JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2024MS004582

Using Deep Learning in Ensemble Streamflow Forecasting: Exploring the Predictive Value of Explicit Snowpack Information

Parthkumar Modi[1,2], Keith Jennings[3], Joseph Kasprzyk[1], Eric Small[4], Cameron Wobus[5], and Ben Livneh[1,2,6]

[1]Department of Civil, Environmental, and Architectural Engineering, University of Colorado Boulder, Boulder, CO, USA, [2]Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder, Boulder, CO, USA, [3]Water Resources Institute, University of Vermont, Burlington, VT, USA, [4]Department of Geological Sciences, University of Colorado Boulder, Boulder, CO, USA, [5]CK Blueshift, LLC, Phoenix, AZ, USA, [6]Western Water Assessment (WWA), University of Colorado Boulder, Boulder, CO, USA

Key Points:

- A new combination of deep learning and Ensemble Streamflow Prediction is evaluated
- Deep learning based seasonal streamflow forecasts perform similarly to Natural Resources Conservation Services statistical forecasts across different lead times
- Forecasts that incorporate comprehensive historical snowpack information perform better than those that exclude snow data

Supporting Information:

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

Correspondence to:

P. Modi,
parthkumar.modi@colorado.edu

Citation:

Modi, P., Jennings, K., Kasprzyk, J., Small, E., Wobus, C., & Livneh, B. (2025). Using deep learning in ensemble streamflow forecasting: Exploring the predictive value of explicit snowpack information. *Journal of Advances in Modeling Earth Systems*, *17*, e2024MS004582. https://doi.org/10.1029/2024MS004582

Received 2 AUG 2024
Accepted 18 FEB 2025

Abstract

The Ensemble Streamflow Prediction (ESP) framework combines a probabilistic forecast structure with process-based models for water supply predictions. However, process-based models require computationally intensive parameter estimation, increasing uncertainties and limiting usability. Motivated by the strong performance of deep learning models, we seek to assess whether the Long Short-Term Memory (LSTM) model can provide skillful forecasts and replace process-based models within the ESP framework. Given challenges in *implicitly* capturing snowpack dynamics within LSTMs for streamflow prediction, we also evaluated the added skill of *explicitly* incorporating snowpack information to improve hydrologic memory representation. LSTM-ESPs were evaluated under four different scenarios: one excluding snow and three including snow with varied snowpack representations. The LSTM models were trained using information from 664 GAGES-II basins during WY1983–2000. During a testing period, WY2001–2010, 80% of basins exhibited Nash-Sutcliffe Efficiency (NSE) above 0.5 with a median NSE of around 0.70, indicating satisfactory utility in simulating seasonal water supply. LSTM-ESP forecasts were then tested during WY2011–2020 over 76 western US basins with operational Natural Resources Conservation Services (NRCS) forecasts. A key finding is that in high snow regions, LSTM-ESP forecasts using simplified ablation assumptions performed worse than those excluding snow, highlighting that snow data do not consistently improve LSTM-ESP performance. However, LSTM-ESP forecasts that explicitly incorporated past years' snow accumulation and ablation performed comparably to NRCS forecasts and better than forecasts excluding snow entirely. Overall, integrating deep learning within an ESP framework shows promise and highlights important considerations for including snowpack information in forecasting.

Plain Language Summary

The Ensemble Streamflow Prediction (ESP) framework generates probabilistic water supply forecasts using process-based models. However, process-based models often face challenges because estimating their parameters is complex and introduces uncertainties. Inspired by emerging deep learning techniques, our study investigates whether the Long Short-Term Memory (LSTM) model can provide skillful forecasts and potentially replace process-based models within ESP. We also explore how including explicit information about snow, which significantly influences water flow, could enhance these forecasts. Our findings indicate that in regions with heavy snowfall, using a simpler representation of snowpack led to less accurate forecasts than those excluding snow information, suggesting that snowpack information does not consistently improve forecast performance. However, LSTM-ESP forecasts incorporating a sophisticated representation of snowpack information performed comparably to current operational forecasts and better than those excluding snowpack information. Integrating deep learning techniques within an ESP could improve water supply forecasts, but careful consideration of how to incorporate snowpack information is crucial for optimal results.

1. Introduction

Water Supply Forecasts (WSFs) from sub-seasonal to seasonal lead times, that is, weeks to months ahead, can aid in water allocation decisions (Chiew et al., 2003; Kaune et al., 2020), crop selection strategies (Mushtaq et al., 2012), water supply controls (Crochemore et al., 2016), managing extremes like droughts and floods

(Amnatsan et al., 2018; Ficchì et al., 2016; Watts et al., 2012), and water market planning (Turner et al., 2017). WSFs are issued at hundreds of points, primarily basin outlets, across the western US near the first of the month between January and June each year. Water at many of these forecast points originates as melt from winter snowpack, emphasizing the role of accurate snow information in supporting the performance of the WSFs (Daly et al., 2000; Li et al., 2017). A popular choice for operational forecasting is the use of Ensemble Streamflow Predictions (ESPs) that use historical weather data and process-based models (Förster et al., 2018; Girons Lopez et al., 2021; Harrigan et al., 2018; Singh, 2016; Wang et al., 2010; Yuan et al., 2016). The historical weather data used by ESPs functions as an analog for seasonal climate forecast uncertainty and serves as a simple estimate of forecast uncertainty (Day, 1985; Troin et al., 2021). Despite the variety of physical representations within process-based models, they suffer from other notable uncertainties arising from simplified model structures and challenges in identifying model parameters. Alternatively, emerging deep learning techniques like the Long Short-Term Memory (LSTM) models learn complex patterns from data using interconnected layers. These models have demonstrated the ability to improve streamflow prediction (Arsenault et al., 2022; Kratzert et al., 2018; Kratzert, Klotz, Herrnegger, et al., 2019; Kratzert, Klotz, Shalev, et al., 2019; Nearing et al., 2021), raising the question about how to effectively leverage deep learning frameworks to enhance WSF performance. Here, we seek to evaluate the potential utility of the emerging deep learning technique, LSTMs, in an ESP framework for water supply forecasting in the western US.

Probabilistic seasonal streamflow forecast information can aid in making risk-based decisions. These often take the form of ESPs that generate an ensemble of equiprobable future streamflow traces with lead times of between 30 and 180 days that is, at sub-seasonal to seasonal timescales. At the core of the ESPs are the initial hydrologic conditions that represent the current state of the hydrologic system and set the starting point for the forecasts. To date, ESPs have relied on process-based models to generate WSFs. However, the reliability and applicability of process-based models used in generating ESP forecasts are limited by several factors. Among these factors are biases in initial hydrologic conditions, which arise due to a lack of knowledge and incomplete process representation of factors controlling initial conditions, which ultimately contribute to inaccuracies in forecasting (DeChant & Moradkhani, 2011). Additional limiting factors include challenges in identifying process-based model parameters, necessitating computationally expensive and potentially ill-constrained calibration (Arheimer et al., 2020; Hirpa et al., 2015; Wood et al., 2016).

Emerging deep learning techniques *learn* complex patterns from data using interconnected layers, enabling them to minimize or avoid the issues facing process-based models. Deep learning is a subset of machine learning and is a powerful method that leverages successive layers of nonlinear transformations (i.e., neural networks) to learn complex predictor mapping from input data to target outputs (Chollet & Chollet, 2021). Recently, multiple investigations have shown improved streamflow performance of deep learning models attributed to large data availability, for for example, CAMELS (Addor et al., 2017; Newman et al., 2014), CONUS404 (Rasmussen et al., 2023), and Caravan (Kratzert et al., 2023), and an implicit accounting of physical relationships through input-output mappings of predictors to predictands (Fleming et al., 2021). These models have been shown to perform equally or better than process-based models in studies including, but not limited to Arsenault et al. (2022), Kratzert, Klotz, Herrnegger, et al. (2019), and Nearing et al. (2021). Considering these performance improvements of deep learning models over process-based models, a parallel surge in streamflow forecasting applications has been driven by the growing exploration of deep learning methodologies (Ibrahim et al., 2022; Ng et al., 2023; Sit et al., 2020). Reasons for the growing interest in deep learning forecasts versus process-based forecasts primarily include lower development and operational costs, the handling of complex tasks with limited domain knowledge, and regionalization performance—extrapolating information from multiple basins to make predictions in individual gaged or ungauged basins (Kratzert, Klotz, Herrnegger, et al., 2019; Kratzert, Klotz, Shalev, et al., 2019).

There are several types of deep learning models. Simple recurrent neural networks retain limited information, rendering them generally unsuitable for hydrological modeling applications (Bengio et al., 1994), where time-tracking of physical components across multiple weeks or months is essential. LSTM is a special kind of recurrent neural network, first proposed by Hochreiter and Schmidhuber (1997), designed to learn long-term dependencies in time-series data. A detailed description of the workings of an LSTM model and its components is provided in Kratzert et al. (2018) and Kratzert, Klotz, Herrnegger, et al. (2019). LSTMs have now been well established for streamflow modeling through the intercomparison of modeling approaches (Kratzert, Klotz, Herrnegger, et al., 2019; Kratzert, Klotz, Shalev, et al., 2019; Mai et al., 2022), the ability to predict

unprecedented extremes (Frame et al., 2022), and regionalization (Ayzel et al., 2020; Kratzert et al., 2018; Nearing et al., 2021). Owing to better streamflow prediction accuracy and regionalization capabilities, LSTMs present a viable alternative to process-based models in an ESP framework by eliminating the need for domain-specific calibration and flexibility in handling diverse domains.

Similar to the utility of initial hydrologic conditions in initializing the process-based ESP forecasts, the LSTM's memory states serve as a useful proxy for generating ESP forecasts. While LSTMs memory states can be conceptually understood as a kind of storage or model state (Kratzert et al., 2018), their utility in generating an ESP forecast requires a comprehensive analysis. As stated earlier, initial hydrologic conditions are crucial for an ESP forecast, with snowpack information identified as the main source of predictive skill across snow-dominated western regions (Koster et al., 2010; Shukla & Lettenmaier, 2011; Wood et al., 2016). Snow serves as a critical source of hydrologic memory, acting as a predictor that impacts the performance of ESP forecasts. While LSTMs trained for streamflow have demonstrated the ability to capture snow dynamics implicitly, there are limitations associated with them since these LSTMs are only trained for streamflow and not explicitly for snow (Kratzert et al., 2018). Hence, identifying the best way to incorporate snow explicitly is a nontrivial endeavor, given the deviation of LSTMs architecture from traditional process-based models.

Motivated by the aforementioned performance of LSTM models compared to process-based models, as well as the limitations associated with process-based ESPs, our study introduces a novel implementation of a LSTM model in an ESP framework. The analysis includes assessing LSTM-ESP forecasts generated by LSTM models, which are trained regionally using information from a large sample of basins across the CONUS domain. A set of four forecast experiments are explored in the new LSTM-ESP framework: one that excludes snow data and three that integrate snow data representations ranging from simple to complex. We seek to understand whether explicitly providing different degrees of snowpack information can improve the LSTM-ESP forecasts. Lastly, we compare the LSTM-ESP with operational forecasts from the NRCS, serving as an operational water supply forecasting benchmark using a consistent methodology over a sufficiently long overlapping period with generated LSTM-ESP forecasts.

2. Methods

We first outline the specifics of the training data and basin selection criteria in Section 2.1. In Section 2.2, we introduce the LSTM by providing details on the model architecture and training process. In Section 2.3, we describe the training protocols of the LSTM models tested here: one model trained without snow information—$LSTM_{NoSnow}$, and one that uses snow information—$LSTM_{Snow}$. This section also describes the implementation of these LSTM models in an ESP framework. Section 2.4 outlines the design of four LSTM-ESP forecast experiments, with one using $LSTM_{NoSnow}$ model and three other experiments using the $LSTM_{Snow}$ model. Section 2.5 provides an overview of forecast performance metrics.

2.1. Study Domain and Training Data

We selected a diverse set of drainage basins across the western US, spanning a range of hydroclimatic characteristics. A total of 664 drainage basins were identified by analyzing the geospatial attributes from the USGS Geospatial Attributes of Gages for Evaluating Streamflow (GAGES-II; Falcone, 2011; Falcone et al., 2010) data set, the Hydro-Climatic Data Network (HCDN; Slack & Landwehr, 1992), and the Catchment Attributes and Meteorology for Large-sample Studies (CAMELS; Addor et al., 2017; Newman et al., 2014) data set. The screening procedure is described in Section 2.1.1, and a map of training basins is shown in Figure S1 in Supporting Information S1.

