https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

Improving Streamflow Simulation through Machine Learning-Powered Data Integration and Its Implications for Forecasting in the Western U.S.

Yuan Yang[1], Ming Pan[1], Dapeng Feng[2], Mu Xiao[1], Taylor Dixon[1], Robert Hartman[3], Chaopeng Shen[4], Yalan Song[4], Agniv Sengupta[1], Luca Delle Monache[1], F. Martin Ralph[1]

[1] Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, University of California San Diego, CA, USA.
[2] Department of Earth System Science, Stanford University, Stanford, CA, USA
[3] Robert K. Hartman Consulting Services, Roseville, CA, USA
[4] Civil and Environmental Engineering, Pennsylvania State University, PA, USA

Correspondence to: Yuan Yang (yuanyangthu@gmail.com) and Ming Pan (m3pan@ucsd.edu)

Abstract. Accurate streamflow forecasts are crucial but remain challenging for the arid Western United States (U.S.). Recently, machine learning methods such as long short-term memory (LSTM) have exhibited high accuracy in streamflow simulation and strong abilities to integrate observations to enhance performance. This study evaluated an LSTM-based data integration approach that incorporates streamflow (Q) and snow water equivalent (SWE) observations to improve streamflow estimations across different lag times (1-10 days, 1-6 months) and timescales (daily and monthly) over hundreds of basins in the Western U.S. Integrating Q at the daily scale provided the greatest improvements, increasing the median Kling-Gupta Efficiency (KGE) of 646 basins from 0.80 to 0.96 when integrating 1-day lagged Q, and remaining at 0.89 even with a 10-day lag. Integrating Q at the monthly scale also enhanced streamflow estimations, though to a lesser extent than at the daily scale, with the median KGE rising from 0.80 to 0.86 when integrating 1-month lagged streamflow. The next most notable improvement resulted from integrating SWE at the monthly scale, where the median KGE improved to 0.86 when integrating 1-month lagged SWE. Furthermore, SWE integration showed greater benefits at the monthly scale in snow-dominated basins during snowmelt season, which was beneficial for spring-summer flow estimations. However, integrating SWE at the daily scale did not show improvements. These results highlight the potential of this LSTM-based data integration approach for both short-term and long-term streamflow forecasting due to its performance, automation and efficiency.

1 Introduction

Accurate, reliable, and easily implementable hydrological forecasts are crucial for Western United States (U.S.), a region characterized by arid conditions and high water demand (Baker et al., 2021; Fleming et al., 2021; Hunt et al., 2022; Pierce et al., 2008). Short-term forecasts aid in flood risk mitigation, while long-term forecasts facilitate water allocation, reservoir operations, hydropower generation, and drought resilience (Broxton et al., 2023; Yaseen et al., 2015). However, this region's

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

complex topography, including deserts, mountains, valleys, and coastal areas, along with its localized climate dynamics, such as atmospheric rivers, monsoons, and seasonal snowpack, pose significant challenges for accurate streamflow forecasting (Zeng et al., 2018).

Operational agencies employ various streamflow forecast practices, tailored to their specific needs and regional characteristics. The U.S. Department of Agriculture Natural Resources Conservation Service (NRCS) utilizes principal component regression (PCR), a statistical model to predict streamflow based on selected predictors (Garen, 1992; Perkins et al., 2009). The National Weather Service (NWS) River Forecast Centers (RFC) developed the Hydrological Ensemble Forecast System (HEFS), which uses the Sacramento Soil Moisture Accounting (SAC-SMA) and SNOW-17 models to generate streamflow forecasts across different timescales (Brown et al., 2014; Demargne et al., 2014). While historically successful, these techniques have become less skillful due to regional climate change and other technical limitations, necessitating potential upgrades or replacements (Fleming and Goodbody, 2019). For instance, the recently developed National Water Model is intended to serve as the basis for the future U.S. streamflow forecasting system (Cosgrove et al., 2024). Additionally, these models require extensive manual expertise for domain-specific implementation, such as subjective predictor selection, careful empirical regression identification, and labor-intensive parameter calibration (Fleming et al., 2021). Moreover, they struggle to ingest new observations to enhance streamflow forecasts without substantial structural modifications, such as recalibrating regressions or integrating data assimilation techniques (Franz et al., 2014; Gichamo and Tarboton, 2019). For example, the California-Nevada RFC (CNRFC) employs a "forecasters-in-the-loop" approach, where forecasters manually adjust predictions as new information becomes available, leveraging their prior experience to enhance forecast accuracy.

With the ever-increasing data availability and large advancements in computing technologies, machine learning (ML) models have emerged as promising alternatives to alleviate these limitations. ML models can automatically extract useful information from complex datasets and generate accurate estimation without requiring extensive knowledge of the underlying physical systems (LeCun et al., 2015; Prasad et al., 2017; Schmidhuber, 2015; Shen, 2018; Shen et al., 2023), thereby reducing the need for manual interventions. Moreover, ML models can easily absorb new datasets during training (Shen, 2018), scale efficiently to multiple catchments (Feng et al., 2020; Kratzert et al., 2018), and extrapolate proficiently to ungauged basins (Feng et al., 2021; Kratzert et al., 2019a). Therefore, a surge in applying ML models for streamflow forecasting has been observed in recent years (Fleming et al., 2021; Nearing et al., 2024). For example, the multi-model machine learning metasystem (M^4) is currently being developed as the next-generation operational forecasting system in NRCS (Fleming and Goodbody, 2019). Among the various ML models, one increasingly popular model is the Long Short-Term Memory (LSTM) network, a specifically designed version of recurrent neural network (RNN) for long-term sequential datasets (Greff et al., 2016; Hochreiter and Schmidhuber, 1997). With its unique structure of memory cells and gating mechanisms, LSTM effectively manages the flow of information over long sequences, enabling the retention of relevant input data while discarding less important information. A growing body of research has demonstrated LSTM's seemingly incomparable performance in streamflow estimation at both daily and monthly scales (Ayana et al., 2023; Cheng et al., 2020; Clark et al., 2024; Dalkilic et al., 2023; Feng et al., 2020, 2021; Frame et al., 2022; Gauch et al., 2021; Kratzert et al., 2019a; Lees et al., 2021; Nearing et al., 2024).

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

Incorporating observations is important to improve streamflow estimation, as it helps adjust model states to better represent actual hydrological conditions (Sabzipour et al., 2023). In the context of LSTM-based models, this can be achieved through methods such as data assimilation (DA) or data integration (DI, Feng et al., 2020; Song et al., 2024), the latter also referred to as "autoregression" in Nearing et al. (2022). Similar to traditional DA in hydrological models, DA in LSTM-based models computes the difference between simulations and observations, and propagates it backward into the model to update the model's internal states. This process relies on inverse procedures, such as variational optimization, and ensemble-based conditional probability estimation, which are not only computationally intensive but also highly sensitive to parameters related to error distributions, regularization coefficients, and resampling procedures (Bannister, 2017; Nearing et al., 2018; Snyder et al., 2008). In contrast, DI directly incorporates observations as inputs and lets LSTM autonomously learn how to optimally utilize this information to enhance estimation. A comparative analysis by Nearing et al. (2022) demonstrated that DI is more accurate and computationally efficient than DA, making it a preferable approach for improving LSTM-based streamflow estimation.

Several studies have demonstrated that directly integrating streamflow observations into the LSTM inputs can significantly improve daily streamflow estimation but only at one or several gauges (Khoshkalam et al., 2023; Le et al., 2019; Sabzipour et al., 2023). Feng et al. (2020), Mangukiya et al. (2023) and Nearing et al. (2022) extended this analysis to large-scale datasets, yet their findings remained constrained to the daily timescale. On the other hand, snow is the primary source of water in the Western U.S., contributing approximately 53% of the total streamflow (Li et al., 2017). Despite its critical role, few studies have investigated the impact of integrating snow observations into LSTM on streamflow estimation. One exception is Thapa et al. (2020), which showed that incorporating snow cover area as an input improved monthly streamflow estimation, though this analysis was limited to only one gauge. Furthermore, different hydrological variables exhibit varying persistence within the water cycle. Snow, for example, has a longer memory effect since it acts as a natural reservoir that stores water during winter and gradually releases water throughout the spring and summer snowmelt season. However, a gap remains in the literature regarding the comprehensive evaluation of how different observations, such as streamflow (Q) and snow water equivalent (SWE), affect streamflow estimation across multiple timescales.

Motivated by the demonstrated performance of LSTM, this study evaluated a flexible LSTM-based data integration approach that incorporates different observations (Q and SWE) to improve streamflow simulations across multiple timescales and hundreds of basins in the Western U.S. This study employed "hindcasting", meaning retrospective simulations were conducted using observed meteorological forcings, rather than weather forecasts. Given that accurate simulations form the foundation of reliable streamflow forecasting, the demonstrated performance of this data integration approach in hindcasting underscores its potential value for forecasting applications. The findings of this study provide critical insights into (1) the effectiveness of LSTM-based data integration for improving streamflow forecasting in the Western U.S. and (2) the different influence of Q and SWE observations on forecast performance across varying timescales.

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

2 Methods

2.1 Data

We selected a total of 646 basins (all dots in Fig. 1a) in the Western U.S. from the U.S. Geological Survey (USGS) Geospatial Attributes of Gages for Evaluating Streamflow II (GAGEII; Falcone, 2011; Falcone et al., 2010) database for model training. Basin selection was based on several criteria, including boundary accuracy, basin area, data length, reservoir influences, and visual inspection (Appendix A). To further investigate the effect of integrating SWE data, we identified a subset of 429 snow-dominated basins (blue dots in Fig. 1a) from the selected 646 basins (Appendix A), while the remaining basins (orange) are classified as rain-dominated.

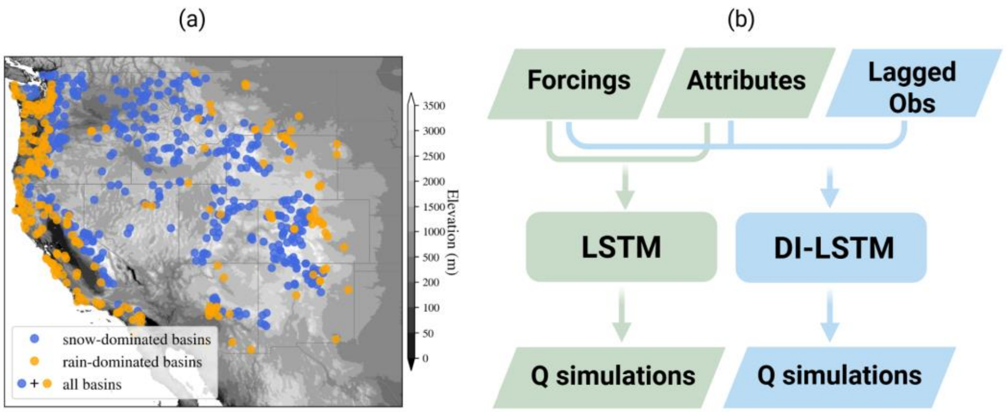

Figure 1. (a) Study basins: blue dots stand for snow-dominated basins, orange dots stand for rain-dominated basins. (b) models: LSTM vs. DI-LSTM model.

