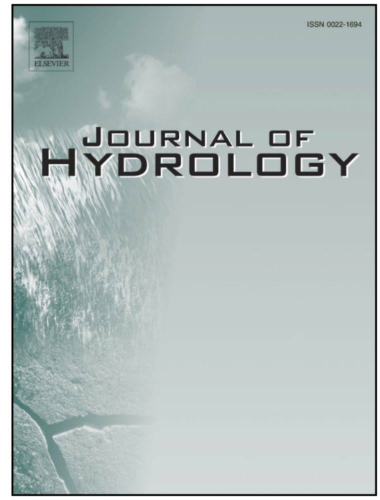
Journal Pre-proofs



Research papers

Improving LSTM hydrological modeling with spatiotemporal deep learning and multi-task learning: a case study of three mountainous areas on the Tibetan Plateau

Bu Li, Ruidong Li, Ting Sun, Aofan Gong, Fuqiang Tian, Mohd Yawar Ali Khan, Guangheng Ni

PII: S0022-1694(23)00343-8
DOI: https://doi.org/10.1016/j.jhydrol.2023.129401
Reference: HYDROL 129401

To appear in: *Journal of Hydrology*

Received Date: 13 April 2022
Revised Date: 10 March 2023
Accepted Date: 11 March 2023

Please cite this article as: Li, B., Li, R., Sun, T., Gong, A., Tian, F., Yawar Ali Khan, M., Ni, G., Improving LSTM hydrological modeling with spatiotemporal deep learning and multi-task learning: a case study of three mountainous areas on the Tibetan Plateau, *Journal of Hydrology* (2023), doi: https://doi.org/10.1016/j.jhydrol.2023.129401

Improving LSTM hydrological modeling with spatiotemporal deep learning and multi-task learning: a case study of three mountainous areas on the Tibetan Plateau

Bu Li[1], Ruidong Li[1], Ting Sun[2], Aofan Gong[1], Fuqiang Tian[1], Mohd Yawar Ali Khan[3], Guangheng Ni[1]

[1]State Key Laboratory of Hydro-science and Engineering, Department of Hydraulic Engineering, Tsinghua University, Beijing 100084, China

[2]Institute for Risk and Disaster Reduction, University College London, London WC1E 6BT, UK

[3]Department of Hydrogeology, Faculty of Earth Sciences, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Correspondence to: G. Ni, ghni@tsinghua.edu.cn

Abstract

Long short-term memory (LSTM) networks have demonstrated their excellent capability in processing long-length temporal dynamics and have proven to be effective in precipitation-runoff modeling. However, the current LSTM hydrological models lack the incorporation of multi-task learning and spatial information, which limits their ability to make full use of meteorological and hydrological data. To address this issue, this study proposes a spatiotemporal deep-learning (DL)-based hydrological model that couples the 2-Dimension convolutional neural network (CNN) and LSTM and introduces actual evaporation (E_a) as an additional training target. The proposed CNN-LSTM model is tested on three large mountainous basins on the Tibetan Plateau, and the results are compared to those obtained from the LSTM-only model. Additionally, a probe method is used to decipher the internal embedding layers of the proposed DL models. The results indicate that both LSTM and CNN-LSTM hydrological models perform well in simulating runoff (Q) and E_a, with Nash-Sutcliffe efficiency coefficients ($NSEs$) higher than 0.82 and 0.95, respectively. The higher $NSEs$ suggest that introducing spatial information into LSTM-only models can improve the overall and peak model performance. Moreover, multi-task simulation with LSTM-only models shows better accuracy in the estimation of Q volume and performance, with $NSEs$ increasing by approximately 0.02. The probe method also reveals that CNN can capture the basin-averaged meteorological values in CNN-LSTM models, while LSTM Q (E_a) models contain the information about the known E_a (Q) process. Overall, this study demonstrates the value of spatial information and multi-task learning in LSTM hydrological modeling and provides a perspective for interpreting the internal embedding layers of DL models.

Highlight

(1) Spatiotemporal DL model enhances LSTM by introducing spatial information.

(2) Multi-task simulation improves LSTM with enhanced performance in estimating Q volume.

(3) Probe method shows CNN captures basin-averaged meteorological values in CNN-LSTM, LSTM Q (E_a) models contain known process.

Keywords

CNN-LSTM; spatiotemporal; multi-task; actual evaporation; Tibetan Plateau

1. Introduction

Hydrological models, either physically-based or data-driven, play vital roles in flood and drought disaster prevention, as well as in water resources management (Blöschl et al. 2019). Recent advances in remote sensing and computational techniques have improved physically-based hydrological models by enhancing their capacity in characterizing hydrodynamic processes at finer scales (Khatakho et al. 2021, Sood and Smakhtin 2015, Tarek et al. 2020): notably, distributed hydrological models (DHMs) can incorporate processes at a sub-basin or other calculation unit scale (Li et al. 2021a). However, the application of DHMs is still limited owing to the unsolved issue in scale mismatch (Blöschl et al. 2019, Blöschl and Sivapalan 1995, Gupta et al. 2008, Hrachowitz et al. 2013, Nearing et al. 2021).

Data-driven models can directly depict the statistical relationship between inputs and outputs without explicit characterization of physical processes (Nearing et al. 2021). Since the 1990s, machine learning (ML) techniques have been widely adopted in hydrological modeling and have proven similar or better performance compared with DHMs (Demirel et al. 2009, Hsu et al. 1995, Kratzert et al. 2018, Lees et al. 2021, Nearing et al. 2021, Yang et al. 2020). In particular, deep learning (DL) models, featuring neural networks, are the most commonly used ML techniques for hydrological modeling (Nearing et al. 2021). They have evolved from original multi-layer perceptron, i.e., artificial neural networks (ANNs, Yang et al. 2020), to more advanced forms with enriched taxonomy, such as convolutional neural networks (CNNs, Jiang et al. 2020), recurrent neural networks (RNNs, Sadeghi Tabas and Samadi 2022), and its variant, long short-term memory networks (LSTM, Lees et al. 2021). ANNs, CNNs, and RNNs are among the earliest DL models with promising performance for hydrological modeling (Demirel et al. 2009, Hsu et al. 1995, Khan et al. 2019, Yang et al. 2020). As feed-forward neural networks, ANNs and CNNs cannot directly represent temporal dynamics and are thus less able to accurately characterize hydrological processes (Kratzert et al. 2018). In contrast, RNNs, which process the input data chronologically by design, can consider temporal dynamics. Although RNNs are more ideal for time series analysis, vanilla RNNs can hardly store sequences over 10 time steps (Bengio et al. 1994) which limits their applicability in modeling slow hydrological processes occurring at larger time scale, such as those related to groundwater, snow, and glacier storage (Kratzert et al. 2018). The recently emerging LSTM models can conquer such weakness with the unique internal gate architectures and have demonstrated superior performance than the vanilla RNNs in time series analysis (Hochreiter and Schmidhuber 1997). Since the first application in hydrological modeling by Kratzert et al. (2018) for a precipitation-runoff (P-Q) simulation in 530 American basins (<2000 km^2), the LSTM P-Q models have been used worldwide (e.g., 669 basins in Great Britain in Lees et al. 2021, Hanjiang River in China in Liu et al. 2021) and have become one of the most powerful tools in P-Q

simulations.

Being promising in hydrological modeling, current DL-based hydrological models still need improvements in three notable aspects (Nearing et al. 2021):

(1) Ability to resolve spatiotemporal features: hydrological processes modeling depends heavily on the spatial patterns of meteorological forcing and underlying surface characteristics. Yang et al. (2020) employed computer vision to resolve spatial features in ANN P-Q modeling and demonstrated that spatial information plays an important role in enhancing model robustness. However, most existing studies about LSTM hydrological modeling almost utilized basin spatially-averaged meteorological data as model inputs (e.g., Jiang et al. 2022, Kratzert et al. 2018, Lees et al. 2021), without fully representing spatial features of inputs for the LSTM hydrological modeling. Coupling 2-D CNN and LSTM is expected to bridge such gap by simultaneously considering both temporal dynamics and spatial features (Miao et al. 2019, Shi et al. 2015) and CNN-LSTM has proven to be promising in different fields (e.g., P nowcasting (Miao et al. 2019, Shi et al. 2015) and water quality forecast (Barzegar et al. 2020, Yang et al. 2021)).

(2) Consideration of multiple hydrological processes: differing from physically-based hydrological models, LSTM-based models simulate the individual hydrological process, such as Q (Feng et al. 2020, Frame et al. 2021), groundwater (Ali et al. 2022, Nourani et al. 2022), and snow water equivalent (Duan and Ullrich 2021) in most studies, but rarely simulate multiple processes simultaneously. It makes LSTM-based hydrological models difficult to explicitly consider the interactions between different hydrological processes and diagnose models based on hydrological theories, such as the water balance equation (Reichstein et al. 2019). Besides, some studies found that introducing additional hydrological processes, such as the actual evaporation (denoted by E_a in this work) process, in calibration for physical-based hydrological models can enhance the Q simulation performance (Herman et al. 2018, Nesru et al. 2020). Therefore, it is beneficial to investigate if considering multiple hydrological processes simultaneously for LSTM-based hydrological models can enhance their capacity in depicting more hydrological processes and thus provide a more comprehensive diagnosis of hydrological variables.

(3) Physical interpretability: due to the "black-box" nature, DL-based hydrological models have no explicit representation of physical processes and thus remain being questioned by some hydrologists (Nearing et al. 2021). To enhance the confidence of users and policymakers in adopting DL-based hydrological models, improvements in the understanding of their physical

interpretability have been attempted recently (Arrieta et al. 2020, Jiang et al. 2022). For example, LSTM hydrological models have proven to learn a generalizable representation of the underlying physical processes. LSTM regional hydrological models outperform DHMs calibrated regionally, and even calibrated for each basin individually (Feng et al. 2020, Kratzert et al. 2019, Sun et al. 2021). Besides, LSTM hydrological models are found to be able to store the hidden information consistent with hydrological knowledge (Jiang et al. 2022, Kratzert et al. 2018, Lees et al. 2022). However, the physical interpretability of DL-models with respect to E_a process–a critical component in the hydrological cycle–is yet to be investigated. Also, the physical concepts of CNN outputs in CNN-LSTM models are still unclear.

