



Research papers

Enhancing streamflow simulation in large and human-regulated basins: Long short-term memory with multiscale attributes

Arken Tursun^{a,b}, Xianhong Xie^{a,b,*}, Yibing Wang^{a,b}, Yao Liu^{a,b}, Dawei Peng^{a,b}, Buyun Zheng^{a,b}

^a State Key Laboratory of Remote Sensing Science, Beijing Normal University, Beijing 100875, China

^b Beijing Engineering Research Center for Global Land Remote Sensing Products, Institute of Remote Sensing Science and Engineering, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China

ARTICLE INFO

Keywords:

LSTM
LISFLOOD
Multiscale static attributes
Streamflow simulation
Human activities
Yellow River Basin

ABSTRACT

Streamflow simulation in human-regulated catchments is a great challenge for both process-based hydrological models and deep learning (DL) methods, mainly because human-regulation rules are difficult to parameterize in these models. In this study, we investigate the roles of river and catchment attributes in DL for streamflow prediction. We evaluate a typical DL method, i.e., long short-term memory (LSTM), and evaluate its performance in 25 large catchments across the Yellow River Basin where human activities are intensive, especially with large numbers of dams and reservoirs influencing streamflow processes. For the LSTM forcing data, we compare two forcing datasets: the Fifth Generation of European Reanalysis (ERA5-Land) and meteorological station-based data. The results show that the LSTM forced by ERA5-Land achieves improved performance, as its mean Kling–Gupta efficiency (KGE) is 0.21 relative to the mean KGE of 0.08 from the meteorological station forced LSTM. Integrating different types of hydrological attributes (catchment and river characteristics) can substantially improve LSTM performance even for catchments with dams and reservoirs. The river-reach attributes show the largest contribution to the LSTM model improvement. Moreover, LSTM with multiscale attributes outperforms a global process-based hydrological model (LISFLOOD) in the middle and lower reaches of the Yellow River Basin. Our study indicates that multiscale attributes are promising pivots for DL methods to improve streamflow prediction in human-regulated basins.

1. Introduction

Streamflow prediction is essential in quantifying the water cycle and for other purposes, including drought monitoring and flood forecasting, especially in human-regulated catchments (Shen, 2018a; Tan and Gan, 2016). It may also provide credible guidelines to manage water resources under the global climate change context (Cho and Kim, 2022; Xie et al., 2022). However, due to the irregular behaviour and dynamic nature of streamflow and spatiotemporal variabilities in precipitation, glacier melt, snowmelt and human activities, streamflow prediction in large and human-regulated catchments remains challenging (Feng et al., 2021; Kratzert et al., 2021).

In past decades, process-based hydrological models have been widely applied to predict streamflow and other related hydrological variables (e.g., snow water equivalent, soil moisture, evapotranspiration) (Elsaadani et al., 2021; Li et al., 2022). The physical states in

process-based hydrological models are critical to interpret hydrological processes. For example, in glacier melt-dominated catchments, temperature is important for seasonal floods; thus, physical models could interpret the relationship among glaciers, temperature, and streamflow (Xie et al., 2022). When calibrating streamflow in process-based models, it is crucial to ensure that other physical states and fluxes are constrained via their physical relations (Li et al., 2022). However, the dreaded problems of parameter equifinality and non-uniqueness may cause large predictive uncertainty (Fan et al., 2019). Despite substantial progress in the parametrization of human activities, process-based models have shown limitations in capturing the nonlinear characteristics in streamflow, and it is still difficult to represent the hydrological processes across human-regulated basins (Coron et al., 2017; Ficchi et al., 2019).

Data-driven methods include simple regressions to complicated neural networks with deep neural layers and thousands of parameters

* Corresponding author at: Beijing Normal University, China.

E-mail address: xianhong@bnu.edu.cn (X. Xie).

<https://doi.org/10.1016/j.jhydrol.2024.130771>

Received 25 May 2023; Received in revised form 28 December 2023; Accepted 29 December 2023

Available online 26 January 2024

0022-1694/© 2024 Elsevier B.V. All rights reserved.

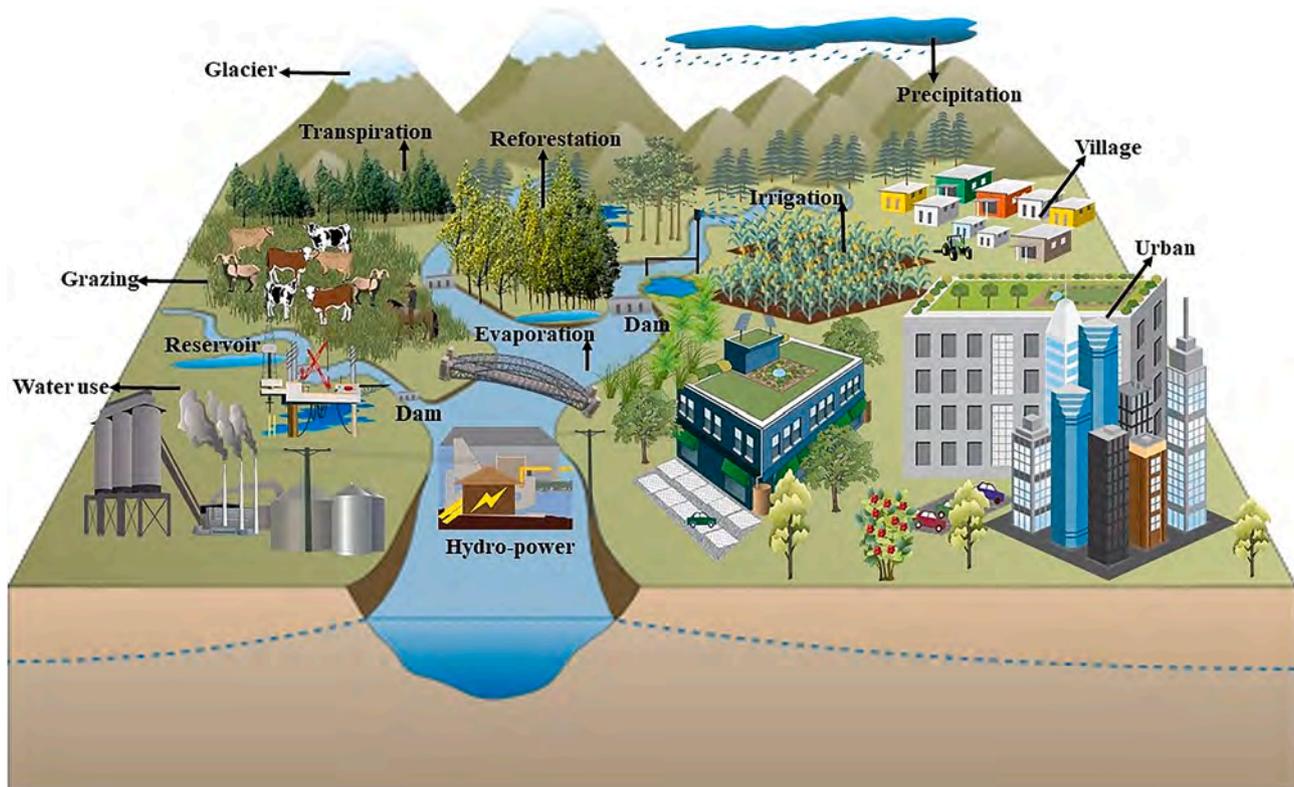


Fig. 1. Diagram of the Yellow River Basin hydrological cycle in the Anthropocene. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(Elsaadani et al., 2021; Li et al., 2022; Li et al., 2022). Deep learning (DL) is a data-driven method and has also been widely used in hydrology (Bennett and Nijssen, 2021; Shen et al., 2021; Yousefi and Toffolon, 2022). Previous studies have shown that a DL model can precisely interpret the patterns in the input–output relationship without explicit knowledge of the physical processes (Moishin et al., 2021; Shen, 2018b; Wright et al., 2022). Long short-term memory (LSTM), for example, is a typical DL method that is able to learn from long time series data (Shen, 2018b), and it is well suited to representing hydrological dynamics forced by different scale memory effects in a large watershed (Feng et al., 2020; Xie et al., 2022). LSTM models trained across many watersheds have been shown to outperform traditional hydrological models and predict well in ungauged basins (PUB) (Arsenault et al., 2023). For example, Kratzert et al. (2019) reported a 531-catchment streamflow prediction study with mean NSE (Nash–Sutcliffe efficiency) values of 0.69 for the LSTM model, 0.64 for the SAC-SMA conceptual hydrological model, and 0.58 for the National Water Model. Furthermore, LSTM has also been used to predict other hydrological variables, such as soil moisture, groundwater and water quality (Li et al., 2022; Yin et al., 2021).

Despite extensive applications, LSTM still needs to be examined in extremely complicated hydrological conditions with sparse datasets. A few studies have used LSTM models to predict time series of streamflow in dammed watersheds (Althoff et al., 2021). For example, Ouyang et al. (2021) evaluated LSTM in 3557 small basins over the contiguous United States (CONUS) and found that LSTM was good in basins with low human activities, but it may render relatively poor performance in the central plains with significant human interventions (e.g., dams and reservoir regulations). Moreover, for human-regulated catchments, Ren et al. (2022) improved the model performance considering the time lag of streamflow for LSTM model training. It is still unclear how to improve LSTM for streamflow prediction in extremely large and human-regulated catchments (Wi and Steinschneider, 2022).

DL models applied in hydrological prediction are generally fed with

forcing data that were produced from meteorological stations (Moishin et al., 2021; Yousefi and Toffolon, 2022). However, the representation of meteorological forcing across large catchments using meteorological station data can be challenging due to the uneven distribution of meteorological conditions (Gauch et al., 2021). To remedy this issue, reanalysis meteorological products may be an option for streamflow prediction (Nevo et al., 2022). For example, Lees et al. (2021) successfully estimated the streamflow and soil moisture using LSTM driven by a global dataset from the land component of the Fifth Generation of European Reanalysis (ERA5-Land) across 669 basins in Great Britain. While favourable results were found for small catchments, it was generally recognized that more intensive evaluations are needed regarding reanalysis data in DL models (Gauch et al., 2021; Kratzert et al., 2021).

Multiscale catchment and river attributes are important in driving hydrological processes in large and human-regulated basins. Catchment static attributes are generally assumed to describe physiographic properties (Feng et al., 2021; Ouyang et al., 2021). However, we contend that using natural catchment attributes alone are insufficient for explaining the intricate hydrological conditions in human-regulated catchments. For example, the Yellow River Basin encompasses numerous reservoirs, dams, and a substantial expanse of cultivated land (Fig. 1). Water withdrawal for crop irrigation, reservoir expansion and large-scale tree restoration projects could directly alter the patterns of streamflow (Jin et al., 2022a; Ni et al., 2022). These human activities pose tough challenges to streamflow prediction. Thus, it is necessary to simulate streamflow by considering different types of static attributes (Hall and Perdigão, 2021; Jin et al., 2022b). For example, the river connectivity status index (CSI), degree of regulation (DOR) and degree of fragmentation (DOF) are influential factors in runoff routing and flow concentration (Grill et al., 2019).