We utilized five forcing variables from CW3E 1-km 1-hourly Meteorological Forcing on NWM Grid (CW3E-Forcing, Pan, 2025) dataset and monthly leaf area index (LAI) climatology (no interannual change) from PROBA-V (Fuster et al., 2020) (Table C1). CW3E-Forcing is generated using an elevation-based downscaling and merging procedure to ingest a series of inputs from different sources with different temporal/spatial resolutions, domains, periods of coverage, and lag times. Key features of this forcing dataset include its long-term record (spanning from 1979 to the present), high resolution (1 km, 1 hour), and national-scale coverage across the conterminous United States. Here, we utilized the aggregated daily retrospective data from 1983 to 2022. Note that in this study, we performed "hindcasting" to show the effectiveness of the DI-LSTM approach, therefore, forecasted forcings were not used.

To inform LSTM about basin rainfall-runoff behaviors, we calculated the top 10 sensitive basin attributes according to Kratzert et al. (2019b), including climate, topography, and soil attributes (Table C1) as additional inputs to train the models. These attributes were static and appended to the forcing data as input for LSTM.

The daily streamflow data, used both as the training target as well as the input of streamflow integration experiments, were obtained directly from the USGS Water Information System.

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

For SWE, we used the daily 4-km gridded SWE data from the University of Arizona dataset (Broxton et al., 2016; Zeng et al., 2018). This dataset is derived through ordinary Kriging interpolation of SWE values from the Snow Telemetry (SNOTEL) sites and further enhanced by incorporating snow depth measurements from thousands of NWS Cooperative Observer Program (COOP) stations (Dawson et al., 2017).

All gridded data were spatially averaged to the basin scale from their original resolutions. All dynamic datasets were aggregated to both daily and monthly timescales to conduct experiments at these two temporal resolutions.

2.2 Modeling

Due to the great potential of LSTM in hydrological modeling, we adopted the LSTM model to investigate the effects of data integration. Additional LSTM details are in Appendix B.

Overall, we trained two types of LSTM models to assess the potential of leveraging lagged observations to improve streamflow estimation (Fig. 1b). The first type is a standard LSTM model that does not perform data integration (DI) and does not use any historical Q or SWE observations. It serves as a valuable benchmark for the comparison against DI-LSTM model and can be written concisely as:

$$Q^{1:t} = LSTM(x^{1:t}, A), \tag{1}$$

Where t is the current time step, x stands for forcings, and A represents static basin attributes.

The second type of model is DI-LSTM, which refers to the incorporation of lagged observations (y) into the model. This model can be concisely written as:

$$Q^{N+1:t} = DI - LSTM(x^{N+1:t}, A, y^{1:t-N}), \tag{2}$$

where N is the lag time step, and y is N-step lagged Q or SWE observations. In other words, we fed a N-step-lagged variable y, and let DI-LSTM decide how to use it to dynamically update both cell and hidden states, as well as the LSTM weights, thereby minimizing the accumulation of compounding errors and achieving a better estimation. The only difference between DI-LSTM model and the standard LSTM is whether lagged observations are incorporated in the inputs. Compared with the complex DA techniques used in conceptual or process-based models, this LSTM-powered DI method is relatively straightforward. Its higher computational efficiency and lower development costs make it a promising candidate for operational implementation.

2.3 Experiments

In this study, we evaluated our DI algorithm with two variables: lagged Q and SWE. Given that the effects of DI are expected to vary across different timescales, we tested the algorithm at both daily and monthly scales across all selected basins. For the daily scale, lag times ranged from 1 to 10 days, while for the monthly scale, lag times from 1 to 6 months were considered. In the following text, we used DI(Q-N) or DI(SWE-N) to denote the integration with Q or SWE from N time steps ago. Additionally, to assess whether integrating SWE has a more pronounced effect in snow-dominated basins, we conducted an

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

additional set of LSTM and DI(SWE-N) experiments specifically for the 429 snow-dominated basins. In total, 52 experiments were conducted in this study. A summary of these experiments is provided in Table 1.

Table 1: Experiments

Time Scale	Lag Time (N)	DI Observations	Training Basins	Experiment Name
Daily	1-10 days	Q	All	Daily DI(Q-N)
		SWE	All & snow-dominated (*)	Daily DI(SWE-N)
Monthly	1-6 months	Q	All	Monthly DI(Q-N)
		SWE	All & snow-dominated (*)	Monthly DI(SWE-N)

* Only used in Sect. 4.2

For each experiment, training data from all selected basins during the 1983-2002 period was used to train LSTM and DI-LSTM models, enabling the network to learn a general understanding of the rainfall-runoff process. The inputs included six meteorological features and 10 static basin attributes (Table C1). The loss function was the Root-Mean-Squared Error (RMSE). Hyperparameters, such as the number of hidden/cell states and the length of the input sequence, were determined through a simple grid search across a range of values. The final configurations (Table C2) were chosen by taking the parameter set that resulted in the best performance. A fast and flexible LSTM framework from the open-source hydroDL repository (Fang et al., 2021) was implemented.

Missing values are common in streamflow data, yet a naive LSTM cannot operate if any of its inputs are missing. To address this limitation in DI(Q) experiments, we initially trained the standard LSTM model by filling in missing data with the mean of the training period and subsequently replaced the missing lagged streamflow data with the corresponding LSTM-modeled streamflow data at the same lag time. To prevent missing target (streamflow) values from influencing the model training, for all experiments, the loss function calculation excluded simulations where the corresponding streamflow observations were missing.

To account for stochasticity in the neural network training, we performed an ensemble of six randomly seeded trainings, and the mean of all six model simulations was used for the model evaluation.

2.4 Evaluation

We evaluated the ensemble mean simulations from two types of models, LSTM and DI-LSTM, for 2003-2022, independent from the training period. The differences between the two kinds of simulations showed the effect of integrating lagged observations. Metrics adopted to evaluate model performance included the modified Kling-Gupta Efficiency (KGE, Kling et al., 2012) and its three component metrics: correlation coefficient (CC, for temporal coherence), relative variability (RV, for bias in variability), and relative bias (RB, for bias in magnitude). The equations of the four metrics are shown in Table C3. We also calculated the percent bias of the top 2% peak flow range (FHV) and the percent bias of the bottom 30% low flow range (FLV) to highlight the performance of the model for peak flows and baseflow, respectively.

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

3 Results

3.1 The effectiveness of DI(Q) at the daily scale

The daily baseline LSTM without any DI already showed a very promising simulation, with a median KGE of 0.80, a median CC of 0.92, a median RV of 0.94, and a median RB of -10.34% during the test period (Fig. 2). Better performance can be seen over more humid regions, and only 12 hyper-arid basins show negative KGE values (Fig. 3). This result, consistent with previous studies, such as Feng et al. (2024), Kratzert et al. (2019b) and Nearing et al. (2024), highlights the ability of a large-scale LSTM model to learn hydrologic behaviors across diverse basins without strong prior structural assumptions.

Overwhelming benefits were observed from integrating lagged streamflow, consistent with previous studies in CONUS (Feng et al., 2020; Nearing et al., 2022), India (Mangukiya et al., 2023) and Canada (Khoshkalam et al., 2023; Sabzipour et al., 2023). Compared to the baseline LSTM, all DI(Q) experiments exhibited significantly improved median values (p <=0.05, Kolmogorov-Smirnov test, Eghbali, 1979; Smirnov, 1948) as well as substantially reduced variability across all metrics (Fig. 2). After integrating the 1-day lagged streamflow, the median KGE, CC, RV and RB improved to 0.96, 0.98, 0.96 and 1.24%, respectively, approaching nearly perfect values. Integrating lagged daily streamflow also improved the relative bias of both low and high flows. Although the median FLV remained largely unchanged, which was already close to zero, the variability of FLV was largely reduced, indicating consistently low values across basins. The underestimation of high flows was significantly reduced, with median values shifting closer to zero and a narrower range of variability. The compaction of FHV was less pronounced than that of FLV, likely due to the shorter timescales of peak flows and their lower dependence on memory compared to low flows. Nevertheless, the benefits of DI(Q) were still noticeable with FHV, demonstrating the role of antecedent conditions in influencing flooding.

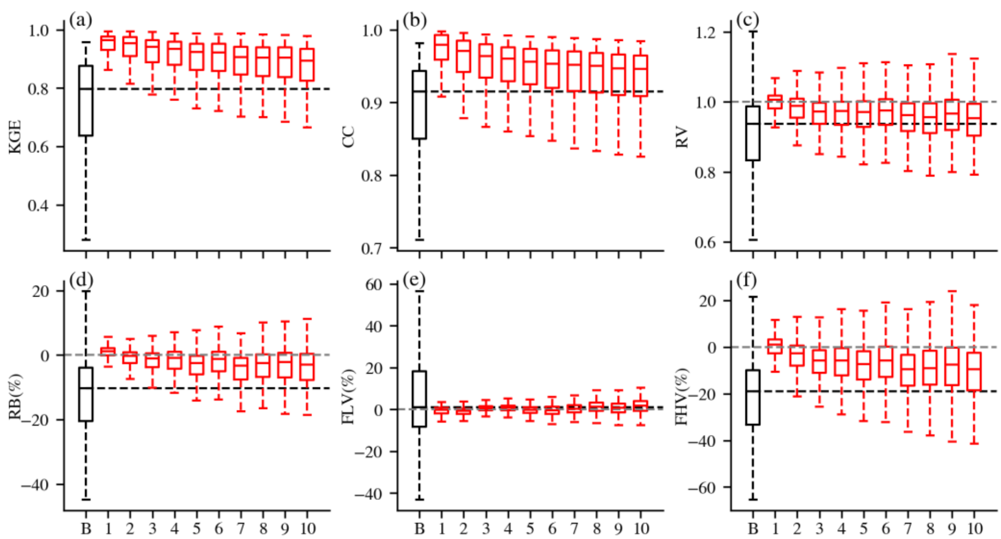

Figure 2. Performance of LSTM (black) and DI(Q-N) (N=1-10) experiments (red) at the daily scale. The "B" on the x-axis stands for baseline LSTM, and N stands for DI(Q-N) experiment. The black horizontal line stands for the median value of the baseline LSTM.

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

The grey horizontal line shows perfect value for RV, RB, FLV, and FHV. The boxplots display the median, 25th/75th percentiles, the lowest datum above Q1 - 1.5*(Q3-Q1) (lower whisker), and the highest datum below Q3 + 1.5*(Q3-Q1) (upper whisker).

Spatially, ubiquitous and heterogeneous benefits from daily DI(Q-N) can be observed over the whole Western U.S. Taking DI(Q-1) as an example, most gauges experienced a boost of 0.1~0.3 in KGE, and about 83% of basins had a KGE larger than 0.9 (Fig. 3). The largest improvements were found in the Rocky Mountains and Sierra Nevada Ranges, where KGE values were boosted from <0.6 to 0.9~1. The spatial pattern of improvements shows a positive correlation with the streamflow autocorrelation, with the strongest benefits in regions with high streamflow autocorrelation (Fig. C1). In several southern basins, utilizing lagged streamflow observations did not improve simulations. One possible explanation is that these are highly arid basins with low streamflow autocorrelation and flash floods (Li et al., 2022; Mangukiya & Sharma, 2025; Saharia et al., 2017). The sudden sharp streamflow peaks in these basins typically persist for less than one day and have little relationship with the previous day's streamflow, limiting the effectiveness of lagged streamflow observations.

