends to decline. Beyond this threshold, there is no significant difference in model performance with an increasing number of dams, with KGE values dropping below 0.3. This decline underscores the limitations of the DPL model in effectively simulating streamflow in human-regulated catchments. Overall, this analysis underscores the robustness of the Multiscale LSTM model with respect to increased damming.

4. Discussion

4.1. Implications for Anthropogenic Signatures at the River-Reach Scale

In regional hydrological modeling, catchment attributes are important in training, and regional models can properly represent the flow dynamics in multiple catchments (Gao et al., 2022; Ibrahim Demir et al., 2022). However, some studies have indicated that LSTM showed poor performance in human-regulated catchments than in natural catchments (Jin et al., 2022; Manh-Hung Le, 2022; Sadler et al., 2022). This is most likely because the streamflow dynamics in human-regulated catchments significantly differ from those in natural catchments. The flow in natural catchments is characterized by significant seasonal patterns, and there exists a critical hydrological cycle in the watershed. By contrast, hydrological processes are significantly altered in human-regulated catchments due to various human activities (Marvel et al., 2019; Yu et al., 2023). Many hydrological signatures are not well represented in catchment attributes, reducing their benefit in achieving favorable model performance in human-regulated catchments.

In this study, we first attempted to use river-reach attributes that sufficiently include river characteristics, such as the river CSI and degree of human regulation (DOR), along with some satellite attributes like mean river width (Width), slope, mean WSE, and maximum width from the SWORD data (Altenau et al., 2021). We demonstrated how to collect river-reach attributes with catchment attributes to form a new type of attribute collection known as Multiscale attributes. We meticulously validated the effectiveness and reliability of these multiscale attributes across both the CAMELS data set and the Yellow River Basin, particularly within human-regulated catchments. The results obtained in this study suggest that anthropogenic similarities at the river-reach scale can be learned by LSTM and DPL models, and that the learned hydrological signatures can represent a wide variety of aspects of irregular streamflow dynamics. This implies that in human-regulated catchments, including catchment attributes alone is insufficient for interpreting complicated hydrological patterns, and users should consider diverse hydrological signatures as an alternative.

4.2. Exploring the Potential of Multiscale LSTM and DPL in Human-Regulated Catchments

In this study, we aim to evaluate the performance of LSTM and DPL models across human-regulated catchments and examine how these models can leverage specific static river attributes at the river-reach scale. We seek to explore the potential of these models by incorporating multiscale anthropogenic signatures. As the hydrology community makes more extensive use of LSTM and DPL models with classical catchment attributes across natural catchments, these models' performance in human-regulated environments remains uncertain (Anderson & Radić, 2022; Bian et al., 2023; Yao et al., 2023). We try to explore the opportunity to merge "data-driven

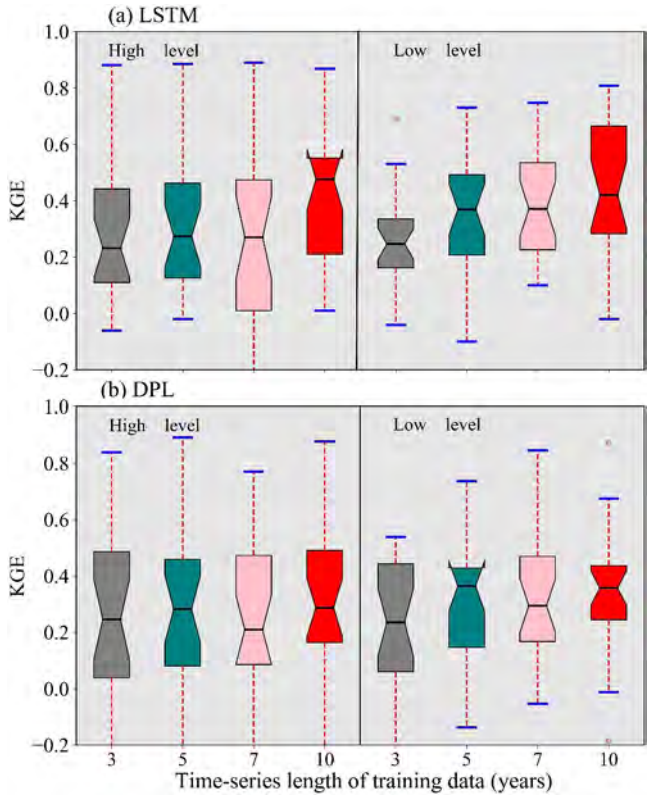

Figure 10. Model performance of training and testing in different human impact periods. High level: trained the models during the low-impact period and subsequently test them to the high-impact period; Low level: trained the models during the high-impact period and subsequently tested them to the low-impact period.

discovery" with "knowledge-driven regularization" offering a promising avenue to streamflow simulation across human-regulated catchments.

We observed that both LSTM and DPL show promising capabilities in capturing long-term hydrological patterns and seamlessly incorporate anthropogenic signatures across the CAMAELS catchments with relatively lower river connectivity status (CSI < 95%). Despite this, LSTM demonstrates robustness in simulating complex hydrological patterns even in catchments with extremely low river connectivity status index (CSI < 50%), establishing a notable correlation between model performance and river attributes (Figure 10a). However, the DPL model faces challenges in accurately representing irregular streamflow patterns observed in catchments affected by human regulation, especially those with multiple reservoirs. It's important to note that the DPL model is a physics-informed machine learning model with regionalized deep-network-based parameterization pipelines (Feng et al., 2022; Feng, Beck, et al., 2023; Feng, Tan, et al., 2023). However, it still encounters difficulties in accurately tracking and adapting to these intricate alterations across the Yellow River Basin. Interestingly, the DPL model with river attributes outperforms both the catchment and multiscale attributes models. This suggests the robustness of the Differentiable, Learnable mechanism of the DPL model in learning anthropogenic signatures.

