Water Resources Research

RESEARCH ARTICLE
10.1029/2025WR039888

A Deep State Space Model for Rainfall-Runoff Simulations

Yihan Wang[1], Lujun Zhang[1], Annan Yu[2], N. Benjamin Erichson[3], and Tiantian Yang[1]

[1]School of Civil Engineering and Environmental Science, University of Oklahoma, Norman, OK, USA, [2]Center for Applied Mathematics, Cornell University, Ithaca, NY, USA, [3]Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Key Points:

- A novel Deep Learning model termed the Frequency Tuned Diagonal State Space Sequence (S4D-FT) is introduced for rainfall-runoff simulations
- S4D-FT generally outperforms the decades-old LSTM at 531 watersheds in CONUS, providing a new DL benchmark for rainfall-runoff simulations
- S4D-FT prevails in watersheds with frequent, prolonged high- and low-flow events of smaller runoffs with snowmelt and intermittent regimes

Supporting Information:

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

Correspondence to:

T. Yang,
tiantian.yang@ou.edu

Citation:

Wang, Y., Zhang, L., Yu, A., Erichson, N. B., & Yang, T. (2025). A deep state space model for rainfall-runoff simulations. *Water Resources Research*, *61*, e2025WR039888. https://doi.org/10.1029/2025WR039888

Received 5 JAN 2025
Accepted 18 NOV 2025

Author Contributions:

Conceptualization: Yihan Wang, Lujun Zhang, N. Benjamin Erichson
Formal analysis: Yihan Wang, Lujun Zhang
Funding acquisition: Tiantian Yang
Investigation: Yihan Wang, Lujun Zhang, Annan Yu, N. Benjamin Erichson, Tiantian Yang
Methodology: Yihan Wang, Lujun Zhang, Annan Yu, N. Benjamin Erichson
Resources: N. Benjamin Erichson
Software: Yihan Wang, Annan Yu, N. Benjamin Erichson
Validation: Yihan Wang, Lujun Zhang

Abstract The classical way of studying the rainfall-runoff processes in the water cycle relies on conceptual or physically-based hydrologic models. Deep learning (DL) has recently emerged as an alternative and blossomed in the hydrology community for rainfall-runoff simulations. However, the decades-old Long Short-Term Memory (LSTM) network remains the benchmark for this task, outperforming newer architectures like Transformers. In this work, we propose a State Space Model (SSM), specifically the Frequency Tuned Diagonal State Space Sequence (S4D-FT) model, for rainfall-runoff simulations. The proposed S4D-FT is benchmarked against the established LSTM and a physically-based Sacramento Soil Moisture Accounting model under in-sample and out-of-sample simulation setups across 531 watersheds in the contiguous United States (CONUS). Results show that S4D-FT is able to outperform the LSTM model across diverse regions under both simulation setups, especially for regions that feature snowmelt-driven or intermittent flow regimes. In contrast, S4D-FT tends to underperform in flashier, high-magnitude flow regimes, likely due to its global state-space convolution computation that emphasizes slow, storage-driven dynamics, which makes it less effective at picking up short bursts and noisy spikes in the data. In summary, our pioneering introduction of the S4D-FT for rainfall-runoff simulations challenges the dominance of LSTM in the hydrology community and expands the arsenal of DL tools available for hydrological modeling.

Plain Language Summary Traditionally, scientists study how rainfall becomes runoff in the water cycle using models based on physical principles. Recently, Artificial Intelligence (AI) and Deep Learning (DL) have emerged as alternative approaches, receiving increased attention in hydrology for simulating rainfall-runoff with notable success. Despite advancements in AI/DL, the Long Short-Term Memory (LSTM) network, a decades-old technique, remains the standard, outperforming newer approaches like Transformers and gradually becoming a go-to DL model for rainfall-runoff simulations. In this study, we introduce the Frequency Tuned Diagonal State Space Sequence (S4D-FT) model, a novel DL architecture distinct from both Transformers and LSTMs, for rainfall-runoff simulations. We tested S4D-FT against the well-established LSTM and a physically-based hydrologic model called the Sacramento Soil Moisture Accounting (Sac-SMA) model across 531 watersheds in the United States. The results show that S4D-FT outperforms LSTM in various regions. Our work introduces the S4D-FT as a new tool for rainfall-runoff simulations, challenging the dominance of LSTM and expanding DL options for hydrological modeling.

1. Introduction

The rainfall-runoff relationship is a fundamental concept in hydrology. It describes how rainfall is transformed into surface runoff through interconnected hydrologic processes, such as infiltration, evapotranspiration, and the exchange of water between surface and subsurface flows (Beven & Kirkby, 1979). Thoroughly understanding these hydrologic processes and subsequently achieving accurate simulations of the rainfall-runoff relationship are critical for proactive flood forecasting and mitigation, efficient agricultural planning, and strategic urban development (Beven, 2012; Knapp et al., 1991; Moradkhani and Sorooshian, 2008).

Physically-based hydrologic models (PBMs), grounded in physical laws that govern hydrologic dynamics, are the standard tools for simulating rainfall-runoff relationships (Beven, 1996). However, the highly nonlinear nature of various hydrologic processes often challenges PBMs, limiting their accuracy in diverse conditions (Beven, 1989; Clark et al., 2017). Consequently, there is a growing need for innovative approaches to address the limitations of PBMs.