In this study, we examine the LSTM model for streamflow simulation in relatively large and human-regulated basins. The primary objectives are (1) to evaluate whether driving inputs from ERA5-land effectively

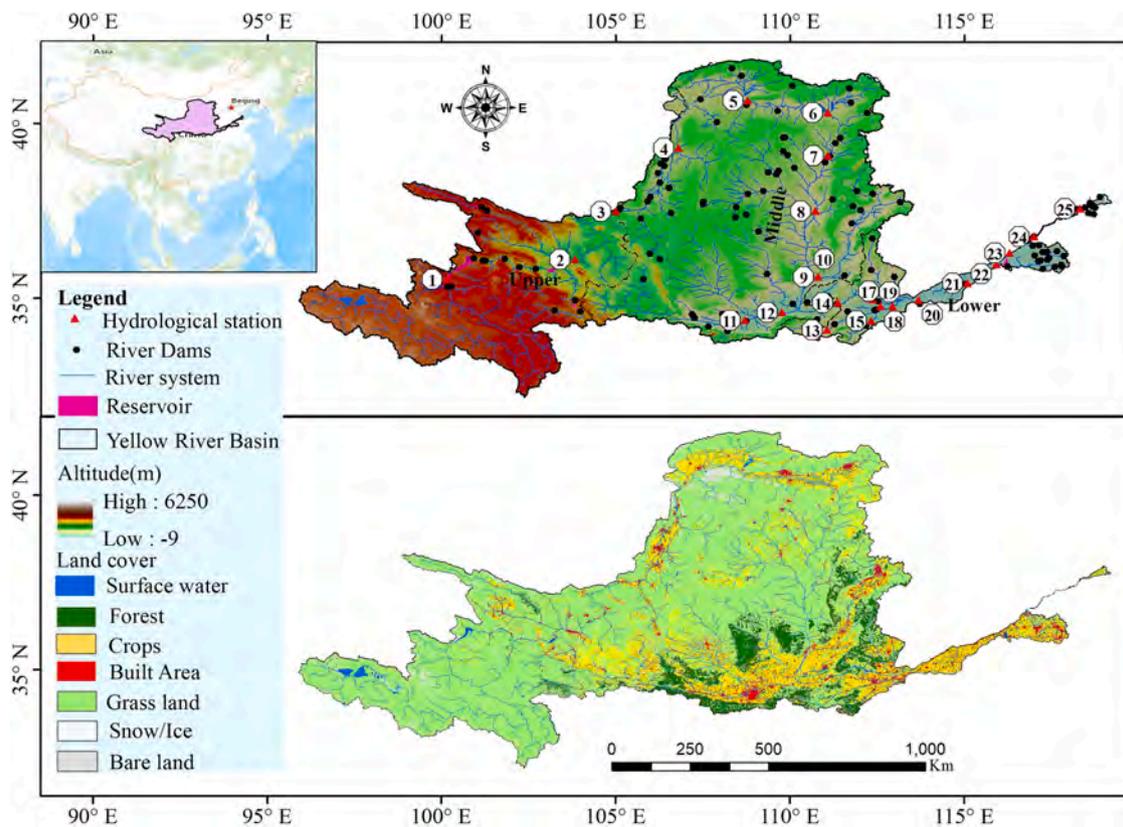


Fig. 2. The location of the stream gauges (above) and land cover map (below).

represent spatial conditions and provide credible streamflow simulations at strongly human-influenced basins; (2) to validate our new hypothesis that multiscale attributes may contain more useful information and should help to distinguish different streamflow behaviours compared to only basin-averaged attributes; and (3) to benchmark the LSTM model with multiscale attributes against a process-based model (LISFLOOD) to assess the ability of the LSTM model to capture the influence of human activities. The multiscale attributes in this study refer to catchment and river attributes, including natural surface conditions and human-regulated signatures.

2. Study area and datasets

2.1. Study area

This study was conducted across the Yellow River Basin in northern China, which is the sixth longest river system worldwide, covering an area of approximately 750,000 km² and supporting the livelihood for 190 million people. The Yellow River originates on the Qinghai-Tibet Plateau region at an elevation of approximately 6,000 m, flows through the Ordos and Loess Plateau and empties 5,450 km downstream into the Bohai Sea (Fig. 2). The average annual precipitation for the Yellow River Basin is approximately 480 mm, although it is unevenly distributed in space and time. Generally, rainfall decreases from southeast to northwest and is strongly seasonal, accumulating mainly between June and September (Liu et al., 2020).

In the past several decades, the demand for water resources has soared dramatically to satisfy the necessities of the growing population. The increasing demand for water resources has led to the construction of reservoirs, dams and canals to store water from the Yellow River (Xu et al., 2022). Indeed, hydrological processes in watersheds were strongly impacted by human activities, such as large-scale water withdrawal for crop irrigation, reservoir expansion and large-scale ecological

restoration projects (Jiang et al., 2022; Xie et al., 2015). Therefore, these anthropogenic interventions probably impact daily streamflow trends across the Yellow River Basin (Jin et al., 2022a).

2.2. Streamflow data

We obtained daily streamflow data for 25 sub-basins in the Yellow River Basin, ranging from 2,450 to 747,000 km², with a median area of 330,000 km², as shown in Fig. 2. The 25 stations were distributed along the main river trajectory. The daily data were for the period of 1981 to 2019, and they were obtained from the Loess Plateau Subcenter, National Earth System Science Data Center, National Science & Technology Infrastructure of China (<https://loess.geodata.cn>). There are missing data from 1998 to 2002, and we did not reconstruct the missing data to avoid uncertainties in reconstruction.

2.3. Meteorological data

Meteorological forcing data used in this study included precipitation, temperature, potential evaporation and wind speed. These data were obtained from ERA5-Land, which produces primary land surface components by the ERA5 atmospheric model (Muñoz-Sabater et al., 2021; Yilmaz, 2023). As a reanalysis dataset, ERA5-land provides a consistent view of the evolution of land variables from 1950 to present, and it renders an accurate description of the climate of the past (Kratzert et al., 2023). The ERA5-Land products have 11 km spatial resolution and hourly temporal frequency (Pelosi et al., 2020; Zhang et al., 2021). Based on the hourly datasets, we calculated daily average forcing variables to predict daily streamflow. We adjusted the time zone of the datasets (GMT + 0) to make a local prediction (GMT + 8). Moreover, we calculated the area-weighted spatial average for each meteorological variable in each catchment.

Table 1
Multiscale attributes used to train the LSTM.

Static variables	Static variable description	Median	Range
Catchment attributes			
area	Catchment area (km ²)	333,755	[2450,746957]
p_mean	Mean daily precipitation (mm d ⁻¹)	1.9	[0.3,2.7]
ari_ix_sav	Global aridity index (%)	58.6	[37,75]
pnv_pc_s08	Potential extent of mixed forest (%)	0.54	[0,81]
pet_mean	Mean daily potential evaporation (mm d ⁻¹)	3.7	[2.3,4.3]
high_prec_dur	Average duration of high precipitation events (days)	1.3	[1.2,1.4]
low_prec_freq	Frequency of low precipitation days	0.57	[0.50,0.77]
swc_pc_s01	Soil water content (%)	58	[30,73]
glc_pc_s11	Forest cover extent (%)	1.45	[0.01,12.6]
cly_pc_sav	Clay fraction in soil (%)	18 %	[10,22]
crp_pc_	Crop land extent (%)	12 %	[0,28]
River attributes			
CSI	River connectivity status index (%)	22	[20,88]
DOR	Degree of regulation (%)	100	[20,100]
DOF	Degree of fragmentation (%)	96	[0,100]
URB	Night light intensity in urban areas (%)	15	[0,100]
USE	Water use for irrigation, industry, municipal (%)	33	[0.3,57]
Num_dams	Number of dams	35	[2,133]

2.4. Catchment and river attributes

2.4.1. Catchment attributes

Catchment attribute datasets include soil, vegetation, land cover and topographic attributes. They were collected from HydroATLAS and daily ERA5-Land time series for global scale catchments (Linke et al., 2019). We computed the spatial join of the HydroATLAS polygons and the catchment boundaries and then derived the catchment attributes as an area-weighted aggregate (Kratzert et al., 2023). Catchment attribute datasets can be grouped into the following categories: hydrology, climatology, physiography, land cover characteristics, soils and geology.

2.4.2. Dam and reservoir datasets

We collected dam and reservoir data from georeferenced global dam and reservoir datasets for bridging attributes and geolocations (GeoDAR) (Wang et al., 2022). These datasets were created by utilizing the Google Maps geocoding application programming interface (API) and multisource inventories. GeoDAR can provide one of the most abundant and spatially resolved global inventories of dams and reservoirs. We calculated the numbers of reservoirs and river dams in every single catchment and added these data into static attributes to improve LSTM-based streamflow modelling.

2.4.3. River connectivity and pressure indicator at the river-reach scale

The Yellow River Basin is host to several hundred significant dams and reservoirs. The downstream effects of these structures, encompassing fragmentation and flow regulation, substantially contribute to the loss of connectivity in the reaches of the Yellow River (Xu et al., 2022). Thus, it is necessary to add river static attributes on the river reach scale, which include a series of information that is able to discriminate among different catchments.

The river-reach attributes utilized in this study originated from the Global Free Flowing River (FFR) dataset, as detailed by Grill et al. (2019). The FFR dataset defines free-flowing rivers based on an extensive literature review and identifies five primary pressure factors that influence river connectivity. These pressure indicators include: (1) Degree of River Fragmentation (DOF), which characterizes the extent and magnitude of reduced longitudinal connectivity in river channels due to

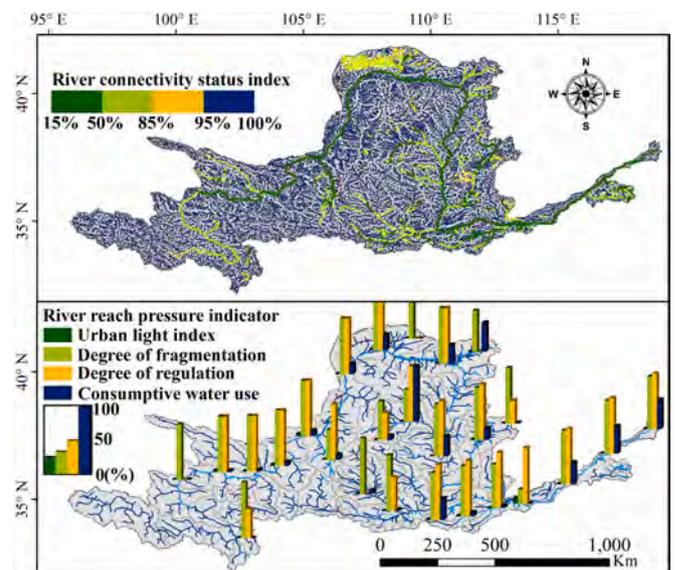


Fig. 3. Connectivity status index (CSI) of river reaches (above) and dominant pressure indicator include Degree of River Fragmentation (DOF), Degree of Regulation (DOR), Consumptive Water Use (USE) and Urban Areas (URB) for the Yellow River reach (below). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

anthropogenic barriers; (2) Degree of Regulation (DOR), which quantifies how the storage of water in dams can alter the natural flow regime downstream, expressed as the percentage of river flow volume that can be withheld in a dam’s reservoir; (3) Sediment Trapping Index (SED), a key driver for morpho-dynamic processes in rivers, shaping the physical template for fluvial ecosystems; (4) Consumptive Water Use (USE), measuring water consumption for purposes such as irrigation, industry, municipal use, and water transfer to other river systems, impacting both lateral and vertical connectivity; (5) Road Density (RDD), serving as a proxy for lateral disconnection from floodplains and the loss of longitudinal connectivity, particularly at intersections with streams and culverts; and (6) Urban Areas (URB), which affect river connectivity by reducing floodplain access due to paving and urban infrastructure, as well as through artificial channelization or levee construction that confines riverbeds or affects meandering. To quantify each pressure factor, six proxy indicators were calculated primarily using global remote sensing data and numerical model outputs, including discharge simulations. Then a weighting model was employed to combine the six pressure indicators. The weights were then applied in a multi-criteria average calculation to derive the river connectivity status index (CSI) for every river-reach. For our analysis, river reaches with a CSI of $\geq 95\%$ were considered to have good connectivity status, while those below 95% were classified as impacted.