4.3. Challenges and Future Work

We meticulously analyzed uncertainties based on the this study and literature across human-regulated catchments (Figure 13). The quality of input data directly affects the model's performance (Aerts et al., 2022; Gauch et al., 2021). Noisy, missing, or insufficient data may hinder the model's ability to capture complex hydrological processes and dynamics (Fang et al., 2022; Feng et al., 2021). Choosing the right model architecture and optimization algorithm is paramount for different hydrological simulation tasks (Li et al., 2023; Qiu et al., 2021). Human activities significantly influence the hydrological characteristics and dynamics of the watershed, thereby increasing the uncertainty of runoff simulations (Tursun et al., 2024a, 2024b). Climate change amplify the uncertainty of hydrological simulations as models need to predict under uncertain future climate conditions (Althoff et al., 2021; Michel et al., 2022; Moosavi et al., 2022). One of the key findings from this study is that the training data period emerges as the most dominant factor affecting model performance relative to other model training options. Notably, selecting training data from both low- and high-human-impact periods led to a substantial enhancement in model performance. This improvement can be attributed to the diverse anthropogenic signatures present across both low- and high-human-impact periods, as opposed to a single period, thereby enriching the training data set and enabling the model to better adapt to varying conditions.

Our exploratory tests revealed that multiscale attributes provide invaluable information for both the LSTM and DPL models across the CAMELS and Yellow River Basin. However, in this study, our primary aim was to discover which anthropogenic signatures the model would learn, and interpret them in terms of their relationships to model performance. Consequently, there were several model training strategies that we did not fully explore, such as evaluating model performance in ungauged basins or regions. Furthermore, while we focus on human-regulated catchments, there is ample opportunity to delve deeper into combining both natural and human-regulated catchments to further explore the robustness of the DPL and LSTM models. By incorporating a broader range of catchment types, we may gain a more comprehensive understanding of how these models perform under varying environmental conditions and anthropogenic influences.

We believe that dynamic attributes at both river and catchment scales are crucial for accurate streamflow simulation, as static attributes alone may not adequately reflect changing hydrological conditions. Therefore, in future studies, it will be imperative to develop methods for integrating multiscale and dynamic attributes into regional hydrological models that are flexible enough to accommodate both human-regulated and natural