The predictors for training LSTM models were selected and employed based on the works of Arsenault et al. (2022), Kratzert, Klotz, Herrnegger, et al. (2019), and Kratzert, Klotz, Shalev, et al. (2019). These predictors (Table 1) include North American Land Data Assimilation System Phase 2 (NLDAS2; Xia et al., 2012) meteorological forcings, aggregated on a daily basis and spatially averaged for each basin from 1/8° (~12-km) grids. These forcings consist of precipitation, average wind speed, 2 m average air temperature, incoming longwave and shortwave radiation, near-surface air pressure, and near-surface vapor pressure. Additionally, the static predictors incorporate basin-wide attributes from GAGES-II that remain constant at each daily time step (U.S. Geological Survey, 2023). In general, the selection of the GAGES-II basin attributes mirrors those utilized in the CAMELS

Table 1
Training Predictors Consisting of Meteorological Forcings (Source: NLDAS2), Static Basin Attributes (Source: GAGES-II), and Snow Data (Source: UA) With Streamflow Data (Source: USGS) as the Predictand

Category	Name	Description
Static	PPTAVG_BASIN	Mean annual precipitation (mm)
	PET	Mean annual potential evapotranspiration (mm)
	T_AVG_BASIN	Average annual air temperature (°C)
	SNOW_PCT_PRECIP	Snow percent of total precipitation estimate
	WDMAX_BASIN	Watershed average of monthly max. number of days of measurable precipitation
	WDMIN_BASIN	Watershed average of monthly min. number of days of measurable precipitation
	PRECIP_SEAS_IND	Precipitation seasonality index (Dingman, 2002; Markham, 1970). Index of how much annual precipitation falls seasonally (high values) or spread out over the year (low values).
	RUNAVE7100	Mean annual total runoff (mm)
	RE	Runoff efficiency = PPTAVG_BASIN/RUNAVE7100
	ELEV_MAX_BASIN	Maximum watershed elevation (m)
	ELEV_MIN_BASIN	Minimum watershed elevation (m)
	DRAIN_SQKM	Watershed drainage area (km^2)
	SLOPE_PCT	Mean watershed slope (%)
	FORESTNLCD06	Watershed percent forest (%)
	PLANTNLCD06	Watershed percent planted/cultivated
	PNV_BAS_PCT	Percentage of the watershed covered by the dominant potential natural vegetation
	ROCKDEPAVE	Average value of total soil thickness examined (in)
	AWCAVE	Average value for the range of available water capacity for the soil layer
	CLAYAVE	Average value of clay content (%)
	SILTAVE	Average value of silt content (%)
	SANDAVE	Average value of sand content (%)
	PERMAVE	Average permeability (in/hr)
	KFACT_UP	Average K-factor for the uppermost soil horizon in each soil component. K-factor is an erodibility factor which quantifies the susceptibility of soil particles to detachment and movement by water.
Meteorological	PRCP	Average daily precipitation (mm/day)
	WIND	Average wind speed (m/s)
	TAS	2 m daily average air temperature (°C)
	SRAD	Incoming shortwave solar radiation (W/m^2)
	LRAD	Incoming longwave solar radiation (W/m^2)
	PRES	Near-Surface Air pressure (Pa)
	VP	Near-Surface Vapor Pressure (Pa)
Snow[a]	SWE	Average Snow Water Equivalent (mm)
Streamflow	SF	Average daily streamflow (mm/day)

Note. The asterisk indicates that the predictor was only included in one of the two trained LSTM models. [a]Snow was the only variable used in $LSTM_{Snow}$ that was not used in $LSTM_{No\text{-}Snow}$.

data set. Daily streamflow estimates from the USGS's National Water Information System (USGS NWIS) were obtained for the USGS stream gages corresponding to the basin outlets (United States Geological Survey, 2024).

We obtain daily snow information from the gridded snow data set, developed at the University of Arizona (Broxton et al., 2019; Zeng et al., 2018—now UA), spatially averaged for each basin from 1/16-degree grids. This data set assimilates snowpack observations from in situ networks like SNOTEL and compares favorably to remote sensing and airborne lidar measurements (Zeng et al., 2018). For a supporting analysis, we also obtained daily SWE from the Natural Resource Conservation Service's SNOTEL network. To handle spatial interpolation from multiple sites, we followed the NRCS's approach that spatially averages snow from all stations inside and

within a certain radius of the basin boundary (i.e., within 40 km of the basin boundary; more details are provided in Text S1 in Supporting Information S1).

2.1.1. Basin Screening Procedure

The basin screening procedure applied here was similar to the CAMELS approach (Addor et al., 2017; Newman et al., 2014) but allowed us to include a slightly larger number of basins relative to CAMELS basins. Both CAMELS and additional basins screened are a subset of the GAGES-II data set that were filtered for minimal human influence, resulting in a majority of basins being headwater basins that fall below a drainage area threshold of 2,500 km^2. Thus, we specifically included basins with minimal anthropogenic influence with drainage areas up to 2,500 km^2 (with an average size of roughly 560 km^2) and a minimum of 30 years of streamflow observations to ensure sufficient data for model training and testing. For additional non-CAMELS basins sourced from GAGES-II, we enforced certain attribute criteria. These included limiting the number of major dams (storage >5,000 acre-feet) to 1 or fewer, the ratio of reservoir storage to average streamflow from 1971 to 2000 to less than 10%, and the GAGES-II hydro-disturbance index is less than 10 (Falcone et al., 2010). To ensure that the basin boundaries and drainage area were accurately represented in the selected basins, additional criteria based on the GAGES-II boundary attributes were employed. The first was basin boundary confidence greater than or equal to 8, that is, how well the basin boundary matches the true drainage area, with value 10 representing high confidence and 2 low confidence. It is important to note that this metric represents a qualitative assessment, not a quantitative one, based on the reliability of the delineation process and its consistency with verified hydrological and geological data (Falcone, 2011). In addition, a percent difference in area less or equal to 10% was used as compared to values reported in the USGS National Water Information System database. The second was a qualitative check of their HUC10 boundaries that are deemed to be at least "reasonable" or "good" based on three factors: (a) how closely the drainage area matches the USGS National Water Information System database, (b) visual comparison of basin boundary to HUC10 boundaries, and (c) streamflow outlet location relative to basin boundary and streamlines (these attributes are further described in Falcone, 2011; Falcone et al., 2010). We chose to work with these specific categories because our focus was on training LSTM models using spatially averaged predictors, which did not require very strict constraints. It should be noted that only 76 basins (out of 664 basins used for training) had official NRCS forecasts available for comparing the experimental forecast skill. These basins are colored in Figure 1 by the ratio of April 1 SWE to water-year to date cumulative precipitation that is derived from gridded snow product and NLDAS2, respectively. The mean annual ratio of April 1 SWE, used here as a proxy for peak SWE (Pagano et al., 2004), to water-year to date cumulative precipitation (SWE/P) is calculated over the water years 1982–2022. This ratio is one way to quantify the degree to which a basin could be considered snowmelt-dominated, with higher values corresponding to more snowmelt-dominated streamflow, and vice versa for lower values. We explore this ratio to ensure that we are incorporating varying snowpack characteristics across the western US. A majority of these basins lie within the US Environmental Protection Agency's snow level III ecoregions labeled in Figure 1.

2.2. Long Short-Term Memory (LSTM) Models

While any neural network can learn non-linear systems, LSTMs excel at handling sequences of inputs over extended periods due to their unique architectures. LSTMs have two key memory states that manage information: the hidden state, which acts as a working memory retaining information from recent events, and the cell state, which functions like a conveyor belt to store and retrieve information from past events. This study adopts a setup similar to Kratzert, Klotz, Herrnegger, et al. (2019) (now onwards "K19") for the experimental setup, as they were shown to work well with CAMELS basins that have minimal anthropogenic influence and have sufficient data for training and testing (data sets are described in Section 2.1). Since we generated forecasts for in-sample basins (i.e., for only basins used in training), we adopted the setup of "LSTM (with statics)" from Kratzert, Klotz, Herrnegger, et al. (2019). This setup only includes hyperparameters and not model parameters or training inputs. Hyperparameters are externally set before the training starts and govern the training process itself, whereas model parameters are learned from the training data during the training process. The hyperparameters used by K19 were not modified due to marginal performance gains observed with varying hyperparameter selection and model architectures (Arsenault et al., 2022). For this study, we elaborate on selected model parameters and hyperparameters crucial to the training process (Table 2 and Figure 2). Specifically, the model parameters consist of weights and biases, whereas hyperparameters include the number of hidden layers, the number of units in each

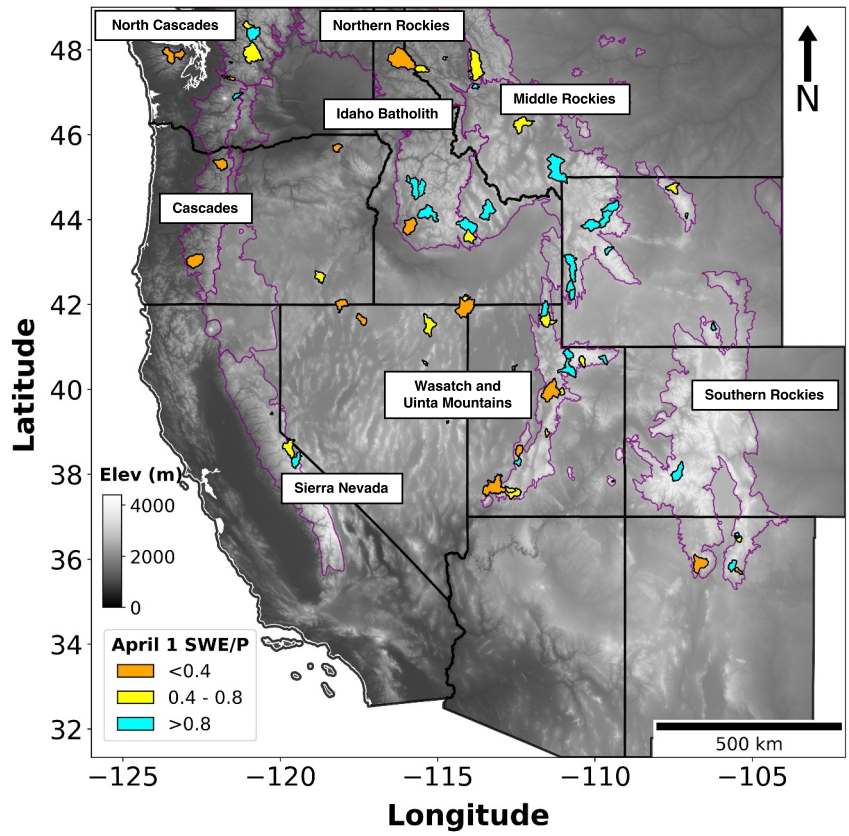


Figure 1. A map of the study domain, comprising 76 USGS drainage basins across the western US colored by the ratio of April 1 SWE to water year-to-date precipitation. The purple boundaries indicate the North American snow ecoregions Level III generated by the US Environmental Protection Agency (US EPA, 2015).

layer, the input sequence length, batch size, dropout rate, number of epochs, optimizer, and batch size. The list of hyperparameters is briefly outlined in Table 2 and explained in the subsequent paragraphs.

Using the K19 model structure, the LSTM includes a single hidden layer comprising 256 units, where units act as computational units through which data flows, and the hidden layer is responsible for learning the intricate structures in data. Additionally, the hidden layer is configured with a dropout rate of 0.4, which involves randomly dropping neurons during training to mitigate overfitting. The input sequence length used is 270 days, which specifies the number of preceding time steps fed into the LSTM to produce streamflow on a given day.

Table 2
The Long Short-Term Memory Hyperparameters Used in This Study

Parameter	Description	Selected value
Number of hidden layers	The number of stacked LSTM layers in the model	1
Number of units	The number of memory cells in each LSTM layer that determine the capacity to learn from the data	256
Input sequence length	The length of preceding time steps fed into the LSTM	270
Batch size	The number of training samples used in one iteration	2,000
Dropout rate	The fraction of the units to drop during training to prevent overfitting	0.4
Number of epochs	The number of times the entire training data set is passed through the model	40
Optimizer	The algorithm used to minimize the loss function	Adam
Learning rate	The step size used by the optimization algorithm to update the model weights	0.001

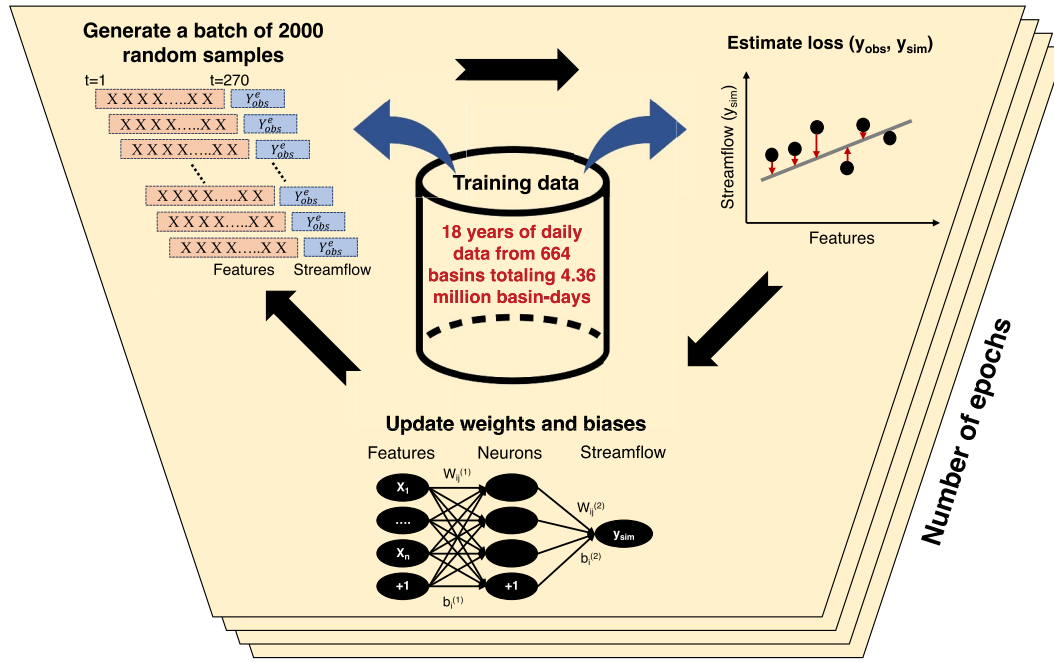


Figure 2. Schematic of Long Short-Term Memory model training for each iteration within an epoch. In each iteration, 2,000 independent random samples are drawn from 18 years of daily data from 664 basins totaling 4.36 million basin-days. Each sample consists of 270 days, that is, input sequence length, of preceding predictors (X) and one target observation (y_{obs}). The loss is computed between observed discharge (y_{obs}) and the network's prediction (y_{sim}). The model parameters, including weights ($w_1 \ldots w_m$) and biases ($b_1 \ldots b_m$), are updated after every iteration. Epoch refers to the complete passing of the entire training data set through the model algorithm once. The weights and biases are model parameters, whereas the batch size, input sequence length, and number of epochs are the hyperparameters.

