



RESEARCH ARTICLE

10.1029/2026JH001385

Mass-Conserving LSTM With Dual States for Streamflow Prediction: Separating Quickflow and Slow Storage

 Saurabh Toraskar¹ , M. Niranjannaik^{1,2}, Abhilash Singh^{1,3} , and Kumar Gaurav¹ 

¹Fluvial Geomorphology and Remote Sensing Laboratory, Department of Earth and Environmental Sciences, Indian Institute of Science Education and Research (IISER) Bhopal, Bhopal, India, ²Department of Earth Sciences, Indian Institute of Technology (IIT) Gandhinagar, Gandhinagar, India, ³School of Mathematics, Faculty of Engineering and Physical Sciences, University of Leeds, Leeds, UK

Key Points:

- The proposed model, MC-LSTM-DS, enforces mass balance and separates short and long memory for streamflow prediction
- Analyzing the internal states of MC-LSTM-DS gives insights into the dominant hydrological processes of the catchment
- MC-LSTM-DS outperforms LSTM and MC-LSTM on CAMELS-IND and matches LSTM while outperforming MC-LSTM on CAMELS-US across multiple metrics

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

A. Singh and K. Gaurav,
a.singh4@leeds.ac.uk;
abhilash.iiserb@gmail.com;
kgaurav@iiserb.ac.in

Citation:

Toraskar, S., Niranjannaik, M., Singh, A., & Gaurav, K. (2026). Mass-conserving LSTM with dual states for streamflow prediction: Separating quickflow and slow storage. *Journal of Geophysical Research: Machine Learning and Computation*, 3, e2026JH001385. <https://doi.org/10.1029/2026JH001385>

Received 3 APR 2026

Accepted 11 SEP 2026

Author Contributions:

Conceptualization: Saurabh Toraskar, Abhilash Singh, Kumar Gaurav

Data curation: M. Niranjannaik

Formal analysis: Saurabh Toraskar

Funding acquisition: Abhilash Singh, Kumar Gaurav

Investigation: Saurabh Toraskar

Methodology: Saurabh Toraskar

Resources: Kumar Gaurav

Abstract We introduce a Mass-Conserving Long Short-Term Memory with Dual States (MC-LSTM-DS) with the intention of separating short- and long-term memory for predicting streamflow. It enforces water balance through a new partition gate on precipitation. We benchmark MC-LSTM-DS against a standard Long Short-Term Memory (LSTM) and the original Mass-Conserving Long Short-Term Memory (MC-LSTM) on CAMELS-IND for 158 basins. Skill score analysis of the three models reveals that enforcing mass conservation degrades performance in the semi-arid and tropical monsoon regions. However, the addition of another cell state in MC-LSTM-DS improves performance over MC-LSTM in these regions. The decomposition of model predictions suggests that the added long cell-state captures slow storage and releases in the model, while the original state tracks quickflow. Their relative contributions vary systematically with climate, providing a hydrological representation of the cell states. The newly introduced partition mechanism captures signatures of different runoff-generating processes. A basin-scale water-balance check suggests additional effective inflows in some monsoon-dominated regions. This highlights the catchments where data or missing process representations (e.g., groundwater) may limit strict closure. We also evaluated our model's cross-regional robustness on the CAMELS-US data set across 531 basins. The performance of MC-LSTM-DS on the CAMELS-US matches that of LSTM and MC-LSTM in Nash-Sutcliffe Efficiency and attains state-of-the-art Kling-Gupta Efficiency and FHV (High Flow Bias). This indicates its potential generalization across regions. This study establishes a binational benchmark and provides an interpretable, mass-conserving deep learning framework for operational streamflow prediction across diverse hydroclimates.

Plain Language Summary The complexity of streamflow, resulting from the interaction of fast (quick-runoff) and slow (storage-driven, delayed-release) processes, makes it difficult to accurately predict. This study proposes a novel physics-aware modeling framework that enforces water-balance constraints while attempting to separate fast and slow hydrologic responses. We evaluate the model across several river basins in India and the United States that span diverse climatic and geological settings, and compare its performance with widely used data-driven baseline models. This study presents three main insights. (a) Adding an extra cell state to the Mass-Conserving LSTM enhances performance in regions where enforcing only mass conservation is not sufficient. (b) The two response pathways make the predictions more interpretable: the fast pathway dominates under wet conditions, while the slow pathway contributes during dry spells, and (c) a simple water-balance check highlights places where additional auxiliary inputs (e.g., groundwater) likely matter, suggesting directions for future improvement. This study concludes that integrating physical reasoning in machine learning models can yield more accurate predictions of streamflow.

1. Introduction

The frequency and intensity of flood and drought events have been reported to increase globally (Chiang et al., 2021; Kreibich et al., 2022; Rohde, 2023). These recurring hydrological extremes raise serious concerns about future trends and highlight the importance of streamflow prediction. The streamflow measurement plays a crucial role in understanding hydrological processes (Mai et al., 2022), the effective management of water resources (Meko & Woodhouse, 2010), and mitigating floods and droughts (Devò et al., 2025; Fleig et al., 2006). Accurate and continuous long-term streamflow time series (Chuphal & Mishra, 2023; Singh Chuphal et al., 2025) are essential for reservoir operations, agricultural practices, flood forecasting, development of early warning

© 2026 The Author(s). *Journal of Geophysical Research: Machine Learning and Computation* published by Wiley Periodicals LLC on behalf of American Geophysical Union.

This is an open access article under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Software: Saurabh Toraskar
Supervision: Abhilash Singh,
Kumar Gaurav
Validation: Saurabh Toraskar,
Abhilash Singh, Kumar Gaurav
Visualization: Saurabh Toraskar
Writing – original draft:
Saurabh Toraskar, M. Niranjannaik
Writing – review & editing:
M. Niranjannaik, Abhilash Singh,
Kumar Gaurav

systems, and design of hydraulic structures (Alfieri et al., 2013; Apel et al., 2004; S. C. Gupta et al., 2015; Hager & Boes, 2014; Zhao et al., 2011).

Streamflow estimation using hydrological models based on physical equations requires extensive calibration and is often influenced by inherent biases and uncertainties in the model parameters (Li et al., 2017; Velazquez et al., 2011). Additionally, the performance of these models depends on the long-term, high-quality observed streamflow data (Balthazar et al., 2024; Gourley & Vieux, 2006). However, streamflow estimation using hydrological models still remains challenging due to the non-linear and complex relationship between hydrological and meteorological variables and catchment attributes (Nanditha & Mishra, 2021).

Recently, data-driven Deep Learning (DL) models such as LSTM have emerged as effective alternatives to the process-based hydrological models (He et al., 2024; Hochreiter & Schmidhuber, 1997; Lin et al., 2021; Xu et al., 2022; Z. Wang et al., 2024). Kratzert et al. (2018) reported that the LSTM-based models outperform the well-calibrated SAC-SMA (Sacramento Soil Moisture Accounting) model. Building upon this foundational work, Kratzert, Klotz, Shalev, et al. (2019) and Kratzert, Klotz, Hernegger, et al. (2019) applied LSTMs to the CAMELS-US (Catchment Attributes and Meteorology for Large-sample Studies) data set (Addor et al., 2017), establishing a benchmark for numerous subsequent studies for rainfall-runoff modeling using DL. Following this, LSTMs became a standard for hydrological modeling (Alizadeh et al., 2021; Feng et al., 2020; Ni et al., 2020; Sabzipour et al., 2023; Sahoo et al., 2019).

Despite the remarkable success of LSTMs in hydrology, they lack the physical reasoning and interpretability (Kim et al., 2021; Nearing et al., 2021). To address this, Hoedt et al. (2021) proposed the Mass-Conserving LSTM (MC-LSTM) model. This incorporates physical conservation principles into the model architecture. The quantity conserved in rainfall-runoff modeling is the amount of precipitation. Despite integrating physical realism into the model, MC-LSTM was only able to outperform LSTM in the top 2% of flows indicated by FHV in the hydrology experiments conducted by Hoedt et al. (2021). Subsequent studies concluded that incorporating physical constraints through the MC-LSTM architecture does not yield substantial improvements in overall performance compared to the standard LSTM (Frame et al., 2023; Y. Wang et al., 2025a) and, in fact, tends to degrade performance during extreme flow conditions (Frame et al., 2022). Recently, Mass Conservation Relaxed LSTM (MCR-LSTM) was proposed, which incorporates a bidirectional gating mechanism to relax the strict mass-conservation constraint and to account for additional mass inputs beyond precipitation (Y. Wang et al., 2025b). However, this model is also not able to outperform LSTM and only demonstrates improved performance over the MC-LSTM in specific regions.

This study proposes a novel mass-conserving recurrent architecture with dual states that combines physical realism with the representational benefits of long and short memories. Building on conservation-aware gating (Hoedt et al., 2021), we introduce a partition mechanism that allocates the conserved input to two complementary reservoirs: a fast “quickflow” and a slow “storage” state. This design preserves strict water-balance accounting while restoring the dual-memory capability that supports the success of classic LSTMs. This addresses the limitations reported for earlier conservation-aware variants (Frame et al., 2022, 2023; Y. Wang et al., 2025a). The resulting model is intended to be both physically constrained and diagnostically interpretable, with states that can be correlated with distinct hydrologic processes.

We evaluate the standard LSTM, MC-LSTM, and the proposed MC-LSTM-DS on two large-sample data sets: India (CAMELS-IND) and the United States (CAMELS-US) (Addor et al., 2017; Mangunkiya et al., 2025). Models are trained as regional networks over all basins and compared under a common setup. We examine spatial skill patterns with respect to Köppen–Geiger climate classes. We interrogate internal mass pathways to test whether the fast and slow states show any signatures of quickflow versus delayed storage release and inspect the partition mechanism. We perform a long-term water-balance check against available evapotranspiration. This study investigates the following research questions:

- Can enforcing mass conservation while restoring dual memory improve streamflow prediction relative to purely data-driven and prior conservation-aware architectures across diverse hydroclimates?
- Do the learned fast and slow states correlate with known hydrological behaviors, how do their relative contributions vary across climate regimes and flow conditions (e.g., peak and low flows), and to what extent does the partitioning mechanism correspond to established runoff-generating processes?

Table 1
Details of the CAMELS-IND Data Set

Attributes category	Variables/Groups	Type	Resolution
Hydrology	Streamflow (observed at gauges)	Catchment-mean	Daily (1980–2020)
Meteorological forcings	Precipitation, Air temperature (Avg, min and max), Solar radiation (long- and short-wave), Relative humidity, Wind speed (Avg, U and V components), Evaporation rates (canopy and surface), Actual evapotranspiration, Potential evapotranspiration, Soil moisture (lvl: 1, 2, 3, 4)	Catchment-mean	Daily (1980–2020)
Catchment attributes	Identifiers (7), Topography (16), Climate indices (42), Hydrological signatures (73), Land use/land cover (13), Soil characteristics (28), Geologic characteristics (7), Anthropogenic factors (25)	Static attributes	Single value for the entire catchment

Note. The data set is divided into three groups according to variable type and temporal resolution. Catchment attributes are organized by descriptor, with the number in parentheses representing the count of variables within each group.

- Where and why does water-balance closure break down in practice, and what does this imply about missing processes or inputs (e.g., groundwater, human influences) needed for physically consistent, generalizable prediction?