To represent the human disturbance at river-reach scale on streamflow, we obtained series of river-reach scale attributes, including the values of DOF, DOR, SED, USE, RDD, URB, and CSI (Table 1). These attributes represent pressure factors from human interferences of river connectivity (Fig. 3). It is important to note that the calculation of river-reach attributes significantly differs from that of catchment attributes. River-reach attributes exclusively consider values at the outlet of each catchment. In this way we deliberately represented the anthropogenic signatures at river-reach scale (Fig. A1. in Appendix A). The geometric dataset of the global river attribute information for every river reach, the values of all pressure indicators (DOF, DOR, SED, USE, RDD and URB)—as well as values for the CSI, are available at <https://doi.org/10.6084/m9.figshare.7688801>.

In this study, we used static attributes from the river reach scale and catchment scale, which we called multiscale static attributes (Table 1). These attributes were chosen to represent the hydrological patterns that

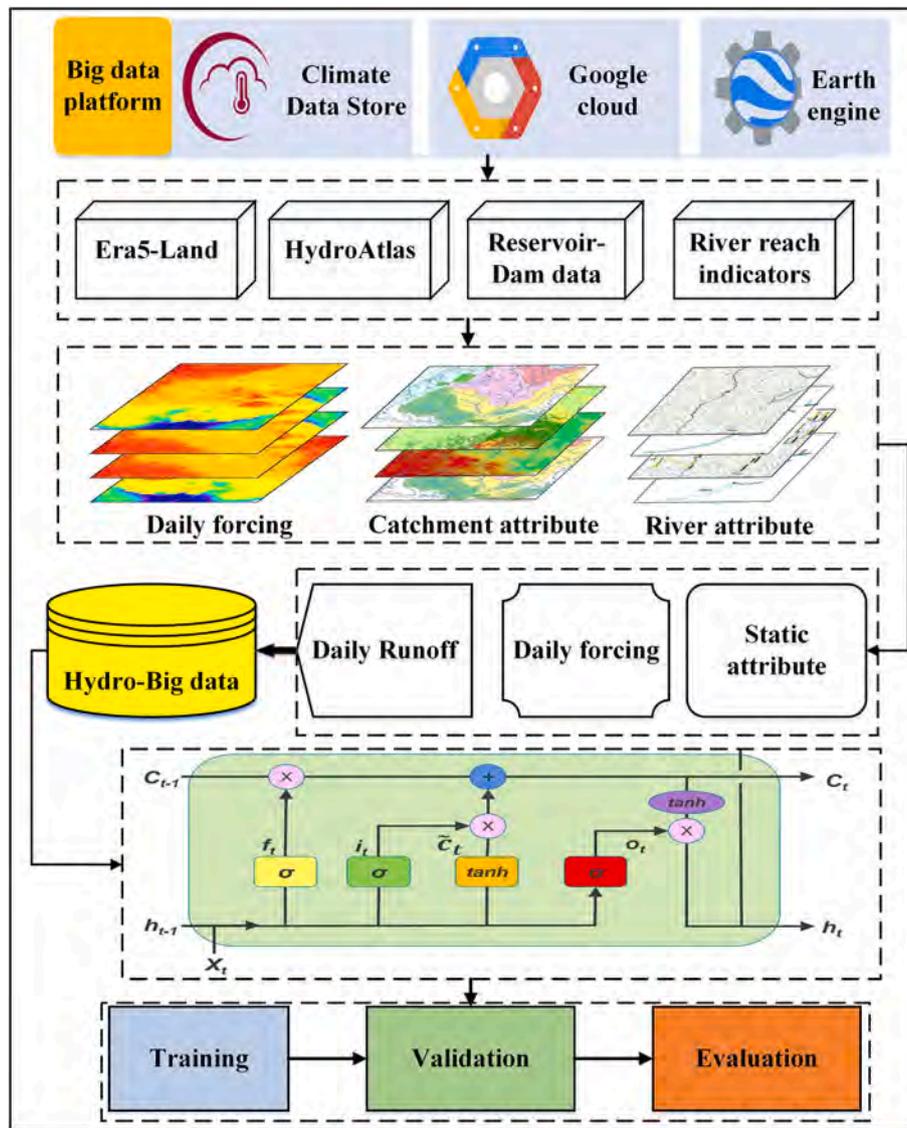


Fig. 4. Graphical description of LSTM-based streamflow modelling across the Yellow River Basin. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

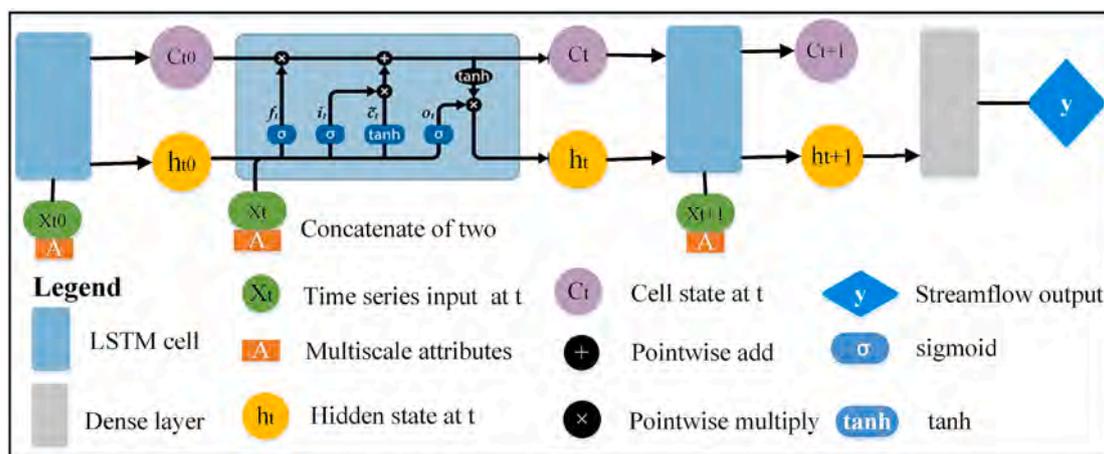


Fig. 5. The standard Long Short-Term Memory (LSTM) cells. The cell state c_{t0} and the hidden state h_{t0} are initialized as zero vectors and then are updated to c_t and h_t with t time series. By feeding the final hidden state h_t into a dense layer, the predicted streamflow y is obtained.

Table 2
List of LSTM model's hyper parameters.

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

LSTM can use to differentiate between catchment streamflow behaviours.

3. Methods

Google Earth Engine (GEE) is a free-to-use cloud service with a large catalogue of geospatial data. On the GEE platform, it is not necessary for users to download and store large amounts of gridded meteorological data, and the platform does not require any specific hardware (Gorelick et al., 2017). We processed the area-weighted spatial average for variables in each catchment in the GEE platform and created hydro-big data from different sources of datasets. After data processing, we trained, validated, and tested the model performance. The flowchart of data processing and LSTM setup and evaluation is shown in Fig. 4.

3.1. Brief description of LSTM

LSTM is able to distinguish valuable information over both short and long-term time horizons (Ha et al., 2021; Vu et al., 2021). It retains an internal state that is updated at every time step by a set of different gates (Fig. 5). At each time step, these gates control the input-state relationship, and only important information passes through the forget gate, while other unnecessary information can be discarded. Whenever information passes through LSTM memory cells, updating actions are performed to determine what old information is discarded and what new information is added (Tiggeloven et al., 2021; Zou et al., 2021).

3.2. Regional LSTM model training

In this study, we employed a LSTM architecture consisting of three layers. The initial two layers were designed with 128 units in their hidden states, and the third layer was a fully connected layer. This connection linked the LSTM output at the final time step to a single output neuron that employed linear activation. The other important hyperparameter was the input sequence length, which determined the range of meteorological input data used by the LSTM to predict the subsequent streamflow. It was prescribed with 365 days to represent a full annual cycle of seasonality, enabling the LSTM model to capture the full annual cycle of streamflow. To simulate daily streamflow, the LSTM model was fed with the last 365 timesteps of meteorological observations. To ensure efficient learning, moreover, we normalized all input features (meteorological variables and multiscale attributes) and the output data (discharge) by subtracting the mean and dividing by the standard deviation. Importantly, the mean and standard deviation for normalization were calculated exclusively from the train-validation period. During testing, the output of the network was retransformed using the normalization parameters from the training period. Detailed information regarding the LSTM model parameters can be found in Table 2. Our LSTM model used the mean square error (MSE) loss as the objective function. We tried other loss functions, such as RMSE, NSE and KGE, while MSE yield the best performance among these loss functions (Lees et al., 2021).

We trained a regional hydrological model (M_θ), which was fed with a single set of parameters θ for all catchments across the Yellow River Basin. Every timestep, the input gate receives both meteorological forcing data ($X_{t,n}$) and static data (A_n , the multiscale attributes can be included in this vector), and they are appended to form a vector of dynamic inputs ($[A_n, X_{t,n}]$). The regional hydrological model M_θ can be expressed as,

$$Y_{t,n} = M_\theta ([A_n, X_{t-k+1:n}, \dots, X_{t,n}]; \theta). \quad (1)$$

This model was used to simulate discharge ($Y_{t,n}$) across the Yellow River Basin. This setup provides an advantage for the LSTM over traditional hydrological models, as the LSTM is able to modify the input gate based on information from time-varying data and provide different parameters at each timestep.

The LSTM model was trained using data from 1 January 1987 to 31 December 2012 and tested from 1 January 2013 to 30 December 2019. We chose the years from 1981 to 1986 for the model validation, which means that we separated time periods for calibration period (1987–2013; train period) and evaluation (2013–2019; test period).

The meteorological forcing data for the LSTM include daily precipitation, surface net solar radiation, potential evapotranspiration, wind speed, mean surface pressure, and mean 2-m surface temperature (dynamic data- $X_{t,n}$) and streamflow (target data- $Y_{t,n}$). We not only prepared individual features describing each catchment's topographic, land-cover, and climatic properties as static inputs but also included river-reach scale static attributes, as shown in Table 1, which represent the pressure level of a river reach from human activities. The river and catchment attributes (static data- $A_{t,n}$) were selected to indicate hydrological information that the LSTM model was able to assess between different catchment rainfall-runoff behaviours.

3.3. Model comparison

We compared LSTM with the process-based model LISFLOOD to assess its performance. This comparison will contribute to a deeper understanding of the potential advantages of well-established global hydrological models like LISFLOOD compared to the emerging multiscale LSTM model. It is important to note that we did not run the LISFLOOD model ourselves in this study. Instead, we used the GloFAS-ERA5 global river discharge reanalysis dataset which is generated from the LISFLOOD model.

3.3.1. LISFLOOD: GloFAS-ERA5 river discharge

The LISFLOOD model, employed as a cornerstone of the GloFAS-ERA5 river discharge reanalysis dataset, is a sophisticated grid-based hydrological model renowned for simulating hydrological processes within catchments (Harrigan et al., 2023). Its comprehensive framework incorporates multiple modules, allowing it to simulate both surface and subsurface process at a grid scale. Through horizontal and vertical water transport mechanisms, LISFLOOD replicates water movement across landscapes and soils. LISFLOOD serves as a foundational model for flood and drought simulation systems (Hales et al., 2022). The meteorological variables used in GloFAS include total precipitation, 2-metre temperature, 2-metre dew temperature, 10-metre U wind component, 10-metre V wind component, downward surface solar radiation, and surface thermal radiation. It has been extensively utilized in diverse applications, spanning large river basins, transnational contexts, and even continental and global scales (Chen et al., 2019; Senent-Aparicio et al., 2021; Swain et al., 2023). LISFLOOD possesses a unique capability to incorporate features that significantly influence river discharge patterns, including lakes, reservoirs, and human water usage. The integration of over 400 large lakes and numerous substantial reservoirs into the GloFAS river network underscores its versatility.

