https://doi.org/10.5194/egusphere-2024-1449
Preprint. Discussion started: 30 May 2024

Robustness of the long short-term memory network in rainfall-runoff prediction improved by the water balance constraint

Qiang Li[1], Tongtiegang Zhao[1]

[1]Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), School of Civil Engineering, Sun Yat-Sen University, Guangzhou, China,

Correspondence to: Tongtiegang Zhao (zhaottg@mail.sysu.edu.cn)

Abstract. While the water balance constraint is fundamental to catchment hydrological models, there is yet no consensus on its role in the long short-term memory (LSTM) network. This paper is concentrated on the part that this constraint plays in the robustness of the LSTM network for rainfall-runoff prediction. Specifically, numerical experiments are devised to examine the robustness of the LSTM and its architecturally mass-conserving variant (MC-LSTM); and the Explainable Artificial Intelligence (XAI) is employed to interrogate how this constraint affects the robustness of the LSTM in learning rainfall-runoff relationships. Based on the Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) dataset, the LSTM, MC-LSTM and EXP-HYDRO models are trained under various amounts of training data and different seeds of parameter initialization over 531 catchments, leading to 95,580 (3×6×10×531) tests. Through large-sample tests, the results show that incorporating the water balance constraint into the LSTM improves the robustness, while the improvement tends to decrease as the amount of training data increases. Under 9 years' training data, this constraint significantly enhances the robustness against data sparsity in 37% (196 in 531) of the catchments and improves the robustness against parameter initialization in 73% (386 in 531) of the catchments. In addition, it improves the robustness in learning rainfall-runoff relationships by increasing the median contribution of precipitation from 45.8% to 47.3%. These results point to the compensation effects between training data and process knowledge on the LSTM's performance. Overall, the in-depth investigations facilitate insights into the use of the LSTM for rainfall-runoff prediction.

https://doi.org/10.5194/egusphere-2024-1449
Preprint. Discussion started: 30 May 2024

1 Introduction

Deep learning (DL) has been increasingly used for rainfall-runoff modelling (Kratzert et al., 2018; Lees et al., 2021; Nearing et al., 2021; Shen, 2018; Tsai et al., 2021). Without explicit descriptions of the underlying physical processes, DL models can easily be set up to directly capture input-output patterns hidden in large datasets (Feng et al., 2020; LeCun et al., 2015). DL models are shown to be effective in simulating complex nonlinear systems across different fields owing to rapid growth of available data and advances in computational capability (LeCun et al., 2015; Reichstein et al., 2019; Wang et al., 2023). In recent years, DL models have gained popularity in hydrology (Frame et al., 2022; Gauch et al., 2021a; Nearing et al., 2021; Kratzert et al., 2018). There are extensive uses of the long short-term memory (LSTM) network (Kratzert et al., 2018), the recurrent neural network (Cai et al., 2022), the gate recurrent unit (Zhang et al., 2021), the sequence-to-sequence model (Xiang et al., 2020) and the encoder-decoder model (Kao et al., 2020; Nearing et al., 2024).

The LSTM network is one of the most important DL models (Feng et al., 2021; Jiang et al., 2022b; Kao et al., 2020; Lees et al., 2022; Razavi, 2021). Due to the recurrent structure and gating mechanism (Hochreiter and Schmidhuber, 1997), the LSTM network can account for not only nonlinear relationships but also temporal dependencies among variables (Jiang et al., 2022b; Read et al., 2019). These inherent capabilities make the LSTM network well suited for modelling hydrological dynamics, especially multi-scale memory effects such as the persistence of soil moisture and the release of water from snowpack (Pokharel et al., 2023; Wi and Steinschneider, 2022). Compared to process-based hydrological models, the LSTM network has been shown to be similarly effective or even better in rainfall-runoff prediction (Gauch et al., 2021a; Lees et al., 2021; Kratzert et al., 2018). There were thorough tests in predictions in ungauged basins (Kratzert et al., 2019a; Yin et al., 2021b), multistep predictions (Kao et al., 2020; Yin et al., 2021a; Xiang et al., 2020), predictions at multiple timescales (Gauch et al., 2021a) and regional modelling (Kratzert et al., 2019b; Feng et al., 2020).

The lack of physical mechanism is a critical issue in the LSTM network (Read et al., 2019; Reichstein et al., 2019; Xie et al., 2021; Zhao et al., 2019). Without explicit physical mechanisms such as the conservation of mass and energy, the LSTM network cannot guarantee causal relationships (Wang et al., 2023; Xie et al., 2021; Frame et al., 2023). Inaccurate and even spurious predictions are possible, in particular when extrapolating the LSTM network beyond training data (Bhasme et al., 2022; Reichstein et al., 2019). This outcome reduces the credibility of the outputs of the LSTM network and limits its applications (Cai et al., 2022; Read et al., 2019; Wang et al., 2023). In the meantime, there exists heavy reliance of the LSTM network on available observations (Read et al., 2019; Xie et al., 2021). Usually, the LSTM network requires a large amount of training data to learn the dynamics so as to achieve robust performance (Gauch et al., 2021b; Kratzert et al., 2019b; Tsai et al., 2021; Yang et al., 2020).

There has been growing attention to the water balance constraint for the LSTM network (Frame et al., 2023, 2022; Hoedt et al., 2021; Pokharel et al., 2023). As to the architecturally mass-conserving variant, i.e., MC-LSTM (Hoedt et al., 2021), the water balance constraint is shown to enhance the accuracy and physical consistency of the regional LSTM (Nearing et al., 2020; Wi and Steinschneider, 2024). Due to the limitations of available data and computational cost, this

https://doi.org/10.5194/egusphere-2024-1449
Preprint. Discussion started: 30 May 2024

model has to be trained for single catchments in certain applications (Wi and Steinschneider, 2022). The robustness against data sparsity is essential in these cases (Feng et al., 2021; Gauch et al., 2021b). Moreover, as parameter initialization plays a part in DL models, the robustness against parameter initialization is also important to the LSTM (Kratzert et al., 2018). Furthermore, the exact role of the water balance constraint in the LSTM's learning of rainfall-runoff relationships remains unknown and requires detailed investigations (Wi and Steinschneider, 2022).

Overall, there is yet no consensus on the part that the water balance constraint plays in the LSTM network (Pokharel et al., 2023). Aiming to bridge the gap, this paper is concentrated on the effects of this constraint on the robustness of the LSTM network at the local scale. Since the robustness refers to the ability to perform consistently across varying conditions (Manure et al., 2023), this paper focuses on the robustness of the LSTM and MC-LSTM from three perspectives. The objectives are to examine (1) the robustness against data sparsity, (2) the robustness against parameter initialization and (3) the robustness in learning rainfall-runoff relationships. To this end, large-sample tests for rainfall-runoff prediction are devised based on the Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) dataset across the contiguous United States. Furthermore, the Explainable Artificial Intelligence (XAI) is utilized to interrogate the rainfall-runoff relationships learned by the LSTM and MC-LSTM.

2 Methods

2.1 LSTM network

The LSTM network takes a recurrent architecture, allowing information to be stored and passed over time steps through the cell state vector ($\boldsymbol{c}^t$) and the hidden state vector ($\boldsymbol{h}^t$) (Hochreiter and Schmidhuber, 1997; Jiang et al., 2022b). At each time step (t), the recurrent unit utilizes the current input ($\boldsymbol{x}^t$) and previous hidden state ($\boldsymbol{h}^{t-1}$) to calculate three gates, the input gate ($\boldsymbol{i}^t$), forget gate ($\boldsymbol{f}^t$) and output gate ($\boldsymbol{o}^t$), which control what new information to add in, what previous information to forget and what current information to output, respectively. Finally, the hidden state ($\boldsymbol{h}^t$) is passes through a head layer to derive the final prediction (y^t). The above processes can be formulated as follows:

$$\begin{cases} \boldsymbol{f}^t = \sigma\left(\mathbf{W}_{\mathrm{xf}}\boldsymbol{x}^t + \mathbf{W}_{\mathrm{hf}}\boldsymbol{h}^{t-1} + \boldsymbol{b}_f\right) \\ \tilde{\boldsymbol{c}}^t = \tanh\left(\mathbf{W}_{\mathrm{xc}}\boldsymbol{x}^t + \mathbf{W}_{\mathrm{hc}}\boldsymbol{h}^{t-1} + \boldsymbol{b}_c\right) \\ \boldsymbol{i}^t = \sigma\left(\mathbf{W}_{\mathrm{xi}}\boldsymbol{x}^t + \mathbf{W}_{\mathrm{hi}}\boldsymbol{h}^{t-1} + \boldsymbol{b}_i\right) \\ \boldsymbol{c}^t = \boldsymbol{f}^t \odot \boldsymbol{c}^{t-1} + \boldsymbol{i}^t \odot \tilde{\boldsymbol{c}}^t \\ \boldsymbol{o}^t = \sigma\left(\mathbf{W}_{\mathrm{xo}}\boldsymbol{x}^t + \mathbf{W}_{\mathrm{ho}}\boldsymbol{h}^{t-1} + \boldsymbol{b}_o\right) \\ \boldsymbol{h}^t = \boldsymbol{o}^t \odot \tanh(\boldsymbol{c}^t) \\ y^t = \boldsymbol{W}_d\boldsymbol{h}^t \end{cases} \tag{1}$$

where $\mathbf{W}$ and $\boldsymbol{b}$ respectively indicate learnable weights and bias parameters to be calibrated during the training procedures. Additionally, σ, $\tanh$ and $\odot$ represent the sigmoid function, the tanh function and the element-wise multiplication, respectively. The internal operation of a standard LSTM network is shown by Fig. S1 in the Supplement.