Deep learning (DL) has emerged as an alternative to PBMs, showing success in capturing the complex, nonlinear patterns in rainfall-runoff simulations. The hydrology community also explores the applicability of DL models in

Visualization: Yihan Wang
Writing – original draft: Yihan Wang
Writing – review & editing: Lujun Zhang, Annan Yu, N. Benjamin Erichson, Tiantian Yang

rainfall-runoff simulations across diverse temporal scales and geospatial locations. For the large-scale studies that focus on the model evaluation in the contiguous United States (CONUS), it is recognized that the decade-old Long Short-Term Memory (LSTM) networks (Hochreiter, 1997) continue to be the best-performing architecture for rainfall-runoff simulations, with even Transformers (Vaswani, 2017) unable to outperform LSTMs (Frame et al., 2022; Kratzert, Klotz, Herrnegger, et al., 2019; Kratzert, Klotz, Shalev, et al., 2019; Liu et al., 2024). Importantly, LSTMs not only excel at in-sample simulation, but also deliver strong out-of-sample simulation performance for the Prediction in Ungauged Basins (PUB; Sivapalan et al., 2003) problems, demonstrating good transferability and generalizability for hydrologic applications (Kratzert, Klotz, Herrnegger, et al., 2019).

In this work, we pioneer the use of a new set of State Space Models (SSMs) (Gu, Goel, & Ré, 2021; Gu, Johnson, et al., 2021) for rainfall-runoff simulations. Since the original invention, the SSMs have achieved state-of-the-art performance across diverse tasks in video, audio, and time-series processing, excelling at long-range sequence modeling while being faster and more memory-efficient than LSTMs and Transformers (Gu & Dao, 2023; Patro & Agneeswaran, 2024). However, to the best of the authors' knowledge, there is no study that has tested out SSMs in simulating the rainfall-runoff processes in the field of hydrology. There is one recent work that applied SSMs in large-scale reservoir simulations, in which the authors proved that the SSMs have superior statistical performance over the traditional LSTMs (Zhang, Yue, et al., 2025). Therefore, in this work, we raise the following three scientific questions: (a) Can SSMs enhance rainfall-runoff simulations and outperform the decades-old LSTM model that has been extensively used and widely accepted as the DL benchmark in the hydrology community? (b) What is the hydrologic capability of SSMs in solving PUB problems? And (c) how can we explain the better or worse of SSM as compared to LSTM in different hydrologic conditions? To answer these questions, we employ the Frequency Tuned Diagonal State Space Sequence (S4D-FT) model (Yu, Lyu, et al., 2024) for rainfall-runoff simulations across 531 watersheds in CONUS and carry out a comprehensive evaluation of the model performance.

Our evaluation follows a three-step approach. Firstly, the overall statistical performance of S4D-FT under both in-sample and out-of-sample (i.e., PUB) setups is compared with the basic S4D (without frequency tuning) (Gu et al., 2022), various existing DL benchmarks, and a traditional PBM (Sacramental Soil Moisture Accounting, Sac-SMA; Anderson & McDonnell, 2005) on CONUS-wide rainfall-runoff simulation tasks. Secondly, the spatial distribution of S4D-FT and LSTM performance across all study watersheds under both setups are illustrated to compare the model performance at different geographic locations. Lastly, a detailed investigation based on the global setup is conducted to analyze the possible factors driving regional variability in S4D-FT's performance across CONUS. With our three-step evaluation, we conclude that the S4D-FT model demonstrates overall better performance, and we identify the conditions under which S4D-FT outperforms or underperforms the LSTM model.

2. Data Sets and Methodology

In this study, we train and test a basic S4D, a variant S4D with frequency tuning (i.e., S4D-FT), and an LSTM model for rainfall-runoff simulations. The methodologies for S4D and S4D-FT are described in Section 2.1. The LSTM follows the implementation of Kratzert, Klotz, Herrnegger, et al. (2019) and Kratzert, Klotz, Shalev, et al. (2019) to match state-of-the-art benchmarks. The experimental design, employed data sets, and training setups are detailed in Section 2.2. The evaluation strategies are described in Section 2.3. Lastly, the methodology of attribution analysis to diagnose why S4D-FT outperforms or underperforms LSTM is described in Section 2.4.

2.1. State Space Models

State Space Models (SSMs; Gu, Goel, & Ré, 2021; Gu, Johnson, et al., 2021) have recently gained increasing attention due to their strong performance in handling long-sequence data, and provide a promising alternative to recurrent neural networks (RNNs). SSMs' architecture enables them to efficiently model complex sequential data while addressing some of the computational and stability challenges commonly faced by RNNs (Erichson et al., 2020, 2022; Rusch & Mishra, 2021). This makes SSMs well-suited for hydrologic applications, such as rainfall-runoff modeling, where long-term dependencies and evolving temporal patterns are critical.