Generating the GloFAS-ERA5 river discharge reanalysis data involve utilizing the LISFLOOD model along with daily surface and sub-surface

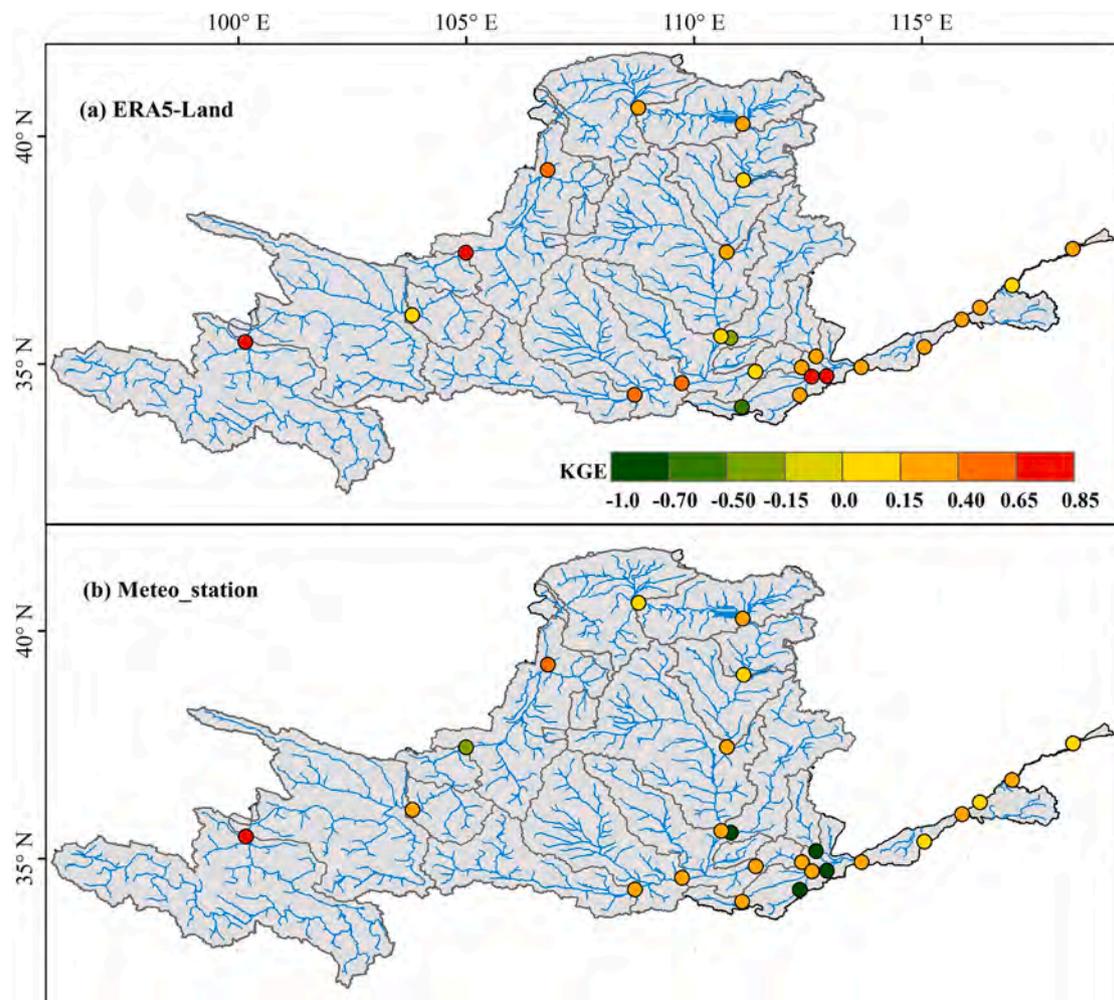


Fig. 6. Comparing the KGE results from ERA5_Land (above) and Meteo_station (below) data forcing.

runoff data from Hydrology Tiled ECMWF Scheme for Surface Exchanges over Land (HTESSEL). The GloFAS-ERA5 river discharge data, a globally comprehensive reanalysis, extends over four decades, starting from 1 January 1979. Notably, the operational nature of the data enables its availability nearly in real-time, 2–5 days after ERA5 data becomes accessible (Harrigan et al., 2020). The dataset's attributes, including its global coverage, resolution, and innovative operational production, cement its significance in enhancing flood forecasting, water resource management, and hydrological research worldwide (Zhao et al., 2022).

3.3.2. Evaluation metrics

In this study, the simulation accuracy of each model is evaluated by statistical error measurements and discharge process error, i.e., Kling–Gupta efficiency (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:

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

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. The ideal value of KGE and its three components (a , b , r) is 1, which means that the model simulation accuracy is good and reliable; Thus, KGE values greater than -0.41 indicate that a model

improves upon the mean flow benchmark – even if the model's KGE value is negative (Knoben et al., 2019). In a more stringent criterion, if the KGE value is lower than 0.4, the simulation is deemed not credible, necessitating further optimization of the model.

4. Results

4.1. Advantage of the ERA5-Land dataset

We compared the LSTM performance forced by ERA5-Land and the meteorological station data (Meteo_station) for estimating daily streamflow. The main objective of this evaluation is to demonstrate the advantage of the ERA5-Land data in driving the LSTM-based streamflow model. We only used meteorological forcing as inputs and did not include any static attributes, which may help to distinguish the difference between the two forcing inputs.

Fig. 6 shows the comparison between ERA5-Land and Meteo_station. It is evident that both models perform well in the source regions. From the middle to the lower reaches, however, a discrepancy in performance becomes apparent. ERA5-Land maintains a consistent level of simulation accuracy in these regions, while Meteo_station exhibits relatively poor performance. For the ERA5-Land driven model, the mean and median KGE scores across the catchments are 0.21 and 0.23, respectively. Notably, only two catchments yield KGE scores below 0, and four catchments achieved KGE scores are over 0.6. In contrast, the Meteo_station forcing model displays a mean KGE of 0.08 and a median KGE of 0.22, with five catchments producing KGE scores below 0 and only one

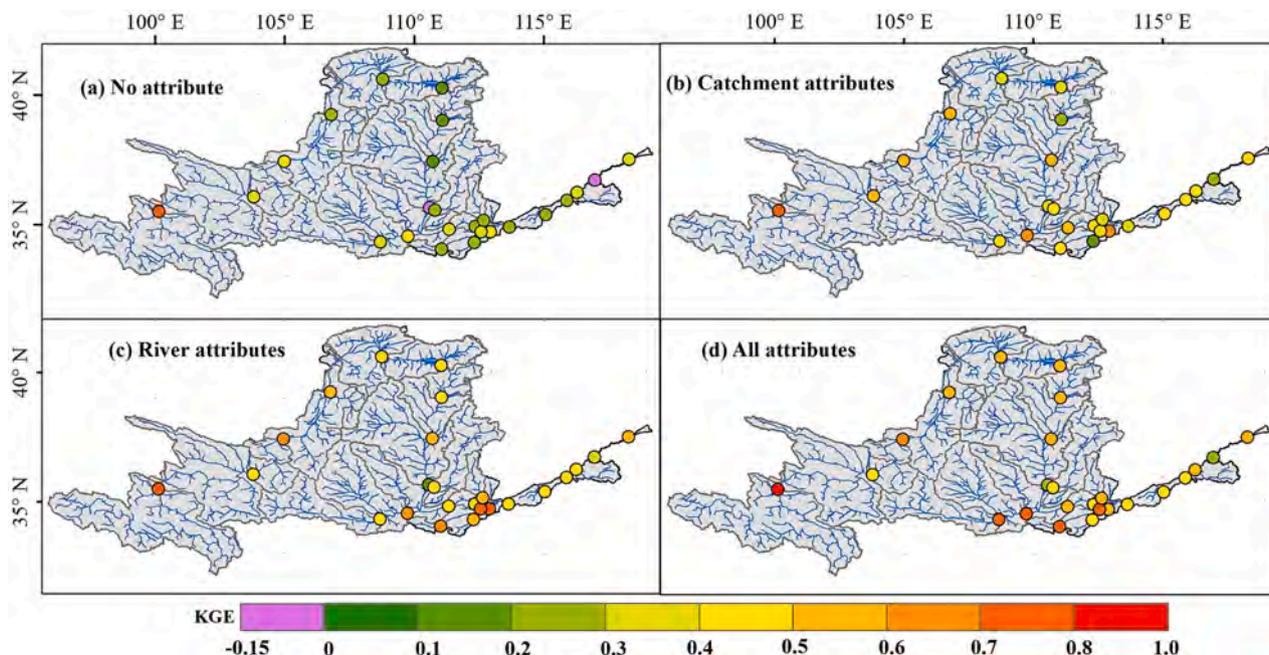


Fig. 7. Spatial KGE patterns for LSTM with vs. without multiscale attributes.

catchment exceeding 0.6. The primary reason for the reduced performance of *Meteo_station* in the middle and lower reaches is likely the increasing size of the basin areas, and the meteorological station data have relatively poor spatial representativeness. Therefore, the *Meteo_station* driving data could not represent the hydrological process across larger basins due to heterogeneity. Overall, the ERA5-Land data contain more useful spatial information to simulate daily streamflow in large and data-limited basins.

4.2. The role of multiscale attributes in model prediction

The multiscale attributes regarding catchment and river reach scale attributes may have different impacts on the performance of LSTM.

Thus, we designed four different scenarios: (1) no-attribute, the LSTM model did not use any catchment or river catchments as input; (2) catchment attribute, the LSTM model was fed with catchment attributes; (3) river attribute, the model was fed with river attributes; and (4) all-attribute, all the catchment and river attributes were used as static attributes in the LSTM model. Please note that in these scenarios we used the same meteorological variables from ERA5-Land to force the LSTM models. The four scenarios can isolate the contribution of each type of attribute in LSTM-based streamflow modelling.

Fig. 7 shows the spatial distribution of KGE for different scenarios. The no-attribute scenario provides relatively poor simulation for most catchments (mean KGE = 0.284 and median = 0.291), except for the source region (Fig. 7a). This scenario struggled to produce acceptable

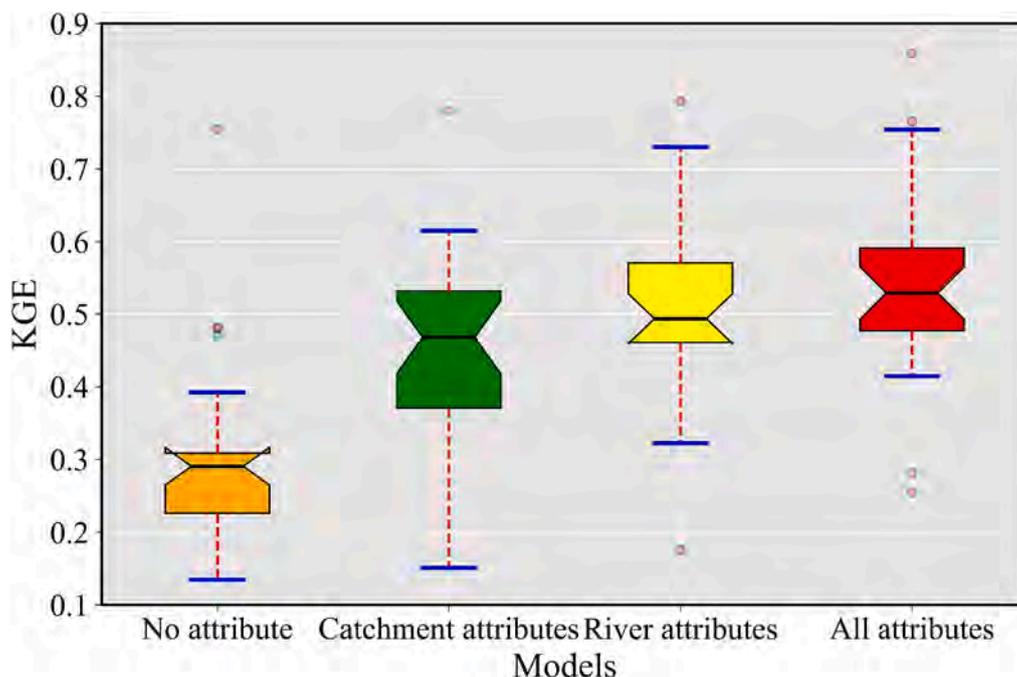


Fig. 8. Boxplots of KGE values for different scenarios.