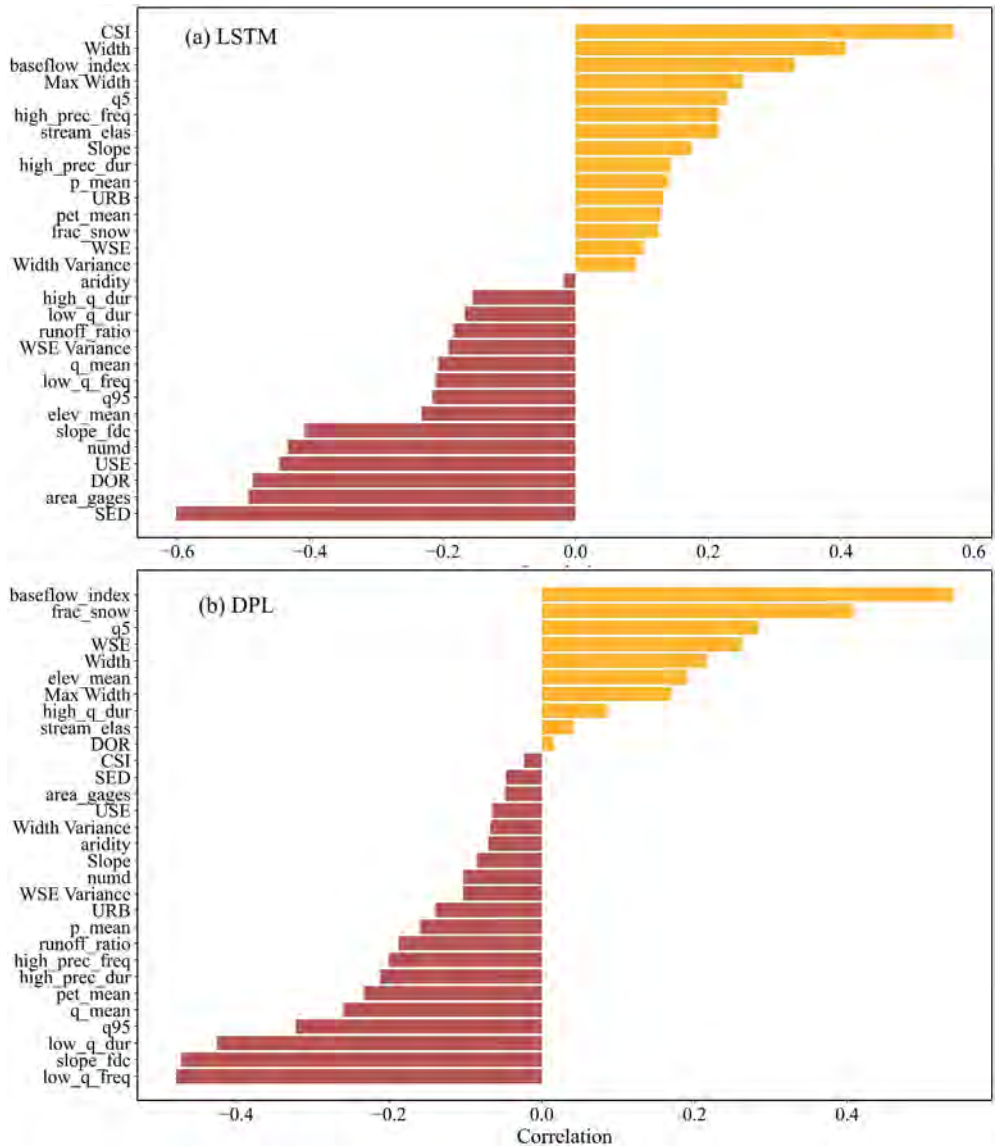

Figure 11. Spearman's rank correlation between the multiscale attributes (rows) and the associated Kling-Gupta Efficiency scores (columns) from the experiments described in Section 3.1 from the Yellow River Basin. The positive correlations are in yellow, and the negative correlations are in red.

catchments across different hydrological patterns. Moreover, combining meteorological data with remote sensing observations, such as, river width, WSE, and slope from the SWOT data (Altenau et al., 2021), can lead to the creation of multimodal data sets. This integration of diverse data sources can provide valuable insights and improve the performance of DL models.

5. Conclusion

In this study, we trained the LSTM and DPL models with multiscale attributes across human-regulated catchments and comprehensively evaluated the models' capability to adapt to varying degrees of human impacts. The results provide new insights into streamflow prediction and other hydrological process simulations. The important findings are summarized as follows.

We collected river attributes such as the river connectivity status index (CSI), degree of regulation (DOR), sediment trapping (SED), as well as river width, slope, WSE from SWORD data to represent anthropogenic signatures.

10.1029/2023WR036853

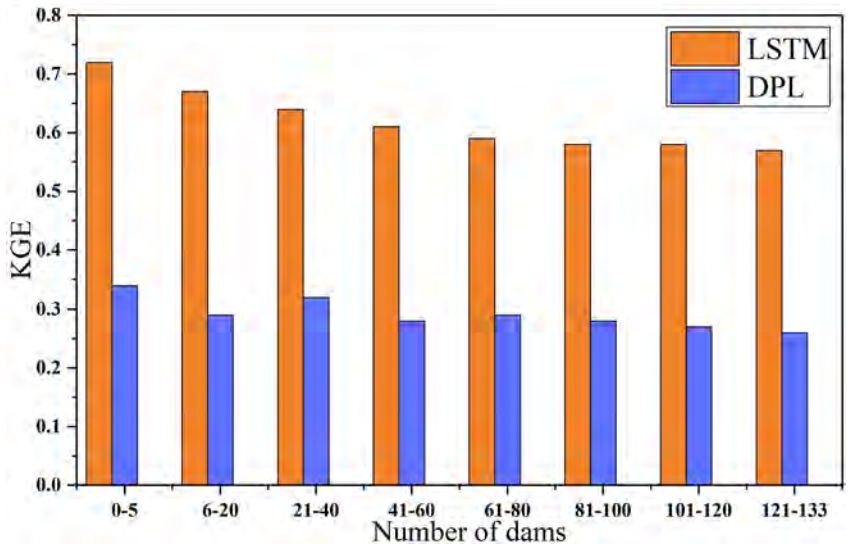

Figure 12. Kling-Gupta Efficiency distributions of the long short-term memory and Differentiable Parameter Learning models with the increase in the river dams in catchments for the Yellow River Basin.

Combining river and catchment attributes proves to be a promising approach to providing the most valuable anthropogenic similarities for the DPL and LSTM models. Even in extremely complicated hydrological conditions, such as those found in the Yellow River Basin, multiscale attributes still provide valuable signatures for the models. Moreover, the training data period emerges as the most dominant factor affecting model performance across human-regulated catchments relative to other hydrological conditions. Selecting training data from both low- and