2.2 Water balance constraint

The Theory-Guided Data Science (TGDS) has presented a new paradigm to incorporate physical constraints into DL models so that their predictions tend to be physically consistent (Karniadakis et al., 2021; Wang et al., 2023; Wi and Steinschneider, 2022; Faghmous et al., 2014; Karpatne et al., 2017). As one of the TGDS strategies, the mass-conserving LSTM (MC-LSTM) is an architecturally mass-conserving variant of the LSTM network (Hoedt et al., 2021). Specifically, the mass conservation constraint is incorporated into the architecture of the LSTM network in order to enforce the water balance constraint in rainfall-runoff prediction (Frame et al., 2023, 2022; Hoedt et al., 2021; Nearing et al., 2021).

The MC-LSTM employs the normalized activation functions and subtracts the output mass from the storage mass to enforce conservation laws in the architecture of the LSTM network. The input variables are classified into mass inputs ($\boldsymbol{x}^t$) and auxiliary inputs ($\boldsymbol{a}^t$). As to mass inputs, the normalized activation functions are used in the input gate ($\boldsymbol{i}^t$) and the forget gate ($\mathbf{R}^t$) to guarantee that the mass is conserved between the mass inputs ($\boldsymbol{x}^t$) and the previous cell state ($\boldsymbol{c}^{t-1}$). Furthermore, the output mass ($\boldsymbol{h}^t$) is subtracted from the total mass ($\boldsymbol{m}^t$) through the output gate ($\boldsymbol{o}^t$) to keep mass conserved between the cell state ($\boldsymbol{c}^t$) and the output mass. Mathematically, the MC-LSTM is described as follows:

$$\boldsymbol{i}^t=\tilde{\sigma}\left(\mathbf{W}_{\mathrm{i}}\boldsymbol{a}^t+\mathbf{U}_{\mathrm{i}}\frac{\boldsymbol{c}^{t-1}}{\|\boldsymbol{c}^{t-1}\|_1}+\mathbf{V}_{\mathrm{i}}\boldsymbol{x}^t+\boldsymbol{b}_i\right) \tag{2}$$

$$\boldsymbol{o}^t=\sigma\left(\mathbf{W}_{\mathrm{o}}\boldsymbol{a}^t+\mathbf{U}_{\mathrm{o}}\frac{\boldsymbol{c}^{t-1}}{\|\boldsymbol{c}^{t-1}\|_1}+\mathbf{V}_{\mathrm{o}}\boldsymbol{x}^t+\boldsymbol{b}_o\right) \tag{3}$$

$$\mathbf{R}^t=\mathrm{R\tilde{e}Lu}\left(\mathbf{W}_{\mathrm{r}}\boldsymbol{a}^t+\mathbf{U}_{\mathrm{r}}\frac{\boldsymbol{c}^{t-1}}{\|\boldsymbol{c}^{t-1}\|_1}+\mathbf{V}_{\mathrm{r}}\boldsymbol{x}^t+\boldsymbol{b}_r\right) \tag{4}$$

$$\boldsymbol{m}^t=\mathbf{R}^t\boldsymbol{c}^{t-1}+\boldsymbol{i}^t\boldsymbol{x}^t \tag{5}$$

$$\boldsymbol{c}^t=(1-\boldsymbol{o}^t)\odot\boldsymbol{m}^t \tag{6}$$

$$\boldsymbol{h}^t=\boldsymbol{o}^t\odot\boldsymbol{m}^t \tag{7}$$

$$y^t=\sum_{i=2}^{n}h_i^t \tag{8}$$

where $\mathbf{W}$, $\mathbf{U}$ and $\mathbf{V}$ represent learnable weights; $\boldsymbol{b}$ denotes the learnable bias parameters; $\tilde{\sigma}$ and $\mathrm{Re\tilde{L}U}$ indicate the normalized sigmoid function and the normalized $\mathrm{Re\tilde{L}U}$ function as Eq. (9) and Eq. (10), respectively. The internal operation of a MC-LSTM network is shown by Fig. S2 in the Supplement.

$$\tilde{\sigma}(i_k) = \frac{\sigma(i_k)}{\sum_k \sigma(i_k)} \tag{9}$$

$$Re\tilde{L}U(i_k) = \frac{max(i_k, 0)}{\sum_k max(i_k, 0)} \tag{10}$$

For unobserved mass sinks, e.g., evapotranspiration, the MC-LSTM takes a subset of the output mass to accumulate the output water that does not convert to runoff. The runoff (y^t) is the sum of the output mass, excluding that subset representing the unobserved mass sinks, as shown by Eq. (8). Accordingly, the internal calculations of the MC-LSTM ensure strictly mass-conservation (here water balance) at any time step, between inputs (here precipitation), outputs (here runoff and other sinks) and cell states (here water storage) (Frame et al., 2023).

2.3 EXP-HYDRO model

The EXP-HYDRO model is employed to benchmark the performances of the LSTM and MC-LSTM. The EXP-HYDRO model is a daily conceptual hydrological model that strictly adheres to water balance (Patil and Stieglitz, 2014). It has two state variables referred to as the snow accumulation bucket (S_0) and the catchment bucket (S_1). The water balance equation is formulated:

$$\begin{cases} \frac{dS_0}{dt} = P_s - M \\ \frac{dS_1}{dt} = P_r + M - ET - Q \end{cases} \tag{11}$$

where M, ET, Q, P_s and P_r are 5 flux variables, representing the snowmelt (mm/day), evapotranspiration (mm/day), streamflow (mm/day), daily snowfall (mm/day) and rainfall (mm/day), respectively. They are calculated by 3 input variables, the daily precipitation (mm/day), temperature (°C) and day length (hour), as shown by Text S1 in the Supplement.

In the analysis, the EXP-HYDRO model is wrapped by the recurrent neural network architecture in the differentiable PyTorch framework (Paszke et al., 2019). The mathematical expressions and learnable parameters are replaced with the physical equations and parameters of the EXP-HYDRO model (Zhong et al., 2023; Jiang et al., 2020). Similar to DL models, its parameters are learnable during the training procedures. In the meantime, the internal calculations follow the physical equations of the EXP-HYDRO model.

2.4 Integrated gradient

The XAI is employed to enhance the transparency and interpretability of black-box DL models (Topp et al., 2023). It improves the understanding of model predictions and facilitates identifications of causality within model architectures. As one of the popular XAI methods, the integrated gradient (IG) (Sundararajan et al., 2017) is developed to interpret the rainfall-runoff predictions from the LSTM (Kratzert et al., 2021; Frame et al., 2021; Jiang et al., 2022a). By means of calculating and integrating the gradients of the model output to the input features, the IG method can trace back the specific contributions of the inputs to the output and assign an importance score to each input feature (Jiang et al., 2022a). The IG score for the input feature x at the ith time step is calculated as:

$$\Phi_{\mathrm{i}}(x) = \left(x_i - x_i'\right) \int_0^1 \frac{\partial f\left(x' + \alpha\left(x - x'\right)\right)}{\partial x_i} d\alpha \quad (12)$$

where $\frac{\partial f\left(x' + \alpha\left(x - x'\right)\right)}{\partial x_i}$ represents the local gradient of the model f at a point interpolated from the baseline input (x', where $\alpha = 0$, denoting the absence of this input feature) to the target input (x, where $\alpha = 1$, representing the original input).

Positive IG score, negative score and IG score close to zero indicate that the input feature contributes to, deteriorates and hardly affects the model output, respectively. Due to the completeness property of the IG method, the model output can be decomposed into the sum of individual contributions of all input features at all time steps, which enables to obtain the contribution of a set of input features by summing their IG scores (Jiang et al., 2022a). For the LSTM and MC-LSTM, a sequence of contributions of the input features are generated for each prediction with the same dimensions as the inputs.

3 Large-sample tests

3.1 Experimental design

The large-sample tests cover daily streamflow observations, catchment attributes and three catchment-averaged daily meteorological forcings for 671 catchments across the contiguous United States over the period from1980 to 2010 (Addor et al., 2017; Newman et al., 2015). The Daymet (Thornton et al., 1997) is chosen as the forcing inputs considering its high spatial resolution (1 km × 1km) and promising forcing quality (Feng et al., 2022; Newman et al., 2015). 531 catchments are selected for a direct comparison with the previous studies (Newman et al., 2017; Kratzert et al., 2019a; Frame et al., 2023). Other catchments that are larger than 2,000 km^2 or show large discrepancies in their areas when calculated using different strategies are removed. In the modelling of the LSTM, MC-LSTM and EXP-HYDRO, the daily runoff is taken as the target variable. The daily precipitation, maximum temperature, minimum temperature, vapor pressure, solar radiation and day length are taken as forcing variables of the LSTM and MC-LSTM. As shown in Fig. 1, three experiments are set up to assess the effects of the water balance constraint on the robustness.

