Water Resources Research

RESEARCH ARTICLE
10.1029/2024WR039792

Locally Relevant Streamflow by Integrating a Land Surface Model Ensemble With a Two-Stage LSTM Post-Processor

Bhanu Magotra[1] and **Manabendra Saharia**[1,2]

[1]Department of Civil Engineering, Indian Institute of Technology Delhi, New Delhi, India, [2]Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, New Delhi, India

Special Collection:
Advancing Interpretable AI/ML Methods for Deeper Insights and Mechanistic Understanding in Earth Sciences: Beyond Predictive Capabilities

Key Points:

- A two-stage Long-Short Term Memory (LSTM) hybrid post-processing scheme is proposed to generate locally relevant streamflow from a process-based land surface model ensemble
- A residual predicting LSTM is sequentially paired with an auto-regressive LSTM, enabling data integration and optimal ensemble weighing
- The nationwide median Kling-Gupta Efficiency for daily streamflow increases from 0.18 to 0.60 with improvements in 208 out of 220 catchments in India

Supporting Information:

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

Correspondence to:

M. Saharia,
msaharia@iitd.ac.in

Citation:

Magotra, B., & Saharia, M. (2026). Locally relevant streamflow by integrating a land surface model ensemble with a two-stage LSTM post-processor. *Water Resources Research*, *62*, e2024WR039792. https://doi.org/10.1029/2024WR039792

Received 22 DEC 2024
Accepted 5 JAN 2026

Author Contributions:

Conceptualization: Bhanu Magotra
Data curation: Bhanu Magotra
Formal analysis: Bhanu Magotra

Abstract Process-based land surface models (LSMs) are widely used for global water cycle and runoff assessments, but when integrated with hydrodynamic models, the streamflow simulations exhibit significant uncertainties in uncalibrated mode, limiting their effectiveness in local hydrology applications. The calibration of LSMs against observed streamflow across large basins and regions is computationally prohibitive and sometimes degrades performance of other variables. In contrast, deep learning models, particularly Long-Short Term Memory (LSTM) networks, have shown promising results in streamflow simulations, but they are often limited by poor reproducibility of other water cycle variables. This study presents a hybrid modeling framework that integrates process-based models with deep learning to improve daily streamflow simulations without requiring basin-specific calibration. The framework is showcased on a national scale using a multi-model hydrologic ensemble from the Indian Land Data Assimilation System (ILDAS). It is integrated with a proposed two-stage post-processor, which pairs a residual error prediction LSTM with an auto-regressive meta-learning LSTM to predict 1-day ahead streamflow. Trained on multi-decadal data from 220 catchments across India, the framework improves Kling-Gupta Efficiency in 208 catchments, raising the national median from 0.18 (uncalibrated) to 0.60. It also reduced peak flow timing error and peak mean absolute percentage error by 25% in 135 catchments. During monsoon and post-monsoon periods, residual error interquartile range (IQR) decreased by 66.3% and 81.7%, respectively. This approach has the potential to integrate LSMs with deep learning for more accurate and locally relevant streamflow predictions, while enhancing other water cycle variables through methods like data assimilation.

Plain Language Summary This study focuses on improving how we predict river streamflow, which is crucial for water management, including flood forecasting. Traditional models, called land surface models (LSMs), simulate the water cycle on a large scale but struggle to accurately predict streamflow at local levels without detailed adjustments. However, these adjustments are difficult and time-consuming. On the other hand, deep learning models like Long-Short Term Memory (LSTM) networks can predict streamflow well but don't explain other important water-related variables, like soil moisture. Our research combines LSMs with deep learning to improve daily streamflow predictions without needing complex adjustments for each area. We tested this method across 220 rivers in India and found it improved streamflow accuracy in 208 of them. This new approach could lead to more accurate water forecasts, helping with flood prevention and better water resource management.

1. Introduction

The process-based land surface models (LSMs) resolve land-atmosphere interaction to provide accurate and consistent estimates of land surface states and fluxes, such as soil moisture, snow cover, evapotranspiration, and runoff. They are a key component in Land Data Assimilation Systems (LDASs) that have been set up in many countries to assist policy makers in land and water resources planning and mitigation of natural hazards. However, LSMs often exhibit significant uncertainties due to imperfect model structures, inadequate process representations, biases in meteorological inputs, and incorrect assumptions about initial conditions and parameters (Jacobs et al., 2008; Moreira et al., 2019; H. Wang et al., 2023; Xia et al., 2019). These uncertainties are particularly evident in streamflow simulations. For example, Lohmann et al. (2004) assessed streamflow and water balance simulations across the continental United States using four land surface models in North American Land Data Assimilation System (NLDAS; Mitchell et al., 2004). Their findings revealed generally poor streamflow estimates across multiple river basins. Several studies have evaluated outputs from Global Land Data Assimilation System (GLDAS; Rodell et al., 2004) across diverse global regions, and consistently identified biases and timing

Funding acquisition: Manabendra Saharia
Investigation: Bhanu Magotra
Methodology: Bhanu Magotra
Project administration: Manabendra Saharia
Resources: Manabendra Saharia
Software: Bhanu Magotra
Supervision: Manabendra Saharia
Validation: Bhanu Magotra
Visualization: Bhanu Magotra
Writing – original draft: Bhanu Magotra
Writing – review & editing: Manabendra Saharia

errors in the simulated streamflow responses (Bai et al., 2016; W. Wang et al., 2016; Zaitchik et al., 2010). Similarly, studies have been conducted in the Indian subcontinent region performing various water balance assessments include streamflow evaluation. (Magotra et al., 2024) performed hydrological assessments across major river basins in India and revealed that the land surface models (LSMs) in the Indian Land Data Assimilation System (ILDAS) tend to overestimate daily streamflow over the region. Additionally, multiple studies have been conducted over the Narmada River Basin which established significant magnitude and timing errors in streamflow, and proposed strategies to correct them (Ossandón et al., 2022; Prakash & Mishra, 2022). Thus, it remains a challenge to obtain locally relevant streamflow from such hydrological models, rendering the system unsuitable for large scale water resources planning and forecasting purposes.

One of the common strategies to improve hydrological simulations is to calibrate the model to reduce uncertainties due to incorrect model parameters. However, calibrating an LSM at multiple basins is computationally prohibitive, requiring substantial computing resources, diverse observational data, and complex strategies (Hirpa et al., 2018; Höge et al., 2018; Mizukami et al., 2017). The computational complexity and time required to calibrate LSMs make them expensive to operate at large scales. Moreover, calibrating LDASs for a single output variable like streamflow often leads to the degradation of other variables, complicating their integration into operational frameworks (Troy et al., 2008; Xia et al., 2018). Additionally, data assimilation can be used to update the model's internal states that eventually result in improved predictions (Chakraborty et al., 2024; Largeron et al., 2020; Mitchell et al., 2004; Reichle, 2008; Reichle & Koster, 2005). However, it becomes computationally expensive to apply data assimilation techniques in large heterogenous regions that require high resolution distributed land surface modeling. Moreover, it is a challenge to apply calibration and data assimilation in catchments that have been heavily influenced by human interventions such as reservoirs (Ye et al., 2014). Therefore, statistical post-processing is seen as a viable strategy to enhance hydrological outputs because it aims to establish a direct relationship between simulated and observed values. This approach seeks to minimize the systematic bias resulting from various sources of uncertainty in the hydrological models as a final step before the outputs are released. Past studies have shown that statistical post-processing has the potential to improve hydrological simulations by applying techniques based on Bayes Theorem such as Bayesian Model Averaging (BMA, (Hoeting et al., 1999; Raftery et al., 2005) and Bayesian processor of ensemble (Marty et al., 2015). The regression based post-processing approach includes quantile regression (Weerts et al., 2010) and General Linear Model Post-Processor (GLMPP; Zhao et al., 2011). Many studies have evaluated alternative approaches that do not require a-priori probability estimations (Evin et al., 2014; Romero-Cuellar et al., 2019; Solomatine & Shrestha, 2009; Zhang & Zhao, 2012).

With recent developments in deep learning, studies have shown the potential of advanced neural networks such as Long Short-Term Memory (LSTM; (Hochreiter & Schmidhuber, 1997) networks to simulate rainfall-runoff process when trained on meteorological data and catchment characteristics (Kratzert et al., 2019; Lees et al., 2021). These studies demonstrate that LSTM models can efficiently ingest large observation data sets while requiring significantly fewer computational resources than traditional methods. However, lack of explainability and the black-box nature of such models is a matter of concern for hydrologists. Hence, a hybrid setting could be well suited for deep-learning models, where models such as LSTMs can act as post processors for process-based hydrological models. Additionally, since a post-processor aims to directly minimize the error between model outputs and observations, a deep learning model such as LSTM is best suited for the task as it can be trained on large data sets to predict and correct the residual error. Recent studies have evaluated such hybrid solutions and have shown that LSTM models as post-processors can significantly improve model simulations. Frame et al. (2021) postprocessed the National Water Model (NWM) using LSTMs and found that post-processors significantly improved NWM outputs in terms of bias and timing. Han and Morrison (2022) performed error predictions on NWM outputs using LSTMs and reduced the bias in hourly runoff at lead times between 1 and 18 hr from a range of −60%–80% to −15%–10%. Liu et al. (2024) developed a national scale hybrid postprocessor to enhance the streamflow outputs from a physically based hydrological model called Danish National Water Resources Model (DKM). They used various model configurations and observed outperformance by LSTM based postprocessors. Overall, past studies have highlighted the potential of deep learning models to learn intrinsic relationships among the drivers of rainfall-runoff processes, both as standalone models and hybrid frameworks. In this study we systematically evaluate a hybrid DL-based post-processing framework for its effectiveness and scalability across the complex and heterogeneous catchments of the Indian subcontinent. This study focuses on a novel strategy designed to enhance six distinct daily streamflow outputs from an LDAS setup. The approach

involves residual error correction at stage 1, followed by the optimal weighing of various ensemble members with observed streamflow moving average for the final output using a combination of meta-learning and autoregressive modeling.