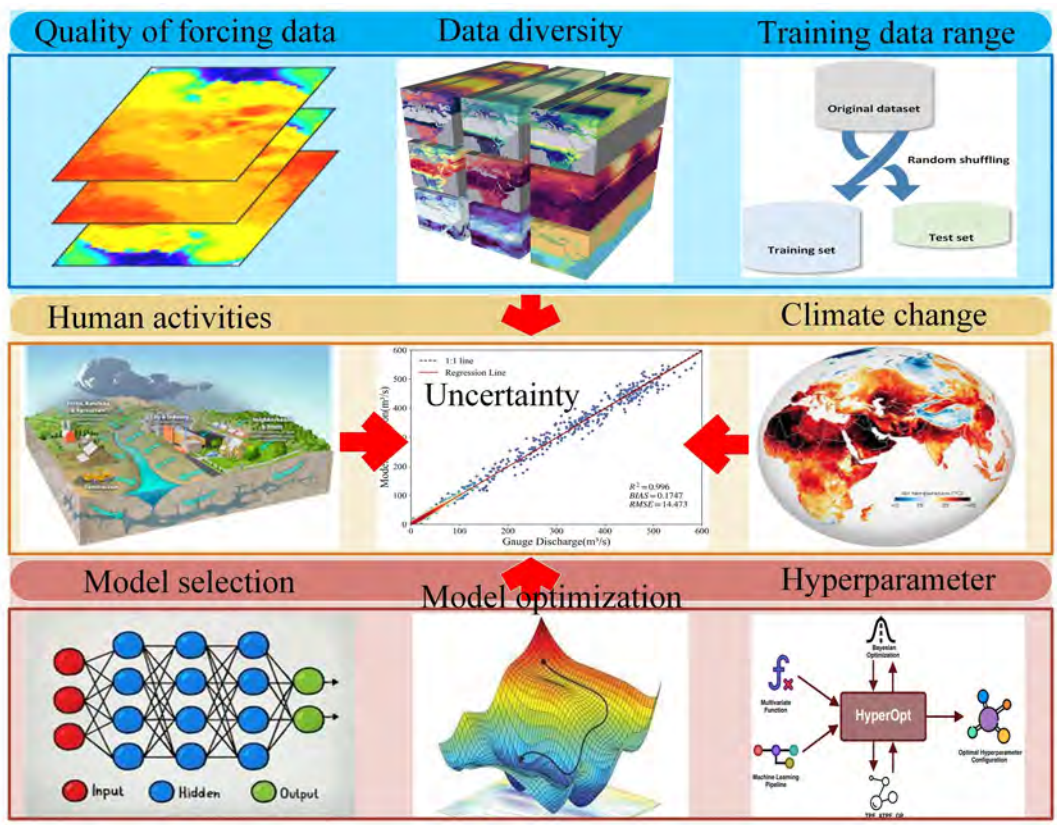

Figure 13. Uncertainty of deep learning-based streamflow modeling across the human-regulated catchments.

high-human-impact periods significantly improves model performance, as the model is extremely sensitive to the selection of the training data period. Additionally, we found that the LSTM model has great potential to adapt to varying degrees of human impacts across human-regulated catchments. Overall, the DPL model can reach nearly state-of-the-art performance across CAMELS catchments with lower human impacts, while its performance significantly deteriorates in extremely complicated hydrological conditions across the Yellow River Basin.

We found that multiscale attributes can provide diverse anthropogenic and hydrological signatures. However, it should be noted that only the appropriate selection of multiscale attributes will maximize their values, and the significance of the hydrological similarities should be highlighted in multiple catchments. We recommend using DOR, DOF, CSI, URB, WSE, width, slope, and river dams to quickly identify anthropogenic similarity. In the future, with a deep understanding of the hydrological process and the development of dynamic attributes at both river and catchment scales, the prediction accuracy of Multiscale LSTM in human-regulated catchments can be further improved.

Appendix A: Streamflow Change

In the early 2000s, significant reservoir construction and water infrastructure projects were launched across the Yellow River Basin, leading to pronounced anthropogenic interventions and impacting streamflow patterns (Figure 7). Consequently, from 1980 to 2000, there was a notable decreasing trend in annual flow, whereas from 2001 to 2014, an increasing trend in annual flow was observed (Figure A1). Therefore, it becomes imperative to

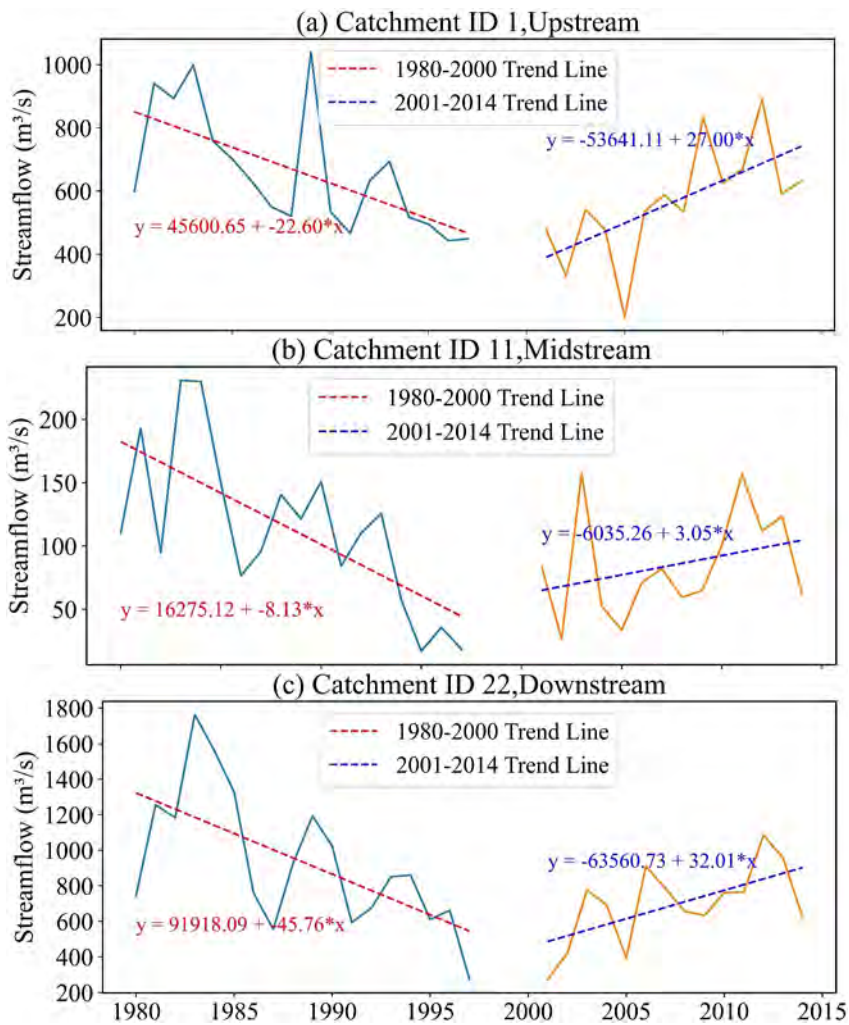

Figure A1. Time series plots of annual flow trends of three different catchments for 1980–2014 for the Yellow River Basin.

divide the time period from 1980 to 2014 into two distinct segments, with the year 2000 serving as the boundary between these periods.