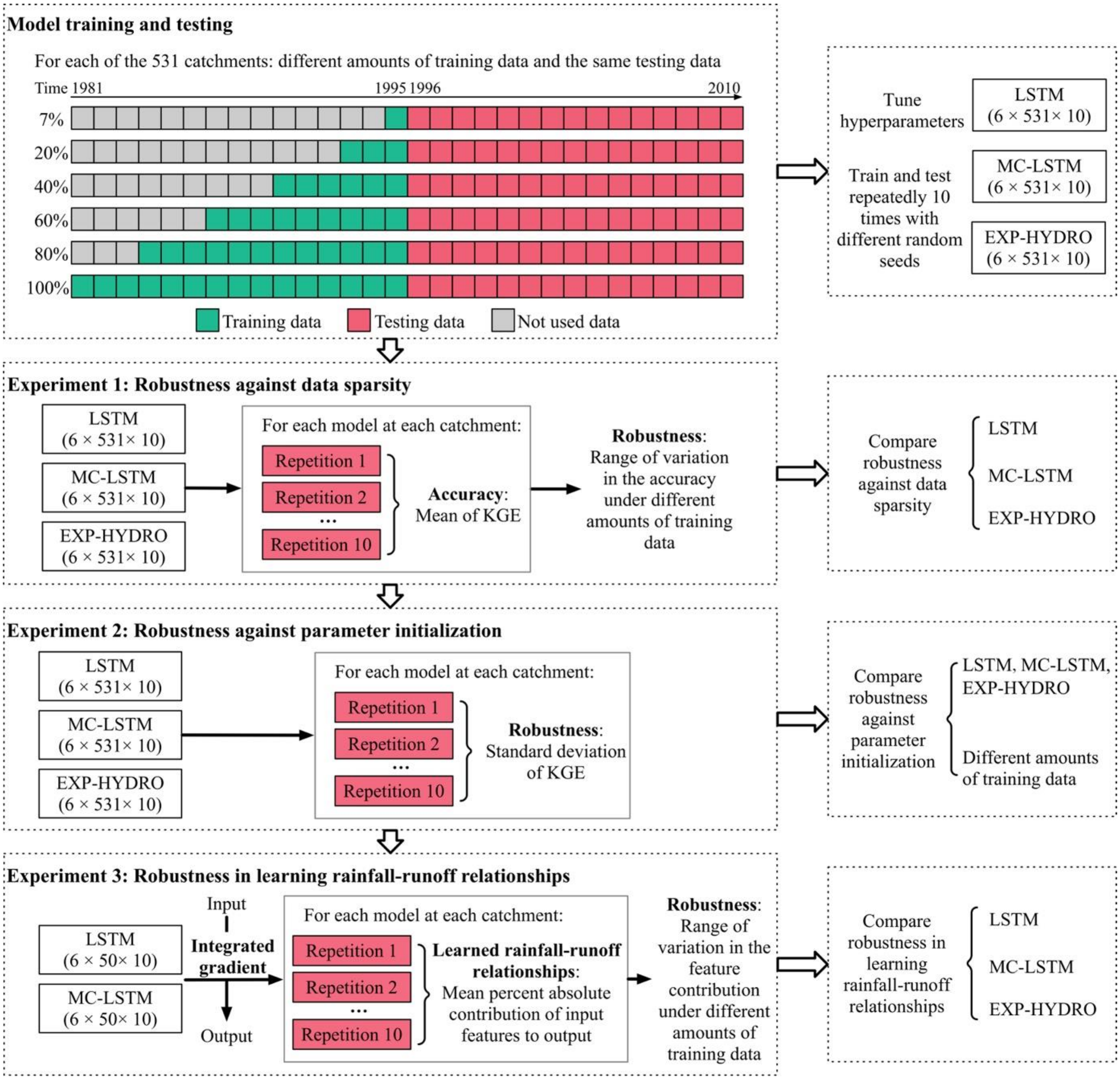

Model training and testing
For each of the 531 catchments: different amounts of training data and the same testing data
Time 1981
1995 1996
2010
7%
20%
40%
60%
80%
100%
🟩 Training data
🟥 Testing data
⬜ Not used data
Tune hyperparameters
LSTM (6 × 531× 10)
Train and test repeatedly 10 times with different random seeds
MC-LSTM (6 × 531× 10)
EXP-HYDRO (6 × 531× 10)
Experiment 1: Robustness against data sparsity
LSTM (6 × 531× 10)
MC-LSTM (6 × 531× 10)
EXP-HYDRO (6 × 531× 10)
For each model at each catchment:
Repetition 1
Repetition 2
...
Repetition 10
Accuracy: Mean of KGE
Robustness: Range of variation in the accuracy under different amounts of training data
Compare robustness against data sparsity
LSTM
MC-LSTM
EXP-HYDRO
Experiment 2: Robustness against parameter initialization
LSTM (6 × 531× 10)
MC-LSTM (6 × 531× 10)
EXP-HYDRO (6 × 531× 10)
For each model at each catchment:
Repetition 1
Repetition 2
...
Repetition 10
Robustness: Standard deviation of KGE
Compare robustness against parameter initialization
LSTM, MC-LSTM, EXP-HYDRO
Different amounts of training data
Experiment 3: Robustness in learning rainfall-runoff relationships
LSTM (6 × 50× 10)
MC-LSTM (6 × 50× 10)
Input
Integrated gradient
Output
For each model at each catchment:
Repetition 1
Repetition 2
...
Repetition 10
Learned rainfall-runoff relationships: Mean percent absolute contribution of input features to output
Robustness: Range of variation in the feature contribution under different amounts of training data
Compare robustness in learning rainfall-runoff relationships
LSTM
MC-LSTM
EXP-HYDRO

Figure 1. Overview of the experimental design.

3.2 Robustness assessments

Experiment 1: Robustness against data sparsity

The robustness against data sparsity is estimated by the range of variation in the accuracy under different amounts of training data (Wang et al., 2023; Read et al., 2019). As the mean performance across the ensemble models with different random seeds is considered as the stable performance, the mean KGE across the 10-member ensemble models is calculated as the accuracy of each catchment. The ranges of variation in accuracy under different amounts of training data are calculated and compared across different models. Larger range of variation implies worse robustness against data sparsity.

Experiment 2: Robustness against parameter initialization

The robustness against parameter initialization is quantified by the standard deviation of KGE across the 10-member ensemble models (Tsai et al., 2021). As the differences among the accuracy of models with different random seeds reflect the robustness against parameter initialization, the standard deviation of KGE across the 10-member ensemble models is calculated for each catchment. The standard deviations of KGE are compared across different models under different amounts of training data. Higher standard deviation of KGE indicates worse robustness against parameter initialization.

Experiment 3: Robustness in learning rainfall-runoff relationships

The robustness in learning rainfall-runoff relationships is assessed by the range of variation in contributions of input features under different amounts of training data. Considering that there exists intense computation, the attention is paid to the 50 catchments with the highest KGE. For the LSTM and MC-LSTM, the contribution of each input feature is calculated using the IG method. The mean percentage of absolute contribution is calculated as the contribution of the input feature (Topp et al., 2023). Larger range of variation in the contributions would suggest worse robustness in learning rainfall-runoff relationships.

3.3 Model training and testing

The period from 1 October 1980 to 30 September 1995 is considered as the training period and the period from 1 October 1995 to 30 September 2010 is considered as the testing period. Sparse training datasets are constructed for each catchment by removing some years' data from the entire training period. This setting can not only avoid the destruction of time dependence but also create real scenarios with limited historical data (Read et al., 2019). The sparse training datasets are set to 7%, 20%, 40%, 60%, 80% and 100% of the complete training data so that the amounts of training data are 1, 3, 6, 9, 12 and 15 years. Repeating the procedures of training and testing independently with different random seeds is used to reduce the uncertainty caused by the random initialization of parameters (Feng et al., 2020; Kratzert et al., 2019a). In total, 95,580 (3×6×10×531) tests are run by performing 10 runs with different random seeds across the 531 catchments.

The input variables are pre-processed. As to the LSTM, the inputs are normalized by removing the mean and scaling by the standard deviation of training data. As to the MC-LSTM, the auxiliary inputs (input variables excluding precipitation) are normalized but the mass input (precipitation) is not. The MC-LSTM is architecturally constrained by the water balance so that the dropout strategy is not utilized. To strike a balance between minimizing the uncertainty caused by different numbers of model parameters and achieving potentially more powerful predictions, the hidden sizes of the LSTM and MC-LSTM networks are set to 50 and 20, respectively. Apart from the above settings, the LSTM, MC-LSTM and EXP-HYDRO models have the same hyperparameters (Table 1) and the same loss function:

$$Loss=\frac{\sum_{n=1}^{N}\left(y_n-\hat{y}_n\right)^2}{\sum_{n=1}^{N}\left(y_n-\bar{y}\right)^2} \tag{13}$$

where N is the number of samples; $\hat{y}_n$ and y_n represent the simulated runoff and its corresponding observation, respectively; $\bar{y}$ is the averaged value of observed runoff. The details of the hyperparameter optimization are provided by Text S2 in the
190 Supplement.

Table 1. Hyperparameters of the LSTM, MC-LSTM and EXP-HYDRO models

Hyperparameter	LSTM	MC-LSTM	EXP-HYDRO
Batch size	256	256	256
Initial learning rate	0.01	0.01	0.01
Learning rate decay	0.3	0.3	0.3
Input time step (day)	365	365	365
Lead time (day)	1	1	1
Hidden size	50	20	-
Dropout rate	0.4	-	-
Epoch	Early stopping	Early stopping	Early stopping
Optimizer	Adam	Adam	Adam

The Kling-Gupta Efficiency (KGE) (Gupta et al., 2009) is used to quantify the performance of the rainfall-runoff
195 prediction:

$$KGE=1-\sqrt{(r-1)^2+(\beta-1)^2+(\gamma-1)^2} \tag{14}$$

There are three components of the KGE, i.e., correlation, bias and variance:

$$r=\frac{\sum_{n=1}^{N}(y_n-\mu_{obs})(\hat{y}_n-\mu_{sim})}{\sqrt{\sum_{n=1}^{N}(y_n-\mu_{obs})^2}\sqrt{\sum_{n=1}^{N}(\hat{y}_n-\mu_{sim})^2}} \tag{15}$$

$$\beta=\frac{\mu_{sim}}{\mu_{obs}} \tag{16}$$

$$\gamma=\frac{\sigma_{sim}}{\sigma_{obs}} \tag{17}$$

where r is the correlation coefficient between simulations and observations; $\hat{y}_n$ and y_n represent the simulated runoff and its corresponding observation, respectively; μ and σ represent the mean and standard deviation of the runoff series, respectively; β represents the ratio between mean simulations and mean observations; γ measures the relative variability in the simulations
200 and observations (Gupta et al., 2009). The value of KGE varies from negative infinity to 1; larger values indicate better performance. Furthermore, the Nash-Sutcliffe efficiency of the three models with the complete training data in this paper are

compared with that in previous studies (Jiang et al., 2020; Kratzert et al., 2018; Patil and Stieglitz, 2014; Newman et al., 2017; Hoedt et al., 2021). As shown by Table S1 in the supplement, the LSTM, MC-LSTM and EXP-HYDRO models in this paper exhibit competitive performances, suggesting the reasonability of the procedures of hyperparameter optimization and model training.

