Text S1: The calculation equations of the 5 flux variables in the EXP-HYDRO model

The EXP-HYDRO model is a conceptual and spatially lumped rainfall-runoff model developed by Patil and Stieglitz (2014). The physical equations and parameters are well introduced and organized by Text S1 in the Supporting Information of Jiang et al. (2020). For easy reading, the calculation equations of the 5 flux variables (M, ET, Q, P_s and P_r) are briefly introduced here.

P_s and P_r are respectively the daily snowfall (mm/day) and rainfall (mm/day), which are estimated by the daily precipitation (P, mm/day) and daily temperature (T, °C) as follows:

$$P_s = \mathrm{fun1}\left(P, T, T_{min}\right) = \begin{cases} 0, & T > T_{min} \\ P, & T \le T_{min} \end{cases} \tag{S1}$$

$$P_r = \mathrm{fun2}\left(P, T, T_{min}\right) = \begin{cases} P, & T > T_{min} \\ 0, & T \le T_{min} \end{cases} \tag{S2}$$

where T_{min} is a parameter representing the temperature threshold below which precipitation falls as snow.

The snowmelt (M, mm/day) is simulated by a simple thermal degree-day model related to T and the snow accumulation bucket (S_0) based on the following equation:

$$M = \mathrm{fun3}\left(S_0, T, D_f, T_{max}\right) = \begin{cases} \min\{S_0, D_f \cdot \left(T - T_{max}\right)\}, & T > T_{max} \text{ and } S_0 > 0 \\ 0 & \text{, otherwise} \end{cases} \tag{S3}$$

where D_f is a parameter denoting the thermal degree-day factor (mm/day/°C); T_{max} is another parameter representing the temperature threshold where the accumulated snow begins to melt.

The evapotranspiration is denoted by ET (mm/day), which is calculated as follows:

$$ET = \mathrm{fun4}\left(S_1, PET, S_{max}\right) = \begin{cases} 0 & S_1 < 0 \\ PET \cdot \left(\dfrac{S_1}{S_{max}}\right) & , 0 \le S_1 \le S_{max} \\ PET & S_1 > S_{max} \end{cases} \tag{S4}$$

where catchment bucket (S_1) denotes its current storage; S_{max} is a parameter representing the storage capacity of the catchment bucket; PET is the potential evapotranspiration (mm/day) estimated by Hamon’s formulation:

$$PET = 29.8 \cdot L_{day} \cdot \frac{e^*(t)}{T + 273.2} \tag{S5}$$

$$e^*(t) = 0.611 \cdot e^{17.3 \cdot T / (T + 237.3)} \tag{S6}$$

where L_{day} is the day length (hour).

The streamflow (Q) is estimated as the sum of the baseflow (Q_b) and the capacity-excess runoff (Q_s), which are respectively expressed as follows:

$$Q_b = \mathrm{fun5}\left(S_1, f, S_{max}, Q_{max}\right) = \begin{cases} 0 & S_1 < 0 \\ Q_{max} \cdot e^{-f \cdot \left(S_{max} - S_1\right)} & ,\ 0 \le S_1 \le S_{max} \\ Q_{max} & S_1 > S_{max} \end{cases} \tag{S7}$$

$$Q_s = \mathrm{fun6}\left(S_1, S_{max}\right) = \begin{cases} 0 & S_1 \le S_{max} \\ S_1 - S_{max}, & S_1 > S_{max} \end{cases} \tag{S8}$$

$$Q = Q_b + Q_s \tag{S9}$$

where f and Q_{max} are two parameters representing the decline rate of runoff (mm^{-1}) and the maximum subsurface runoff (mm/day), respectively.

Text S2: Hyperparameters optimization

There are two categories of hyperparameters to be optimized, including hyperparameters of model structure (such as hidden layer and hidden size) and training process (such as learning rate and batch size) (Li et al., 2024). To strike a balance between model performance and time cost, 50 catchments are selected randomly. For each catchment, the first 14 years (from 1 October 1980 to 30 September 1994) of the entire training period is set as the training period and the last year (from 1 October 1994 to 30 September 1995) is set as the validation period for hyperparameters optimization. Models are trained using the Adam optimizer and the early stopping strategy. For each hyperparameter setting, three repetitions with different random seeds are used to initialize the parameters. The mean Nash-Sutcliffe efficiency (NSE) on validation period over the 3 repetitions represents the validation performance. Hyperparameters setting with the highest median NSE scores on validation period over the 50 catchments is chosen as the optimal hyperparameters.