2. Data Sets

We used the CAMELS-IND (Mangukiya et al., 2025) meteorological forcings and watershed attributes data set for 472 watersheds of Peninsular India available at <https://zenodo.org/records/14999580>. It provides long-term, consistent data that is readily available. The CAMELS-IND consists of daily observed streamflow data from the India Water Resources Information System, and daily meteorological forcings: precipitation, air temperature (avg, min, and max), solar radiation (long- and short-wave), relative humidity, wind speed and its horizontal velocity components (U and V), evaporation rates from canopy and soil, actual and potential evapotranspiration, and soil moisture at 4 levels, which are constructed using multiple data sources (Table 1). The CAMELS-IND also consists of the static catchment attributes, including diverse descriptors covering location and topography, climate and hydrological signatures, anthropogenic influences, landuse/landcover, soil, and geological characteristics.

We have also used the CAMELS-US (Addor et al., 2017), which is a comprehensive data set of 671 catchments across CONUS (Contiguous United States) used in a large number of studies for benchmarking of novel architectures (Botterill & McMillan, 2023; Kratzert, Klotz, Shalev, et al., 2019; Hoedt et al., 2021; Y. Wang et al., 2025b). We have utilized the extended North American Land Data Assimilation System (NLDAS) forcings (Kratzert, 2019) for this study, which is consistent with previous studies (Frame et al., 2022, 2023; Y. Wang et al., 2025b).

3. Methodology

3.1. Data Processing

We have removed the static features from the time series that had missing values. The categorical column `low_prec_timing` is removed because it had a single unique value (non-informative). We also dropped `ghi_stn_id`, `cwc_site_name`, `river_basin`, and `cwc_river` as they are identifiers/nominal labels, not physical attributes. Retaining them would risk site-specific data leakage and memorization in a regional model. The

remaining categorical columns are one-hot encoded to use as the inputs. To reduce dimensionality and redundancy in the resultant static feature set, we applied principal component analysis and retained the components that explain 99% of the variance. This has reduced the representation from 114 variables to 59 principal components.

The dynamic variables `wind_u`, `wind_v`, and `t_avg` are removed to mitigate multicollinearity (their information is largely captured by retained wind and temperature variables). We noticed that `pet` (mm/day) has substantial missing values. We dropped this column and instead used `pet_gleam`, which has complete coverage. The target variable `streamflow` ($\text{m}^3 \text{s}^{-1}$) is converted to mm day^{-1} using the catchment area `cwc_area` (km^2).

3.2. Model Architectures

3.2.1. LSTM and MC-LSTM

The LSTM unit is composed of a cell and a hidden state. They contain information across time steps and a set of gating mechanisms that regulate the flow of information (Hochreiter & Schmidhuber, 1997). The gating mechanisms comprise a forget gate (discarding irrelevant cell state content based on the previous hidden state and the inputs), an input gate (admitting new information to the cell state), and an output gate (exposing the updated state as the next hidden state). The mathematical formulation of LSTM is provided in Section S1 in Supporting Information S1.

The MC-LSTM enforces mass conservation by making several architectural changes in the original LSTM (Hochreiter & Schmidhuber, 1997). The core idea of the MC-LSTM is to treat memory cells as mass (water) storage. The basic gating mechanisms of MC-LSTM are the redistribution gate (R_t), the input gate (i_t), and the output gate (o_t). The MC-LSTM requires one of the inputs to be specified as mass input x_t , representing the conserved quantity (precipitation in our case). All variables, including mass input, form the auxiliary inputs a_t . We have used a time-dependent redistribution gate, which learns to distribute the existing mass at each time step. The input gate distributes the incoming mass into memory cells, and the output gate determines the mass that is to be subtracted. The mathematical formulation of the gates is given by Equations 1–3

$$R_t = \widehat{RELU}\left(W_r \cdot a_t + U_r \cdot \frac{c_{t-1}}{\|c_{t-1}\|_1} + b_r\right) \quad (1)$$

$$i_t = \hat{\sigma}\left(W_i \cdot a_t + U_i \cdot \frac{c_{t-1}}{\|c_{t-1}\|_1} + b_i\right) \quad (2)$$

$$o_t = \sigma\left(W_o \cdot a_t + U_o \cdot \frac{c_{t-1}}{\|c_{t-1}\|_1} + b_o\right) \quad (3)$$

where W_r , U_r , b_r , W_i , U_i , b_i , W_o , U_o , b_o are learnable model parameters. $\widehat{RELU}$, $\hat{\sigma}$ are normalized ReLU and σ functions respectively. The normalization of these functions is necessary to ensure the mass is conserved. The forward pass of the MC-LSTM unit is formulated as Equations 4–6

$$m_t^{\text{tot}} = R_t \cdot c_{t-1} + i_t \cdot x_t \quad (4)$$

$$c_t = (1 - o_t) m_t^{\text{tot}} \quad (5)$$

$$h_t = o_t \odot m_t^{\text{tot}} \quad (6)$$

where c_t is the memory cell state and c_{t-1} is the memory cell state from the previous time step. m_t^{tot} is the total mass of the current time step resulting from the summation of redistributed mass ($R_t \cdot c_{t-1}$) and distributed input mass ($i_t \cdot x_t$). h_t is the outgoing vector, which is subtracted from m_t^{tot} to balance the mass. Figure 1a illustrates the architecture of MC-LSTM.

3.2.2. MC-LSTM With Dual States (MC-LSTM-DS)

Motivation: The main motivation for proposing this architecture is to better mimic the characteristics of real-world catchment systems. The rationale is similar to the reason behind the success of the LSTM architecture,

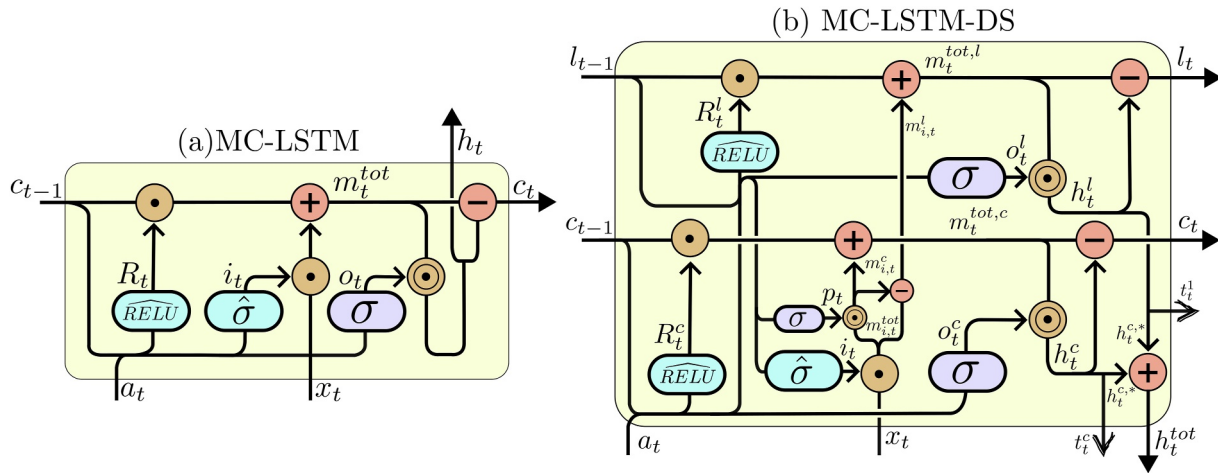


Figure 1. (a) Architecture of Mass-Conserving Long Short-Term Memory (MC-LSTM) (adapted from Hoedt et al. (2021)). (b) Architecture of the proposed MC-LSTM with Dual States (MC-LSTM-DS). σ represents the sigmoid activation layer. $\hat{\sigma}$ and $\widehat{RELU}$ depict the normalized sigmoid and normalized RELU activations, $\odot$ and $\cdot$ represent the element-wise multiplication, and dot product respectively. The addition and subtraction are element-wise accordingly.

which employs two separate memory mechanisms that allow it to capture both short-term dynamics and long-term dependencies in hydrological processes. In addition, the proposed design aims to retain the mass conservation property of the MC-LSTM framework, as it is an important direction for improving the robustness and interpretability of data-driven hydrological models.

Hydrological intuition: The intuition behind the proposed architecture can be described as follows. Both cell states are represented as vectors with dimensionality equal to the hidden size. These two vectors correspond to different hydrological processes, while each element can be interpreted as a conceptual “bucket” representing a sub-process. The redistribution gate can be interpreted as an operator that accounts for internal changes occurring within the catchment between successive time steps. Under the bucket analogy, the redistribution gate governs the transfer of water among sub-catchment processes while conserving the total mass. After this internal redistribution, the catchment is ready for the input of the current time step. First, the input gate distributes the scalar rainfall input across the hidden dimension, producing a vector representation whose elements sum to the original rainfall amount. The partition gate determines the fraction of water in each element allocated to the short-term state vector, with the remaining portion assigned to the long-term state, thereby conserving mass. Finally, the output gates of the cell states determine the output water from their respective processes.

The gates of the short-term cell state receive all the input features, as well as the normalized values of the previous short-term cell state as inputs. The normalized values reflect the hydrological understanding that catchment dynamics are strongly influenced by the current state of the system. Similarly, the gates of the long-term cell state take as inputs all the input features, along with the normalized values of both the previous long-term and short-term cell states. Incorporating information from the short-term state is motivated by the fact that short-term processes are typically associated with near-surface dynamics. Since incoming rainfall interacts first with these near-surface processes, they exert a strong influence on the evolution of the longer-term storage components.

Mathematical formulation: This new cell state is introduced with the expectation of capturing long-term dependencies and is denoted as l . The architecture of this new MC-LSTM-DS includes two redistribution gates (R_t^c and R_t^l), two output gates (o_t^c and o_t^l), one input gate (i_t), and one new partition gate (p_t). The redistribution gate reallocates the mass across elements of the same vector. The gate R_t^c takes a_t and the normalized previous cell state c_{t-1} as inputs. In contrast, the gate R_t^l takes a_t , the normalized c_{t-1} , and the normalized long-term state l_{t-1} as inputs. The input gate distributes the incoming mass into the memory cells using the same inputs as R_t^l , after which the partition gate calculates the amount of mass to be added ($m_{i,t}$) to the redistributed cell state c mass ($R_t^c \cdot c_{t-1}$), resulting in $m_t^{tot,c}$. The remaining mass $m_{i,t}$ is added to the redistributed cell state l ($R_t^l \cdot c_{t-1}$) resulting in $m_t^{tot,l}$. The inputs for output gates are the same as their redistribution counterparts. The output gates o_t^c and o_t^l

regulate the fraction of mass removed from their respective cell states at time t , producing the outputs h_t^c and h_t^l . For both of these output vectors, the first element corresponds to the “trash cell,” t_t^c and t_t^l , which are discarded and not utilized for prediction. The remaining elements of the vectors are summed to obtain the functional components of the outputs of each cell state, which are denoted as $h_t^{c,*}$ and $h_t^{l,*}$. The final output at time t , h_t^{tot} , is then computed as the sum of these two functional components.

$$C_{t,*} = U_*^c \cdot \frac{c_{t-1}}{\|c_{t-1}\|_1} + U_*^l \cdot \frac{l_{t-1}}{\|l_{t-1}\|_1} \quad (* \in \{r, i, p, o\}) \quad (7)$$