A schematic of the LSTM training process is shown in Figure 2. As a first step, the weights and biases are initialized using the Xavier uniform distribution (Glorot & Bengio, 2010). Following the initialization, a subset of 2,000 samples (i.e., batch size—an LSTM hyperparameter) is randomly extracted in every iteration from all available training samples to make predictions. Similar to K19, the training is done regionally, that is, training samples from 664 basins across the CONUS are used. Each sample consists of one streamflow observation on a given day, representing the value of the dependent variable, alongside the input from a sequence of the preceding 270 days, such that we construct a "sequence to value" LSTM prediction. More details on trained LSTM models and training data are provided in Sections 2.1 and 2.3.1, respectively.

Since streamflow observations on a particular day are solely contingent upon predictors from the preceding 270 days, a batch can encompass random time steps or random basins devoid of any chronological ordering requirement (Kratzert et al., 2018). This batch configuration is possible because static basin attributes are included as inputs, informing the model about the characteristics of the basin. During each iteration, the predictors (X) are passed through the model, which comprises weights (w) and biases (b), to generate predictions (y_{sim}) and compute the error, also known as loss relative to observations (y_{obs}). Subsequently, this loss is employed to update the model parameters through back-propagation algorithms (Figure 2). Since the training encompasses basins with varying hydroclimatic characteristics, the training loss function is a basin-averaged Nash-Sutcliffe Efficiency (NSE) that is modified to normalize the mean squared error within each basin using the variance of discharge observations (Kratzert, Klotz, Herrnegger, et al., 2019; Kratzert, Klotz, Shalev, et al., 2019). This is done to avoid overweighting large and humid basins within the training loss function.

Similar to the calibration process in traditional process-based models, where parameters are calibrated based on objective criteria, LSTM model parameters (specifically weights and biases) are updated iteratively using a training loss function. However, the iteration concept differs between both models. In process-based models, each iteration corresponds to a full model run during calibration, with parameters updated after each iteration. In contrast, in LSTM models, parameter updates occur after every epoch, where one epoch refers to the complete

passing of all training data through the model algorithm once. For example, if the entire training data comprises 100,000 samples and the batch size is 2000, one epoch would comprise 50 iterations, for example, 100,000/2,000. Consequently, within each epoch, each sample out of all samples is selected without replacement until all samples have been utilized at least once (Kratzert et al., 2018). This study uses 40 epochs for training with a single seed and an Adam optimizer, which offers more adaptability and efficiency as compared to the Stochastic Gradient Descent (Ruder, 2016). We chose 40 epochs based on the experiments performed by Arsenault et al. (2022), Kratzert et al. (2018), and Kratzert, Klotz, Herrnegger, et al. (2019). We did not repeat the training process for multiple seeds as the impact on the performance was minimal (Table 2; Kratzert, Klotz, Shalev, et al., 2019).

2.3. Experimental Setup

In Section 2.3.1, we first describe the training process of two distinct LSTM models: one trained without snow information—$LSTM_{NoSnow}$—and one that uses snow information—$LSTM_{Snow}$. This is followed by a comparison with the reproduced K19 model. In Section 2.3.2, we outline the process of running these LSTM models in an ESP framework.

2.3.1. LSTM Model Training

Two distinct LSTM models were trained. The $LSTM_{NoSnow}$ used only meteorological forcings and basin attributes and no snow information, providing a baseline for comparison. The $LSTM_{Snow}$ was trained using meteorological forcings, basin attributes, as well as gridded snow data. LSTM models were trained from WY1983–2000 with a predictand of daily USGS streamflow data at the basin outlets. Following the model training, we evaluated the performance of both LSTM models against K19 and observations in years WY2001–2010 over 493 basins, which represents the subset of the CAMELS basins that are common between K19 and this study. It should be noted that we reproduced the K19 model ("LSTM—with statics") based on Kratzert, Klotz, Herrnegger, et al. (2019) with a single seed. The LSTM-ESP forecasts from all the regional models were only generated in recent years, WY2011–2020, due to the widespread availability of forecasts for comparison (further described in Section 2.4). It is important to highlight that the ESP forecasts, which generate daily streamflow, were accumulated to April–July water supply volumes, a primary component of the analysis.

2.3.2. Ensemble Streamflow Predictions

In general, ESP forecasts generated on April 1 hold significant operational importance. This is because April 1 historically serves as a surrogate for the timing of peak SWE conditions and provides near-maximum predictive information (Pagano et al., 2004). However, anticipated future reductions in snow may diminish the predictive efficacy of April 1 (Livneh & Badger, 2020; P. A. Modi et al., 2022). Hence, for this study, we include three forecast dates—February 1, March 1, and April 1 and one evaluation period—April–July. The forecast period extends from the forecast date through the end of July. While most of our analysis focuses on April 1 as the forecast date, we also included two additional forecast dates: February 1 and March 1. These forecast dates were selected because they are commonly used in operational forecasting (Pagano et al., 2009) and accommodate for regional variations in the timing of peak SWE across our study basins (Musselman et al., 2021).

For ESP forecasts on a given forecast date, the LSTM simulation begins at the start of the water year, that is, October 1, using true meteorological forcings to obtain LSTM's memory states on the forecast date. Using these memory states on the forecast date and historical meteorological forcings, an ensemble of streamflow traces is produced in the forecast period as a function of the current hydroclimatic state and retrospective weather conditions (Day, 1985; Troin et al., 2021). The result is a daily probabilistic hydrologic forecast ranging from 30 days up to 180 days from the forecast date that uses the spread in historical data from the past ∼20–30 years as an analog for the uncertainty in meteorological conditions after the forecast date.

A conceptual illustration of an ESP forecast generated on April 1 using LSTM is shown in Figure 3. The thick red line indicates the model run before the forecast date using "true" meteorological forcings. Ensemble streamflow traces (shown in light red) are generated using the memory states on April 1 (shown in blue) and historical meteorological forcings from the past 23 years. For example, the streamflow prediction for a given day in the forecast period will be generated using 270 days (i.e., input sequence length) of preceding predictors that consist of observed meteorological forcings prior to April 1 and historical meteorological forcings from April 1 until the day of prediction.

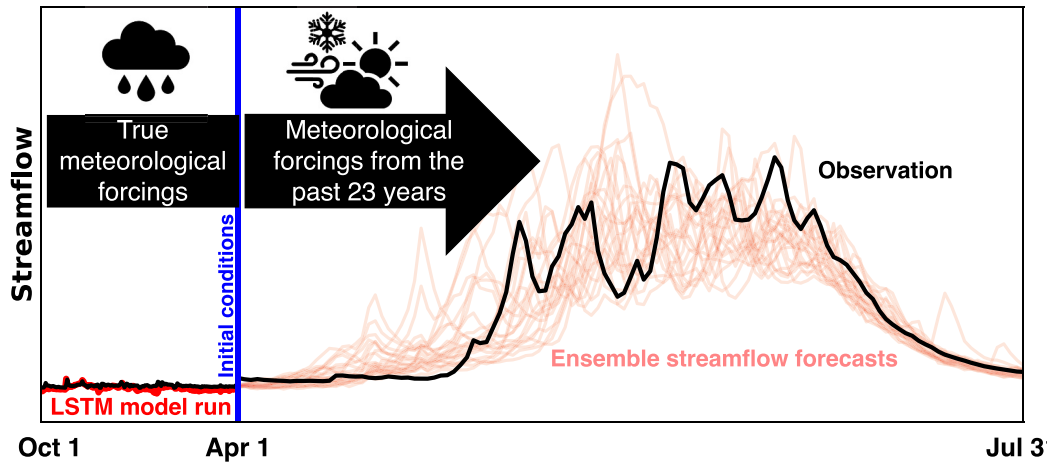


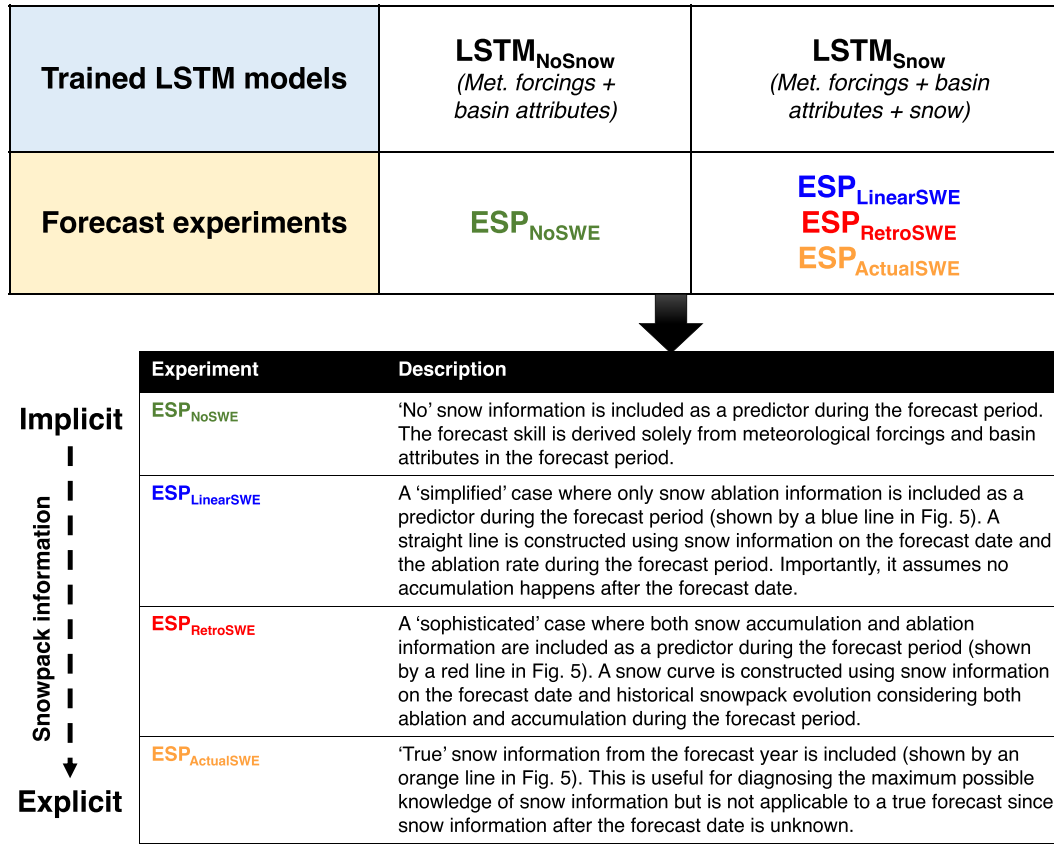
Figure 3. Illustration of an LSTM-ESP forecast issued on April 1. The thick red line depicts the Long Short-Term Memory model run before the forecast date using "true" meteorological forcings starting from October 1. Using the memory states on April 1 (shown in blue) and historical meteorological forcings from the past 23 years, ensemble streamflow forecasts are generated (shown in light red). Data based on USGS basin 13313000—Johnson Creek, ID for the forecast year 2011. It should be noted that the *x*-axis shown here is not uniform and represents the LSTM-ESP conceptually.

2.4. Forecast Experiments

We designed four forecast experiments to examine the performance of LSTM-ESP forecasts under varying levels of snowpack information. These experiments ranged from providing no direct snowpack information (*implicit*) to directly supplying true snowpack information (*explicit*). In the implicit setup, the LSTM received only meteorological observations, relying on the model's ability to essentially map inputs onto outputs, by inferring the impact of how snow dynamics inherently affect the target variable of streamflow. While using only meteorological forcings implicitly captures snowpack dynamics within LSTMs for streamflow prediction, ESPs would be limited by the lack of known snowpack information on and beyond the forecast date, which impacts their ability to represent hydrologic memory after the forecast date. Conversely, in the explicit setup, the LSTM was provided with direct snowpack data, enabling it to incorporate this information as an additional predictor. By including explicit snowpack information on and beyond the forecast date, we aim to provide LSTM-ESP with different degrees of snowpack information and a better representation of hydrologic memory. These include known snowpack information on the forecast date and assumptions about snow evolution after the forecast date as a way to boost the representation of hydrologic memory that is commensurate with the physical hydrological system.

