# Can Transfer Learning Improve Hydrological Predictions in the Alpine Regions?

Yingying Yao[a*], Yufeng Zhao[a], Xin Li[b], Dapeng Feng[c], Chaopeng Shen[c], Chuankun Liu[d], Xingxing Kuang[e], Chunmiao Zheng[e,f]

[a] Department of Earth and Environmental Science, School of Human Settlements and Civil Engineering, Xi'an Jiaotong University, Xi'an, China.

[b] National Tibetan Plateau Data Center, State Key Laboratory of Tibetan Plateau Earth System Science (LATPES), Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing, China.

[c] Civil and Environmental Engineering, Pennsylvania State University, University Park, PA, USA.

[d] Sichuan Academy of Environmental Policy and Planning, Chengdu, China.

[e] State Environmental Protection Key Laboratory of Integrated Surface Water-Groundwater Pollution Control, School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen, China.

[f] EIT Institute for Advanced Study, Ningbo, China.

*Corresponding authors. E-mail addresses: yaoyy27@xjtu.edu.cn (Yingying Yao)

**Abstract:** The Tibetan Plateau (TP) is an important Asian water tower for livelihood, irrigation, hydropower, and downstream ecosystems. Hydrological predictions in the TP have long been limited by sparse and discontinuous observations. Transfer learning (TL) technology may improve hydrological predictions by pre-training deep learning (DL) models on data-rich areas and then applying them to data-limited areas. However, the extent to which the DL and TL models work in alpine regions and the physical knowledge they provide remain unclear. Models were pretrained using data from 671 catchments across the U.S. and fine-tuned using data from four basins around the TP. Our results show that streamflow data with different temporal resolutions (monthly or daily) have little effect on discharge predictions when using DL. The number of discharge observations needed to enable acceptable performance of the DL models (Nash-Sutcliffe Efficiency coefficient > 0.6) depends on the hydrological characteristics of the catchments, in particular, how closely they conform to a general rainfall-runoff system. Climate forcing data determine the performance of streamflow prediction, whereas other attributes (i.e., soil and geology) have less significant impacts on prediction. The effectiveness of the DL and TL models is limited because groundwater contributes significantly to river discharge. This study provides an updated understanding of the application of DL and TL to quantify hydrological changes in the global cryosphere environment.

**Keywords**: Tibetan Plateau, deep learning, transfer learning, streamflow prediction

## 1. Introduction

The Tibetan-Himalayan region is the world's highest plateau and contains the headwaters of Asia's large rivers (e.g., Changjiang, Yellow River, Brahmaputra, Mekong), and thus acts as a water tower providing water resources for irrigation, livelihood, and ecosystem habitat of densely populated countries downstream (Liniger & Weingartner, 1998; Pritchard, 2019; Viviroli & Weingartner, 2008; Yao & Khan, 2022). The Tibetan Plateau (TP) is one of the most sensitive areas in the world to climate change, and amplified rising temperatures and precipitation significantly affect the hydrological cycle and influence future freshwater availability (Hock et al., 2019; Immerzeel et al., 2020). River discharge is a composite indicator that reflects changes in the climatic driving force and cryo-hydrogeological conditions. However, discharge simulation and prediction using process-based hydrological models have long been limited by difficulties in characterizing and parameterizing complex topography with multiple cryo-hydrogeological attributes (e.g., permafrost and glaciers) and calibrating spare streamflow observations (McDonnell et al., 2007). Moreover, parameters from a well-calibrated process-based model in one basin may not be directly applicable to another (Kratzert et al., 2019; Ma et al., 2021). Therefore, a generic modeling approach is needed for characterizing complexity and heterogeneity that can be learned and applied to alpine regions.

Data-driven hydrological models based on machine learning (ML) and deep learning (DL) algorithms have demonstrated a strong ability to simulate and predict

hydrological variables (Chadalawada et al., 2020; Herath et al., 2021; Jiang et al., 2022; Shen, 2018; Shen et al., 2018), including groundwater level predictions (Cai et al., 2022; Cai et al., 2021). In particular, the time-series prediction algorithm of long short-term memory (LSTM), a special type of recurrent neural network, is advantageous in handling rainfall-runoff simulations (Kratzert et al., 2019). LSTM has been employed as an effective tool for predicting river discharge from catchment to basin scales (Feng et al., 2020; Fu et al., 2020). Furthermore, the discharge predicted by LSTM can help determine the processes and obtain a hydrological understanding that process-based models fail to conceptualize. For instance, process-based models can be run and converged if the model domain is in a closed basin and all water balance components are known, whereas LSTM does not have these restrictions (Lees et al., 2021). Thus, LSTM can help modelers determine the water balance. These advantages are particularly important for vast glaciated and permafrost alpine regions, where the understanding of hydrological processes is too limited to apply process-based models.

Currently, most algorithm tests and applications in streamflow prediction using LSTM utilize available catchment datasets with a unified format of meteorological forcing and geographic attributes. Such datasets include the Catchment Attributes and Meteorology for Large-sample Studies (CAMELS), which contains data from 671 catchments across the United States (Addor et al., 2017; Newman et al., 2014), and CAMELS-GB, which includes data from 518 catchments across Great Britain (Lees et al., 2021). Although LSTM can be applied in data-scarce regions, its performance is

constrained by discontinuous observations (Bowes et al., 2019; Liang et al., 2018). Transfer learning (TL), which migrates hydrological knowledge learned from one region to another, has great potential for improving predictions in data-scarce regions (Gauch et al., 2021). By pre-training models on available datasets with abundant observations, LSTM can master the underlying dynamics and subsequently fine-tune the target dataset. Ma et al. (2021) demonstrated improved performance of TL among data-dense (664 basins in Great Britain), moderately data-dense (49 basins in Chile), and data-scarce areas (five basins in China).

Machine learning models are currently used for spatial identification and interpolation in alpine regions (Ding et al., 2016; Tong et al., 2021; Wang et al., 2020; Xia et al., 2018; Xu et al., 2018). However, rainfall-runoff simulations using machine learning models are seldomly realized in ungauged alpine regions because of limited data availability and concerns regarding permafrost, snow, glaciers, and other factors. Our objective was to determine if the LSTM and TL algorithms could serve as models for describing physical processes in the alpine regions, and also to examine their forecast ability with respect to cryospheric attributes (e.g., glaciers, permafrost). Data availability includes issues on data resolution and observed length, which indicate that most streamflow measurements were recorded monthly and only recorded for a few years in alpine regions. The permafrost, snow, and glaciers are geographical attributes with higher spatial uncertainty and largely unknown effects on streamflow.

In this study, we examined the effects of the amount of river discharge

observations and geographical attributes on streamflow prediction based on the data-rich CAMELS dataset. We then integrated TL and LSTM to simulate streamflow in three typical alpine basins with different dominating attributes across the TP region, which allowed us to fill large data gaps: the Yarlung Zangbo Basin in the southern TP joint to the Himalayan mountains, the river discharge of which is impacted by seasonal frozen ground, snow, and glaciers; the headwater basin of the Yellow River in the central TP, the river discharge of which is impacted by continuous permafrost but without glaciers; and the headwater basin of the Heihe River in the northern TP, the river discharge of which is impacted by groundwater. Another watershed, the middle stream of the Yellow River Basin, which is affected by human activities, was selected for comparison with natural alpine regions. In addition to filling data gaps with high accuracy, this study aims to provide insights into runoff simulation and prediction using machine learning and transfer learning in ungauged alpine and cryosphere regions affected by glaciers and permafrost.

## 2. Data and Methods

### 2.1 Data

To examine the effects of data availability and attributes on discharge prediction, the original CAMELS dataset was used (Addor et al., 2017). This dataset comprises 671 natural basins across the United States (Figure 1a). The CAMELS dataset contains 58 static attributes that describe the geographical properties of the basin topography

(e.g., elevation and slope), climatic indices (e.g., mean daily precipitation), hydrological signatures (e.g., mean daily streamflow), land cover, soil, and geology (Addor et al., 2017). Daily climate forcing (e.g., temperature, precipitation, radiation, and humidity), streamflow data between 1984 and 2014 (Newman et al., 2015), and detailed descriptions of the datasets can be obtained in the reference (Newman et al., 2014).

To examine the performance of the integrated LSTM and TL, we selected three representative alpine basins with a total of 18 catchments from north to south of the Tibet-Himalayan region, all of which are regarded as natural without disturbance from human activities. The first is the Qilian Mountains (QM), the headwater region of the Heihe River Basin in the northern margin of the TP (Figure 1c), where mountain block recharge has the same direction as streams to the outlets (Yao et al., 2017). The second is the headwater of the Yellow River Basin (HYRB) in the central TP (Figure 1d). Permafrost accounts for over 87% of this area, and the meltwater from glaciers to streams accounts for less than 2% (Ma et al., 2019); thus, it is an ideal study area to quantify the effects of thawing permafrost on streams without considering the glaciers. The third is the Yarlung Zangbo Basin (YZB) between the TP and the Himalayan mountains (Figure 1e), where the upstream region in the central Himalayas is dry, with an annual precipitation of less than 300 mm covered with continuous permafrost, while the downstream region in the eastern Himalayas is impacted by a monsoon climate with annual precipitation over 1,000 mm and is affected by glaciers (Yao et al., 2021). The

fourth area, which is strongly influenced by human activities (Sun et al., 2020), is the middle stream of the Yellow River Basin in the Loess Plateau (MYRB). MYRB was chosen as a human-impact contrast for the other three alpine regions.