The gating mechanisms are formulated using Equations 8–13. Substitute Equation 7 in Equations 9–11 and 13

$$R_t^c = \widehat{RELU}\left(W_r \cdot a_t + U_r^c \cdot \frac{c_{t-1}}{\|c_{t-1}\|_1} + b_r\right) \quad (8)$$

$$R_t^l = \widehat{RELU}(W_r \cdot a_t + C_{t,r} + b_r) \quad (9)$$

$$i_t = \sigma(W_i \cdot a_t + C_{t,i} + b_i) \quad (10)$$

$$p_t = \sigma(W_p \cdot a_t + C_{t,p} + b_p) \quad (11)$$

$$o_t^c = \sigma\left(W_o \cdot a_t + U_o^c \cdot \frac{c_{t-1}}{\|c_{t-1}\|_1} + b_o\right) \quad (12)$$

$$o_t^l = \sigma(W_o \cdot a_t + C_{t,o} + b_o) \quad (13)$$

The forward pass is defined using Equations 14–21:

$$m_{i,t}^c = i_t \cdot x_t \odot p_t \quad (14)$$

$$m_{i,t}^l = i_t \cdot x_t \odot (1 - p_t) \quad (15)$$

For Equations 16–20 $j \in \{c, l\}$:

$$m_t^{\text{tot},j} = R_t^j \cdot c_{t-1}^j + m_{i,t}^j \quad (16)$$

$$c_t^j = (1 - o_t^j) \odot m_t^{\text{tot},j} \quad (17)$$

$$h_t^j = o_t^j \odot m_t^{\text{tot},j} \quad (18)$$

$$t_t^j = h_t^{j,(1)} \quad (19)$$

$$h_t^{j,*} = \sum_{i=2}^n h_t^{j,(i)} \quad (20)$$

$$h_t^{\text{tot}} = h_t^{c,*} + h_t^{l,*} \quad (21)$$

3.3. Evaluation Metrics

Model performance is assessed using a multi-criteria framework comprising eight metrics to provide a comprehensive evaluation. We calculate the Nash-Sutcliffe Efficiency (NSE), which is the primary metric used in hydrological predictions (Nash & Sutcliffe, 1970). Further, we compute the Kling-Gupta Efficiency (KGE) and the α -NSE and β -NSE decompositions as formulated in H. V. Gupta et al. (2009). We also report the Pearson correlation coefficient (r). To study models' performance in high and low flow regimes, we computed FHV (High flow bias calculated for the top 2%) and FLV (Low flow bias calculated for the bottom 30%) (Yilmaz et al., 2008). We also include the Peak-Timing metric (Kratzert et al., 2020), and Peak-MAPE (Frame et al., 2022) to gauge the

performance of models in predicting peaks. Detailed formulation of these metrics is provided in Section S2 in Supporting Information S1.

3.4. Experimental Setup

To evaluate the model's performance, we have selected only those gauge stations that had more than 70% data record available. This filter resulted in a total of 158 stations. We have used the data from 1 January 1993 to 31 December 2010 for model training, 1 January 2011 to 31 December 2018 for validation, and 1 January 1980 to 31 December 1992 for testing purposes. The placement of the test period prior to the training period follows such large-sample hydrological studies (Kratzert, Klotz, Shalev, et al., 2019; Hoedt et al., 2021; Y. Wang et al., 2025b). Additionally, this period contained fewer data gaps in CAMELS-IND, providing a more continuous streamflow record and enabling a more robust evaluation of model performance. We have trained the model for all the catchments together (regional model) following Kratzert et al. (2018, 2024).

For all the models, the sequence length was set to 365 to capture the characteristics of the full annual hydrological cycle (Frame et al., 2022; Hoedt et al., 2021; Mangukiya et al., 2023); the batch size was set to 256; and the learning rate was set to 0.01 and reduced to 0.001 after 20 epochs. For LSTM, we tried the hidden sizes of 64 (Val NSE: 0.57), 128 (Val NSE: 0.56), and 256 (Val NSE: 0.56) while keeping all the other parameters constant. The hidden size was set to 64 for LSTM due to its better performance. The hidden size for MC-LSTM and MC-LSTM-DS was set to 64 due to its proven better performance (Hoedt et al., 2021; Y. Wang et al., 2025b). The LSTM is trained for 50 epochs to ensure proper convergence. The MC-LSTM and MC-LSTM-DS are trained for 30 epochs. To avoid overfitting, we selected the model parameters from the epoch with the highest validation NSE. We pre-normalized all of the input variables and the target streamflow (mm/day) for LSTM. The normalization strategy employed was standard normalization, where the mean (μ) and standard deviation (σ) were computed from the training data set. These statistics were subsequently used to scale the test samples, ensuring consistency in feature scaling while preventing data leakage. In the case of the MC-LSTM and MC-LSTM-DS, all the variables except the mass input, that is, precipitation, and the target variable, streamflow, are pre-normalized. This step is essential for the models to enforce mass conservation. The loss function is set to the basin-averaged NSE (Kratzert, Klotz, Shalev, et al., 2019), optimized using the AdamW optimizer (Loshchilov & Hutter, 2017).

The models are run for 5 different seeds to minimize the stochasticity of weight initialization. Ensembles are then constructed by averaging predictions from the best k runs ($k = 1$ to 5). The best k runs are selected according to their validation NSE. The ensemble with the optimal number of runs (k), corresponding to the highest validation NSE, is selected, and its performance is reported for both the validation and test periods. Additionally, we carry out a one-sided Wilcoxon Signed-Rank Test on the NSE and KGE values across basins to establish the statistical significance of improvements in the performance of MC-LSTM-DS against MC-LSTM and LSTM. This non-parametric paired test evaluates whether the median difference in NSE between two models is greater than zero, thereby testing the hypothesis that one model systematically outperforms the other. The test is applied across all basins. Statistical significance is established when the corresponding p -value falls below the chosen significance threshold of 5% ($p < 0.05$).

To assess the relative performance of the proposed model, MC-LSTM-DS, against MC-LSTM and LSTM, we analyze the skill scores based on both NSE and KGE. In addition, the skill of MC-LSTM relative to LSTM is also evaluated. The skill score is computed following the formulation of Y. Wang et al. (2025b), as given in Equations 22 and 23. A skill score of 1 indicates perfect agreement with observations, a value of 0 denotes performance equivalent to the reference model, and negative values indicate that the model performs worse than the reference. The base map corresponds to the Köppen–Geiger climate zones (Beck et al., 2023), enabling an assessment of whether the observed performance improvements or degradations are influenced by climatic controls.

$$NSE_{SSmodel}^{ref} = \frac{NSE_{model} - NSE_{ref}}{1 - NSE_{ref}} \quad (22)$$

$$KGE_{SSmodel}^{ref} = \frac{KGE_{model} - KGE_{ref}}{1 - KGE_{ref}} \quad (23)$$

We evaluate the validity of the assumed mass balance framework, wherein precipitation is treated as the sole input, and identify regions where this assumption fails to hold in real-world catchments. We further analyze the

Table 2

Evaluation Metrics of LSTM, MC-LSTM, and MC-LSTM-DS on the Validation and Testing Data

Metric	Validation			Testing		
	LSTM	MC-LSTM	MC-LSTM-DS	LSTM	MC-LSTM	MC-LSTM-DS
NSE	0.59	0.58	0.59	0.57	0.64	0.65
KGE	0.58	0.57	0.56	0.43	0.54	0.56
Pearson-r	0.8	0.81	0.81	0.81	0.85	0.84
α -NSE	0.84	0.91	0.93	0.58	0.68	0.69
β -NSE	0.02	0.01	0.03	−0.1	−0.07	−0.06
FHV	−15.83	−10.27	−7.05	−40.49	−31.39	−29.54
FLV	−1.21	−71.39	−55.28	53.16	−13.72	21.4
Peak-Timing	0.62	0.62	0.6	0.62	0.55	0.54
Peak-MAPE	49.11	47.92	47.94	59.83	54	50.55

Note. The best value of each metric on the validation set is highlighted in bold, and those corresponding to the best performance on the testing set are shown in blue.

internal states of the MC-LSTM-DS model to assess whether they capture hydrological signatures characteristic of distinct climatic regimes. In addition, the behavior of the partitioning gate is examined to better understand the internal routing strategies learned by the model and to assess whether these patterns are consistent with known runoff-generation processes. Finally, to assess the generalization capability of the proposed model, its performance is compared against established physics-based, mass-conserving, and purely data-driven approaches on the CAMELS-US data set.

4. Results and Discussion

4.1. Model Performance

Table 2 reports the performance of LSTM, MC-LSTM, and MC-LSTM-DS on the validation and testing sets. Results on the validation set are reported for completeness and are not analyzed in depth. For the primary evaluation metric, NSE, the LSTM exhibits a slight decrease in performance on the testing set as compared to the validation set, which is expected. However, both mass-conserving LSTM variants show a notable improvement in testing performance as compared to the validation set, indicating stronger generalization capabilities. On the testing set, both the MC-LSTM and the MC-LSTM-DS demonstrate a significant performance improvement relative to standard LSTM. This finding aligns with recent evidence showing improved performance on CAMELS-IND when mass constraints are included (Sahu et al., 2025). Further shifting the focus to our proposed architecture, MC-LSTM-DS, we can observe that it outperforms MC-LSTM in all the metrics except Pearson-r and FLV. On the testing set, MC-LSTM-DS achieves superior KGE with no decline from the validation set, whereas LSTM exhibits a substantial drop and MC-LSTM shows a minor decrease, highlighting the robustness of MC-LSTM-DS. The NSE is further decomposed into α and β components to provide a more diagnostic assessment of the model. α -NSE shows the model's ability to reproduce the observed variability of flow, with 1 being the best, and β -NSE represents the bias in the simulated mean flow, where the value closer to 0 is better. The proposed model performs better in both α and β -NSE. MC-LSTM-DS shows an improvement in high flow bias (FHV), which improves when mass conservation is enforced (Hoedt et al., 2021). There is a decrease in FLV when compared to MC-LSTM, but still a significant increase when compared to LSTM. To assess model performance on peaks, we use the Peak-Timing metric and Peak-MAPE. The Peak-Timing measures the average time lag (in days) between observed and simulated peaks, whereas the Peak-MAPE quantifies the average percentage error in peak magnitudes. In both the Peak-Timing and Peak-MAPE, a value closer to 0 is better. A key finding is that enforcing mass constraints improves the prediction of both peak timing and magnitude. Moreover, MC-LSTM-DS outperforms other models on both metrics.

Table 3 presents the results of the Wilcoxon signed-rank test. At the 5% significance level, MC-LSTM-DS significantly outperforms both MC-LSTM and LSTM across basins for NSE and KGE. While the performance gains over MC-LSTM are relatively modest, they are consistently observed across basins and are statistically significant. In contrast, the improvements over LSTM are substantially larger for both metrics. The very

Table 3

Results of the One-Sided Wilcoxon Signed-Rank Test Comparing MC-LSTM-DS Against Other Baseline Models' Performance Across CAMELS-IND Basins

Metric	Baseline model	p (MC-LSTM-DS > model)	Median difference
NSE	MC-LSTM	2.5×10^{-7}	0.01
	LSTM	8.9×10^{-18}	0.06
KGE	MC-LSTM	4.4×10^{-10}	0.02
	LSTM	1.3×10^{-16}	0.11

low p -values reported in Table 3 provide strong evidence that the observed performance gains are unlikely to have arisen from random variation. The larger gains in KGE suggest that the proposed architecture improves not only predictive accuracy but also the overall agreement between observed and simulated streamflow dynamics.

To not rely solely on the median value of NSE and to visualize the distribution of model performance across basins, we have plotted the smoothed Cumulative Distribution Function (CDF) of all three models (Figure 2). In the CDF plot of NSE, a right-shifted curve indicates superior performance, as it reflects that a larger proportion of basins achieve higher NSE values, signifying improved predictive skill and consistency. We can clearly observe the performance improvement resulting from the enforcement of mass conservation (Figure 2). Furthermore, the MC-LSTM-DS curve lies consistently to the right of the MC-LSTM across the entire distribution, indicating its superior overall performance among all models.