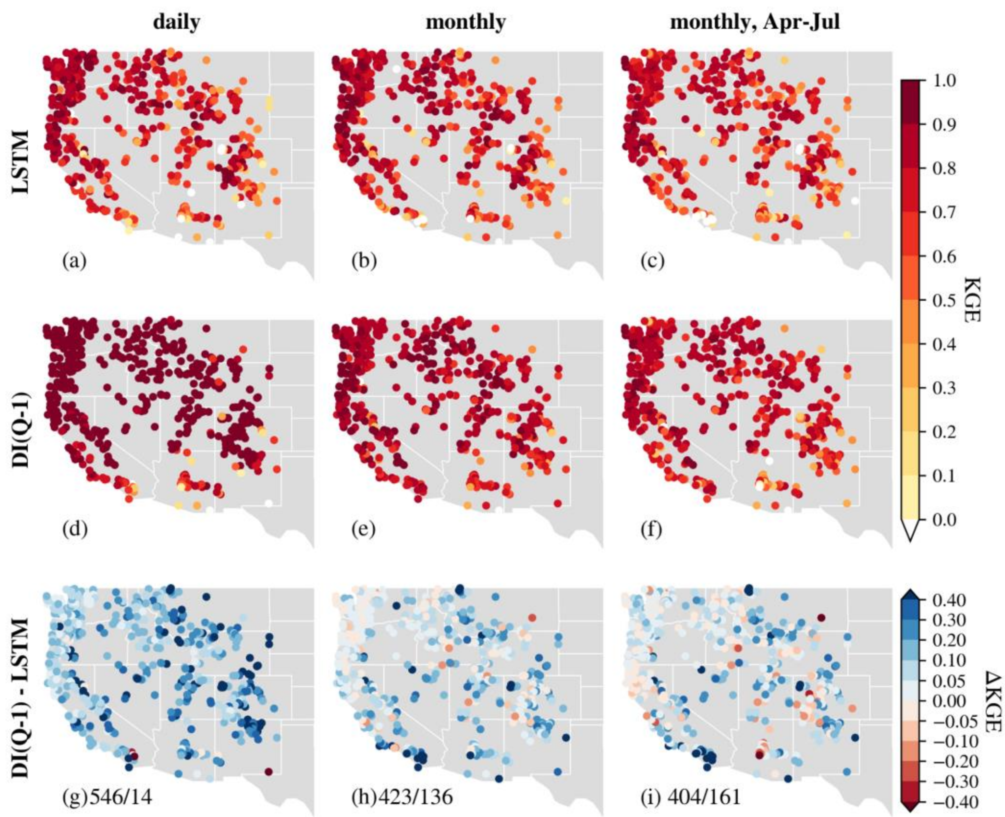

Figure 3. Comparison of KGE spatial patterns over the Western U.S. for experiments at the daily scale (left), monthly scale (middle) and monthly scale but only evaluation for April to July (right). From top to bottom: (a-c) LSTM, (d-f) DI(Q-1), (g-i) $\Delta KGE = KGE_{DI(Q-1)} - KGE_{LSTM}$. N1/N2 on (g-i) stands for the number of basins where DI(Q-1)/LSTM performs better, respectively.

In general, more recent observations typically contribute more to predictive improvements (Cheng et al., 2020; Sabzipour et al., 2023). The benefits of daily DI(Q) gradually decayed as N increased, with a corresponding widening of metric variability

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

(Fig. 2). This gradual decay of DI(Q) benefits, to a certain extent, reflects the memory length of hydrological processes (Feng et al., 2020; Sabzipour et al., 2023). However, even in the DI(Q-10) experiment, the median KGE, CC, RV and RB remained at 0.89, 0.95, 0.95 and -3.00%, respectively, still outperforming the baseline LSTM. This demonstrates that integrating streamflow from 10 days ago remains valuable for daily streamflow simulations. Accordingly, if implemented in a forecasting mode, the results suggest that near real-time streamflow observations could be leveraged to enhance streamflow forecast across these basins in the Western U.S. up to 10 days in advance.

3.2 The effectiveness of DI(Q) at the monthly scale

At the monthly scale, the baseline LSTM simulated streamflow well, achieving a median KGE of 0.80, quite similar to the daily-scale results. This consistency in performance across temporal resolutions aligns with findings from Yao et al. (2023), indicating that the standard LSTM is largely unaffected by changes in temporal resolution. Integrating lagged streamflow observations from 1 to 6 months ago also significantly improved model performance, yielding higher median values and reduced variability across all metrics. Monthly DI(Q-1) achieved a median KGE of 0.86 (Fig. 4a) and enhanced simulations in about 76% of basins (Fig. 3). Even DI(Q-6) exhibited a higher median KGE (0.83) and a smaller spread, showing the advantage of integrating monthly streamflow. However, the improvements at the monthly scale were less pronounced than those at the daily scale. This was expected since the monthly streamflow autocorrelation is usually weaker (Fig. C1), and lagged streamflow provides reduced predictive value.

Effective water management in the Western U.S. depends heavily on spring-summer (April-July) streamflow volume forecasts, commonly referred to as seasonal Water Supply Forecasts (WSFs). To assess model performance during this critical period, we evaluated streamflow from April to July. When evaluated specifically for the April-July period, LSTM performed slightly worse than the full-year analysis, with a median KGE of 0.76. However, integrating lagged monthly streamflow significantly contributes to better performance, with higher median KGE values for monthly DI(Q-1) and monthly DI(Q-6) (0.81 and 0.78, respectively) as well as reduced variability (Fig. 4c). The improvements for the April- July flow were slightly smaller than those for the year-round flow, likely because loss functions used in monthly DI(Q) experiments were optimized for year-round flow rather than being specifically tailored to the April-July period.

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

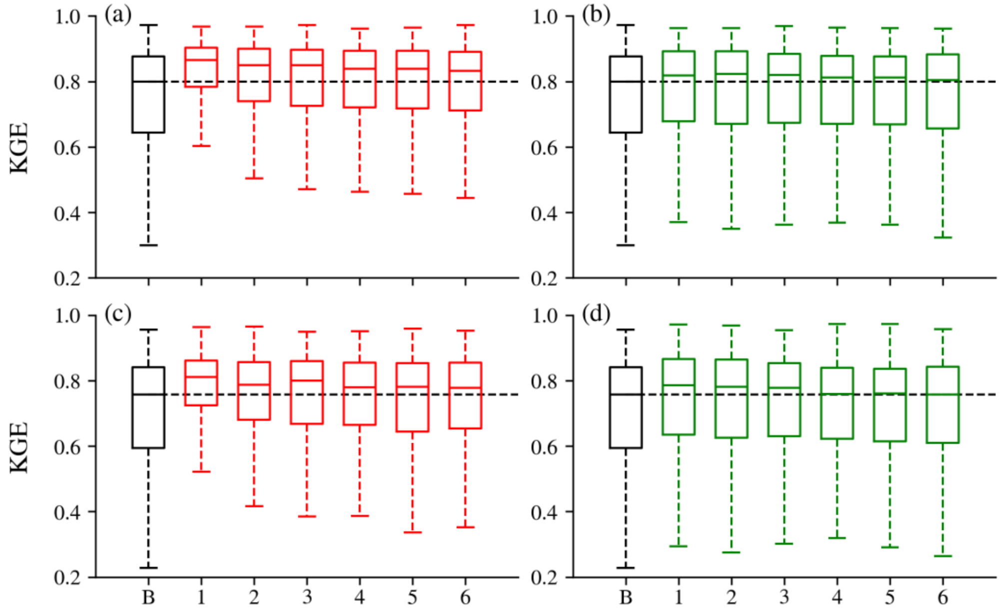

Figure 4. KGE boxplots for DI(Q-N) (left) and DI(SWE-N) (right) at monthly scale (top) and monthly scale but only evaluation from April to July (bottom).

245 **3.3 The effectiveness of DI(SWE) at the daily scale**

In contrast to daily DI(Q), integrating lagged SWE data at the daily scale did not improve streamflow simulations in terms of KGE (Fig. 5). This outcome aligns with expectations, as snow-related processes typically have a longer memory effect. Moreover, temperature, one of the model inputs, partially reflects snow dynamics, which the LSTM can effectively leverage through its memory states to estimate streamflow. However, significant improvements were still observed in CC and RV, indicating that DI(SWE) can enhance temporal dynamics and reduce variability biases. The overestimation was reduced, particularly during low-flow conditions, while underestimation worsened, leading to poorer RB medians. This increased underestimation may stem from the prevalence of seasonal snowpack in most basins, where abundant days with zero SWE values could introduce bias when integrated into the model. Additionally, the quality of the SWE dataset itself likely plays a role. Further investigation, such as utilizing SWE data from Airborne Snow Observatory (ASO, Painter et al., 2016) or snow course, is needed to better understand the underestimation issue.

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

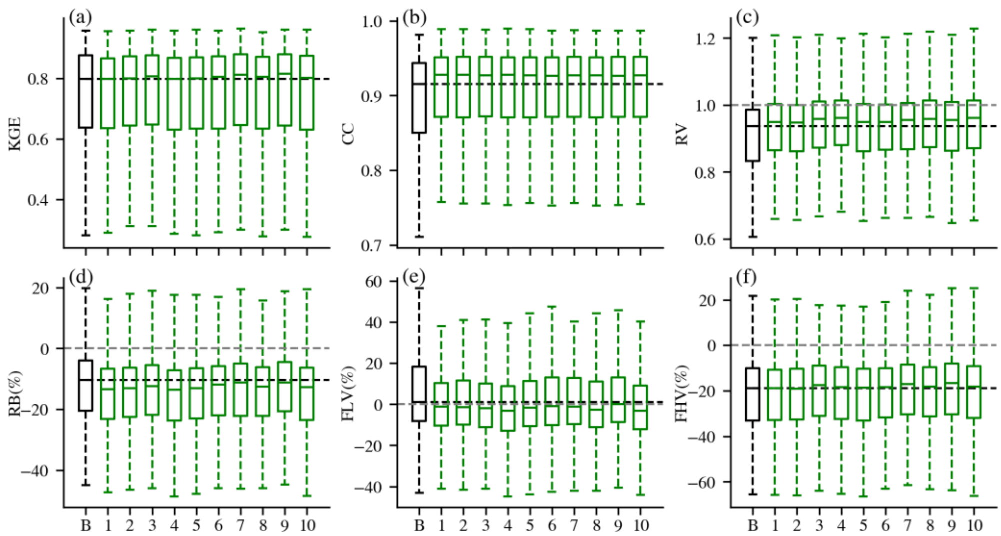

Figure 5. Performance of LSTM (black) and DI(SWE-N)(N=1-10) experiments (green) at the daily scale. The "B" on the x-axis stands for baseline LSTM, and N stands for the DI(SWE-N) experiment. The black horizontal line stands for the median value of the baseline LSTM. The grey horizontal line shows perfect value for RV, RB, FLV, and FHV.