In this work, we aim to overcome these deficiencies by developing a spatiotemporal DL-based hydrological model by coupling 2-D CNN and LSTM (CNN-LSTM) and introducing multi-task learning. The potential of simultaneous multi-task (MT) learning in DL-based hydrological models is also investigated by involving E_a process as additional learning target. Besides, we also advance the understanding of physical interpretability of DL-based models by extracting the meteorological and hydrological processes hidden in the proposed LSTM and CNN-LSTM models.

In the remainder of this paper, we first describe the proposed DL-based model by introducing the basic architecture of CNN and LSTM and physical interpretability method (Sec. 2), then evaluate the model performance in three large mountainous basins on the Tibetan Plateau with comparison to the LSTM hydrological models (Sec. 3); we also explore physical interpretations of this model with respect to the hydrological and meteorological processes (Sec. 4).

2. Methods

2.1 Model development

We propose a DL-based hydrological model (Figure 1a) by coupling 2-D CNN (Figure 1b) and LSTM (Figure 1c) to utilize their respective advantages: the former for hierarchical spatial feature extraction while the latter for learning long temporal dependencies. The proposed model can use 2-D spatial meteorological and underlying surface data as input and predicts hydrological processes with daily Q and E_a as output. We note, however, only meteorological data–daily P and mean temperature–are used in this work as inclusion of surface characteristics demonstrate minimal improvement in model performance (not shown). The model can perform either single-task or multi-task learning by setting training targets: the former simulates the individual hydrological process, while the latter focuses on two or more processes simultaneously. Below we focus on the model design and description of key components; more technical details refer to Appendix A.

2.1.1 Convolutional neural networks (CNNs)

CNNs (Figure 1b), are a particular type of feed-forward neural network, including the input, convolutional, pooling, and full connection layer (LeCun et al. 1998). The convolutional layer, the core of CNNs, uses convolutional kernels to extract information from various N-Dimensions model inputs. We utilize 2-D CNN to capture the spatial information of meteorological data. The convolutional layer processes meteorological data by reducing the spatial size (width and height) of inputs (features), increasing the channel number, and generating the 1-dimension sequence finally. Figure 1b takes a study basin as the example to illustrate the data dimensions in internal layers of the CNNs in this study and more details refer to Appendix A.

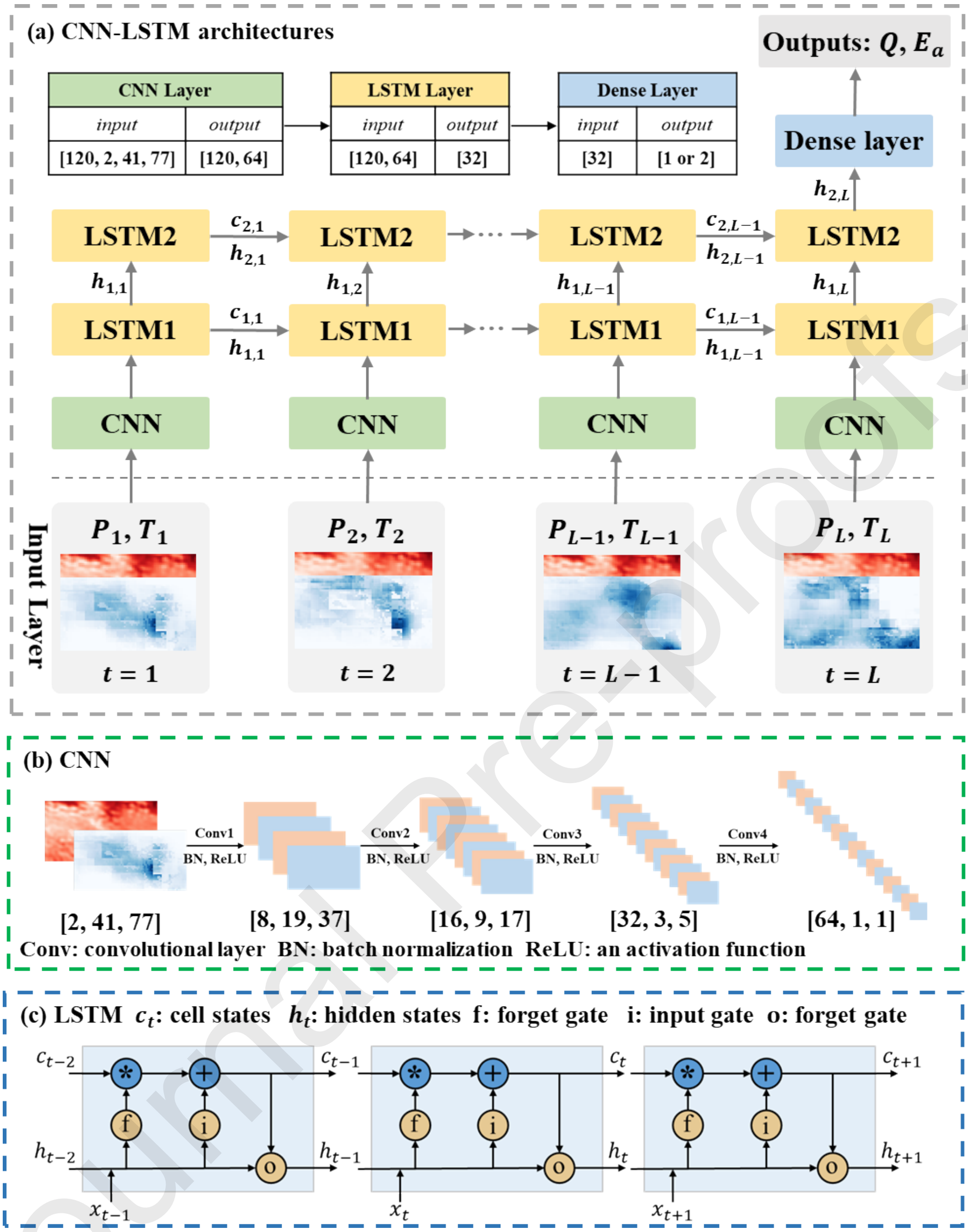


Figure 1. (a) CNN-LSTM model architectures and the dimensions of data before and after each layer (take the Yellow as an example and other basins are shown in Appendix A, same with Panel (**b**)) in this study. (b) The workflow of the CNN model and the dimensions of data in the internal layers of CNN model. (c) The internals of LSTM cells. The meaning of each unexplained variable is stated in the text.

2.1.2 Long short-term memory (LSTM) networks

The LSTM models (Figure 1c) are designed to alleviate the weakness of the vanilla RNNs in processing long-length temporal dynamics (Hochreiter and Schmidhuber 1997, Sherstinsky 2020). The success of LSTM lies in the coordination of memory cells (cell states; c_t in Figure 1c) and hidden cells (hidden states; h_t in Figure 1c) in the internal architecture that capture the slow and quick evolution processes, respectively. Besides, three gates (i.e., input, forget, and output) are designed to control the information in each cell to be stored, removed, and passed, respectively. These architectures are beneficial for LSTM models to process long-length temporal dynamics. More detailed descriptions of LSTM within the context of hydrological modeling refer to Kratzert et al. (2018).

2.1.3 Multi-task learning

Multi-task (MT) learning is to set multiple tasks as optimization targets in a DL-based model (Caruana 1997). The training can benefit from the enriched representations of MT information, thus enhancing the performance of each task using the information enveloped in the other tasks. Besides, MT learning can achieve higher efficiency and less over-fitting than single-task (ST) learning because it can lead the model to a more general feature representation preferred by multiple related tasks (Li et al. 2022, 2023). The MT learning is introduced to investigate the effect of multi-hydrological-process learning in LSTM-based hydrological modeling in this study.

2.1.4 Coupling between CNN and LSTM

To utilize the respective advantage of CNN and LSTM models, we develop a spatiotemporal DL-based hydrological model by coupling CNN and LSTM models (Figure 1a). At each time step, the CNN-LSTM coupling is realized via two stages:

1) the CNN reduces the 2-D gridded meteorological input into a 1-D sequence and feed it to the LSTM1 layer;

2) the hidden states of the LSTM1 layer are input to the LSTM2 layer and cell and hidden states of two LSTM layers are then fed to the corresponding layers at the next time step, respectively.

After all processing through the CNN-LSTM coupled layer prior to time step L, the outputs of the LSTM2 layer at the last time step are passed to the dense layer to obtain the predictions: learned hydrological variables (Figure 1a). Depending on the target mode–ST or MT–the DL model may produce different number of output variables. ST model sets one hydrological process as optimization target while MT model sets two or more hydrological processes as optimization targets by MT learning.

2.2 Physical interpretability method

The internal embedding layers of the DL hydrological models contain a large amount of data that is not explicitly interpretable. These data may conceal some untrained internal hydrological variables. For example, the internal embedding layers of Q models may contain the information about E_a process. This study utilizes probes–regression models that map the internal embedding layers of trained models to untrained hydrological variables (Hewitt and Liang 2019, Lees et al. 2022)–to test whether trained DL models can learn the known but untrained hydrological variables and examine the internal representation and further physical interpretability of models. The simplest form of probes is a linear regression (LR) model that connects the learned embedding layers to a given output. This study explores untrained hydrological variables from proposed LSTM and CNN-LSTM models based on LR models and detailed experiment design refers to Sec. 4.1.