Firstly, the hyperparameters of model structure are optimized. Based on the model structure of Kratzert et al. (2018) and the results that a one-layer LSTM network is qualified to capture rainfall-runoff responses of a catchment (Kratzert et al., 2019, 2021), the number of hidden layers is set to 1; the range of the hidden size is set to [20，40，50，60, 80, 100]. The range of dropout rate is set to [0.2, 0.4, 0.6, 0.8]. Following other hyperparameters of Kratzert et al. (2018), the LSTM is developed with input sequence for the past T = 365 d, the mini-batch size of 512, the drop-out rate of 0.1 and the Adam optimizer with the learning rate of 0.0001. Secondly, the hyperparameters of training process is fine-tuned based on the optimal hyperparameters in the first step. The LSTM network is tuned with different batch sizes (128, 256, 512), different learning rates (0.1, 0.01, 0.001, 0.0001) and different learning rate decay (0.1, 0.3, 0.5, 0.7).

The optimal hyperparameters of the LSTM are shown by Table 1 in the main text. In order to compromise between maximum reducing the uncertainty caused by different numbers of model parameters and achieving potentially more powerful predictions, the hidden size of the MC-LSTM network is set to 50. The numbers of parameters for the MC-LSTM and LSTM differ by less than 0.1%. As the EXP-HYDRO model is a process-based model, there is no need for the DL wrapped EXP-HYDRO model to normalize their input variables or to set the hidden size and dropout rate. Excluding the hidden size and dropout rate, the MC-LSTM and EXP-HYDRO models have the same hyperparameters as the LSTM, as shown by Table 1. While optimizing the hyperparameters of the MC-LSTM can obtain better performance, the MC-LSTM uses some of the hyperparameters of the LSTM directly, instead of optimizing them separately. Furthermore, the sensitivity analysis of model hyperparameters is devised based on model hyperparameters from Frame et al. (2023, 2022). The hidden sizes of the LSTM and MC-LSTM are 256 and 64, respectively. The results of the sensitivity analysis are presented by Fig. S3 to S8 in the Supplement.

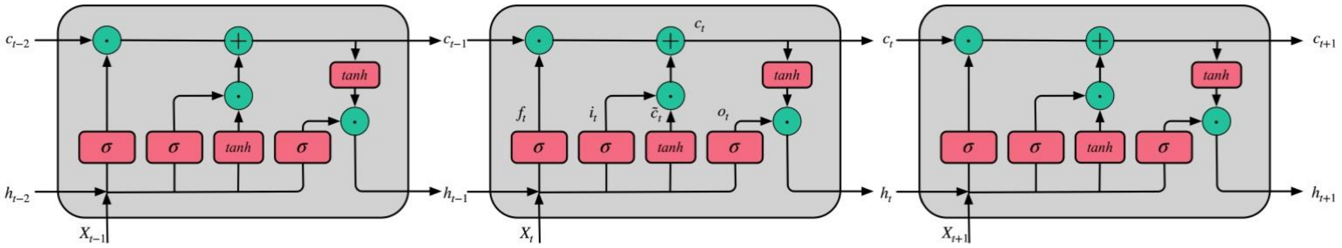

Figure S1. The internal operation of a standard LSTM network.