4.2. Spatial Distribution of Skill Scores

Figures 3a–3f presents the pairwise skill scores for NSE and KGE comparing LSTM, MC-LSTM, and MC-LSTM-DS across individual catchments. Before interpreting the spatial patterns, it is important to note that NSE is more sensitive to high-flow conditions, whereas KGE provides a more balanced assessment of model performance across the entire hydrograph (H. V. Gupta et al., 2009). Figures 3a and 3b show the skill scores of MC-LSTM relative to LSTM. We observe that MC-LSTM does not improve upon LSTM in semi-arid (BSh) regions, where a degradation in performance is evident, especially in terms of KGE. Similar behavior is also observed in the tropical monsoon regions of southern India. Notable improvements occur in tropical savanna climates. Figures 3c and 3d present the NSE and KGE skill scores of MC-LSTM-DS relative to LSTM, and the spatial pattern remains largely consistent with that observed for MC-LSTM, suggesting that the degradation in

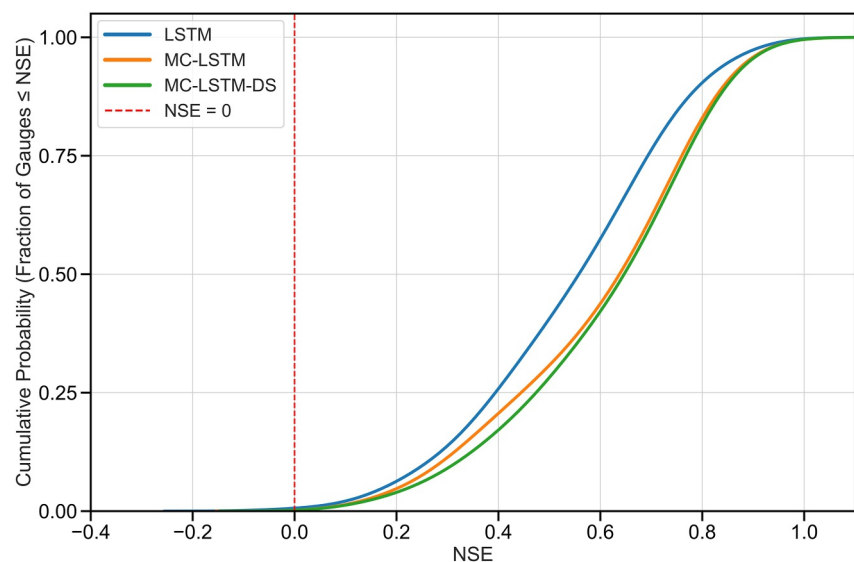


Figure 2. Cumulative distribution function of Nash-Sutcliffe efficiency (NSE) across the basins for long short-term memory (blue), mass-conserving long short-term memory (orange), and MC-LSTM-DS (green). Vertical line in red (dashed) shows zero NSE.

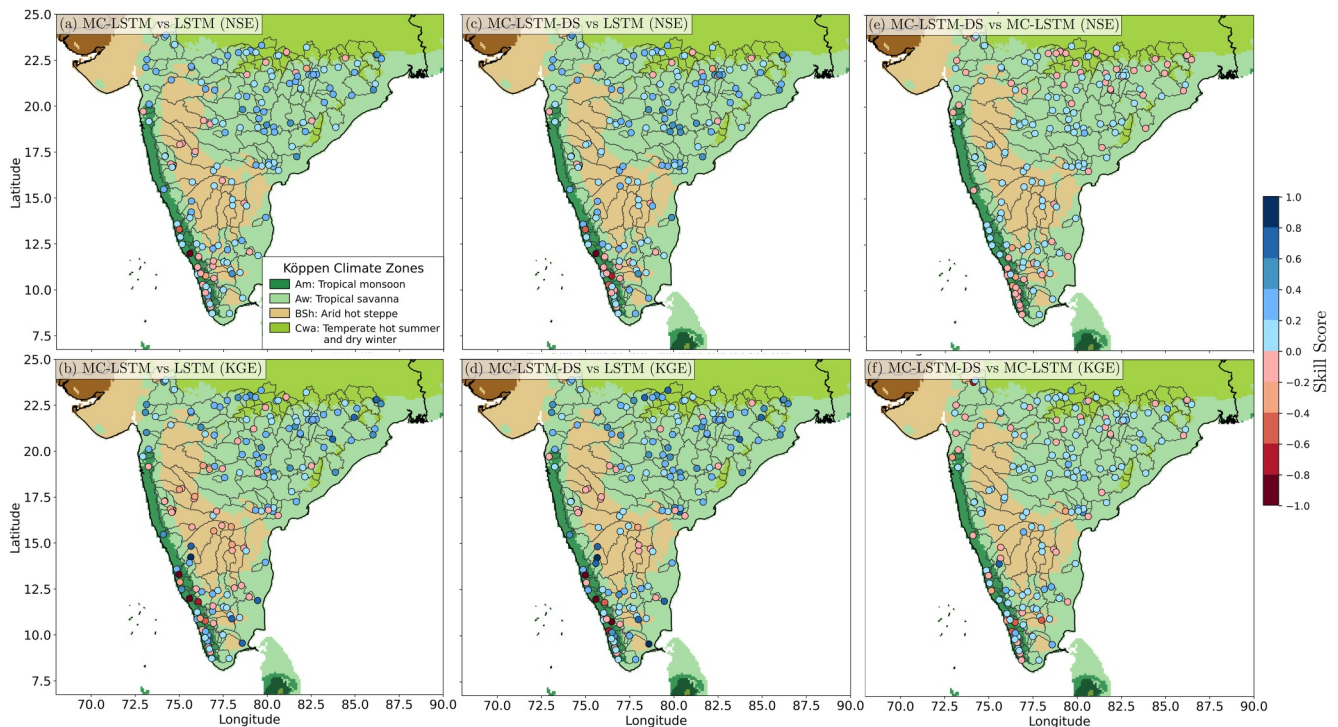


Figure 3. Circles in different colors show the spatial distribution of pairwise skill scores (Nash-Sutcliffe Efficiency and Kling-Gupta Efficiency shown in top and bottom panels, respectively) for mass-conserving long short-term memory (MC-LSTM) versus long short-term memory (LSTM), MC-LSTM-DS versus LSTM, and MC-LSTM-DS versus MC-LSTM. Different colors in the base map of Peninsular India show different climate zones according to Köppen–Geiger climate classification (Source: Beck et al. (2023)).

performance in certain regions likely arises from the constraints imposed by strict mass conservation within the model structure. Catchments in the Western Ghats region are generally small and characterized by steep topographic gradients and are in the tropical monsoon climate, whereas those in central India are relatively dry and underlain by highly fractured Deccan Trap formations that can facilitate subsurface flow across catchment boundaries and lie in the semi-arid region. These characteristics align with the factors identified by Fan (2019) as increasing the likelihood of catchment leakage, including small catchment size, steep topographic and climatic gradients, permeable subsurface materials, and dry climatic conditions. Finally, Figures 3e and 3f compare the proposed MC-LSTM-DS model against MC-LSTM, highlighting regions where the additional cell state improves performance while still preserving mass conservation. Improvements are mainly observed in tropical savanna and semi-arid regions; slower processes have a vital contribution to streamflow dynamics, supporting our hypothesis that the additional cell state enhances the representation of slower hydrological processes. In humid subtropical (Cwa) climates, the model shows a degradation in NSE but not in KGE, suggesting that while the overall hydrograph characteristics remain reasonably well captured, the model may struggle to reproduce peak flows. In tropical savanna regions, the model performance remains mixed, which may reflect the influence of catchment heterogeneity and the fact that hydrological responses in these basins are not solely governed by climatic controls but are also shaped by variations in physiographic and hydrological characteristics.

4.3. Investigating the Cell States of MC-LSTM-DS

To investigate the behavior of both cell states (c and l), we analyze their respective trash cells (defined as the first element of the mass m_{out} vector, supposed to account for non-streamflow outputs) and the functional component (which is the sum of all the elements of m_{out} except the trash cells). We obtain the final prediction of the model by adding the functional components of both c and l . We hypothesize that the c cell state represents fast or short-term hydrological processes within a catchment. Accordingly, its trash cell may correspond to immediate water losses such as evapotranspiration, evaporation from open water bodies, and anthropogenic consumption (e.g., irrigation). The functional component likely represents the surface or near-surface rapid runoff. In contrast, the l cell state is

proposed to capture slow or long-term processes, with its trash cell potentially representing groundwater seepage into adjacent catchments and the functional component reflecting the contribution of groundwater to streamflow.

We selected four representative catchments, one from each of the four major Köppen–Geiger climate zones (Am, Aw, BSh, and Cwa) (Figures 4 and 5), based on three criteria. First, each catchment lies entirely within a single climate zone, ensuring that its hydrological response is representative of that climatic regime. Second, the selected catchments exhibit relatively high predictive performance, as indicated by their NSE values. Third, the performance gains of MC-LSTM-DS over both MC-LSTM and LSTM are substantial for these catchments, particularly in terms of NSE and KGE. For the Am climate zone, a catchment with performance comparable to MC-LSTM and substantially better than LSTM was selected, as none satisfied all three criteria. The corresponding performance metrics are provided in Section S4 in Supporting Information S1. Rather than averaging across multiple catchments within a climate zone, we analyzed individual time series to preserve the relative temporal dynamics of the cell states, as averaging could obscure meaningful patterns due to differences in their value ranges. The analysis focuses exclusively on the internal states of the model to examine whether the learned dynamics vary systematically across climate zones. The specified quantities were plotted for a single model run (seed = 42), enabling a detailed interpretation of the model's behavior. Furthermore, only a subset of the testing period, spanning from 1 January 1985 to 31 December 1990, was considered to improve visual clarity. In each subplot, the top panel shows the contribution of both cell states to the model prediction (functional components) along with the corresponding trash cell values, the middle panel displays only the functional components, and the bottom panel presents the trash cell values.

Tropical Monsoon (Am): Figure 4a shows the states of gauge 14005, which exhibits an NSE of 0.87. The trash cell values are not clearly visible in the top panel; however, the bottom panel reveals that they are approximately an order of magnitude smaller. This suggests that the contributions from outputs other than streamflow are minimal, which is consistent with the high runoff ratio in this region (Figure 5c of Mangukiy et al. (2025)). Runoff ratio is the ratio of long-term mean daily flow to long-term mean daily precipitation. Furthermore, the functional components exhibit large magnitudes and strong fluctuations due to the small size of the catchment and the tropical characteristics of the catchment. Finally, in the bottom plot, the trash cell values of the c cell state remain relatively constant during the non-monsoon season. This behavior is likely associated with groundwater outflows (Fan, 2019).

Tropical Savanna (Aw): Gauge 3008 (Figure 4b), located in a tropical savanna region, exhibits an NSE of 0.77. A notable characteristic is the comparatively higher magnitude of the trash cell values, which is consistent with the relatively low runoff ratio (0.214) (Mangukiy et al., 2025). Furthermore, the region is characterized by a high density of medium to large dams (211), of which approximately 94.6% are primarily used for irrigation (Mangukiy et al., 2025). These dams likely contribute to the observed peaks post-monsoon in the trash cell values of the c (fast) cell state, reflecting regulated releases for irrigation. An additional peak is observed during the summer months, which reflects high evapotranspiration rates typical of tropical savanna climates, where pronounced wet and dry seasons drive strong seasonal contrasts in hydrological processes.