Spatially, most improvements were observed in the Rocky Mountains (Fig. 6), where deeper snowpack usually exists and flow is dominated by snow. To further investigate whether the effect of integrating lagged SWE varies across different snowpacks, we evaluated model performance separately over rain-dominated basins (orange dots in Fig. 1a) and snow-dominated basins (blue dots in Fig. 1b). Figures 7a and 7e present the KGE values of the LSTM model, while Figures 7b and 7f show the KGE differences between DI(SWE) and LSTM at the daily scale for both types of basins. The baseline LSTM performed better in snow-dominated basins, with a higher median KGE of 0.80 (compared to 0.77 for rain-dominated basins) and smaller variability (Fig. 7). In terms of KGE differences, snow-dominated basins showed no obvious improvement, with a median ΔKGE of zero, while more rain-dominated basins exhibited negative ΔKGE after integrating lagged SWE. These rain-dominated basins are mainly located on the west side of the Cascade Mountains, the eastern slope of the Rocky Mountains, and the Southwest, where snowmelt is less dominant and rainfall contributes significantly to streamflow. Consequently, utilizing lagged SWE data did not show an impact on streamflow; instead, adding more zero SWE values into the LSTM model led to increased underestimation, ultimately degrading performance.

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

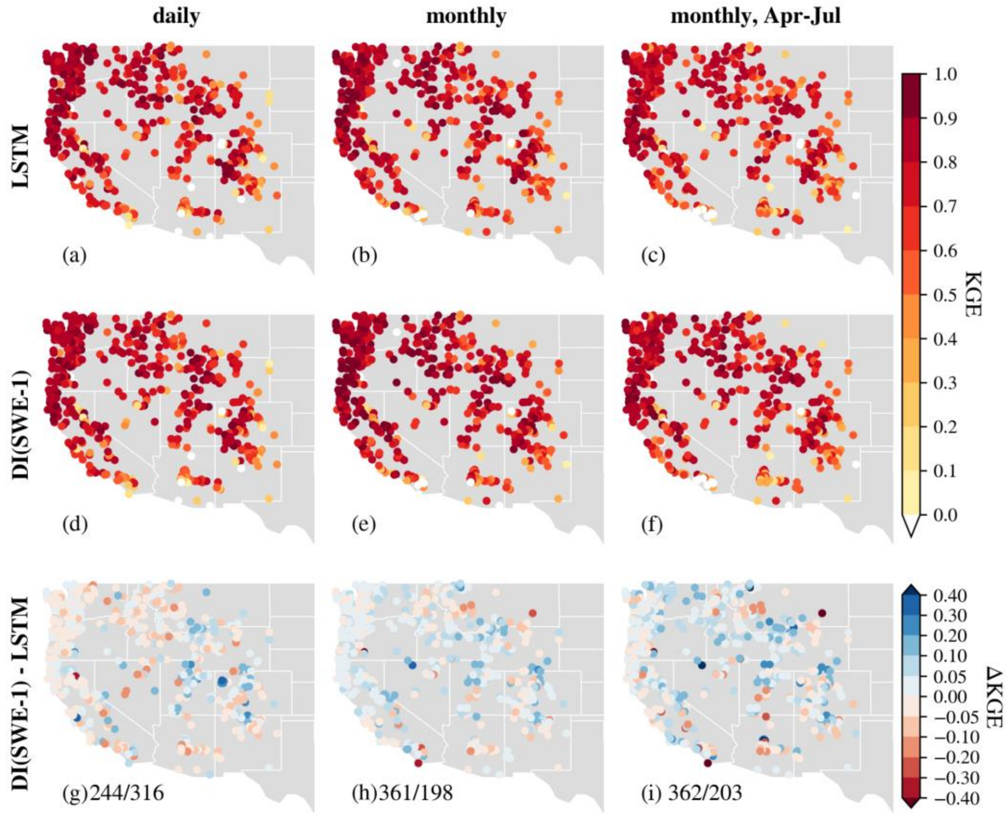

Figure 6. Comparison of KGE spatial patterns over the Western U.S. at the daily scale (left), monthly scale (middle) and monthly scale but only evaluation for April to July (right). From top to bottom: (a-c) LSTM, (d-f) DI(SWE-1), (g-i) $\Delta KGE = KGE_{DI(SWE-1)} - KGE_{LSTM}$. N1/N2 on (g-i) stands for the number of basins where DI(SWE-1)/LSTM performs better, respectively.

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

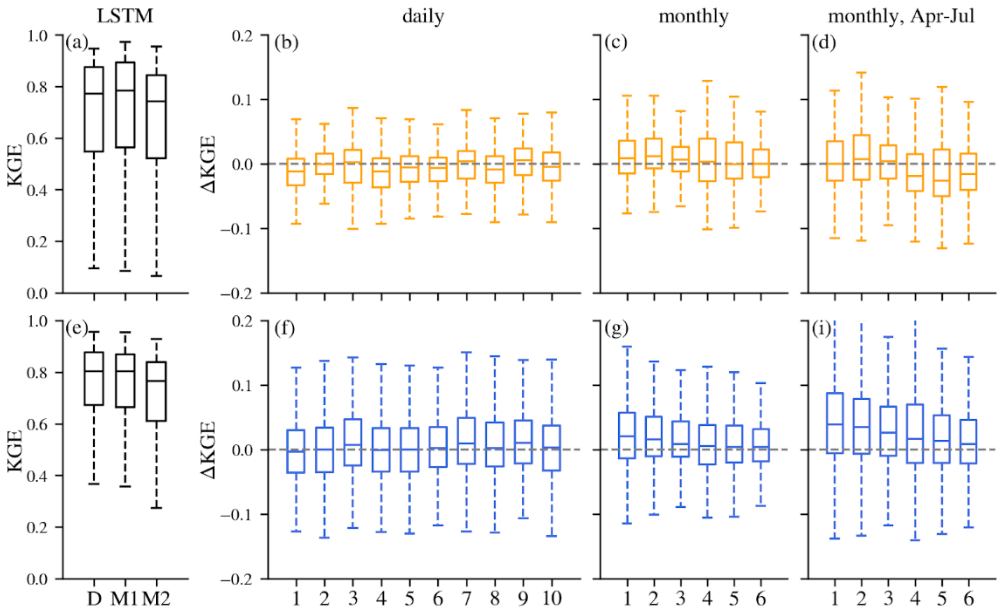

Figure 7. Comparison of KGE over rain-dominated basins (top) and snow-dominated basins (bottom). (a) and (e), black boxplots stand for KGE of baseline LSTM, and D, M1, M2 on the x-axis stand for the results of daily scale, monthly scale and monthly scale but evaluation only for April to July. (b-d) and (f-i) colored boxplots stand for KGE difference between DI(SWE-N) and LSTM ($\Delta KGE = KGE_{DI(SWE-N)} - KGE_{LSTM}$) at the daily, monthly, and monthly scale but evaluation only for April to July. The grey horizontal lines are zero.

Considering the delayed effect of snow processes on streamflow generation, we further investigated the effect of integrating SWE from different seasons (accumulation and snowmelt) on streamflow. Figure 8 shows the metric differences between DI(SWE) and LSTM during accumulation and snowmelt seasons. The spreads (i.e., interquartile ranges) of metric differences were wider during snowmelt season, indicating that integrating lagged SWE data had a greater effect (both deterioration and improvement) on streamflow estimation during this season (Fig. 8). The percentage of basins with positive ΔCC increased from 52-57% during accumulation season to 67-71% during snowmelt season. The median values of ΔCC were noticeably higher during snowmelt season compared to accumulation season (Fig. 8b). More improvements were also observed in RV during snowmelt season, with more basins showing RV values closer to 1 (negative |RV-1|) and larger negative median Δ|RV-1| (Fig. 8c). However, larger Δ|RB| were also observed during the snowmelt season. As a result, when considering the comprehensive metric, KGE, no noticeable difference in median ΔKGE was found between the two seasons.

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

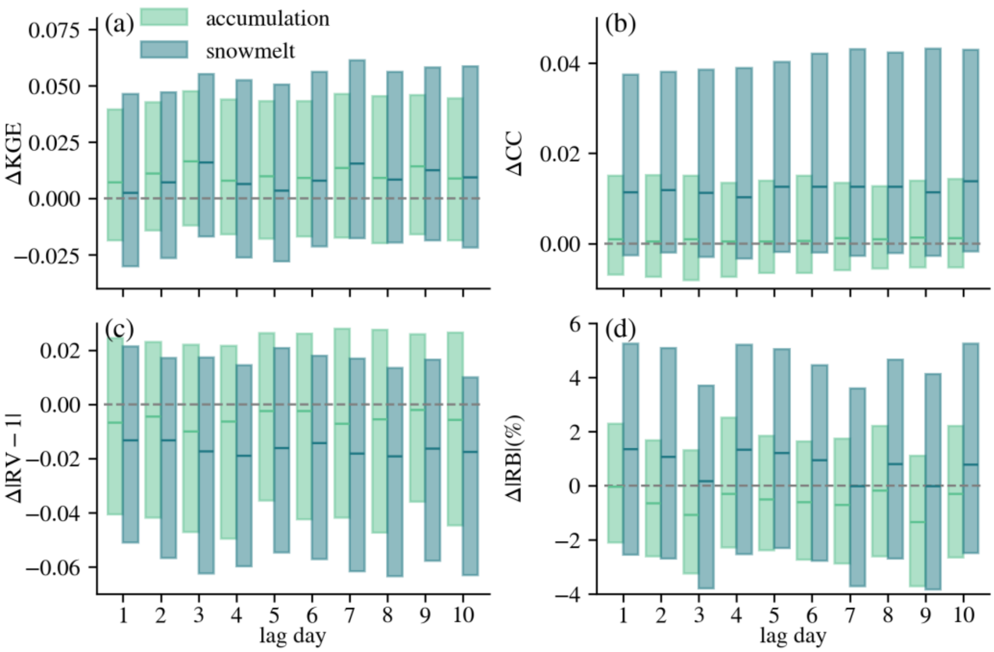

Figure 8. Metric differences between DI(SWE-N) and LSTM over snow accumulation and snowmelt seasons (difference in KGE, CC, |RV-1|, and |RB|). Only median and interquartile range (25th ~75th) are shown here. N stands for DI(SWE-N) experiment. The grey horizontal lines show zero.

3.4 The effectiveness of DI(SWE) at the monthly scale

Due to the long memory of snow processes in the hydrological cycle, integrating lagged SWE at the monthly scale provided benefits to streamflow simulation, as evidenced by slightly higher median KGE values as well as smaller spreads (Fig. 4). For instance, integrating lagged SWE from one month ago led to improved KGE in about 65% of basins (Fig. 6), with the median KGE increasing from 0.80 to 0.82. Similar improvements were also observed when evaluating spring-summer (April-July) streamflow.