We demonstrate that accurate localized streamflow response can be generated from ILDAS in hydrologically diverse catchments using a DL-based post-processor. We used two Long Short-Term Memory (LSTM) model layers, termed residual error correction layer and meta-learning layer. These layers are trained on 220 catchments across the Indian subcontinent using dynamic meteorological forcings and static catchment physical attributes. The residual error correction layer is trained to predict the residual error (observed minus simulated) in ILDAS streamflow output by learning the complex patterns between residuals and upstream sources of information such as meteorological forcings and catchment characteristics. The meta learning layer integrates the residual-corrected ensemble streamflow with recent streamflow observations and generates a deterministic daily streamflow time-series for each catchment produced from an optimal combination of the six corrected daily streamflow series and auto-regressed streamflow observations. The novelty of this two-stage framework lies in its decoupling of residual error correction from the final ensemble combination. Instead of attempting a single model to capture all complexities, it dedicates an initial stage to correcting systematic errors (residuals) in the base streamflow simulations. These corrected simulations are then fed into a separate meta-learning stage that optimally blends these independent estimates with recent observations, achieving a more robust and accurate localized flow prediction.

Overall, the objectives of this study are to:

1. Quantify the skill of a hybrid deep learning framework to generate locally relevant, accurate, and reliable streamflow outputs from a process-based model ensemble.
2. Present a comprehensive framework for operational streamflow forecasting over the Indian subcontinent with residual error correction and real-time data integration capabilities.

The paper is organized as follows: Section 2 describes the data sets used in this study. It also briefly explains the LSTM model and the methodology involved and evaluation of the results. In Section 3, results are presented along with relevant discussion. Finally, Section 4 provides the conclusions of the study and future work.

2. Data and Methods

We integrate the state-of-the-art LSTM models with a hydrologic-hydrodynamic setup over the Indian subcontinent, called the Indian Land Data Assimilation System (ILDAS). ILDAS generates 0.1° daily outputs for various water balance components, including soil moisture, runoff, ground water, and streamflow (Magotra et al., 2024). It operates in an ensemble mode using two land surface models and three forcing data sets, resulting in an output stream of six time series at streamflow gauging stations (see Section 2.1). However, in its current form, ILDAS is uncalibrated and without any explicit representation of water management practices such as reservoirs and irrigation, exhibiting large errors in the daily streamflow and hence, rendering it unsuitable for operational water resources management. Calibrating the ILDAS for individual catchments holds considerable potential to improve performance, but it requires large computations and complicated strategies to achieve nationwide calibrated parameters. Alternatively, our proposed hybrid framework leverages the computational efficiency of DL models and physical reliability of LSMs.

2.1. Hydrological Simulations From Indian Land Data Assimilation System (ILDAS)

The Indian Land Data Assimilation System (ILDAS; Magotra et al., 2024) is setup over the Indian subcontinent, which is built on NASA's Land Information System Framework (LISF). ILDAS incorporates two land surface models: Noah-MP 3.6 (Niu et al., 2011) and the Catchment Land Surface Model (CLSM; Koster et al., 2000), along with the HyMAP routing model (Getirana et al., 2012). Three different meteorological forcings were employed: (a) Indian Meteorological Department (IMD; Pai et al., 2014) data, (b) Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2; Gelaro et al., 2017), and (c) Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS; Funk et al., 2015). This combination resulted in an ensemble of six outputs, representing the two models combined with the three forcings. The ensemble outputs were generated at a spatial resolution of 0.1° and a daily timestep on a study domain extended from 68°E to 98°E and 5.5°N to 37.5°N, over a 41-year period from 1980 to 2021. Validation of the outputs was conducted using a

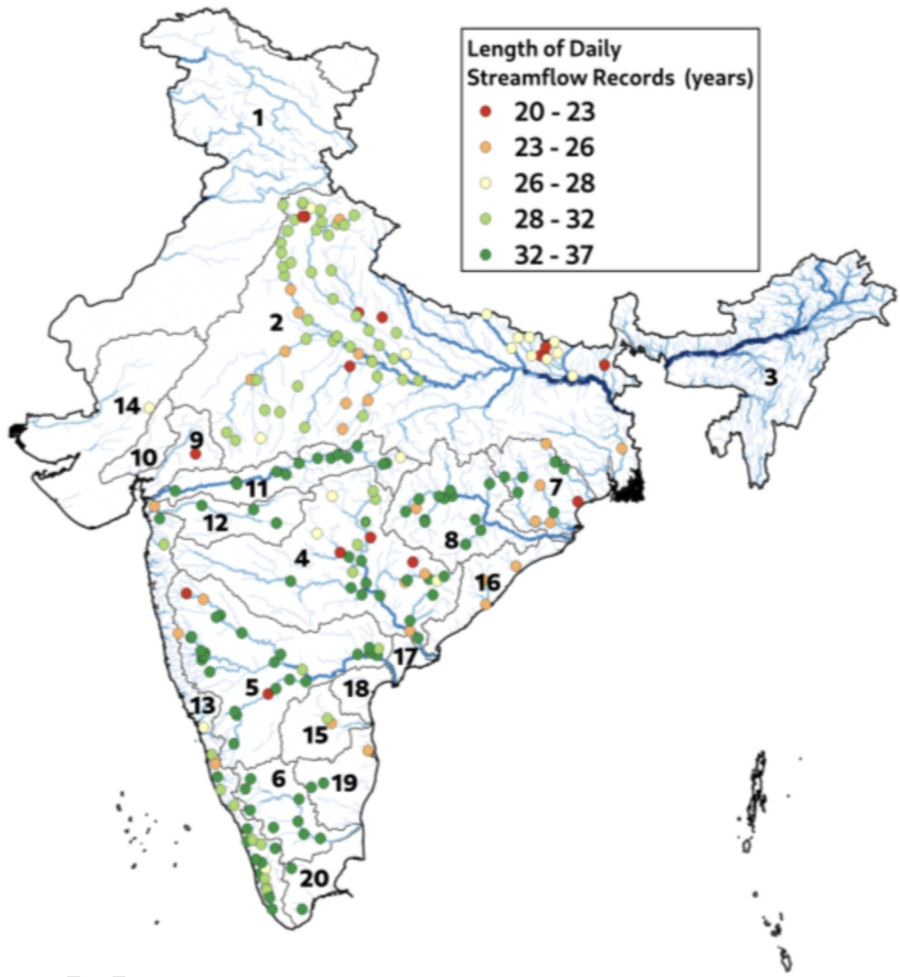

Figure 1. Spatial distribution of 220 streamflow gauge stations in India, characterized by their data length and major river basins. Basin names are provided in Table S2 in Supporting Information S1, corresponding to the numbered labels on the map.

range of error metrics, hydrological signatures, and climatological assessments. While this ensemble approach offers a degree of uncertainty mitigation by accounting for variations in model structure and meteorological inputs, significant biases were observed, particularly in daily streamflow estimates (Magotra et al., 2024). These biases were most pronounced in arid and anthropogenically impacted regions.

2.2. Observed Streamflow

The observed data set comprises daily streamflow measurements from 220 gauge locations across mainland India, with each location providing at least 20 years of continuous data. These observations were obtained from various governmental agencies through both public domain sources and official requests. To ensure data integrity and consistency, the records underwent thorough scrutiny for any discrepancies and were standardized for subsequent analysis. Missing data points were imputed using the day's climatological average, derived from historical records available up to that point. Catchment watersheds corresponding to each gauge location were delineated programmatically, treating each gauge station as a pour-point. This was achieved using the Multi-Error-Removed Improved-Terrain DEM (MERIT DEM; Yamazaki et al., 2017). The periods for training, validation, and testing were tailored for each basin, reflecting the varying lengths of the time series available for the different gauges. The locations of all gauge stations used in this study are depicted in Figure 1.

2.3. Meteorological and Geophysical Data Sets

We incorporated seven daily meteorological forcing time series and 20 geophysical attributes; all spatially averaged over the catchment areas (detailed in Table S1 in Supporting Information S1). Precipitation and related variables were sourced from the Indian Meteorological Department's (IMD) 0.25° daily gridded data set (see Pai et al., 2014), while temperature characteristics were derived from IMD's 1° daily temperature gridded data set (Srivastava et al., 2009). Additional meteorological variables were computed using the Indian Monsoon Data Assimilation and Analysis (IMDAA) data set (Rani et al., 2021), with values averaged over the catchments. For elevation and terrain-related variables, we employed theMERIT DEM. Soil properties were sourced from the SoilGrids data set (Hengl et al., 2017) and the HiHydroSoil v2.0, both of which provide global raster data at a 250-m resolution. Further, we incorporated data from the Moderate Resolution Imaging Spectroradiometer (MODIS) Leaf Area Index (LAI; Myneni et al., 2021) and evapotranspiration estimates from the ERA5 reanalysis data set (Hersbach et al., 2020). Land use and land cover characteristics were extracted from the National Remote Sensing Centre's (NRSC) LULC 30 m data set. Additionally, gauge locations were classified based on the presence or absence of upstream reservoirs, determined through visual inspection of Google Earth imagery. Out of 220 gauge locations, 86 were identified as the ones with upstream reservoirs.

2.4. Long Short-Term Memory Networks

LSTM (Long Short-Term Memory) networks are a variant of Recurrent Neural Networks (RNNs), first introduced by Hochreiter and Schmidhuber (1997). They were specifically designed to address the issues of exploding and vanishing gradients that are common in traditional RNNs, enabling LSTM models to effectively capture and learn long-term temporal dependencies. This makes them particularly suitable for recognizing patterns over extended time periods, such as vegetation seasonality, snowmelt, and groundwater dynamics. LSTM models retain information from previous time steps through memory cells, which function similarly to system states in dynamic models. The network processes a sequence of inputs in the form of a time series, denoted as $x = [x_1,\ldots, x_T]$ where x represents a vector of input features at each time step t. The LSTM structure and its operations can be described through the following set of equations, which outline the interactions between memory cells and gates that control the flow of information across time steps.

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma\left(W_f x_t + U f h_{t-1} + b f\right) \tag{2}$$

$$g_t = \tan h(W g x_t + U g h_{t-1} + b g) \tag{3}$$

$$o_t = \sigma(W_o x_t + U_o + b_o) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{5}$$

$$h_t = o_t \odot \tan h(c_t) \tag{6}$$

where i_t, f_t, and o_t are the input gate, the forget gate, and the output gate, respectively, g_t is the cell input and x_t is the network input at time step t ($1 \le t \le T$), h_{t-1} is the recurrent input and c_{t-1} the cell state from the previous time step. Figure 2 represents a schematic of a single LSTM cell. At the start, the hidden states and the cell states are initialized as a vector of zeros. W, U, and b are calibrated parameters specific to each gate, as denoted by the subscripts. The architecture involves two activation functions, the sigmoid, σ $(\cdot)$, and hyperbolic tangent, tan$h$ $(\cdot)$. $\odot$ denotes the element-wise multiplication. The c_t acts like the memory of the system which gets altered by the forget gate and a combination of input gate and cell update. Essentially, the forget gate controls which information needs to be deleted, and input gate with cell update governs the information that needs to be added to the cell state memory. Finally, the output gate controls the flow of information from cell state to output layer.