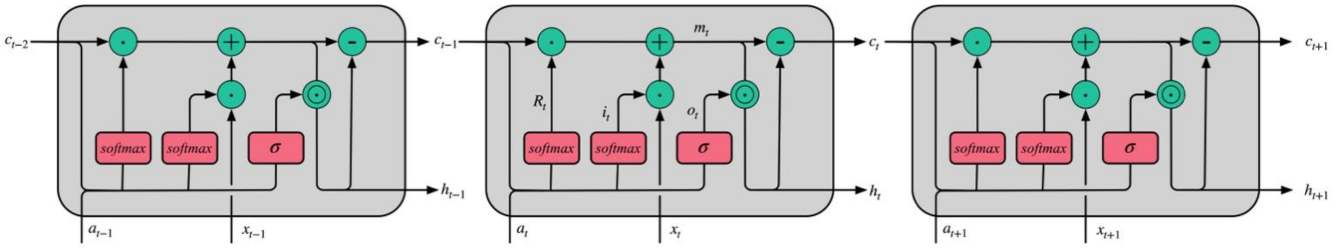

Figure S2. The internal operation of a MC-LSTM network.

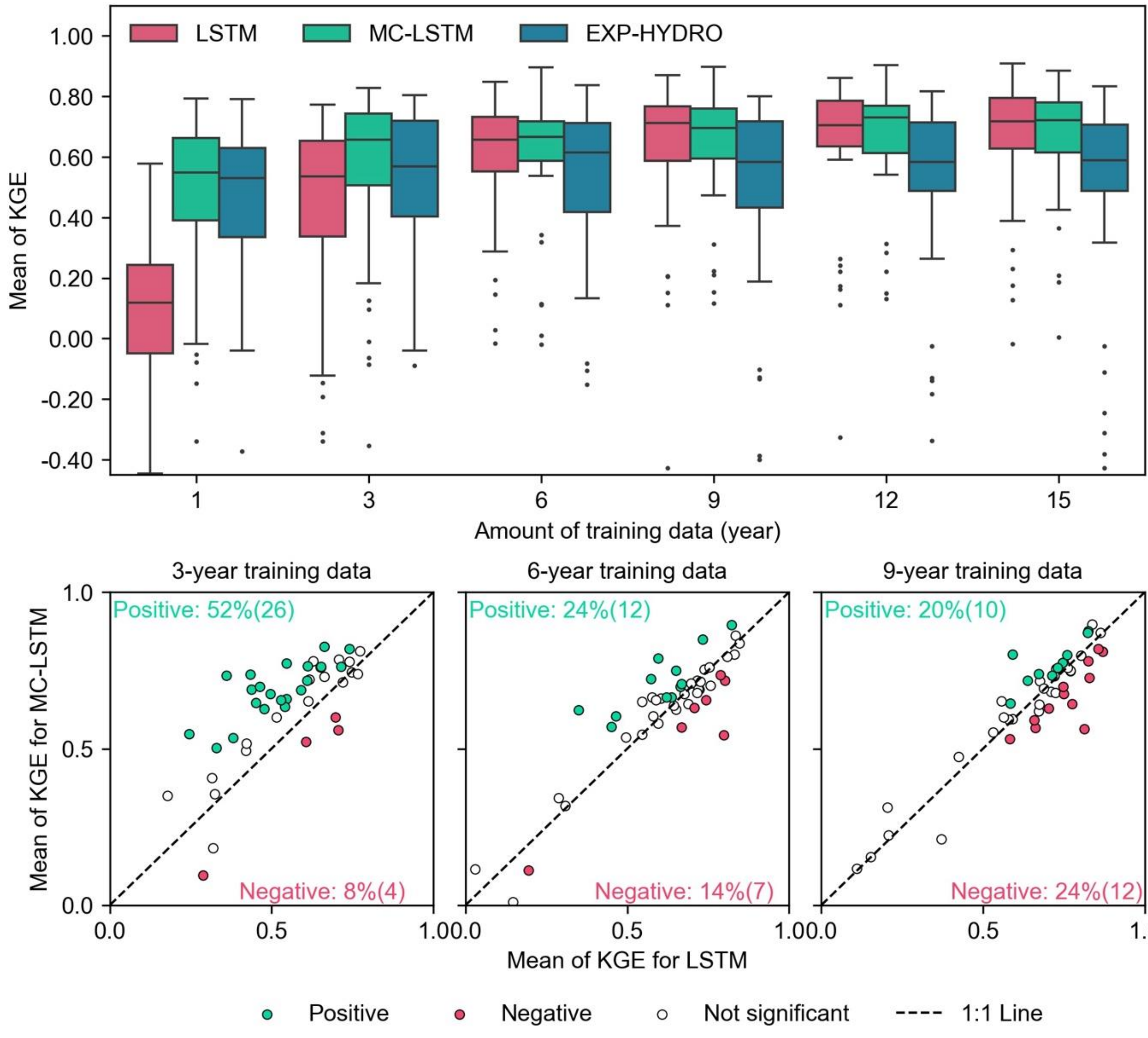

Figure S3. As for Fig. 2, but for the LSTM and MC-LSTM with hidden sizes of 256 and 64, respectively, in 50 randomly selected catchments.