We collected climate data from 30 meteorological stations in 22 catchments between 1956 and 2019 from the National Meteorological Information Center (https://data.cma.cn/), including daily precipitation, evaporation, air and ground temperatures, and radiation. Streamflow data from 20 gauging stations were consolidated between 1956 and 2016. All climate and streamflow data were processed in a unified format according to CAMELS. The Nash-Sutcliffe Efficiency coefficient (NSE) was used to evaluate the goodness-of-fit of streamflow, and the percent bias of the top 2% peak flow range (FHV) and the percent bias of the bottom 30% low flow range (FLV) were used to evaluate the performance in simulating high and low flow.

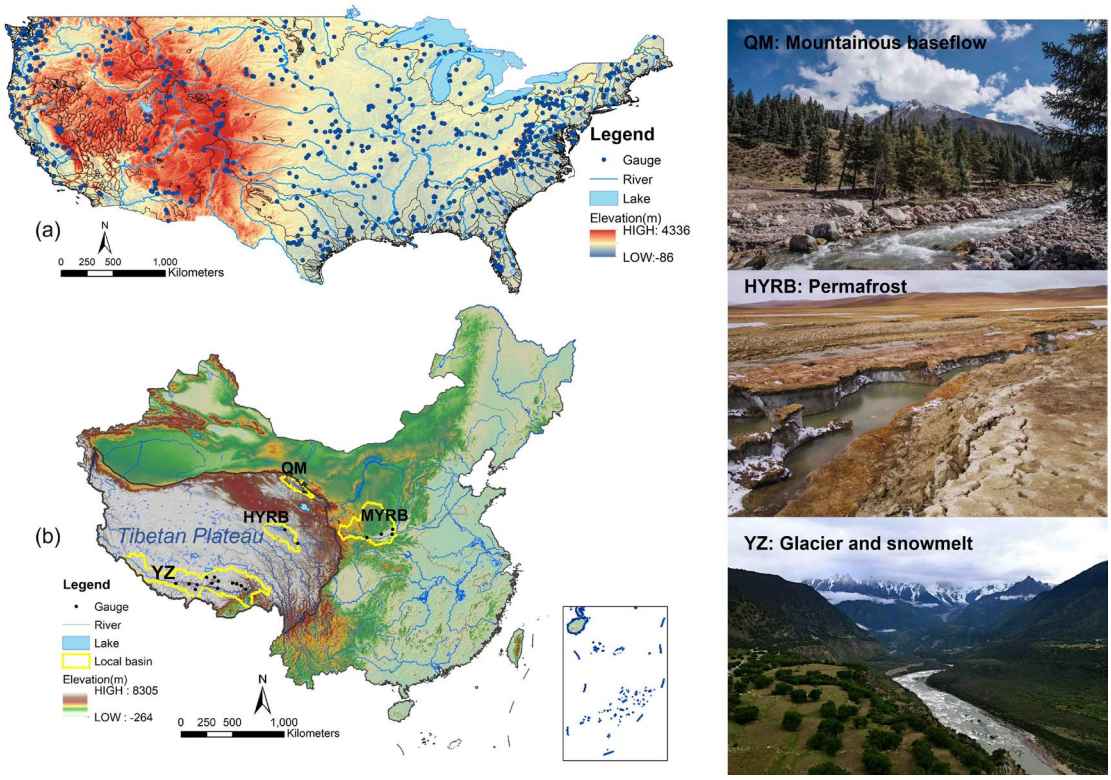



**Figure 1. Study area and topographic features for long short-term memory (LSTM) and transfer learning (TL)**. (a) Map showing the 671 catchments across the entire United States included in the CAMELS dataset. (b) Map of the Chinese river systems in the Tibetan Plateau and three representative watersheds spanning the Tibetan Plateau, including the Qilian Mountains (QM), the headwater of the Yellow River Basin (HYRB), the Yarlung Zangbo Basin (YZB), as well as the middle stream of the Yellow River Basin (MYRB). Photographs on the right show distinct landscapes and hydrogeology characterized by forests and groundwater in the QM, permafrost in the HYRB, and glaciers in the YZB.

## 2.2 Experiments on the effects of data availability and attributes on discharge prediction

In this study, data availability refers to the resolution and length of streamflow data observations. Because a large part of the streamflow measurements and water balance evaluations in alpine regions are only publicly available at a monthly scale, the effectiveness of LSTM based on different resolutions must be determined. We hypothesized that the model performance on a monthly scale would be better than that on a daily scale because the monthly streamflow variations were smoother to fit. To test this hypothesis, we ran experiments on all 671 catchments of the CAMELS dataset at monthly and daily scales. For the monthly simulation, the forcing data resolution was aggregated to monthly. The training period for both the monthly and daily tests was between 1985 and 1995, and the test period was between 1995 and 2005.

The length (observed number of years) of streamflow data is difficult to ensure for alpine regions; however, DL models require large amounts of data. Thus, we expect to evaluate the extent to which the observed streamflow data can support the simulation and attain an acceptable performance. An NSE value of 0.6 can be regarded as the threshold for good performance (Gupta & Kling, 2011). For each catchment in the CAMELS dataset, we identified the length (years) of the streamflow data that resulted in an NSE of 0.6.

Alpine regions are characterized by permafrost, snow, and glaciers, which are akin to the land cover and soil properties in the LSTM model. Although the current datasets

187 do not include these alpine attributes, we can quantify the extent by which these
188 attributes affect the NSE and spatially identify the most sensitive attributes. For each
189 attribute, we ran the experiments using CAMELS data with and without labeling and
190 computed the differences in the NSE. All attributes were tested and the most sensitive
191 attribute (the largest △NSE) was identified for each catchment.
192
193 **2.3 Integrated TL and LSTM**
194 The details of LSTM and TL are shown in Figure 2 (Feng et al., 2020; Ma et al.,
195 2021). LSTM is a recurrent neural network that learns from time-series data, and has
196 "cell states" and "gates" (such as "forget gates", "input gates", and "output gates") in
197 the hidden states retained as weights from previous time steps to determine if these
198 weights can be memorized or forgotten in the next step. Figure 2a shows the
199 architecture of the LSTM method. We concatenated forcing data, attributes, and the
200 observed streamflow $y^*$ as inputs ($I^t$) in the tth time step. $I^t$ was transformed into $x^t$
201 through the input linear layer, then the LSTM unit was entered. The LSTM unit output
202 $h^t$ and the output linear layer transform it into the simulated streamflow ($y^t$). The
203 detailed structure of the LSTM unit is shown in Figure 2b. The equation of the forward
204 pass of the LSTM method is shown in Figure 2c.
205 TL construction included two major steps: pre-training with source data from
206 CAMELS using LSTM and extracting the weights, and then transferring the extracted
207 weights to local simulations using LSTM for the four river basins. The weights were

provided from pre-training CAMELS data for initializing local simulations, and the TL was able to freeze the initialized weights when training with local data or allowed parts of the weights to change depending on local training. In this study, we set three options, TL_A, TL_B, and TL_C, to represent the degree of transfer learning (Figure 2d). The option TL_A transferred all the initialized weights from CAMELS to the local dataset and froze them, except for the weights for the input and output linear layers during the local training. Option TL_B froze the weights from the input linear layer but allowed the weights of the recurrent connections from the hidden states to be updated. Option TL_C allowed all weights transferred from CAMELS to be updated during local training. Options TL_A and TL_C served as the highest and lowest degrees of transfer learning, respectively.

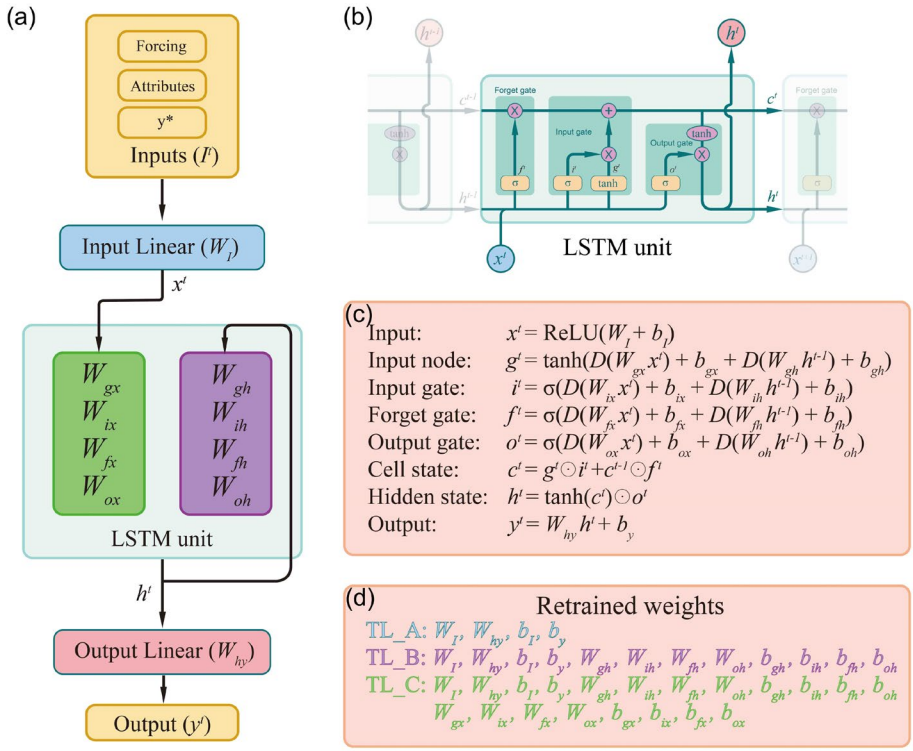



**Figure 2. Structure of the long short-term memory (LSTM) and transfer learning**

**(TL) models**. (a) The structure of the LSTM and TL models included an input linear layer, an LSTM unit, and an output linear layer. (b) Detailed structure of the LSTM unit. (c) The equations of the forward pass of the LSTM model. (d) The weights which were allowed to be updated during the training period in TL_A, TL_B, and TL_C.