Figure 1 provides an overview of an SSM-based model processing pipeline, specifically the Diagonal State Space Sequence model (S4D) employed in this study (details provided in later sections). Raw input feature(s) of length L are first projected into a higher-dimensional H-channel representation to align with the model's working

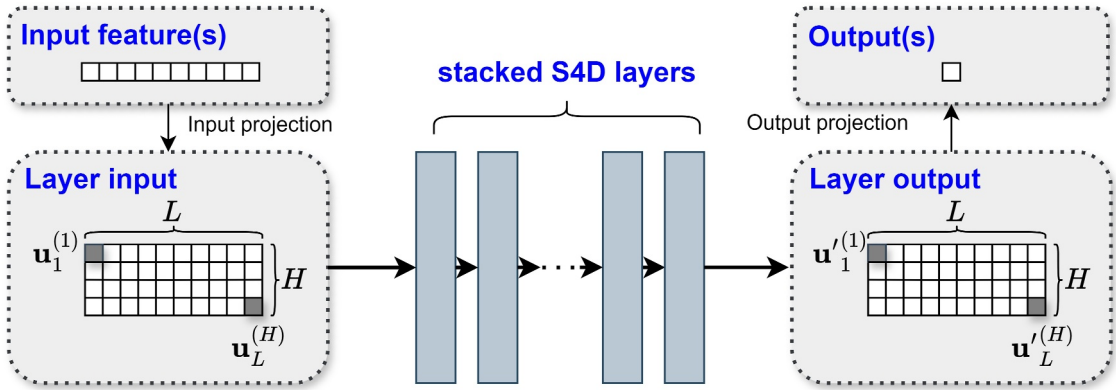

Figure 1. A processing pipeline of SSM variant of S4D. Raw input features of length L are projected to an H-channel representation (layer input), passed through a stack of S4D layers. The layer output from the last S4D layer is then passed through a final output projection (readout) and mapped to the model output.

dimensionality. The projected input is then passed through a stack of S4D layers, each of which performs state space computations parameterized by a set of learned state matrices. The output from the final S4D layer is passed through output projection to map the internal representation back to the prediction target(s).

Following this overview, we introduce the theoretical foundation (Section 2.1.1) and the practical implementation (Section 2.1.2) for SSMs. The specific SSM variants, namely S4D and S4D-FT employed in this study are introduced in Section 2.1.3, together with their mathematical properties that enable high computational efficiency and long-range memory.

2.1.1. Continuous-Time State Space Formulation as Theoretical Foundation

The foundation of SSMs is built upon continuous-time linear time-invariant (LTI) systems, which provide a structured approach for capturing relationships between inputs, outputs, and latent states over time. These relationships can be represented by the following equations:

$$\boldsymbol{x}'(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) \tag{1}$$

$$\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{D}\boldsymbol{u}(t) \tag{2}$$

where $\boldsymbol{u}(t) \in \mathbb{C}^m$, $\boldsymbol{y}(\mathrm{t}) \in \mathbb{C}^p$, and $\boldsymbol{x}(t) \in \mathbb{C}^n$ are the inputs, outputs, and latent states at time t, respectively. Here, m is the number of channels after input projection, p is the dimension of output, and n represents the per-channel state dimension. $\boldsymbol{x}'(t)$ is the derivative of $\boldsymbol{x}(t)$ with respect to time in the continuous form. The matrices $\boldsymbol{A} \in \mathbb{C}^{n\times n}$, $\boldsymbol{B} \in \mathbb{C}^{n\times m}$, $\boldsymbol{C} \in \mathbb{C}^{p\times n}$, and $\boldsymbol{D} \in \mathbb{C}^{m\times p}$ are the trainable parameters. Each matrix serves a distinct role in the model: $\boldsymbol{A}$ defines how the state evolves over time, $\boldsymbol{B}$ determines how inputs influence the state, $\boldsymbol{C}$ maps the state to the output, and $\boldsymbol{D}$ directly relates inputs to outputs.

2.1.2. Discretization With Trainable Sampling Intervals for Practical Implementation

For practical implementation of SSMs, the continuous-time equations (Equations 1 and 2) need to be discretized using a sampling interval. Usually, this interval is determined by the data interval (e.g., daily or hourly time steps based on data availability) and is held fixed throughout model training and inference. However, real-world processes operate across a broad spectrum of temporal scales. For example, in streamflow dynamics, rapid runoff events last hours, while slow groundwater-driven baseflow persists weeks or months. Consequently, using a fixed sampling interval may limit a model's ability to capture the full temporal patterns present in the modeling system.

SSMs address such limitations through a trainable sampling interval Δt, which allows the model to adapt its internal update frequency independently of the fixed data interval. Specifically, SSMs incorporate multiple parallel processing units (referred to as channels), with each channel h having its own sampling interval:

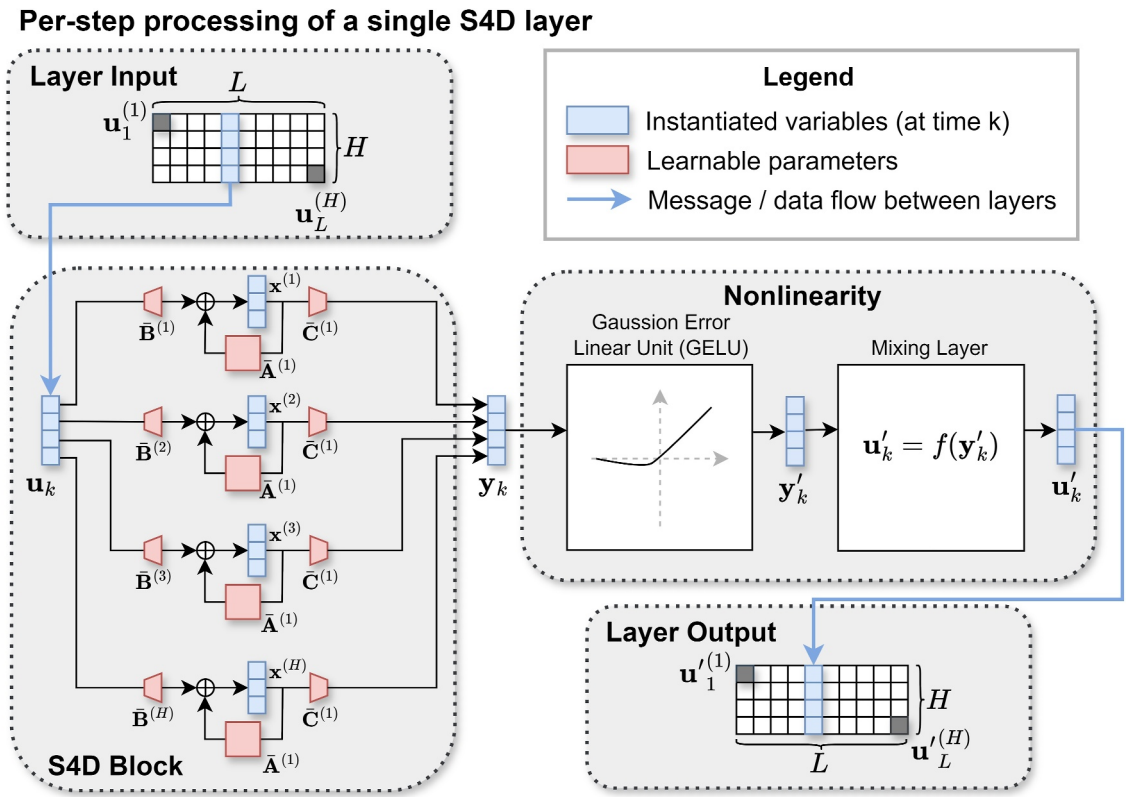

Figure 2. S4D(-FT) layer mechanics. For a single S4D layer, at each time step k, the input $\boldsymbol{u}_k$ (with H channels and sequence length L) is fed into the S4D block, consisting of H parallel LTI systems. Each S4D block is parameterized with time-discretized learnable matrices $\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}}, \overline{\boldsymbol{C}}$, and $\overline{\boldsymbol{D}}$. The output $\boldsymbol{y}_k$ passes through a Gaussian Error Linear Unit (GELU) activation function and a mixing layer consisting of Gated Linear Unit (GLU) and 1D convolution to introduce nonlinearity. The resulting output $\boldsymbol{u}'_k$ is then passed to the next S4D layer.

$$\Delta t_h = \exp(\ell_h), \tag{3}$$

where the scalar $\ell_h \in \mathbb{R}$ is a trainable parameter. Each ℓ_h is initialized from $\ell_h \sim \text{Uniform}[\ln(\Delta t_{\min}), \ln(\Delta t_{\max})]$, where $\Delta t_{\min}$ and $\Delta t_{\max}$ are user-specified bounds. During training, all ℓ_h are optimized jointly via gradient descent. By replacing a single fixed Δt with per-channel and trainable intervals Δt_h, the SSM can automatically allocate some channels to fast dynamics (small Δt_h) and others to slow dynamics (large Δt_h), enabling the model to represent processes across diverse time scales within a single unified framework.

To illustrate the concept, Figure 2 shows the state space computation performed within a single S4D layer, that is, one of the stacked layers in Figure 1. The layer comprises a linear S4D block (multiple LTI systems) followed by a within-layer nonlinearity component. The "Layer Input" in Figure 2 corresponds to the projected multi-channel input representation introduced earlier in Figure 1. Specifically, each row of the layer input $\boldsymbol{u}$ corresponds to one of the H state-space channels, and each column corresponds to a time step within a sequence length L. For clarity, the top-left shaded grid entry $\boldsymbol{u}_1^{(1)}$ represents the first time step in the first channel, whereas the bottom-right shaded grid entry $\boldsymbol{u}_L^{(H)}$ represents the final time step in the last channel.

As shown in the "S4D Block" box in Figure 2, at each time step k, the vector input $\boldsymbol{u}_k$ is simultaneously processed by multiple parallel LTI systems, each associated with a separate channel. Each channel maintains its own state-space representation. The system matrices for channel $h = 1, 2, \ldots, H$, denoted $\overline{\boldsymbol{A}}^{(h)}$, $\overline{\boldsymbol{B}}^{(h)}$, $\overline{\boldsymbol{C}}^{(h)}$, and $\overline{\boldsymbol{D}}^{(h)}$, are obtained by discretizing the continuous-time matrices $\boldsymbol{A}^{(h)}$, $\boldsymbol{B}^{(h)}$, $\boldsymbol{C}^{(h)}$, and $\boldsymbol{D}^{(h)}$ using the channel's learned sampling interval Δt_h and the zero-order-hold (ZOH) method (Gu et al., 2022):