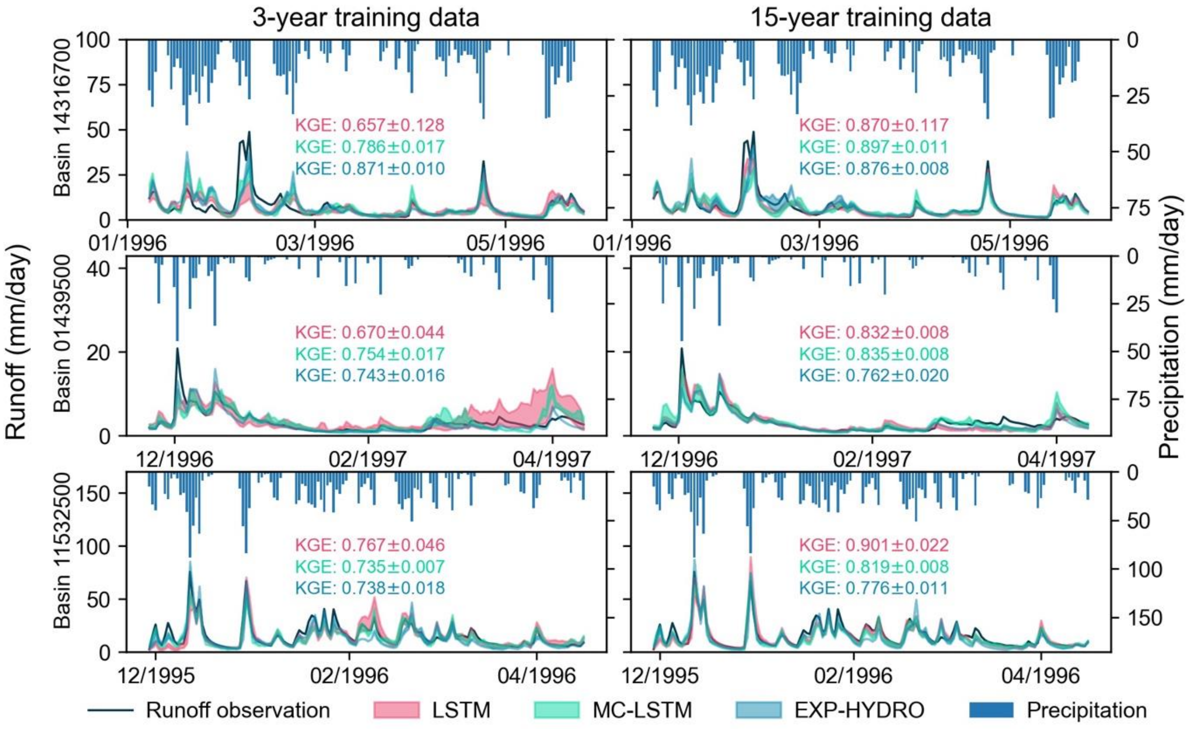

Figure S4. As for Fig. 3, but for the LSTM and MC-LSTM with hidden sizes of 256 and 64, respectively.