We conducted three experiments to test the effects of the source data and evaluate the sensitivity of the local attributes to the performance of the models. For the first experiment, we pre-trained the models using CAMELS data of all 671 basins and selected 531 basins, eliminating basins with substantial area discrepancies in the geospatial fabric among different USGS groups and basins with areas over 2,000 $km^2$ (Addor et al., 2017; Dzhamalov et al., 2017). We then compared the performances using transfer parameters from 671 basins and 531 basins. For the second experiment, we tested the effects of the daily and monthly resolutions of the source data on local performance after transfer. We transferred the parameters from both the daily and monthly resolutions of the source data to the monthly local streamflow predictions for all four basins. In the third experiment, we quantified the sensitivity of the local attributes to the local performance for each sub-basin of the tested basins. The hyperparameters of LSTM and TL were manually adjusted based on previous studies (Feng et al., 2020; Ma et al., 2021). The LSTM hidden size was 256, the number of training epochs was 1000, the Mini-batching size was 128, and the length of training instances were 365 for daily model and 12 for monthly model.

**Table 1. Overview of simulation employed in LSTM and TL models for CAMELS and the four tested basins**

| Model/Parameters | CAMELS | QM | HYRB | YZB | MHYB |
|---|---|---|---|---|---|
| Training period | 1985–1995 | 1990–1995 | 1981–1989 | 1981–1990 | 1971–1982 |
| Testing period | 1995–2005 | 1996–2001 | 2007–2016 | 1991–2000 | 1983–1986 |
| Length of training | 10 years | 6 years | 9 years | 10 years | 12 years |
| Mini-batching size | 128 | 128 | 128 | 128 | 128 |
| LSTM hidden size | 256 | 256 | 256 | 256 | 256 |
| Number of training epochs | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |

## 2.4 Baseflow calculation

Baseflow is part of streamflow that originates from delayed sources such as groundwater. To further evaluate the performances of the LSTM and TL models in areas with groundwater baseflow contributions, we used a digital filtering approach to separate the baseflow from the total river discharge. The algorithm from Eckhardt (Eckhardt, 2005) was used to separate the baseflow of all sub-basins of the four river basins in this study; this algorithm has been demonstrated to have the best performance in 1,815 catchments across the contiguous United States (Xie et al., 2020). The baseflow index (BFI) is the ratio between the baseflow and streamflow. The correlation coefficient (r) between the BFI and NSE of deep learning was computed. The equation used for the Eckhardt method is as follows:

$$b_k = \frac{\alpha(1-B_{max})b_{k-1} + (1-\alpha)B_{max}Q_k}{1-\alpha B_{max}}, \quad (1)$$

where $Q_k$ is the kth total streamflow, $b_k$ is the kth baseflow, α is the recession constant set as 0.975 (Eckhardt, 2008), $B_{max}$ is the max BFI, which was determined according to the field investigation of hydrogeological conditions. This can be estimated by using the following equation:

$$b_{k-1} = \frac{b_k}{\alpha}(b_k \le Q_k). \quad (2)$$

A backward iterative operation based on Equation (2) was performed on the daily streamflow, and $B_{max}$ was obtained by dividing the maximum possible total baseflow by the total streamflow (Xie et al., 2020). For the sub-basins where $B_{max}$ was less than 0.8, we set it to 0.8 to obtain better results (Xie et al., 2020).

# 3. Results

## 3.1 Effects of data availability

Figure 3 shows the results of the comparison between daily and monthly streamflow simulations. The streamflow resolution does not impact the prediction; on average, slightly better predictions were obtained for coarse predictions (Figure 3d), with a median NSE of 0.71 and 0.72 for daily and monthly resolutions, respectively. In spatial NSE, the monthly simulation performs better than daily simulation, particularly in areas represented by higher NSE values (Figure 3e, 22%>19% for NSE over 0.8). Among all basins included in CAMELS, 47% of catchments showed a better performance for monthly simulations (Figure 3f).

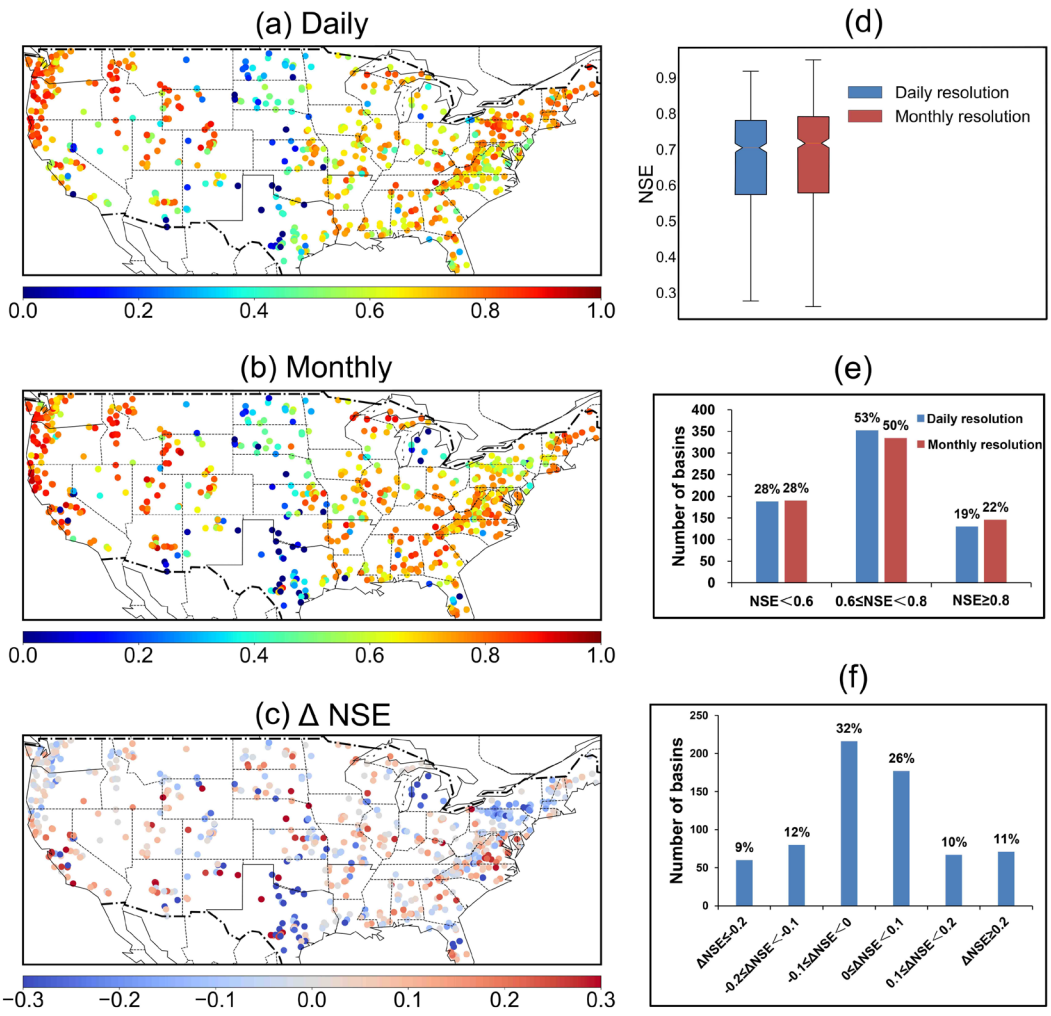
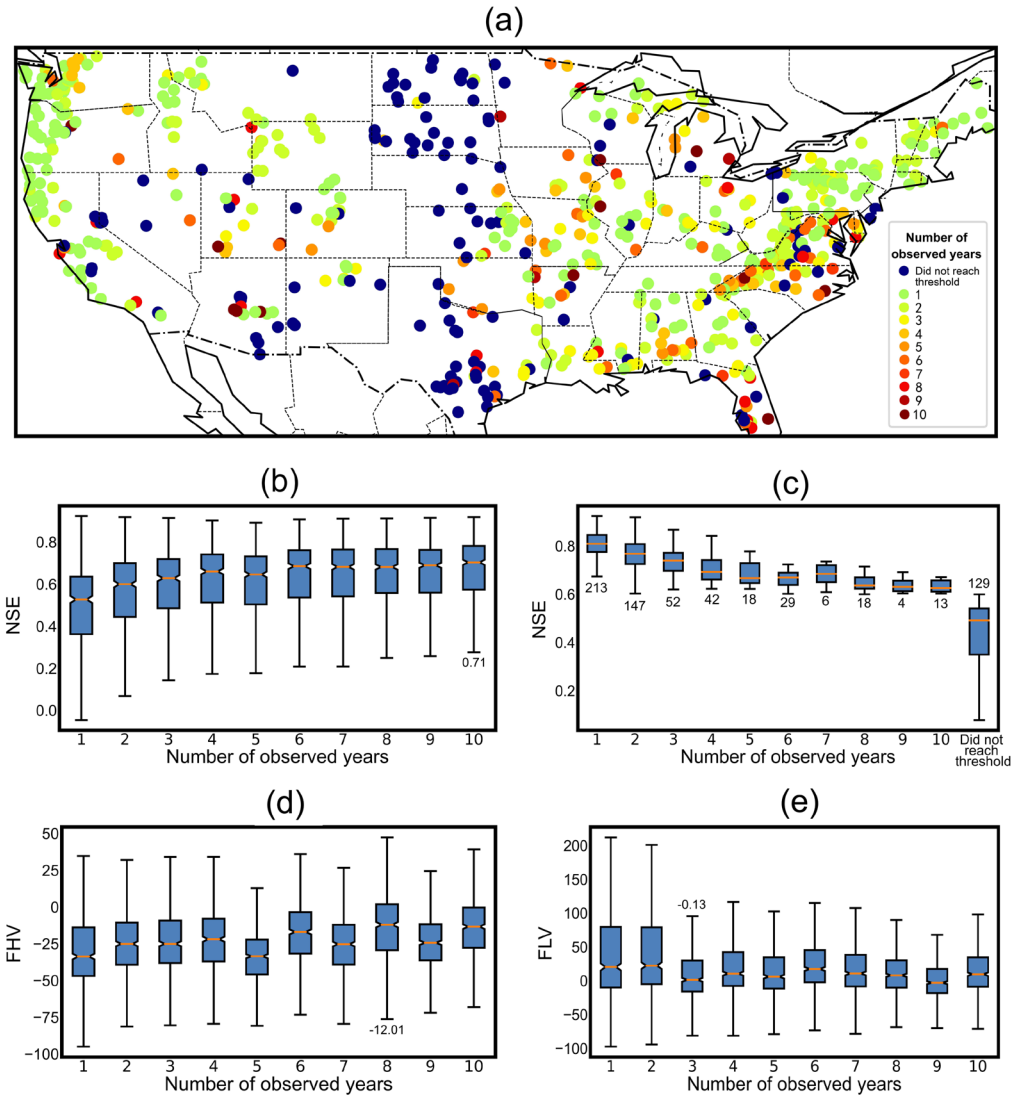