$$\overline{\boldsymbol{A}}^{(h)} = \exp\left(\Delta t_h \boldsymbol{A}^{(h)}\right) \tag{4}$$

$$\overline{\boldsymbol{B}}^{(h)} = \left(\boldsymbol{A}^{(h)}\right)^{-1}\left(\exp\left(\Delta t_h \boldsymbol{A}^{(h)}\right) - I\right) \cdot \boldsymbol{B}^{(h)} \tag{5}$$

$$\overline{\boldsymbol{C}}^{(h)} = \boldsymbol{C}^{(h)} \tag{6}$$

$$\overline{\boldsymbol{D}}^{(h)} = \boldsymbol{D}^{(h)} \tag{7}$$

At each time step k, the model updates the state in every channel according to the following discretized system:

$$\boldsymbol{x}_k = \overline{\boldsymbol{A}}\boldsymbol{x}_{k-1} + \overline{\boldsymbol{B}}\boldsymbol{u}_k \tag{8}$$

$$\boldsymbol{y}_k = \overline{\boldsymbol{C}}\boldsymbol{x}_{k-1} + \overline{\boldsymbol{D}}\boldsymbol{u}_k \tag{9}$$

Although LTI systems (i.e., the "S4D Block" in Figure 2) are linear, SSMs gain the ability to capture complex, nonlinear relationships by stacking multiple LTI systems and connecting them with nonlinear transformations, creating a deep model. As shown in the "Nonlinearity" box in Figure 2, for each S4D layer, the output $\boldsymbol{y}_k$ from the S4D block undergoes a series of nonlinear transformations before being passed to the next layer. Specifically, the output first passes through a Gaussian Error Linear Unit (GELU; Hendrycks & Gimpel, 2016) activation function applied independently to each channel. GELU is followed by a mixing layer for cross-channel interactions. In this work, the mixing layer consists of a Gated Linear Unit (GLU; Dauphin et al., 2017) followed by a lightweight 1D convolution, together learning to emphasize or suppress features and increase representational capacity. The resulting vector $\boldsymbol{u}'_k$ serves as the input to the subsequent S4D layer.

2.1.3. S4D and S4D-FT

In this study, we employ two specific variants of SSMs for rainfall-runoff simulations, termed Diagonal State Space Sequence (S4D) and Frequency Tuned S4D (S4D-FT). The S4D model simplifies the architecture by setting $\boldsymbol{A}$ to be diagonal and configuring the LTI system for single-input/single-output (SISO) operations, where $m = p = 1$. To enhance S4D's adaptability to various temporal patterns, the S4D-FT, adopted from Yu, Lyu, et al. (2024), introduces frequency tuning that further tunes the intrinsic bias that comes from the distribution of the eigenvalues of $\boldsymbol{A}$ in the Laplace domain. To clarify, "frequency" here refers to the rate at which a signal changes over time. Low-frequency components represent slowly varying trends, such as gradual shifts or persistent long-term cycles, whereas high-frequency components correspond to rapidly changing features, including sharp transitions, brief oscillations or bursts.

Frequency tuning is implemented by explicitly rescaling both the real and imaginary parts of $\boldsymbol{A}$ through learnable hyperparameters, denoted as α_r (real-part scaling) and α_i (imaginary part scaling):

$$\boldsymbol{A} = -\exp\left(\alpha_r \cdot \log \boldsymbol{A}_{\text{real}}^{\text{base}} + i \cdot \left(\alpha_i \cdot \boldsymbol{A}_{\text{imag}}^{\text{base}}\right)\right) \tag{10}$$

Though numerous other SSM variants exist (Agarwal et al., 2023; Hasani et al., 2022; Smith et al., 2022; Yu, Mahoney, & Erichson, 2024), Yu, Lyu, et al. (2024) shows that the S4D-FT remains highly competitive when the frequency bias is tuned at initialization. In S4D-FT, α_r affects the decay rate of state dynamics, thereby controlling the model's effective memory length. Lower values of α_r retains longer memory, whereas higher values emphasize short-memory response. On the other hand, α_i affects the oscillatory behaviors of S4D-FT to capture high-frequency patterns in the input sequence. Smaller values of α_i suppress oscillatory behavior, leading to smooth hydrographs, whereas larger values of α_i amplify oscillations and allow the model to track high-frequency signals such as streamflow associated with flashy storms. Consequently, this frequency tuning provides a flexible mechanism to enhance SSM performance across diverse temporal domains.