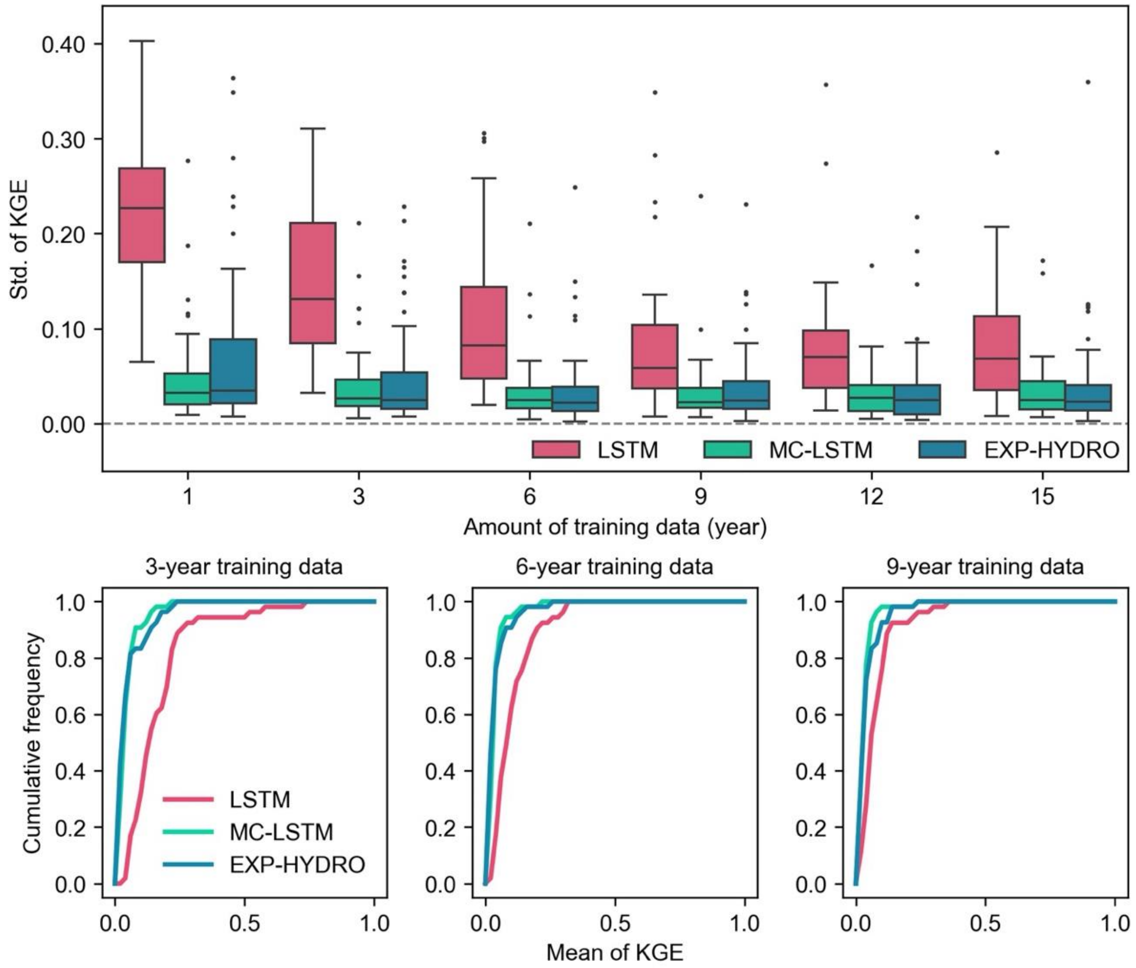

Figure S5. As for Fig. 4, but for the LSTM and MC-LSTM with hidden sizes of 256 and 64, respectively, in 50 randomly selected catchments.