**Figure 3. Comparison of Nash-Sutcliffe Efficiency coefficient (NSE) between daily and monthly observed streamflow records.** (a) Daily NSE; (b) Monthly NSE; (c) △NSE = NSE (Monthly) − NSE (daily); (d) Average NSE for daily and monthly resolution; (e) Comparison of average daily and monthly NSE at different intervals and number of basins; (f) △NSE at different intervals and number of basins.

Figure 4 shows the data requirements (input sequence length and observation data length) for the streamflow simulation at a daily resolution. Our results show that 31.74% of catchments among all the basins included in the CAMELS dataset require only one

285 year of input data to achieve good simulation results (NSE > 0.6), but 19.23% of
286 catchments cannot achieve good simulation results even with ten years of data. For all
287 671 catchments, the simulation results were positively correlated with the time-series
288 length of the input data (Figure 4b). In general, catchments that required fewer years of
289 input data to achieve good simulation results (NSE > 0.6) simulated better results than
290 those that required a longer time-series for input data (Figure 4c). Increasing the number
291 of years of input time-series data also slightly improved the FHV and FLV (Figure 4d
292 and 4e).




293

294 **Figure 4. Effect of time-series length on Nash-Sutcliffe Efficiency coefficient (NSE).**

295 (a) Distribution of length of data series (in years) required to achieve an NSE of 0.6 or

296 higher (When the NSE does not reach 0.6 even after using 10 years of observed data, it

297 is referred to as "did not reach threshold"); (b) Box plots of NSE obtained from

298 simulations using 1 to 10 years of observed data for all watersheds; (c) Box plots for

299 the best performance of selected catchments which required the shortest lengths of

300 observed data that resulted in NSE values greater than 0.6; (d) Effects of length of

observed data on the percent bias of the top 2% peak flow range (FHV); (e) Effects of length of observed data on the percent bias of the bottom 30% low flow range (FLV).

### 3.2 Effect of forcing data and attributes

Figure 5 shows how the forcing data and attributes affected the streamflow simulation. A detailed description of all forcing and attribute parameters is provided in Table S1. Precipitation (prcp) was the most important forcing data of the input data for the simulation results. Model simulations that did not use precipitation as input data reduced the median NSE by 0.39 (Figure 5a). The overall effect of rainfall on runoff simulation performance increased from west to east. $\triangle NSE < 0.6$ was observed in parts of the southwestern (i.e., New Mexico) and central (i.e., Kansas) regions. Figure 5b shows the distribution of $\triangle NSE$, between simulations with and without the most sensitive driving forcing and attribute for each catchment, excluding precipitation. In addition to rainfall, day length (dayl) is another sensitive factor that resulted in median NSE changing from 0.71 to 0.69, especially in the central United States. With the exception of precipitation, the lack of input had little effect on the simulation results (Figure 5c). Figure 5d shows the number of catchments to which the forcing data or attributes were the most sensitive. Combined with Figure 5a and 5b, Figure 5d shows that land surface and hydrogeological attributes as inputs (e.g., geol, soil conductivity, and porosity) have a broad influence on catchments in the east and west coast regions. Using these inputs can obtain better NSE even with less data, although the magnitude

of this influence is insignificant, with median NSE changing from 0.71 to 0.68. The absence of precipitation data significantly underestimated the simulation of FHV but had little effect on FLV (Figure 5e and 5f).

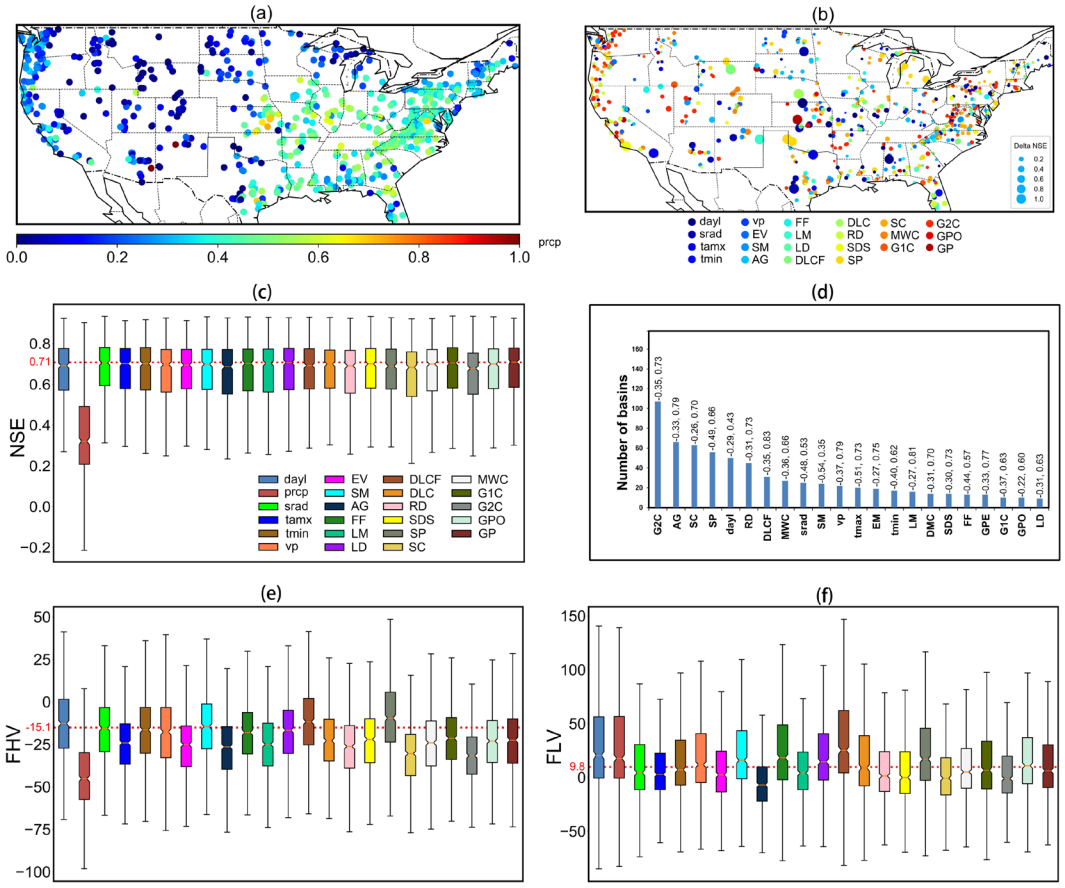



**Figure 5. Effect of climate driving forcing and attributes on simulation performance.** (a) Distribution of △Nash-Sutcliffe Efficiency coefficient (NSE) between simulations with and without precipitation input; (b) Distribution of △NSE between simulations with and without most sensitive attributes for each catchment; (c) Box plots of NSE without each indicated attribute; (d) Number of catchments that are most affected by an attribute; (e) Box plots of the percent bias of the top 2% peak flow range (FHV) without the indicated attribute; (f) Box plots of the percent bias of the

bottom 30% low flow range (FLV) without the indicated attribute. The red dashed lines in the (c), (e), and (f) plots indicate the averaged NSE, FHV, and FLV that contain all drives and attribute inputs as a contrast.

### 3.3 Performance of integrated TL and LSTM in alpine regions

We evaluated the simulated performances of LSTM and TL using a combination of three options (TL_A, TL_B, and TL_C) and two types of datasets (all 671 catchments and 531 selected catchments from CAMELS), and the results are shown in Figure 6. The best performances of QM, HYRB, YZB, and MYRB for monthly streamflow simulations had maximum NSE of 0.80, 0.31, 0.82, and 0.64, respectively, corresponding to the TL_B, TL_B, TL_A, and TL_A options, respectively. Except for YZB, the best performances for the other basins were obtained by transferring the selected 531 catchments instead of using the entire dataset. This suggests that a better overall performance for alpine basins is achieved by selecting better simulated original catchments for TL simulations in the tested basins and by directly using the original data (i.e., initialized weights) for the TL simulation. However, the simulation performances for FHV and FLV showed a different trend than that of the NSE. The maximum NSE of QM and HYRB for daily streamflow simulations decreased to 0.65 and 0.41, respectively, which was worse than the monthly simulation. The best performance for both the QM and HYRB was obtained using option TL_A on the 531 selected catchments.