3. Model evaluation

We evaluate the performance of proposed CNN-LSTM model in three large mountainous basins on the Tibetan Plateau (TP; Sec. 3.1). To assess the applicability of CNN-LSTM in resolving spatial information in hydrological modeling, an LSTM-only model is also configured as a benchmarking baseline: differing from the CNN-LSTM model, the basin spatially-averaged meteorological data at each time step are directly input to the LSTM-only model without the involvement of CNN model (detailed in Kratzert et al. 2018). Also, ST and MT experiments (Sec. 3.2) are designed to evaluate the MT learning performance of CNN-LSTM and LSTM models and the influence of MT on DL-based hydrological modeling. In this study, the model inputs are daily total P and mean air temperature (T) in line with similar studies (e.g., Jiang et al. 2022). In addition to Q, we also select the E_a–an important hydrological processes for which data are available– as the training target to evaluate the effect of multi-task learning in LSTM modeling.

3.1 Study area and data

3.1.1 Study area

The Tibetan Plateau (TP; Figure 2a) is the highest plateau in the world, known as the “Roof of the World”, or the “Third Pole”. Given the high altitude and vast glaciers of TP, along with the many mighty rivers (e.g., Yangtze, Lancang/Mekong, Yarlung Zangbo/Brahmaputra, and Yellow, among others) that provide enormous water sources for downstream livelihoods and agricultural irrigation (Huss et al. 2017, Immerzeel et al. 2010, Nan et al. 2021, Schaner et al. 2012, Wang et al. 2021, Zhang et al. 2013), TP is also considered the “Water tower of Asia”. However, owing to the incompleteness of knowledge of complex alpine hydrological processes, it is challenging to adequately model hydrological processes by DHMs in the TP (Li et al. 2019b, Nan et al. 2021).

Table 1. Basic facts of the three study basins. "DEM" represents Digital Elevation Model.

Basins	**Area (km^2)**	**DEM range (m)**	**Average annual *P* (mm)**	**Average annual *Q* (10^9 m^3)**
Yellow	123,000	2,656-6,253	510	20
Yangtze	139,000	3,516-6,575	460	16
Lancang	91,000	1,243-6,334	830	32

To systematically evaluate the performance of the LSTM and CNN-LSTM hydrological model in such a challenging environment, we select source regions of three rivers (Figure 2a)–the Yellow River (Figure 2b), the Yangtze River (Figure 2c), and the Lancang River (Figure 2d)–characterized as mountainous areas as the study basins. The landforms of all three basins undulate greatly with elevation variability greater than 3,000 m. The annual average P of three basins ranges from 460 to 830 mm. The detailed description of the three study basins is shown in Table 1. Yellow, Yangtze, and Lancang are used hereinafter to denote their respective source regions if not specified otherwise.

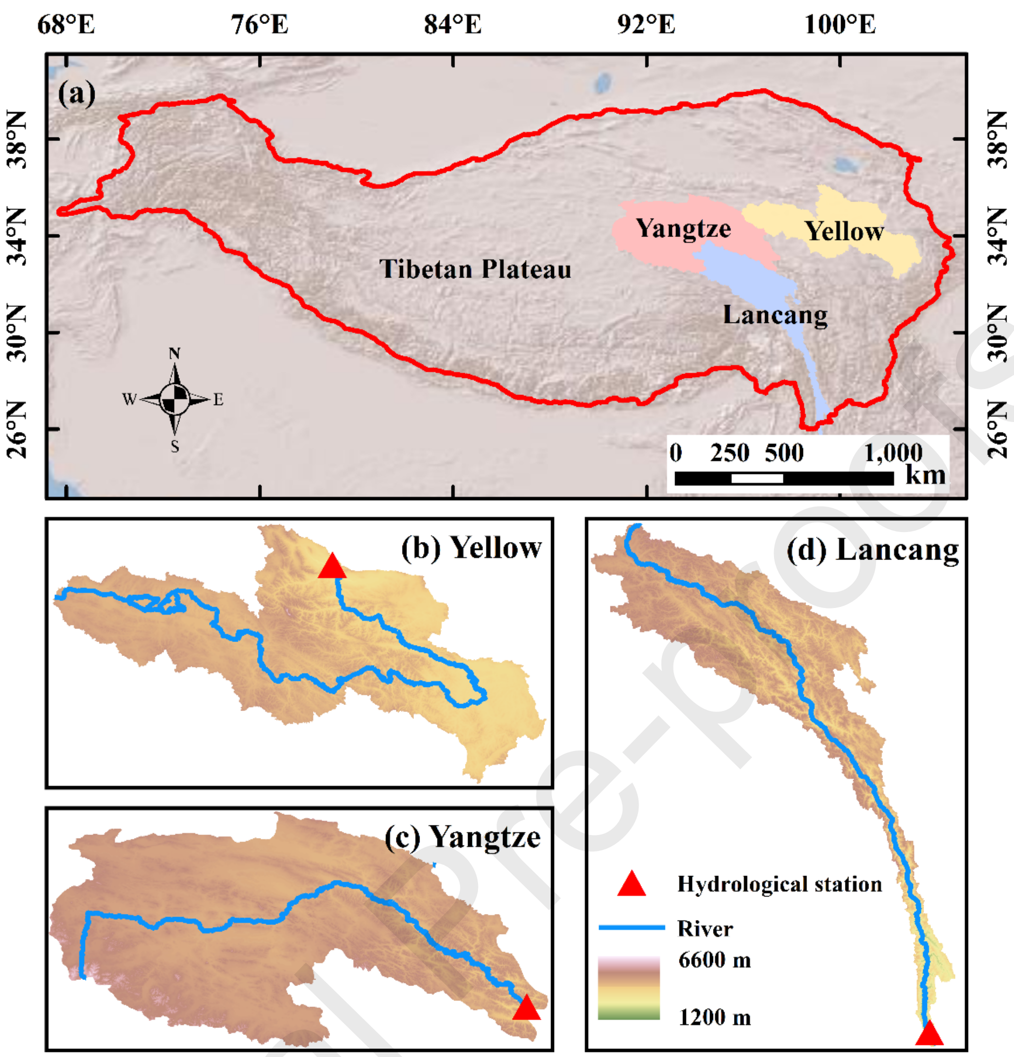


Figure 2. (a) The terrain of the Tibetan Plateau and the location of three study basins. The terrain and CNN-LSTM models input range of the Yellow River source region (b), the Yangtze River source region (c), and the Lancang River source region (d).

3.1.2 Data

Due to the limited availability of in situ observations in the study area, we utilize either the optimal remote sensing or reanalysis datasets for different input variables as follows:

(1) Precipitation (*P*): Multi-Source Weighted-Ensemble Precipitation (MSWEP) V2.2 with 0.1° spatial and 3h temporal resolution (Beck et al. 2017, Beck et al. 2019).

(2) Air temperature (T): The air temperature at 2m AGL (T2) from the fifth generation of ECMWF atmospheric reanalysis of the global climate (ERA5) reanalysis dataset with 0.1° spatially and 1h temporal resolution (Hersbach et al. 2020)

Table 2. The length of the training, evaluation, and testing periods in three study basins.

Basins	Training period	Evaluation period	Testing period
Yellow	1983-2004	2006-2009	2011-2014
Yangtze	1983-2004	2006-2009	2011-2014
Lancang	1988-2004	2005-2007	2008-2010

While for output/evaluation dataset, we use the following datasets:

(1) Evaporation (E_a): Global Land Evaporation Amsterdam Model (GLEAM) v3.5a E_a dataset with 0.25° spatial and 1-day temporal resolution (Martens et al. 2017, Miralles et al. 2011). The basin spatially-averaged daily E_a are calculated as the model learning target.

(2) Runoff (Q): the daily in situ measurements collected at three hydrological stations (Figure 1) provided by local water agencies.

All the above datasets are pre-processed into daily resolution in line with the model time step and spilt into three periods for training, evaluation, and testing (Table 2).

3.2 Model evaluation experiment

3.2.1 Experiment design

To test the performance of LSTM and CNN-LSTM models for single-task (ST) and multi-task (MT) modes, a total of 18 scenarios are set up: 3 basins (Yellow, Yangtze, Lancang) × 2 models (LSTM and CNN-LSTM) × 3 tasks (Q, E_a, $Q + E_a$). Task Q and E_a are for the ST experiment and $Q + E_a$ for MT.

3.2.2 Evaluation metrics

This study utilizes Nash–Sutcliffe efficiency (NSE; Equation 1; J.E.Nash and

J.V.Sutcliffe 1970) and its three decompositions (Gupta et al. 2009), namely, the correlation coefficient (r; Eq. 2), the variance bias (α; Eq. 3), and the total volume bias (β; Eq. 4) to evaluation metrics:

$$NSE = 1 - \frac{\sum_{t=1}^{T}(V_{sim} - V_{obs})^2}{\sum_{t=1}^{T}(V_{obs} - \overline{V_{obs}})^2} \quad (1)$$

$$r = \frac{\sum_{t=1}^{T}(V_{sim} - \overline{V_{sim}})(V_{obs} - \overline{V_{obs}})}{\sqrt{\sum_{t=1}^{T}(V_{sim} - \overline{V_{sim}})^2 \sum_{t=1}^{T}(V_{obs} - \overline{V_{obs}})^2}} \quad (2)$$

$$\alpha = \frac{\sigma_{sim}}{\sigma_{obs}} \quad (3)$$

$$\beta = \frac{(\mu_{sim} - \mu_{obs})}{\sigma_{obs}} \quad (4)$$

where V_{sim} (V_{obs}), σ_{sim} (σ_{obs}), and μ_{sim} (μ_{obs}) are the simulated (observed) values, standard deviations, and means, respectively. The variance (α) and total volume bias (β) measure the error in the standard deviation and the average values, respectively.

Besides, we calculate the peak, middle, and low values bias of the results grouped by the exceedance probability of observations to assess the range-specific model performance (Yilmaz et al. 2008):

$$BIV = \frac{\sum_{i=1}^{I}(V_{sim,i} - V_{obs,i})}{\sum_{i=1}^{I} V_{obs,i}} \quad (5)$$

where I is one of P, M, and L that represent the peak (0–0.02 exceedance probabilities), middle (0.3-0.7 exceedance probabilities), and low (0.7–1.0 exceedance probabilities) values, respectively.