2.5. Experimental Setup

Initially, we conducted hyperparameter tuning using a sample of 50 randomly selected catchments. Employing the k-fold technique, we divided these catchments into five equal sets, training on 40 catchments and testing on

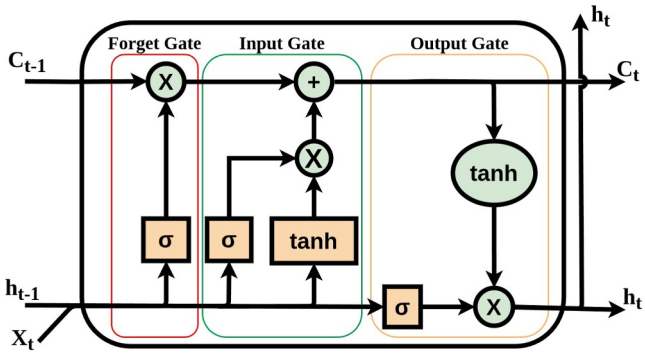

Figure 2. A schematic of the LSTM cell, showing the long-term cell state (c_t), hidden state (h_t), and the three gates: forget (f_t), input (i_t) and output (o_t).

the remaining 10 in each fold. We focused on four hyperparameters: dropout rate, number of hidden layers, sequence length, and batch size. The specific parameters chosen for tuning are detailed in Table 1. A five-member ensemble model was trained for 50 epochs using five years of data, split into training, validation, and testing sets in a 3:1:1 ratio. In total, we tested 700 different combinations and trained 3,500 models. The optimal hyperparameter values were determined based on the highest mean ensemble Kling-Gupta Efficiency (KGE) value.

The raw ILDAS streamflow ensemble output is post-processed using the methodology shown in Figure 3. The details of various model configurations used in the study are summarized in Table 2. Initially, individual LSTM models are used to correct the residual errors in each of the six streamflow members. The residual error, representing a collective effect of upstream uncertainties in ILDAS's inputs, model structure, and process representation, is the difference between observed and simulated daily streamflow values. This first LSTM layer, termed the Residual Correcting layer (ILDAS-rLSTM), was trained using the same meteorological forcings as ILDAS, along with catchment attributes, with the residual error as the target variable. The ILDAS streamflow was then corrected using the predicted residual values from the ILDAS-rLSTM. The ILDAS-rLSTM model extracts new information from the training data that the process-based ILDAS might have missed due to its simplified process representations and inherent limitations. By learning from the discrepancies between observed and simulated streamflow, ILDAS-rLSTM identifies patterns and relationships in the data that are not captured by ILDAS. This allows ILDAS-rLSTM to account for complex, non-linear interactions and localized phenomena that ILDAS may not adequately represent, thereby refining the streamflow predictions.

Subsequently, the second LSTM layer, known as the Meta Learning layer (ILDAS-mLSTM), was trained on a 3-day historical moving average of the observed streamflow and the residual-corrected streamflow outputs, with observed streamflow as the target variable. To prevent look-ahead bias and ensure sequential training, the training periods for the ILDAS-rLSTM and ILDAS-mLSTM were kept distinct. While the ILDAS-rLSTM focused on minimizing the absolute bias in ILDAS streamflow outputs, the ILDAS-mLSTM enhanced the relevance of the streamflow predictions by integrating past observed data and applying optimal weights to the six corrected outputs.

The ILDAS-mLSTM model improves the deterministic accuracy of the streamflow predictions beyond what can be achieved with a simple equally weighted ensemble mean. By leveraging historical observed information, the ILDAS-mLSTM captures temporal dependencies and fine-tunes the streamflow outputs to reflect local conditions more accurately. This data integration and sophisticated model weighing enable the ILDAS-mLSTM to produce more precise and contextually relevant streamflow forecasts, thereby improving the overall reliability of the hybrid post-processor. The results were compared at each gauge point using two benchmark models: an ILDAS ensemble mean output (ILDAS) and a purely independent LSTM streamflow model trained on 220 catchments using meteorological forcing and catchment attributes (LSTM). These comparisons were made to assess the performance improvements brought by our hybrid post-processing framework. Table 2 summarizes the various model configurations applied in the study.

Table 1
Hyperparameters and Their Values Tested During Hyperparameter Tuning

	Hyper parameter	Values
1	Dropout Rate	[0,0.25,**0.4**,0.5]
2	Hidden Layer Size	[**64**,96,128,160,192,224,256]
3	Sequence Length	[30,90,180,270,**365**]
4	Batch Size	[64,128,256,**512**,1024]

Note. The bold values represent the final selection for training the models.

In ILDAS-rLSTM, we solely focused on reducing the residual error (magnitude) and hence, the mean squared error (MSE) was selected as loss function as it penalizes the large errors because of the squared term. For ILDAS-mLSTM, we used normalized mean absolute error (NMAE) obtained by dividing MAE by standard deviation of the observed data:

$$\text{nmae} = \left(\frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{n} \right) \div \sigma_{\text{obs}}$$

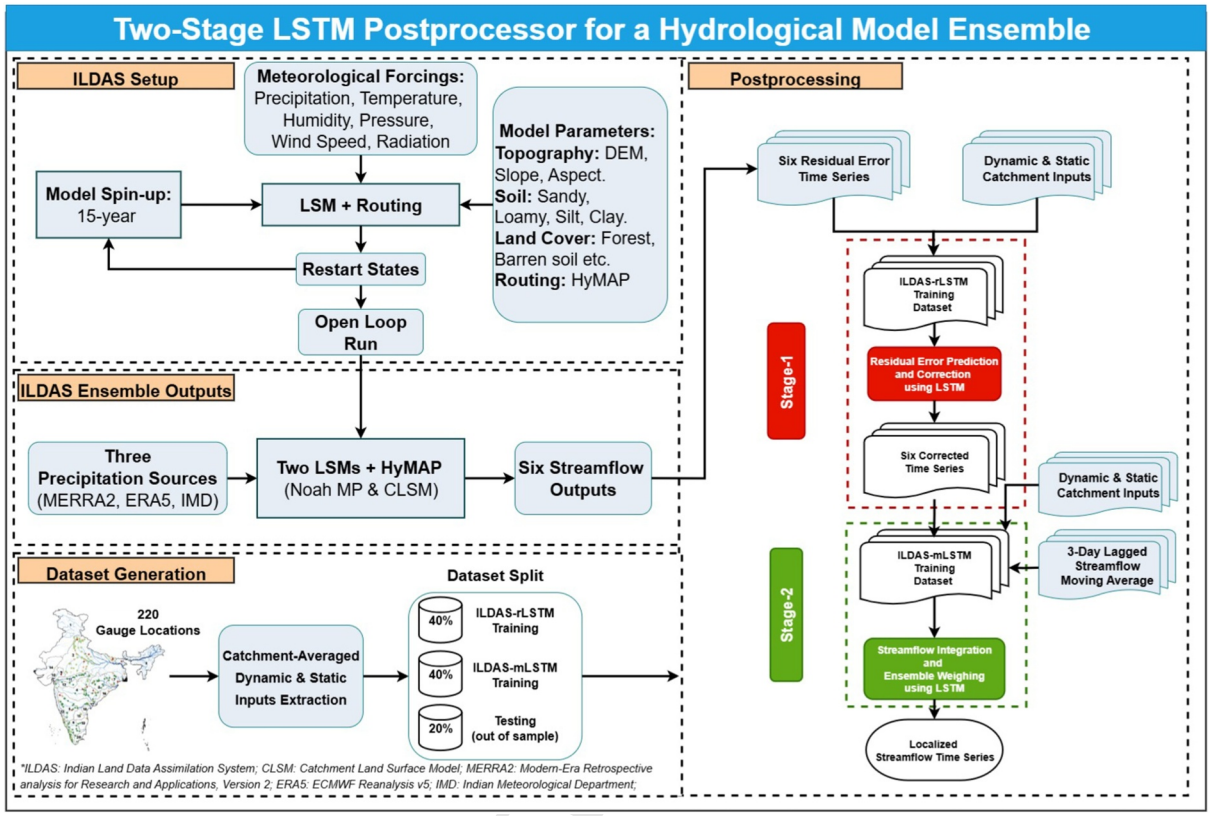

Figure 3. An illustration of the methodology adopted in this study.

where,

y_i = observed value
$\hat{y}_i$ = simulated value
n = number of observations
σ_{obs} = standard deviation of observed timeseries

Normalizing the MAE ensures that the model treats the error terms equally for wet and dry catchments, hence ensuring a more uniform spatial learning. The ADAM optimizer (Kingma & Ba, 2017) employed for model training. The ILDAS-rLSTM had 6 different variants that were trained to predict residual error in streamflow output from two LSMs and three precipitation sources. The ILDAS-mLSTM was trained as a single model to predict observed daily streamflow. Each LSTM was trained for 30 epochs and had 10 members with random initial weights, resulting in a total of 70 models. To prevent overfitting and improve robustness, we employed a dropout rate of 0.4, which means that 40% of the weights were randomly set to zero during training. The hidden layer size was set to 64, with a batch size of 512 and a sequence length of 365 days. The model performance was evaluated by comparing the ensemble mean of the predicted time series during the testing period against the observed time series for the same period. The "NeuralHydrology" Python package (Kratzert et al., 2022) was used for training the models. The testing was performed using customized python scripts. All experiments were conducted on a desktop workstation equipped with an RTX 4090 GPU with 24GB of VRAM and Intel Core i9 13900KF CPU.