At first glance, the SSM structure may resemble an RNN, leading one to question its effectiveness. However, SSMs have several unique advantages that address key limitations of RNNs and LSTMs, particularly in handling long-range dependencies. RNNs and LSTMs are based on recurrent operations. The sequential nature of these models often results in slow training. SSMs, on the other hand, are able to process sequences in parallel, which

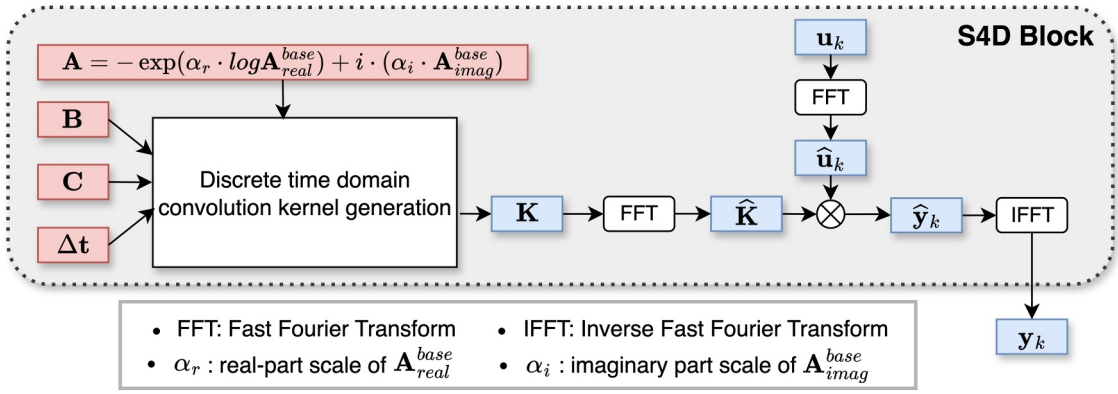

Figure 3. Illustration of S4D time-discretized convolution kernel computation. In basic S4D (without frequency tuning), a discrete-time convolution kernel $\boldsymbol{K}$ is generated from the continuous state-space matrices ($\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D}$) and sampling intervals Δt. In the frequency-tuning S4D (S4D-FT), the real and imaginary parts of the base $\boldsymbol{A}$ matrix are rescaled using two predefined parameters α_r and α_i respectively to adjust the model's frequency bias. For efficient convolution computation, the resulting time-domain convolution kernel $\boldsymbol{K}$ is transformed into the frequency domain ($\hat{\boldsymbol{K}}$) by Fast Fourier Transform (FFT). The input sequence, also transformed from time domain to frequency domain ($\widehat{\boldsymbol{u_k}}$), is multiplied elementwise with $\hat{\boldsymbol{K}}$, and the result is converted back to the time domain by Inverse FFT to compute the convolution output $\boldsymbol{y}_k$ efficiently.

reduces computing time and avoids issues such as exploding and vanishing gradients. Specifically, SSMs reformulate the system's state evolution as a convolution between the input sequence $\boldsymbol{u}$ and an analytically derived kernel $\overline{\boldsymbol{K}}$:

$$\boldsymbol{y} = \boldsymbol{u} * \overline{\boldsymbol{K}}, \tag{11}$$

where $\overline{\boldsymbol{K}} = (\overline{\boldsymbol{CB}}, \overline{\boldsymbol{CAB}}, \ldots, \overline{\boldsymbol{CA}^{L-1}\boldsymbol{B}})$.

The convolution kernel $\overline{\boldsymbol{K}}$ is a combination of damped exponential and sinusoidal modes. If discretized with small sampling interval $\Delta \boldsymbol{t}$, the damped sinusoids fade only a little at each time step, preserving information over long horizon. $\overline{\boldsymbol{K}}$ can be computed either in the time domain (Smith et al., 2022) or frequency domain (Parnichkun et al., 2024; Yu, Lyu, et al., 2024). In this study, the convolution operation is conducted in the frequency domain, as shown in Figure 3. Specifically, the input sequence and the kernel are transformed via Fast Fourier Transform (FFT), multiplied element-wise, and then transformed back to the time domain using the inverse FFT (IFFT). This approach reduces the computational cost for long sequences and enables fully parallel processing across all channels and time steps.

2.2. Experimental Design and Computational Cost

To compare the performance of the basic S4D, S4D-FT and LSTM, we first train and test all three models on 531 watersheds simultaneously, referred to as the "global setup" following the terminology in Kratzert, Klotz, Herrnegger, et al. (2019). Model training configuration aligns with existing LSTM and Transformer benchmarks (Frame et al., 2022; Liu et al., 2024). Specifically, the DL models are trained using long-term hydrometeorological time series and catchment attributes of 531 unimpaired watersheds across CONUS from the Catchment Attributes and MEteorology for Large-sample Studies (CAMELS; Addor et al., 2017; Newman et al., 2015) as well as the corresponding streamflow measurements from the United States Geological Survey (USGS). In consistent with previous studies, we use 32 input variables (5 hydrometeorological variables from North American Land Data Assimilation System (NLDAS; Xia et al., 2012) and 27 static catchment attributes, detailed in Table S1 in Supporting Information S1) from 10/1/1999 to 9/30/2008 for training, and 10/1/1989 to 9/30/1999 for testing. All DL models are trained using the Adam optimizer in a sequence-to-one setting with a 365-day look-back window. To ensure robustness, each model is trained with multiple random seeds, resulting in an eight-member ensemble.