3.3 Evaluation results

3.3.1 Performance of CNN-LSTM models

In general, CNN-LSTM Q models work remarkably well in all three study basins with $NSEs > 0.89$ (Table 3) (Moriasi et al. 2007) and can accurately capture peak of Q with appropriate magnitudes and timing (Figure 3). The results are compared to some traditional hydrological models in other relevant studies, and it is indicated that CNN-LSTM Q models outperform them (see detailed in Appendix B). Besides, simulated E_a processes agree well with GLEAM data with $NSEs$ of 0.97 in all three study basins (Table 4; Figure 4-5). These results demonstrate that the proposed CNN-LSTM model performs favorably and thus is a reasonable approach to simulate hydrological processes in large mountainous basins.

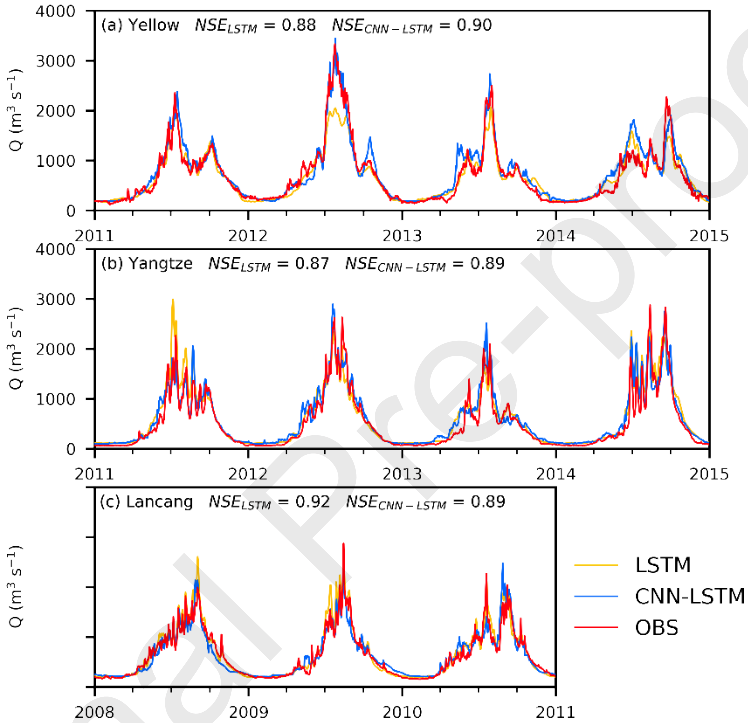


Figure 3. The comparison of simulated (ST Q models) and observed Q processes in the test period at three hydrological stations.

Table 3. The evaluation metrics of LSTM and CNN-LSTM hydrological models for Q simulation. LSTM (MT) and CNN-LSTM (MT) represent the LSTM and CNN-LSTM Q (multi-task) models, respectively.

	NSE	r	α	β	BPV	BMV	BLV

Yellow	LSTM	0.88	0.95	0.80	-0.06	-0.31	0.02	0.12
	CNN-LSTM	0.90	0.95	1.01	0.11	-0.00	0.17	0.16
	LSTM MT	0.90	0.95	0.85	-0.04	-0.29	0.03	0.10
	CNN-LSTM MT	0.86	0.94	0.98	0.15	-0.13	0.22	0.32
Yangtze	LSTM	0.87	0.94	1.01	0.12	-0.15	0.24	0.58
	CNN-LSTM	0.89	0.95	0.99	0.11	-0.11	0.30	0.63
	LSTM MT	0.89	0.95	0.97	0.04	-0.19	0.07	0.56
	CNN-LSTM MT	0.82	0.94	1.09	0.22	-0.06	0.53	0.55
Lancang	LSTM	0.92	0.97	1.11	0.07	0.04	0.01	0.07
	CNN-LSTM	0.89	0.95	0.97	0.00	-0.11	0.02	0.18
	LSTM MT	0.93	0.96	0.99	-0.02	-0.04	-0.03	0.13
	CNN-LSTM MT	0.82	0.94	1.14	0.17	-0.01	0.09	0.32

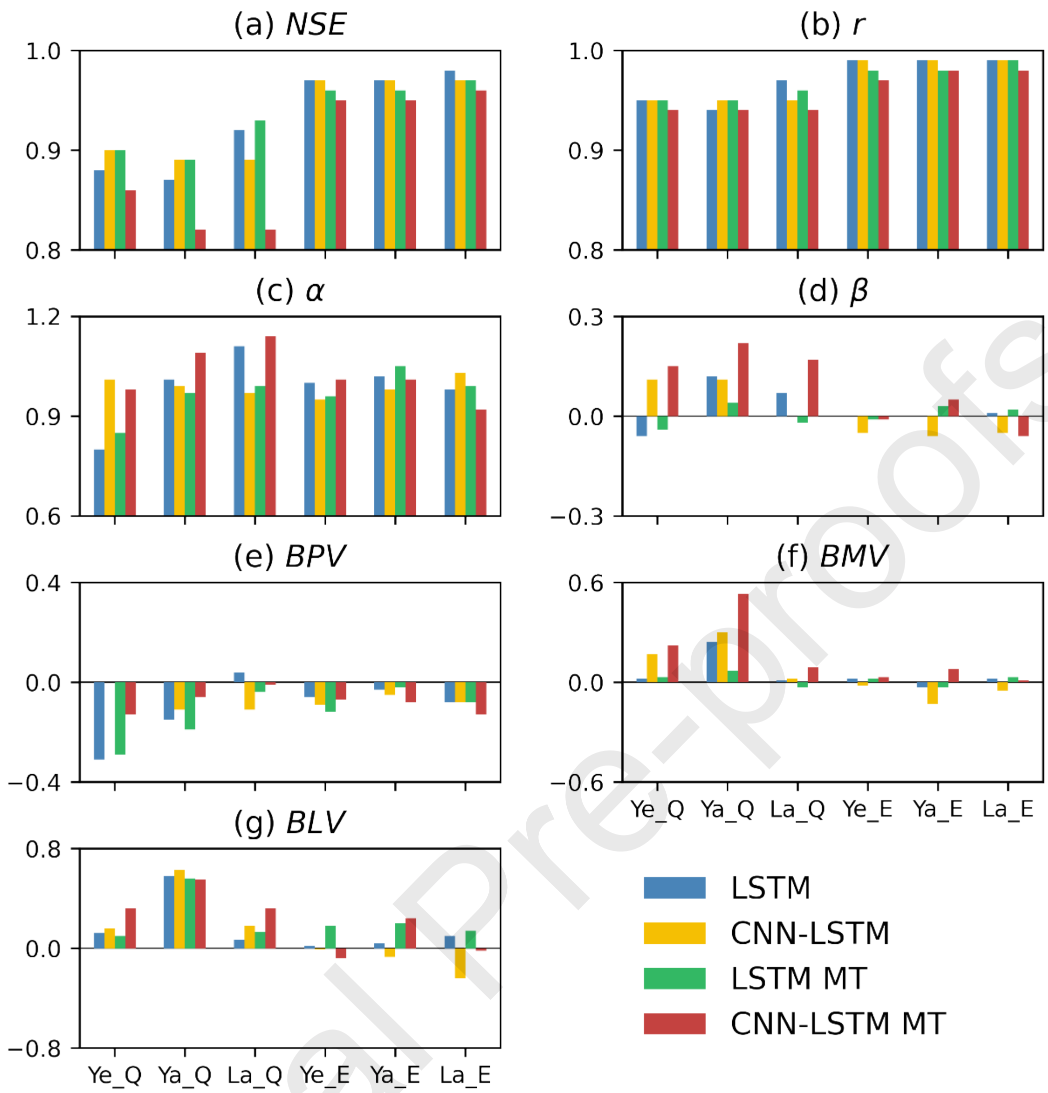


Figure 4. The evaluation metrics results of different DL-based hydrological models. "LSTM" ("CNN-LSTM") and "LSTM MT" ("CNN-LSTM MT") stand for LSTM (CNN-LSTM) ST and MT models, respectively. "Ye_Q", "Ye_E", "Ya_Q", "Ya_E", "La_Q", and "La_E" denote the Q and E_a simulation results in the Yellow, Yangtze, and Lancang, respectively.

Table 4. The evaluation metrics of LSTM and CNN-LSTM hydrological models for E_a simulation. LSTM (MT) and CNN-LSTM (MT) represent the LSTM and CNN-LSTM E_a (multi-task) models, respectively.

Basins	Models	NSE	r	α	β	BPV	BMV	BLV

Yellow	LSTM	0.97	0.99	1.00	0.00	-0.06	0.02	0.02
	CNN-LSTM	0.97	0.99	0.95	-0.05	-0.09	-0.02	-0.01
	LSTM MT	0.96	0.98	0.96	-0.01	-0.12	0.02	0.18
	CNN-LSTM MT	0.95	0.97	1.01	-0.01	-0.07	0.03	-0.08
Yangtze	LSTM	0.97	0.99	1.02	0.00	-0.03	-0.03	0.04
	CNN-LSTM	0.97	0.99	0.98	-0.06	-0.05	-0.13	-0.07
	LSTM MT	0.96	0.98	1.05	0.03	-0.02	-0.03	0.20
	CNN-LSTM MT	0.95	0.98	1.01	0.05	-0.08	0.08	0.24
Lancang	LSTM	0.98	0.99	0.98	0.01	-0.08	0.02	0.10
	CNN-LSTM	0.97	0.99	1.03	-0.05	-0.08	-0.05	-0.24
	LSTM MT	0.97	0.99	0.99	0.02	-0.08	0.03	0.14
	CNN-LSTM MT	0.96	0.98	0.92	-0.06	-0.13	0.01	-0.02