Arid Hot Steppe (BSh): Gauge 4057 (Figure 5a) represents the characteristic of the Arid climate. It exhibits an NSE of 0.51, which is relatively high for this climate. The much higher relative magnitude of trash cells to the functional components reflects the very low runoff ratio of 0.06. The most significant observation in this plot is seen in the middle panel, in which we can see that the contribution of l (long term) to the streamflow is higher. This indicates that streamflow is primarily storage-controlled, where infiltrated water is gradually released from soil and groundwater. Generation of rapid runoff is limited. This behavior is typical of semi-arid catchments with episodic rainfall and significant subsurface storage. The trash cells peak post-monsoon, possibly due to the usage of water for irrigation. All the states reach almost zero during the summer months, indicative of the drought-like conditions these regions experience during summer.

Temperate Hot Summer and Dry Winter (Cwa): Gauge 12043 (Figure 5b) exhibits an NSE of 0.79 and represents a Cwa climate. This region is characterized by strong seasonality, with moderately high rainfall and hot summers. The pronounced seasonality is reflected in the comparable magnitude of the functional components and trash cells. Furthermore, storage is largely depleted during the summer months, leading to near-zero values across most time series, except for a few years.

Overall, the analysis demonstrates that the internal dynamics of the model exhibit systematic variations across different climate regimes. The relative contributions of the functional components and trash cells, as well as the

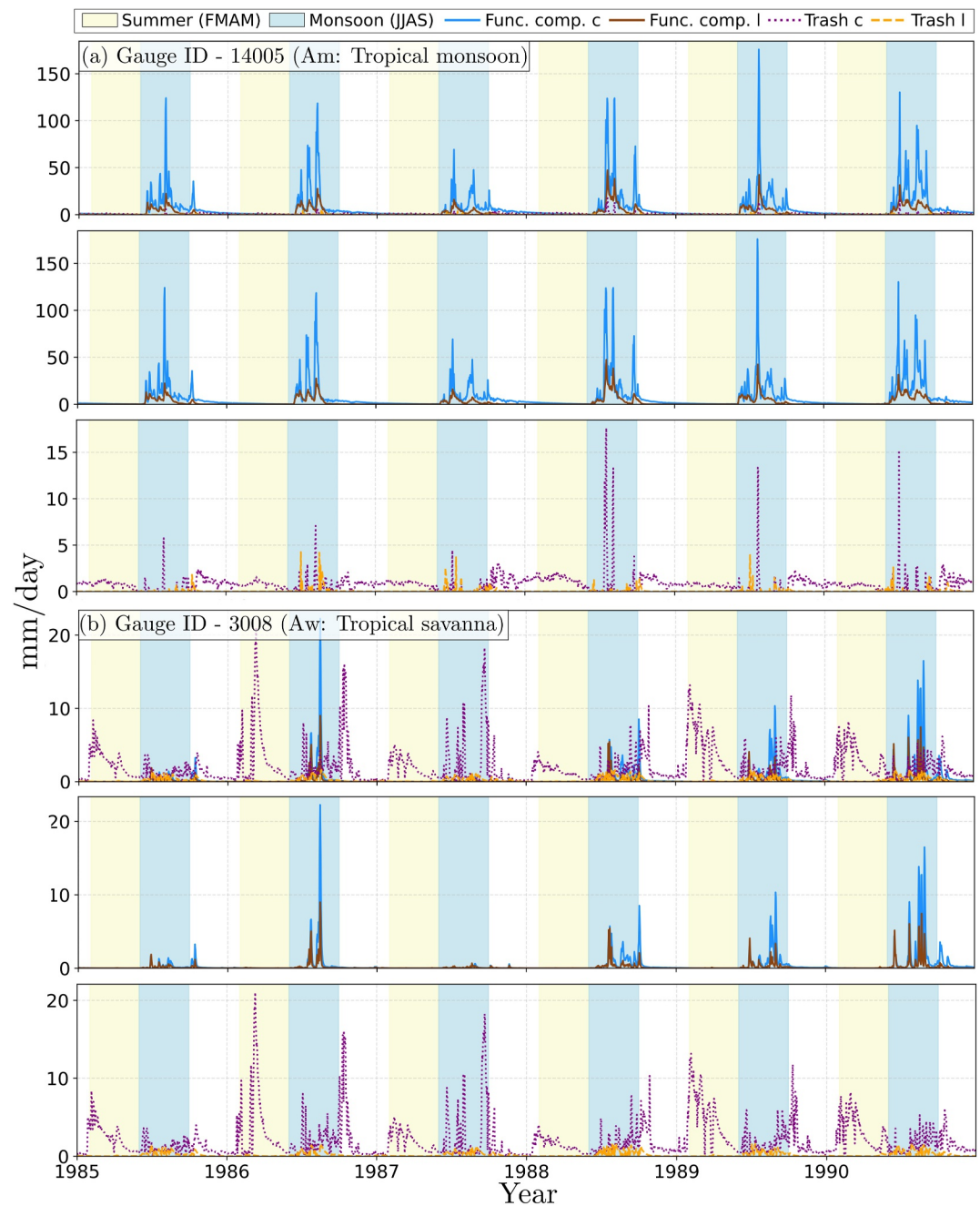


Figure 4. Time series of outgoing masses (streamflow and other outputs) from both cell states (c and l correspond to fast and slow states, respectively) for two representative catchments, Gauge IDs: 14005 and 3008. The catchment IDs 14005 and 3008 are in the Am (Tropical Monsoon) and Aw (Tropical Savanna) Köppen–Geiger climate zones, respectively. For improved visual clarity, the time series are shown for a subset of the testing period (1 January 1985–31 December 1990). The solid lines in blue and brown show the contributions of the fast and slow cell states to streamflow, respectively. The dotted lines in purple and orange are the non-streamflow outputs of the fast and slow cell states (trash cells). For each catchment, the top panel for both catchments presents all four quantities, the middle panel shows only the functional components, and the bottom panel displays the trash cell components, respectively.

dominance of fast (c) versus slow (l) processes, align well with known hydrological characteristics such as runoff ratios, storage behavior, and seasonality. These findings suggest that the model is able to learn meaningful process representations, with its internal states reflecting climate-dependent hydrological behavior.

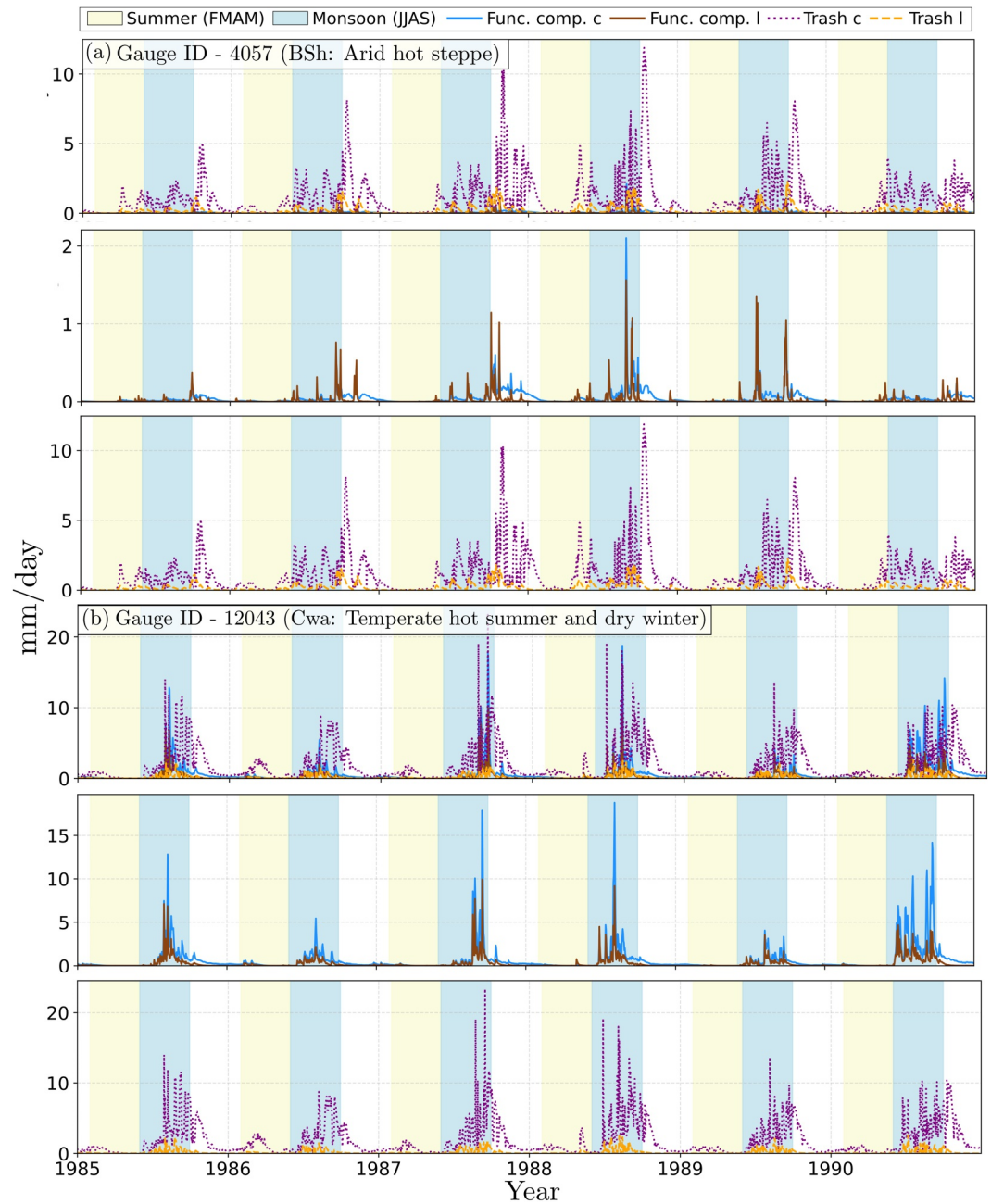


Figure 5. Time series of outgoing masses (streamflow and other outputs) from both cell states (c and l correspond to fast and slow states, respectively) for two representative catchments, Gauge IDs: 4057 and 12043. The catchment IDs 4057 and 12043 are in the BSh (Arid hot steppe) and Cwa (Temperate hot summer and dry winter) Köppen–Geiger climate zones, respectively. For improved visual clarity, the time series are shown for a subset of the testing period (1 January 1985–31 December 1990). The solid lines in blue and brown show the contributions of the fast and slow cell states to streamflow, respectively. The dotted lines in purple and orange are the non-streamflow outputs of the fast and slow cell states (trash cells). For each catchment, the top panel for both catchments presents all four quantities, the middle panel shows only the functional components, and the bottom panel displays the trash cell components, respectively.

4.4. Inspection of Learned Partition Mechanisms

The newly introduced partition gate p_i allocates incoming mass (precipitation) between the short-term and long-term storage components. In this subsection, we examine the behavior of the partitioning mechanism to assess whether it captures physically meaningful runoff generation processes.