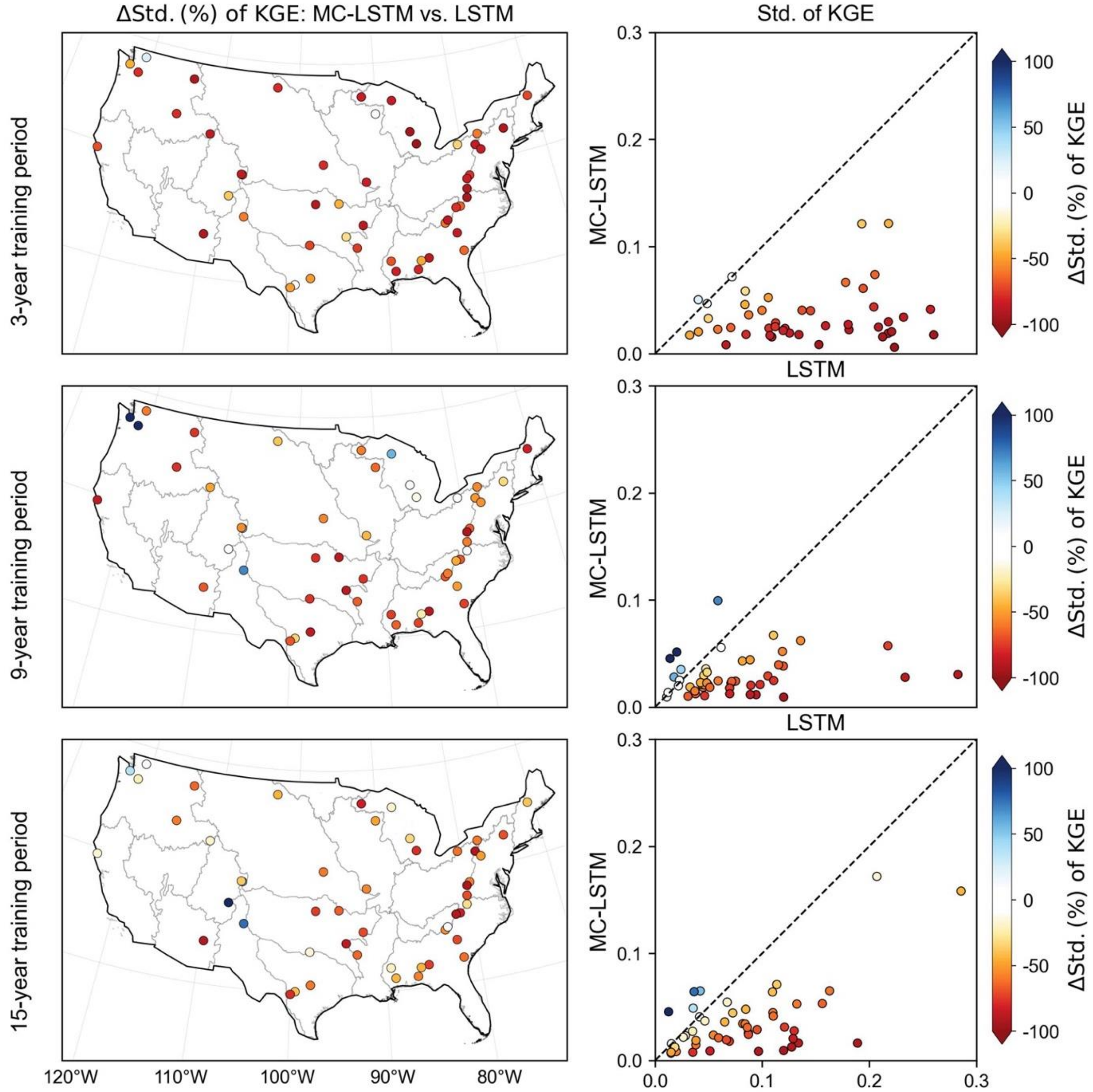

Figure S6. As for Fig. 5, but for the LSTM and MC-LSTM with hidden sizes of 256 and 64, respectively, in 50 randomly selected catchments.