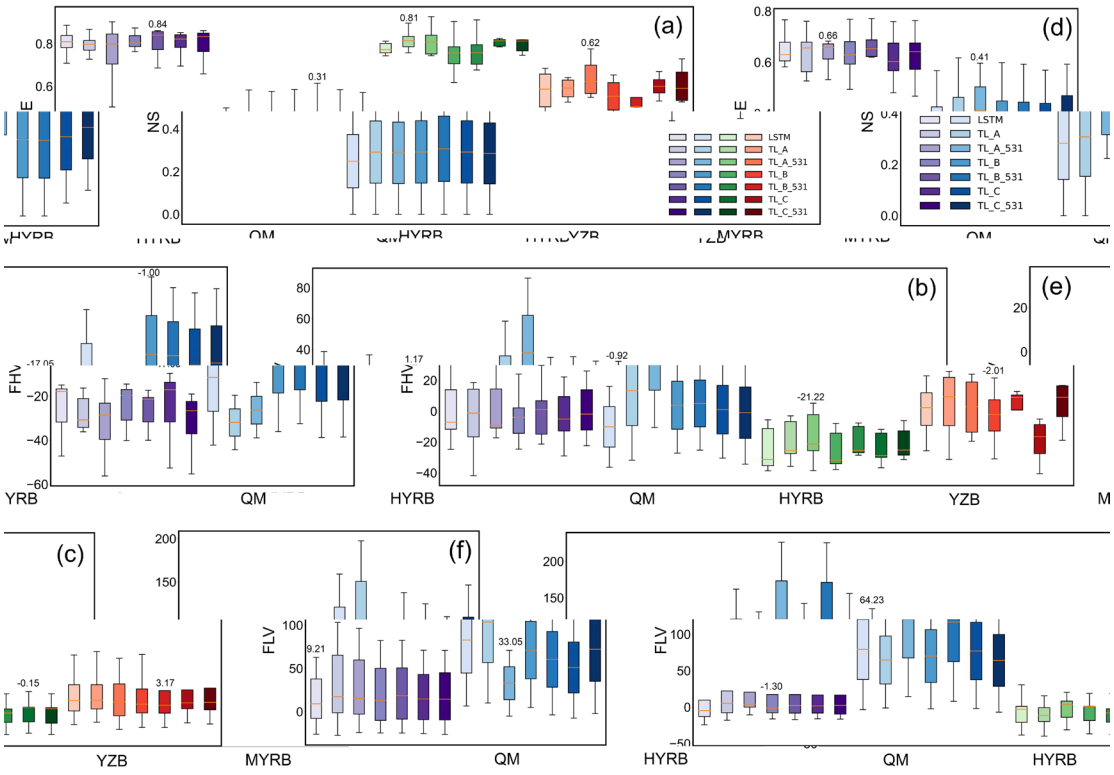



**Figure 6. Simulation performance of the long short-term memory (LSTM) and transfer learning (TL) models on four test river basins.** (a) Box plots of Nash-Sutcliffe Efficiency coefficient (NSE) for monthly streamflow simulations of the Qilian Mountains (QM), the headwater of the Yellow River Basin (HYRB), the Yarlung Zangbo Basin (YZB) and the middle stream of the Yellow River Basin (MYRB) under LSTM different TL options; (b) Box plots of the percent bias of the top 2% peak flow range (FHV) of monthly streamflow simulations; (c) Box plots of the percent bias of the bottom 30% low flow range (FLV) of monthly streamflow simulations; (d), (e), and (f) Box plots of NSE, FHV and FLV of daily streamflow simulations for QM and HYRB. The values shown indicate the best simulation results among the LSTM and six TL options.

Furthermore, we examined the spatial pattern of the performance of the streamflow simulations for the sub-basins of the four tested basins, and the results are shown in Figure 7 and Table 2. The mean NSE of the monthly and daily simulated streamflow of all sub-basins using the LSTM models alone were 0.60 and 0.46, respectively, which showed a poorer performance than TL models with monthly and daily NSE of 0.76 and 0.61, respectively. The NSE of HYRB was less than 0.6, which was caused by dried-up conditions upstream; therefore, no color is shown in Figure 7(b) for the monthly results in the pattern. Overall, for all the tested basins, the streamflow simulation for the downstream sub-basins was somewhat better than that for the middle and upper reaches. The simulated performance of streamflow in the alpine basins of QM and YZB is apparently better than that of the MYRB, which is disturbed by human activities.

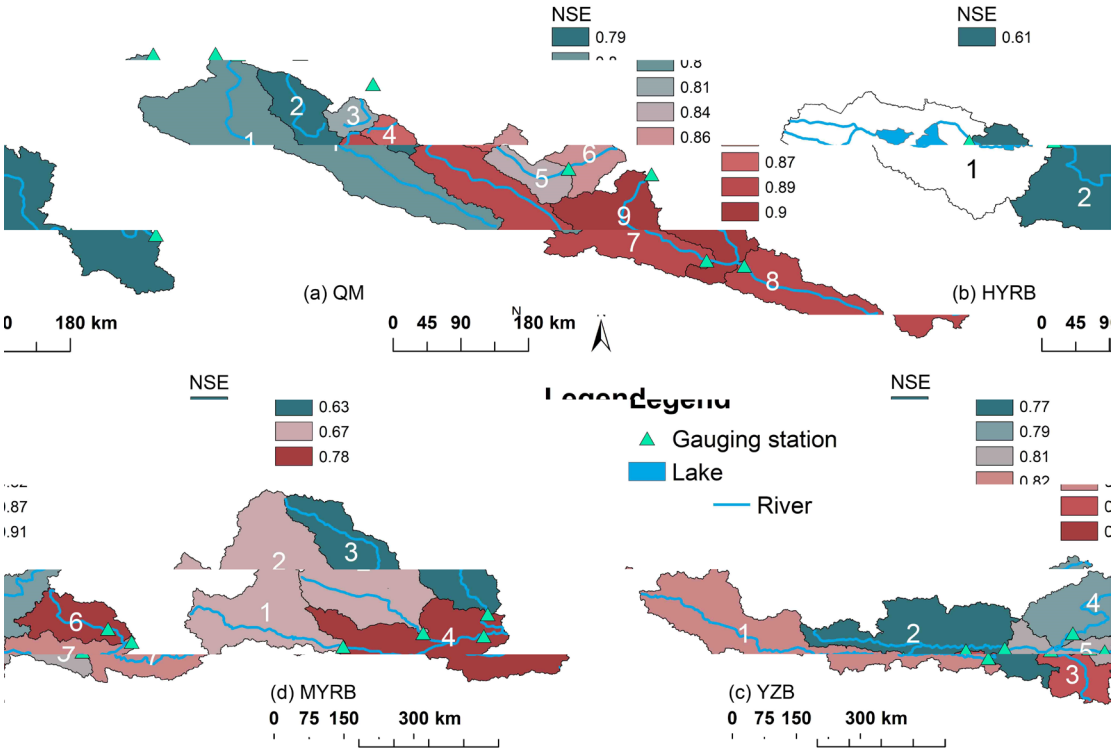


**Figure 7. Nash-Sutcliffe Efficiency coefficient (NSE) patterns for sub-basins**. (a)–(d) NSE patterns in the Qilian Mountains (QM), the headwater of the Yellow River Basin (HYRB), the Yarlung Zangbo Basin (YZB) and the middle stream of the Yellow River Basin (MYRB). The numbers identify different sub-basins from upstream to downstream, and the colors indicate the maximum NSE value for each sub-basin. The NSE of subbasin 1 in HYRB, indicated by the white color, was below 0.6.

**Table 2. Maximum NSE value (best performance) for each sub-basin by monthly and daily resolution**

| Basin | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| QM (monthly) | 0.80 | 0.79 | 0.81 | 0.87 | 0.84 | 0.86 | 0.89 | 0.89 | 0.90 |
| HYRB (monthly) | 0 | 0.61 | | | | | | | |
| YZB (monthly) | 0.82 | 0.77 | 0.87 | 0.79 | 0.81 | 0.91 | 0.82 | | |
| MYRB (monthly) | 0.67 | 0.67 | 0.63 | 0.78 | | | | | |
| QM (daily) | 0.63 | 0.53 | 0.68 | | 0.69 | | 0.69 | 0.66 | 0.77 |
| HYRB (daily) | 0.22 | 0.62 | | | | | | | |

We also compared the difference in the performance of streamflow simulations using the TL and Soil & Water Assessment Tool (SWAT) (Figure 8). The outlet streamflows at the QM (Yingluoxia Station) (Zhang et al., 2016) and YZB (Nuxia Station) simulated by SWAT (Deng et al., 2023) were collected for comparison with our results during the test period. The TL_A_531 and TL_C_531 models provided the best predictions of streamflow for the Yingluoxia and Nuxia stations, with the maximum NSE. Comparing the results from the TL model to SWAT, it can be seen that

the TL simulation is generally similar to SWAT but more effective for low flow, as shown by the FLV values.