4 Results

4.1 Robustness against data sparsity

The KGE values across the 531 catchments under different amounts of training data are illustrated in Fig. 2. As the amount of training data increases from 1 to 15 years, the LSTM network benefits the most. By contrast, the MC-LSTM network is less affected and leads to higher accuracy than the LSTM does. This result suggests that incorporating the water balance constraint into the LSTM enhances the robustness against data sparsity. Under 3 years' training data, the median of the mean of KGE for the LSTM across the 531 catchments is 0.493 and the median for the MC-LSTM is 0.580. According to the two-sided Wilcoxon signed-rank test at the significance level of 0.05, the mean of KGE for the MC-LSTM is significantly higher than that for the LSTM in 55% (291 in 531) of the catchments. On the other hand, when there are more than 6 years' training data, the LSTM network exhibits similar accuracy to the MC-LSTM and EXP-HYDRO models. Under 9 years' training data, the median of the mean of KGE for the LSTM across the 531 catchments is 0.649 and the median for the MC-LSTM is 0.660. The mean of KGE for the MC-LSTM is significantly higher than that for the LSTM in 37% (196 in 531) of the catchments.

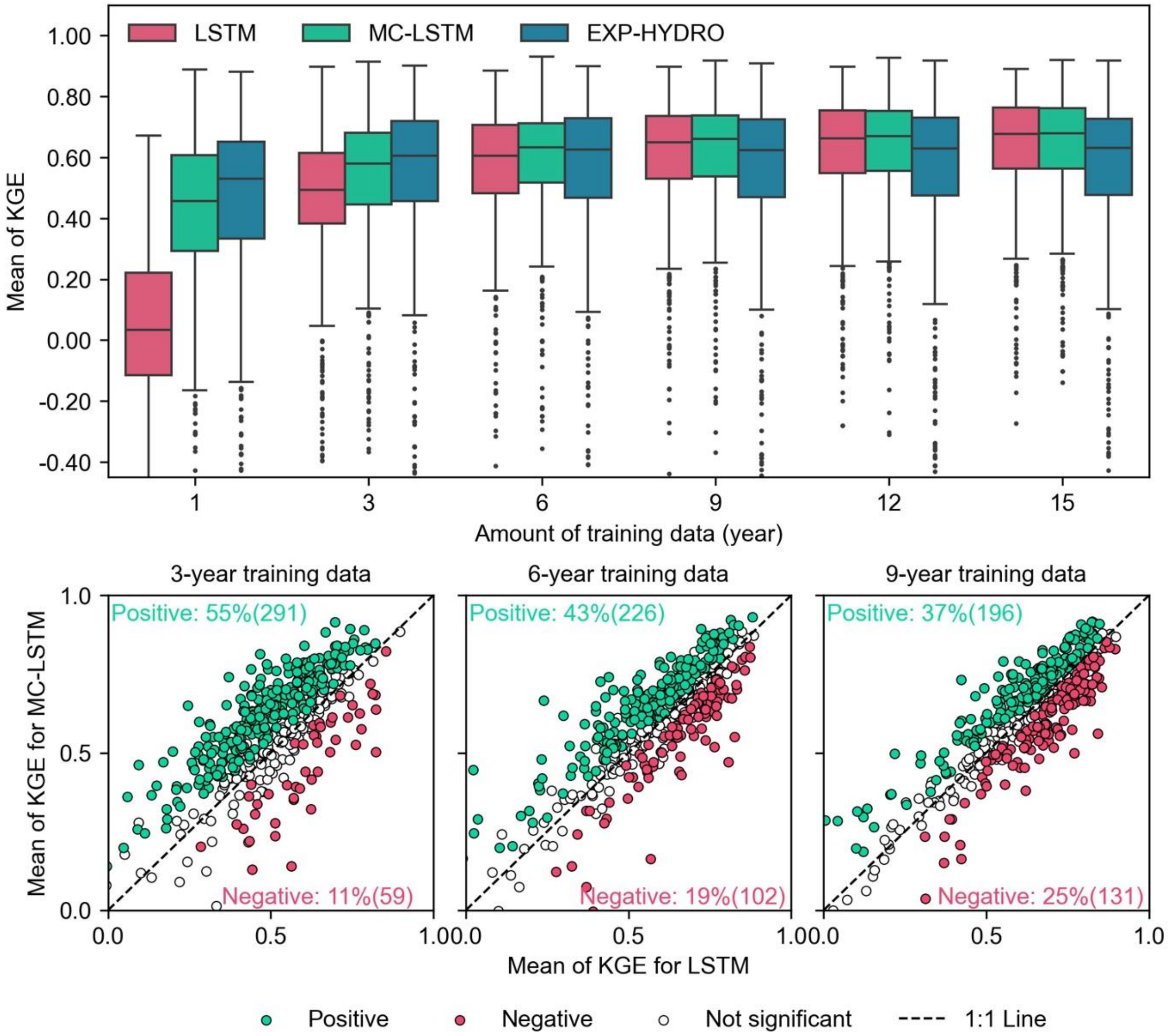


Figure 2. Mean values of KGE for the LSTM, MC-LSTM and EXP-HYDRO models across the 531 catchments. Positive, Negative and Not significant denote catchments where the accuracy is significantly improved, deteriorated and not significantly affected by incorporating the water balance constraint, respectively.

225

In the testing period, the hydrographs for three case study catchments are shown in Fig. 3. When there are only 3 years' training data, the LSTM exhibits the widest uncertainty range. As the amount of training data increases from 3 to 15 years, the accuracy of the LSTM increases and the uncertainty reduces; the accuracy and uncertainty for the MC-LSTM and EXP-HYDRO models change marginally. This outcome indicates that the incorporation of the water balance constraint into the LSTM strengthens the robustness against data sparsity. In the catchment 14316700, the MC-LSTM and EXP-HYDRO models outperform the LSTM. When there are 3 years' and 15 years' training data, the mean values of KGE for the MC- 230

LSTM are respectively 0.780 and 0.903; the mean values for the LSTM are respectively 0.678 and 0.872. In the catchment 01439500 and 11532500, the MC-LSTM and EXP-HYDRO models outperform the LSTM when there are 3 years' training data. The mean values of KGE for the MC-LSTM and LSTM are respectively 0.747 and 0.586 in the catchment 01439500. Under 15 years' training data, the LSTM exhibits similar accuracy and uncertainty to the MC-LSTM in the catchment 01439500; the LSTM presents similar accuracy but more uncertainty than the MC-LSTM in the catchment 11532500.

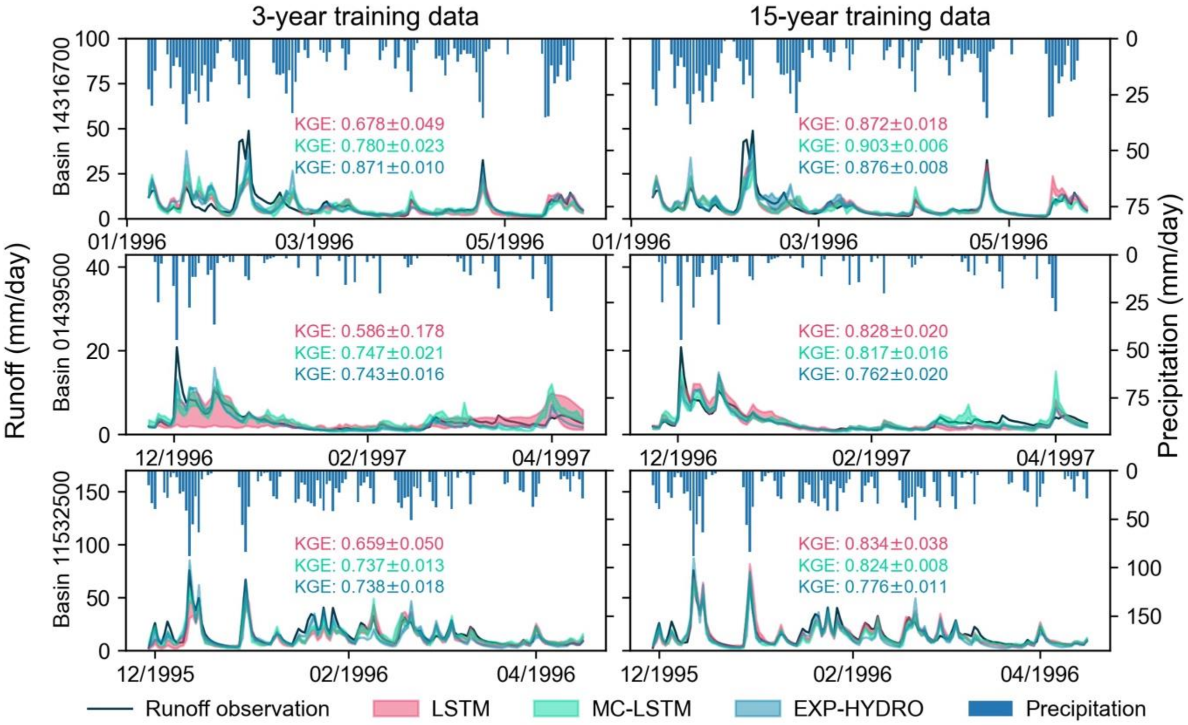


Figure 3. Hydrographs for three catchments under 3 and 15 years' training data. The bands of runoff represent the uncertainty ranges of 10-member ensemble models under different random seeds.