2.6. Evaluation Criteria

The results were evaluated using Kling Gupta Efficiency (KGE; Gupta et al., 2009) as the primary metric, with its three components, namely, correlation coefficient (r), variability ratio(α) and bias(β). The calculation of KGE is expressed as:

Table 2
Details of Various Model Configurations Used in the Experimental Setup

S. No.	Model name	Specifications	Outputs
1	ILDAS	**Model:** NoahMP 3.6 and CLSM Fortuna v2.5 **Inputs:** MERRA-2, CHIRPS, IMD* **Simulation:** Indian subcontinent, 0.1°, Daily, 1980-2023	One Six-member ensemble of multiple land surface variables including daily streamflow
2	LSTM	**Model:** Physically Informed LSTM **Inputs:** IMD* and IMDAA* Meteorological forcing, static catchment-averaged geophysical attributes **Simulation:** Trained on 20+ years of daily streamflow (80:20 split), 220 gauge locations	One 10-member ensemble of daily streamflow timeseries at each gauge location
3	ILDAS-rLSTM	**Model:** Six LSTM models trained to correct residual error individually in each of the six-member ILDAS ensemble **Inputs:** Six individual ILDAS streamflow members, ensemble member specific meteorological forcing, static catchment-averaged geophysical attributes **Simulation:** Trained on 40% of daily streamflow data, 220 gauge locations	Six 10-member ensembles of corrected ILDAS streamflow timeseries at each gauge location
4	ILDAS-mLSTM	**Model:** One LSTM model trained to optimally combine six corrected ILDAS ensemble members from ILDAS-rLSTM with observed streamflow **Inputs:** Six residual-corrected streamflow series, 3-day lagged observed streamflow moving average, static catchment-averaged geophysical attributes **Simulation:** Sequentially trained after ILDAS-rLSTM on the next 40% of daily streamflow data, 220 gauge locations	One 10-member ensemble of daily streamflow timeseries at each gauge location

*IMD: Indian Meteorological Department; *IMDAA: Indian Monsoon Data Assimilation and Analysis.

$$\mathrm{KGE} = 1 - \sqrt{S_r[r-1]^2 + S_\alpha[\alpha-1]^2 + S_\beta[\beta-1]^2}$$

where:

$$r = \frac{\sum_{i=1}^{n}\left(Q_{\mathrm{obs},i} - \overline{Q_{\mathrm{obs}}}\right)\left(Q_{\mathrm{sim},i} - \overline{Q_{\mathrm{sim}}}\right)}{\sqrt{\sum_{i=1}^{n}\left(Q_{\mathrm{obs},i} - \overline{Q_{\mathrm{obs}}}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Q_{\mathrm{sim},i} - \overline{Q_{\mathrm{sim}}}\right)^2}}$$

$$\alpha = \frac{\sigma_{\mathrm{sim}}}{\sigma_{\mathrm{obs}}}$$

$$\beta = \frac{\overline{Q_{\mathrm{sim}}}}{\overline{Q_{\mathrm{obs}}}}$$

where:

$Q_{\mathrm{obs},i}$ = Observed flow at timestep i
$Q_{\mathrm{sim},i}$ = Simulated flow at timestep i
σ_{obs} = standard deviation of observed data
σ_{sim} = standard deviation of simulated data
$\overline{Q_{\mathrm{obs}}}$ = mean of observed data
$\overline{Q_{\mathrm{sim}}}$ = mean of observed data

S_r, S_α, and S_β are scaling factors for the three components respectively, that can be specified by the user. We also decomposed KGE into its three components to assess the correlation, variability, and volumetric bias of various

models. The three components of the KGE provide insight into different aspects of a model's performance. Correlation (Pearson-r) reflects the agreement in timing between simulated and observed values; the variability ratio (α) captures the statistical variability, and bias (β) highlights any systematic bias. A KGE score of 1 represents perfect alignment between the simulated and observed values, while a score farther from 1 indicates worse performance. Depending on the study's goals, the scaling factors can be adjusted to prioritize one or more components of the KGE (Mizukami et al., 2019). In this study, we aimed for a balanced assessment of performance, so we set all three scaling factors to 1.0.

We used Kling Gupta Efficiency Skill Score (KGESS, Hirpa et al., 2018) to provide an objective assessment of progressive gain in skill for various model configurations. KGESS is calculated as:

$$\text{KGESS} = \frac{\text{KGE}_{\text{perf}} - \text{KGE}_{\text{base}}}{1 - \text{KGE}_{\text{base}}}$$

where KGE_{perf} and KGE_{base} are the catchment-wise KGEs for new and base LSTM models, respectively. In the denominator, the KGE_{base} is subtracted from the highest possible KGE value, that is, 1. We also calculated peak timing error and peak bias error that explain the time difference and the magnitude difference between observed and simulated peak flows, respectively.

We also calculated residuals error in different seasons and the Interquartile Range (IQR) of the residuals. Residual errors represent the difference between observed and simulated streamflow values. While the mean or median of residuals can indicate overall bias, the IQR offers a non-parametric measure of the spread or dispersion of these errors, indicating the consistency and precision of the model's predictions. A smaller IQR signifies that the majority of the model's errors are clustered closely around the median, indicating more consistent and reliable performance across a range of flow conditions.

Peak flow events are critical for water resource management, especially in flood forecasting. The modeling of extreme events is critical in an operational forecasting framework. At each gauge location, peak flow events were identified using the python package Scipy's "find peaks" algorithm, where the observed time series was analyzed for all peaks. The algorithm identifies peaks above a magnitude cutoff which was kept at 90th percentile of the observed data. Peaks with a prominence less than the standard deviation of the observed time series were discarded. Additionally, peaks occurring within 100 steps (days) of each other were removed, retaining only the most significant ones. For the simulated data, corresponding peaks were identified within a 3-day window of the observed peaks. The peak timing error (in days) and the peak mean absolute percentage error (MAPE) were calculated for each gauge location.

$$\text{MAPE} = \sum_{i=1}^{n} \left| \frac{\text{Obs}_i - \text{Sim}_i}{\text{Obs}_i} \right| \times 100$$

We then counted the number of stations where peak timing improved (i.e., a reduction in peak timing error) and where peak MAPE decreased. Additionally, we employed a 2 × 2 contingency matrix to calculate one of four outcomes:

1. Hits (H): The event was observed and predicted.
2. False Alarms (FA): The event was predicted but not observed.
3. Misses (M): The event was observed but not predicted.
4. Correct Negatives (CN): The event was neither observed nor predicted.

We calculated three key metrics to evaluate the overall performance of the models in capturing extreme flow events such as floods:

$$\text{Critical Success Index (CSI)} = \frac{H}{H + \text{FA} + M}$$

$$\text{False Alarm Ratio (FAR)} = \frac{\text{FA}}{H + \text{FA}}$$

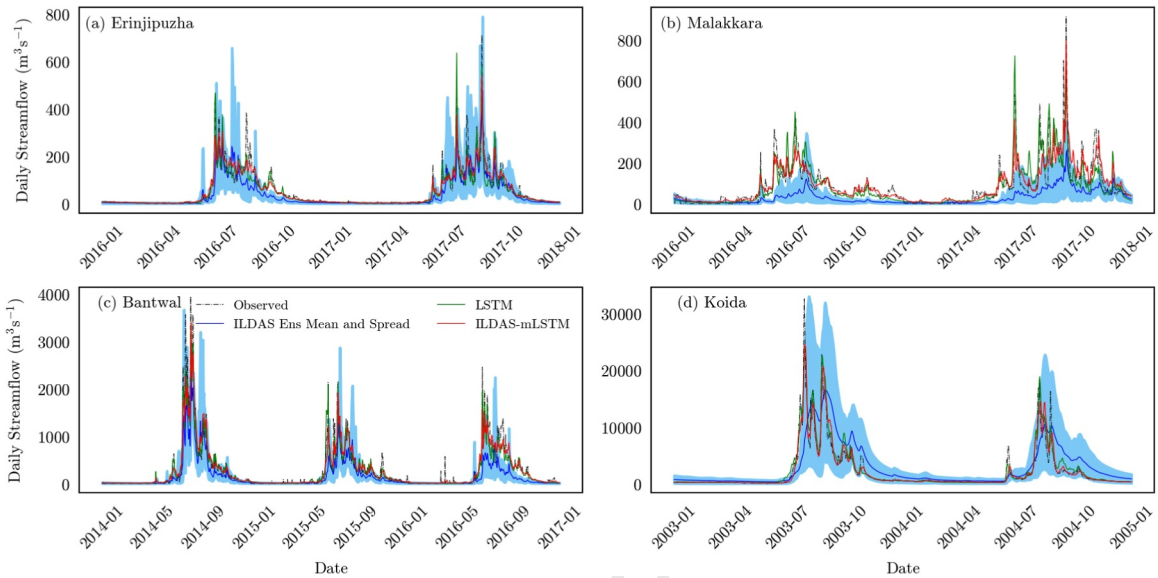

Figure 4. A hydrograph showing the agreement between ILDAS ensemble, LSTM post-processed, and observed daily streamflow at four different gauge locations, showcasing varying streamflow regimes (low, medium, and high flows).

$$\text{Probability of Detection (POD)} = \frac{H}{H + M}$$

3. Results and Discussion

3.1. Performance of Hybrid LSTM Models

Figure 4 illustrates the daily streamflow hydrographs at four different gauge locations, showcasing varying streamflow regimes (low, medium, and high flows). We compared observed values with those simulated by ILDAS, LSTM, and ILDAS-mLSTM. Notably, the ILDAS streamflow significantly deviates from the observed data, both in timing and magnitude. In contrast, the post-processed streamflow from ILDAS-mLSTM demonstrates remarkably improved performance, closely matching the patterns of observed streamflow. The LSTM model also performs better than ILDAS, even matching ILDAS-mLSTM's performance in several segments of the hydrographs. To further evaluate performance and relative improvements across 220 gauge locations, we've calculated various metrics in the sections that follow.