From the start of the water year until the forecast date, for example, before April 1 (or Mar. 1 or Feb 1), all experiments use common predictors like "true" meteorological forcings and basin attributes, while the snow forecast experiments also use "true" snow information. Then, "during" the forecast period, that is, after April 1 (or Mar. 1 or Feb. 1), we use meteorological forcings from earlier historical years as is common in ESP, as well as basin attributes, and the snow experiments use three different treatments of snow as predictors, depending on the experiment. These treatments of snow combine the "true" snowpack conditions up to and including the forecast date, with snowpack data after the forecast date that come from the same historical years as the ESP meteorological forcing. These forecast experiments are defined based on the degree of historical snow information used—ESP_{NoSWE}, $ESP_{LinearSWE}$, $ESP_{RetroSWE}$, and $ESP_{ActualSWE}$. Figure 4 provides more details, and Figure 5a provides a graphical depiction of April 1 ESP forecasts. Importantly, only in the ESP_{NoSWE} experiment uses the trained $LSTM_{NoSnow}$ model, while the latter three experiments ($ESP_{LinearSWE}$, $ESP_{RetroSWE}$, and $ESP_{ActualSWE}$) utilize the $LSTM_{Snow}$ trained model.

The ESP_{NoSWE} experiment employs the $LSTM_{NoSnow}$ model (Section 2.3.1) and operates under the assumption that no historical snowpack information is available during the forecast period. This is shown in Figure 5a by the green zero SWE line that indicates no snow information was included during the forecast period. Snow treatment in this experiment is considered *implicit* because it may indirectly infer the impact of snow from the LSTM's "memory states" and relationships between streamflow and available historical meteorological forcings rather than directly incorporating snow as a predictor.



Experiment	Description
ESP_{NoSWE}	'No' snow information is included as a predictor during the forecast period. The forecast skill is derived solely from meteorological forcings and basin attributes in the forecast period.
$ESP_{LinearSWE}$	A 'simplified' case where only snow ablation information is included as a predictor during the forecast period (shown by a blue line in Fig. 5). A straight line is constructed using snow information on the forecast date and the ablation rate during the forecast period. Importantly, it assumes no accumulation happens after the forecast date.
$ESP_{RetroSWE}$	A 'sophisticated' case where both snow accumulation and ablation information are included as a predictor during the forecast period (shown by a red line in Fig. 5). A snow curve is constructed using snow information on the forecast date and historical snowpack evolution considering both ablation and accumulation during the forecast period.
$ESP_{ActualSWE}$	'True' snow information from the forecast year is included (shown by an orange line in Fig. 5). This is useful for diagnosing the maximum possible knowledge of snow information but is not applicable to a true forecast since snow information after the forecast date is unknown.

Figure 4. Chart of trained Long Short-Term Memory models and their corresponding forecast experiments, including names and descriptions. A graphical portrayal of the treatment of snow for each experiment is provided in Figure 5.

The $ESP_{LinearSWE}$, $ESP_{RetroSWE}$, and $ESP_{ActualSWE}$ forecast experiments employ the $LSTM_{Snow}$ model (Section 2.3.1) to assess the impact of differing levels of historical snowpack information on forecast performance. Explicitly using snow information during the forecast period means that the generated forecasts will not only depend on LSTM's memory states and historical meteorological forcings but also on the snow information provided during the forecast period in the form of a predictor. Just as ESP models do not know the true meteorology during the forecast period, they also do not know the true snow information (i.e., SWE values). Thus, we test multiple types of snow data to see how they affect forecast performance.

The $ESP_{LinearSWE}$ applies the known SWE information on the forecast date with average linear ablation rates computed based on historical data, representing a simplified baseline case. In the case of $ESP_{LinearSWE}$, for a given ensemble year, the ablation rate is estimated by taking the ratio of total ablation and the total number of days the ablation was greater than zero during the forecast period. These estimates are calculated for each ensemble year from UA snow data. A straight line is constructed for each ensemble year by using the SWE on the forecast date (from the forecast year) as an initial intercept and ablation rate (from the ensemble year) as the slope (shown by multiple blue lines in Figure 5a). The Linear SWE assumes that no accumulation happens after the forecast date and that ablation occurs at a constant rate making it highly simplified and it's use solely as a diagnostic tool to understand the potential value of limited snow information follows from earlier studies (Barnhart et al., 2016; Evan & Eisenman, 2021; Trujillo & Molotch, 2014). We expect the linear ablation assumption to be particularly disadvantageous in high snow regions with significant snow accumulation after the forecast date (e.g., after April 1 in high snow regions).

The $ESP_{RetroSWE}$ is more sophisticated, since it integrates the known SWE information on the forecast date (from the forecast year) with explicit accumulation and ablation rates after the forecast date from individual historical years (shown by multiple red lines in Figure 5a). The accumulation rate is estimated similarly to the ablation rate, that is, by taking the ratio of total accumulation and the total number of days where accumulation was greater than

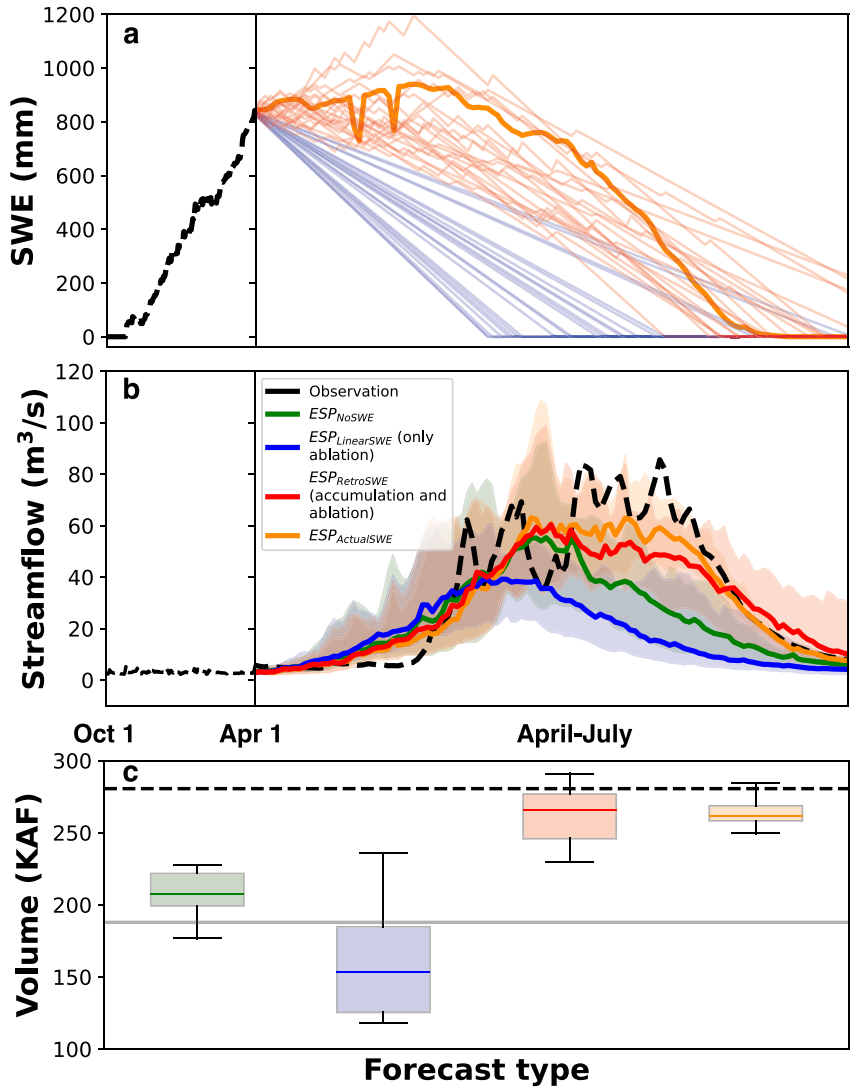


Figure 5. Demonstration of forecast experiment design: (a) historical snowpack information used as a predictor in the form of $ESP_{LinearSWE}$ (only ablation), $ESP_{RetroSWE}$ (accumulation and ablation), and $ESP_{ActualSWE}$, (b) demonstration of median Ensemble Streamflow Prediction (ESP) forecasts on April 1 corresponding to each forecast experiment, and (c) cumulative volumes from each ESP forecast with a dashed black line and a solid gray line representing observed and climatological volumes respectively. Data for this example is based on USGS basin 13313000—Johnson Creek, Idaho for the forecast year 2011.

zero during the forecast period. A snow curve is constructed for each ensemble year by using SWE on the forecast date as an initial intercept and then tracing the ensemble year's snow curve with accumulation and ablation rates. We trace this curve by systematically tracking changes in the snowpack, that is, when ablation happens, the ablation rate is applied, and when accumulation happens, the accumulation rate is applied. It should be noted that we cannot use the historical SWE curves directly since that would create a discontinuity in SWE values on the forecast date.

The $ESP_{ActualSWE}$ experiment uses the "true" snowpack evolution from the respective forecast year that occurred during the forecast period (solid orange line, Figure 5a). However, this experiment is not a forecast per se, because snowpack information beyond the forecast date would be unknown. In addition, unlike the $ESP_{LinearSWE}$ and $ESP_{RetroSWE}$ experiments, the $ESP_{ActualSWE}$ experiment presents a single SWE curve (from the forecast year) as a predictor during the forecast period alongside common predictors regardless of the ensemble year.

We employ these treatments of snow information in the forecast experiments to determine the extent to which they enhance or degrade forecasts with respect to the ESP_{NoSWE} experiment. The daily ESP forecasts generated from these forecast experiments are illustrated in Figure 5b alongside the observed streamflow. The forecasted daily streamflow from each experiment is further cumulated to seasonal volumes (i.e., April–July—evaluation period) and used for comparison with operational forecasts (Figure 5c). It is important to note that the purpose of Figure 5c is solely to illustrate how the AMJJ volume changes with varying levels of snowpack information. The evaluation of forecast performance is addressed in detail in later Sections 3.2–3.4. Additionally, these forecast volumes are shown for the year 2011, which experienced much above-normal snowpack, characterized by late snowmelt and elevated streamflow. This hydrological anomaly (a) explains the deviation between the observed streamflow and the climatological average, and (b) may also account for why most of the models underpredicted streamflow in Figure 5c.

2.4.1. Forecast Comparison

We use Natural Resources Conservation Services (NRCS) statistical forecasts as a proxy for operational forecasts over the study watersheds for benchmarking purposes. These forecasts were chosen since they are methodologically consistent across all regions and easily accessible for a larger number of basins and years. The NRCS employs a Principal Component Regression model. This model is usually modified to retain the principal components (Garen, 1992; Lehner et al., 2017) and uses predictors like SWE and accumulated precipitation from SNOTEL and antecedent streamflow USGS to predict seasonal streamflow volumes (i.e., April–July in this study).

NRCS forecasts include five forecasted exceedance probabilities at 90%, 70%, 50%, 30%, and 10%. To clarify, 90% means there is a 90% chance that the observed streamflow volume will exceed this forecast value and a 10% chance that it will be less than this forecast value. They are generated for multiple forecast points and different evaluation periods (e.g., Apr–Jul, Mar–June, May–Aug, etc.). In order to make LSTM-ESP forecasts comparable, the same five probabilities of exceedance were obtained from the LSTM-ESP ensemble.

2.5. Performance Metrics

The Nash Sutcliffe Efficiency (NSE) was used to quantify streamflow prediction accuracy of the different LSTM models. The NSE ranges from negative infinity to 1, with 1 indicating perfect agreement between the simulated and observed values, and values closer to 0 indicating poorer performance.

We employed three forecast performance metrics, drawing from those widely adopted operationally and recently used in the Bureau of Reclamation's WSF challenge (DrivenData, 2023). The Normalized Root Mean Square Error (NRMSE, in %) was used to analyze the skill of seasonal volumes from the forecasts against the corresponding observed streamflow volumes. The RMSE was normalized by the median of observed streamflow volumes and values closer to 0 indicate better forecast performance.

Internal Coverage refers to the proportion of the observed values that fall within the ensemble. This metric determines whether the ensemble spread adequately captures the expected interannual variability in the forecasts or if it is too narrow. Well-calibrated forecasts are expected to have a coverage value of 0.8 or higher to ensure forecast uncertainty is well-represented (DrivenData, 2023). The equation for interval coverage is shown in Equation 1 as:

$$\mathrm{IC} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\left(\hat{y}_{0.1,i} \le y_{obs} \le \hat{y}_{0.9,i}\right)} \tag{1}$$

Where $\mathbb{1}$ is an indicator function that assigns a value of 1 if the condition is satisfied or 0 otherwise. y_i represents the actual observed value for the ith observation, $\hat{y}_{0.1,i}$ and $\hat{y}_{0.9,i}$ represents the predicted value for the 0.1 and 0.9 quantiles of the ith observation and n represents the total number of observations.

An additional NRMSE skill score (SS_{NRMSE}) quantifies forecast skill relative to the skill of a reference climatological forecast, which is the mean seasonal observed volume from WY1982–2005. The climatology would have a skill score of 0, whereas a perfect forecast would result in a skill score of 1 (Equation 2).

$$SS_{NRMSE} = 1 - \frac{NRMSE_{forecast}}{NRMSE_{climatology}} \quad (2)$$

Lastly, a one-sided Mann-Whitney U test was also performed to test whether the performance from different forecast experiments is equal for any forecast metric. The non-parametric hypothesis test was chosen over a parametric Student's paired *t*-test as it performs well with non-normally distributed data. Statistical significance was reported at the 95% confidence level ($p \leq 0.05$).