Appendix B: Multiscale Attributes

In our study, we carefully derived the attributes at both catchment and river scales to enhance the capabilities of the LSTM model in representing complex hydrological processes. For catchment-level attributes, our approach involved performing a spatial join between the HydroATLAS polygons and the catchment boundaries. Subsequently, we derived catchment attributes by computing area-weighted aggregates, ensuring a comprehensive representation of catchment characteristics (Figure B1). To effectively capture the influence of human disturbances on the streamflow at the river-reach scale, we compiled a series of river-reach attributes.

It is important to note that the method employed for calculating river-reach attributes differs significantly from that used for catchment attributes. Rather than calculating an average value for all river-reaches within a catchment, we opted to consider the river-reach attributes specifically at the outlet of each catchment, as depicted in Figure B1. This approach allowed us to precisely capture the anthropogenic signatures at the river-reach scale. Notably, we focused on CSI values derived from the outlet reach of each catchment, offering a targeted representation of the connectivity status at this critical juncture.

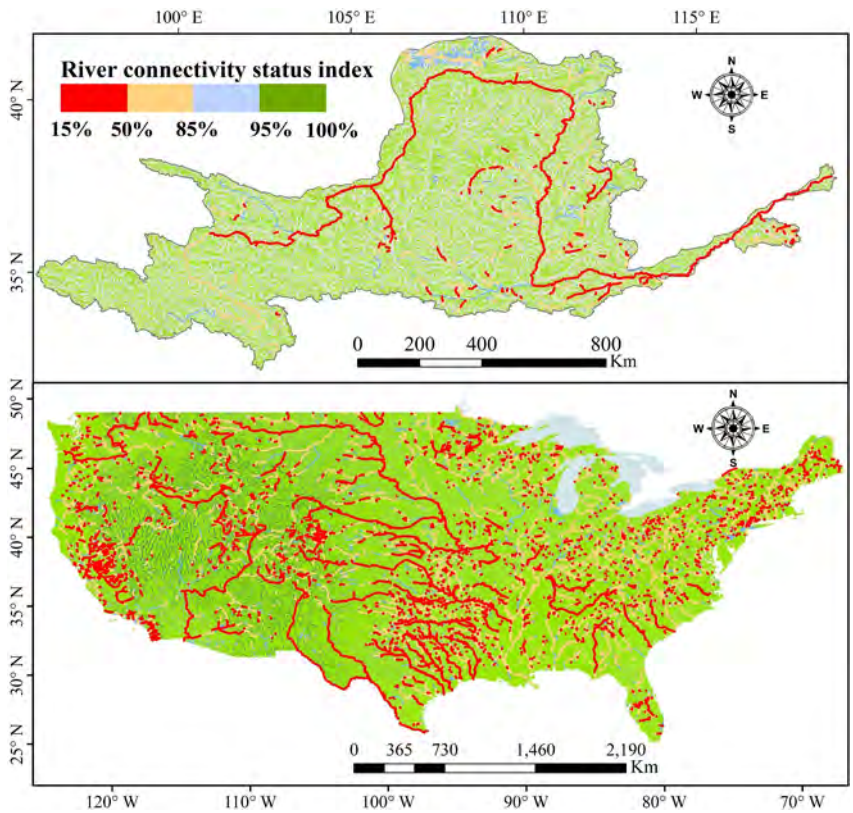

Figure B1. The connectivity status index of river reaches across the Yellow River Basin (above) and United states (below).

Data Availability Statement

The Multiscale attributes, results and codes are available at https://zenodo.org/records/11112700 (Tursun, 2024).

Acknowledgments
This study was supported by a Grant from the National Natural Science Foundation of China (Grant 42271021).
Acknowledgment for the data support from "Loess Plateau SubCenter, National Earth System Science Data Center, National Science & Technology Infrastructure of China (http://loess.geodata.cn)".

References

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

Aerts, J. P. M., Hut, R. W., van de Giesen, N. C., Drost, N., van Verseveld, W. J., Weerts, A. H., & Hazenberg, P. (2022). Large-sample assessment of varying spatial resolution on the streamflow estimates of the wflow_sbm hydrological model. *Hydrology and Earth System Sciences*, *26*(16), 4407–4430. https://doi.org/10.5194/hess-26-4407-2022

Altenau, E. H., Pavelsky, T. M., Durand, M. T., Yang, X., Frasson, R. P. D. M., & Bendezu, L. (2021). The surface water and ocean topography (SWOT) mission river database (SWORD): A global river network for satellite data products. *Water Resources Research*, *57*(7). https://doi.org/10.1029/2021wr030054

Althoff, D., Rodrigues, L. N., & Silva, D. D. D. (2021). Addressing hydrological modeling in watersheds under land cover change with deep learning. *Advances in Water Resources*, *154*, 103965. https://doi.org/10.1016/j.advwatres.2021.103965

Anderson, S., & Radić, V. (2022). Evaluation and interpretation of convolutional long short-term memory networks for regional hydrological modelling. *Hydrology and Earth System Sciences*, *26*(3), 795–825. https://doi.org/10.5194/hess-26-795-2022