Figure 5 compares the daily streamflow residual error across all basins for four models, analyzed by season in the Indian subcontinent. The temporal variation in residual error magnitude aligns with the region's precipitation patterns throughout the year. The monsoon season (June–September) exhibits the largest residual values, which persist into the post-monsoon period (October–November). Notably, the uncalibrated ILDAS model shows the highest interquartile range (IQR) for residuals at 187.94 m^3/s, while ILDAS-mLSTM significantly reduces the residual IQR to 63.23 m^3/s, representing a 66.3% decrease. In the post-monsoon season, ILDAS-mLSTM further reduces the IQR by 81.7%. These results demonstrate the superior ability of the ILDAS-mLSTM model to capture seasonal variability in streamflow residuals, particularly during periods of high precipitation. This is because ILDAS, in its current state, does not explicitly incorporate information regarding reservoir characteristics, operating rules, or other human interventions that influence downstream flow. In contrast, the post-processed streamflow from ILDAS-mLSTM demonstrated significantly improved performance, notably reducing the Interquartile Range (IQR) of residuals during these regulated periods. This improvement is attributed to the learning capabilities of the LSTM. By training on recent observed streamflow data, the LSTM learns the complex, non-linear relationships that govern the actual observed streamflow, including the effects of upstream streamflow management. Therefore, while not explicitly including reservoir operations, ILDAS-mLSTM effectively adjusts its simulated flow to align with the managed observed flow, providing a more accurate representation in regulated basins. However, during the winter (December–February) and pre-monsoon (March–May) seasons, the LSTM

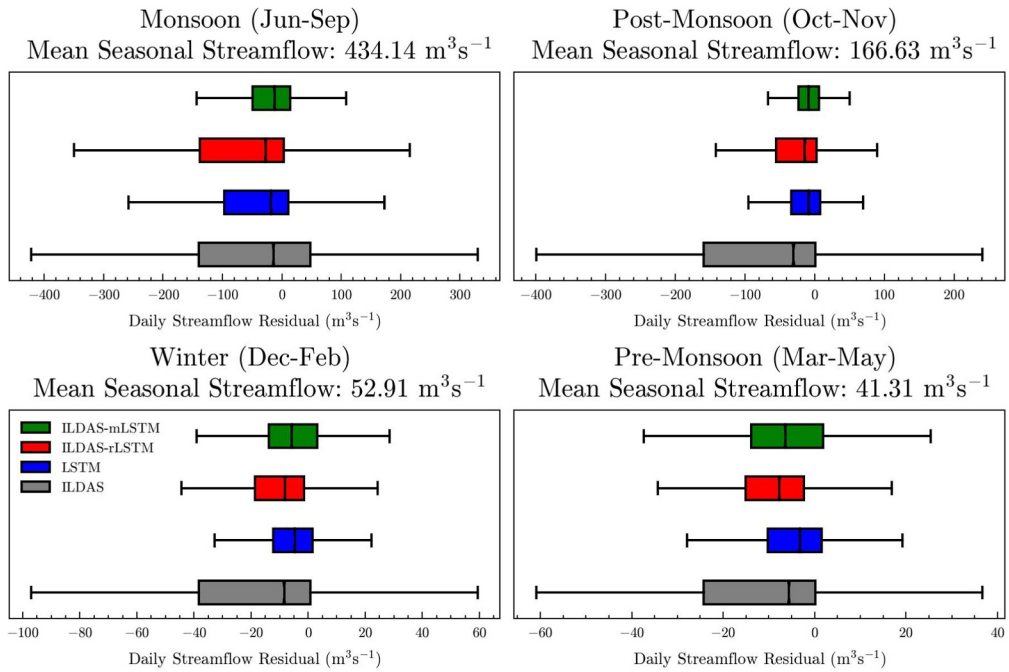

Figure 5. Comparison of daily streamflow residual for different seasons in each model. The box plots correspond to each of the four models. The seasons correspond to the Indian subcontinent climate as per following: Monsoon (June, July, August, September), Post-Monsoon (October, November), Winter (December, January, February), Pre-Monsoon (March, April, May).

slightly outperforms ILDAS-mLSTM, while ILDAS remains the poorest performing model. This can be attributed to the nature of the ILDAS-mLSTM framework, which effectively performs a kind of autoregressive modeling by utilizing observed streamflow moving average in addition to ensemble weighing to generate streamflow predictions. This approach proves most significant and beneficial during periods of high variability, such as the monsoon and post-monsoon seasons. Conversely, in the typically drier winter and pre-monsoon seasons with less flow variability, LSTM is able to perform slightly better.

Figure 6 presents the cumulative distributions of various error metrics for each model configuration across 220 gauge locations. The evaluations were conducted using testing data, which varies by gauge station and generally includes the last two to 3 years of available data at each site. Among all models, ILDAS consistently performs the worst across all metrics, showing low correlation and high bias in daily streamflow predictions, which is expected due to it being uncalibrated. The nationwide median KGE for ILDAS is 0.18, with decomposed metrics yielding values of 0.61 for correlation, 1.061 for the variability ratio, and 1.45 for bias. The LSTM model, trained without ILDAS inputs, achieves a high nationwide median KGE of 0.59, driven by a strong correlation of 0.83. Its variability ratio and bias are 0.98 and 1.14, respectively. While the ILDAS-rLSTM shows notable improvement over ILDAS with a KGE of 0.43, it still underperforms compared to LSTM. This suggests that the potential advantages of deep learning-based residual error prediction and correction within an LDAS may be limited, and purely data-driven models could outperform simplistic hybrid approaches such as ILDAS-rLSTM in the same catchments. The ILDAS-mLSTM model, which is trained using the outputs of the ILDAS-rLSTM and incorporates recent observations, shows clear advantages by generating an optimally weighted deterministic streamflow time series. It achieves the highest KGE among all models at 0.60, with a significant improvement in bias correction (beta = 1.04). Although its Pearson-r (0.79) and variability ratio (0.752) are slightly lower than those of LSTM, the reduced error bar spread indicates greater consistency. This highlights the benefits of a hybrid framework that integrates physical information during training, yielding more consistent results with lower uncertainty compared to purely data-driven models.

After establishing that ILDAS-mLSTM showed the greatest improvement in predicting daily streamflow across the country, we calculated the KGESS to quantify the relative skill gain. Figure 7 illustrates the distribution of

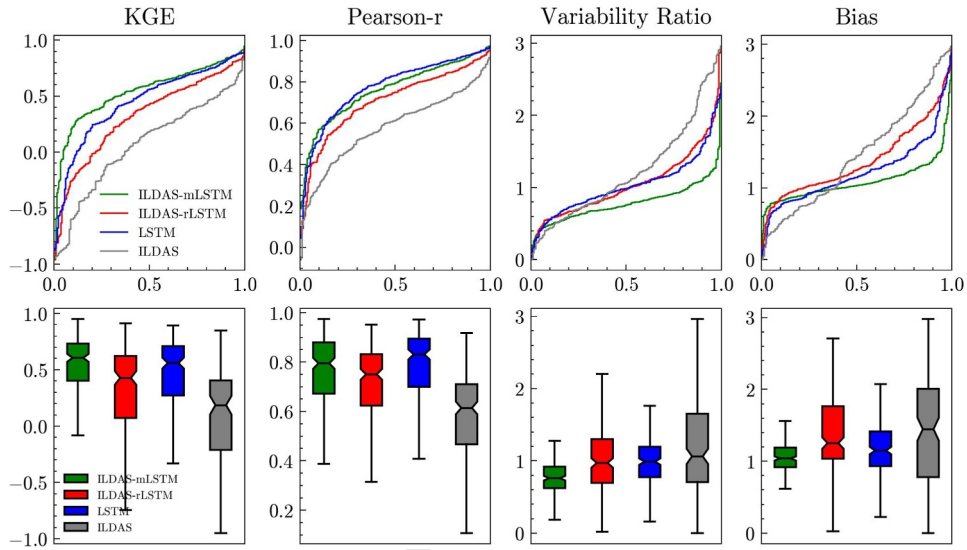

Figure 6. Comparison of nationwide error metrics across 220 gauge stations for all model configurations. The evaluation is performed using daily streamflow observations considering all seasons (annual). The evaluation period varied for each individual gauge station due to non-uniform observation records. Each column corresponds to KGE and its three components while colors represent the four models. The top row shows cumulative distribution of error metrics, and bottom row represents box plots for each metric.

KGESS for ILDAS-mLSTM compared to each model, along with the number of gauge locations where performance improved or declined. In the comparison between ILDAS-mLSTM and ILDAS, KGE improved in 208 out of 220 basins, with the enhancement observed uniformly across all regions of India. When compared to LSTM, ILDAS-mLSTM improved KGE in 136 basins, primarily in central India, where catchments are characterized by a high number of man-made structures, such as reservoirs. This suggests that incorporating localized observations as a 3-day lagged moving average allowed the model to better capture streamflow dynamics and

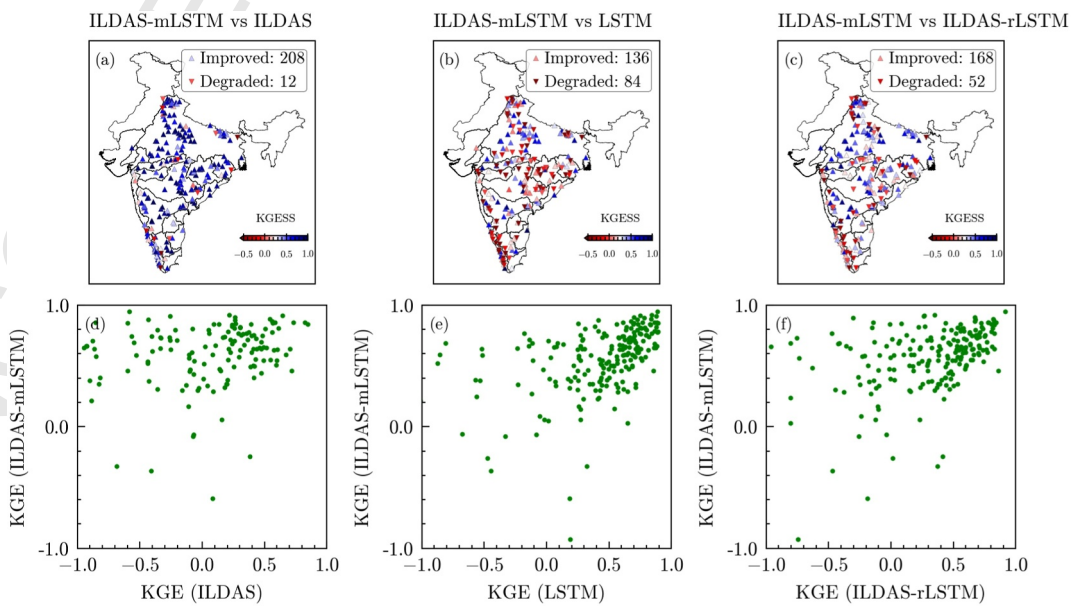

Figure 7. Distribution of KGESS by comparing KGE for various models (ILDAS, LSTM and ILDAS-rLSTM) with ILDAS-mLSTM. ILDAS-mLSTM is the best performing model, and hence, the KGESS calculated to understand the scale of improvement it made over the other models. The inset value denotes the number of gauges where KGESS is positive (improved KGE) or negative (degraded KGE).