The hyperparameters for S4D-FT are manually tuned by trial-and-error, using a randomly selected 10% of the training data as the validation set. The finalized hyperparameters are applied to the basic S4D model where

Table 1
Training Efficiency and Peak Memory (per Seed) for S4D-FT Versus LSTM on a Single NVIDIA L40S GPU

Cost (per seed)	S4D-FT	LSTM
GPU time per epoch	0.2 hr	0.42 hr
Total training time	10 hr	12.5 hr
Peak memory	4.8 GB	2.3 GB

applicable. In addition, we provide a sensitivity analysis to further evaluate the robustness of S4D-FT. Specifically, each hyperparameter of S4D-FT is systematically varied within a defined range while all others are held constant at their final selected values. The tested ranges, descriptions, as well as the final selected values of the S4D and S4D-FT hyperparameters are provided in Table S3 in Supporting Information S1. The results of the sensitivity analysis are shown in Figures S1 and S2 in Supporting Information S1. For LSTM, we adopt the same hyperparameters as reported in earlier benchmark work without further tuning (detailed in Table S2 in Supporting Information S1) to ensure comparability with the best scores reported by previous studies.

In addition to the global setup, we also evaluate S4D-FT's performance under the "PUB setup," following the terminology and experimental design in Kratzert, Klotz, Herrnegger, et al. (2019). The evaluation of S4D-FT under the PUB setup could reveal its generalizability in ungauged basins relative to the well-established LSTM. Under the PUB setup, S4D-FT and LSTM are evaluated out of sample both spatially and temporally using the k-fold cross-validation approach ($k = 12$). Specifically, all 531 basins are randomly separated into 12 folds. For each fold, each model is trained on the training-period data from watersheds in the 11 in-sample folds and evaluated on the testing-period data from watersheds in the held-out fold. Hyperparameters for the PUB setup training are directly reused from the global setup for both S4D-FT and LSTM without modification.

We measure the computational cost for LSTM and S4D-FT on a single NVIDIA-L40S GPU, as shown in Table 1. For the global setup, S4D-FT completes one training epoch in approximately 12 min, whereas LSTM needs about 25 min per epoch. The total training time sums to approximately 10 GPU-hours per seed for S4D-FT (train for 50 epochs) and 12.5 hr for LSTM per seed (train for 30 epochs). For the PUB setup (12-fold cross-validation), both models require 12 times the GPU time relative to the global setup, resulting in an estimated total training time of 120 GPU-hours for S4D-FT and 150 GPU-hours for LSTM per seed. S4D-FT's higher training efficiency stems from its use of global convolution via FFT to process all time steps simultaneously, in contrast to LSTM's slower sequential recurrence (as detailed in Section 2.1). However, it comes with higher memory demands. Specifically, S4D-FT occupies approximately 4.8 GB of GPU memory during training, more than twice that of LSTM, which uses around 2.3 GB.

2.3. Evaluation

The performance of the proposed S4D and S4D-FT is evaluated from two perspectives. Firstly, the overall statistical accuracy is comprehensively evaluated using six statistical metrics under both the global setup and the PUB setup. These metrics include the Pearson-r correlation, Nash-Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970), Kling-Gupta Efficiency (KGE; Gupta et al., 2009), percent bias in flow duration curve high-segment (top 2%) volume (FHV; Yilmaz et al., 2008), percent bias in flow duration curve low-segment (lowest 30%) volume (FLV; Yilmaz et al., 2008), as well as the overall percentage bias (PBias). The detailed formulation of each metric is provided in Table 2. For broader context and comprehensive comparison, we also include statistical values of other popular DL models benchmarked on the CAMELS data set, namely the Mass Conserving (MC) LSTM (Frame et al., 2022) and Transformers (Liu et al., 2024), as well as a physical Sac-SMA model.

Secondly, we present the spatial distribution of S4D-FT's performance compared to LSTM for a clearer and more direct comparison for watersheds at different geospatial locations under both the global setup and the PUB setup. We employ NSE and KGE skill scores to illustrate the simulation accuracy of S4D-FT relative to LSTM. Further details on the computation of NSE and KGE skill scores can also be found in Table 2. We do not conduct spatial performance analysis for the remaining DL models or the physically-based Sac-SMA, as these models have been shown to perform less effectively than the LSTM (Frame et al., 2022; Kratzert, Klotz, Herrnegger, et al., 2019; Kratzert, Klotz, Shalev, et al., 2019; Liu et al., 2024).

2.4. Attribution Analysis

To investigate the reason why S4D-FT outperforms or underperforms LSTM from a hydrologic perspective, a diagnostic attribution analysis is conducted based on the results from the global setup. In the attribution analysis, all study watersheds are divided into two groups. Group 1 includes watersheds where S4D-FT consistently