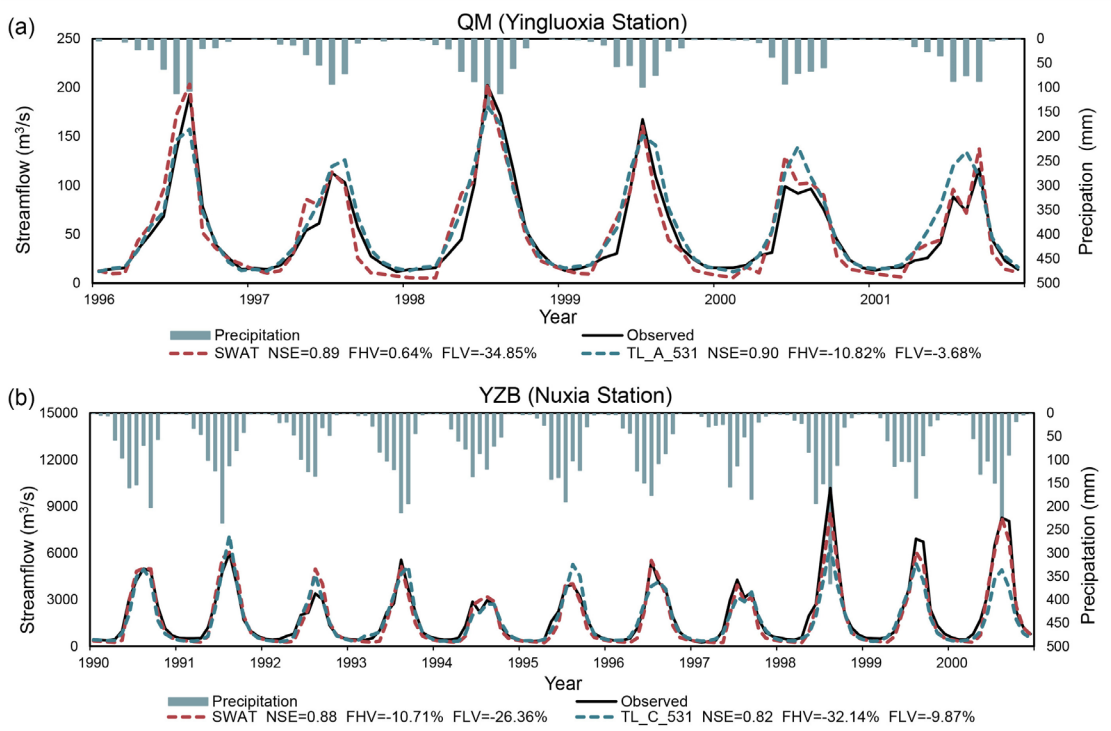



**Figure 8. Observed and simulated streamflow by transfer learning (TL) and Soil & Water Assessment Tool (SWAT) models.** (a) Qilian Mountains (QM) Yingluoxia station; (b) Yarlung Zangbo Basin (YZB) Nuxia station.

### 3.4 Factors controlling streamflow simulation efficiency by DL and TL in the TP

In Figure 9, we show how the most sensitive sub-basin parameters differ and how they impact runoff simulations for the three alpine headwater basins and one middle-stream oasis basin. In each sub-basin, the TL model with the best fit was retrained with the inclusion and removal of a parameter, and the most sensitive parameter with the greatest effect on the NSE was then determined. Temperature (tmin, tmax) was the

primary factor affecting the runoff simulation performance for QM in the northern TP and HYRB in the northeast. In the QM, climate forcing on temperature affects runoff simulations, with △NSE ranging from 0.06 to 0.27, while in the lower sub-basins of the HYRB, the temperature affected △NSE by 0.1. In YZB in the southern portion of the TP, the main parameters affecting the basins were soil classification (OF), soil conductivity (SC), and permafrost area fraction (APF). Outside the TP, land use and precipitation were the main factors affecting the runoff simulations for MYRB, with precipitation affecting △NSE by greater than 0.8.

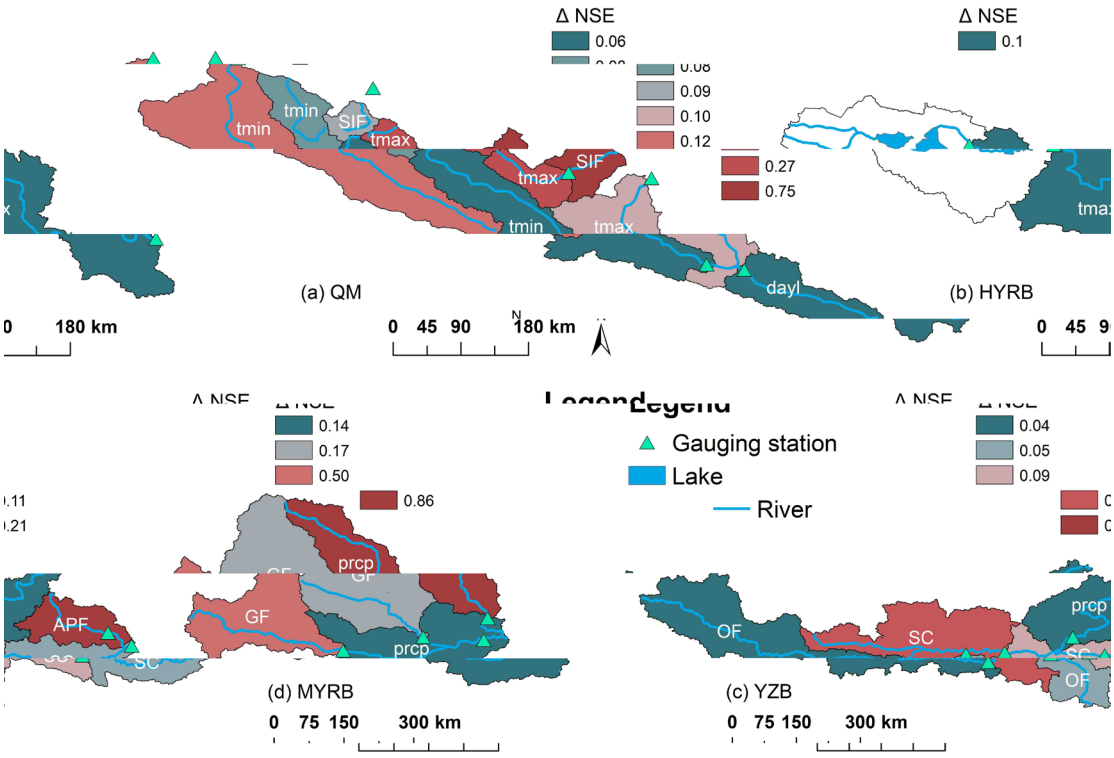



**Figure 9. △Nash-Sutcliffe Efficiency coefficient (NSE) patterns observed for the most sensitive factor among climate forcing and attribute parameters in each sub-basin.** (a)–(d) △NSE patterns in the Qilian Mountains (QM), the headwater of the Yellow River Basin (HYRB), the Yarlung Zangbo Basin (YZB) and the middle stream

of the Yellow River Basin (MYRB). The abbreviations shown represent the parameters that affected runoff simulations the most for each sub-basin. △NSE is indicated by color.

The NSE of the LSTM and various TL models were calculated for each of the four tested basins, and the BFI was calculated for each sub-basin based on the observed data. Figure 10 shows that NSE value obtained by LSTM and TL models is negatively correlated with BFI in general ($r_1$ = −0.56). This negative relationship was more pronounced in areas where precipitation was greater than 40 mm ($r_2$ = −0.87) (Figure 10 (a)). The performances of the LSTM and TL models were significantly and negatively correlated with the BFI in HYRB, YZB, and MYRB, which are glacier-dominated, permafrost-dominated, and human activity-dominated basins, respectively, as shown in Figure 10 (b). The results indicate that LSTM and TL perform well for areas with rainfall-runoff mechanisms and are less effective in areas in which groundwater serves as a controlling factor in generating runoff.

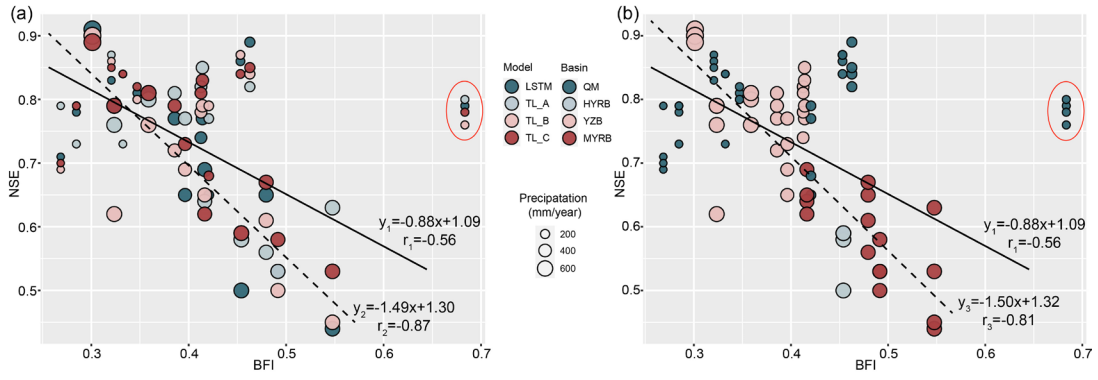



**Figure 10. The relationship between baseflow index (BFI) and Nash-Sutcliffe Efficiency coefficient (NSE).** (a) The different colors represent the long short-term

memory (LSTM) and three transfer learning (TL) models; $y_1$ represents the linear fit to all data points except for the red circled points; $y_2$ represents the linear fit to data points with precipitation greater than 400 mm. (b) The different colors represent the four river basins; $y_1$ represents the linear fit to all data points except for the red circled points; $y_2$ represents the linear fit to data points excluding those from the Qilian Mountains (QM). The diameter of the circle represents the amount of precipitation. HYRB, headwater of the Yellow River Basin; YZB, Yarlung Zangbo Basin; MYRB, middle stream of the Yellow River Basin.