3. Results

We first compare the historical model performance from both LSTM models with observations, as well as relative to the performance of the K19 model (Section 3.1). Following this, we examine the forecast skill from LSTM-ESP forecast experiments, initially comparing among themselves on April 1 (Section 3.2) and, subsequently, comparing selected forecast experiments with NRCS forecasts issued on April 1 (Section 3.3). We lastly compare this forecast performance across different lead times alongside NRCS (Section 3.4).

3.1. Evaluating Historical LSTM Model Performance: The Impact of Snow on Streamflow Prediction Accuracy

We first compared the daily NSE (through the entire year) of $LSTM_{NoSnow}$ and $LSTM_{Snow}$ models with the K19 model during the testing period, WY2001–2010, which is separate from the training period, WY1983–2000. Approximately 80% of the 493 common basins with the K19 study showed an average daily NSE of 0.5 or greater with both models (Figures 6a and 6b). Both LSTM models ($LSTM_{NoSnow}$ and $LSTM_{Snow}$) showed statistically higher NSE (median—0.72, 0.74) than K19 model (median—0.70) based on a one-sided Mann-Whitney U test ($p \leq 0.05$). These modest performance gains may be a result of adding more basins in the model training (531 basins in K19 vs. 664 basins in this study). Furthermore, we assessed the NRMSE of total April–July streamflow volumes, accumulated from daily streamflow. Approximately 50% of basins showed an NRMSE of 25% or less for both LSTM models (Figures 6c and 6d), and only the $LSTM_{Snow}$ model (median—23%) showed statistically lower NRMSE compared to the K19 model (median—25%) based on a one-sided Mann-Whitney U test ($p \leq 0.05$).

3.2. Intercomparison of Skill From Forecast Experiments on April 1

April 1 forecast skill was evaluated across 76 basins during the period WY2011–2020. These 76 basins were categorized into low (SWE/P < 0.4), moderate (0.4 ≤ SWE/P ≤ 0.8), and high snow regions (SWE/P > 0.8). Figure 7 compares NRMSE, Internal Coverage, and SS_{NRMSE} across the range of April 1 SWE/P ratios. Surprisingly, the ESP_{NoSWE} experiment showed reasonable NRMSE in high snow regions (median—21%), though it performed less effectively in regions with low (median—28%) to moderate snow (median—32%). However, lower Internal Coverage (median ∼ 0.4 across all categories) and lower skill scores (median ∼ 25% across all categories) suggest ESP_{NoSWE} does not adequately capture the variability in the forecasts.

Conversely, the $ESP_{LinearSWE}$ experiment, which relies on simplified ablation rates, showed poor performance in the high snow regions, with median NRMSE reaching up to 30% and median Internal Coverage of 0.2. This finding highlights the negative impact of using a simplified representation of snow on the forecast skill in high snow regions. However, $ESP_{LinearSWE}$ demonstrated better NRMSE in low (median—21%) to moderate (median—24%) snow regions, most likely attributable to the majority of snow accumulation occurring before April 1, making ablation information more sufficient for forecasting purposes. These improvements in low and moderate snow regions are evidenced by better Internal Coverage and skill scores compared to ESP_{NoSWE}.

In colder regions like the Idaho Batholiths, Middle Rockies, and Wasatch and Uinta Mountains (Figure 1), significant snow accumulation occurs for up to 40 days after April 1 (Musselman et al., 2017). The $ESP_{RetroSWE}$ experiment, which relies on both historical ablation and accumulation rates, showed a median NRMSE of 15% and skill score of 50%, an improvement in high snow regions compared to ESP_{NoSWE} and $ESP_{LinearSWE}$.

As expected, the $ESP_{ActualSWE}$ experiment, which uses true snow information during the forecast period, generally showed lower NRMSE (median—18%, 22%, 17%) and higher skill scores (median—55, 59, 45) across three SWE/P categories (Figures 7a and 7c) but demonstrated a lower Internal Coverage (median—0.5, 0.4, 0.3)

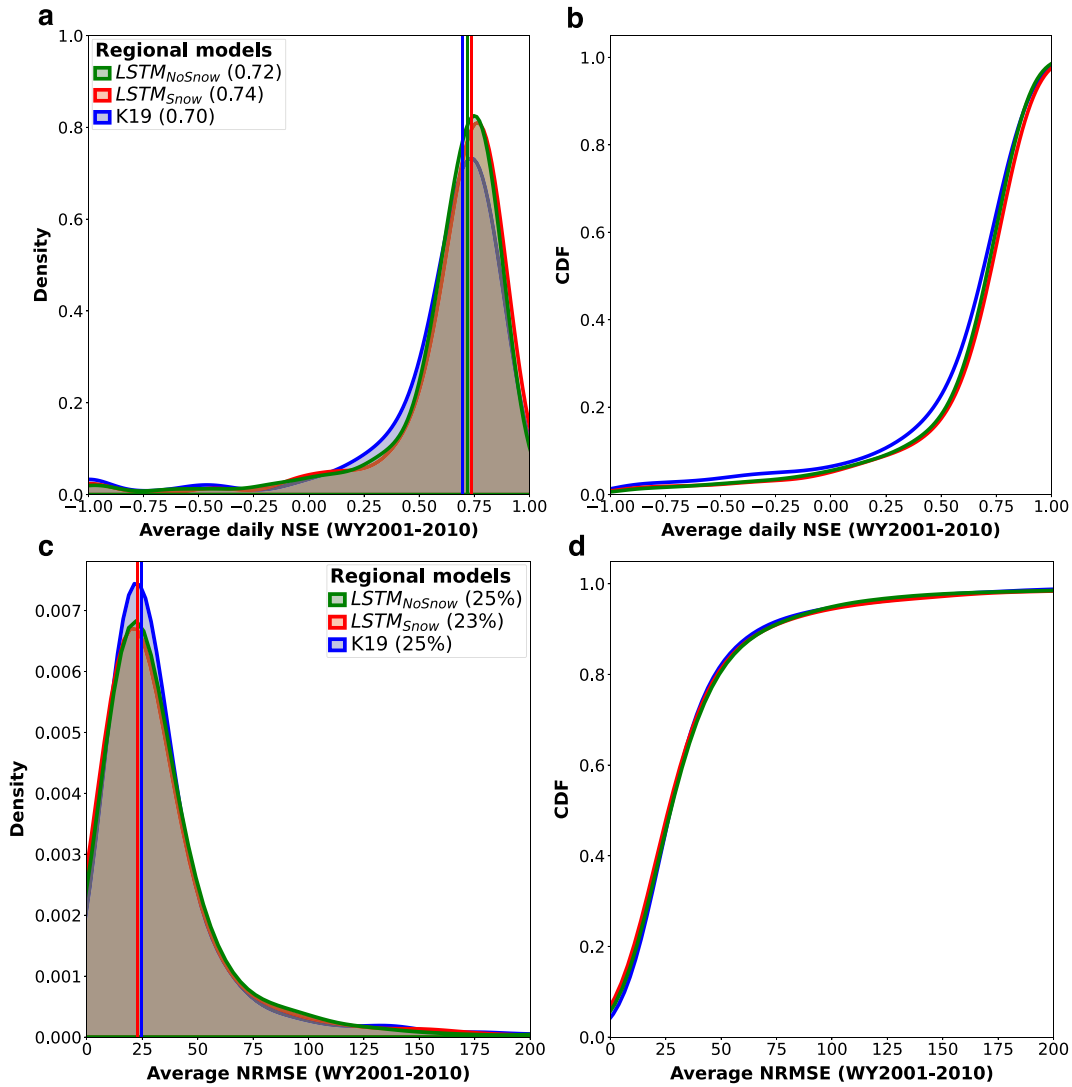


Figure 6. (a–b) Daily Nash-Sutcliffe Efficiency and (c–d) Normalized Root Mean Square Error of the total April–July streamflow volumes compared through the entire year during the testing period, WY2001–2010. Comparison shown for the common 493 basins for $LSTM_{NoSnow}$ and $LSTM_{Snow}$ models alongside the K19 model. The brackets in the legend indicate median values for the respective metric.

in general. This discrepancy arises because, in the $ESP_{ActualSWE}$ experiment, the SWE remains constant across all ensemble years throughout the forecast period (Figure 5a and as described in Section 2.4). It should be noted that this is not the case in $ESP_{LinearSWE}$ and $ESP_{RetroSWE}$ experiments where the SWE varies across ensemble members (Figure 5a).

In general, the interannual variability in the forecast was best captured by the $ESP_{RetroSWE}$ (Figure 7b), and it performed statistically better ($p \leq 0.05$) than ESP_{NoSWE} and $ESP_{LinearSWE}$ experiments across all SWE/P categories. Overall, this comparison demonstrated the importance of the quality of snowpack information in generating LSTM-ESP streamflow forecasts across the western US.

3.3. Comparison of LSTM-ESP Forecasts With NRCS Forecasts on April 1

As shown in Section 3.2, the $ESP_{RetroSWE}$ experiment yielded overall better forecast performance across snow-based experiments and was chosen for further comparison to benchmark against ESP_{NoSWE} and NRCS forecasts. ESP_{NoSWE} and $ESP_{RetroSWE}$ exhibited NRMSE (median across all SWE/P categories—25%, 20%) and SS_{NRMSE} (median across all SWE/P categories—29, 50) comparable to the NRCS forecasts (NRMSE—18%, SS_{NRMSE} -

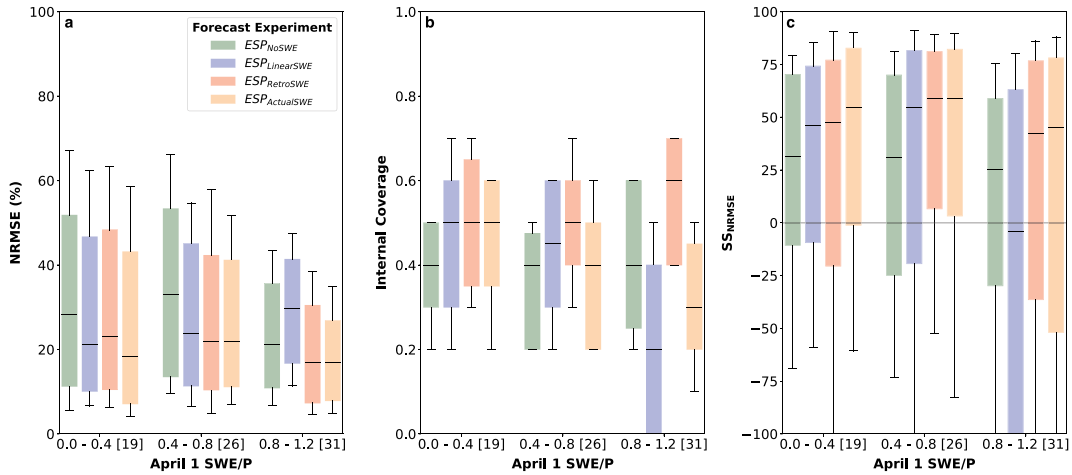


Figure 7. April 1 forecast performance over the period WY2011–2020 from all forecast experiments (ESP_{NoSWE}, $ESP_{LinearSWE}$, $ESP_{RetroSWE}$, and $ESP_{ActualSWE}$) across 76 basins for the (a) Normalized Root Mean Square Error, (b) Internal Coverage, and (c) SS_{NRMSE}. The boxplot whiskers represent the 15th and 85th percentiles, and the square brackets (*x*-axis) indicate the number of basins in each SWE/P category.

40), with the $ESP_{RetroSWE}$ experiment performing statistically similar ($p \leq 0.05$) to NRCS across all SWE/P categories (Figures 8a and 8c). The ESP_{NoSWE} performs reasonably well but demonstrates higher NRMSE and lower skill scores than the NRCS and $ESP_{RetroSWE}$ forecasts. As shown in Figure 8b, the NRCS demonstrates a wider ensemble spread (i.e., high Internal Coverage) than both LSTM-ESP forecasts, that is, NRCS better captures the observations within the forecast range. This is attributable to number of ensemble years, architectural differences and input data between the models, as well as subjective manual hydrologist intervention to the NRCS forecasts (Fleming et al., 2021; Garen, 1992). We also compared the $ESP_{RetroSWE}$ and NRCS forecasts spatially. The performance of the $ESP_{RetroSWE}$ was generally similar (white dots in Figures 8d and 8f) to the NRCS, albeit with some instances of mixed outcomes (red and blue dots in Figures 8d and 8f). Moreover, the $ESP_{RetroSWE}$ illustrated a narrow ensemble spread across most of the basins (red dots in Figure 8e).

Importantly, we investigated whether the snow-based LSTM-ESP findings were specific to the UA SWE product or not. To this end, we repeated all snow-based forecast experiments ($ESP_{LinearSWE}$, $ESP_{RetroSWE}$, and $ESP_{ActualSWE}$) using SNOTEL observations for each basin instead of UA snow while keeping all other aspects of the experiment identical. The resulting comparison is provided in Text S1 in Supporting Information S1. Similar forecast skill responses were seen overall (Figure S2 in Supporting Information S1), without any clear regional pattern. Limited basin-wise disparities were seen and are likely attributable to inherent differences between station-based snow measurements and gridded data. Overall, this comparison suggests the findings reported here would not be quantitatively different if using either UA or SNOTEL snow data.