4.2 Robustness against parameter initialization

Across the 531 catchments, the standard deviation of KGE for the LSTM, MC-LSTM and EXP-HYDRO models is summarized in Fig. 4. Overall, the standard deviation of KGE for the three models decreases as the amount of training data increases. As the amount increases from 1 to 15 years, the LSTM network witnesses the largest reduction in the standard deviation of KGE; and the median value across the 531 catchments decreases from 0.146 to 0.037. By contrast, the MC-LSTM and EXP-HYDRO models exhibit slighter reductions in the median from 0.044 to 0.028 and from 0.044 to 0.020, respectively. In the meantime, the LSTM network leads to higher standard deviation of KGE than the MC-LSTM and EXP-

HYDRO models do under all the different amounts of training data. When there are 3 years' training data, the median of the standard deviation of KGE for the LSTM across the 531 catchments is 0.073 and the median for the MC-LSTM is 0.036. When there are 9 years' training data, the median values for the LSTM and MC-LSTM are respectively 0.038 and 0.029. These results indicate that the water balance constraint enhances the robustness of the LSTM against parameter initialization.

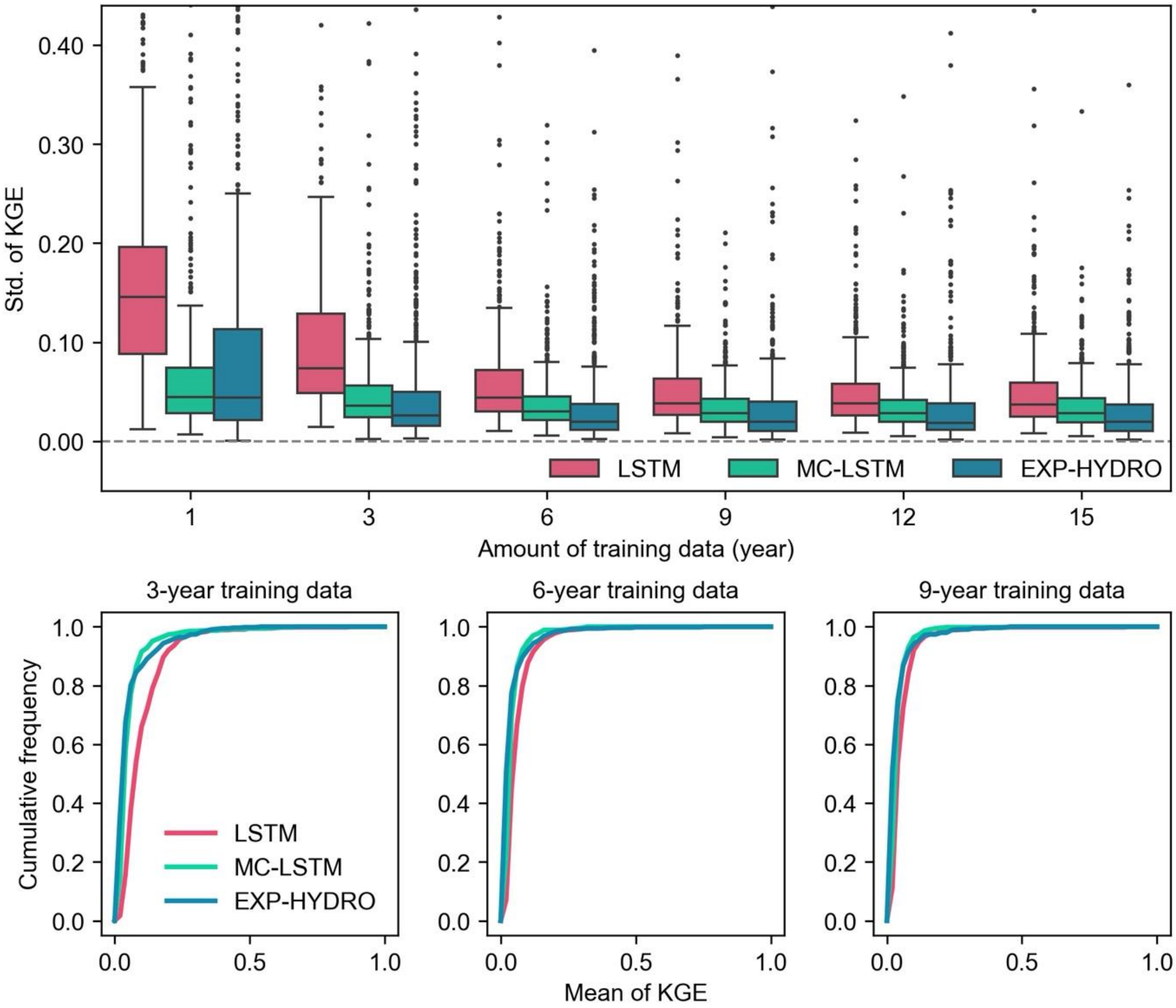
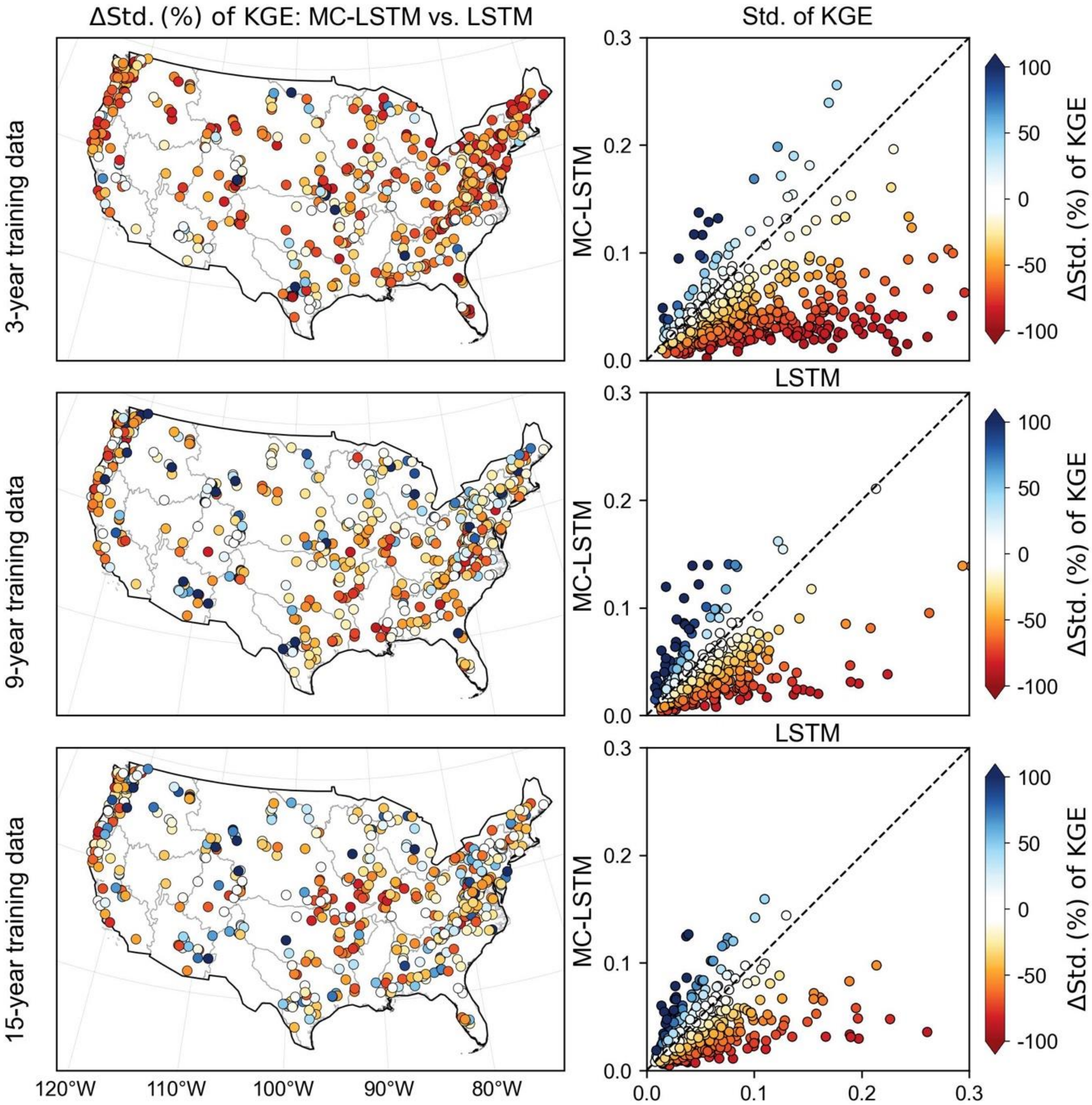


Figure 4. Standard deviation of KGE for the LSTM, MC-LSTM and EXP-HYDRO models under different amounts of training data across the 531 catchments.