## 4. Discussion

Because there is a limited number of gauges and poor data resolution of river discharges in the alpine region, we assessed how these factors might affect streamflow simulation performance by models using the CAMELS dataset. By separately training daily and monthly streamflow data, we found that data resolution had a negligible effect on the runoff simulation for 70% of the catchments. When the discharge observations are limited and the simulation aims to evaluate the annual or seasonal water balance, DL models can be quickly trained using monthly streamflow data (at least 10 times faster than training with daily streamflow) to obtain better performance. Our validation of length of streamflow data revealed that more than a third of the total catchments can achieve an NSE of more than 0.6 using only one third of the observed streamflow as

training data, and the NSE could be further improved as more data becomes available. In contrast, nearly 20% of the catchments did not achieve acceptable simulation results (NSE < 0.6), despite being trained on all of the available data. This is inconsistent with the common perception that a data-driven model depends primarily on data and that more data can enhance the simulation. Gauch et al. (2021) demonstrated that multi-basin training is important instead of using only long periods of data. Our results further demonstrate that the number of data required depends on the hydrological characteristics of the catchments, namely, how closely they conform to a general rainfall-runoff mechanism. Additionally, the length of the data series does not have a significant effect on the peak and low flows, with overall trends of low peak flow simulation and slightly higher low flow simulation.

Using both climate-driven data and geographical attributes, we tested the impact of these different input parameters on the runoff simulations. Rainfall is the driving force that has the greatest impact on results of the runoff simulations. Precipitation affected the runoff simulations mainly at peak flows (25% reduction in the average simulation performance) and had no significant impact at low flows, as reflected by DL. This is in accordance with the hydrological rule that runoff is primarily derived from rainfall during the wet season, whereas the dry season runoff primarily originates from groundwater discharge as baseflow (Dzhamalov et al., 2017; Yao et al., 2021). In addition to precipitation, subsurface characteristics, including geological setting, soil properties, and land use/vegetation cover impact the runoff simulation performance in

474 over one-third of the catchments. The hydrological conductivity and porosity of the soil
475 are similar parameters that impact runoff simulation by influencing surface infiltration
476 and lateral flow. Previous studies have shown the importance of the storage capacity
477 and hydraulic conductivity of unsaturated soil and saturated aquifers using physical-
478 based hydrological models at the regional scale or site experiments at the field scale
479 (Mirbabaei et al., 2021; Vannier et al., 2016; Vannier et al., 2014). Our study reveals the
480 extent to which various subsurface characteristics affect runoff production on a
481 subcontinental scale. Thus, although the data-driven model does not incorporate
482 physical processes, the results obtained from a large amount of data can be used to
483 characterize physical processes.

484 Because of the scarcity of data in alpine regions, simulations and predictions of
485 river discharge are commonly used by physical models and are sometimes combined
486 with data-driven models. By transferring the trained parameters from the selected
487 catchments of CAMELS to the test basins, we found that the integrated LSTM and TL
488 was more effective in the alpine river basins, particularly for those that had not dried
489 up (i.e., QM and YZB). It has been shown that TL is effective for simulating discharge
490 in alpine basins with cryospheric characteristics with some degree of confidence. We
491 also observed that the driving elements played a dominant role in simulations of
492 cryospheric river discharge, whereas the attribute elements had a limited impact. In
493 alpine river discharge modeling, confusion often arises from the effects of glaciers and
494 permafrost. Climate and geological attributes are difficult to transfer to basins without

the presence of these elements. Physical-based models and field experiments show that glaciers provide roughly 35% of the annual total discharge of the Panj River basin as well as effectively buffering runoff both in wet and dry years (Pohl et al., 2017). Moreover, the permafrost coverage of a catchment can have a significant effect on river runoff, and subsurface flows can be as rapid as surface flow through highly conductive peat layers and soil pipes (Quinton & Marsh, 1999; Song et al., 2020). In the TP, approximately 66–86% of total spring runoff is derived from thawing of surface soil, inducing saturation-excess runoff, while 14–34% comes from subsurface interflow. In response to rising soil temperatures, the active layer has a significant effect on the pattern and composition of runoff generation (Genxu et al., 2017). The DL approach does not directly explore these processes, but reacts to their mechanisms through parameter sensitivity in different sub-basins. In the QM and HYRB, where permafrost has a strong influence, temperature is an important driver of runoff, as shown in Figure 9. With glacial influence, soil classification and infiltration become the main factors influencing the discharge in the YZB. Furthermore, a negative correlation was found between the effectiveness of DL and TL and the magnitude of groundwater contribution to river discharge. This indicates that the effectiveness of the DL and TL models is limited in areas in which groundwater is an important source of river discharge. Therefore, the DL and TL models can be combined with more accurate climate forcing data to improve the accuracy of runoff simulations, while attribute factors can be mined to provide mechanistic insight.

The use of DL and TL for hydrological simulation in the cryosphere has not yet been fully adopted; however, it is foreseen to become an important method for mechanism exploration and quick hydrological prediction in the near future, as more high-precision forcing data become available. Meanwhile, the establishment of further systematic hydrological observatories is required to obtain more flow data for future hydrological studies in alpine basins. It is also essential to develop a deeper understanding of flow mechanisms using physical models. Based on this knowledge, it is possible to promote the integration of data-driven and physical models.

## 5. Conclusions

To determine whether TL can be used to predict streamflow in alpine areas and the mechanistic information it can provide, we first examined the effects of the resolution and length of river discharge observations on the training model, and the effect of geographic attributes on streamflow predictions based on the data-rich CAMELS dataset. We then integrated TL and LSTM to simulate the streamflow in three typical alpine basins across the TP and a non-alpine basin for comparison. This study revealed the following major findings:

1. Streamflow simulations were slightly affected by the resolution of discharge observations for training model in 70% of the catchments. The quick processes of training and prediction by DL using monthly streamflow data resulted in a slightly better performance than daily simulations. The

requirements of river discharge measurements for training the DL model depend on the hydrological characteristics of the catchments, particularly how closely they conform to a general rainfall-runoff mechanics.

2. Rainfall as an input variable had the largest impact on runoff simulation results. As evidenced by DL, rainfall had a significant impact on the simulated runoff simulations only at peak flows (25% reduction in average simulation performance). In addition to rainfall, subsurface characteristics such as geological setting, soil properties, and land use/vegetation cover influence runoff simulation performance in over one-third of the catchments. Although the data-driven model does not incorporate physical processes, the results generated from a large amount of data can be used to inform the physical processes.
3. The combination of LSTM and TL is effective for simulating discharge in alpine river basins, particularly for those that do not have dried up streams (e.g., QM and YZB). With some degree of confidence, TL is effective in modelling alpine basins that exhibit cryospheric characteristics. Simulations of cryospheric runoff rely heavily on climate-forcing data, whereas attribute elements play a less important role. The effectiveness of DL and TL models is limited in areas where groundwater contributes significantly to river discharge. Consequently, the DL approach does not explore these processes directly but rather reacts to them based on parameter sensitivity in different

sub-basins.

DL and TL have been demonstrated to be useful for predicting river discharge in alpine basins, even those that involve glaciers and permafrost. For alpine regions where precipitation rather than groundwater dominates runoff, river discharge simulations can be effectively improved by using high-resolution driving climate data. Our study provides an updated understanding of the effectiveness and future prospects of DL and TL for simulating hydrological variations in the North and South Pole regions and mountainous alpine regions in the context of climate warming.

## Acknowledgements

Xingxing Kuang, Yingying Yao, and Chunkun Liu received support from the National Natural Science Foundation of China (Grants No.92047202, No. 41901023 and No. 42107096). We thank the two anonymous reviewers whose comments helped to improve the quality of the manuscript.

## Data Availability Statement

The CAMELS test dataset was obtained from https://gdex.ucar.edu/dataset/camels.html (Addor et al., 2017; Newman et al., 2014). Daily climate datasets provided by the National Meteorological Information Center (http://data.cma.cn/en) were used in this study. River discharge observations on the Tibetan Plateau were obtained from the Hydrology and Water Resources Bureau of the

Tibet Autonomous Region through a restricted use agreement as per government regulations. For those interested in using observational streamflow data, the same procedure must be followed.

## Supplementary Material

Table S1. Summary of the forcing and attribute variables that were used as input to the LSTM and TL models

Table S2. NSE, FHV and FLV values of all 22 subbasins for different LSTM and TL models

## References


Addor, N., Newman, A. J., Mizukami, N., et al. (2017). The CAMELS data set: catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences, 21*(10), 5293-5313. doi:10.5194/hess-21-5293-2017

Bowes, B. D., Sadler, J. M., Morsy, M. M., et al. (2019). Forecasting groundwater table in a flood prone coastal city with long short-term memory and recurrent neural networks. *Water, 11*(5), 1098.

Cai, H., Liu, S., Shi, H., et al. (2022). Toward improved lumped groundwater level predictions at catchment scale: Mutual integration of water balance mechanism and deep learning method. *Journal of Hydrology, 613*, 128495.

Cai, H., Shi, H., Liu, S., et al. (2021). Impacts of regional characteristics on improving the accuracy of groundwater level prediction using machine learning: The case of central eastern continental United States. *Journal of Hydrology: Regional Studies, 37*, 100930.