3.4. Comparison of LSTM-ESP Forecasts With NRCS Forecasts Across Different Lead Times

Given the interest from water resource managers and government agencies in assessing skill across different forecast lead times, we compared the ESP_{NoSWE} and $ESP_{RetroSWE}$ experiments to the NRCS forecasts at three distinct lead times (i.e., Feb 1, Mar 1, and Apr 1). Consistent with our expectations, the performance (i.e., NRMSE and SS_{NRMSE}) from both LSTM-ESP forecasts as well as NRCS usually improved up to 15% from longer (February) to shorter (April) lead times (Figures 9a and 9c). The $ESP_{RetroSWE}$ forecast showed statistically better performance (i.e., NRMSE and SS_{NRMSE}—$p \leq 0.05$) than both NRCS and ESP_{NoSWE} forecasts (Figures 9a and 9c), particularly across February and March whereas the NRCS showed statistically better Internal Coverage ($p \leq 0.05$) than both LSTM-ESP forecasts (Figure 9b). Commensurate with our earlier findings, we see lower Internal Coverage from both LSTM based ESPs across all lead times. The disparities in forecast metrics were largely consistent, showing improvement at shorter lead times. However, the ESP_{NoSWE}'s Internal Coverage on April 1 was dramatically smaller compared to Feb 1 and Mar 1 forecasts, possibly due to lack of explicit snowpack information. This analysis was constrained to only 53 out of the 76 basins due to the availability of NRCS forecasts across the different lead times.

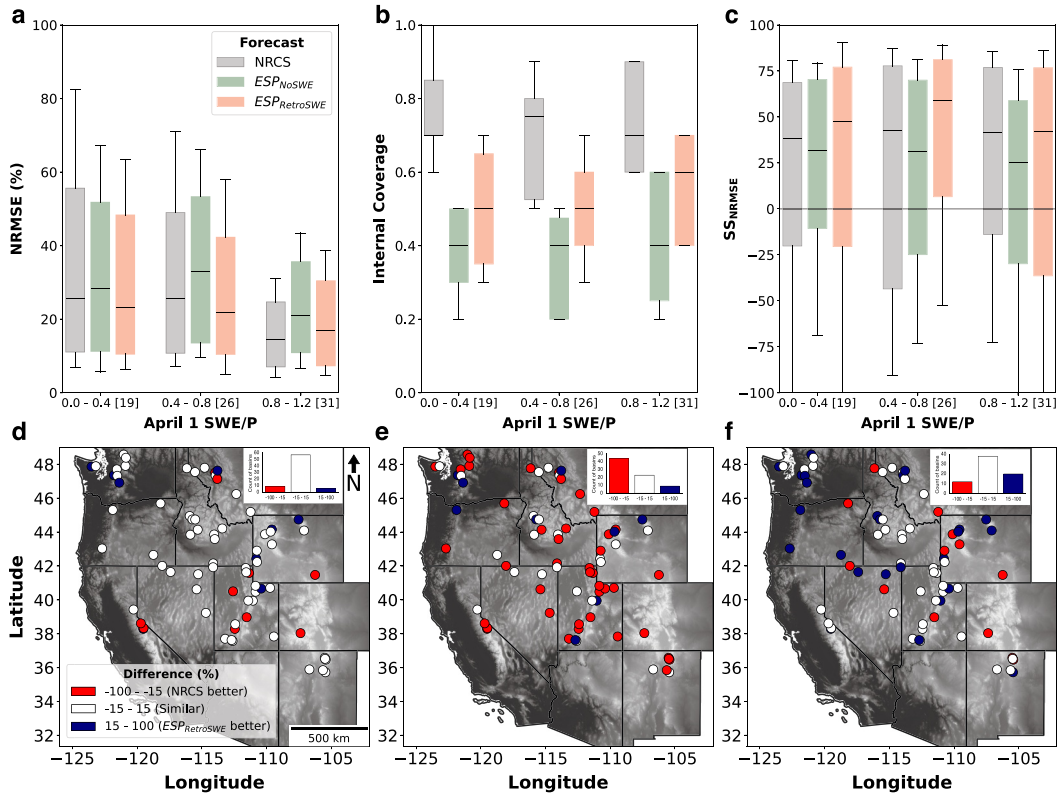


Figure 8. Normalized Root Mean Square Error, Internal Coverage, and SS_{NRMSE} compared on April 1 across WY2011–2020 from ESP_{NoSWE} and $ESP_{RetroSWE}$ experiments and Natural Resources Conservation Services (NRCS) across 76 basins (a–c). The spatial plots (d–f) represent the difference between $ESP_{RetroSWE}$ and NRCS forecasts for each performance metric. The boxplot (a–c) whiskers represent the 15th and 85th percentiles, and the square brackets (*x*-axis) indicate the number of basins in each SWE/P category.

4. Discussion

This study was motivated by recent literature showing that LSTM performance could consistently exceed that of process-based models (Ng et al., 2023). Our analysis demonstrated the viability of LSTM in an ESP framework. LSTM-ESP forecasts were developed by leveraging LSTM's regionalization and structural abilities using a combination of predictors, including historical meteorological forcings, basin attributes, and snowpack

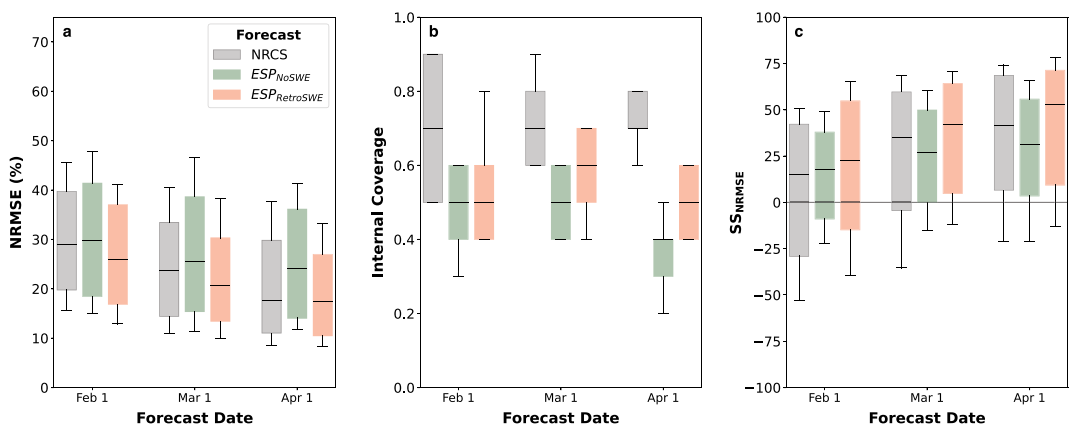


Figure 9. Normalized Root Mean Square Error (NRMSE), Internal Coverage, and SS_{NRMSE} compared at different lead times in WY2011–2020 from Natural Resources Conservation Services, ESP_{NoSWE} and $ESP_{RetroSWE}$ forecasts across 53 basins. The whiskers on the box plots represent the 15th and 85th percentiles.

information. Two LSTM models (i.e., $LSTM_{NoSnow}$ and $LSTM_{Snow}$) were trained regionally by maximizing available training inputs across the CONUS, resulting in better predictor mapping across the different hydrological processes. Using regionally trained LSTM models provided flexibility in application compared to a process-based model that requires basin-wise parameter calibration or regionalization using donor basin in the case of an ungauged basin (Arsenault et al., 2022).

The model performance from the $LSTM_{NoSnow}$ and $LSTM_{Snow}$ models was compared with the K19 model, given the similarities of data sets used in both studies and to test the performance of our chosen underlying model architecture. Despite incorporating a larger number of basins into training both LSTM models compared to K19, only minor performance differences were observed with the generated hindcasts (Figure 6). These differences also meant that including snow as an additional predictor did not significantly improve LSTM model performance. This finding is attributed to both the utilization of a sufficiently diverse set of basins with varying hydroclimatological regimes in the K19 model and inclusion of highly sensitive basin traits that exert a significant influence on the streamflow, for example, drainage area, average basin precipitation—Ciulla & Varadharajan, 2023), in our model and in the K19 model.

We designed the forecast experiments to examine the overall capability of LSTM in an ESP framework and to assess the impact of different degrees of snowpack information on forecast skill. Notably, the ESP_{NoSWE} forecast exhibited better performance in high snow regions but showed reduced performance in low and moderate snow regions (Figure 7). This may suggest that the implicit role of snow was only captured by the ESP_{NoSWE} forecast for regions where it had a more consistent effect on streamflow. On the contrary, the $ESP_{LinearSWE}$ experiment led to lower performance in high snow regions while still providing reasonable skill in low and moderate snow regions. This disparity between ESP_{NoSWE} and $ESP_{LinearSWE}$ highlights the fact that omitting snowpack information may be better than incorporating simplified representations of snow during the forecast period, particularly in the high snow regions. Given the reasonable skill demonstrated by the ESP_{NoSWE} experiment, it could serve as a viable forecasting alternative in situations where snow observations are scarce or unavailable. Subsequently, the $ESP_{RetroSWE}$, accounting for both accumulation and ablation during the forecast period demonstrated improved performance across all regions in comparison to both ESP_{NoSWE} and $ESP_{LinearSWE}$ experiments (Figure 7). The $ESP_{ActualSWE}$ experiment did not have a large enough ensemble spread to consistently capture the observations (as explained in Sections 2.4 and 3.2), even though it did provide a good benchmark for comparing the other LSTM-ESP forecast experiments. Overall, these various forecast experiments demonstrated and confirmed that forecast skill is significantly influenced by the level of "known" predictor information during the forecast period (i.e., snowpack in this study). However, disparities between forecasts (ESP_{NoSWE} and $ESP_{LinearSWE}$) also suggest that snow data does not consistently improve LSTM-ESP performance.

All the forecast experiments perform slightly worse in low and moderate snow regions compared to high snow regions. Sparse precipitation and snowpack observations in high-elevation regions (Henn et al., 2018) and minimal snowmelt contribution to streamflow hinder the model's ability to predict streamflow accurately (Kratzert et al., 2018). It should be noted that the small sample size of basins raises the possibility that the results might be due to random variation, as suggested by the law of small numbers. Furthermore, streamflow in these regions may be influenced by factors (e.g., climate indices) not captured well by the predictors optimized for snow-dominated basins used in this study, highlighting the need for future testing to better understand these dynamics. However, as the LSTM model is not process-based, these differences observed in this study cannot be attributed to specific physical mechanisms. One way we address this limitation is by training the model over hydro-climatologically diverse basins across the CONUS, aiming to improve its reliability in low and moderate regions. However, region-specific adjustments (Hosseini et al., 2024) or grouping-based training strategies (Yu et al., 2024) were not performed here, although these might further improve model predictions in low and moderate snow regions.

To analyze the potential utility of LSTM-based ESP forecasts, we compared the ESP_{NoSWE} and $ESP_{RetroSWE}$ against the official NRCS forecasts. We found that $ESP_{RetroSWE}$ forecast showed comparable performance to NRCS but lacked a wider ensemble spread as compared to the NRCS forecasts (Figure 8b). This meant that the ensemble spread from these forecasts did not capture the interannual variability during the WY2011–2020 as well as the NRCS forecasts. These differences arise from disparities in model structures, input data, number of ensemble years, and, in particular, the manual forecast intervention noted in Section 3.3 (Fleming et al., 2021; Garen, 1992) to expand ensemble spread in the NRCS forecasts. However, it should be noted that one of the

disadvantages of a wider ensemble spread is greater uncertainty levels and low confidence during decision-making. To verify the robustness of the snow-based ESP experiments, we reran our experiments using the SNOTEL data set instead of the UA data set, which are in situ observations assimilated as part of the gridded UA data set. Despite exhibiting some basin-wise differences, the SNOTEL data set showed a similar response to the UA data set (Figure S2 in Supporting Information S1). This confirms that the overall findings of the paper were qualitatively similar, irrespective of the snow product used. Lastly, we tested the ESP_{NoSWE} and $ESP_{RetroSWE}$ forecasts at two additional lead times (i.e., Feb-1 and Mar-1 in this study). The forecast differences from LSTM-ESP experiments were largely consistent at different lead times except for the Internal Coverage, which showed a dramatic decline on April 1 with respect to Feb. 1 and Mar 1 forecasts (Figure 9). It is important to recognize that the reasons why certain types of snow treatments perform better than others still remain uncertain and require further diagnosis. Given that these data are integrated into LSTM's hidden state, which some propose serves as an analog for storage (Kratzert et al., 2018), it is not possible to conclusively determine whether the model is improving its hydrologic representation of snow processes via different treatments of snow.