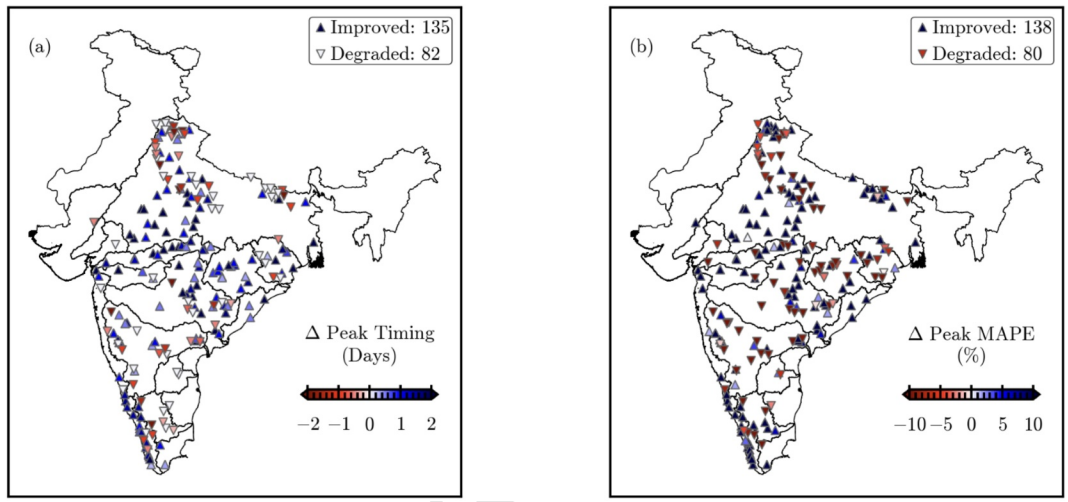

Figure 8. Comparison of peak timing and peak magnitude errors in ILDAS versus ILDAS-mLSTM. The improvement at each location is highlighted by an upward triangle and is color coded by the magnitude of the improvement.

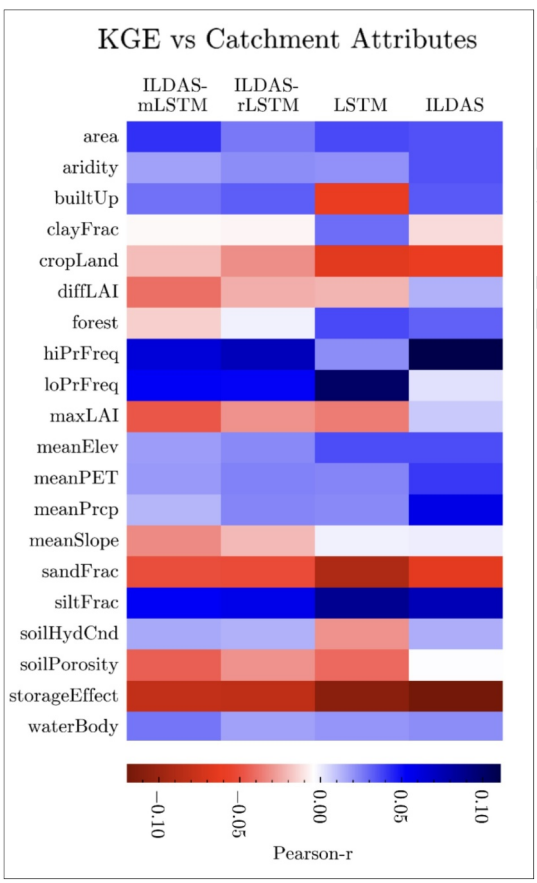

Figure 9. Correlations between KGE of various models and static catchment attributes. We calculated Pearson correlation to understand the role of various catchment properties in the model's overall performance. The positive (blue) value of Peason-r signifies an increase in the model performance with a higher value of the catchment property and vice versa.

reduce prediction bias. A similar pattern emerged in the comparison between ILDAS-mLSTM and ILDAS-rLSTM, where KGE improved at 168 stations, further reinforcing the benefit of integrating local observations and dynamically weighting ensemble outputs instead of simply averaging them.

3.2. Peak Flow Analysis

Table 4 presents a comprehensive evaluation of the models' performance in simulating peak flow events leading to flood conditions. The metrics calculated include Peak Mean Absolute Percentage Error (Peak MAPE), Peak Timing Error, Critical Success Index (CSI), False Alarm Ratio (FAR), and Probability of Detection (POD). Among all models, ILDAS-mLSTM demonstrates the most favorable performance in terms of Peak MAPE, achieving the lowest value at 44%. This indicates its superior accuracy in predicting the magnitude of flood peaks. Furthermore, ILDAS-mLSTM exhibits the lowest FAR at 0.09, signifying that it produces the least false flood predictions. While LSTM shows the highest POD at 0.87, followed closely by ILDAS-rLSTM (0.88), ILDAS-mLSTM also achieves a strong POD of 0.82, indicating its robust ability to identify actual flood events. Regarding peak timing, LSTM has the lowest timing error (0.73) followed by ILDAS-mLSTM (0.80). In contrast, the standalone ILDAS model consistently performs poorly, exhibiting the highest Peak MAPE (65%) and the lowest CSI (0.63). These results collectively highlight the significant improvements offered by the hybrid framework in simulating the magnitude and occurrence of extreme high flow events.

Figure 8 shows the distribution of improvement in peak timing and peak MAPE for ILDAS-mLSTM versus ILDAS. Our results showed that ILDAS-mLSTM improved peak timing at 135 gauge locations and reduced peak MAPE at 138 locations. As with the KGESS distribution, the greatest improvements were observed in the catchments of central India, as well as in the gauges along the west-flowing rivers of Kerala. The improvement in peak flow simulation signifies the potential of hybrid post-processing solutions for operational flood forecasting.

Table 3
A Brief Overview of Performance for All Models

S. No.	Model	Median KGE	Median KGE skill score	Percent stations where KGESS > 0	Daily streamflow residual error IQR (m^3s^{-1}) Monsoon (June–September) 431.14 m^3/s	Post-monsoon (October–November) 166.63 m^3/s	Winter (December–February) 52.91 m^3/s	Pre-monsoon (March–May) 41.31 m^3/s
1	ILDAS	0.18	-	-	187.94	160.07	39.09	24.41
2	LSTM	0.56	0.66	90.45%	107.81	41.33	**13.76**	**11.77**
3	ILDAS-rLSTM	0.42	0.56	91.36%	141.27	57.94	17.18	12.77
4	ILDAS-mLSTM	**0.60**	**0.74**	**94.54%**	**63.23**	**29.27**	16.93	15.71

Note. The highest scores are highlighted in bold. The months mentioned below the season name represent the season duration, followed by the mean seasonal daily streamflow value in m^3/s. Uncalibrated ILDAS is the worst performing model in terms of nationwide KGE and residual error IQR (inter quartile range) for all seasons. The LSTM-based post-processing model (ILDAS-mLSTM) outperforms other models by improving KGE in 94.54% locations and reducing the residual error IQR in monsoon and post-monsoon season. The base LSTM model slightly outperforms ILDAS-mLSTM in residual error IQR in two out of the four seasons.

Furthermore, we evaluated the correlation between static catchment attributes and the KGE of each model (Figure 9). For ILDAS, performance improves with larger, wetter catchments, but degrades in catchments with increased storage effects, agricultural activity, and dryness. In contrast, LSTM shows poorer performance in areas with higher built-up area, sandy soils, and reservoirs. Both ILDAS-rLSTM and ILDAS-mLSTM slightly improve performance in these challenging conditions, though storage effects remain negatively correlated with KGE for all models. This suggests that reservoir-altered flows require more advanced modeling approaches. Additionally, attributes such as mean elevation, mean precipitation, max LAI, and diffLAI (see Annexure 1) show weaker correlations in deep learning models compared to ILDAS, indicating that ILDAS captures these signals more effectively.

3.3. Discussion

In this study, we introduced a novel two-stage post-processor by disaggregating the problem in two parts: first, it specifically addresses and corrects residuals within each individual ensemble member using deep learning, rather than globally. Subsequently, the second stage integrates these corrected ensemble members with observations using deep learning instead of conventional methods such as averaging. This enables a more localized and precise streamflow output, as the model can learn to prioritize reliable ensemble members, adapt to varying hydrological conditions, and capture dependencies specific to different locations. Moreover, due to its modular nature, this framework can also be applied to the post-processing of other hydrological variables beyond streamflow.

We assessed multiple LSTM configurations to improve the performance of ILDAS on a national scale. Initially, LSTMs were trained to predict residuals within each daily streamflow ensemble member. This was followed by a second LSTM model, designed to combine the residual-corrected streamflow ensemble members with the observed streamflow moving average into a deterministic streamflow time series. The results were compared against uncalibrated ILDAS outputs and a standalone LSTM model. Overall, the meta-learning stage (ILDAS-mLSTM) exhibited superior performance across overall assessment and during peak flow events. The ILDAS-rLSTM did not outperform the LSTM standalone, a finding which is consistent with our earlier study where a LSTM trained on catchment attributes, meteorological and physical model outputs did not outperform the LSTM trained just on the catchment attributes and meteorological data (Magotra et al., 2025). The ILDAS-mLSTM showed the highest median KGE and median KGESS for 220 gauge locations (Table 3). The performance improvements exhibited by the hybrid framework are consistent with the earlier studies (Frame et al., 2021; Hunt et al., 2022; Liu et al., 2024; Tang et al., 2023). The ILDAS-mLSTM model demonstrates a superior ability to capture seasonal variability in streamflow residuals, particularly during high precipitation seasons (Figure 5). During the monsoon, ILDAS-mLSTM reduces the residual IQR from ILDAS's

Table 4
An Overview of Peak Event Analysis Showing Various Metrics Such as Peak MAPE, Peak Timing Error, Critical Success Index (CSI), False Alarm Ratio (FAR), and Probability of Detection (POD)

Model	Peak MAPE	Peak timing error	CSI	FAR	POD
ILDAS	65%	0.90	0.63	0.11	0.77
LSTM	47%	0.73	0.73	0.14	0.87
ILDAS-rLSTM	54%	0.80	0.71	0.18	0.88
ILDAS-mLSTM	44%	0.80	0.72	0.09	0.82

Note. The ILDAS-mLSTM exhibits lowest Peak MAPE and FAR while LSTM has slightly better peak timing and CSI.