Chadalawada, J., Herath, H., & Babovic, V. (2020). Hydrologically informed machine learning for rainfall-runoff modeling: A genetic programming-based toolkit for automatic model induction. *Water Resources Research, 56*(4), e2019WR026933.

Deng, Y., Yao, Y., Zhao, Y., et al. (2023). Impact of climate change on the long-term

water balance in the Yarlung Zangbo Basin. *Frontiers in Earth Science, 11*, 380.

Ding, J., Li, F., Yang, G., et al. (2016). The permafrost carbon inventory on the Tibetan Plateau: a new evaluation using deep sediment cores. *Global Change Biology, 22*(8), 2688-2701.

Dzhamalov, R., Safronova, T., & Telegina, E. (2017). Annual distribution of river runoff with estimated contribution of winter low-water season. *Water Resources, 44*(6), 785-792.

Eckhardt, K. (2005). How to construct recursive digital filters for baseflow separation. *Hydrological Processes, 19*(2), 507-515.

Eckhardt, K. (2008). A comparison of baseflow indices, which were calculated with seven different baseflow separation methods. *Journal of Hydrology, 352*(1-2), 168-173.

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). doi:10.1029/2019wr026793

Fu, M., Fan, T., Ding, Z. a., et al. (2020). Deep learning data-intelligence model based on adjusted forecasting window scale: application in daily streamflow simulation. *IEEE Access, 8*, 32632-32651.

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. doi:10.1016/j.envsoft.2020.104926

Genxu, W., Tianxu, M., Juan, C., et al. (2017). Processes of runoff generation operating during the spring and autumn seasons in a permafrost catchment on semi-arid plateaus. *Journal of Hydrology, 550*, 307-317.

Gupta, H. V., & Kling, H. (2011). On typical range, sensitivity, and normalization of Mean Squared Error and Nash-Sutcliffe Efficiency type metrics. *Water Resources Research, 47*(10 %@ 0043-1397).

Herath, H. M. V. V., Chadalawada, J., & Babovic, V. (2021). Hydrologically informed machine learning for rainfall–runoff modelling: towards distributed modelling. *Hydrology and Earth System Sciences, 25*(8), 4373-4401.

Hock, R., Rasul, G., Adler, C., et al. (2019). *High Mountain Areas: In: IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*. Retrieved from

Immerzeel, W. W., Lutz, A. F., Andrade, M., et al. (2020). Importance and vulnerability of the world's water towers. *Nature, 577*(7790), 364-369. doi:10.1038/s41586-019-1822-y

643 Jiang, S., Zheng, Y., Wang, C., et al. (2022). Uncovering flooding mechanisms across
644 the contiguous United States through interpretive deep learning on representative
645 catchments. *Water Resources Research, 58*(1), e2021WR030185.

646 Kratzert, F., Klotz, D., Shalev, G., et al. (2019). Benchmarking a catchment-aware
647 long short-term memory network (LSTM) for large-scale hydrological modeling.
648 *Hydrology and Earth System Sciences Discussions*, 1-32.

649 Lees, T., Buechel, M., Anderson, B., et al. (2021). Benchmarking data-driven rainfall–
650 runoff models in Great Britain: a comparison of long short-term memory
651 (LSTM)-based models with four lumped conceptual models. *Hydrology and*
652 *Earth System Sciences, 25*(10), 5517-5534.

653 Liang, C., Li, H., Lei, M., et al. (2018). Dongting lake water level forecast and its
654 relationship with the three gorges dam based on a long short-term memory
655 network. *Water, 10*(10), 1389.

656 Liniger, H., & Weingartner, R. (1998). Mountains and freshwater supply. *Unasylva*
657 *(FAO)*.

658 Ma, K., Feng, D., Lawson, K., et al. (2021). Transferring Hydrologic Data Across
659 Continents – Leveraging Data-Rich Regions to Improve Hydrologic Prediction
660 in Data-Sparse Regions. *Water Resources Research, 57*(5).
661 doi:10.1029/2020wr028600

662 Ma, Q., Jin, H., Victor F, B., et al. (2019). Impacts of degrading permafrost on
663 streamflow in the source area of Yellow River on the Qinghai-Tibet Plateau,
664 China. *Advances in Climate Change Research, 10*(4), 225-239.

665 McDonnell, J. J., Sivapalan, M., Vaché, K., et al. (2007). Moving beyond
666 heterogeneity and process complexity: A new vision for watershed hydrology.
667 *Water Resources Research, 43*(7). doi:10.1029/2006wr005467

668 Mirbabaei, S. M., Shabanpour, M., van Dam, J., et al. (2021). Observation and
669 simulation of water movement and runoff in a coarse texture water repellent soil.
670 *Catena, 207*, 105637.

671 Newman, A., Sampson, K., Clark, M., et al. (2014). A large-sample watershed-scale
672 hydrometeorological dataset for the contiguous USA. *UCAR/NCAR, doi, 10*,
673 D6MW2F4D.

674 Newman, A. J., Clark, M. P., Sampson, K., et al. (2015). Development of a large-
675 sample watershed-scale hydrometeorological data set for the contiguous USA:
676 data set characteristics and assessment of regional variability in hydrologic
677 model performance. *Hydrology and Earth System Sciences, 19*(1), 209-223.
678 doi:10.5194/hess-19-209-2015

679 Pohl, E., Gloaguen, R., Andermann, C., et al. (2017). Glacier melt buffers river runoff
680 in the P amir M ountains. *Water Resources Research, 53*(3), 2467-2489.

681 Pritchard, H. D. (2019). Asia's shrinking glaciers protect large populations from
682 drought stress. *Nature, 569*(7758), 649-654. doi:10.1038/s41586-019-1240-1

683 Quinton, W., & Marsh, P. (1999). A conceptual framework for runoff generation in a
684 permafrost environment. *Hydrological Processes, 13*(16), 2563-2581.

685 Shen, C. (2018). A transdisciplinary review of deep learning research and its
686 relevance for water resources scientists. *Water Resources Research, 54*(11),
687 8558-8593.

688 Shen, C., Laloy, E., Elshorbagy, A., et al. (2018). HESS Opinions: Incubating deep-
689 learning-powered hydrologic science advances as a community. *Hydrology and*
690 *Earth System Sciences, 22*(11), 5639-5656. doi:10.5194/hess-22-5639-2018

691 Song, C., Wang, G., Mao, T., et al. (2020). Linkage between permafrost distribution
692 and river runoff changes across the Arctic and the Tibetan Plateau. *Science China*
693 *Earth Sciences, 63*(2), 292-302.

694 Sun, P., Wu, Y., Gao, J., et al. (2020). Shifts of sediment transport regime caused by
695 ecological restoration in the Middle Yellow River Basin. *Science of The Total*
696 *Environment, 698*, 134261. doi:https://doi.org/10.1016/j.scitotenv.2019.134261

697 Tong, C., Wang, H., Magagi, R., et al. (2021). Spatial Gap-Filling of SMAP Soil
698 Moisture Pixels Over Tibetan Plateau via Machine Learning Versus Geostatistics.
699 *IEEE journal of selected topics in applied earth observations and remote*
700 *sensing, 14*, 9899-9912.

701 Vannier, O., Anquetin, S., & Braud, I. (2016). Investigating the role of geology in the
702 hydrological response of Mediterranean catchments prone to flash-floods:
703 Regional modelling study and process understanding. *Journal of Hydrology, 541*,
704 158-172.

705 Vannier, O., Braud, I., & Anquetin, S. (2014). Regional estimation of catchment-scale
706 soil properties by means of streamflow recession analysis for use in distributed
707 hydrological models. *Hydrological Processes, 28*(26), 6276-6291.

708 Viviroli, D., & Weingartner, R. (2008). “Water towers”—A global view of the
709 hydrological importance of mountains. In *Mountains: Sources of water, sources*
710 *of knowledge* (pp. 15-20): Springer.

711 Wang, T., Yang, D., Yang, Y., et al. (2020). Permafrost thawing puts the frozen carbon
712 at risk over the Tibetan Plateau. *Science Advances, 6*(19), eaaz3513.

713 Xia, J., Ma, M., Liang, T., et al. (2018). Estimates of grassland biomass and turnover
714 time on the Tibetan Plateau. *Environmental Research Letters, 13*(1), 014020.

715 Xie, J., Liu, X., Wang, K., et al. (2020). Evaluation of typical methods for baseflow
716 separation in the contiguous United States. *Journal of Hydrology, 583*, 124628.

717 Xu, Y., Knudby, A., Shen, Y., et al. (2018). Mapping monthly air temperature in the
718 Tibetan Plateau from MODIS data based on machine learning methods. *IEEE*
719 *journal of selected topics in applied earth observations and remote sensing,*
720 *11*(2), 345-354.

721 Yao, Y., & Khan, Z. A. (2022). Predicting Pakistan's next flood. *Science, 378*(6619),
722 483-483. doi:doi:10.1126/science.ade7824

723 Yao, Y., Zheng, C., Andrews, C., et al. (2017). What Controls the Partitioning between
724 Baseflow and Mountain Block Recharge in the Qinghai-Tibet Plateau?
725 *Geophysical Research Letters, 44*(16), 8352-8358.

726 Yao, Y., Zheng, C., Andrews, C. B., et al. (2021). Role of Groundwater in Sustaining
727 Northern Himalayan Rivers. *Geophysical Research Letters, 48*(10).
728 doi:10.1029/2020gl092354

729 Zhang, A., Liu, W., Yin, Z., et al. (2016). How will climate change affect the water
730 availability in the Heihe River Basin, Northwest China? *Journal of*
731 *Hydrometeorology, 17*(5), 1517-1542.
732