Arsenault, R., Martel, J., 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

Bennett, A., & Nijssen, B. (2021). Deep learned process parameterizations provide better representations of turbulent heat fluxes in hydrologic models. *Water Resources Research*, *57*(5). https://doi.org/10.1029/2020wr029328

Bian, L., Qin, X., Zhang, C., Guo, P., & Wu, H. (2023). Application, interpretability and prediction of machine learning method combined with LSTM and LightGBM-a case study for runoff simulation in an arid area. *Journal of Hydrology*, *625*, 130091. https://doi.org/10.1016/j.jhydrol.2023.130091

Botterill, T. E., & Mcmillan, H. K. (2023). Using machine learning to identify hydrologic signatures with an encoder–decoder framework. *Water Resources Research*, *59*(3). https://doi.org/10.1029/2022wr033091

Chen, X., Wang, S., Gao, H., Huang, J., Shen, C., Li, Q., et al. (2022). Comparison of deep learning models and a typical process-based model in glacio-hydrology simulation. *Journal of Hydrology*, *615*, 128562. https://doi.org/10.1016/j.jhydrol.2022.128562

Déry, S. J., Hernández-Henríquez, M. A., Stadnyk, T. A., & Troy, T. J. (2021). Vanishing weekly hydropeaking cycles in American and Canadian rivers. *Nature Communications*, *12*(1), 7154. https://doi.org/10.1038/s41467-021-27465-4

Du, X., Silwal, G., & Faramarzi, M. (2022). Investigating the impacts of glacier melt on stream temperature in a cold-region watershed: Coupling a glacier melt model with a hydrological model. *Journal of Hydrology*, *605*, 127303. https://doi.org/10.1016/j.jhydrol.2021.127303

Durand, M., Gleason, C. J., Pavelsky, T. M., Prata De Moraes Frasson, R., Turmon, M., David, C. H., et al. (2023). A framework for estimating global river discharge from the surface water and ocean topography satellite mission. *Water Resources Research*, *59*(4). https://doi.org/10.1029/2021wr031614

Fang, K., Kifer, D., Lawson, K., Feng, D., & Shen, C. (2022). The data synergy effects of time-series deep learning models in hydrology. *Water Resources Research*, *58*(4). https://doi.org/10.1029/2021wr029583

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). 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). https://doi.org/10.1029/2022wr032404

Feng, D., Tan, Z., & He, Q. (2023). Physics-informed neural networks of the Saint-Venant equations for downscaling a large-scale river model. *Water Resources Research*, *59*(2). https://doi.org/10.1029/2022wr033168

Ficklin, D. L., Abatzoglou, J. T., Robeson, S. M., Null, S. E., & Knouft, J. H. (2018). Natural and managed watersheds show similar responses to recent climate change. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(34), 8553–8557. https://doi.org/10.1073/pnas.1801026115

Gao, S., Zhang, S., Huang, Y., Han, J., Luo, H., Zhang, Y., & Wang, G. (2022). A new seq2seq architecture for hourly runoff prediction using historical rainfall and runoff as input. *Journal of Hydrology*, *612*, 128099. https://doi.org/10.1016/j.jhydrol.2022.128099

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

Gavahi, K., Foroumandi, E., & Moradkhani, H. (2023). A deep learning-based framework for multi-source precipitation fusion. *Remote Sensing of Environment*, *295*, 113723. https://doi.org/10.1016/j.rse.2023.113723

Gholizadeh, H., Zhang, Y., Frame, J., Gu, X., & Green, C. T. (2023). Long short-term memory models to quantify long-term evolution of streamflow discharge and groundwater depth in Alabama. *Science of the Total Environment*, *901*, 165884. https://doi.org/10.1016/j.scitotenv.2023.165884

Grill, G., Lehner, B., Thieme, M., Geenen, B., Tickner, D., Antonelli, F., et al. (2019). Mapping the world's free-flowing rivers. *Nature*, *569*(7755), 215–221. https://doi.org/10.1038/s41586-019-1111-9

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), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Ha, S., Liu, D., & Mu, L. (2021). Prediction of Yangtze River streamflow based on deep learning neural network with El Niño–Southern Oscillation. *Sci Rep-Uk*, *11*(1), 11738. https://doi.org/10.1038/s41598-021-90964-3

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

Ibrahim Demir, Z. X. B. D., Xiang, Z., Demiray, B., & Sit, M. (2022). Waterbench: A large-scale benchmark dataset for data-driven streamflow forecasting. *Earth System Science Data*, *14*(12), 5605–5616. https://doi.org/10.5194/essd-2022-52

Jiang, F., Xie, X., Wang, Y., Liang, S., Zhu, B., Meng, S., et al. (2022). Vegetation greening intensified transpiration but constrained soil evaporation on the loess plateau. *Journal of Hydrology*, *614*, 128514. https://doi.org/10.1016/j.jhydrol.2022.128514

Jiang, P., Shuai, P., Sun, A., Mudunuru, M. K., & Chen, X. (2023). Knowledge-informed deep learning for hydrological model calibration: An application to coal creek watershed in Colorado. *Hydrology and Earth System Sciences*, *27*(14), 2621–2644. https://doi.org/10.5194/hess-27-2621-2023

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

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., 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., et al. (2023). Caravan - A global community dataset for large-sample hydrology. *Scientific Data*, *10*(1), 61. https://doi.org/10.1038/s41597-023-01975-w