187.94 m^3/s to just 63.23 m^3/s (Table 3). This improvement further extends into the post-monsoon season, with an 81.7% reduction in IQR. Conversely, during the drier winter (December–February) and pre-monsoon (March–May) seasons, the standalone LSTM model slightly outperforms ILDAS-mLSTM, while ILDAS remains the poorest performer. This can be attributed to the nature of the ILDAS-mLSTM framework, which utilizes an observed streamflow moving average for autoregressive correction, resulting in higher performance gains during periods of high flow variability such as monsoons. In drier periods, the simpler LSTM standalone method may suffice. ILDAS-mLSTM demonstrates the most favorable performance in predicting flood peak magnitudes, achieving the lowest Peak Mean Absolute Percentage Error (MAPE) at 44%. It also produces the fewest false flood predictions, with the lowest False Alarm Ratio (FAR) at 0.09. While LSTM shows a slightly higher Probability of Detection (POD) at 0.87 and lower peak timing error (0.73), ILDAS-mLSTM remains favorable with a strong POD of 0.82 and a peak timing error of 0.80. In contrast, the ILDAS model consistently performs poorly in peak events analysis, exhibiting the highest Peak MAPE (65%) and the lowest Critical Success Index (CSI) (0.63). These results underscore the significant improvements offered by the hybrid framework in accurately simulating both the magnitude and occurrence of extreme high-flow events, which are critical for flood forecasting and water resource management.

One of the limitations of this study is that our evaluation exclusively relies on a temporal split for testing the model's performance. Given that our approach incorporates a moving average of observed streamflow, a necessary component for the meta-learning stage, the postprocessor can be applied to the catchments where observations are available. Future enhancements to the modeling framework will focus on increasing both generalizability and applicability to ungauged regions. One key improvement involves substituting ground observations with bias-corrected satellite-based discharge data at ungauged locations (Andreadis et al., 2025; Riggs et al., 2023). This would enable localized streamflow predictions in areas lacking traditional in situ measurements, aligning the framework with modern remote-sensing hydrological monitoring capabilities. Alternatively, the ILDAS-mLSTM model could be trained without incorporating the observed streamflow series, yielding just an optimally weighted ensemble output that is independent of local ground truth data integration. Since this output is not based on site-specific observations, it provides a transferable ensemble prediction suitable for use in ungauged basins. Another potential limitation of the current implementation is the use of the 3-day simple moving average instead of the observed values directly in the ILDAS-mLSTM. The raw observation data exhibited some unrealistic and erroneous spikes (e.g., sudden, non-hydrological shifts from low to extremely high values, and back to low next day), which were interpreted as measurement errors. A simple moving average was employed to effectively mitigate the undue influence of these localized, high-magnitude errors but it could lead to removal of important signals that LSTM could otherwise learn from. Therefore, more sophisticated imputation techniques could be used in future research to enhance data quality while preserving the necessary hydrological signal. Additionally, our analysis was restricted to 220 specific gauge locations due to the practical limitations imposed by the availability of reliable observed streamflow data. Moreover, the test periods varied across individual basins due to the non-uniform availability of observational data at each station. This constraint means our findings are specifically validated only at these discrete catchments. Future research could significantly benefit from extending the number of gauge locations for training and analysis of data. A critical aspect of applying deep learning models like LSTMs, is understanding model interpretability. While our study successfully demonstrates the enhanced predictive capabilities of the hybrid deep learning framework for daily streamflow, we acknowledge that a comprehensive interpretation of the internal mechanisms of these complex models must be explored. For instance, it is important to analyze the precise contribution and interaction of various input features within the LSTM framework, beyond simply assessing the role of individual catchment attributes in overall model performance. We recognize this as a limitation and emphasize the need for future research to delve deeper into the interpretability of such deep learning models in hydrological forecasting.

4. Conclusions

In this study, we aimed to produce locally relevant streamflow from a process-based land surface model ensemble using a hybrid deep learning framework. The framework worked in two stages: (a) residual error correction in daily streamflow, and (b) auto-regression and optimal selection of final output from residual corrected streamflow and recent observations. Given the difficulties in calibrating large scale LSMs, we presented an alternative approach to improve the streamflow predictions in ILDAS by leveraging observed data at a fraction of computing power that is generally required in traditional approaches. The hybrid deep learning framework consisted of two

LSTM layers, called ILDAS-rLSTM and ILDAS-mLSTM, working in tandem to process the raw streamflow from ILDAS and generating locally relevant streamflow at 220 gauge locations across the Indian subcontinent. We also evaluated the performance of our approach with a pure data-driven base LSTM model, labeled as LSTM. The ILDAS-mLSTM proved to be the most effective modeling approach, securing a superior overall evaluation performance relative to the ILDAS, LSTM, and ILDAS-rLSTM. This conclusion is supported by a comprehensive assessment, which included error metrics, specifically the Kling-Gupta Efficiency (KGE) and its components, peak Mean Absolute Percentage Error (MAPE), and peak timing error, alongside peak event analysis metrics like Probability of Detection (POD), False Alarm Ratio (FAR), and the Critical Success Index (CSI). The final outputs were spatially consistent and accurate compared to raw ILDAS streamflow. However, a few limitations of this study include the lack of uniform testing and training periods across all gauge stations, as well as the relatively short length of testing data (2–3 years). Additionally, several training variables, such as land use/land cover, vegetation type, and waterbody fraction, were treated as static during the LSTM training. Future work could explore incorporating dynamic schemes for these variables, which may further improve the predictive performance of the models. In summary, the conclusions of this study are as follows:

1. The nationwide median KGE increased from 0.18 to 0.60 for raw versus final outputs, with 94.5% gauge locations showing improvement in daily streamflow simulations.
2. The post-processed streamflow by ILDAS-mLSTM exhibited 66.3% and 81.7% lesser residual error inter quartile range (IQR) for monsoon and post-monsoon periods, effectively capturing the seasonal patterns during high precipitation periods in the region.
3. The peak flow event analysis showed timing and peak mean absolute percentage errors in post-processed streamflow reduced by approximately 25%.

This study advances hydrological science by integrating physical modeling with deep learning, demonstrating the potential of hybrid approaches to generate locally relevant insights from large-scale land data assimilation systems. These frameworks can be applied to various land surface model outputs, offering scalable and practical methods that can be seamlessly integrated into operational forecasting models.

Conflict of Interest

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

Data Availability Statement

The data sets used in this study are available from the following sources:

- Streamflow: Central Water Commission, India, and India WRIS, https://indiawris.gov.in/wris/#/timeseriesdata
- IMD precipitation and temperature: https://www.imdpune.gov.in/
- MERIT DEM: http://hydro.iis.u-tokyo.ac.jp/~yamadai/MERIT_DEM/
- NeuralHydrology: https://github.com/neuralhydrology/neuralhydrology
- MODIS LAI: https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/MOD15A2H
- SoilGrids: https://soilgrids.org/
- HiHydroSoil v2.0: https://gee-community-catalog.org/projects/hihydro_soil/
- NRSC LU/LC: https://bhuvan.nrsc.gov.in/
- ERA-5: ECMWF, https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5

Acknowledgments
This research was conducted in the HydroSense lab (https://hydrosense.iitd.ac.in/) of IIT Delhi, and the authors acknowledge the IIT Delhi High Performance Computing facility for providing computational and storage resources. Dr. Manabendra Saharia gratefully acknowledges financial support for this work through grants from Ministry of Earth Sciences/IITM Pune Monsoon Mission III (RP04574) and Ministry of Earth Sciences (RP04741). The authors gratefully acknowledge the Central Water Commission (CWC), National Water Informatics Centre (NWIC), and the Ministry of Jal Shakti (MoJS) for providing the streamflow data sets used in this study. The authors also acknowledge ISRO NRSC for providing access to 1:50000 LU/LC data set.

References

Andreadis, K. M., Coss, S. P., Durand, M., Gleason, C. J., Simmons, T. T., Tebaldi, N., et al. (2025). A first look at river discharge estimation from SWOT satellite observations. *Geophysical Research Letters*, *52*(9), e2024GL114185. https://doi.org/10.1029/2024GL114185

Bai, P., Liu, X., Yang, T., Liang, K., & Liu, C. (2016). Evaluation of streamflow simulation results of land surface models in GLDAS on the Tibetan Plateau. *Journal of Geophysical Research: Atmospheres*, *121*(20), 12180–12197. https://doi.org/10.1002/2016JD025501

Chakraborty, A., Saharia, M., Chakma, S., Kumar Pandey, D., Niranjan Kumar, K., Thakur, P. K., et al. (2024). Improved soil moisture estimation and detection of irrigation signal by incorporating SMAP soil moisture into the Indian Land Data Assimilation System (ILDAS). *Journal of Hydrology*, *638*, 131581. https://doi.org/10.1016/j.jhydrol.2024.131581

Evin, G., Thyer, M., Kavetski, D., McInerney, D., & Kuczera, G. (2014). Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. *Water Resources Research*, *50*(3), 2350–2375. https://doi.org/10.1002/2013WR014185

Frame, J. M., Kratzert, F., Raney, A., II., Rahman, M., Salas, F. R., & Nearing, G. S. (2021). Post-processing the national water model with long short-term memory networks for streamflow predictions and model diagnostics. *JAWRA Journal of the American Water Resources Association*, *57*(6), 885–905. https://doi.org/10.1111/1752-1688.12964

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., et al. (2015). The climate hazards infrared precipitation with stations—A new environmental record for monitoring extremes. *Scientific Data*, *2*(1), 150066. https://doi.org/10.1038/sdata.2015.66

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., et al. (2017). The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

Getirana, A. C. V., Boone, A., Yamazaki, D., Decharme, B., Papa, F., & Mognard, N. (2012). The Hydrological Modeling and Analysis Platform (HyMAP): Evaluation in the Amazon Basin. *Journal of Hydrometeorology*, *13*(6), 1641–1665. https://doi.org/10.1175/JHM-D-12-021.1

Han, H., & Morrison, R. R. (2022). Improved runoff forecasting performance through error predictions using a deep-learning approach. *Journal of Hydrology*, *608*, 127653. https://doi.org/10.1016/j.jhydrol.2022.127653

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., et al. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One*, *12*(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hirpa, F. A., Salamon, P., Beck, H. E., Lorini, V., Alfieri, L., Zsoter, E., & Dadson, S. J. (2018). Calibration of the Global Flood Awareness System (GloFAS) using daily streamflow data. *Journal of Hydrology*, *566*, 595–606. https://doi.org/10.1016/j.jhydrol.2018.09.052

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial.