A number of limitations should be noted with this study. We recognize that a comparison with process-based ESP forecasts generated by the River Forecast Centers (RFC) might be regarded as more appropriate for this study. However, due to the limited availability of operational ESP forecasts (starting in 2015) for our study basins, as well as inconsistent methodologies across regions, we chose to use the NRCS forecasts. These inconsistent methodologies often arise due to technical heterogeneity (e.g., the use of different hydrologic models), stakeholder needs, and challenges due to diverse regional differences (NOAA, 2024). Frameworks such as the Community Hydrologic Prediction System were proposed to enhance operational forecasting by offering modular, flexible, scalable workflows for integrating forecasts across diverse hydrologic conditions (Restrepo et al., 2010). Importantly, it should be noted that the differences in forecast volumes between NRCS and RFC results are minor in the context of the overall forecast uncertainty (Lukas & Payton, 2020). Development of regional data sets for large-scale hydrological analysis, such as CAMELS (Addor et al., 2017; Newman et al., 2014), CONUS404 (Rasmussen et al., 2023), and Caravan (Kratzert et al., 2023) have been useful in producing deep learning forecasts. However, these data sets typically have their own limitations, including a lack of common standards for intercomparison, a lack of uncertainty estimates for assessing data reliability, and a lack of characterization of human intervention (Addor et al., 2020). Recent literature has also tackled the question of physical interpretability and time-tracking diagnostic information by employing physics-guided or theory-guided ML architectures. For instance, Feng et al. (2022a, 2022b) exhibited a differentiable, learnable, process-based model that approached the performance of the LSTM model and also provided the outputs of untrained variables like soil moisture, groundwater storage, and snowpack that can be useful for diagnosis and understanding the physics from big data. Recent advances in LSTM architecture, for example, by Hoedt et al. (2021), is designed to follow mass conservation laws, providing a solution to overcome the lack of physical constraints. While such deep-learning architectures improve interpretability and respect physical laws, they still do not reach the level of model process descriptions or model physical structures (Feng, Liu, et al., 2022).

With respect to the experimental design, there exist additional limitations. We only tested a single deep learning model—LSTM, which could be replaced by other neural networks like Gated Recurrent Units (GRUs; Cho et al., 2014), transformers (Vaswani et al., 2017) or physics-guided ML architecture (Feng, Beck, et al., 2022; Feng, Liu, et al., 2022; Hoedt et al., 2021). Subsequently, testing different sets of LSTM hyperparameters, higher resolution meteorological forcings (e.g., CONUS404), a larger training period, and including only snow-dominated basins might show improvement in the overall LSTM model performance and corresponding ESP forecasts. Despite assessing the impact of different snow products, we only tested three treatments of snow (Linear, Retrospect, and Actual), and there are several other ways the snowpack information can be used to guide such ESP forecasts.

Furthermore, the $ESP_{LinearSWE}$ and $ESP_{RetroSWE}$ experiments contain simplifications that limit their skill in comparison to the $ESP_{ActualSWE}$ case. First, they use "estimated" snow information in the forecast experiments, which does not have the same range of variability and will have different properties than the full complexity of snowpack dynamics. Second, the Snow LSTM model was trained using "true" snow information and it did not contain the same types of estimates used in some of the forecast experiments, potentially leading to gaps in predictive accuracy. The impact of these constraints could be evaluated by training additional LSTM-ESP models using estimated snow information. The relatively large computational effort to conduct such a retraining falls outside the scope of this analysis.

Future avenues for wide acceptance and viability of such deep learning based seasonal streamflow forecasts include a systematic comparison with operational models (Mai et al., 2022) and the use of physics-guided or theory-guided DL approaches (Feng, Beck, et al., 2022; Feng, Liu, et al., 2022; Hoedt et al., 2021; Shen et al., 2023). The potential for enhancing the hydrological interpretability of deep-learning models is evident, offering significant scope for the advancement of streamflow forecasting through architectural modifications (De La Fuente et al., 2023) and the development of hybrid models (Feng, Liu, et al., 2022; Slater et al., 2023).

5. Conclusions

This study developed deep learning based ESP streamflow forecasts using the LSTM model across western US basins with the aim to evaluate the potential utility of these emerging techniques in generating seasonal streamflow forecasts. The bulk of our analysis used two LSTM models: one that did not use snow as a predictor ($LSTM_{NoSnow}$), and another that used snow as a predictor ($LSTM_{Snow}$), both trained regionally. This regional training was done to maximize the information available for training and produce computationally efficient ESP forecasts. The performance of these LSTM models was evaluated and compared to a published model, K19 (Kratzert, Klotz, Herrnegger, et al., 2019), owing to the similarity of the data sets and similar model architecture. The comparison showed comparable performance, where approximately 80% of the tested basins showed an NSE of 0.5 or more, and 50% showed an NRMSE of 25% or less, confirming our LSTM would have satisfactory utility for generating seasonal streamflow forecasts. Notably, the inclusion of snow as a predictor, in addition to the meteorological forcings and basin attributes, did not significantly improve the model performance, demonstrating the LSTM's abilities to incorporate snowpack information implicitly based on the information provided.

The forecast experiments were designed to explore the viability of LSTM in an ESP framework and assess the impact of snowpack information in guiding these LSTM-ESP forecasts. These forecast experiments were compared internally and with the NRCS operational forecasts that are widely used in operational settings. In general, the results showed that the LSTM-ESP forecast, explicitly incorporating previous years' snow accumulation and ablation ($ESP_{RetroSWE}$), showed comparable performance to the NRCS operational forecasts with the exception of lacking a comparably wide ensemble spread. However, particularly in high snow regions, ESP forecasts incorporating a simplified representation of snowpack information ($ESP_{LinearSWE}$), which only included ablation information, performed poorer as compared to those excluding snowpack information entirely (ESP_{NoSWE}). This disparity between $ESP_{LinearSWE}$ and ESP_{NoSWE} also suggest that snow data does not consistently improve the performance of LSTM-ESP forecasts.

A consistent trend was observed with LSTM ESPs across all lead times (February–April), with performance showing improvement at shorter lead times. The disparities and biases between LSTM-ESP forecasts and NRCS operational forecasts stem from differences in model architectures, input data, the number of ensemble years, and manual adjustments. Overall, this study underscores the prospective utility of employing deep learning models to generate seasonal streamflow forecasts and how information, such as those from snow, can influence forecast skill in such predictions.

Data Availability Statement

All data products used in the analysis are publicly available. A total of 664 GAGES-II basins are selected following screening criteria that ensure minimal upstream regulation and continuous data availability for at least 30 years. The meteorological forcings, basin attributes, snow and streamflow data are obtained from NLDAS2 (Xia et al., 2009, 2012), GAGES-II (U.S. Geological Survey, 2023), UA (Broxton et al., 2019) and the US Geological Survey streamflow gages (United States Geological Survey, 2024) respectively. NRCS forecast data and SNOTEL snowpack observations are downloaded from the National Water and Climate Center portal (United States et al., 2024). The Modi and Livneh (2024) data set provides the source code, training data, and model runs for the LSTM model used in this research.

Acknowledgments
We acknowledge funding support from the NOAA Grant NA20OAR4310420 Identifying Alternatives to Snow-based Streamflow Predictions to Advance Future Drought Predictability and NSF Grant BCS 2009922 Water-Mediated Coupling of Natural-Human Systems: Drought and Water Allocation Across Spatial Scales. We gratefully acknowledge constructive and insightful suggestions from two anonymous reviewers, which improved the clarity of the original manuscript. We are extremely grateful for all the computing resources provided by the University of Colorado Boulder Research Computing.

References

Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K., & Mendoza, P. A. (2020). Large-sample hydrology: Recent progress, guidelines for new datasets and grand challenges. *Hydrological Sciences Journal*, *65*(5), 712–725. https://doi.org/10.1080/02626667.2019.1683182

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Amnatsan, S., Yoshikawa, S., & Kanae, S. (2018). Improved forecasting of extreme monthly reservoir inflow using an analogue-based forecasting method: A case study of the Sirikit Dam in Thailand. *Water*, *10*(11), 1614. https://doi.org/10.3390/w10111614

Arheimer, B., Pimentel, R., Isberg, K., Crochemore, L., Andersson, J. C. M., Hasan, A., & Pineda, L. (2020). Global catchment modelling using World-Wide HYPE (WWH), open data, and stepwise parameter estimation. *Hydrology and Earth System Sciences*, *24*(2), 535–559. https://doi.org/10.5194/hess-24-535-2020

Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F., & Mai, J. (2022). *Continuous streamflow prediction in ungauged basins: Long Short-Term Memory Neural Networks clearly outperform hydrological models* (preprint). *Hydrometeorology/Modelling approaches*. https://doi.org/10.5194/hess-2022-295

Ayzel, G., Kurochkina, L., Kazakov, E., & Zhuravlev, S. (2020). Streamflow prediction in ungauged basins: Benchmarking the efficiency of deep learning. *E3S Web of Conferences*, *163*, 01001. https://doi.org/10.1051/e3sconf/202016301001

Barnhart, T. B., Molotch, N. P., Livneh, B., Harpold, A. A., Knowles, J. F., & Schneider, D. (2016). Snowmelt rate dictates streamflow. *Geophysical Research Letters*, *43*(15), 8006–8016. https://doi.org/10.1002/2016GL069690

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166. https://doi.org/10.1109/72.279181

Broxton, P., Zeng, X., & Dawson, N. (2019). Daily 4 km gridded SWE and snow depth from assimilated in-situ and modeled data over the conterminous US, version 1 [Dataset]. *NASA National Snow and Ice Data Center Distributed Active Archive Center*. https://doi.org/10.5067/0GGPB220EX6A

Chiew, F. H. S., Zhou, S. L., & McMahon, T. A. (2003). Use of seasonal streamflow forecasts in water resources management. *Journal of Hydrology*, *270*(1–2), 135–144. https://doi.org/10.1016/S0022-1694(02)00292-5

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. (Version 2). *arXiv*. https://doi.org/10.48550/ARXIV.1409.1259

Chollet, F., & Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.

Ciulla, F., & Varadharajan, C. (2023). *A network Approach for multiscale catchment Classification using traits* (preprint). *Catchment hydrology/Modelling approaches*. https://doi.org/10.5194/egusphere-2023-1675

Crochemore, L., Ramos, M.-H., & Pappenberger, F. (2016). Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrology and Earth System Sciences*, *20*(9), 3601–3618. https://doi.org/10.5194/hess-20-3601-2016

Daly, S. F., Davis, R., Ochs, E., & Pangburn, T. (2000). An approach to spatially distributed snow modelling of the Sacramento and San Joaquin basins, California. *Hydrological Processes*, *14*(18), 3257–3271. https://doi.org/10.1002/1099-1085(20001230)14:18<3257::aid-hyp199>3.0.co;2-z

Day, G. N. (1985). Extended streamflow forecasting using NWSRFS. *Journal of Water Resources Planning and Management*, *111*(2), 157–170. https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157)

DeChant, C. M., & Moradkhani, H. (2011). Improving the characterization of initial condition for ensemble streamflow prediction using data assimilation. *Hydrology and Earth System Sciences*, *15*(11), 3399–3410. https://doi.org/10.5194/hess-15-3399-2011

De La Fuente, L. A., Ehsani, M. R., Gupta, H. V., & Condon, L. E. (2023). Towards interpretable LSTM-based modelling of hydrological systems. https://doi.org/10.5194/egusphere-2023-666

Dingman, S. L. (2002). *Physical hydrology* (2nd ed.). Prentice Hall.

DrivenData. (2023). Water supply forecast Rodeo: Forecast stage. Retrieved from https://www.drivendata.org/competitions/259/reclamation-water-supply-forecast/

Evan, A., & Eisenman, I. (2021). A mechanism for regional variations in snowpack melt under rising temperature. *Nature Climate Change*, *11*(4), 326–330. https://doi.org/10.1038/s41558-021-00996-w

Falcone, J. A. (2011). GAGES-II: Geospatial attributes of gages for evaluating streamflow. Retrieved from https://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml

Falcone, J. A., Carlisle, D. M., Wolock, D. M., & Meador, M. R. (2010). GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States. *Ecology*, *91*(2), 621. https://doi.org/10.1890/09-0889.1

Feng, D., Beck, H., Lawson, K., & Shen, C. (2022). The suitability of differentiable, learnable hydrologic models for ungauged regions and climate change impact assessment (preprint). Catchment hydrology/Modelling approaches. https://doi.org/10.5194/hess-2022-245

Feng, D., Liu, J., Lawson, K., & Shen, C. (2022). Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. *Water Resources Research*, *58*(10), e2022WR032404. https://doi.org/10.1029/2022WR032404

Ficchì, A., Raso, L., Dorchies, D., Pianosi, F., Malaterre, P.-O., Van Overloop, P.-J., & Jay-Allemand, M. (2016). Optimal operation of the multireservoir system in the seine river basin using deterministic and ensemble forecasts. *Journal of Water Resources Planning and Management*, *142*(1), 05015005. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000571

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

Förster, K., Hanzer, F., Stoll, E., Scaife, A. A., MacLachlan, C., Schöber, J., et al. (2018). Retrospective forecasts of the upcoming winter season snow accumulation in the Inn headwaters (European Alps). *Hydrology and Earth System Sciences*, *22*(2), 1157–1173. https://doi.org/10.5194/hess-22-1157-2018

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., et al. (2022). Deep learning rainfall–runoff predictions of extreme events. *Hydrology and Earth System Sciences*, *26*(13), 3377–3392. https://doi.org/10.5194/hess-26-3377-2022

Garen, D. C. (1992). Improved techniques in regression-based streamflow volume forecasting. *Journal of Water Resources Planning and Management*, *118*(6), 654–670. https://doi.org/10.1061/(ASCE)0733-9496(1992)118:6(654)

Girons Lopez, M., Crochemore, L., & Pechlivanidis, I. G. (2021). Benchmarking an operational hydrological model for providing seasonal forecasts in Sweden. *Hydrology and Earth System Sciences*, *25*(3), 1189–1209. https://doi.org/10.5194/hess-25-1189-2021