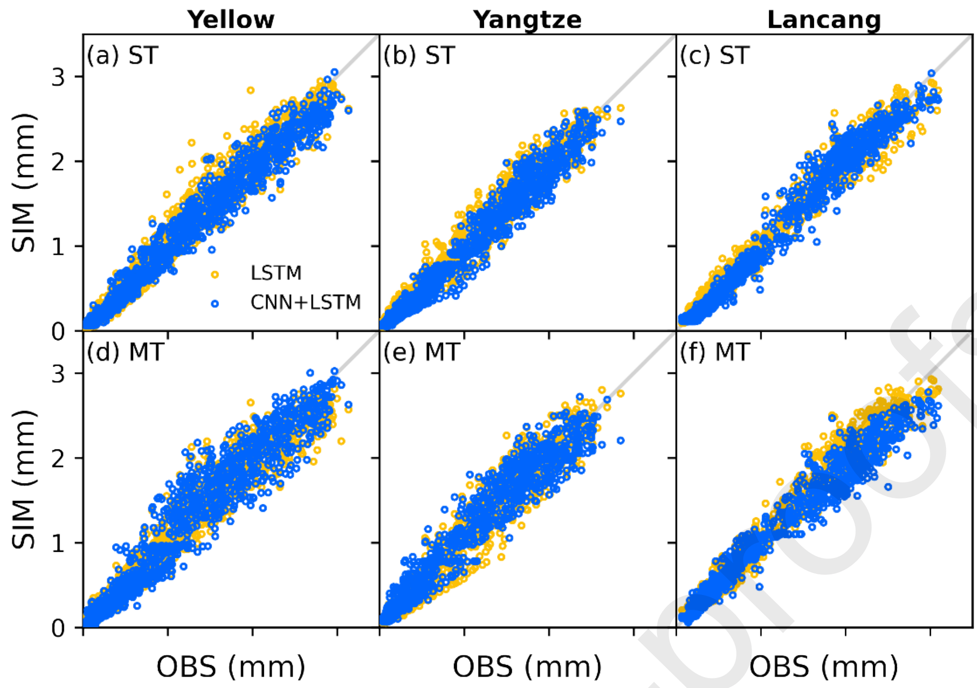


Figure 5. The comparison in daily E_a between simulations (SIM in y axis) and GLEAM (OBS in x axis) under two learning modes (ST and MT) in the test period at three study basins.

3.3.2 The capacity of 2-D CNN in extracting spatial characteristics

To evaluate the capacity of 2-D CNN in extracting spatial characteristics, we developed LSTM Q and E_a models as benchmarking baselines. High corrections ($NSEs$ more than 0.87 of Q and 0.97 of E_a; Table 3-4) indicate the good performance of LSTM models in hydrological simulation. By comparing the performance of LSTM and CNN-LSTM Q models, it is found that CNN-LSTM models outperform LSTM models in overall Q simulation with higher $NSEs$ of Q in the Yellow and Yangtze (Table 3). Also, evaluation results of three biases metrics indicate that both LSTM and CNN-LSTM Q models underestimate the peak Q processes ($BPV < 0$) but overestimate the median and low Q processes ($BMV, BLV > 0$) in the Yellow and Yangtze. But the peak Q underestimation of CNN-LSTM models is mitigated compared with that of LSTM models, with BPV from -0.31 and -0.15 (LSTM) to 0.00 and -0.11 (CNN-LSTM) in the Yellow and Yangtze, respectively. In contrast, the overestimation of simulated median and low Q processes by CNN-LSTM Q models are stronger than that by LSTM Q models in all three study basins. We conclude that the introduction of 2-D CNN into LSTM Q models has a two-sided effect: it can enhance the overall model performance as well as the capacity in simulating peaking Q processes, while may work slightly poorly in modeling Q in the median and low ranges.

These effects can be explained by the different LSTM inputs. In the CNN-LSTM models, the inputs to LSTM are the feature vectors (size is 1×64 at each time step; Table A3) extracted from 2-D spatial meteorological data by 2-D CNN; whereas those are the spatially-averaged P and T data within study basins in the LSTM-only models. Based on the remarkable performance of both LSTM and CNN-LSTM models, we assume that the LSTM inputs of CNN-LSTM models (i.e., CNN outputs) contain the information about basin spatially-averaged P and T data (discussed later in Sec. 4.2.1). Besides, we infer that more information hidden in the CNN outputs, capturing the peak runoff features, contribute to enhancing the downstream LSTM performance; while remaining redundant information might hurt downstream LSTM performance. Noting that there is one exception: in the Lancang, CNN-LSTM Q model does not outperform LSTM models in terms of overall and high Q : this could be due to different hydrological characteristics of the testing period and shorter training dataset (Table 2). All three years of the testing dataset in the Lancang can be categorized as normal or dry years without extreme Q processes (e.g., 2012 in the Yellow)–that might not benefit from the capacity of 2-D CNN in modeling peaking Q processes. Besides, we also infer that shorter training dataset might not contain enough information to train the optimal parameters for the CNN-LSTM model, whose training parameters are larger than the LSTM-only model. It causes the CNN-LSTM model does not outperform LSTM model in the Lancang. On the other hand, both LSTM and CNN-LSTM E_a models perform well without significant differences between two models in all three study basins (Table 4; Figure 4-5). It demonstrates that LSTM is a remarkable approach for average E_a of the basin modeling and introducing 2-D CNN has little impact on it, further suggesting that average P and T play vital roles in E_a modeling and CNN outputs might contain the information about it.

3.3.3 The effect of multi-task learning

To investigate the applicability of MT learning in training LSTM and CNN-LSTM models, we configured them with Q and E_a as the simultaneous learning targets in addition to those single-task (ST) experiments targeting at only Q or E_a. The simulations by MT models agree well with observations in all three study basins with favorable $NSEs$ of Q (0.89 by LSTM-MT and 0.82 by CNN-LSTM-MT) (Figure 6). Besides, both LSTM and CNN-LSTM MT models achieve remarkable performance in predicting E_a in all three study basins (Figure 5; Table 4). These results demonstrate that both LSTM and CNN-LSTM models are capable of simulating multiple hydrological processes simultaneously.

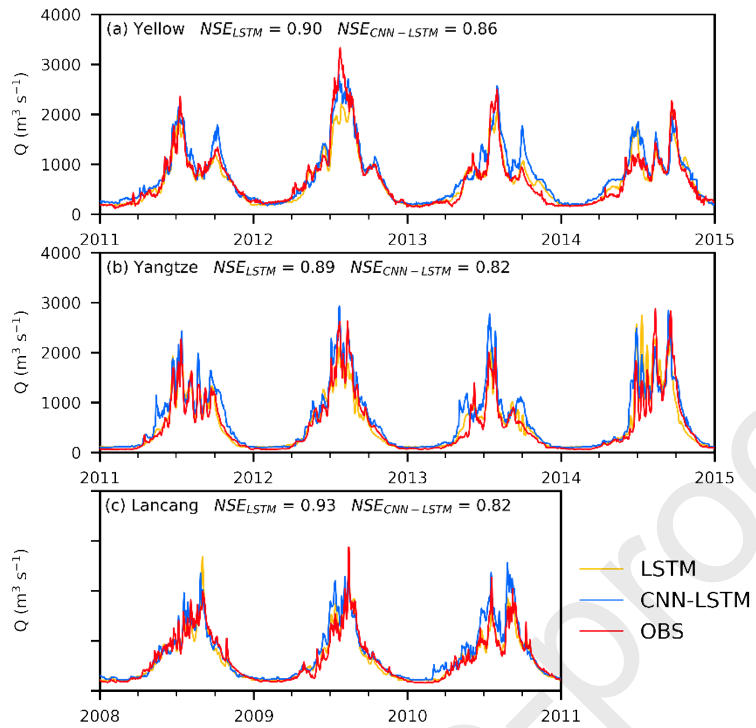


Figure 6. The comparison of simulated (MT models) and observed Q processes in the test period at three hydrological stations.

Furthermore, compared with LSTM Q models, the LSTM MT ones achieve higher $NSEs$ and lower total volume biases (β), suggesting that introducing E_a as additional training target can enhance the Q performance of LSTM models by reducing the total volume biases. It is consistent with the effect of introducing the E_a process in calibration for physical-based hydrological models (Herman et al. 2018, Nesru et al. 2020). It further motivates us to assume that the internal states of LSTM hydrological models, similar to physical-based hydrological models, might encompass multiple untrained physical hydrological processes. Thus, we put forward the hypothesis that E_a (Q) process information can be easily reconstructed from internal states of trained LSTM Q (E_a) models without the aforementioned complicated training procedures. This hypothesis will be tested in Sec. 4.2.2.

In contrast, the Q performance of CNN-LSTM MT models declines significantly compared with their ST counterparts in all three study basins. Significant performance differences between ST and MT might be explained by the variant LSTM inputs in CNN-LSTM models with different training references, compared with the invariant LSTM inputs in LSTM models (Sec. 2.3). Besides, the redundant information contained in the CNN outputs might make it difficult to simulate multiple hydrological processes. On the other hand, differences in E_a performance between ST and MT mode of both LSTM and CNN-LSTM are not significant, indicating that MT learning in the LSTM and CNN-LSTM models has minimal effect on E_a modeling.

4. Physical interpretability of DL hydrological models

This section further explores the physical interpretability of CNN and LSTM models by using the LR model (Sec. 2.2) to look into potential hydrological and meteorological information CNN and LSTM may capture (Figure 7).

4.1 Experiment design

According to the above discussions, we design two experiments to test two hypotheses as follows:

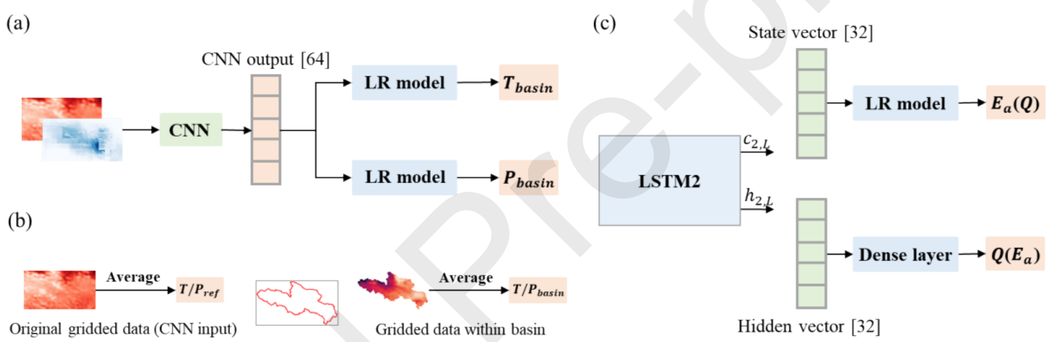


Figure 7. An overview of the physical interpretability exploration based on the LR model in this study. (a) LR models are used to map CNN output to spatially-averaged meteorological data within the basin; (b) The difference between original grided data and grided data within the basin; (c) LR models map the state vector of LSTM Q (E_a) models to untrained E_a (Q) processes.