The differences between the LSTM and MC-LSTM in terms of robustness across the 531 catchments are shown in Fig. 5. Overall, the MC-LSTM network exhibits smaller standard deviation of KGE than the LSTM network. When the models are trained with 3, 9 and 15 years' data, the standard deviation of KGE for the MC-LSTM is smaller than that for the LSTM in 450 (85%), 386 (73%) and 366 (69%) catchments, respectively. The basins where the MC-LSTM is more robust than the LSTM scatter throughout the contiguous United States without specific patterns of spatial distribution. These results suggest that incorporating the water balance constraint into the LSTM network improves the robustness against parameter initialization. In the meantime, as the amount of training data increases from 3 to 15 years, the number of catchments where the MC-LSTM is more robust than the LSTM decreases from 450 to 366. As shown in Fig.4 and Fig.5, the differences in the standard deviation of KGE between the MC-LSTM and LSTM also decrease as the amount of training data increases. The implication is that increasing training data can compensate for the lack of robustness in the LSTM network due to the lack of the water balance constraint.



270

Figure 5. Differences in the standard deviation of KGE between the MC-LSTM and LSTM across the 531 catchments.

4.3 Robustness in learning rainfall-runoff relationships

For the 50 case study catchments, the contributions of input features are showcased in Fig. 6. Overall, precipitation 275 contributes the most to the LSTM and MC-LSTM networks among the six input features. As the amount of training data increases from 1 year to 15 years, the median contribution of precipitation to the LSTM increases substantially; by contrast, the median contribution to the MC-LSTM changes marginally. Under 3 years' training data, the median contributions of precipitation to the LSTM and the MC-LSTM are respectively 39.6% and 49.4%. Under 9 years' training data, the median contributions of precipitation to the LSTM and the MC-LSTM are respectively 45.8% and 47.3.%. These results indicate that the water balance constraint enhances the robustness in learning rainfall-runoff relationships. As for the other five input features, solar radiation and vapor pressure make remarkable contributions to the MC-LSTM. Specifically, under 15 years' training data, the median contributions of solar radiation and vapor pressure to the MC-LSTM are respectively 26.0% and 24.4%. This difference implies that the water balance constraint could make the MC-LSTM tend to estimate evapotranspiration based on the energy budget (Wi and Steinschneider, 2024).

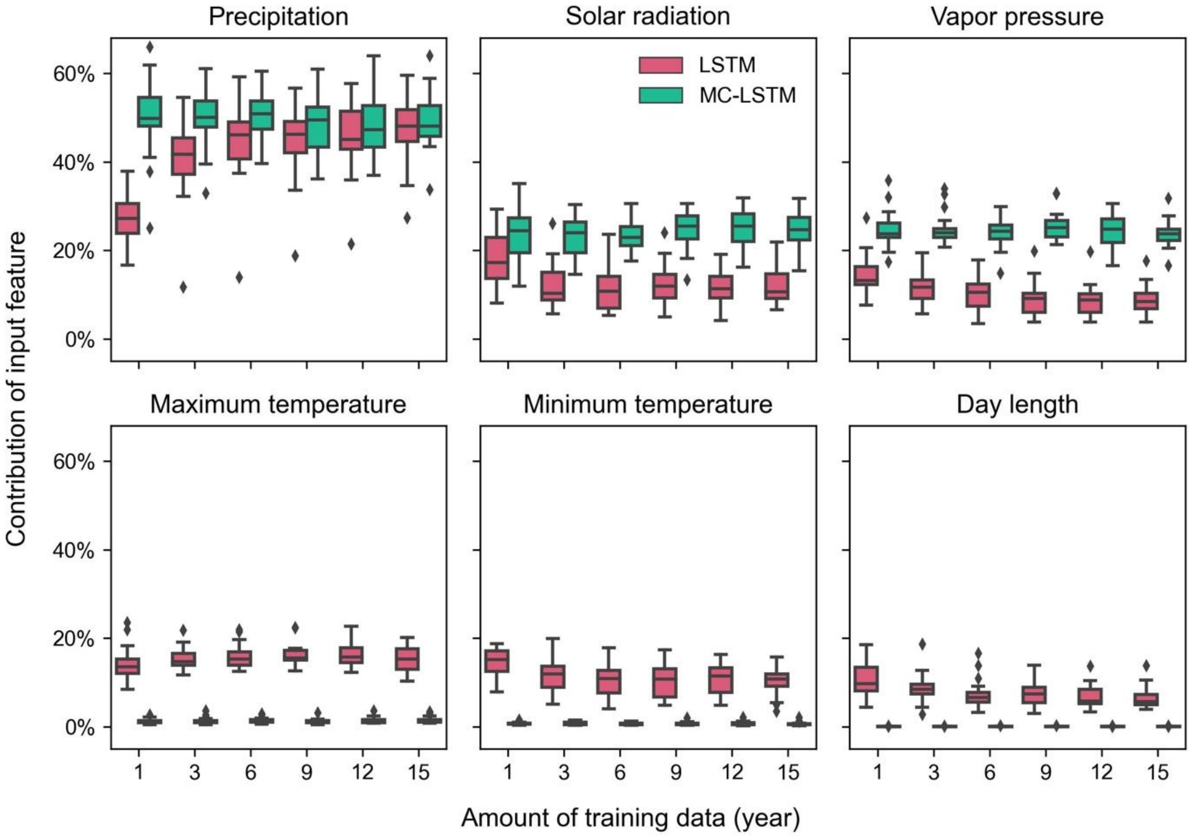


Figure 6. Contributions of input features to the LSTM and MC-LSTM networks for 20 case study catchments.

For the catchment 14316700, the cumulative contributions and seasonal mean contributions of input features are presented in Fig. 7. When there are 6 years' and 15 years' training data, the contributions of all input features to the LSTM scatter throughout the lookback period and exhibit much variability; by contrast, the contributions to the MC-LSTM are concentrated in the last 100 days and remain more stable (Fig. 7a). Among the six input features, precipitation contributes the most to the LSTM and MC-LSTM. Compared with the LSTM, the MC-LSTM focus more on the information from solar radiation and vapor pressure (Fig. 7b). As the amount of training data increases from 6 year to 15 years, the contribution of precipitation increases from 38.4% to 46.4% for the LSTM but changes little for the MC-LSTM. These results indicate that the water balance constraint improves the robustness in learning rainfall-runoff relationships. Under 15 years' training data, the contributions of precipitation to the LSTM and MC-LSTM are more than 41.1% in winter, spring and autumn but less than 35.0% in summer. This result aligns with the temporal patterns of precipitation in this catchment that there is more precipitation from November to April but less precipitation in summer. The implication is that both the LSTM and MC-LSTM may have learned the rainfall-runoff relationships.

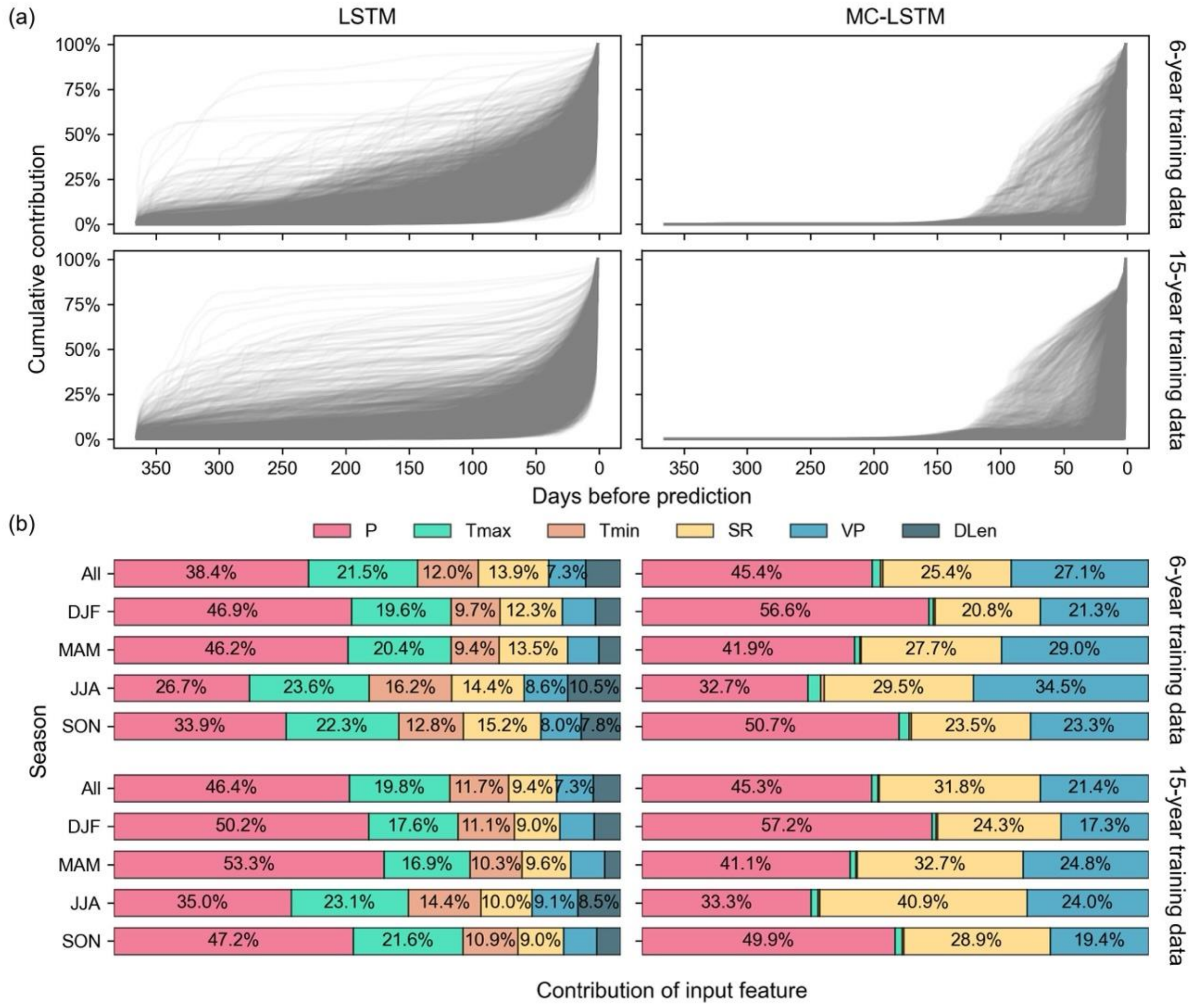


Figure 7. (a) Cumulative contributions and (b) seasonal mean contributions of input features to the LSTM and MC-LSTM networks for catchment 14316700. P, Tmax, Tmin, SR, VP and DLen are short for daily precipitation, daily maximum temperature, daily minimum temperature, solar radiation, vapor pressure and day length, respectively.