Höge, M., Wöhling, T., & Nowak, W. (2018). A primer for model selection: The decisive role of model complexity. *Water Resources Research*, *54*(3), 1688–1715. https://doi.org/10.1002/2017WR021902

Hunt, K. M. R., Matthews, G. R., Pappenberger, F., & Prudhomme, C. (2022). Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States. *Hydrology and Earth System Sciences*, *26*(21), 5449–5472. https://doi.org/10.5194/hess-26-5449-2022

Jacobs, C. M. J., Moors, E. J., Ter Maat, H. W., Teuling, A. J., Balsamo, G., Bergaoui, K., et al. (2008). Evaluation of European Land Data Assimilation System (ELDAS) products using in situ observations. *Tellus A: Dynamic Meteorology and Oceanography*, *60*(5), 1023–1037. https://doi.org/10.1111/j.1600-0870.2008.00351.x

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. arXiv. https://doi.org/10.48550/arXiv.1412.6980

Koster, R. D., Suarez, M. J., Ducharne, A., Stieglitz, M., & Kumar, P. (2000). A catchment-based approach to modeling land surface processes in a general circulation model: 1. Model structure. *Journal of Geophysical Research*, *105*(D20), 24809–24822. https://doi.org/10.1029/2000JD900327

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., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Benchmarking a Catchment-Aware Long Short-Term MemoryNetwork (LSTM) for large-scale hydrological modeling (preprint). In *Global Hydrology/Modelling Approaches*. https://doi.org/10.5194/hess-2019-368

Largeron, C., Dumont, M., Morin, S., Boone, A., Lafaysse, M., Metref, S., et al. (2020). Toward snow cover estimation in mountainous areas using modern data assimilation methods: A review. *Frontiers in Earth Science*, *8*, 325. https://doi.org/10.3389/feart.2020.00325

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., & Dadson, S. J. (2021). Benchmarking data-driven rainfall–runoff models in Great Britain: A comparison of long short-term memory (LSTM)-based models with four lumped conceptual models. *Hydrology and Earth System Sciences*, *25*(10), 5517–5534. https://doi.org/10.5194/hess-25-5517-2021

Liu, J., Koch, J., Stisen, S., Troldborg, L., & Schneider, R. (2024). A national-scale hybrid model for enhanced streamflow estimation—Consolidating a physically based hydrological model with long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, *28*(13), 2871–2893. https://doi.org/10.5194/hess-28-2871-2024

Lohmann, D., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., et al. (2004). Streamflow and water balance intercomparisons of four land surface models in the North American Land Data Assimilation System project. *Journal of Geophysical Research*, *109*(D7), D07S91. https://doi.org/10.1029/2003JD003517

Magotra, B., Prakash, V., Saharia, M., Getirana, A., Kumar, S., Pradhan, R., et al. (2024). Towards an Indian Land Data Assimilation System (ILDAS): A coupled hydrologic-hydraulic system for water balance assessments. *Journal of Hydrology*, *629*, 130604. https://doi.org/10.1016/j.jhydrol.2023.130604

Magotra, B., Saharia, M., & Dhanya, C. T. (2025). Improved streamflow simulations in hydrologically diverse basins using physically-informed deep learning models. *Hydrological Sciences Journal*, *70*(5), 775–788. https://doi.org/10.1080/02626667.2025.2458545

Marty, R., Fortin, V., Kuswanto, H., Favre, A.-C., & Parent, E. (2015). Combining the Bayesian processor of output with Bayesian model averaging for reliable ensemble forecasting. *Journal of the Royal Statistical Society - Series C: Applied Statistics*, *64*(1), 75–92. https://doi.org/10.1111/rssc.12062

Mitchell, K. E., Lohmann, D., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., et al. (2004). The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *Journal of Geophysical Research*, *109*(D7), D07S90. https://doi.org/10.1029/2003JD003823

Mizukami, N., Clark, M. P., Newman, A. J., Wood, A. W., Gutmann, E. D., Nijssen, B., et al. (2017). Towards seamless large-domain parameter estimation for hydrologic models. *Water Resources Research*, *53*(9), 8020–8040. https://doi.org/10.1002/2017WR020401

Moreira, A. A., Ruhoff, A. L., Roberti, D. R., Souza, V. D. A., da Rocha, H. R., & Paiva, R. C. D. (2019). Assessment of terrestrial water balance using remote sensing data in South America. *Journal of Hydrology*, *575*, 131–147. https://doi.org/10.1016/j.jhydrol.2019.05.021

Myneni, R., Knyazikhin, Y., & Park, T. (2021). MODIS/terra leaf area index/FPAR 8-day L4 global 500m SIN grid V061 [Dataset]. *NASA EOSDIS Land Processes DAAC*. https://doi.org/10.5067/MODIS/MOD15A2H.061

Niu, G. Y., Yang, Z. L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., et al. (2011). The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research*, *116*(12), D12109. https://doi.org/10.1029/2010JD015139

Ossandón, Á., S, N. J., Mendoza, P. A., Rajagopalan, B., & Mishra, V. (2022). A Bayesian hierarchical framework for postprocessing daily streamflow simulations across a river network. *Journal of Hydrometeorology*, *23*(6), 947–963. https://doi.org/10.1175/JHM-D-21-0167.1

Pai, D. S., Sridhar, L., Rajeevan, M., Sreejith, O. P., Satbhai, N. S., & Mukhopadhyay, B. (2014). Development of a new high spatial resolution (0.25° × 0.25°) long period (1901-2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region (Vol. *65*, pp. 1–18). https://doi.org/10.54302/mausam.v65i1.851

Prakash, V., & Mishra, V. (2022). Soil moisture and streamflow data assimilation for streamflow prediction in the Narmada River basin. https://doi.org/10.1175/JHM-D-21-0139.1

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*(5), 1155–1174. https://doi.org/10.1175/mwr2906.1

Rani, S. I., T, A., George, J. P., Rajagopal, E. N., Renshaw, R., Maycock, A., et al. (2021). IMDAA: High resolution satellite-era reanalysis for the Indian monsoon Region. *Journal of Climate*, 1–78. https://doi.org/10.1175/JCLI-D-20-0412.1

Reichle, R. H. (2008). Data assimilation methods in the Earth sciences. *Advances in Water Resources*, *31*(11), 1411–1418. https://doi.org/10.1016/j.advwatres.2008.01.001

Reichle, R. H., & Koster, R. D. (2005). Global assimilation of satellite surface soil moisture retrievals into the NASA Catchment land surface model. *Geophysical Research Letters*, *32*(2). https://doi.org/10.1029/2004GL021700

Riggs, R. M., Allen, G. H., Wang, J., Pavelsky, T. M., Gleason, C. J., David, C. H., & Durand, M. (2023). Extending global river gauge records using satellite observations. *Environmental Research Letters*, *18*(6), 064027. https://doi.org/10.1088/1748-9326/acd407

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., et al. (2004). The global land data assimilation system. *Bulletin of the American Meteorological Society*, *85*(3), 381–394. https://doi.org/10.1175/BAMS-85-3-381

Romero-Cuellar, J., Abbruzzo, A., Adelfio, G., & Francés, F. (2019). Hydrological post-processing based on approximate Bayesian computation (ABC). *Stochastic Environmental Research and Risk Assessment*, *33*(7), 1361–1373. https://doi.org/10.1007/s00477-019-01694-y

Solomatine, D. P., & Shrestha, D. L. (2009). A novel method to estimate model uncertainty using machine learning techniques. *Water Resources Research*, *45*(12), W00B11. https://doi.org/10.1029/2008WR006839

Srivastava, A. K., Rajeevan, M., & Kshirsagar, S. R. (2009). Development of a high resolution daily gridded temperature data set (1969–2005) for the Indian region. *Atmospheric Science Letters*, *10*(4), 249–254. https://doi.org/10.1002/asl.232

Tang, S., Sun, F., Liu, W., Wang, H., Feng, Y., & Li, Z. (2023). Optimal postprocessing strategies with LSTM for global streamflow prediction in ungauged basins. *Water Resources Research*, *59*(7), e2022WR034352. https://doi.org/10.1029/2022WR034352

Troy, T. J., Wood, E. F., & Sheffield, J. (2008). An efficient calibration method for continental-scale land surface modeling. *Water Resources Research*, *44*(9), W09411. https://doi.org/10.1029/2007WR006513

Wang, H., Huo, X., Duan, Q., Liu, R., & Luo, S. (2023). Uncertainty quantification for the Noah-MP land surface model: A case study in a grassland and sandy soil region. *Journal of Geophysical Research: Atmospheres*, *128*(20), e2023JD038556. https://doi.org/10.1029/2023JD038556

Wang, W., Cui, W., Wang, X., & Chen, X. (2016). Evaluation of GLDAS-1 and GLDAS-2 forcing data and Noah model simulations over China at the monthly scale. *Journal of Hydrometeorology*, *17*(11), 2815–2833. https://doi.org/10.1175/JHM-D-15-0191.1

Weerts, A. H., Winsemius, H. C., & Verkade, J. S. (2010). Estimation of predictive hydrological uncertainty using quantile regression: Examples from the national flood forecasting system (England and Wales). https://doi.org/10.5194/hessd-7-5547-2010

Xia, Y., Hao, Z., Shi, C., Li, Y., Meng, J., Xu, T., et al. (2019). Regional and global land data assimilation systems: Innovations, challenges, and prospects. *Journal of Meteorological Research*, *33*(2), 159–189. https://doi.org/10.1007/s13351-019-8172-4

Xia, Y., Mocko, D. M., Wang, S., Pan, M., Kumar, S. V., Peters-Lidard, C. D., et al. (2018). Comprehensive evaluation of the Variable Infiltration Capacity (VIC) model in the North American land data assimilation system. *Journal of Hydrometeorology*, *19*(11), 1853–1879. https://doi.org/10.1175/JHM-D-18-0139.1

Ye, A., Duan, Q., Yuan, X., Wood, E. F., & Schaake, J. (2014). Hydrologic post-processing of MOPEX streamflow simulations. *Journal of Hydrology*, *508*, 147–156. https://doi.org/10.1016/j.jhydrol.2013.10.055

Zaitchik, B. F., Rodell, M., & Olivera, F. (2010). Evaluation of the global land data assimilation system using global river discharge data and a source-to-sink routing scheme. *Water Resources Research*, *46*(6), W06507. https://doi.org/10.1029/2009WR007811

Zhang, X., & Zhao, K. (2012). Bayesian neural networks for uncertainty analysis of hydrologic modeling: A comparison of two schemes. *Water Resources Management*, *26*(8), 2365–2382. https://doi.org/10.1007/s11269-012-0021-5

Zhao, L., Duan, Q., Schaake, J., Ye, A., & Xia, J. (2011). A hydrologic post-processor for ensemble streamflow predictions. *Advances in Geosciences*, *29*, 51–59. https://doi.org/10.5194/adgeo-29-51-2011