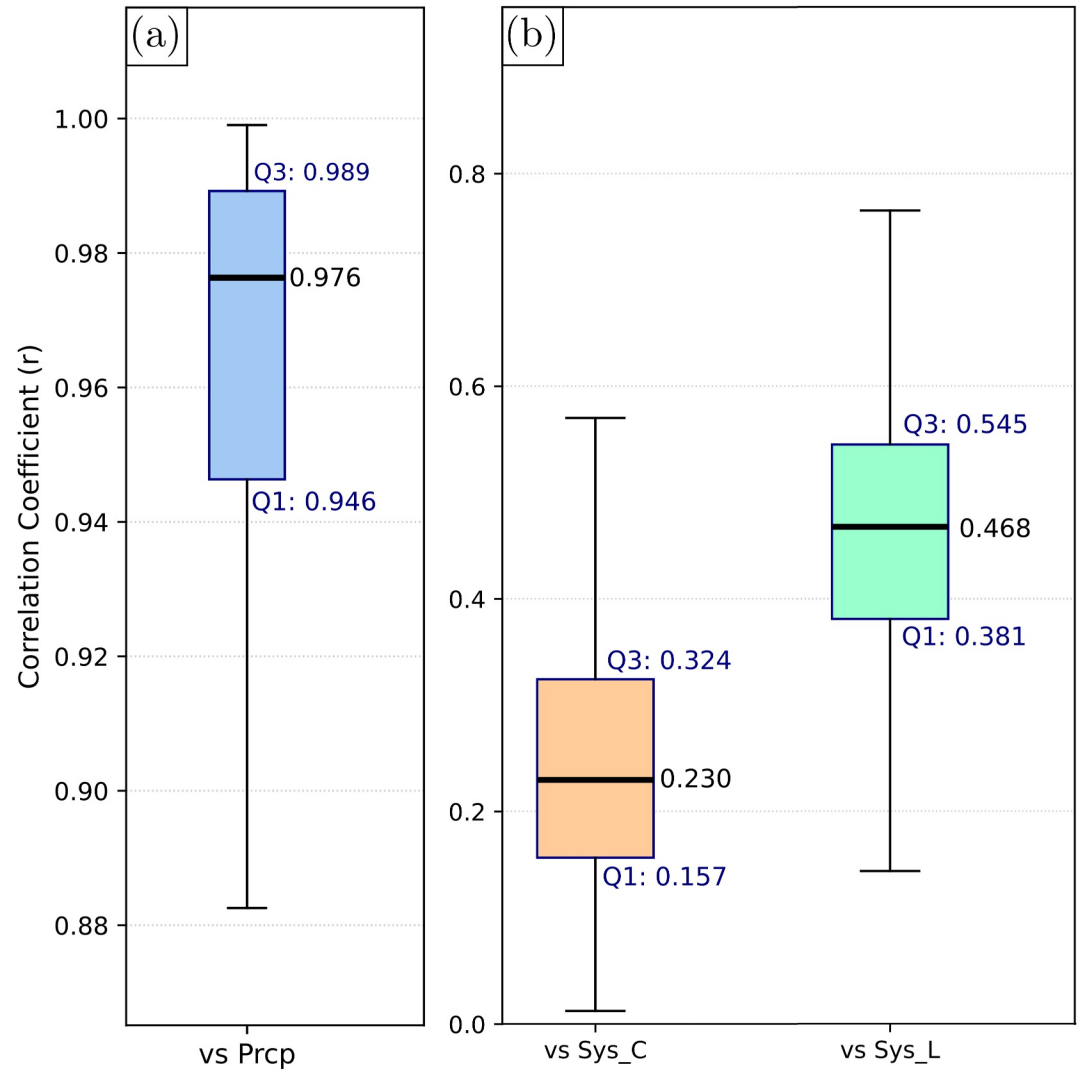


Figure 6. Box plots of correlation coefficients across all basins between the difference in mass allocated to the short-term and long-term storage components ($\sum_{j=1}^n m_{i,t}^{c,(j)} - \sum_{j=1}^n m_{i,t}^{l,(j)}$) and (a) Precipitation and (b) Total mass in the short-term and long-term cell states prior to precipitation.

First, we compute the correlation coefficient between daily precipitation and the difference between the mass allocated to the short-term and long-term storage components ($\sum_{j=1}^n m_{i,t}^{c,(j)} - \sum_{j=1}^n m_{i,t}^{l,(j)}$) across all basins. The distribution of these correlations is shown as a box plot in Figure 6a. From a process-based perspective, higher precipitation is expected to favor rapid runoff generation, leading to a relatively larger fraction of water being routed to the short-term storage compared to the long-term storage. Consistent with this expectation, Figure 6a shows strongly positive correlations, with a median value of $r = 0.976$ and the interquartile range exceeding 0.94. To verify that this relationship is not solely a mathematical consequence of the partitioning formulation, we examined the variability of the partitioning term ($2p_t - 1$), which exhibits substantial variability (mean = 0.40, standard deviation = 0.38), indicating that ($2p_t - 1$) is not approximately constant through time. This indicates that the model increases the relative contribution of short-term storage under higher precipitation. This behavior is consistent with characteristics of infiltration-excess overland flow.

To investigate the representation of saturation-excess overland flow, we compute the correlation between the difference ($\sum_{j=1}^n m_{i,t}^{c,(j)} - \sum_{j=1}^n m_{i,t}^{l,(j)}$) and the total mass present in the short-term and long-term cell states prior to precipitation. These storage components are assumed to represent the pre-event catchment wetness conditions.

As shown in Figure 6b, the partitioning difference is positively correlated with both storage components, with a stronger correlation for long-term storage ($r = 0.468$) than for short-term storage ($r = 0.23$). This suggests that the partitioning mechanism is primarily influenced by the slower-varying storage component, while transient storage fluctuations play a secondary role. Under wetter conditions, characterized by elevated long-term storage, the model allocates a relatively larger fraction of incoming mass to the short-term component. This behavior is consistent with saturation-excess overland flow, where high antecedent storage limits available storage capacity and promotes rapid runoff generation.

4.5. Analysis of the Mass Conservation of MC-LSTM-DS

The mass conservation of MC-LSTM-DS conserves mass in a manner where the precipitation (ppt) is the only inflow, and the outputs of both the cell states consist of the trash cells and functional components (Streamflow). These relations are formulated in Equations 24–26.

$$\text{PPT} = \text{TC}^{\text{tot}} + \text{Streamflow}^{\text{tot}} \quad (24)$$

$$\text{TC}^{\text{tot}} = \text{TC}^c + \text{TC}^l \quad (25)$$

$$\text{Streamflow}^{\text{tot}} = h_t^{c,*} + h_t^{l,*} \quad (26)$$

In a real-world catchment, over long time periods, the total water entering the catchment is equal to the amount leaving. Under the assumption that precipitation is the only input, TC^{tot} should account for the total amount of water aside from streamflow leaving the catchment, which can be either in the form of evapotranspiration (ET) or groundwater outflows (GW_{out}) as given in Equation 27.

$$\text{TC}^{\text{tot}} = \text{ET} + \text{GW}_{\text{out}} \quad (27)$$

If our assumption is true, then $\text{TC}^{\text{tot}} - \text{ET} \geq 0$, in the case that it is not ($\text{TC}^{\text{tot}} - \text{ET} < 0$), it violates mass conservation and implies the presence of mass inputs other than precipitation.

We have done a water balance analysis similar to Y. Wang et al. (2025a), where the relative percentage difference (RD_{TC}) between the TC^{tot} and ET is calculated. Given that actual evapotranspiration (AET; `aet_gleam`) data are available in our data archive, we adopt this variable to represent ET. We also calculate the bias percentage to aid our analysis. The formulations of RD_{TC} and Bias are given in Equations 28 and 29:

$$RD_{TC} = \frac{\overline{\text{TC}^{\text{tot}}} - \overline{\text{AET}}}{\overline{\text{AET}}} \times 100 \quad (28)$$

$$\text{Bias} = \frac{\overline{\text{sim}} - \overline{\text{obs}}}{\overline{\text{obs}}} \times 100 \quad (29)$$

where $\overline{\text{TC}^{\text{tot}}}$, $\overline{\text{AET}}$, $\overline{\text{sim}}$, and $\overline{\text{obs}}$ are the long-term means of total trash cell values, actual evaporation, and simulated and observed streamflow.

Figure 7 shows the RD_{TC} and bias for five representative catchments corresponding to the four major climatic zones. Gauges located to the left of the red line (Bias = 0) indicate systematic underestimation by the model, whereas those to the right indicate overestimation. A negative RD_{TC} suggests that the observed evapotranspiration exceeds the trash cell values (sum of all outputs other than streamflow, including evapotranspiration). In contrast, a positive RD_{TC} indicates that the model predicts a higher sum of non-streamflow outputs than actual evapotranspiration.

Figure 7 shows that gauges in the tropical monsoon region exhibit strongly negative values of RD_{TC} along with a clear negative bias. Negative RD_{TC} suggests that the actual evapotranspiration exceeds the model's trash cell. One possible explanation for this mismatch is the presence of additional water sources, such as groundwater inflows, that help sustain high evapotranspiration. Since precipitation is the only input provided to the model, it cannot account for these extra contributions. As a result, the model remains constrained in representing the full water

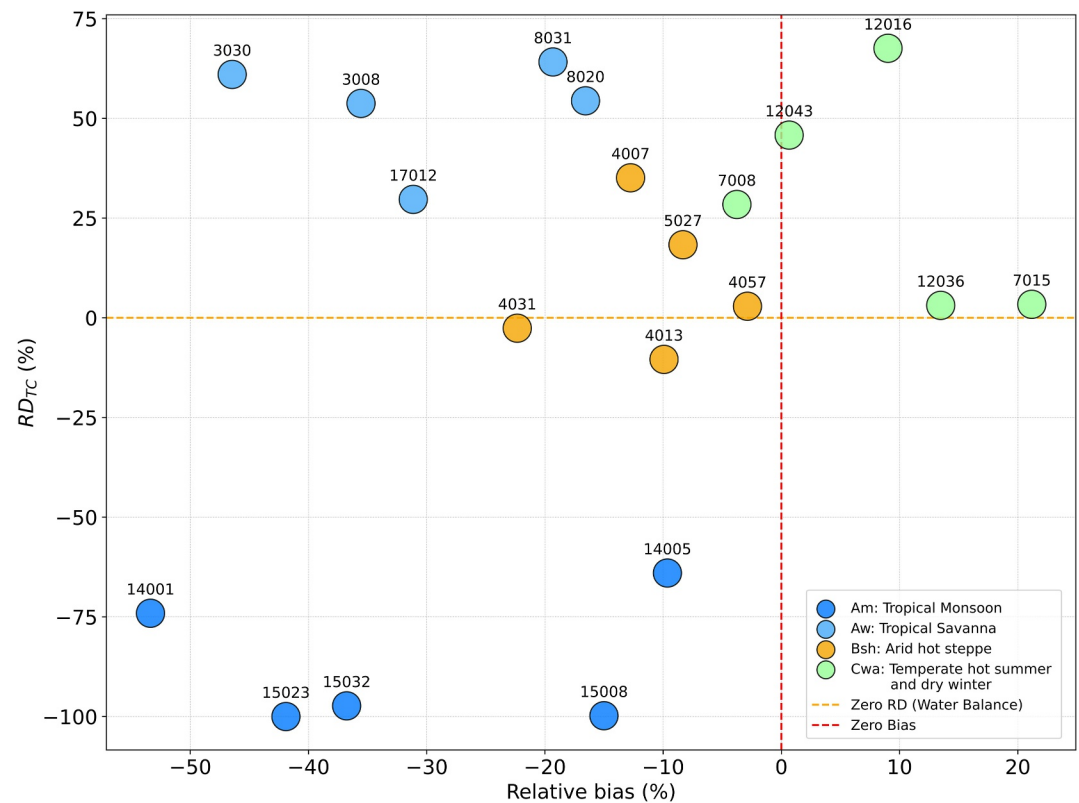


Figure 7. Scatter plot of RD_{TC} (%) and relative bias (%) of five representative catchments from each of the four main climate zones (Am, Aw, BSh, Cwa). The points have the labels of their respective gauge IDs. $RD_{TC} = 0$ and Bias = 0 are depicted using orange and red dashed lines, respectively.

balance of the catchment. This limitation also leads to lower simulated streamflow and ultimately to a systematic underestimation of the observed discharge.