(1) The CNN of trained CNN-LSTM models can capture spatially-averaged meteorological data within basin (T_{basin}/P_{basin}) from gridded meteorological data (CNN input; Figure 7a-b). To test this hypothesis, we construct LR models to map the CNN output of three CNN-LSTM (Q, E_a, and MT) models (Figure 7a) to spatially-averaged precipitation and temperature data within basins (P_{basin} and T_{basin}) and then use the correlation coefficient r between LR model output and spatially-averaged meteorological data within basins (namely r_{CNN}) as indicator to examine their correlation (Figure 7a). In addition, we calculate the r of spatial averages of meteorological data within

the basin and those of the original gridded data (r_{ref}) as the reference indicator (Figure 7b). By comparing r_{CNN} and r_{ref}, we can thus infer if the CNN can extract basin-scale characteristics of meteorological input data or not.

(2) E_a (Q) process information can be reconstructed from internal states of trained LSTM Q (E_a) models without the aforementioned complicated training procedures (Figure 7c). Similar as Lees et al. (2022), LR models are constructed to map the cell states vector–the memory of LSTM E_a (Q) models–to the untrained Q (E_a) processes and examine the resultant NSE to justify if LSTM is capable of inferring untrained hydrological information. For this, we obtain 6 sets of NSE values: 2 relations ($E_a \rightarrow Q$ and $Q \rightarrow E_a$) × 3 basins (Yellow, Yangtze, and Lancang).

4.2 The physical interpretability results

4.2.1 The physical interpretability of CNN outputs

As the meteorological data within a studied basin is essentially a subset of the original gridded dataset (Figure 7b), it is expected the spatial average of the former datasets (i.e., P_{basin} and T_{basin}) can be highly correlated to those of the latter (i.e., P_{ref} and T_{ref}), in particular when their spatial extents are comparable (e.g. the Yangtze river basin in Figure 2c). And our estimation does suggest high P and T correlation for all three basins ($r_{ref} > 0.97$) except the lower P correlation for the Lancang ($r_{ref} < 0.80$; Table 5). Thus, we select the Lancang basin as the study area to examine the capacity of CNN in inferring basin-specific information from the original gridded dataset. Our results indicate that the spatial averages of P and T within the Lancang can be extracted from the CNN outputs in high fidelity: r_{CNN} of all configurations are much larger than r_{ref} for P (cf. > 0.87 vs. -0.50; Table 5) and close to 1 for T, which supports our first hypothesis. Furthermore, it is also interesting to see the difference in r_{CNN} between different configurations: for P, a lower r_{CNN} is produced by the evapotranspiration-targeted model, suggesting those runoff-targeted models perform better in extracting precipitation-related information; while a reverse pattern is found for T. Such results are in line with the common hydrological understanding that P has a more significant impact on Q while T is a strong control of E_a.

Table 5. The r_{CNN} (three CNN-LSTM models: Q: CNN-Q, E_a: CNN-E, $Q + E_a$: CNN-Q+E) and r_{ref} of P and T in the Lancang.

	r_{CNN}	r_{ref}

	CNN_Q	CNN_E	CNN_Q+E	
P	0.96	0.94	0.97	0.79
T	0.98	1.00	1.00	1.00

4.2.2 The physical interpretability of LSTM cell states

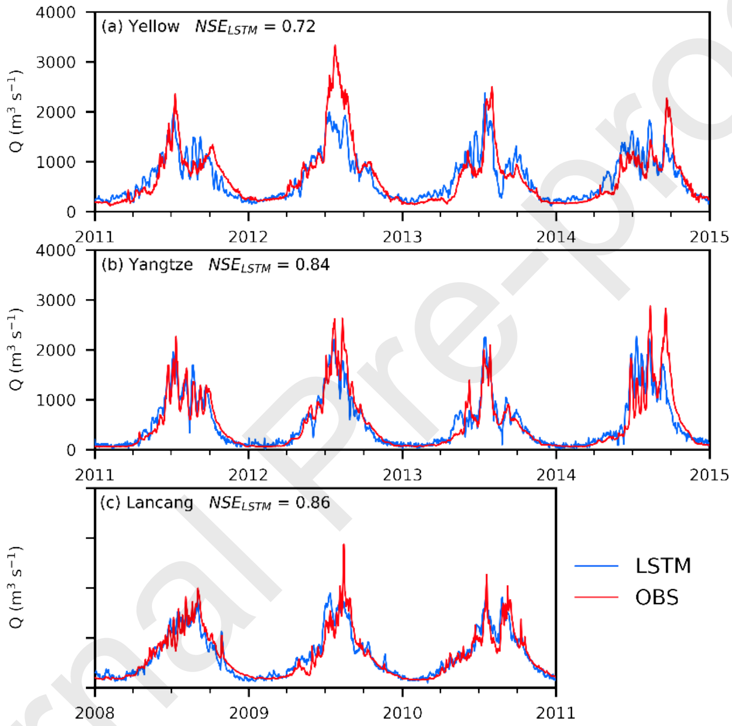


Figure 8. The comparison of simulated (extracted from the cell states of LSTM E_a models) and observed Q processes in the test period at three hydrological stations.

High correlation is found between the LSTM-inferred and observed Q ($NSEs$ of 0.72, 0.84, and 0.86 in the Yellow, Yangtze, and Lancang; Figure 8), suggesting the remarkable capacity of LSTM in inferring the Q-related processes solely from Ea. Likewise, the inferred E_a processes also agree favorably with GLEAM E_a data with $NSEs$ more than 0.93 in all three study basins (Figure 9). Such results support our second hypothesis that LSTM Q (E_a) models contain information about untrained E_a (Q) processes without prior training procedures.

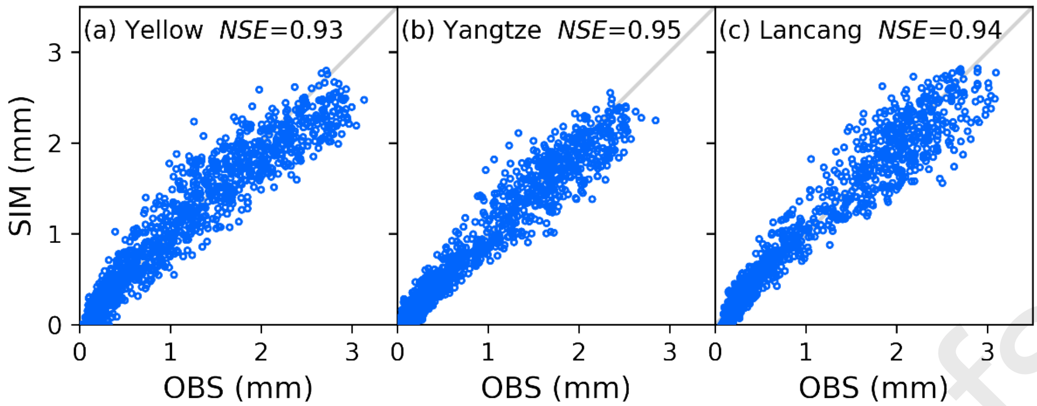


Figure 9. The comparison of simulated (extracted from the cell states of LSTM Q models) and observed E_a processes in the test period at three hydrological stations.

5. Conclusion and limitation

In this study, we develop an integrated spatiotemporal DL-based model by coupling CNN and LSTM for hydrological simulation and evaluate it in the source regions of the Yellow River, the Yangtze River, and the Lancang River. Besides, we employ a simple linear regression method to explore the physical interpretability of CNN and LSTM in DL-based hydrological models to improve our understanding of DLM-inferred hydrological processes. Our main findings are as follows:

(1) Both LSTM and CNN-LSTM models perform favorably with NSEs more than 0.87 (Q) and 0.97 (E_a) in all three study basins. The CNN-LSTM Q models achieve better performance than LSTM ones in modeling Q. Therefore, we recommend the CNN-LSTM models for hydrological modeling and flooding forecasts in large mountainous basins.

(2) Both LSTM and CNN-LSTM multi-task models work well for Q (NSEs > 0.82) and E_a (NSEs > 0.95) simulation and LSTM multi-task models outperform LSTM Q models in all three study basins. This finding demonstrates that introducing E_a as an additional training target in the LSTM Q model can enhance the model performance and further research can be conducted to evaluate the potential of LSTM models for more and even whole hydrological modeling in the future.

(3) The internal cells of CNN and LSTM contain the hydrological and meteorological information consistent with our knowledge. CNN outputs in CNN-LSTM models can capture the basin-specific characteristics from 2-D gridded meteorological data of larger domains, while the internal cells of

LSTM Q (E_a) models may infer the E_a (Q) processes. These findings advance the understanding of the physical interpretability of DL-based hydrological models.

Being promising in predicting hydrological processes, this model has several limitations. First, unlike the observed runoff data, the GLEAM dataset–as the E_a training target–is of high accuracy but subject to some biases in the Tibetan Plateau (Li et al. 2019a). It might lead trained E_a models to a slight deviation from the true E_a process. Besides, this model is only evaluated in the three large mountainous basins on the Tibetan Plateau due to the limitation of computational resources. Future research will be carried out to evaluate the performance of proposed CNN-LSTM model in more basins, develop CNN-LSTM regional hydrological models using more underlying surface data, and explore more physical information hidden in the internal cells of DL-based hydrological models.