5 Discussion

The improvements of the robustness of the LSTM network due to the water balance constraint are assessed by considering different model hyperparameters. In the supplement, the results across 50 randomly selected catchments are presented. As shown in Figs. S3 and S4, the incorporation of the water balance constraint evidently improves the robustness against data sparsity in 20% (10 in 50) of the catchments when there are 9 years' training data. As shown in Figs. S5 and S6, the MC-LSTM is more robustness against parameter initialization in 86% (43 in 50) of the catchments when there are 9 years' training data, respectively. As shown in Figs. S7 and S8, incorporating this constraint improves the robustness in learning rainfall-runoff relationships by increasing the median contribution of precipitation from 40.5% to 50.6% when there are 9 years' training data. These results show that the robustness improved by the water balance constraint exhibits similar patterns under different hyperparameters, which highlights the reliability of the results in this paper.

TGDS strategies, such as incorporating physical constraints into model architectures (Hoedt et al., 2021; Li et al., 2024), reconfiguring loss functions with physical penalties (Wang et al., 2023; Yang et al., 2020) and pretraining models with synthetic data (Xie et al., 2021; Zhang et al., 2022), have previously been shown to be effective in enhancing robustness against data sparsity (Karniadakis et al., 2021). In this paper, the incorporation of the water balance constraint into the architecture of the LSTM is shown to enhance the robustness against data sparsity and parameter initialization at the local scale. In the meantime, the XAI method is employed to understand the exact role of this constraint in the LSTM, which can complement and explain the effects of this constraint on the robustness from the perspective of learning rainfall-runoff relationships (Jiang et al., 2022a). Large-sample tests help to understand model limitations and draw reliable conclusions from a broad perspective (Addor et al., 2020; Huang et al., 2023).

For DL models, physical constraints can be effective for local models but be less effective for regional models (Frame et al., 2023, 2022; Xie et al., 2021). This outcome can be attributed to the fact that pure DL models are more flexible to capture patterns inside observations that may be of inconsistent water balance, compared to models strictly constrained by the water balance (Kratzert et al., 2024; Frame et al., 2023; Beven, 2020). Besides, catchments with similar characteristics may have similar rainfall-runoff relationships of which DL models can take advantage (Xie et al., 2021; Bertola et al., 2023). Recent studies have illustrated that the LSTM performs better when trained with data from a large amount of hydrologically diverse catchments than with data from a single catchment (Kratzert et al., 2024). Large-sample hydrology is thus expected to enhance the performances of DL models in predictions of extreme events and projections under climate change (Bertola et al., 2023; Wi and Steinschneider, 2022, 2024; Gupta et al., 2014).

6 Conclusions

This paper is concentrated on the beneficial part that the water balance plays in the robustness of the LSTM network for rainfall-runoff prediction. That is, large-sample tests based on the CAMELS dataset are performed to assess the robustness of the LSTM and MC-LSTM from three perspectives, i.e., data sparsity, parameter initialization and learning rainfall-runoff relationships. In the meantime, the IG method is used to interrogate the rainfall-runoff relationships learned by the LSTM and MC-LSTM. The results highlight that incorporating the water balance constraint into the LSTM improves the robustness and that the improvement decreases as the amount of training data increases. One finding is that the robustness against data sparsity is significantly enhanced by this constraint in 37% (196 in 531) of the catchments when there are 9 years' training data. Another finding is that the robustness against parameter initialization is improved by this constraint in 73% (386 in 531) of the catchments when there are 9 years' training data. In addition, it is found that the robustness in learning rainfall-runoff relationships is enhanced by this constraint, resulting in an increase in the median contribution of precipitation from 45.8% to 47.3% when there are 9 years' training data. These results are associated with the compensation effects between data and knowledge on performances of DL models. Considering data from a large amount of hydrologically diverse catchments, the improved robustness of TGDS models deserves scrutiny. The in-depth investigations of this paper facilitate insights into the use of the LSTM network for rainfall-runoff modelling.

Data availability. The CAMELS dataset is download from https://gdex.ucar.edu/dataset/camels.html (Addor et al., 2017).

Author contribution. TZ and QL designed the experiments. QL carried them out. QL and TZ developed the model code and performed the simulations. TZ and QL prepared the manuscript.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Acknowledgements. This paper is supported by the National Natural Science Foundation of China (2023YFF0804900 and 52379033) and the Guangdong Provincial Department of Science and Technology (2019ZT08G090).

References

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrol. Earth Syst. Sci., 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, 2017.

Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K., and Mendoza, P. A.: Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges, Hydrol. Sci. J., 65, 712–725, https://doi.org/10.1080/02626667.2019.1683182, 2020.

Bertola, M., Blöschl, G., Bohac, M., Borga, M., Castellarin, A., Chirico, G. B., Claps, P., Dallan, E., Danilovich, I., Ganora, D., Gorbachova, L., Ledvinka, O., Mavrova-Guirguinova, M., Montanari, A., Ovcharuk, V., Viglione, A., Volpi, E., Arheimer, B., Aronica, G. T., Bonacci, O., Čanjevac, I., Csik, A., Frolova, N., Gnandt, B., Gribovszki, Z., Gül, A., Günther, K., Guse, B., Hannaford, J., Harrigan, S., Kireeva, M., Kohnová, S., Komma, J., Kriauciuniene, J., Kronvang, B., Lawrence, D., Lüdtke, S., Mediero, L., Merz, B., Molnar, P., Murphy, C., Oskoruš, D., Osuch, M., Parajka, J., Pfister, L., Radevski, I., Sauquet, E., Schröter, K., Šraj, M., Szolgay, J., Turner, S., Valent, P., Veijalainen, N., Ward, P. J., Willems, P., and Zivkovic, N.: Megafloods in Europe can be anticipated from observations in hydrologically similar catchments, Nat. Geosci., 16, 982–988, https://doi.org/10.1038/s41561-023-01300-5, 2023.

Beven, K.: Deep learning, hydrological processes and the uniqueness of place, Hydrol. Process., 34, 3608–3613, https://doi.org/10.1002/hyp.13805, 2020.

Bhasme, P., Vagadiya, J., and Bhatia, U.: Enhancing predictive skills in physically-consistent way: physics informed machine learning for hydrological processes, J. Hydrol., 615, 128618, https://doi.org/10.1016/j.jhydrol.2022.128618, 2022.

Cai, H., Liu, S., Shi, H., Zhou, Z., Jiang, S., and Babovic, V.: Toward improved lumped groundwater level predictions at catchment scale: Mutual integration of water balance mechanism and deep learning method, Journal of Hydrology, 613, 128495, https://doi.org/10.1016/j.jhydrol.2022.128495, 2022.

Faghmous, J. H., Banerjee, A., Shekhar, S., Steinbach, M., Kumar, V., Ganguly, A. R., and Samatova, N.: Theory-guided data science for climate change, Computer, 47, 74–78, https://doi.org/10.1109/MC.2014.335, 2014.

Feng, D., Fang, K., and Shen, C.: Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales, Water Resour. Res., 56, e2019WR026793, https://doi.org/10.1029/2019WR026793, 2020.

Feng, D., Lawson, K., and Shen, C.: Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data, Geophys. Res. Lett., 48, e2021GL092999, https://doi.org/10.1029/2021GL092999, 2021.

Feng, D., Liu, J., Lawson, K., and Shen, C.: Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy, Water Resour. Res., 58, e2022WR032404, https://doi.org/10.1029/2022WR032404, 2022.

Frame, J. M., Kratzert, F., Raney II, A., Rahman, M., Salas, F. R., and Nearing, G. S.: 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, 885–905, https://doi.org/10.1111/1752-1688.12964, 2021.

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

Frame, J. M., Kratzert, F., Gupta, H. V., Ullrich, P., and Nearing, G. S.: On strictly enforced mass conservation constraints for modelling the rainfall-runoff process, Hydrol. Process., 37, e14847, https://doi.org/10.1002/hyp.14847, 2023.

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., and Hochreiter, S.: Rainfall–runoff prediction at multiple timescales with a single long short-term memory network, Hydrol. Earth Syst. Sci., 25, 2045–2062, https://doi.org/10.5194/hess-25-2045-2021, 2021a.

Gauch, M., Mai, J., and Lin, J.: The proper care and feeding of CAMELS: how limited training data affects streamflow prediction, Environ. Modell. Softw., 135, 104926, https://doi.org/10.1016/j.envsoft.2020.104926, 2021b.

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

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., and Andréassian, V.: Large-sample hydrology: a need to balance depth with breadth, Hydrol. Earth Syst. Sci., 18, 463–477, https://doi.org/10.5194/hess-18-463-2014, 2014.

Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural Comput., 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G. S., Hochreiter, S., and Klambauer, G.: MC-LSTM: mass-conserving LSTM, in: Proceedings of the 38th International Conference on Machine Learning, International Conference on Machine Learning, 4275–4286, 2021.