The benefits of DI(SWE) at the monthly scale gradually declined as N increased, reflecting the decreasing persistence of snow in the hydrological cycle and its diminishing predictive value over longer lag periods (Fig. 4). However, DI(SWE-6) still showed some improvements, with slightly higher 25th and 75th percentiles and smaller interquartiles, despite an almost unchanged median. This suggests that integrating SWE data from six months ago remains informative for streamflow simulation. Therefore, if implemented in a forecasting mode, the findings suggest that near real-time SWE observations have the potential to enhance monthly streamflow forecasts up to six months in advance.

The benefits of DI(SWE) at the monthly scale were more pronounced in snow-dominated basins compared to rain-dominated basins (Fig. 7c and 7g). This improvement difference became even more evident when evaluating streamflow from April to July, the primary snowmelt season (Fig. 7d and 7i), further emphasizing the greater impact of DI(SWE) in snow-dominated basins during snowmelt season.

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

4 Discussions

4.1 Comparison of integrating different observations at different timescales

Figure 9 summarizes the median KGE values for all experiments at different timescales, as shown in Fig. 2, 4, and 5, separately. The benefits of different integration experiments can be roughly ranked as follows:

daily DI(Q) > monthly DI(Q) > monthly DI(SWE) > daily DI(SWE)

It is counterintuitive that even at the monthly scale and during April-July period, integrating lagged streamflow observations provided greater improvements than integrating SWE, despite snow being a key predictor of spring-summer flow in the snow-dominated Western U.S. (Fleming et al., 2024; Koster et al., 2010; Shukla and Lettenmaier, 2011; Wood et al., 2016). This may be because LSTM already inherently captures snow-related information through its memory states and learned relationships between streamflow and historical meteorological forcings (e.g., precipitation and temperature) (Feng et al., 2020; Jiang et al., 2022; Modi et al., 2025). Therefore, explicitly integrating lagged SWE as an additional predictor offers limited extra value for streamflow estimation.

In the monthly-scale analysis, DI(Q) yielded slightly greater improvements when evaluated over the entire year, whereas DI(SWE) showed a marginally larger enhancement in spring-summer flow estimates when integrating lagged SWE from 1–3 months prior.

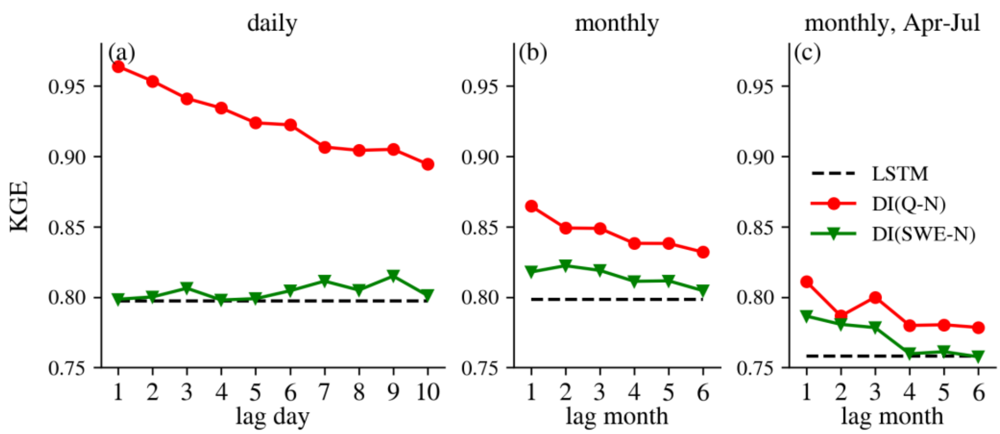

Figure 9. Median KGE values of all experiments at the daily scale (left), monthly scale(middle) and monthly scale but only evaluation for April to July (right). N on the x-axis stands for DI(Q-N) or DI(SWE-N) experiment.

4.2 Comparison of DI(SWE) between snow-dominated basins and all basins

From the above analysis, we found that DI(SWE) experiments showed greater improvements when evaluated over snow-dominated basins. To further explore this, we conducted the same DI(SWE) experiments exclusively trained over snow-dominated basins to determine if additional gains could be achieved. As expected, training the models (both LSTM and DI(SWE)) over a more homogeneous group of basins provided higher performance (Fig. C2). Figure 10 shows the median ΔKGE between DI(SWE) and the corresponding baseline LSTM over all basins and snow-dominated basins. Similar to daily

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

DI(SWE) trained over all basins, daily DI(SWE) trained exclusively over snow-dominated basins did not enhance streamflow estimation and even slightly degraded performance. However, at the monthly scale, DI(SWE) improved streamflow estimations for both the whole year flow and April-July flow. This improvement became more pronounced for the April-July period, reinforcing the finding that integrating SWE has a larger effect on streamflow estimation over snow-dominated basins during snowmelt season.

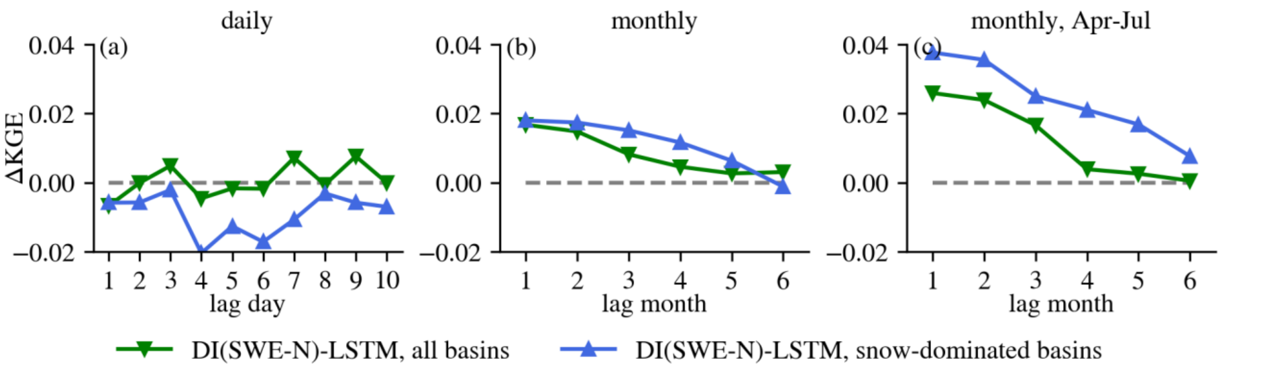

Figure 10. Median ΔKGE between DI(SWE) and LSTM over all basins (green) and snow-dominated basins (blue). From left to right are results at the daily scale, monthly scale and monthly scale but evaluation for April to July. N on the x-axis stands for DI(SWE-N) experiment.

4.3 Potential operational forecast applications and limitations

ML is gaining popularity in hydrology research and operational communities. This trend is driven by several key factors, including its easy implementation without substantial development and operational costs, strong model performance, ability to handle complex prediction tasks, and flexible model structure to adapt new datasets as additional predictors during training. Moreover, ML enables automated and objective modeling, minimizing the need for extensive manual interventions and subjective decision-making (Fleming et al., 2021, 2024; Modi et al., 2025).

This study evaluated the performance of an LSTM-powered data integration model that integrates lagged Q and SWE observations across various lag times at both daily and monthly scales. The pronounced improvements observed in the hindcasting mode highlight its potential for forecasting applications. In forecasting mode, recent observations can be incorporated into the LSTM model to dynamically update hydrological conditions, reducing the initialization errors compared to models that rely solely on forecasted forcings. In this framework, the “lag time” in hindcasting corresponds to the “lead time” in forecasting mode. In other words, by integrating recent Q or SWE data into the LSTM model, this approach could enhance streamflow forecasts in the Western U.S. with lead times of 1–10 days at the daily scale and 1–6 months at the monthly scale. Given its demonstrated effectiveness, flexibility, and automation, this data integration framework hold promises for real-time hydrological forecasting, offering valuable applications in water resource management.

Despite much promise, the DI-LSTM approach would have certain limitations when applied to operational streamflow forecasting. First, the improvements demonstrated in this study may be less pronounced in real-world forecasting applications.

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

Here, a "hindcasting" mode was used, leveraging observed meteorological forcings to evaluate the effective of DI-LSTM for streamflow simulations, thereby providing an upper bound on potential performance gains. However, operational forecast systems rely on predicted forcings, which inherently contain significant uncertainties that impact streamflow forecasts. Additionally, the accuracy of weather forecasts is expected to decay with increasing lead time, further diminishing the DI-LSTM predictive skill gains for longer lead time. Therefore, further research is necessary to assess the performance of DI-LSTM in an operational setting using actual forecasted meteorological inputs. Moreover, collaboration with the meteorological community is essential to improving the accuracy of forcing predictions. Second, this study provides deterministic streamflow estimation with limited uncertainty analysis. Uncertainty is inherent in all aspects of hydrological modeling, and its estimation is critical for actionable hydrological forecasts (Fang et al., 2020; Klotz et al., 2022). To address uncertainty due to random initial weights and biases, this study employed six repeated runs with different random seeds. However, uncertainties related to model inputs and observational data for model training were not explicitly considered. Recent studies have introduced various methods to quantify uncertainty in ML-based models for different uncertainty sources, such as Markov Chain Monte Carlo, variational inference, Monte Carlo dropout, Mixture density networks and ensemble techniques (Abdar et al., 2021). Future work should further explore uncertainty quantification to enhance forecast reliability and underpin decision-making in water resources management.

5 Conclusion

Based on LSTM, we evaluated a flexible data integration approach (DI-LSTM) incorporating different observations, e.g., Q and SWE, across multiple lag times at both daily and monthly scales over hundreds of basins in the Western U.S. By comparing DI-LSTM with the baseline LSTM, we assessed the impact of integrating lagged observations on streamflow estimations. The key findings in the Western U.S. are summarized as follows:

(1) The baseline LSTM without integrating any lagged observations already showed strong predictive capability in the Western U.S., achieving a median KGE of 0.80 at both daily and monthly scales.

(2) Integrating Q at the daily scale yielded the most substantial improvements, with significantly improved median values and reduced spread across all performance metrics. The median KGE across 646 basins increased to 0.96 with the integration of 1-day lagged streamflow and remained at 0.89 even with a 10-day lag. Integrating Q at the monthly scale also improved streamflow estimations, though to a lesser extent, with the median KGE increasing from 0.80 to 0.86 when integrating streamflow from 1 month ago.

(3) Integrating lagged SWE at the monthly scale led to better accuracy, whereas its integration at the daily scale did not improve streamflow estimations. This finding reflects the long-term memory of snow processes in the hydrological cycle, which extends beyond short timescales.

(4) The benefits of integrating SWE were more pronounced in snow-dominated basins during the snowmelt season, highlighting its value for improving spring-summer flow estimations.

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

(5) Overall, the benefits of integrating different observations at different timescales for streamflow estimations can be roughly ranked as follows: daily DI(Q) > monthly DI(Q) > monthly DI(SWE) > daily DI(SWE).

Due to its strong predictive performance, automation without the need for extensive domain-specific customization, and flexibility to ingest additional observations, the DI-LSTM approach demonstrates large potential for short-term (e.g., 1-10 days) and long-term (1-6 months) operational streamflow forecasts in the Western U.S. However, further studies, such as using real forecasted forcing data, are needed to assess its performance under realistic forecasting conditions.