Appendix A: Architecture and parameters of LSTM and CNN-LSTM models

This study developed 3 (basins: Yellow, Yangtze, Lancang) × 2 (model: LSTM, CNN+LSTM) × 3 (objective: runoff, evaporation, runoff + evaporation) = 18 DL-based hydrological models. The loss function is the mean-squared error (MSE), and the optimizer is Adaptive Moment Estimation (Adam). The batch size is 32, and the epoch is 400. The detailed parameters are as follows.

Table A1. The parameters of LSTM hydrological models

parameters	range of parameter values	final parameter values		
		Yellow	Yangtze	Lancang
Hidden states	16, 32, 64, 128, 256	32	32	32
Length of the input sequence	10, 20, 30×n (n=1, 2, …, 12)	90	180	120
Number of LSTM layer	1, 2	2	2	2

Table A2. The parameters of LSTM models in CNN-LSTM hydrological models

parameters	range of parameter values	final parameter values		
		Yellow	Yangtze	Lancang
Hidden states	16, 32, 64, 128, 256	32	32	32
Length of the input sequence	10, 20, 30×n (n=1, 2, …, 12)	60	120	90

Number of LSTM layer	1, 2	2	2	2

Table A3. The parameters of CNN models in CNN-LSTM hydrological models

Yellow		**Yangtze**		**Lancang**	
size	**Architectures**	**size**	**Architectures**	**size**	**architectures**
41×77	*conv*, 5×5, s=2, 8	34×71	*conv*, 5×5, s=2, 8	84×59	*conv*, 5×5, s=2, 8
19×37	*conv*, 5×5, s=2, 16	16×35	*conv*, 5×5, s=2, 16	40×28	*conv*, 5×5, s=2, 16
9×17	*conv*, 5×5, s=2×3, 32	7×17	*conv*, 5×5, s=2×3, 32	19×12	*conv*, 5×5, s=3×2, 32
3×5	*conv*, 3×5, s=1, 64	2×5	*conv*, 2×5, s=1, 64	5×4	*conv*, 5×4, s=1, 64
1×1		1×1		1×1	

Appendix B: The simulated daily and monthly runoff results of three study basins in other relevant studies

To compare the proposed CNN-LSTM model with traditional hydrological models, we collected simulated daily and monthly runoff results using hydrological models in relevant studies (Table B1). Although the simulation periods and model inputs are different from this study, the higher NSE results can demonstrate that the CNN-LSTM models outperform traditional hydrological models in runoff simulation.

Table B1. The simulated daily and monthly runoff NSE results of three study basins in other relevant studies

Basins	**Models**	**Simulation periods**	**NSE results**	**Reference**

Yellow	SWAT	2014-2018	0.71	(Xie et al. 2020)
Yellow	SPHY	2000-2016	0.76	(Zhang et al. 2022)
Yangtze	WEB-DHM-SF	1987-2016	0.62	(Qi et al. 2019)
Yangtze	SWAT	2001-2016	0.75 (monthly)	(Ahmed et al. 2022)
Lancang	SWAT	1981-2015	0.71 (monthly)	(Li et al. 2021b)

Code and data availability. Precipitation, air temperature, and actual evaporation data used in this research study are openly available and the sources are mentioned in Sec. 3.1.2. Model outputs and code are available by request to the corresponding author. The observed runoff data are not publicly available for legal/ethical reasons.

Acknowledgments

This work was funded by National Natural Science Foundation of China (92047301) and National Key Research and Development Project of China (2018YFA0606002).

References

Ahmed, N., Wang, G., Booij, M.J., et al. (2022). Separation of the impact of landuse/landcover change and climate change on runoff in the upstream area of the Yangtze River, China. Water resources management 36(1), 181-201.

Ali, A.S.A., Ebrahimi, S., Ashiq, M.M., et al. (2022). CNN-Bi LSTM neural network for simulating groundwater level. Environ Eng 8, 1-7.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58, 82-115.

Barzegar, R., Aalami, M.T. and Adamowski, J. (2020). Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. Stochastic Environmental Research and Risk Assessment 34(2), 415-433.

Beck, H.E., van Dijk, A.I.J.M., Levizzani, V., et al. (2017). MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. Hydrology and Earth System Sciences 21(1), 589-615.

Beck, H.E., Wood, E.F., Pan, M., et al. (2019). MSWEP V2 Global 3-Hourly 0.1° Precipitation: Methodology and Quantitative Assessment. Bulletin of the American Meteorological Society 100(3), 473-500.

Bengio, Y., Simard, P. and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks 5(2), 157-166.

Blöschl, G., Bierkens, M.F.P., Chambel, A., et al. (2019). Twenty-three unsolved problems in hydrology (UPH) – a community perspective. Hydrological Sciences Journal 64(10), 1141-1158.

Blöschl, G. and Sivapalan, M. (1995). Scale issues in hydrological modelling: a review. Hydrological Processes 9(3-4), 251-290.

Caruana, R. (1997). Multitask learning. Machine learning 28(1), 41-75.

Demirel, M.C., Venancio, A. and Kahya, E. (2009). Flow forecast by SWAT model and ANN in Pracana basin, Portugal. Advances in Engineering Software 40(7), 467-473.

Duan, S. and Ullrich, P. (2021). A comprehensive investigation of machine learning models for estimating daily snow water equivalent over the Western US. Earth and Space Science Open Archive.

Feng, D., Fang, K. and Shen, C. (2020). Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks With Data Integration at Continental Scales. Water Resources Research 56(9), e2019WR026793.

Frame, J.M., Kratzert, F., Raney, A., et al. (2021). Post-Processing the National Water Model with Long Short-Term Memory Networks for Streamflow Predictions and Model Diagnostics. Journal of the American Water Resources Association 57(6), 885-905.

Gupta, H.V., Kling, H., Yilmaz, K.K., et al. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology 377(1-2), 80-91.

Gupta, H.V., Wagener, T. and Liu, Y. (2008). Reconciling theory with observations: elements of a diagnostic approach to model evaluation. Hydrological Processes: An International Journal 22(18), 3802-3813.

Herman, M.R., Nejadhashemi, A.P., Abouali, M., et al. (2018). Evaluating the role of evapotranspiration remote sensing data in improving hydrological modeling predictability. Journal of Hydrology 556, 39-49.

Hersbach, H., Bell, B., Berrisford, P., et al. (2020). The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society 146(730), 1999-2049.

Hewitt, J. and Liang, P. (2019). Designing and Interpreting Probes with Control Tasks. Proceedings of the 2019 Con.

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

Hrachowitz, M., Savenije, H.H.G., Blöschl, G., et al. (2013). A decade of Predictions in Ungauged Basins (PUB)—a review. Hydrological Sciences Journal 58(6), 1198-1255.

Hsu, K.l., Gupta, H.V. and Sorooshian, S.J.W.r.r. (1995). Artificial neural network modeling of the rainfall-runoff process. Water Resources Research 31(10), 2517-2530.

Huss, M., Bookhagen, B., Huggel, C., et al. (2017). Toward mountains without permanent snow and ice. Earths Future 5(5), 418-435.

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

J.E.Nash and J.V.Sutcliffe (1970). River flow forecasting through conceptual models part I — A discussion of principles. Journal of Hydrology 10(3), 282-290.

Jiang, S., Zheng, Y. and Solomatine, D. (2020). Improving AI System Awareness of Geoscience Knowledge: Symbiotic Integration of Physical Approaches and Deep Learning. Geophysical Research Letters 47(13).

Jiang, S., Zheng, Y., Wang, C., et al. (2022). Uncovering Flooding Mechanisms Across the Contiguous United States Through Interpretive Deep Learning on Representative Catchments. Water Resources Research 58(1), e2021WR030185.

Khan, M.Y.A., Tian, F., Hasan, F., et al. (2019). Artificial neural network simulation for prediction of suspended sediment concentration in the River Ramganga, Ganges Basin, India. International Journal of Sediment Research 34(2), 95-107.

Khatakho, R., Talchabhadel, R. and Thapa, B.R. (2021). Evaluation of different precipitation inputs on streamflow simulation in Himalayan River basin. Journal of Hydrology 599.

Kratzert, F., Klotz, D., Brenner, C., et al. (2018). Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrology and Earth System Sciences 22(11), 6005-6022.

Kratzert, F., Klotz, D., Shalev, G., et al. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrology and Earth System Sciences 23(12), 5089-5110.

LeCun, Y., Bottou, L., Bengio, Y., et al. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278-2324.

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

Lees, T., Reece, S., Kratzert, F., et al. (2022). Hydrological concept formation inside long short-term memory (LSTM) networks. Hydrology Earth System Sciences 26(12), 3079-3101.

Li, B., Zhou, X., Ni, G., et al. (2021a). A multi-factor integrated method of calculation unit delineation for hydrological modeling in large mountainous basins. Journal of Hydrology 597, 126180.

Li, R., Huang, H., Yu, G., et al. (2021b). Contributions of climatic variation and human activities to streamflow changes in the Lancang-Mekong River Basin. Resources Science 43(12), 2428-2441.

Li, R., Sun, T., Tian, F., et al. (2022). SHAFTS (v2022. 3): a deep-learning-based Python package for Simultaneous extraction of building Height And FootprinT from Sentinel Imagery. Geoscientific Model Development Discussions, 1-42.

Li, R., Sun, T., Tian, F., et al. (2023). SHAFTS (v2022.3): a deep-learning-based Python package for simultaneous extraction of building height and footprint from sentinel imagery. Geoscientific Model Development 16(2), 751-778.

Li, X., Long, D., Han, Z., et al. (2019a). Evapotranspiration estimation for Tibetan plateau headwaters using conjoint terrestrial and atmospheric water balances and multisource remote sensing. Water Resources Research 55(11), 8608-8630.

Li, Z., Feng, Q., Li, Z., et al. (2019b). Climate background, fact and hydrological effect of multiphase water transformation in cold regions of the Western China: A review. Earth-Science Reviews 190, 33-57.