The top-left region is mainly associated with the tropical savanna zone. The high positive values of RD_{TC} indicate that the model simulates trash cell fluxes that are larger than the evapotranspiration, suggesting an overestimation of residual outflows. This could point to additional loss pathways, but combined with the simultaneous negative bias in streamflow shows that the model may not be accurately capturing the actual outflow dynamics. Instead, it appears to release too much water through the trash cell, leaving less water available for streamflow simulation and resulting in systematic underestimation (negative bias). A likely reason for this behavior is the presence of a large number of medium to large dams in these catchments (Mangukiya et al., 2025). These structures regulate river flow and modify natural storage and release processes, which are not explicitly included in the model, leading to higher trash cell values and poorer streamflow estimates. The gauges in the remaining two climatic regions are largely centered around zero, indicating satisfactory water conservation and model prediction.

4.6. Cross-Region Validation Using CAMELS-US

The CAMELS-US data set (Addor et al., 2017) has been extensively employed in previous studies for evaluating DL models, particularly LSTM-based architectures and their variants. It serves as an established benchmark for rainfall-runoff modeling. To test the generalizability and robustness of the MC-LSTM-DS model, we also evaluated the model on the CAMELS-US data set. To ensure fair comparison with prior work, we employed the same input features, temporal split, basin subset, and hyperparameter settings as Y. Wang et al. (2025b). The details are given in Section S3 in Supporting Information S1. We utilize the evaluation metrics of NSE, KGE, Pearson-r, FHV, and FLV.

Table 4 presents the performance of MC-LSTM-DS, MC-LSTM, and LSTM on the testing set of the CAMELS-US data set. We evaluated these models using the same experimental setup. In addition, we report the

Table 4

Evaluation Metrics of SAC-SMA, LSTM, MC-LSTM, MCR-LSTM and MC-LSTM-DS on the CAMELS-US Data Set

Metrics	SAC-SMA	LSTM	MC-LSTM	MCR-LSTM	MC-LSTM-DS
NSE	0.67	0.72	0.71	0.71	0.72
KGE	0.67	0.73	0.71	0.72	0.74
Pearson-r	0.83	0.86	0.86	0.86	0.87
FHV (%)	−22.75	−17.51	−19.03	−16.88	−15.68
FLV (%)	48.53	−4.48	−50.13	−29.66	−7.70

Note. The values of SAC-SMA, LSTM, MC-LSTM, and MCR-LSTM are directly taken from Y. Wang et al. (2025b). The best value for each metric is highlighted in blue, and the second best is underlined.

performance metrics of the SAC-SMA and MCR-LSTM as presented in Y. Wang et al. (2025b). Their study employs the SAC-SMA model with parameters pre-calibrated by Frame et al. (2022). We report only the median values, as they are commonly used in hydrological studies for their robustness to outliers. For NSE, we observe that the proposed MC-LSTM-DS achieves the same performance as LSTM. However, KGE offers a more robust and balanced evaluation of model performance than NSE. With respect to the Pearson-r, KGE, and FHV, we observe that MC-LSTM-DS outperforms all the models, giving it the current state-of-the-art status for overall streamflow prediction. Moreover, its superior performance in FHV demonstrates its state-of-the-art capability for flood prediction. In terms of the FLV metric, MC-LSTM-DS performs marginally below the standard LSTM but substantially better than the other mass-conserving models, confirming that the additional cell state effectively captures low flows along with high flows as intended.

To assess the statistical significance of the performance differences, we applied the Wilcoxon signed-rank test to basin-wise NSE and KGE scores, with the results summarized in Table 5. The test indicates that MC-LSTM-DS significantly outperforms MC-LSTM for both metrics, whereas no statistically significant difference is observed between MC-LSTM-DS and LSTM at the 5% significance level. These findings suggest that the proposed architecture consistently improves upon MC-LSTM while achieving predictive performance comparable to that of LSTM on the CAMELS-US data set.

Although MC-LSTM-DS does not significantly outperform LSTM, it represents a meaningful advancement. The proposed model consistently outperforms both MC-LSTM and MCR-LSTM across the evaluated performance metrics. In particular, it achieves substantially lower FLV values than the other mass-conserving models, reaching performance comparable to that of LSTM. Moreover, MC-LSTM-DS attains predictive skill comparable to LSTM while simultaneously enforcing mass conservation and providing a framework for interpreting model behavior through physically meaningful internal states. These results suggest that the proposed architecture successfully balances predictive performance, physical consistency, and interpretability.

5. Limitations and Future Work

This study represents a meaningful advancement in the application of DL for streamflow prediction. However, several limitations need to be discussed. The decomposition of streamflow contributions into the two internal states is not unique, and the learned states should therefore be interpreted as latent representations that are consistent with, rather than direct observations of, underlying hydrological processes. While the proposed MC-LSTM-DS model

Table 5

Results of the One-Sided Wilcoxon Signed-Rank Test Comparing MC-LSTM-DS Against Other Baseline Models' Performance Across Basins

Metric	Baseline model	p (MC-LSTM-DS > model)	Median difference
NSE	MC-LSTM	4.71×10^{-10}	0.01
	LSTM	5.81×10^{-2}	0.0003
KGE	MC-LSTM	9.89×10^{-19}	0.02
	LSTM	1.98×10^{-1}	0.005

consistently outperforms both MC-LSTM and LSTM on the Indian data set, the analyses presented throughout the study suggest groundwater contributions to streamflow. Although the additional cell state in MC-LSTM-DS is designed to capture slower storage and release processes, it does not explicitly represent groundwater inflows or their dynamics. This omission may limit the model's ability to fully account for subsurface hydrological processes, particularly in regions where baseflow contributions are significant. Future work should therefore focus on developing mechanisms to explicitly incorporate groundwater data and its temporal dynamics into the modeling framework, which could further enhance both predictive performance and physical interpretability.

The CAMELS-US data set is used in this study primarily for benchmarking purposes. The main focus of the study is the Indian subcontinent, and all results and interpretations are specific to this region. Consequently, the findings do not generalize to other regions such as CONUS, where additional hydrological processes, particularly snowmelt, play an important role in streamflow generation. Applying the same analysis in such regions may therefore lead to different conclusions. Future studies could extend this framework to regions like CONUS to explicitly account for snow-driven hydrological dynamics. More broadly, owing to the reproducible and open-source nature of the proposed framework, the model can be readily implemented in other geographical settings, enabling similar analyses across diverse hydro-climatic regimes.

6. Conclusion

This study proposes a novel architecture, MC-LSTM-DS, which has two cell states and conservation-aware gating. We evaluated our model on the CAMELS-IND data set against the MC-LSTM and standard LSTM. The spatial distribution of skill scores shows that, although enforcing mass conservation may degrade performance in semi-arid regions, MC-LSTM-DS improves upon MC-LSTM in these regions. Investigation of the cell states was carried out to interpret the representations learned by the model across different climatic zones. Inspecting the partition mechanism reveals that it captures signatures of different runoff-generating processes. Further, to check how mass conservation holds in the real world, we carry out a water balance analysis. Finally, to prove the generalizability of our model, we evaluate it on the benchmark CAMELS-US data set. The main findings of this study can be summarized as follows:

1. MC-LSTM-DS outperforms standard LSTM and other mass-conserving LSTM variants on CAMELS-IND on most of the considered metrics and matches the performance of LSTM on the CAMELS-US data set with a superior FLV than other mass-conserving LSTMs. This shows that restoring the dual memory while also enforcing mass conservation helps the models to be physically consistent while providing state-of-the-art predictions.
2. Investigation of the cell states reveals that the c state suggests fast, short-term hydrologic responses, while the l state represents slower, long-term processes. The relative magnitudes show consistency with expected hydrologic behavior in different climate zones, like efficient runoff generation in tropical basins, strong wet-dry seasonality in savannas, and high water losses and groundwater dependence in semi-arid catchments. Also, the partition mechanisms show similarity to infiltration excess and saturation excess overland flow behavior.
3. The assumed water balance breaks down in practice in the tropical catchments. In the tropical monsoon catchments, the water balance analysis indicates additional water inputs likely originating from groundwater inflows. While in the tropical savanna, the model fails to effectively capture the water losses from the catchments, probably owing to increased anthropogenic influences.

To conclude, the models and analyses proposed in this study could be further extended to test in other study regions. Combining physical realism with DL, this study lays the groundwork for more generalizable and interpretable hydrological prediction frameworks, which could be implemented across the globe.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Availability Statement

The input data used in this study are sourced from publicly available data sets. The CAMELS-IND data set (Mangukiyi et al., 2025), which includes the meteorological forcings, catchment attributes, and streamflow

observations, can be downloaded from <https://zenodo.org/records/14999580>. For the CAMELS-US (Newman et al., 2022), the NLDAS extended forcings can be downloaded from <https://www.hydroshare.org/resource/0a68bfd7ddf642a8be9041d60f40868c/> and the whole data set from <https://zenodo.org/records/15529996>. All model-training code was developed using the NeuralHydrology Python library (Kratzert et al., 2022). The MC-LSTM-DS model (Toraskar et al., 2026) integrated into the NeuralHydrology code-base, along with the run results, ensemble outputs, analysis scripts, and usage instructions, is publicly available through the Zenodo repository <https://zenodo.org/records/17881538>.

Acknowledgments

We acknowledge IISER Bhopal and University of Leeds for providing institutional support. ST acknowledges the Department of Science and Technology (DST), Government of India, for INSPIRE-SHE (Scholarship for Higher Education). The authors express their gratitude to the developers of the NeuralHydrology Python library for developing and maintaining the codebase for public use. We would like to acknowledge Anusandhan National Research Foundation, New Delhi, India, for funding through: CRG/2022/008185.