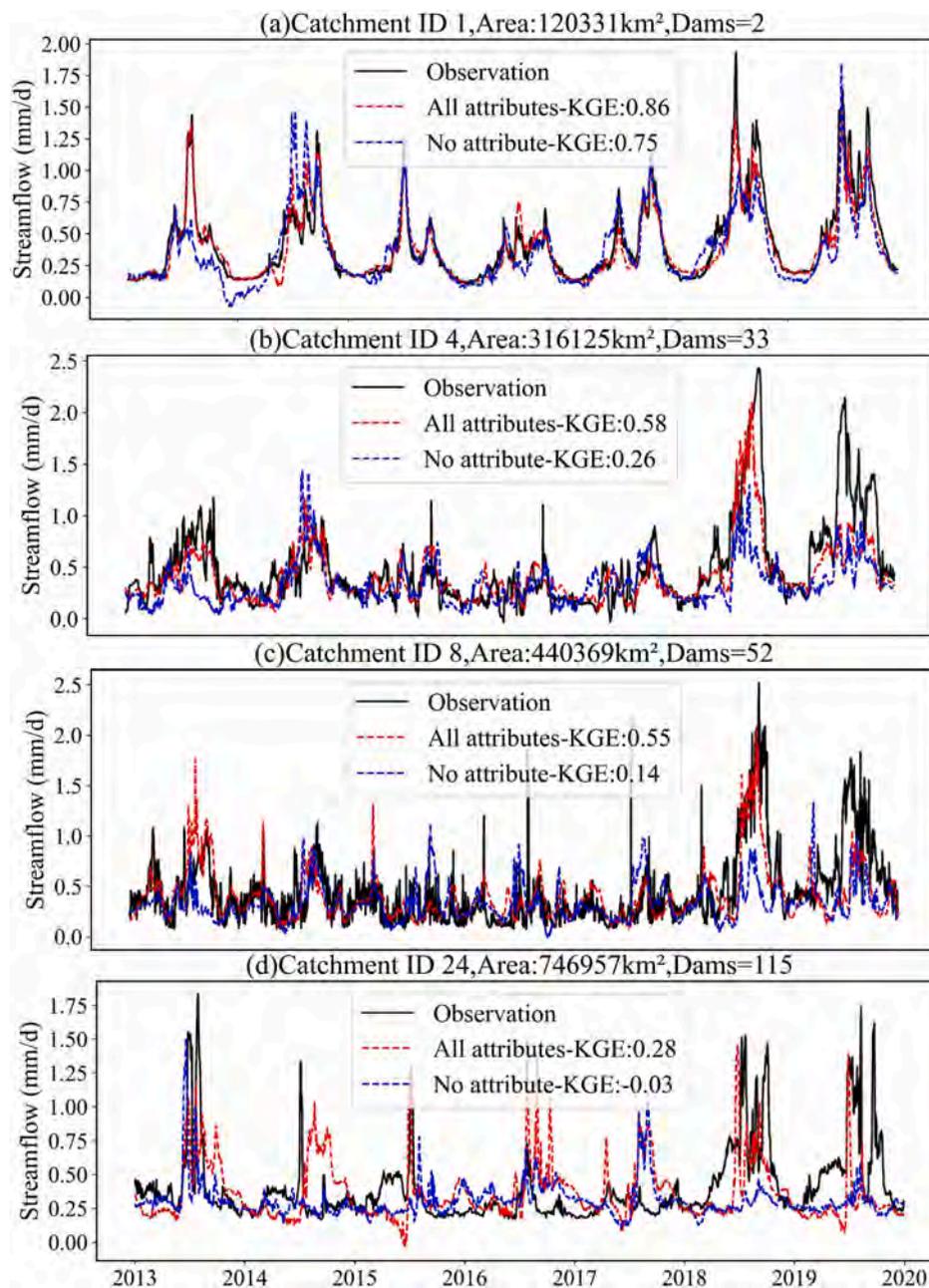


Fig. 9. Time series of data and model predictions from models with no attribute, catchment attributes, river attributes and all attributes for discharge simulation for the test period.

simulations in catchments with relatively significant anthropogenic activities in the middle and lower reaches. In contrast, as shown in Fig. 7b, there is significant improvement in the scenario with the catchment attribute, which implies that these catchment static attributes could provide more valuable information for learning hydrological behaviours (mean KGE = 0.449 and median = 0.452). Although there has been an improvement, the catchment attribute scenario still exhibits poor performance in certain catchments within the middle and lower reaches.

For the scenario with river attributes (Fig. 7c), we can clearly see that the river static attributes are helpful in improving model performance in large and human-regulated catchments (mean KGE = 0.514 and median = 0.521). When all the catchment and river attributes were used as static attributes, LSTM achieved the best performance among the four scenarios (Fig. 7d). In particular, it reproduced streamflow processes well in the middle and lower river reaches for larger catchments (catchment

areas range from 400,000 to 750,000 km²), where large anthropogenic impulses (e.g., reservoir and dam management, agricultural irrigation, urbanization) impacted the hydrological cycle.

Fig. 8 shows the statistics of KGE of the four scenarios for all catchments. The LSTM with the all attributes (i.e., the multiscale attributes) outperformed the other models. The performance of LSTM with river static attributes nearly approaches all-attribute LSTM model performances, which indicates that river static attributes mainly contributed to improving model performance. Catchment static attributes provided lower performance than both all-attribute and river static attribute models, and most of the KGE values were less than 0.45. The LSTM with no attribute model struggled to produce reliable simulations, and the value of KGE in most catchments was less than 0.30. It is necessary to assess the impacts of different types of static attributes on model performance. Δ KGE is the difference between a reference model (no attribute) and the other three different scenarios. The mean

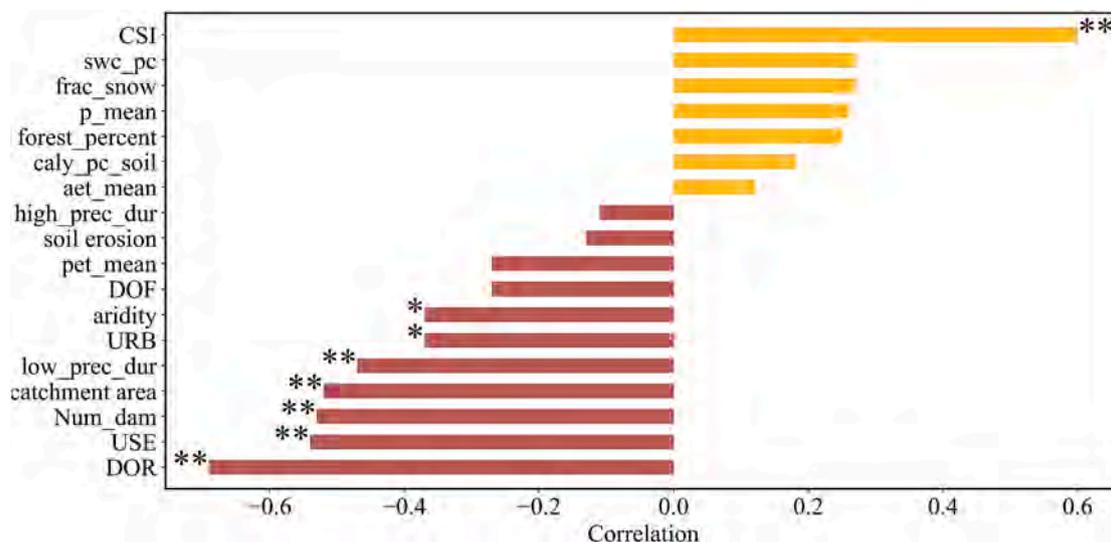


Fig. 10. Static features (rows) and their Spearman’s rank correlation coefficient with model (columns) KGE scores. The positive correlations are in yellow, and the negative correlations are in red. The asterisks “***” and “**” indicate that the correlation is significant at the $P = 0.01$ and $P = 0.05$ level, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

differences (Δ_{meanKGE}) between the no attribute and all-attribute LSTM models are most significant ($\Delta_{\text{meanKGE}} = 0.262$), and the differences for river static attributes ($\Delta_{\text{meanKGE}} = 0.232$) and catchment static

attributes ($\Delta_{\text{meanKGE}} = 0.165$). While the mean performance demonstrates significant differences, due to the presence of worse performing stations, the median differences are smaller for the all-attribute LSTM

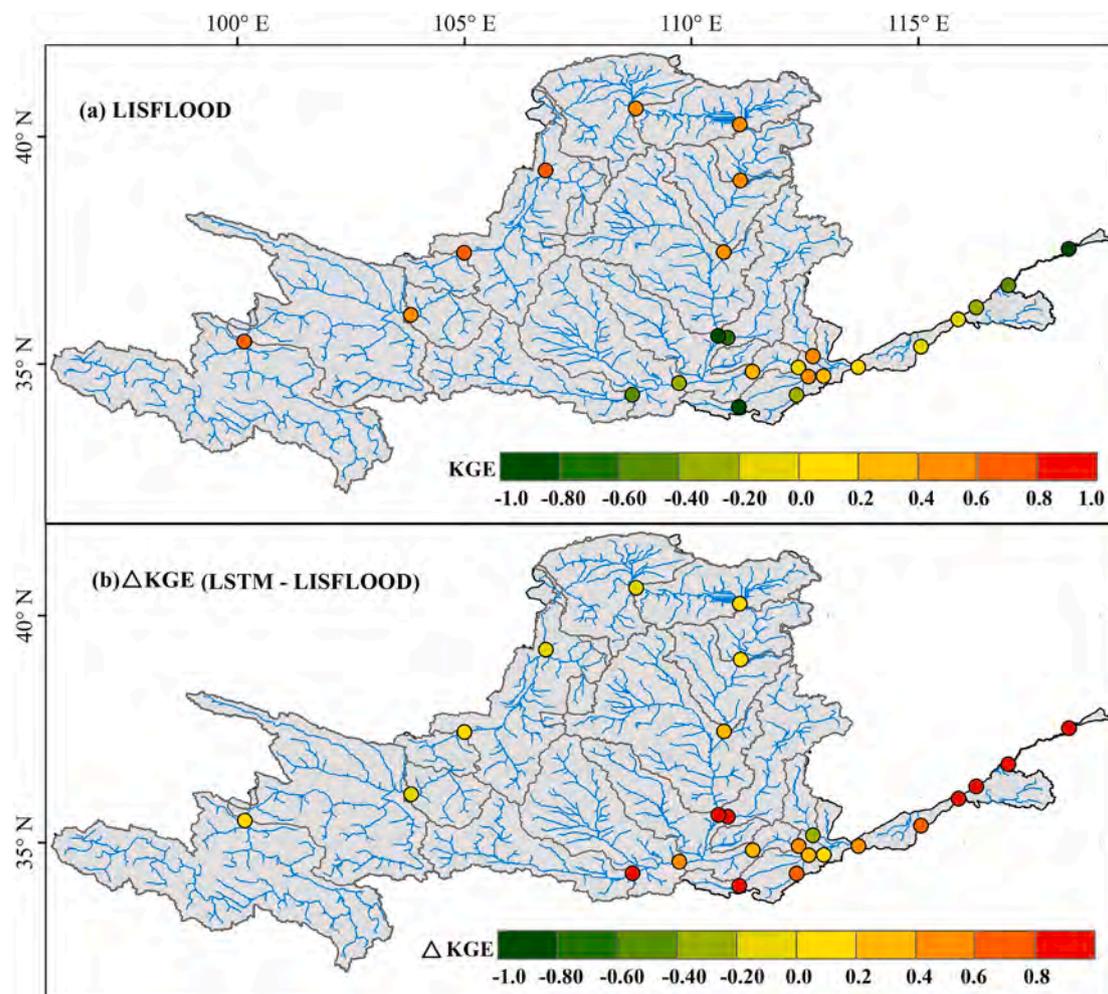


Fig. 11. Spatial KGE patterns of LISFLOOD (above) and the Δ KGE relative to the reference model.