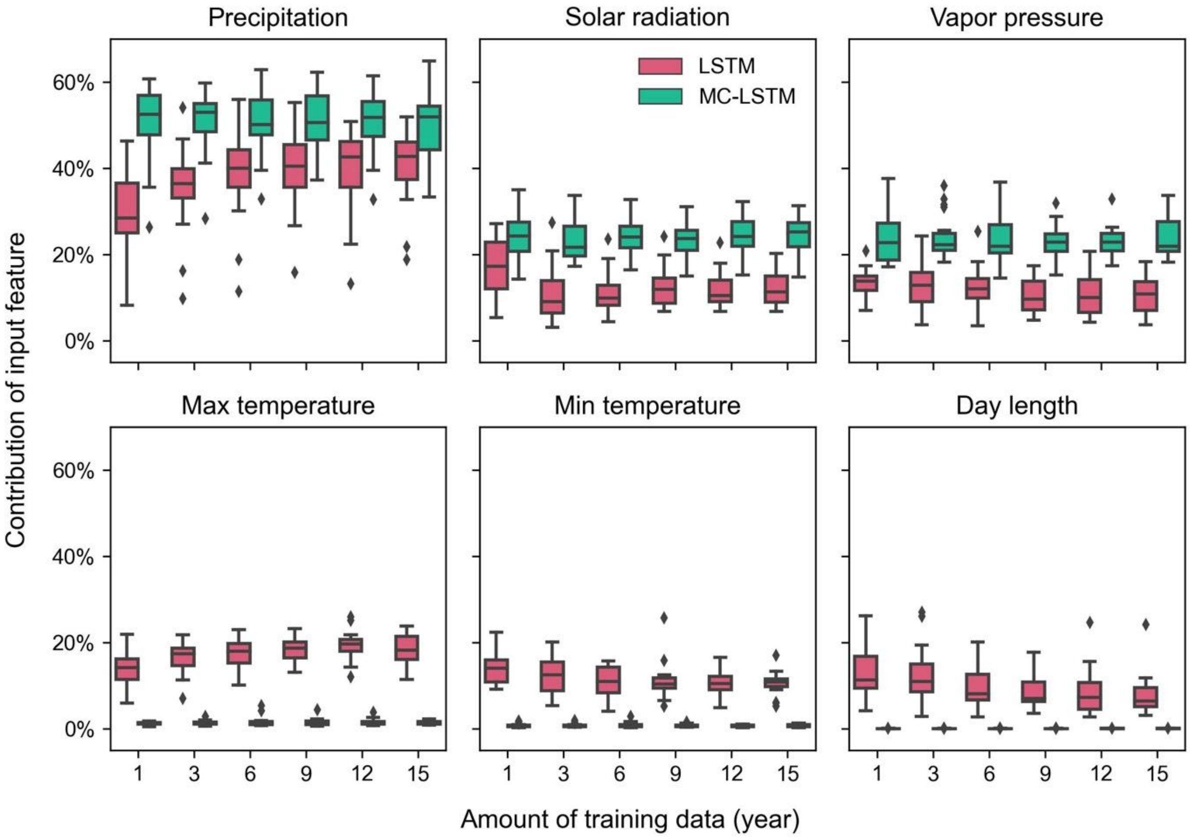

Figure S7. As for Fig. 6, but for the LSTM and MC-LSTM with hidden sizes of 256 and 64, respectively, in 24 case study catchments with the highest KGE.