Appendix A: Training basin selection and snow-dominated basin selection

We performed a screening to identify suitable training basins in the Western US by implementing the following procedure:

1) *Basin area*: Only basins within the range of 50-5,000 km^2 were selected. Basins smaller than 50 km^2 were discarded due to probable artificial boundaries. The maximum area threshold was applied since channel routing effects become apparent at the daily scale in larger basins (Gericke and Smithers, 2014).

2) *Data length*: only basins with at least 10-yr data during the training period (1983-2002) were selected to ensure sufficient data for training.

3) *Reservoir influences*: To minimize the effect of river regulation by dams or reservoirs, only basins with degree of regulation (DOR) no greater than 0.1 were selected (Ouyang et al., 2021). The DOR is defined as the ratio of total reservoir capacity within a basin to the mean annual cumulative discharge, with total reservoir capacity data sourced from GAGEII.

4) *Visual inspection*: Since some data are collected manually, they may contain errors in reported discharge values. We excluded basins with potentially erroneous discharge records, such as those with an unreasonably high magnitude far exceeding precipitation or with abrupt, dramatic differences between time intervals.

Snow-dominated basins were further selected based on the following two criteria:

1) *Corr*(maxSWE, Qtot) > *Corr*(Ptot, Qtot)

2) *Corr*(maxSWE, Qtot) > 0.1,

Here, *Corr* stands for correlation. MaxSWE refers to the maximum snow water equivalent (SWE) from the previous October to the following April for each Water Year (October to September). Qtot represents the total streamflow volume from April to July for each Water Year, while Ptot indicates the total precipitation over the same period.

Appendix B: LSTM model

LSTM introduces “memory cells” and “gates” to keep and filter information. Cell states allow information to be stored over long time periods, which is desirable for modeling processes such as snow accumulation and snowmelt. The input, forget and output gates control the flow of information, controlling what to let in, what to forget, and what to output from the system,

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

respectively. These gates are all trained automatically and simultaneously, using input data to predict the target variable. The forward propagation equations of the LSTM model are described by the following equations:

Input transformation: $\boldsymbol{x}^t = \boldsymbol{ReLU}(\boldsymbol{W}_I \boldsymbol{I}^t + \boldsymbol{b}_I)$, (A1)

Input node: $\boldsymbol{g}^t = \tanh(\mathcal{D}(\boldsymbol{W}_{gx}\boldsymbol{x}^t) + \mathcal{D}(\boldsymbol{W}_{gh}\boldsymbol{h}^{t-1}) + \boldsymbol{b}_g)$, (A2)

Input gate: $\boldsymbol{i}^t = \boldsymbol{\sigma}(\mathcal{D}(\boldsymbol{W}_{ix}\boldsymbol{x}^t) + \mathcal{D}(\boldsymbol{W}_{ih}\boldsymbol{h}^{t-1}) + \boldsymbol{b}_i)$, (A3)

Forget gate: $\boldsymbol{f}^t = \boldsymbol{\sigma}(\mathcal{D}(\boldsymbol{W}_{fx}\boldsymbol{x}^t) + \mathcal{D}(\boldsymbol{W}_{fh}\boldsymbol{h}^{t-1}) + \boldsymbol{b}_f)$, (A4)

Output gate: $\boldsymbol{o}^t = \boldsymbol{\sigma}(\mathcal{D}(\boldsymbol{W}_{ox}\boldsymbol{x}^t) + \mathcal{D}(\boldsymbol{W}_{oh}\boldsymbol{h}^{t-1}) + \boldsymbol{b}_o)$, (A5)

Cell state: $\boldsymbol{s}^t = \boldsymbol{g}^t \odot \boldsymbol{i}^t + \boldsymbol{s}^{t-1} \odot \boldsymbol{f}^t$, (A6)

Hidden state: $\boldsymbol{h}^t = \tanh(\boldsymbol{s}^t) \odot \boldsymbol{o}^t$, (A7)

Output: $\boldsymbol{y}^t = \boldsymbol{W}_{hy}\boldsymbol{h}^t + \boldsymbol{b}_y$ (A8)

Where $\boldsymbol{I}^t$ reprensents the raw input to the model, $\boldsymbol{x}^t$ represents the input vector to the LSTM cell. $\boldsymbol{ReLU}$ is the rectified linear unit, $\boldsymbol{\sigma}$ is the sigmoidal function, $\odot$ is the element-wise multiplication operator, $\mathcal{D}$ is the dropout operator. $\boldsymbol{W}$ and $\boldsymbol{b}$ with different subscripts represent the gate-specific network weights and bias parameters, respectively. $\boldsymbol{g}^t$ is the output of the input node, $\boldsymbol{i}^t$, $\boldsymbol{f}^t$, $\boldsymbol{o}^t$ are the input, forget, and output gates, respectively; $\boldsymbol{h}^t$ represents the hidden states, $\boldsymbol{s}^t$ represents the memory cell states and $\boldsymbol{y}^t$ represents the predicted output.

Appendix C

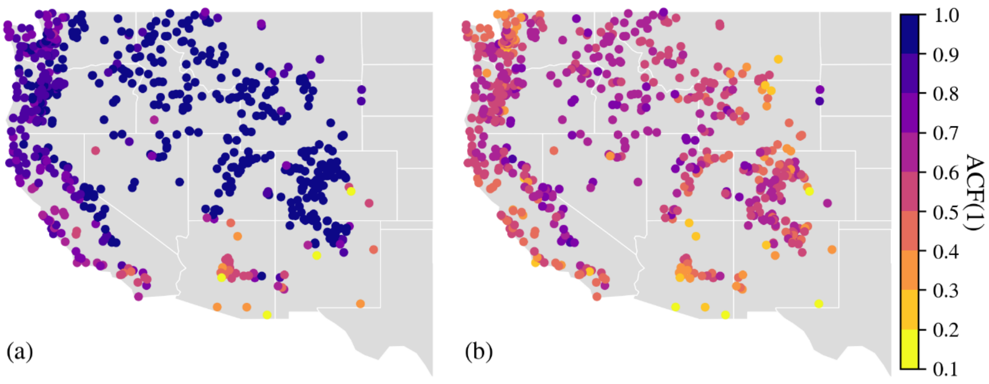

Figure C1. Spatial distribution of (a) 1-day-lag and (b) 1-month-lag autocorrelation function of streamflow (ACF(1)).

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

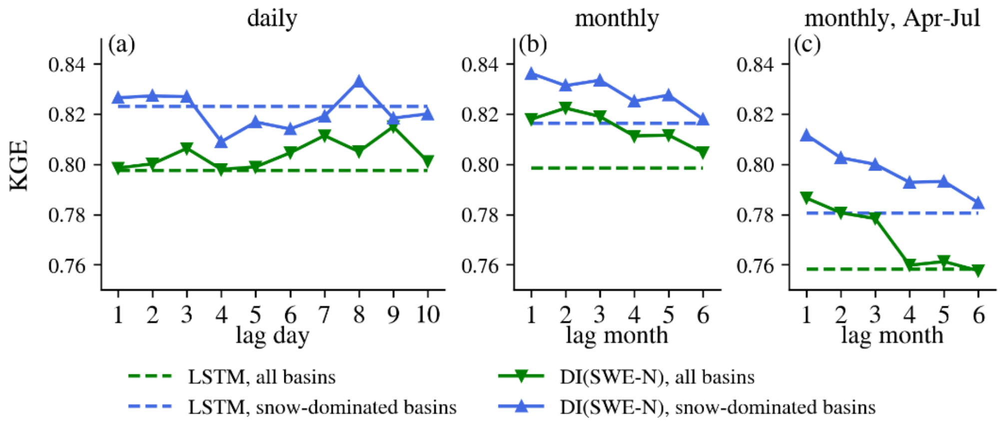

Figure C2. Median KGE values of DI(SWE-N) over all basins (green) and snow-dominated basins (blue). From left to right are results at the daily scale, monthly scale and monthly scale but only evaluation for April to July. N on the x-axis stands for DI(SWE-N) experiments.

Table C1: Summary of the forcing data and attribute variables used in this study.

	Variable	Data Source	Units
Forcing	Daily precipitation	MSWEP V2.80 (Beck et al., 2019)	mm/d
	Daily maximum temperature	ERA5 (Hersbach et al., 2018)	°C
	Daily minimum temperature		°C
	Daily mean surface downwelling shortwave		W/m^2
	Daily mean 10m wind		m/s
	Monthly LAI climatology	PROBA-V LAI (Fuster et al., 2020)	-
Attributes	Mean daily precipitation	MSWEP V2.80	mm/d
	High precipitation duration - the average duration of high precipitation events (number of consecutive days ⩾ 5 times mean daily precipitation)		days
	Fraction of precipitation falling as snow (i.e., on days colder than 0 °C)	MSWEP V2.80 and ERA5	-
	Aridity - P/PET, where PET is estimated by the Hargreaves (1994) method		-
	Frozen days - days colder than 0 °C	ERA5	days
	Area	basin boundary file	km^2

	Mean elevation	GMTED (Amatulli et al., 2018a)	m above sea level
	Mean slope		°
	Geological permeability	GLHYMPS V2 (Huscroft et al., 2018)	m2
	Soil sand content	SoilGrids (Hengl et al., 2017)	%

Table C2. Hyperparameters for the LSTM or DI-LSTM model

Hyperparameter	Daily Scale	Monthly Scale
Length of training instances	365	48
Mini-batching size	100	50
LSTM dropout rate	0.5	0.5
LSTM hidden size	256	256
Number of training epochs	300	300
Number of stacked LSTM layer	1	1

Table C3. The definition of KGE and its three component metrics.

Metric	Equation	Perfect Value
CC	$\mathrm{CC} = \frac{cov(Q_o, Q_m)}{\sigma_{Q_o} \cdot \sigma_{Q_m}}$	1
RV	$\mathrm{RV} = \frac{\sigma_{Q_m}/\mu_{Q_m}}{\sigma_{Q_o}/\mu_{Q_o}}$	1
RB	$\mathrm{RB} = \frac{\sum_1^N Q_{m,i} - \sum_1^N Q_{o,i}}{\sum_i^N Q_{o,i}} \times 100$	0
KGE	$\mathrm{KGE} = 1 - \sqrt{(CC-1)^2 + RB^2 + (\mathrm{RV}-1)^2}$	1

Note, Q_o, Q_m represent streamflow observations and simulations, respectively. cov, σ and μ represent covariance, standard deviation and mean, respectively.

Code and Data Availability. The source codes for LSTM-based rainfall-runoff simulations are from hydroDL, which is available at: https://zenodo.org/record/5015120 (Fang et al., 2021).

CW3E-Forcing is available at: https://www.reachhydro.org/home/records/1-km-conus-forcing (Pan, 2025). The PROBA-V LAI is available at: https://land.copernicus.eu/global/products/lai. Elevation data from GMTED is available at: https://doi.pangaea.de/10.1594/PANGAEA.867115 (Amatulli et al., 2018b). Geological permeability from GLHYMPS V2 is available at: https://borealisdata.ca/dataset.xhtml?persistentId=doi%3A10.5683/SP2/TTJNIU. Soil sand content data from SoilGrids is available at: https://soilgrids.org/.