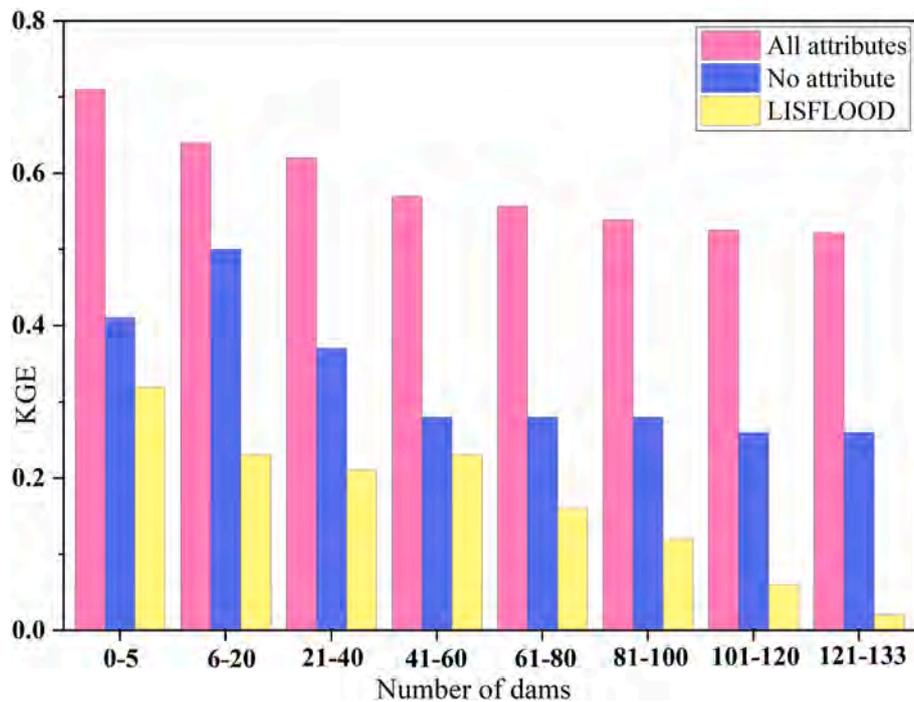


Fig. 12. KGE distribution with the increase in river dams in catchments.

model ($\Delta_{\text{median}}\text{KGE} = 0.234$) and the catchment static attribute LSTM model ($\Delta_{\text{median}}\text{KGE} = 0.155$), and the median difference is still larger for the river static attribute model ($\Delta_{\text{median}}\text{KGE} = 0.237$). Both summaries (median, mean) demonstrate that river-reach scale static attributes play critical roles in streamflow simulation in extremely complicated hydrological conditions.

To further analyse the model performance under different scenarios, we present several catchment hydrographs where catchments are impacted by human interventions at different levels. For catchment ID 1 (area = 120,331 km² and number of dams = 2), the KGE value was 0.86 for the all-attribute model and 0.75 for the no-attribute model; both models were able to capture the streamflow dynamics in catchments with relatively low human impact (Fig. 9a). For catchment ID 4 (area = 316,125 km², dams = 33) and ID 8 (area = 440,369 km², dams = 52), the KGE values for all-attribute scenarios were 0.58 and 0.56, respectively. In contrast, the KGE values were 0.26 and 0.14 for the no attribute scenarios, respectively. Presumably, the performance of the no attribute model continuously decreased due to significant human interventions (Fig. 9b and 9c). For extremely large catchments (area = 746,957 km², dams = 115), the KGE value decreases from 0.28 in all-attribute models to -0.03 in the scenario without attributes (Fig. 9d). Therefore, the model with all attributes is able to learn from complicated streamflow dynamics, but the no attribute model fails to reproduce the streamflow across large and human-regulated catchments.

To identify the contribution of the river and catchment attributes to the model performance, we calculated the Spearman's rank correlation of KGE and each static attribute. As shown in Fig. 10, the static attributes for degree of regulation (DOR, -0.69), number of dams in catchments (Num_dam, -0.53), degree of water use (USE, -0.54) and all anthropogenic attributes present a negative correlation with the model KGE scores. In contrast, the river connectivity status index (CSI, 0.6) is positively correlated with the model KGE scores. Human activities substantially alter the hydrological cycles in large basins, so including river static attributes in LSTM can improve streamflow prediction.

4.3. Benchmarking LSTM with LISFLOOD

We compare LSTM with LISFLOOD to distinguish the advantage of

LSTM. Please note here the LSTM and the LISFLOOD were forced with the same meteorological forcing data from ERA5-Land. Fig. 11a shows spatial KGE patterns from the LISFLOOD model. The LISFLOOD model performed poorly with negative KGEs in 12 catchments, which are characterized by large areas ($\geq 350,000$ km²) with relatively strong human impacts. The mean and median KGE from LISFLOOD are approximately 0.03, 0.01, respectively. Spatially, the LISFLOOD model struggled to produce credible simulations across large and human-regulated catchments, particularly in lower reaches. In the majority of these catchments, negative KGE values were observed.

Fig. 11b shows the difference of KGE between LSTM with all attributes and LISFLOOD, i.e., ΔKGE ($\text{LSTM}_{\text{all attribute}} - \text{LISFLOOD}$). We can see that ΔKGE is generally larger in the lower reaches than in the other areas. The mean and median ΔKGE s are approximately 0.53. Therefore, the spatial patterns of ΔKGE clearly indicate that the LSTM model with all attributes exhibits better performance.

Fig. 12 shows the comparison between the LSTM and the LISFLOOD models at dammed catchments, which was represented as the relationship between the number of dams and the KGE values across the 25 catchments across the Yellow River Basin. It is worth noting that as there are a few dams in catchments, there are no significant differences in the performances of the three different models. Which implies that both LISFLOOD and no-attribute LSTM models are able to provide reliable streamflow simulation in catchments with low anthropogenic impact. However, as the number of dams increases, the river connectivity will decrease, leading to an increase in the pressure level of river reaches. Therefore, the KGE scores of the no-attribute LSTM and LISFLOOD models dropped significantly due to the increase in river dams. When the number of dams in a catchment is close to 120, the performance of the all-attribute (i.e., the multiscale attribute) LSTM may degrade, but it still maintains relatively stable simulation accuracy with KGE over 0.54. Please note the improvement for the all-attribute LSTM model is most likely due to the incorporation of anthropogenic attributes at the catchment and river scales. This is especially meaningful for future water management because streamflow estimation at human-regulated catchments was previously a source of large errors for LSTM streamflow modelling.

5. Discussion

In this study, we examined two different sources of forcing data for LSTM streamflow simulations. The results clearly suggest that a significant difference exists for ERA5-Land and Meteo_station across the Yellow River Basin. The distribution of meteorological stations is uneven in the Yellow River Basin, and few stations are available in the upper and middle reaches. The lack of sufficient spatial data will degrade the performance of the Meteo_station driving model in the middle and lower reaches (Fig. 5). However, the ERA5-Land data have been widely applied in different climate-hydrological studies and are able to represent more sufficient spatial information for land surface meteorological forcing (Kratzert et al., 2023). Therefore, LSTM forced by ERA5-Land shows better performance in extremely large and data-limited catchments. As the catchment area increased, the Meteo_station forced LSTM model struggled to produce credible simulations. In contrast, the ERA5-Land forced model shows good performance in the lower reaches across large catchments. This advantage evidenced that ERA5-Land data represent meteorological patterns across large catchments. It is worth emphasizing that ERA5-Land data could be an attractive option in large and data-limited areas for deep learning-based hydrological models.

It is a difficult task to simulate streamflow in large and human-regulated catchments because we are still not clear about the purpose and regulation rules of reservoirs (Althoff et al., 2021). Even for process-based hydrological models, coupling various human-regulation impacts (e.g., agricultural irrigation, dam and reservoir regulations) is a great challenge (Xie and Cui, 2011), and their performance is generally spoiled due to uncertainties from human activities. Therefore, it is necessary to describe these human interventions for LSTM models. We used the river connectivity index (CSI) and other pressure indicators in the river reach. As expected, LSTM with river static attributes provides better performance than the model with catchment static attributes. It is worth mentioning that treating river static attributes as catchment attributes is a strong assumption that could significantly improve the model performance in human-regulated catchments. The Yellow River Basin's complex dynamics underscore the significance of capturing human-regulated changes in river systems that affect streamflow patterns. It is essential to indicate that the value of river attributes isn't solely based on the number of attributes but on their ability to capture relevant and meaningful information in a human-regulated context. The main findings in this study imply that river static attributes are helpful for providing reliable streamflow simulations in large human-regulated catchments. Thus, it is important to contain river static attributes to train regional models in strongly human-regulated and large catchments.

The main results for this study show that across the 25 catchments, LSTM with different static attributes was able to outperform the global hydrological model (i.e., LISFLOOD) in almost all catchments (Fig. 11). For the upstream catchments, the LISFLOOD model nearly reached the performance of the LSTM model. However, the LISFLOOD model is unable to represent human-regulated streamflow processes, so it struggles to provide reliable simulations in the mid-lower reaches, where most catchments are strongly human regulated.

Uncertainty in LSTM-based streamflow modelling arises from various sources. The quality and quantity of streamflow and meteorological variables may impact the model performance (Botterill and McMillan, 2023; Klotz et al., 2022; Moosavi et al., 2022). Model structure and hyperparameter tuning are critical aspects, as variations in hyperparameters like the number of hidden layers, units, loss function, optimization methods and learning rates can significantly impact model performance (Gauch et al., 2021). The setup of the training period and data splitting for validation are also essential, and inadequate representation of hydrological conditions during training may lead to poor generalization (Kratzert et al., 2019; Manh-Hung Le, 2022). Therefore, this study deliberately selected the training period from 1987 to 2012. This period spans a time before and after the year 2000, during which

the human impact on the Yellow River Basin significantly increased. By incorporating both low and significant human impacts into the training period, we equipped the LSTM model with the ability to capture a wide range of hydrological behaviours. Climate variability, model overfitting, and the limited interpretability of LSTM models also contribute to uncertainty (Althoff et al., 2021). Furthermore, the model transferability across regions and the dynamic nature of human impacts on hydrological systems introduce additional challenges (Arsenault et al., 2023; Kratzert et al., 2021; Ma et al., 2021). Managing these uncertainties necessitates a rigorous approach to data handling, model development, and ultimately enhancing the reliability of LSTM-based streamflow predictions.