Kumar, M. T., & Rao, M. C. (2023). Studies on predicting soil moisture levels at Andhra Loyola College, India, using SARIMA and LSTM models. *Environmental Monitoring and Assessment*, *195*(12), 1426. https://doi.org/10.1007/s10661-023-12080-1

Li, G., Zhu, H., Jian, H., Zha, W., Wang, J., Shu, Z., et al. (2023). A combined hydrodynamic model and deep learning method to predict water level in ungauged rivers. *Journal of Hydrology*, *625*, 130025. https://doi.org/10.1016/j.jhydrol.2023.130025

Li, Q., Zhu, Y., Shangguan, W., Wang, X., Li, L., & Yu, F. (2022). An attention-aware LSTM model for soil moisture and soil temperature prediction. *Geoderma*, *409*, 115651. https://doi.org/10.1016/j.geoderma.2021.115651

Liu, Y., Xie, X., Tursun, A., Wang, Y., Jiang, F., & Zheng, B. (2023). Surface water expansion due to increasing water demand on the loess plateau. *Journal of Hydrology: Regional Studies*, *49*, 101485. https://doi.org/10.1016/j.ejrh.2023.101485

Ma, K., Feng, D., Lawson, K., Tsai, W. P., Liang, C., Huang, X., et al. (2021). Transferring hydrologic data across continents – Leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. *Water Resources Research*, *57*(5). https://doi.org/10.1029/2020wr028600

Ma, Y., Montzka, C., Bayat, B., & Kollet, S. (2021). Using long short-term memory networks to connect water table depth anomalies to precipitation anomalies over Europe. *Hydrology and Earth System Sciences*, *25*(6), 3555–3575. https://doi.org/10.5194/hess-25-3555-2021

Manh-Hung Le, H. K. S. A. (2022). Streamflow estimation in ungauged regions using machine learning: Quantifying uncertainties in geographic extrapolation. *Hydrology and Earth System Sciences*. https://doi.org/10.5194/hess-2022-320

Marvel, K., Cook, B. I., Bonfils, C. J. W., Durack, P. J., Smerdon, J. E., & Williams, A. P. (2019). Twentieth-century hydroclimate changes consistent with human influence. *Nature*, *569*(7754), 59–65. https://doi.org/10.1038/s41586-019-1149-8

Michel, A., Schaefli, B., Wever, N., Zekollari, H., Lehning, M., & Huwald, H. (2022). Future water temperature of rivers in Switzerland under climate change investigated with physics-based models. *Hydrology and Earth System Sciences*, *26*(4), 1063–1087. https://doi.org/10.5194/hess-26-1063-2022

Moosavi, V., Gheisoori Fard, Z., & Vafakhah, M. (2022). Which one is more important in daily runoff forecasting using data driven models: Input data, model type, preprocessing or data length? *Journal of Hydrology*, *606*, 127429. https://doi.org/10.1016/j.jhydrol.2022.127429

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., et al. (2021). Era5-land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, *13*(9), 4349–4383. https://doi.org/10.5194/essd-13-4349-2021

Nearing, G. S., Klotz, D., Frame, J. M., Gauch, M., Gilon, O., Kratzert, F., et al. (2022). Technical note: Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks. *Hydrology and Earth System Sciences*, *26*(21), 5493–5513. https://doi.org/10.5194/hess-26-5493-2022

Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., et al. (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

Ng, K. W., Huang, Y. F., Koo, C. H., Chong, K. L., El-Shafie, A., & Najah Ahmed, A. (2023). A review of hybrid deep learning applications for streamflow forecasting. *Journal of Hydrology*, *625*, 130141. https://doi.org/10.1016/j.jhydrol.2023.130141

Ni, Y., Yu, Z., Lv, X., Qin, T., Yan, D., Zhang, Q., & Ma, L. (2022). Spatial difference analysis of the runoff evolution attribution in the Yellow River Basin. *Journal of Hydrology (Amsterdam)*, *612*, 128149. https://doi.org/10.1016/j.jhydrol.2022.128149

Ouyang, W., Lawson, K., Feng, D., Ye, L., Zhang, C., & Shen, C. (2021). Continental-scale streamflow modeling of basins with reservoirs: Towards a coherent deep-learning-based strategy. *Journal of Hydrology*, *599*, 126455. https://doi.org/10.1016/j.jhydrol.2021.126455

Qiu, R., Wang, Y., Rhoads, B., Wang, D., Qiu, W., Tao, Y., & Wu, J. (2021). River water temperature forecasting using a deep learning method. *Journal of Hydrology*, *595*, 126016. https://doi.org/10.1016/j.jhydrol.2021.126016

Riggs, R. M., Allen, G. H., David, C. H., Lin, P., Pan, M., Yang, X., & Gleason, C. (2022). Rodeo: An algorithm and google earth engine application for river discharge retrieval from landsat. *Environmental Modelling & Software*, *148*, 105254. https://doi.org/10.1016/j.envsoft.2021.105254

Sadler, J. M., Appling, A. P., Read, J. S., Oliver, S. K., Jia, X., Zwart, J. A., & Kumar, V. (2022). Multi-task deep learning of daily streamflow and water temperature. *Water Resources Research*, *58*(4). https://doi.org/10.1029/2021wr030138

Shen, C., Chen, X., & Laloy, E. (2021). Editorial: Broadening the use of machine learning in hydrology. *Frontiers in Water*, *3*. https://doi.org/10.3389/frwa.2021.681023