References

- Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The camels data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, 21(10), 5293–5313. <https://doi.org/10.5194/hess-21-5293-2017>
- Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., & Pappenberger, F. (2013). Glofas—global ensemble streamflow forecasting and flood early warning. *Hydrology and Earth System Sciences*, 17(3), 1161–1175. <https://doi.org/10.5194/hess-17-1161-2013>
- Alizadeh, B., Bafti, A. G., Kamangir, H., Zhang, Y., Wright, D. B., & Franz, K. J. (2021). A novel attention-based lstm cell post-processor coupled with Bayesian optimization for streamflow prediction. *Journal of Hydrology*, 601, 126526. <https://doi.org/10.1016/j.jhydrol.2021.126526>
- Apel, H., Thieken, A. H., Merz, B., & Blöschl, G. (2004). Flood risk assessment and associated uncertainty. *Natural Hazards and Earth System Sciences*, 4(2), 295–308. <https://doi.org/10.5194/nhess-4-295-2004>
- Balthazar, L. D., Miranda, F., Cândido, V. B., Capriles, P., Moraes, M., Ribeiro, C. M., et al. (2024). Long-term natural streamflow forecasting under drought scenarios using data-intelligence modeling. *Water Cycle*, 5, 266–277. <https://doi.org/10.1016/j.watcyc.2024.07.001>
- Beck, H. E., McVicar, T. R., Vergopolan, N., Berg, A., Lutsko, N. J., Dufour, A., et al. (2023). High-resolution (1 km) Köppen-Geiger maps for 1901–2099 based on constrained cmip6 projections. *Scientific Data*, 10(1), 724. <https://doi.org/10.1038/s41597-023-02549-6>
- Botterill, T. E., & McMillan, H. K. (2023). Using machine learning to identify hydrologic signatures with an encoder–decoder framework. *Water Resources Research*, 59(3), e2022WR033091. <https://doi.org/10.1029/2022wr033091>
- Chiang, F., Mazdiyasn, O., & AghaKouchak, A. (2021). Evidence of anthropogenic impacts on global drought frequency, duration, and intensity. *Nature Communications*, 12(1), 2754. <https://doi.org/10.1038/s41467-021-22314-w>
- Chuphal, D. S., & Mishra, V. (2023). Hydrological model-based streamflow reconstruction for Indian sub-continental river basins, 1951–2021. *Scientific Data*, 10(1), 717. <https://doi.org/10.1038/s41597-023-02618-w>
- Devó, P., Basso, S., & Marani, M. (2025). Estimation of extreme floods using a statistical and conceptual model of the hydrological response. *Water Resources Research*, 61(5), e2024WR038667. <https://doi.org/10.1029/2024wr038667>
- Fan, Y. (2019). Are catchments leaky? *Wiley Interdisciplinary Reviews: Water*, 6(6), e1386. <https://doi.org/10.1002/wat2.1386>
- Feng, D., Fang, K., & Shen, C. (2020). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Resources Research*, 56(9), e2019WR026793. <https://doi.org/10.1029/2019wr026793>
- Fleig, A. K., Tallaksen, L. M., Hisdal, H., & Demuth, S. (2006). A global evaluation of streamflow drought characteristics. *Hydrology and Earth System Sciences*, 10(4), 535–552. <https://doi.org/10.5194/hess-10-535-2006>
- Frame, J. M., Kratzert, F., Gupta, H. V., Ullrich, P., & Nearing, G. S. (2023). On strictly enforced mass conservation constraints for modelling the rainfall-runoff process. *Hydrological Processes*, 37(3), e14847. <https://doi.org/10.1002/hyp.14847>
- Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., et al. (2022). Deep learning rainfall-runoff predictions of extreme events. *Hydrology and Earth System Sciences*, 26(13), 3377–3392. <https://doi.org/10.5194/hess-26-3377-2022>
- Gourley, J. J., & Vieux, B. E. (2006). A method for identifying sources of model uncertainty in rainfall-runoff simulations. *Journal of Hydrology*, 327(1–2), 68–80. <https://doi.org/10.1016/j.jhydrol.2005.11.036>
- Gupta, S. C., Kessler, A. C., Brown, M. K., & Zvomuya, F. (2015). Climate and agricultural land use change impacts on streamflow in the upper midwestern United States. *Water Resources Research*, 51(7), 5301–5317. <https://doi.org/10.1002/2015wr017323>
- Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1–2), 80–91. <https://doi.org/10.1016/j.jhydrol.2009.08.003>
- Hager, W. H., & Boes, R. M. (2014). Hydraulic structures: A positive outlook into the future. *Journal of Hydraulic Research*, 52(3), 299–310. <https://doi.org/10.1080/00221686.2014.923050>
- He, M., Jiang, S., Ren, L., Cui, H., Qin, T., Du, S., et al. (2024). Streamflow prediction in ungauged catchments through use of catchment classification and deep learning. *Journal of Hydrology*, 639, 131638. <https://doi.org/10.1016/j.jhydrol.2024.131638>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G. S., et al. (2021). Mc-lstm: Mass-conserving lstm. In *International conference on machine learning* (pp. 4275–4286).
- Kim, T., Yang, T., Gao, S., Zhang, L., Ding, Z., Wen, X., et al. (2021). Can artificial intelligence and data-driven machine learning models match or even replace process-driven hydrologic models for streamflow simulation? A case study of four watersheds with different hydro-climatic regions across the conus. *Journal of Hydrology*, 598, 126423. <https://doi.org/10.1016/j.jhydrol.2021.126423>
- Kratzert, F. (2019). Camels extended nldas forcing data [Dataset]. *HydroShare*. <https://doi.org/10.4211/hs.0a68bfd7ddf642a8be9041d60f40868c>
- Kratzert, F., Gauch, M., Klotz, D., & Nearing, G. (2024). Hess opinions: Never train a long short-term memory (lstm) network on a single basin. *Hydrology and Earth System Sciences*, 28(17), 4187–4201. <https://doi.org/10.5194/hess-28-4187-2024>
- Kratzert, F., Gauch, M., Nearing, G., & Klotz, D. (2022). Neuralhydrology — A python library for deep learning research in hydrology. *Journal of Open Source Software*, 7(71), 4050. <https://doi.org/10.21105/joss.04050>
- Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall-runoff modelling using long short-term memory (lstm) networks. *Hydrology and Earth System Sciences*, 22(11), 6005–6022. <https://doi.org/10.5194/hess-22-6005-2018>
- Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, 55(12), 11344–11354. <https://doi.org/10.1029/2019wr026065>
- Kratzert, F., Klotz, D., Hochreiter, S., & Nearing, G. S. (2020). A note on leveraging synergy in multiple meteorological datasets with deep learning for rainfall-runoff modeling. *Hydrology and Earth System Sciences Discussions*, 1–26.
- Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (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. <https://doi.org/10.5194/hess-23-5089-2019>

- Kreibich, H., Van Loon, A. F., Schröter, K., Ward, P. J., Mazzoleni, M., Sairam, N., et al. (2022). The challenge of unprecedented floods and droughts in risk management. *Nature*, 608(7921), 80–86. <https://doi.org/10.1038/s41586-022-04917-5>
- Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., & Di, Z. (2017). A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *Wiley Interdisciplinary Reviews: Water*, 4(6), e1246. <https://doi.org/10.1002/wat2.1246>
- Lin, Y., Wang, D., Wang, G., Qiu, J., Long, K., Du, Y., et al. (2021). A hybrid deep learning algorithm and its application to streamflow prediction. *Journal of Hydrology*, 601, 126636. <https://doi.org/10.1016/j.jhydrol.2021.126636>
- Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Mai, J., Craig, J. R., Tolson, B. A., & Arsenaault, R. (2022). The sensitivity of simulated streamflow to individual hydrologic processes across North America. *Nature Communications*, 13(1), 455. <https://doi.org/10.1038/s41467-022-28010-7>
- Mangunkiya, N. K., Kanneganti, B. K., Dey, P., Sharma, S., Bejagam, V., Mujumdar, P. P., & Sharma, A. (2025). Camels-ind: Hydrometeorological time series and catchment attributes for 472 catchments in peninsular India. *Zenodo*. <https://doi.org/10.5281/zenodo.14999580>
- Mangunkiya, N. K., Sharma, A., & Shen, C. (2023). How to enhance hydrological predictions in hydrologically distinct watersheds of the Indian subcontinent? *Hydrological Processes*, 37(7), e14936. <https://doi.org/10.1002/hyp.14936>
- Meko, D. M., & Woodhouse, C. A. (2010). Application of streamflow reconstruction to water resources management. In *Dendroclimatology: Progress and prospects* (pp. 231–261). Springer.
- Nanditha, J., & Mishra, V. (2021). On the need of ensemble flood forecast in India. *Water Security*, 12, 100086. <https://doi.org/10.1016/j.wasec.2021.100086>
- Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, 10(3), 282–290. [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6)
- Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research*, 57(3), e2020WR028091. <https://doi.org/10.1029/2020wr028091>
- Newman, A. J., Sampson, K., Clark, M., Bock, A., Viger, R., Blodgett, D., et al. (2022). Camels: Catchment attributes and meteorology for large-sample studies. *Zenodo*. <https://doi.org/10.5065/D6MW2F4D>
- Ni, L., Wang, D., Singh, V. P., Wu, J., Wang, Y., Tao, Y., & Zhang, J. (2020). Streamflow and rainfall forecasting by two long short-term memory-based models. *Journal of Hydrology*, 583, 124296. <https://doi.org/10.1016/j.jhydrol.2019.124296>
- Rohde, M. M. (2023). Floods and droughts are intensifying globally. *Nature Water*, 1(3), 226–227. <https://doi.org/10.1038/s44221-023-00047-y>
- Sabzipour, B., Arsenaault, R., Troin, M., Martel, J.-L., Brissette, F., Brunet, F., & Mai, J. (2023). Comparing a long short-term memory (lstm) neural network with a physically-based hydrological model for streamflow forecasting over a Canadian catchment. *Journal of Hydrology*, 627, 130380. <https://doi.org/10.1016/j.jhydrol.2023.130380>
- Sahoo, B. B., Jha, R., Singh, A., & Kumar, D. (2019). Long short-term memory (lstm) recurrent neural network for low-flow hydrological time series forecasting. *Acta Geophysica*, 67(5), 1471–1481. <https://doi.org/10.1007/s11600-019-00330-1>
- Sahu, G., Mangunkiya, N. K., & Sharma, A. (2025). Does mc-lstm model improve the reliability of streamflow prediction in human-influenced watersheds? *Journal of Hydrology*, 665, 134711. <https://doi.org/10.1016/j.jhydrol.2025.134711>
- Singh Chuphal, D., Thirumalai, K., & Mishra, V. (2025). Recent drying of the ganga river is unprecedented in the last 1,300 years. *Proceedings of the National Academy of Sciences*, 122(40), e2424613122. <https://doi.org/10.1073/pnas.2424613122>
- Toraskar, S., Niranjannaik, M., Singh, A., & Gaurav, k. (2026). Mass-conserving lstm with dual states for streamflow prediction: Separating quickflow and slow storage. *Zenodo*. <https://doi.org/10.5281/zenodo.17881538>
- Velazquez, J. A., Ancil, F., Ramos, M.-H., & Perrin, C. (2011). Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures. *Advances in Geosciences*, 29, 33–42. <https://doi.org/10.5194/adgeo-29-33-2011>
- Wang, Y., Zhang, L., Erichson, N. B., & Yang, T. (2025a). Investigating the streamflow simulation capability of a new mass-conserving long short-term memory (Mc-lstm) model across the contiguous United States. *Journal of Hydrology*, 658, 133161. <https://doi.org/10.1016/j.jhydrol.2025.133161>
- Wang, Y., Zhang, L., Erichson, N. B., & Yang, T. (2025b). A mass conservation relaxed (mcr) lstm model for streamflow simulation across conus. *Water Resources Research*, 61(8), e2024WR039131. <https://doi.org/10.1029/2024wr039131>
- Wang, Z., Xu, N., Bao, X., Wu, J., & Cui, X. (2024). Spatio-temporal deep learning model for accurate streamflow prediction with multi-source data fusion. *Environmental Modelling and Software*, 178, 106091. <https://doi.org/10.1016/j.envsoft.2024.106091>
- Xu, W., Chen, J., Zhang, X. J., Xiong, L., & Chen, H. (2022). A framework of integrating heterogeneous data sources for monthly streamflow prediction using a state-of-the-art deep learning model. *Journal of Hydrology*, 614, 128599. <https://doi.org/10.1016/j.jhydrol.2022.128599>
- Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the nws distributed hydrologic model. *Water Resources Research*, 44(9). <https://doi.org/10.1029/2007wr006716>
- Zhao, T., Cai, X., & Yang, D. (2011). Effect of streamflow forecast uncertainty on real-time reservoir operation. *Advances in Water Resources*, 34(4), 495–504. <https://doi.org/10.1016/j.advwatres.2011.01.004>