This study focused on the application of LSTM with multiscale attribute formulations in human-regulated catchments. Natural nested catchments would be expected to provide more information between sites than the regulated catchments, thereby improving the performance of the LSTM. However, daily streamflow patterns in neighbouring catchments can vary significantly due to the influence of local factors and human activities across the Yellow River Basin. There may be limited common information between these nested catchments. Therefore, estimating downstream catchments solely based on upstream catchments is challenging. The benefit of nesting catchments requires further investigation and could be a subject of future research. Moreover, we noticed there is a surge of interest in hybrid hydrological modelling techniques, such as the differentiable parameter learning model (DPL), which has shown great promise (Feng et al., 2022; Kraft et al., 2022). Most techniques have been applied in small catchments with relatively low anthropogenic impacts. Therefore, we suggest that in future studies, it is necessary to evaluate their performance in human-regulated catchments, and there is still room to improve model performance in extremely large catchments.

6. Conclusions

In this study, we address the question of whether LSTM can produce credible streamflow prediction in large human-regulated basins. The results of this study provide three main conclusions:

(i) The LSTM model with the ERA5-Land forcing can provide better performance than that with the Meteo_station forcing in data-sparse and strongly human-influenced regions. ERA5-Land has the advantage of generalizing spatial patterns across large and data-limited catchments, and it is helpful to distinguish catchment-based hydrological processes and could provide valuable information for LSTM-based hydrological modelling.

(ii) The LSTM model with river static attributes provides better streamflow prediction than that with catchment static attributes, and the LSTM model combining river and catchment static attributes could significantly improve the model performance. This suggests that the multiscale attributes could be valuable to improve rainfall-runoff prediction in large and human-regulated catchments.

(iii) Both LSTM without attributes and the LISFLOOD model perform well in the upper regions with slight human interventions, but their performance is not acceptable in the middle and lower regions due to strong human interventions. The LSTM model with river attributes can improve this situation, and the improvement is particularly significant in catchments with large numbers of dams and reservoirs.

The strengths of LSTM with river static attributes over catchment static attributes highlight the power of learning ability in extremely complicated conditions (e.g., extremely large catchments, significant anthropogenic interventions, and scarce data). However, in some catchments, the LSTM model still struggles to produce acceptable performance due to the lack of explicit information related to human activities. These challenges stem from the large number of dams and reservoirs in the Yellow River Basin, each with its unique set of regulation rules, making it difficult to discern and incorporate these intricate regulatory processes. Additionally, quantifying water use for purposes

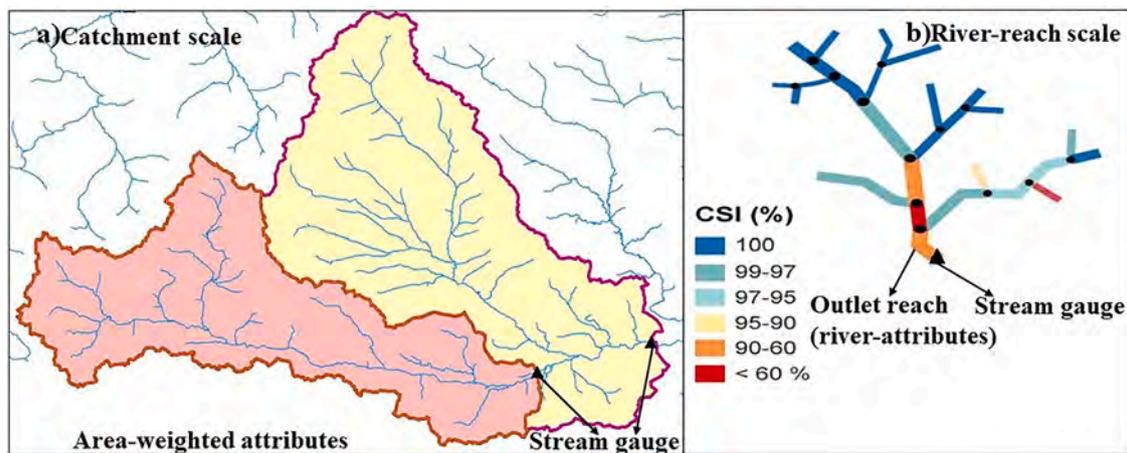


Fig. A1. Multiscale attributes for catchment and river.

like irrigation and extraction from rivers is a complex task, further contributing to the model's difficulties in these catchments. In future studies, it is necessary to develop a high-dimensional embedding layer that allows for more dynamic or high-scaled static attributes to represent complicated hydroclimate conditions in the human-regulated basins. Moreover, coupling river attributes in hybrid hydrological modelling may be a good attempt to improve processes-based and DL models for streamflow prediction in complex river catchments.

CRedit authorship contribution statement

Arken Tursun: Data processing, software, computation, writing original draft. **Xianhong Xie:** Conceptualization, supervision, writing-reviewing and editing. **Yibing Wang:** Data processing. **Yao Liu:** Data processing. **Dawei Peng:** Data processing. **Buyun Zheng:** Discussion.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used are partially confidential due to the data policy restriction in the data holder institutions.

Acknowledgement

This study was supported by grants from the National Natural Science Foundation of China (No. 41971030) and the Open Fund of State Key Laboratory of Remote Sensing Science and Beijing Engineering Research Center for Global Land Remote Sensing Products (No. OF202202). Acknowledgement for the data support from "Loess Plateau SubCenter, National Earth System Science Data Center, National Science & Technology Infrastructure of China. (<http://loess.geodata.cn>)". We would like to thank Frederik Kratzert for releasing "neuralhydrology" codebase (Frederik Kratzert, 2019).

Appendix

In this study we derived attributes at catchment and river scales, which would help LSTM model to represent hydrological process. For catchment attributes, we computed the spatial join of the HydroATLAS polygons and the catchment boundaries and then derived the catchment attributes as an area-weighted aggregate. To represent the human

disturbance at river-reach scale on streamflow, we obtained series of river-reach scale attributes, including the values of DOF, DOR, SED, USE, RDD, URB, and CSI. These attributes represent pressure factors from human interferences of river connectivity. It should be noted that the way of calculation of river-reach attributes significantly different from catchment attributes. We did not calculate average value of all river-reaches within the catchment, but considered the river-reach attributes at the outlet of each catchment (Fig. A1). In this way we precisely represented the anthropogenic signatures at river-reach scale. As shown in Fig. A1, CSI (connectivity status index) values were estimated for each individual river reach. We only considered the CSI values in the outlet reach of the catchment.