Huang, Z., Zhao, T., Tian, Y., Chen, X., Duan, Q., and Wang, H.: Reliability of ensemble climatological forecasts, Water Resour. Res., 59, e2023WR034942, https://doi.org/10.1029/2023WR034942, 2023.

Jiang, S., Zheng, Y., and Solomatine, D.: Improving AI system awareness of geoscience knowledge: symbiotic integration of physical approaches and deep learning, Geophys. Res. Lett., 47, e2020GL088229, https://doi.org/10.1029/2020GL088229, 2020.

Jiang, S., Bevacqua, E., and Zscheischler, J.: River flooding mechanisms and their changes in Europe revealed by explainable machine learning, Hydrol. Earth Syst. Sci., 26, 6339–6359, https://doi.org/10.5194/hess-26-6339-2022, 2022a.

Jiang, S., Zheng, Y., Wang, C., and Babovic, V.: Uncovering flooding mechanisms across the contiguous United States through interpretive deep learning on representative catchments, Water Resour. Res., 58, https://doi.org/10.1029/2021WR030185, 2022b.

Kao, I.-F., Zhou, Y., Chang, L.-C., and Chang, F.-J.: Exploring a long short-term memory based encoder-decoder framework for multi-step-ahead flood forecasting, J. Hydrol., 583, 124631, https://doi.org/10.1016/j.jhydrol.2020.124631, 2020.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L.: Physics-informed machine learning, Nat Rev Phys, 3, 422–440, https://doi.org/10.1038/s42254-021-00314-5, 2021.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., and Kumar, V.: Theory-guided data science: a new paradigm for scientific discovery from data, IEEE Trans. Knowl. Data Eng., 29, 2318–2331, https://doi.org/10.1109/TKDE.2017.2720168, 2017.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall-runoff modelling using long short-term memory (LSTM) networks, Hydrol. Earth Syst. Sci., 22, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018, 2018.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward improved predictions in ungauged basins: exploiting the power of machine learning, Water Resour. Res., 55, 11344–11354, https://doi.org/10.1029/2019WR026065, 2019a.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrol. Earth Syst. Sci., 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019b.

Kratzert, F., Klotz, D., Hochreiter, S., and Nearing, G. S.: A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall-runoff modeling, Hydrol. Earth Syst. Sci., 25, 2685–2703, https://doi.org/10.5194/hess-25-2685-2021, 2021.

Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS Opinions: Never train an LSTM on a single basin, Hydrol. Earth Syst. Sci. Discuss., 2024, 1–19, https://doi.org/10.5194/hess-2023-275, 2024.

LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, Nature, 521, 436–444, https://doi.org/10.1038/nature14539, 2015.

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

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

Li, L., Dai, Y., Wei, Z., Shangguan, W., Zhang, Y., Wei, N., and Li, Q.: Enforcing Water Balance in Multitask Deep Learning Models for Hydrological Forecasting, J. Hydrometeorol., 25, 89–103, https://doi.org/10.1175/JHM-D-23-0073.1, 2024.

Manure, A., Bengani, S., and S, S.: Robustness and reliability, in: Introduction to Responsible AI: Implement Ethical AI Using Python, edited by: Manure, A., Bengani, S., and S, S., Apress, Berkeley, CA, 133–158, 2023.

Nearing, G., Kratzert, F., Klotz, D., Hoedt, P.-J., Klambauer, G., Hochreiter, S., and Gupta, H.: A deep learning architecture for conservative dynamical systems: application to rainfall-runoff modeling, in: AI for Earth Sciences Workshop, NeurIPS 2020, 2020.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T. Y., Weitzner, D., and Matias, Y.:

Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, https://doi.org/10.1038/s41586-024-07145-1, 2024.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What role does hydrological science play in the age of machine learning?, Water Resour. Res., 57, e2020WR028091, https://doi.org/10.1029/2020WR028091, 2021.

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

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

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems, 32, 2019.

Patil, S. and Stieglitz, M.: Modelling daily streamflow at ungauged catchments: what information is necessary?, Hydrol. Process., 28, 1159–1169, https://doi.org/10.1002/hyp.9660, 2014.

Pokharel, S., Roy, T., and Admiraal, D.: Effects of mass balance, energy balance, and storage-discharge constraints on LSTM for streamflow prediction, Environ. Modell. Softw., 166, 105730, https://doi.org/10.1016/j.envsoft.2023.105730, 2023.

Razavi, S.: Deep learning, explained: fundamentals, explainability, and bridgeability to process-based modelling, Environ. Modell. Softw., 144, 105159, https://doi.org/10.1016/j.envsoft.2021.105159, 2021.

Read, J. S., Jia, X., Willard, J., Appling, A. P., Zwart, J. A., Oliver, S. K., Karpatne, A., Hansen, G. J. A., Hanson, P. C., Watkins, W., Steinbach, M., and Kumar, V.: Process-guided deep learning predictions of lake water temperature, Water Resour. Res., 55, 9173–9190, https://doi.org/10.1029/2019WR024922, 2019.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven earth system science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Shen, C.: A transdisciplinary review of deep learning research and its relevance for water resources scientists, Water Resour. Res., 54, 8558–8593, https://doi.org/10.1029/2018WR022643, 2018.

Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic Attribution for Deep Networks, in: Proceedings of the 34th International Conference on Machine Learning, International Conference on Machine Learning, 3319–3328, 2017.

Thornton, P. E., Running, S. W., and White, M. A.: Generating surfaces of daily meteorological variables over large regions of complex terrain, J. Hydrol., 190, 214–251, https://doi.org/10.1016/S0022-1694(96)03128-9, 1997.

Topp, S. N., Barclay, J., Diaz, J., Sun, A. Y., Jia, X., Lu, D., Sadler, J. M., and Appling, A. P.: Stream Temperature Prediction in a Shifting Environment: Explaining the Influence of Deep Learning Architecture, Water Resour. Res., 59, e2022WR033880, https://doi.org/10.1029/2022WR033880, 2023.

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., and Shen, C.: From calibration to parameter learning: harnessing the scaling effects of big data in geoscientific modeling, Nat. Commun., 12, 5988, https://doi.org/10.1038/s41467-021-26107-z, 2021.

Wang, Y., Wang, W., Ma, Z., Zhao, M., Li, W., Hou, X., Li, J., Ye, F., and Ma, W.: A deep learning approach based on physical constraints for predicting soil moisture in unsaturated zones, Water Resour. Res., 59, e2023WR035194, https://doi.org/10.1029/2023WR035194, 2023.

Wi, S. and Steinschneider, S.: Assessing the physical realism of deep learning hydrologic model projections under climate change, Water Resour. Res., 58, e2022WR032123, https://doi.org/10.1029/2022WR032123, 2022.

Wi, S. and Steinschneider, S.: On the need for physical constraints in deep learning rainfall–runoff projections under climate change: a sensitivity analysis to warming and shifts in potential evapotranspiration, Hydrol. Earth Syst. Sci., 28, 479–503, https://doi.org/10.5194/hess-28-479-2024, 2024.

Xiang, Z., Yan, J., and Demir, I.: A Rainfall-Runoff Model With LSTM-Based Sequence-to-Sequence Learning, WATER RESOUR RES, 56, e2019WR025326, https://doi.org/10.1029/2019WR025326, 2020.

Xie, K., Liu, P., Zhang, J., Han, D., Wang, G., and Shen, C.: Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships, J. Hydrol., 603, 127043, https://doi.org/10.1016/j.jhydrol.2021.127043, 2021.

Yang, S., Yang, D., Chen, J., Santisirisomboon, J., Lu, W., and Zhao, B.: A physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data, J. Hydrol., 590, 125206, https://doi.org/10.1016/j.jhydrol.2020.125206, 2020.

Yin, H., Zhang, X., Wang, F., Zhang, Y., Xia, R., and Jin, J.: Rainfall-runoff modeling using LSTM-based multi-state-vector sequence-to-sequence model, J. Hydrol., 598, 126378, https://doi.org/10.1016/j.jhydrol.2021.126378, 2021a.

Yin, H., Guo, Z., Zhang, X., Chen, J., and Zhang, Y.: Runoff predictions in ungauged basins using sequence-to-sequence models, J. Hydrol., 603, 126975, https://doi.org/10.1016/j.jhydrol.2021.126975, 2021b.

Zhang, D., Wang, D., Peng, Q., Lin, J., Jin, T., Yang, T., Sorooshian, S., and Liu, Y.: Prediction of the outflow temperature of large-scale hydropower using theory-guided machine learning surrogate models of a high-fidelity hydrodynamics model, J. Hydrol., 606, 127427, https://doi.org/10.1016/j.jhydrol.2022.127427, 2022.

Zhang, J., Chen, X., Khan, A., Zhang, Y., Kuang, X., Liang, X., Taccari, M. L., and Nuttall, J.: Daily runoff forecasting by deep recursive neural network, J. Hydrol., 596, 126067, https://doi.org/10.1016/j.jhydrol.2021.126067, 2021.

Zhao, W. L., Gentine, P., Reichstein, M., Zhang, Y., Zhou, S., Wen, Y., Lin, C., Li, X., and Qiu, G. Y.: Physics-constrained machine learning of evapotranspiration, Geophys. Res. Lett., 46, 14496–14507, https://doi.org/10.1029/2019GL085291, 2019.

Zhong, L., Lei, H., and Gao, B.: Developing a Physics-Informed Deep Learning Model to Simulate Runoff Response to Climate Change in Alpine Catchments, Water Resour. Res., 59, e2022WR034118, https://doi.org/10.1029/2022WR034118, 2023.