The daily streamflow data from USGS is available at: https://waterdata.usgs.gov/nwis. UA SWE dataset: https://climate.arizona.edu/data/UA_SWE/DailyData_4km/. The reservoir storage information is from GAGEII attributes: https://pubs.usgs.gov/publication/70046617 (Falcone, 2011).

Author Contribution. YY: initial idea, modeling, analysis, visualization, and writing. MP: initial idea, conceptualization and project administration. YY took the lead in the preparation of the manuscript, but all the authors contributed.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors thank the developer of datasets used in this research for their efforts in creating and sharing valuable resources. Part of the analysis was conducted using Delta, managed by National Center for Supercomputing Applications at University of Illinois Urbana-Champaign.

Financial support. This research has been supported by the National Oceanic and Atmospheric Administration (NOAA) Cooperative Institute for Research on Hydrology (CIROH) through the NOAA Cooperative Agreement with The University of Alabama, NA22NWS4320003.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., and Nahavandi, S.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Inf. Fusion, 76, 243–297, https://doi.org/10.1016/j.inffus.2021.05.008, 2021.

Amatulli, G., Domisch, S., Tuanmu, M.-N., Parmentier, B., Ranipeta, A., Malczyk, J., and Jetz, W.: A suite of global, cross-scale topographic variables for environmental and biodiversity modeling, Sci. Data, 5, 180040, https://doi.org/10.1038/sdata.2018.40, 2018a.

Amatulli, G., Domisch, S., Tuanmu, M.-N., Parmentier, B., Ranipeta, A., Malczyk, J., and Jetz, W.: A suite of global, cross-scale topographic variables for environmental and biodiversity modeling, links to files in GeoTIFF format, PANGAEA, https://doi.org/10.1594/PANGAEA.867115, 2018b.

Ayana, Ö., Kanbak, D. F., Kaya Keleş, M., and Turhan, E.: Monthly streamflow prediction and performance comparison of machine learning and deep learning methods, Acta Geophys., 71, 2905–2922, https://doi.org/10.1007/s11600-023-01023-6, 2023.

Baker, S. A., Rajagopalan, B., and Wood, A. W.: Enhancing Ensemble Seasonal Streamflow Forecasts in the Upper Colorado River Basin Using Multi-Model Climate Forecasts, JAWRA J. Am. Water Resour. Assoc., 57, 906–922, https://doi.org/10.1111/1752-1688.12960, 2021.

Bannister, R. N.: A review of operational methods of variational and ensemble-variational data assimilation, Q. J. R. Meteorol. Soc., 143, 607–633, https://doi.org/10.1002/qj.2982, 2017.

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Dijk, A. I. J. M. van, McVicar, T. R., and Adler, R. F.: MSWEP V2 Global 3-Hourly 0.1° Precipitation: Methodology and Quantitative Assessment, Bull. Am. Meteorol. Soc., 100, 473–500, https://doi.org/10.1175/BAMS-D-17-0138.1, 2019.

Brown, J. D., Wu, L., He, M., Regonda, S., Lee, H., and Seo, D.-J.: Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 1. Experimental design and forcing verification, J. Hydrol., 519, 2869–2889, https://doi.org/10.1016/j.jhydrol.2014.05.028, 2014.

Broxton, P. D., Dawson, N., and Zeng, X.: Linking snowfall and snow accumulation to generate spatial maps of SWE and snow depth, Earth Space Sci., 3, 246–256, https://doi.org/10.1002/2016EA000174, 2016.

Broxton, P. D., Van Leeuwen, W. J. D., Svoma, B. M., Walter, J., and Biederman, J. A.: Subseasonal to seasonal streamflow forecasting in a semiarid watershed, JAWRA J. Am. Water Resour. Assoc., 59, 1493–1510, https://doi.org/10.1111/1752-1688.13147, 2023.

Cheng, M., Fang, F., Kinouchi, T., Navon, I. M., and Pain, C. C.: Long lead-time daily and monthly streamflow forecasting using machine learning methods, J. Hydrol., 590, 125376, https://doi.org/10.1016/j.jhydrol.2020.125376, 2020.

Clark, S. R., Lerat, J., Perraud, J.-M., and Fitch, P.: Deep learning for monthly rainfall–runoff modelling: a large-sample comparison with conceptual models across Australia, Hydrol. Earth Syst. Sci., 28, 1191–1213, https://doi.org/10.5194/hess-28-1191-2024, 2024.

Cosgrove, B., Gochis, D., Flowers, T., Dugger, A., Ogden, F., Graziano, T., Clark, E., Cabell, R., Casiday, N., Cui, Z., Eicher, K., Fall, G., Feng, X., Fitzgerald, K., Frazier, N., George, C., Gibbs, R., Hernandez, L., Johnson, D., Jones, R., Karsten, L., Kefelegn, H., Kitzmiller, D., Lee, H., Liu, Y., Mashriqui, H., Mattern, D., McCluskey, A., McCreight, J. L., McDaniel, R., Midekisa, A., Newman, A., Pan, L., Pham, C., RafieeiNasab, A., Rasmussen, R., Read, L., Rezaeianzadeh, M., Salas, F., Sang, D., Sampson, K., Schneider, T., Shi, Q., Sood, G., Wood, A., Wu, W., Yates, D., Yu, W., and Zhang, Y.: NOAA's National Water Model: Advancing operational hydrology through continental-scale modeling, JAWRA J. Am. Water Resour. Assoc., 60, 247–272, https://doi.org/10.1111/1752-1688.13184, 2024.

Dalkilic, H. Y., Kumar, D., Samui, P., Dixon, B., Yesilyurt, S. N., and Katipoğlu, O. M.: Application of deep learning approaches to predict monthly stream flows, Environ. Monit. Assess., 195, 705, https://doi.org/10.1007/s10661-023-11331-5, 2023.

Dawson, N., Broxton, P., and Zeng, X.: A New Snow Density Parameterization for Land Data Initialization, J. Hydrometeorol., 18, 197–207, https://doi.org/10.1175/JHM-D-16-0166.1, 2017.

Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H. D., Fresch, M., Schaake, J., and Zhu, Y.: The Science of NOAA's Operational Hydrologic Ensemble Forecast Service, Bull. Am. Meteorol. Soc., 95, 79–98, https://doi.org/10.1175/BAMS-D-12-00081.1, 2014.

Eghbali, H. J.: K-S Test for Detecting Changes from Landsat Imagery Data, IEEE Trans. Syst. Man Cybern., 9, 17–23, https://doi.org/10.1109/TSMC.1979.4310069, 1979.

Falcone, J. A.: GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow, https://doi.org/10.5066/P96CPHOT, 2011.

Falcone, J. A., Carlisle, D. M., Wolock, D. M., and Meador, M. R.: GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States, Ecology, 91, 621–621, https://doi.org/10.1890/09-0889.1, 2010.

Fang, K., Kifer, D., Lawson, K., and Shen, C.: Evaluating the Potential and Challenges of an Uncertainty Quantification Method for Long Short-Term Memory Models for Soil Moisture Predictions, Water Resour. Res., 56, e2020WR028095, https://doi.org/10.1029/2020WR028095, 2020.

Fang, K., Shen, C., and Feng, D.: mhpi/hydroDL: MHPI-hydroDL, Zenodo, https://doi.org/10.5281/zenodo.5015120, 2021.

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

Feng, D., Lawson, K., and Shen, C.: Mitigating Prediction Error of Deep Learning Streamflow Models in Large Data-Sparse Regions With Ensemble Modeling and Soft Data, Geophys. Res. Lett., 48, e2021GL092999, https://doi.org/10.1029/2021GL092999, 2021.

Feng, D., Beck, H., de Bruijn, J., Sahu, R. K., Satoh, Y., Wada, Y., Liu, J., Pan, M., Lawson, K., and Shen, C.: Deep dive into hydrologic simulations at global scale: harnessing the power of deep learning and physics-informed differentiable models (δHBV-globe1.0-hydroDL), Geosci. Model Dev., 17, 7181–7198, https://doi.org/10.5194/gmd-17-7181-2024, 2024.

Fleming, S. W. and Goodbody, A. G.: A Machine Learning Metasystem for Robust Probabilistic Nonlinear Regression-Based Forecasting of Seasonal Water Availability in the US West, IEEE Access, 7, 119943–119964, https://doi.org/10.1109/ACCESS.2019.2936989, 2019.

Fleming, S. W., Garen, D. C., Goodbody, A. G., McCarthy, C. S., and Landers, L. C.: Assessing the new Natural Resources Conservation Service water supply forecast model for the American West: A challenging test of explainable, automated, ensemble artificial intelligence, J. Hydrol., 602, 126782, https://doi.org/10.1016/j.jhydrol.2021.126782, 2021.

Fleming, S. W., Rittger, K., Oaida Taglialatela, C. M., and Graczyk, I.: Leveraging Next-Generation Satellite Remote Sensing-Based Snow Data to Improve Seasonal Water Supply Predictions in a Practical Machine Learning-Driven River Forecast System, Water Resour. Res., 60, e2023WR035785, https://doi.org/10.1029/2023WR035785, 2024.

https://doi.org/10.5194/egusphere-2025-1708
Preprint. Discussion started: 30 April 2025

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

Franz, K. J., Hogue, T. S., Barik, M., and He, M.: Assessment of SWE data assimilation for ensemble streamflow predictions, J. Hydrol., 519, 2737–2746, https://doi.org/10.1016/j.jhydrol.2014.07.008, 2014.

Fuster, B., Sánchez-Zapero, J., Camacho, F., García-Santos, V., Verger, A., Lacaze, R., Weiss, M., Baret, F., and Smets, B.: Quality Assessment of PROBA-V LAI, fAPAR and fCOVER Collection 300 m Products of Copernicus Global Land Service, Remote Sens., 12, 1017, https://doi.org/10.3390/rs12061017, 2020.

Garen, D. C.: Improved Techniques in Regression-Based Streamflow Volume Forecasting, J. Water Resour. Plan. Manag., 118, 654–670, https://doi.org/10.1061/(ASCE)0733-9496(1992)118:6(654), 1992.

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

Gericke, O. J. and Smithers, J. C.: Review of methods used to estimate catchment response time for the purpose of peak discharge estimation, Hydrol. Sci. J., 59, 1935–1971, https://doi.org/10.1080/02626667.2013.866712, 2014.

Gichamo, T. Z. and Tarboton, D. G.: Ensemble Streamflow Forecasting Using an Energy Balance Snowmelt Model Coupled to a Distributed Hydrologic Model with Assimilation of Snow and Streamflow Observations, Water Resour. Res., 55, 10813–10838, https://doi.org/10.1029/2019WR025472, 2019.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J.: LSTM: A search space odyssey, IEEE Trans. Neural Netw. Learn. Syst., 28, 2222–2232, https://doi.org/10.1109/TNNLS.2016.2582924, 2016.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, PloS One, 12, e0169748, https://doi.org/10.1371/journal.pone.0169748, 2017.