Liu, Y., Zhang, T., Kang, A., et al. (2021). Research on Runoff Simulations Using Deep-Learning Methods. Sustainability 13(3), 1336.

Martens, B., Miralles, D.G., Lievens, H., et al. (2017). GLEAM v3: satellite-based land evaporation and root-zone soil moisture. Geoscientific Model Development 10(5), 1903-1925.

Miao, Q., Pan, B., Wang, H., et al. (2019). Improving Monsoon Precipitation Prediction Using Combined Convolutional and Long Short Term Memory Neural Network. Water 11(5), 977.

Miralles, D.G., Holmes, T.R.H., De Jeu, R.A.M., et al. (2011). Global land-surface evaporation estimated from satellite-based observations. Hydrology and Earth System Sciences 15(2), 453-469.

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., et al. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the Asabe 50(3), 885-900.

Nan, Y., He, Z., Tian, F., et al. (2021). Can we use precipitation isotope outputs of isotopic general circulation models to improve hydrological modeling in large mountainous catchments on the Tibetan Plateau? Hydrology and Earth System Sciences 25(12), 6151-6172.

Nearing, G.S., Kratzert, F., Sampson, A.K., et al. (2021). What Role Does Hydrological Science Play in the Age of Machine Learning? Water Resources Research 57(3), e2020WR028091.

Nesru, M., Shetty, A. and Nagaraj, M.K. (2020). Multi-variable calibration of hydrological model in the upper Omo-Gibe basin, Ethiopia. Acta Geophysica 68(2), 537-551.

Nourani, V., Khodkar, K. and Gebremichael, M. (2022). Uncertainty assessment of LSTM based groundwater level predictions. Hydrological Sciences Journal 67(5), 773-790.

Qi, J., Wang, L., Zhou, J., et al. (2019). Coupled snow and frozen ground physics improves cold region hydrological simulations: an evaluation at the upper Yangtze River Basin (Tibetan Plateau). Journal of Geophysical Research: Atmospheres 124(23), 12985-13004.

Reichstein, M., Camps-Valls, G., Stevens, B., et al. (2019). Deep learning and process understanding for data-driven Earth system science. Nature 566(7743), 195-204.

Sadeghi Tabas, S. and Samadi, S. (2022). Variational Bayesian dropout with a Gaussian prior for recurrent neural networks application in rainfall–runoff modeling. Environmental Research Letters 17(6).

Schaner, N., Voisin, N., Nijssen, B., et al. (2012). The contribution of glacier melt to streamflow. Environmental Research Letters 7(3), 034029.

Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Physica D: Nonlinear Phenomena 404.

Shi, X., Chen, Z., Wang, H., et al. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems 28.

Sood, A. and Smakhtin, V. (2015). Global hydrological models: a review. Hydrological Sciences Journal 60(4), 549-565.

Sun, A.Y., Jiang, P., Mudunuru, M.K., et al. (2021). Explore Spatio-Temporal

Learning of Large Sample Hydrology Using Graph Neural Networks. Water Resources Research 57(12), e2021WR030394.

Tarek, M., Brissette, F.P. and Arsenault, R. (2020). Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America. Hydrology and Earth System Sciences 24(5), 2527-2544.

Wang, T., Zhao, Y., Xu, C., et al. (2021). Atmospheric dynamic constraints on Tibetan Plateau freshwater under Paris climate targets. Nature Climate Change 11(3), 219-225.

Xie, P., Zhuo, L., Yang, X., et al. (2020). Spatial-temporal variations in blue and green water resources, water footprints and water scarcities in a large river basin: A case for the Yellow River basin. Journal of Hydrology 590, 125222.

Yang, S., Yang, D., Chen, J., et al. (2020). A physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data. Journal of Hydrology 590, 125206.

Yang, Y., Xiong, Q., Wu, C., et al. (2021). A study on water quality prediction by a hybrid CNN-LSTM model with attention mechanism. Environ Sci Pollut Res Int 28(39), 55129-55139.

Yilmaz, K.K., Gupta, H.V. and Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. Water Resources Research 44(9), W09417.

Zhang, L., Su, F., Yang, D., et al. (2013). Discharge regime and simulation for the upstream of major rivers over Tibetan Plateau. Journal of Geophysical Research: Atmospheres 118(15), 8500-8518.

Zhang, T., Li, D. and Lu, X. (2022). Response of runoff components to climate change in the source-region of the Yellow River on the Tibetan plateau. Hydrological Processes 36(6), e14633.

Table 1. Basic facts of the three study basins. "DEM" represents Digital Elevation Model.

Basins	**Area (km^2)**	**DEM range (m)**	**Average annual *P* (mm)**	**Average annual *Q* (10^9 m^3)**

Yellow	123,000	2,656-6,253	510	20
Yangtze	139,000	3,516-6,575	460	16
Lancang	91,000	1,243-6,334	830	32

715 **Table 2.** The length of the training, evaluation, and testing periods in three study
716 basins.

Basins	Training period	Evaluation period	Testing period
Yellow	1983-2004	2006-2009	2011-2014
Yangtze	1983-2004	2006-2009	2011-2014
Lancang	1988-2004	2005-2007	2008-2010

717

Table 3. The evaluation metrics of LSTM and CNN-LSTM hydrological models for Q simulation. LSTM (MT) and CNN-LSTM (MT) represent the LSTM and CNN-LSTM Q (multi-task) models, respectively.

		NSE	r	α	β	BPV	BMV	BLV
Yellow	LSTM	0.88	0.95	0.80	-0.06	-0.31	0.02	0.12
	CNN-LSTM	0.90	0.95	1.01	0.11	-0.00	0.17	0.16
	LSTM MT	0.90	0.95	0.85	-0.04	-0.29	0.03	0.10
	CNN-LSTM MT	0.86	0.94	0.98	0.15	-0.13	0.22	0.32
Yangtze	LSTM	0.87	0.94	1.01	0.12	-0.15	0.24	0.58
	CNN-LSTM	0.89	0.95	0.99	0.11	-0.11	0.30	0.63
	LSTM MT	0.89	0.95	0.97	0.04	-0.19	0.07	0.56
	CNN-LSTM MT	0.82	0.94	1.09	0.22	-0.06	0.53	0.55
Lancang	LSTM	0.92	0.97	1.11	0.07	0.04	0.01	0.07
	CNN-LSTM	0.89	0.95	0.97	0.00	-0.11	0.02	0.18
	LSTM MT	0.93	0.96	0.99	-0.02	-0.04	-0.03	0.13
	CNN-LSTM MT	0.82	0.94	1.14	0.17	-0.01	0.09	0.32

Table 4. The evaluation metrics of LSTM and CNN-LSTM hydrological models for E_a simulation. LSTM (MT) and CNN-LSTM (MT) represent the LSTM and CNN-LSTM E_a (multi-task) models, respectively.

Basins	Models	NSE	r	α	β	BPV	BMV	BLV
Yellow	LSTM	0.97	0.99	1.00	0.00	-0.06	0.02	0.02
	CNN-LSTM	0.97	0.99	0.95	-0.05	-0.09	-0.02	-0.01
	LSTM MT	0.96	0.98	0.96	-0.01	-0.12	0.02	0.18
	CNN-LSTM MT	0.95	0.97	1.01	-0.01	-0.07	0.03	-0.08
Yangtze	LSTM	0.97	0.99	1.02	0.00	-0.03	-0.03	0.04
	CNN-LSTM	0.97	0.99	0.98	-0.06	-0.05	-0.13	-0.07
	LSTM MT	0.96	0.98	1.05	0.03	-0.02	-0.03	0.20
	CNN-LSTM MT	0.95	0.98	1.01	0.05	-0.08	0.08	0.24
Lancang	LSTM	0.98	0.99	0.98	0.01	-0.08	0.02	0.10
	CNN-LSTM	0.97	0.99	1.03	-0.05	-0.08	-0.05	-0.24
	LSTM MT	0.97	0.99	0.99	0.02	-0.08	0.03	0.14
	CNN-LSTM MT	0.96	0.98	0.92	-0.06	-0.13	0.01	-0.02

Table 5. The r_{CNN} (three CNN-LSTM models: Q: CNN-Q, E_a: CNN-E, $Q + E_a$: CNN-Q+E) and r_{ref} of P and T in the Lancang.

	r_{CNN}			r_{ref}
	CNN_Q	**CNN_E**	**CNN_Q+E**	
P	0.96	0.94	0.97	0.79
T	0.98	1.00	1.00	1.00

Declaration of interests

☒ 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.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Long short-term memory (LSTM) networks have demonstrated their excellent capability in processing long-length temporal dynamics and have proven to be effective in precipitation-runoff modeling. However, the current LSTM hydrological models lack the incorporation of multi-task learning and spatial information, which limits their ability to make full use of meteorological and hydrological data. To address this issue, this study proposes a spatiotemporal deep-learning (DL)-based hydrological model that couples the 2-Dimension convolutional neural network (CNN) and LSTM and introduces actual evaporation (E_a) as an additional training target. The proposed CNN-LSTM model is tested on three large mountainous basins on the Tibetan Plateau, and the results are compared to those obtained from the LSTM-only model. Additionally, a probe method is used to decipher the internal embedding layers of the proposed DL models. The results indicate that both LSTM and CNN-LSTM hydrological models perform well in simulating runoff (Q) and E_a, with Nash-Sutcliffe efficiency coefficients ($NSEs$) higher than 0.82 and 0.95, respectively. The higher $NSEs$ suggest that introducing spatial information into LSTM-only models can improve the overall and peak model performance. Moreover, multi-task simulation with LSTM-only models shows better accuracy in the estimation of Q volume and performance, with $NSEs$ increasing by approximately 0.02. The probe method also reveals that CNN can capture the basin-averaged meteorological values in CNN-LSTM models, while LSTM Q (E_a) models contain the information about the known E_a (Q) process. Overall, this study demonstrates the value of spatial information and multi-task learning in LSTM hydrological modeling and provides a perspective for interpreting the internal embedding layers of DL models.