References

- Althoff, D., Rodrigues, L.N., Silva, D.D.D., 2021. Addressing hydrological modeling in watersheds under land cover change with deep learning. *Adv. Water Resour.* 154, 103965. <https://doi.org/10.1016/j.advwatres.2021.103965>.
- 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. *Hydrol. Earth Syst. Sc.* 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 Resour. Res.* 57 (5). <https://doi.org/10.1029/2020WR029328>.
- Botterill, T.E., Mcmillan, H.K., 2023. Using machine learning to identify hydrologic signatures with an encoder–decoder framework. *Water Resour. Res.* 59 (3). <https://doi.org/10.1029/2022WR033091>.
- Chen, K., Guo, S., Wang, J., Qin, P., He, S., Sun, S., Naeini, M.R., 2019. Evaluation of glosas-seasonal forecasts for cascade reservoir impoundment operation in the upper yangtze river. *Water-Sui* 11 (12), 2539. <https://doi.org/10.3390/w11122539>.
- Cho, K., Kim, Y., 2022. Improving streamflow prediction in the wrf-hydro model with lstm networks. *J. Hydrol.* 605, 127297. <https://doi.org/10.1016/j.jhydrol.2021.127297>.
- Coron, L., Thirel, G., Delaigue, O., Perrin, C., Andréassian, V., 2017. The suite of lumped gr hydrological models in an r package. *Environ. Modell. Softw.* 94, 166–171. <https://doi.org/10.1016/j.envsoft.2017.05.002>.
- Elsaadani, M., Habib, E., Abdelhameed, A.M., Bayoumi, M., 2021. Assessment of a spatiotemporal deep learning approach for soil moisture prediction and filling the gaps in between soil moisture observations. *Frontiers in Artificial Intelligence* 4. <https://doi.org/10.3389/frai.2021.636234>.
- Fan, Y., Clark, M., Lawrence, D.M., Swenson, S., Band, L.E., Brantley, S.L., Brooks, P.D., Dietrich, W.E., Flores, A., Grant, G., Kirchner, J.W., Mackay, D.S., McDonnell, J.J., Milly, P.C.D., Sullivan, P.L., Tague, C., Ajami, H., Chaney, N., Hartmann, A., Hazenberg, P., Mcnamara, J., Pelletier, J., Perket, J., Rouholahnejad Freund, E., Wagener, T., Zeng, X., Beighley, E., Buzan, J., Huang, M., Livneh, B., Mohanty, B.P., Nijssen, B., Safeeq, M., Shen, C., Verseveld, W., Volk, J., Yamazaki, D., 2019. Hillslope hydrology in global change research and earth system modeling. *Water Resour. Res.* 55 (2), 1737–1772. <https://doi.org/10.1029/2018WR023903>.
- 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 Resour. Res.* 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. *Geophys. Res. Lett.* 48 (14). <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 Resour. Res.* 58 (10). <https://doi.org/10.1029/2022WR032404>.
- Ficchi, A., Perrin, C., Andréassian, V., 2019. Hydrological modelling at multiple sub-daily time steps: model improvement via flux-matching. *J. Hydrol.* 575, 1308–1327. <https://doi.org/10.1016/j.jhydrol.2019.05.084>.
- Frederik Kratzert, D.K.G.S., 2019. Benchmarking a catchment-aware long short-term mem network (Lstm) for large-scale hydrological modeling. *Hydrol Earth Syst Sc.* <https://doi.org/10.5194/hess-2019-368>.
- Gauch, M., Mai, J., Lin, J., 2021. The proper care and feeding of camels: how limited training data affects streamflow prediction. *Environ. Modell. Softw.* 135, 104926. <https://doi.org/10.1016/j.envsoft.2020.104926>.
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google earth engine: planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* 202, 18–27. <https://doi.org/10.1016/j.rse.2017.06.031>.
- Grill, G., Lehner, B., Thieme, M., Geenen, B., Tickner, D., Antonelli, F., Babu, S., Borrelli, P., Cheng, L., Crochetiere, H., Ehalt Macedo, H., Filgueiras, R., Goichot, M., Higgins, J., Hogan, Z., Lip, B., McClain, M.E., Meng, J., Mulligan, M., Nilsson, C., Olden, J.D., Opperman, J.J., Petry, P., Reidy Liermann, C., Sáenz, L., Salinas-Rodríguez, S., Schelle, P., Schmitt, R.J.P., Snider, J., Tan, F., Tockner, K., Valdujo, P. H., van Soesbergen, A., Zarfl, C., 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. *J. Hydrol.* 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). <https://doi.org/10.1038/s41598-021-90964-3>.
- Hales, R.C., Nelson, E.J., Souffront, M., Gutierrez, A.L., Prudhomme, C., Kopp, S., Ames, D.P., Williams, G.P., Jones, N.L., 2022. Advancing global hydrologic modeling with thegeoglossw ecmwf streamflow service. *J. Flood Risk Manag.* <https://doi.org/10.1111/jfr3.12859>.
- Hall, J., Perdigão, R.A.P., 2021. Who is stirring the waters? *Science* 371 (6534), 1096–1097. <https://doi.org/10.1126/science.abg6514>.
- Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., Barnard, C., Cloke, H., Pappenberger, F., 2020. Glofas-era5 operational global river discharge reanalysis 1979–present. *Earth Syst. Sci. Data* 12 (3), 2043–2060. <https://doi.org/10.5194/essd-12-2043-2020>.
- Harrigan, S., Zsoter, E., Cloke, H., Salamon, P., Prudhomme, C., 2023. Daily ensemble river discharge reforecasts and real-time forecasts from the operational global flood awareness system. *Hydrol. Earth Syst. Sc.* 27 (1), 1–19. <https://doi.org/10.5194/hess-27-1-2023>.
- Jiang, F., Xie, X., Wang, Y., Liang, S., Zhu, B., Meng, S., Zhang, X., Chen, Y., Liu, Y., 2022. Vegetation greening intensified transpiration but constrained soil evaporation on the loess plateau. *J. Hydrol.* 614, 128514. <https://doi.org/10.1016/j.jhydrol.2022.128514>.
- 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. *J. Hydrol.: Reg. Stud.* 42, 101119. <https://doi.org/10.1016/j.ejrh.2022.101119>.
- Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., Nearing, G., 2022. Uncertainty estimation with deep learning for rainfall-runoff modeling. *Hydrol. Earth Syst. Sc.* 26 (6), 1673–1693. <https://doi.org/10.5194/hess-26-1673-2022>.
- Knoben, W.J.M., Freer, J.E., Woods, R.A., 2019. Technical note: inherent benchmark or not? Comparing nash–sutcliffe and kling–gupta efficiency scores. *Hydrol. Earth Syst. Sc.* 23 (10), 4323–4331. <https://doi.org/10.5194/hess-23-4323-2019>.
- Kraft, B., Jung, M., Körner, M., Koirala, S., Reichstein, M., 2022. Towards hybrid modeling of the global hydrological cycle. *Hydrol. Earth Syst. Sc.* 26 (6), 1579–1614. <https://doi.org/10.5194/hess-26-1579-2022>.
- 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 Resour. Res.* 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. *Hydrol Earth Syst Sc* 25 (5), 2685–2703. <https://doi.org/10.5194/hess-25-2685-2021>.
- Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., Matias, Y., 2023. Caravan – a global community dataset for large-sample hydrology. *Sci. Data* 10 (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. *Hydrol. Earth Syst. Sc.* 25 (10), 5517–5534. <https://doi.org/10.5194/hess-25-5517-2021>.
- Li, H., Ke, C., Zhu, Q., Li, M., Shen, X., 2022a. A deep learning approach to retrieve cold-season snow depth over arctic sea ice from AMSR2 measurements. *Remote Sens. Environ.* 269, 112840. <https://doi.org/10.1016/j.rse.2021.112840>.
- Li, Y., Lv, C., Chang, J., Wang, Z., Kong, C., 2022d. A bayesian data assimilation method to enhance the time sequence prediction ability of data-driven models. *AIP Adv.* 12 (10), 105021. <https://doi.org/10.1063/5.0119688>.
- Li, J., Miao, C., Zhang, G., Fang, Y.H., Shangguan, W., Niu, G.Y., 2022b. Global evaluation of the noah-mp land surface model and suggestions for selecting parameterization schemes. *J. Geophys. Res. Atmos.* 127 (5). <https://doi.org/10.1029/2021JD035753>.
- Li, Q., Zhu, Y., Shangguan, W., Wang, X., Li, L., Yu, F., 2022c. An attention-aware lstm model for soil moisture and soil temperature prediction. *Geoderma* 409, 115651. <https://doi.org/10.1016/j.geoderma.2021.115651>.
- 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. *Sci. Data* 6 (1). <https://doi.org/10.1038/s41597-019-0300-6>.
- Liu, Y., Yang, Z., Lin, P., Zheng, Z., Xie, S., 2020. Comparison and evaluation of multiple land surface products for the water budget in the yellow river basin. *J. Hydrol.* 584, 124534. <https://doi.org/10.1016/j.jhydrol.2019.124534>.
- Ma, K., Feng, D., Lawson, K., Tsai, W.P., Liang, C., Huang, X., Sharma, A., Shen, C., 2021. Transferring hydrologic data across continents – leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. *Water Resour. Res.* 57 (5). <https://doi.org/10.1029/2020WR028600>.
- Manh-Hung Le, H.K.S.A., 2022. Streamflow estimation in ungauged regions using machine learning: quantifying uncertainties in geographic extrapolation. *Hydrol Earth Syst Sc.* <https://doi.org/10.5194/hess-2022-320>.
- Moishin, M., Deo, R.C., Prasad, R., Raj, N., Abdulla, S., 2021. Designing deep-based learning flood forecast model with convlstm hybrid algorithm. *IEEE Access* 9, 50982–50993. <https://doi.org/10.1109/ACCESS.2021.3065939>.
- Moosavi, V., Gheisoori Fard, Z., Vafakah, M., 2022. Which one is more important in daily runoff forecasting using data driven models: input data, model type, preprocessing or data length? *J. Hydrol.* 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., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D.G., Piles, M., Rodríguez-Fernández, N.J., Zsoter, E., Buontempo, C., Thépaut, J., 2021. Era5-land: a state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* 13 (9), 4349–4383. <https://doi.org/10.5194/essd-13-4349-2021>.
- 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., Shavit, I., Yuklea, R., Royz, M., Giladi, N., Peled Levi, N., Reich, O., Gilon, O., Maor, R., Timnat, S., Shechter, T., Anisimov, V., Gigi, Y., Levin, Y., Moshe, Z., Ben-Haim, Z., Hassidim, A., Matias, Y., 2022. Flood forecasting with machine learning models in an operational framework. *Hydrol. Earth Syst. Sc.* 26 (15), 4013–4032. <https://doi.org/10.5194/hess-26-4013-2022>.
- 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. *J. Hydrol.* 599, 126455. <https://doi.org/10.1016/j.jhydrol.2021.126455>.
- Pelosi, A., Terribile, F., D Urso, G., Chirico, G., 2020. Comparison of era5-land and uerra mescan-surfex reanalysis data with spatially interpolated weather observations for the regional assessment of reference evapotranspiration. *Water-Sui* 12 (6), 1669. <https://doi.org/10.3390/w12061669>.
- Ren, Y., Zeng, S., Liu, J., Tang, Z., Hua, X., Li, Z., Song, J., Xia, J., 2022. Mid- to long-term runoff prediction based on deep learning at different time scales in the upper yangtze river basin. *Water-Sui* 14 (11), 1692. <https://doi.org/10.3390/w14111692>.
- Senent-Aparicio, J., Blanco-Gómez, P., López-Ballesteros, A., Jimeno-Sáez, P., Pérez-Sánchez, J., 2021. Evaluating the potential of glofas-era5 river discharge reanalysis data for calibrating the swat model in the grande san miguel river basin (el salvador). *Remote Sens.-Basel* 13 (16), 3299. <https://doi.org/10.3390/rs13163299>.
- Shen, C., 2018a. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resour. Res.* 54 (11), 8558–8593. <https://doi.org/10.1029/2018WR022643>.
- Shen, C., 2018b. Deep learning: a next-generation big-data approach for hydrology. *Eos* 99. <https://doi.org/10.1029/2018EO095649>.
- Shen, C., Chen, X., Laloy, E., 2021. Editorial: broadening the use of machine learning in hydrology. *Frontiers. Water* 3. <https://doi.org/10.3389/frwa.2021.681023>.
- Swain, S., Mishra, S.K., Pandey, A., Srivastava, P.K., Nandi, S., 2023. Characterization and assessment of hydrological droughts using glofas streamflow data for the narmada river basin, india. *Environ. Sci. Pollut. R.* <https://doi.org/10.1007/s11356-023-27036-8>.
- Tan, X., Gan, T.Y., 2016. Contribution of human and climate change impacts to changes in streamflow of canada. *Sci Rep-Uk* 5 (1). <https://doi.org/10.1038/srep17767>.
- Thomas Lees, S.R.F.K., 2021. Hydrological concept formation inside long short-term memory (Lstm) networks. *Hydrol Earth Syst Sc.* <https://doi.org/10.5194/hess-2021-566>.
- 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). <https://doi.org/10.1038/s41598-021-96674-0>.
- Vu, M.T., Jardani, A., Massei, N., Fournier, M., 2021. Reconstruction of missing groundwater level data by using long short-term memory (Lstm) deep neural network. *J. Hydrol.* 597, 125776. <https://doi.org/10.1016/j.jhydrol.2020.125776>.
- Wang, J., Walter, B.A., Yao, F., Song, C., Ding, M., Maroof, A.S., Zhu, J., Fan, C., Mcalister, J.M., Sikder, S., Sheng, Y., Allen, G.H., Crétaux, J., Wada, Y., 2022. Geodar: georeferenced global dams and reservoirs dataset for bridging attributes and geolocations. *Earth Syst. Sci. Data* 14 (4), 1869–1899. <https://doi.org/10.5194/essd-14-1869-2022>.

- Wi, S., Steinschneider, S., 2022. Assessing the physical realism of deep learning hydrologic model projections under climate change. *Water Resour. Res.* 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>.
- Xie, X., Cui, Y., 2011. Development and test of swat for modeling hydrological processes in irrigation districts with paddy rice. *J. Hydrol.* 396 (1), 61–71. <https://doi.org/10.1016/j.jhydrol.2010.10.032>.
- Xie, X., Liang, S., Yao, Y., Jia, K., Meng, S., Li, J., 2015. Detection and attribution of changes in hydrological cycle over the three-north region of china: climate change versus afforestation effect. *Agr. Forest Meteorol.* 203, 74–87. <https://doi.org/10.1016/j.agrformet.2015.01.003>.
- Xie, J., Liu, X., Bai, P., Liu, C., 2022a. Rapid watershed delineation using an automatic outlet relocation algorithm. *Water Resour. Res.* 58 (3). <https://doi.org/10.1029/2021WR031129>.
- Xie, J., Liu, X., Tian, W., Wang, K., Bai, P., Liu, C., 2022b. Estimating gridded monthly baseflow from 1981 to 2020 for the contiguous us using long short-term memory (lstm) networks. *Water Resour. Res.* 58 (8). <https://doi.org/10.1029/2021WR031663>.
- 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. *J. Hydrol.* 609, 127750. <https://doi.org/10.1016/j.jhydrol.2022.127750>.
- Yilmaz, M., 2023. Accuracy assessment of temperature trends from era5 and era5-land. *Sci. Total Environ.* 856, 159182. <https://doi.org/10.1016/j.scitotenv.2022.159182>.
- Yin, W., Fan, Z., Tangdamrongsub, N., Hu, L., Zhang, M., 2021. Comparison of physical and data-driven models to forecast groundwater level changes with the inclusion of grace – a case study over the state of victoria, australia. *J. Hydrol.* 602, 126735. <https://doi.org/10.1016/j.jhydrol.2021.126735>.
- Yousefi, A., Toffolon, M., 2022. Critical factors for the use of machine learning to predict lake surface water temperature. *J. Hydrol.* 606, 127418. <https://doi.org/10.1016/j.jhydrol.2021.127418>.
- Zhang, R., Li, L., Zhang, Y., Huang, F., Li, J., Liu, W., Mao, T., Xiong, Z., Shangguan, W., 2021. Assessment of agricultural drought using soil water deficit index based on era5-land soil moisture data in four southern provinces of china. *Agriculture* 11 (5), 411. <https://doi.org/10.3390/agriculture11050411>.
- Zhao, T., Chen, Z., Tu, T., Yan, D., Chen, X., 2022. Unravelling the potential of global streamflow reanalysis in characterizing local flow regime. *Sci. Total Environ.* 838, 156125. <https://doi.org/10.1016/j.scitotenv.2022.156125>.
- Zou, G., Zhang, B., Yong, R., Qin, D., Zhao, Q., 2021. Fdn-learning: urban pm2.5-concentration spatial correlation prediction model based on fusion deep neural network. *Big Data Res.* 26, 100269. <https://doi.org/10.1016/j.bdr.2021.100269>.