Hersbach, H., de Rosnay, P., Bell, B., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Alonso-Balmaseda, M., Balsamo, G., Bechtold, P., Berrisford, P., Bidlot, J.-R., de Boisséson, E., Bonavita, M., Browne, P., Buizza, R., Dahlgren, P., Dee, D., Dragani, R., Diamantakis, M., Flemming, J., Forbes, R., Geer, A., Haiden, T., Hólm, E., Haimberger, L., Hogan, R., Horányi, A., Janiskova, M., Laloyaux, P., Lopez, P., Munoz-Sabater, J., Peubey, C., Radu, R., Richardson, D., Thépaut, J.-N., Vitart, F., Yang, X., Zsótér, E., and Zuo, H.: Operational global reanalysis: progress, future directions and synergies with NWP, https://doi.org/10.21957/TKIC6G3WM, 2018.

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

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

Huscroft, J., Gleeson, T., Hartmann, J., and Börker, J.: Compiling and Mapping Global Permeability of the Unconsolidated and Consolidated Earth: GLobal HYdrogeology MaPS 2.0 (GLHYMPS 2.0), Geophys. Res. Lett., 45, 1897–1904, https://doi.org/10.1002/2017GL075860, 2018.

Jiang, S., Zheng, Y., Wang, C., and Babovic, V.: Uncovering Flooding Mechanisms Across the Contiguous United States Through Interpretive Deep Learning on Representative Catchments, Water Resour. Res., 58, e2021WR030185, https://doi.org/10.1029/2021WR030185, 2022.

Khoshkalam, Y., Rousseau, A. N., Rahmani, F., Shen, C., and Abbasnezhadi, K.: Applying transfer learning techniques to enhance the accuracy of streamflow prediction produced by long Short-term memory networks with data integration, J. Hydrol., 622, 129682, https://doi.org/10.1016/j.jhydrol.2023.129682, 2023.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, J. Hydrol., 424–425, 264–277, https://doi.org/10.1016/j.jhydrol.2012.01.011, 2012.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G.: Uncertainty estimation with deep learning for rainfall–runoff modeling, Hydrol. Earth Syst. Sci., 26, 1673–1693, https://doi.org/10.5194/hess-26-1673-2022, 2022.

Koster, R. D., Mahanama, S. P. P., Livneh, B., Lettenmaier, D. P., and Reichle, R. H.: Skill in streamflow forecasts derived from large-scale estimates of soil moisture and snow, Nat. Geosci., 3, 613–616, https://doi.org/10.1038/ngeo944, 2010.

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

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, Water Resour. Res., 55, 11344–11354, https://doi.org/10.1029/2019WR026065, 2019a.

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

Le, X.-H., Ho, H. V., Lee, G., and Jung, S.: Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting, Water, 11, 1387, https://doi.org/10.3390/w11071387, 2019.

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

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

Li, D., Wrzesien, M. L., Durand, M., Adam, J., and Lettenmaier, D. P.: How much runoff originates as snow in the western United States, and how will that change in the future?, Geophys. Res. Lett., 44, 6163–6172, https://doi.org/10.1002/2017GL073551, 2017.

Li, Z., Gao, S., Chen, M., Gourley, J. J., Liu, C., Prein, A. F., and Hong, Y.: The conterminous United States are projected to become more prone to flash floods in a high-end emissions scenario, Commun. Earth Environ., 3, 1–9, https://doi.org/10.1038/s43247-022-00409-6, 2022.

Mangukiya, N. K. and Sharma, A.: Deep Learning-Based Approach for Enhancing Streamflow Prediction in Watersheds With Aggregated and Intermittent Observations, Water Resour. Res., 61, e2024WR037331, https://doi.org/10.1029/2024WR037331, 2025.

Mangukiya, N. K., Sharma, A., and Shen, C.: How to enhance hydrological predictions in hydrologically distinct watersheds of the Indian subcontinent?, Hydrol. Process., 37, e14936, https://doi.org/10.1002/hyp.14936, 2023.

Modi, P., Jennings, K., Kasprzyk, J., Small, E., Wobus, C., and Livneh, B.: Using Deep Learning in Ensemble Streamflow Forecasting: Exploring the Predictive Value of Explicit Snowpack Information, J. Adv. Model. Earth Syst., 17, e2024MS004582, https://doi.org/10.1029/2024MS004582, 2025.

Nearing, G., Yatheendradas, S., Crow, W., Zhan, X., Liu, J., and Chen, F.: The Efficiency of Data Assimilation, Water Resour. Res., 54, 6374–6392, https://doi.org/10.1029/2017WR020991, 2018.

Nearing, G., Klotz, D., Frame, J. M., Gauch, M., Gilon, O., Kratzert, F., Sampson, A. K., Shalev, G., and Nevo, S.: Technical note: Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks, Hydrol. Earth Syst. Sci., 26, 5493–5513, https://doi.org/10.5194/hess-26-5493-2022, 2022.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T. Y., Weitzner, D., and Matias, Y.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, https://doi.org/10.1038/s41586-024-07145-1, 2024.

Ouyang, W., Lawson, K., Feng, D., Ye, L., Zhang, C., and Shen, C.: Continental-scale streamflow modeling of basins with reservoirs: Towards a coherent deep-learning-based strategy, J. Hydrol., 599, 126455, https://doi.org/10.1016/j.jhydrol.2021.126455, 2021.

Painter, T. H., Berisford, D. F., Boardman, J. W., Bormann, K. J., Deems, J. S., Gehrke, F., Hedrick, A., Joyce, M., Laidlaw, R., Marks, D., Mattmann, C., McGurk, B., Ramirez, P., Richardson, M., Skiles, S. M., Seidel, F. C., and Winstral, A.: The Airborne Snow Observatory: Fusion of scanning lidar, imaging spectrometer, and physically-based modeling for mapping snow water equivalent and snow albedo, Remote Sens. Environ., 184, 139–152, https://doi.org/10.1016/j.rse.2016.06.018, 2016.

Pan, M.: CW3E 1-km 1-hourly meteorological forcing on NWM grid. Center for Western Weather and Water Extremes (CW3E), https://doi.org/10.5281/zenodo.14714512, 2025.

Perkins, T. R., Pagano, T. C., and Garen, D. C.: Innovative operational seasonal water supply forecasting technologies, J. Soil Water Conserv., 64, 15A-17A, https://doi.org/10.2489/jswc.64.1.15A, 2009.

Pierce, D. W., Barnett, T. P., Hidalgo, H. G., Das, T., Bonfils, C., Santer, B. D., Bala, G., Dettinger, M. D., Cayan, D. R., Mirin, A., Wood, A. W., and Nozawa, T.: Attribution of Declining Western U.S. Snowpack to Human Effects, J. Clim., 21, 6425–6444, https://doi.org/10.1175/2008JCLI2405.1, 2008.

Prasad, R., Deo, R. C., Li, Y., and Maraseni, T.: Input selection and performance optimization of ANN-based streamflow forecasts in the drought-prone Murray Darling Basin region using IIS and MODWT algorithm, Atmospheric Res., 197, 42–63, https://doi.org/10.1016/j.atmosres.2017.06.014, 2017.

Sabzipour, B., Arsenault, R., Troin, M., Martel, J.-L., Brissette, F., Brunet, F., and Mai, J.: Comparing a long short-term memory (LSTM) neural network with a physically-based hydrological model for streamflow forecasting over a Canadian catchment, J. Hydrol., 627, 130380, https://doi.org/10.1016/j.jhydrol.2023.130380, 2023.

Saharia, M., Kirstetter, P.-E., Vergara, H., Gourley, J. J., Hong, Y., and Giroud, M.: Mapping Flash Flood Severity in the United States, J. Hydrometeorol., 18, 397–411, https://doi.org/10.1175/JHM-D-16-0082.1, 2017.

Schmidhuber, J.: Deep learning in neural networks: An overview, Neural Netw., 61, 85–117, https://doi.org/10.1016/j.neunet.2014.09.003, 2015.

Shen, C.: A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists, Water Resour. Res., 54, 8558–8593, https://doi.org/10.1029/2018WR022643, 2018.

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, F., Kifer, D., Li, L., Liu, X., Ren, W., Zheng, Y., Harman, C. J., Clark, M., Farthing, M., Feng, D., Kumar, P., Aboelyazeed, D., Rahmani, F., Song, Y., Beck, H. E., Bindas, T., Dwivedi, D., Fang, K., Höge, M., Rackauckas, C., Mohanty, B., Roy, T., Xu, C., and Lawson, K.: Differentiable modelling to unify machine learning and physical models for geosciences, Nat. Rev. Earth Environ., 1–16, https://doi.org/10.1038/s43017-023-00450-9, 2023.

Shukla, S. and Lettenmaier, D. P.: Seasonal hydrologic prediction in the United States: understanding the role of initial hydrologic conditions and seasonal climate forecast skill, Hydrol. Earth Syst. Sci., 15, 3529–3538, https://doi.org/10.5194/hess-15-3529-2011, 2011.

Smirnov, N.: Table for Estimating the Goodness of Fit of Empirical Distributions, Ann. Math. Stat., 19, 279–281, https://doi.org/10.1214/aoms/1177730256, 1948.

Snyder, C., Bengtsson, T., Bickel, P., and Anderson, J.: Obstacles to High-Dimensional Particle Filtering, Mon. Weather Rev., 136, 4629–4640, https://doi.org/10.1175/2008MWR2529.1, 2008.

Song, Y., Tsai, W.-P., Gluck, J., Rhoades, A., Zarzycki, C., McCrary, R., Lawson, K., and Shen, C.: LSTM-Based Data Integration to Improve Snow Water Equivalent Prediction and Diagnose Error Sources, J. Hydrometeorol., 25, 223–237, https://doi.org/10.1175/JHM-D-22-0220.1, 2024.

Thapa, S., Zhao, Z., Li, B., Lu, L., Fu, D., Shi, X., Tang, B., and Qi, H.: Snowmelt-Driven Streamflow Prediction Using Machine Learning Techniques (LSTM, NARX, GPR, and SVR), Water, 12, 1734, https://doi.org/10.3390/w12061734, 2020.

Wood, A. W., Hopson, T., Newman, A., Brekke, L., Arnold, J., and Clark, M.: Quantifying Streamflow Forecast Skill Elasticity to Initial Condition and Climate Prediction Skill, J. Hydrometeorol., 17, 651–668, https://doi.org/10.1175/JHM-D-14-0213.1, 2016.

Yao, Y., Zhao, Y., Li, X., Feng, D., Shen, C., Liu, C., Kuang, X., and Zheng, C.: Can transfer learning improve hydrological predictions in the alpine regions?, J. Hydrol., 625, 130038, https://doi.org/10.1016/j.jhydrol.2023.130038, 2023.

Yaseen, Z. M., El-shafie, A., Jaafar, O., Afan, H. A., and Sayl, K. N.: Artificial intelligence based models for stream-flow forecasting: 2000–2015, J. Hydrol., 530, 829–844, https://doi.org/10.1016/j.jhydrol.2015.10.038, 2015.

Zeng, X., Broxton, P., and Dawson, N.: Snowpack Change From 1982 to 2016 Over Conterminous United States, Geophys. Res. Lett., 45, 12,940-12,947, https://doi.org/10.1029/2018GL079621, 2018.