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (pp. 249–256). JMLR Workshop and Conference Proceedings. Retrieved from https://proceedings.mlr.press/v9/glorot10a.html

Harrigan, S., Prudhomme, C., Parry, S., Smith, K., & Tanguy, M. (2018). Benchmarking ensemble streamflow prediction skill in the UK. *Hydrology and Earth System Sciences*, *22*(3), 2023–2039. https://doi.org/10.5194/hess-22-2023-2018

Henn, B., Newman, A. J., Livneh, B., Daly, C., & Lundquist, J. D. (2018). An assessment of differences in gridded precipitation datasets in complex terrain. *Journal of Hydrology*, *556*, 1205–1219. https://doi.org/10.1016/j.jhydrol.2017.03.008

Hirpa, F. A., Fagbemi, K., Afiesimam, E., Shuaib, H., & Salamon, P. (2015). Saving lives: Ensemble-based early warnings in developing nations. In Q. Duan, F. Pappenberger, J. Thielen, A. Wood, H. L. Cloke, & J. C. Schaake (Eds.), *Handbook of hydrometeorological ensemble forecasting* (pp. 1–22). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40457-3_43-1

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

Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G., et al. (2021). MC-LSTM: Mass-Conserving LSTM. *arXiv*. https://doi.org/10.48550/arXiv.2101.05186

Hosseini, F., Prieto, C., & Álvarez, C. (2024). Hyperparameter optimization of regional hydrological LSTMs by random search: A case study from Basque Country, Spain. *Journal of Hydrology*, *643*, 132003. https://doi.org/10.1016/j.jhydrol.2024.132003

Ibrahim, K. S. M. H., Huang, Y. F., Ahmed, A. N., Koo, C. H., & El-Shafie, A. (2022). A review of the hybrid artificial intelligence and optimization modelling of hydrological streamflow forecasting. *Alexandria Engineering Journal*, *61*(1), 279–303. https://doi.org/10.1016/j.aej.2021.04.100

Kaune, A., Chowdhury, F., Werner, M., & Bennett, J. (2020). The benefit of using an ensemble of seasonal streamflow forecasts in water allocation decisions. *Hydrology and Earth System Sciences*, *24*(7), 3851–3870. https://doi.org/10.5194/hess-24-3851-2020

Koster, R. D., Mahanama, S. P. P., Livneh, B., Lettenmaier, D. P., & Reichle, R. H. (2010). Skill in streamflow forecasts derived from large-scale estimates of soil moisture and snow. *Nature Geoscience*, *3*(9), 613–616. https://doi.org/10.1038/ngeo944

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, *55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., et al. (2023). Caravan - A global community dataset for large-sample hydrology. *Scientific Data*, *10*(1), 61. https://doi.org/10.1038/s41597-023-01975-w

Lehner, F., Wood, A. W., Llewellyn, D., Blatchford, D. B., Goodbody, A. G., & Pappenberger, F. (2017). Mitigating the impacts of climate nonstationarity on seasonal streamflow predictability in the U.S. Southwest. *Geophysical Research Letters*, *44*(24), 12208–12217. https://doi.org/10.1002/2017GL076043

Li, D., Wrzesien, M. L., Durand, M., Adam, J., & Lettenmaier, D. P. (2017). How much runoff originates as snow in the western United States, and how will that change in the future? *Geophysical Research Letters*, *44*(12), 6163–6172. https://doi.org/10.1002/2017gl073551

Livneh, B., & Badger, A. M. (2020). Drought less predictable under declining future snowpack. *Nature Climate Change*, *10*(5), 452–458. https://doi.org/10.1038/s41558-020-0754-8

Lukas, J., & Payton, E. (2020). Colorado river basin climate and hydrology: State of the science. https://doi.org/10.25810/3HCV-W477

Mai, J., Shen, H., Tolson, B. A., Gaborit, É., Arsenault, R., Craig, J. R., et al. (2022). The great lakes runoff intercomparison project phase 4: The great lakes (GRIP-GL). *Hydrology and Earth System Sciences*, *26*(13), 3537–3572. https://doi.org/10.5194/hess-26-3537-2022

Markham, C. G. (1970). Seasonality of precipitation in the United States. *Annals of the Association of American Geographers*, *60*(3), 593–597. https://doi.org/10.1111/j.1467-8306.1970.tb00743.x

Modi, P., & Livneh, B. (2024). Long short term memory simulations and code for 664 basins in the ensemble streamflow prediction framework (LSTM-ESP). *Zenodo*. https://doi.org/10.5281/ZENODO.14213154

Modi, P. A., Small, E. E., Kasprzyk, J., & Livneh, B. (2022). Investigating the role of snow water equivalent on streamflow predictability during drought. *Journal of Hydrometeorology*, *23*(10), 1607–1625. https://doi.org/10.1175/JHM-D-21-0229.1

Mushtaq, S., Chen, C., Hafeez, M., Maroulis, J., & Gabriel, H. (2012). The economic value of improved agrometeorological information to irrigators amid climate variability: The economic value of improved agrometeorological information. *International Journal of Climatology*, *32*(4), 567–581. https://doi.org/10.1002/joc.2015

Musselman, K. N., Addor, N., Vano, J. A., & Molotch, N. P. (2021). Winter melt trends portend widespread declines in snow water resources. *Nature Climate Change*, *11*(5), 418–424. https://doi.org/10.1038/s41558-021-01014-9

Musselman, K. N., Clark, M. P., Liu, C., Ikeda, K., & Rasmussen, R. (2017). Slower snowmelt in a warmer world. *Nature Climate Change*, *7*(3), 214–219. https://doi.org/10.1038/nclimate3225

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research*, *57*(3). https://doi.org/10.1029/2020WR028091

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2014). Development of a large-sample watershed-scale hydrometeorological dataset for the contiguous USA: Dataset characteristics and assessment of regional variability in hydrologic model performance (preprint). Catchment hydrology/Modelling approaches. https://doi.org/10.5194/hessd-11-5599-2014

Ng, K. W., Huang, Y. F., Koo, C. H., Chong, K. L., El-Shafie, A., & Najah Ahmed, A. (2023). A review of hybrid deep learning applications for streamflow forecasting. *Journal of Hydrology*, *625*, 130141. https://doi.org/10.1016/j.jhydrol.2023.130141

NOAA. (2024). *River forecast Centers*. National Oceanic and Atmospheric Administration. Retrieved November 21, 2024, from https://www.noaa.gov/jetstream/rfcs

Pagano, T., Garen, D., & Sorooshian, S. (2004). Evaluation of official western U.S. Seasonal water supply outlooks, 1922–2002. *Journal of Hydrometeorology*, *5*(5), 896–909. https://doi.org/10.1175/1525-7541(2004)005<0896:EOOWUS>2.0.CO;2

Pagano, T. C., Garen, D. C., Perkins, T. R., & Pasteris, P. A. (2009). Daily updating of operational statistical seasonal water supply forecasts for the western U.S. *JAWRA Journal of the American Water Resources Association*, *45*(3), 767–778. https://doi.org/10.1111/j.1752-1688.2009.00321.x

Rasmussen, R. M., Chen, F., Liu, C. H., Ikeda, K., Prein, A., Kim, J., et al. (2023). CONUS404: The NCAR–USGS 4-km long-term regional hydroclimate reanalysis over the CONUS. *Bulletin of the American Meteorological Society*, *104*(8), E1382–E1408. https://doi.org/10.1175/BAMS-D-21-0326.1

Restrepo, P., Roe, J., Dietz, C., & Werner, M. (2010). Hydrologic forecasting at the US National weather Service in the 21st Century: Transition from the NWS River Forecast System (NWSRFS) to the community hydrologic prediction system (CHPS). https://www.researchgate.net/publication/234395144_Hydrologic_Forecasting_at_the_US_National_Weather_Service_in_the_21st_Century_Transition_from_the_NWS_River_Forecast_System_NWSRFS_to_the_Community_Hydrologic_Prediction_System_CHPS

Ruder, S. (2016). An overview of gradient descent optimization algorithms (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1609.04747

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., et al. (2023). Differentiable modeling to unify machine learning and physical models and advance Geosciences (Version 2). https://doi.org/10.48550/ARXIV.2301.04027

Shukla, S., & Lettenmaier, D. P. (2011). Seasonal hydrologic prediction in the United States: Understanding the role of initial hydrologic conditions and seasonal climate forecast skill. *Hydrology and Earth System Sciences*, *15*(11), 3529–3538. https://doi.org/10.5194/hess-15-3529-2011

Singh, S. K. (2016). Long-term streamflow forecasting based on ensemble streamflow prediction technique: A case study in New Zealand. *Water Resources Management*, *30*(7), 2295–2309. https://doi.org/10.1007/s11269-016-1289-7

Sit, M., Demiray, B. Z., Xiang, Z., Ewing, G. J., Sermet, Y., & Demir, I. (2020). A comprehensive review of deep learning applications in hydrology and water resources. *Water Science and Technology*, *82*(12), 2635–2670. https://doi.org/10.2166/wst.2020.369

Slack, J. R., & Landwehr, J. M. (1992). *Hydro-climatic data network: A US Geological Survey streamflow data set for the United States for the study of climate variations, 1874–1988*. USGS Open-File Report 92-129. US Geological Survey.

Slater, L. J., Arnal, L., Boucher, M.-A., Chang, A. Y.-Y., Moulds, S., Murphy, C., et al. (2023). Hybrid forecasting: Blending climate predictions with AI models. *Hydrology and Earth System Sciences*, *27*(9), 1865–1889. https://doi.org/10.5194/hess-27-1865-2023

Troin, M., Arsenault, R., Wood, A. W., Brissette, F., & Martel, J.-L. (2021). Generating ensemble streamflow forecasts:A review of methods and approaches over the past 40 years. *Water Resources Research*, *57*(7), e2020WR028392. https://doi.org/10.1029/2020WR028392

Trujillo, E., & Molotch, N. P. (2014). Snowpack regimes of the Western United States. *Water Resources Research*, *50*(7), 5611–5623. https://doi.org/10.1002/2013WR014753

Turner, S. W. D., Bennett, J. C., Robertson, D. E., & Galelli, S. (2017). Complex relationship between seasonal streamflow forecast skill and value in reservoir operations. *Hydrology and Earth System Sciences*, *21*(9), 4841–4859. https://doi.org/10.5194/hess-21-4841-2017

United States Geological Survey. (2024). USGS water data for the nation [Dataset]. *U.S. Geological Survey*. https://doi.org/10.5066/F7P55KJN

United States, US Department of Agriculture, National Resource Conservation Service, & National Water and Climate Center. (2024). Air and water database [Dataset]. https://nwcc-apps.sc.egov.usda.gov/

US EPA, O. (2015). Level III and IV ecoregions of the continental United States [data and tools]. Retrieved May 30, 2024, from https://www.epa.gov/eco-research/level-iii-and-iv-ecoregions-continental-united-states

U. S. Geological Survey. (2023). GAGES-II: Geospatial attributes of gages for evaluating streamflow [Dataset]. *U.S. Geological Survey*. https://doi.org/10.5066/P96CPHOT

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. (Version 7). *arXiv*. https://doi.org/10.48550/ARXIV.1706.03762

Wang, Q., Hapuarachchi, P., & Pagano, T. (2010). Continuous rainfall-runoff model comparison and short-term daily streamflow forecast skill evaluation. https://doi.org/10.4225/08/58542C672DD2C

Watts, G., Christierson, B. V., Hannaford, J., & Lonsdale, K. (2012). Testing the resilience of water supply systems to long droughts. *Journal of Hydrology*, *414–415*, 255–267. https://doi.org/10.1016/j.jhydrol.2011.10.038

Wood, A. W., Hopson, T., Newman, A., Brekke, L., Arnold, J., & Clark, M. (2016). Quantifying streamflow forecast skill elasticity to initial condition and climate prediction skill. *Journal of Hydrometeorology*, *17*(2), 651–668. https://doi.org/10.1175/JHM-D-14-0213.1

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L., et al. (2012). Continental-scale water and energy flux analysis and validation for North American Land data assimilation system project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow: Validation of model-simulated streamflow. *Journal of Geophysical Research*, *117*(D3). https://doi.org/10.1029/2011JD016051

Xia, Y., NCEP/EMC, et al. (2009). NLDAS primary forcing data L4 hourly 0.125 x 0.125 degree, version 002 [Dataset]. *NASA Goddard Earth Sciences Data and Information Services Center*. https://doi.org/10.5067/6J5LHHOHZHN4

Yu, Q., Jiang, L., Schneider, R., Zheng, Y., & Liu, J. (2024). Deciphering the mechanism of better predictions of regional LSTM models in ungauged basins. *Water Resources Research*, *60*(7), e2023WR035876. https://doi.org/10.1029/2023WR035876

Yuan, X., Ma, F., Wang, L., Zheng, Z., Ma, Z., Ye, A., & Peng, S. (2016). An experimental seasonal hydrological forecasting system over the Yellow River basin – Part 1: Understanding the role of initial hydrological conditions. *Hydrology and Earth System Sciences*, *20*(6), 2437–2451. https://doi.org/10.5194/hess-20-2437-2016

Zeng, X., Broxton, P., & Dawson, N. (2018). Snowpack change from 1982 to 2016 over conterminous United States. *Geophysical Research Letters*, *45*(23). https://doi.org/10.1029/2018GL079621