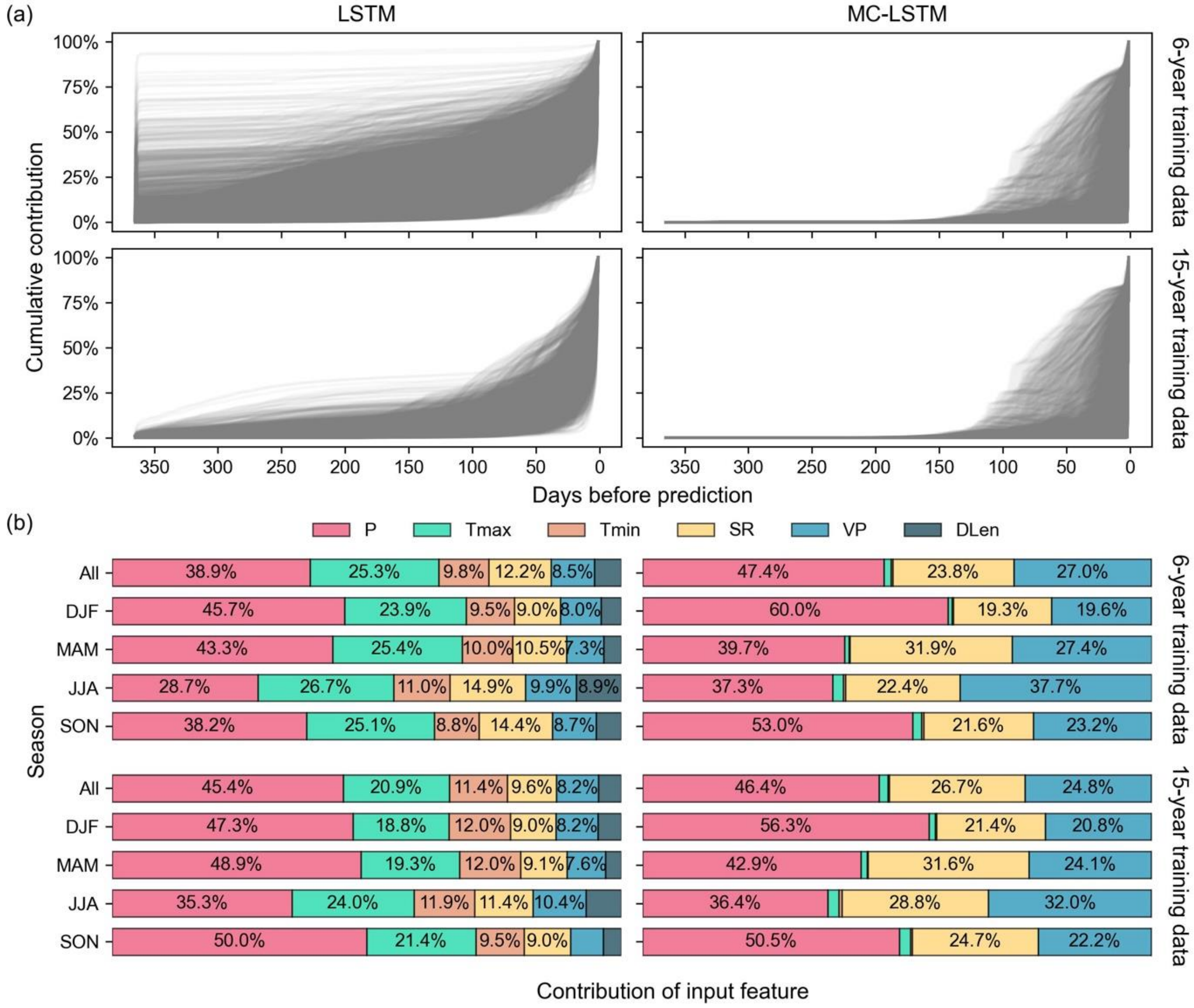

Figure S8. As for Fig. 7, but for the LSTM and MC-LSTM with hidden sizes of 256 and 64, respectively.

Table S1. Comparison of daily NSE statistics across the CAMELS catchments.

Model	Scale	Count of catchments	Dataset	Daily NSE statistics			Source
				median	mean	Proportion for NSE ≥0.55	
LSTM	Local	531	CAMELS	0.67	0.63	76%	This paper
MC-LSTM	Local	531	CAMELS	0.63	0.59	71%	This paper
EXP-HYDRO*	Local	531	CAMELS	0.49	0.42	40%	This paper
LSTM	Local	569	CAMELS	0.60	0.52	61.5%	Jiang et al. (2020)
EXP-HYDRO*	Local	569	CAMELS	0.48	-0.16	38.3%	Jiang et al. (2020)
LSTM	Local	241	CAMELS	0.65	0.63	NA	Kratzert et al. (2018)
EXP-HYDRO	Local	756	HCDN	NA	NA	~43% (>0.6)	Patil and Stieglitz (2014)
VIC	Local	531	CAMELS	0.57-0.59	NA	~56%	Newman et al. (2017)
LSTM	Regional	447	CAMELS	0.737	NA	NA	Hoedt et al. (2021)
MC-LSTM	Regional	447	CAMELS	0.726	NA	NA	Hoedt et al. (2021)

Local: Single model is trained for single catchment; Regional: Single model is trained for multiple catchments

HCDN: Hydro-Climate Data Network; VIC: Variable Infiltration Capacity model

EXP-HYDRO: Deep learning wrapped EXP-HYDRO model; NA: not available*

References

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.

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.

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.

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, 2019.

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.

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.

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.

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.