Singh, N. K., & Basu, N. B. (2022). The human factor in seasonal streamflows across natural and managed watersheds of north America. *Nature Sustainability*, *5*(5), 397–405. https://doi.org/10.1038/s41893-022-00848-1

Thomas Lees, S. R. F. K. (2021). Hydrological concept formation inside long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*. https://doi.org/10.5194/hess-2021-566

Tian, H., Wang, P., Tansey, K., Zhang, J., Zhang, S., & Li, H. (2021). An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong Plain, PR China. *Agricultural and Forest Meteorology*, *310*, 108629. https://doi.org/10.1016/j.agrformet.2021.108629

Tiggeloven, T., Couasnon, A., van Straaten, C., Muis, S., & Ward, P. J. (2021). Exploring deep learning capabilities for surge predictions in coastal areas. *Sci Rep-Uk*, *11*(1), 17224. https://doi.org/10.1038/s41598-021-96674-0

Tu, T., Comte, L., & Ruhi, A. (2023). The color of environmental noise in river networks. *Nature Communications*, *14*(1), 1728. https://doi.org/10.1038/s41467-023-37062-2

Tursun, A. (2024). Streamflow prediction in human-regulated catchments using multiscale LSTM modeling with anthropogenic similarities. *Zenodo*. https://doi.org/10.5281/zenodo.11112700

Tursun, A., Xie, X., Wang, Y., Liu, Y., Peng, D., Rusuli, Y., & Zheng, B. (2024a). Reconstruction of missing streamflow series in human-regulated catchments using a data integration LSTM model. *Journal of Hydrology: Regional Studies*, *52*, 101744. https://doi.org/10.1016/j.ejrh.2024.101744

Tursun, A., Xie, X., Wang, Y., Liu, Y., Peng, D., & Zheng, B. (2024b). Enhancing streamflow simulation in large and human-regulated basins: Long short-term memory with multiscale attributes. *Journal of Hydrology*, *630*, 130771. https://doi.org/10.1016/j.jhydrol.2024.130771

Wang, J., Walter, B. A., Yao, F., Song, C., Ding, M., Maroof, A. S., et al. (2022). Geodar: Georeferenced global dams and reservoirs dataset for bridging attributes and geolocations. *Earth System Science Data*, *14*(4), 1869–1899. https://doi.org/10.5194/essd-14-1869-2022

Wang, W., Zhang, Y., Geng, X., & Tang, Q. (2021). Impact classification of future land use and climate changes on flow regimes in the yellow river source region, China. *Journal of Geophysical Research: Atmospheres*, *126*(13). https://doi.org/10.1029/2020jd034064

Wei, X., Wang, G., Schmalz, B., Hagan, D. F. T., & Duan, Z. (2023). Evaluation of transformer model and self-attention mechanism in the Yangtze River basin runoff prediction. *Journal of Hydrology: Regional Studies*, *47*, 101438. https://doi.org/10.1016/j.ejrh.2023.101438

Wi, S., & Steinschneider, S. (2022). Assessing the physical realism of deep learning hydrologic model projections under climate change. *Water Resources Research*, *58*(9). https://doi.org/10.1029/2022wr032123

Wright, L. G., Onodera, T., Stein, M. M., Wang, T., Schachter, D. T., Hu, Z., & Mcmahon, P. L. (2022). Deep physical neural networks trained with backpropagation. *Nature*, *601*(7894), 549–555. https://doi.org/10.1038/s41586-021-04223-6

Xu, J., Ma, Z., Yan, S., & Peng, J. (2022). Do era5 and era5-land precipitation estimates outperform satellite-based precipitation products? A comprehensive comparison between state-of-the-art model-based and satellite-based precipitation products over mainland China. *Journal of Hydrology*, *605*, 127353. https://doi.org/10.1016/j.jhydrol.2021.127353

Xu, M., Wang, G., Wang, Z., Hu, H., Kumar Singh, D., & Tian, S. (2022). Temporal and spatial hydrological variations of the yellow river in the past 60 years. *Journal of Hydrology*, *609*, 127750. https://doi.org/10.1016/j.jhydrol.2022.127750

Yao, Z., Wang, Z., Wang, D., Wu, J., & Chen, L. (2023). An ensemble CNN-LSTM and GRU adaptive weighting model based improved sparrow search algorithm for predicting runoff using historical meteorological and runoff data as input. *Journal of Hydrology*, *625*, 129977. https://doi.org/10.1016/j.jhydrol.2023.129977

Yousefi, A., & Toffolon, M. (2022). Critical factors for the use of machine learning to predict lake surface water temperature. *Journal of Hydrology*, *606*, 127418. https://doi.org/10.1016/j.jhydrol.2021.127418

Yu, S., Li, W., Zhou, L., Yu, X., Zhang, Q., & Shen, Z. (2023). Human disturbances dominated the unprecedentedly high frequency of yellow river flood over the last millennium. *Science Advances*, *9*(8), eadf8576. https://doi.org/10.1126/sciadv.adf8576

Zhang, Y., Ragettli, S., Molnar, P., Fink, O., & Peleg, N. (2022). Generalization of an encoder-decoder LSTM model for flood prediction in ungauged catchments. *Journal of Hydrology*, *614*, 128577. https://doi.org/10.1016/j.jhydrol.2022.128577

Zhang, Y., Zheng, H., Zhang, X., Leung, L. R., Liu, C., Zheng, C., et al. (2023). Future global streamflow declines are probably more severe than previously estimated. *Nature Water*, *1*(3), 261–271. https://doi.org/10.1038/s44221-023-00030-7