Table 2
Evaluation Metrics and Their Formulations, Ranges, and Optimal Values

Metric	Formulation	Range	Optima
Pearson-r	$\frac{\sum_{i=1}^{n}\left(\left(Q_{\mathrm{sim},i}-\overline{Q_{\mathrm{sim}}}\right)\left(Q_{\mathrm{obs},i}-\overline{Q_{\mathrm{obs}}}\right)\right)}{\sqrt{\sum_{i=1}^{n}\left(Q_{\mathrm{sim},i}-\overline{Q_{\mathrm{sim}}}\right)^{2}\sum_{i=1}^{n}\left(Q_{\mathrm{obs},i}-\overline{Q_{\mathrm{obs}}}\right)^{2}}}$	(-inf, 1]	1
Nash-Sutcliffe Efficiency (NSE)	$1-\frac{\sum_{i=1}^{n}\left(Q_{\mathrm{obs},i}-Q_{\mathrm{sim}}\right)^{2}}{\sum_{i=1}^{n}\left(Q_{\mathrm{obs},i}-\overline{Q_{\mathrm{obs}}}\right)^{2}}$	(-inf, 1]	1
Kling-Gupta Efficiency (KGE)	$1-\sqrt{(CC-1)^{2}+\left(\frac{\overline{Q_{\mathrm{sim}}}}{\overline{Q_{\mathrm{obs}}}}-1\right)^{2}+\left(\frac{\sigma_{\mathrm{sim}}}{\sigma_{\mathrm{obs}}}-1\right)^{2}}$	(-inf, 1]	1
Percent bias in flow duration curve high-segment volume (FHV)	$\frac{\sum_{h=1}^{H}\left(Q_{\mathrm{sim},h}-Q_{\mathrm{obs},h}\right)}{\sum_{h=1}^{H}Q_{\mathrm{obs},h}}\times 100$	(-inf, inf)	0
Percent bias in flow duration curve low-segment volume (FLV)	$\frac{-\sum_{l=1}^{L}\left[\log(Q_{\mathrm{sim},l})-\log(Q_{\mathrm{sim},L})\right]-\sum_{l=1}^{L}\left[\log(Q_{\mathrm{obs},l})-\log(Q_{\mathrm{obs},L})\right]}{\sum_{l=1}^{L}\left[\log(Q_{\mathrm{obs},l})-\log(Q_{\mathrm{obs},L})\right]}\times 100$	(-inf, inf)	0
Percentage bias (PBias)	$\frac{\sum_{i=1}^{n}Q_{\mathrm{sim},i}-\sum_{i=1}^{n}Q_{\mathrm{obs},i}}{\sum_{i=1}^{n}Q_{\mathrm{obs},i}}\times 100$	(-inf, inf)	0
NSE skill score (model relative to reference)	$\frac{\mathrm{NSE}_{\mathrm{model}}-\mathrm{NSE}_{\mathrm{ref}}}{1-\mathrm{NSE}_{\mathrm{ref}}}$	(-inf, 1]	1
KGE skill score (model relative to reference)	$\frac{\mathrm{KGE}_{\mathrm{model}}-\mathrm{KGE}_{\mathrm{ref}}}{1-\mathrm{KGE}_{\mathrm{ref}}}$	(-inf, 1]	1
FHV improvement (model relative to reference)	$\frac{\lvert\mathrm{FHV}_{\mathrm{ref}}\rvert-\lvert\mathrm{FHV}_{\mathrm{model}}\rvert}{100}$	(-inf, inf)	inf
Pearson-r improvement (model relative to reference)	$\lvert\mathrm{Pearsonr}_{\mathrm{ref}}-1\rvert-\lvert\mathrm{Pearson\ r}_{\mathrm{model}}-1\rvert$	(-inf, inf)	inf
PBias improvement (model relative to reference)	$\frac{\lvert\mathrm{PBias}_{\mathrm{ref}}\rvert-\lvert\mathrm{PBias}_{\mathrm{model}}\rvert}{100}$	(-inf, inf)	inf
Percentage difference in hydrologic signature h between watersheds Group 2 and Group 1	$\frac{\overline{h_{\mathrm{Group\ 2}}}-\overline{h_{\mathrm{Group\ 1}}}}{\overline{h_{\mathrm{Group\ 1}}}}\times 100$	(-inf, inf)	0

Note. n: length of testing sample time series. H: the number of flow indices that fall within the exceedance probability of 0.02 (i.e., top 2% of flow volumes). L: the number of flow indices that fall within the 30% low-flow segment (i.e., 0.7–1.0 exceedance probability range).

outperforms LSTM, with both positive NSE and KGE skill scores. Group 2 includes the remaining watersheds with negative NSE and/or KGE skill scores, suggesting that S4D-FT does not completely outperform LSTM.

The attribution analysis includes two parts. First, we identify which statistical aspects of S4D-FT's simulation drive NSE and KGE improvements or deteriorations. We compute correlations between NSE and KGE skill scores and improvements in additional evaluation metrics (FHV, Pearson-r, and PBias) across 531 study watersheds. To further validate our findings, we present simulated and observed hydrographs from the testing period for two representative watersheds from each group (i.e., one where S4D-FT performs well and one poorly). Formulations for computing FHV, Pearson-r, and PBias improvements are also presented in Table 2.

The second part of the attribution analysis is to investigate how watershed characteristics influence S4D-FT's performance. Specifically, we consider a series of hydrologic signatures as indicators of streamflow behavior to identify what types of streamflow S4D-FT are good at or bad at. To further link the model performance with physical drivers, we also examine the climate, soil, and vegetation properties presented in the CAMELS data set. The selected watershed characteristics with technical descriptions are provided in Table S4 in Supporting Information S1. For the selected hydrologic, climate, soil, and vegetation characteristics, we calculate percentage differences (formulation presented in Table 2) between Groups 1 and 2, and assess the correlations with NSE and KGE skill scores for each group.

3. Results

3.1. Overall Performance of SSMs and Existing Benchmarks

The statistical performance of S4D-FT, basic S4D, and other benchmarks (Sac-SMA, LSTM, MC-LSTM, Transformers, and Modified Transformers) for rainfall-runoff simulations across CONUS under both the global and PUB setups is summarized in Table 3. Among these models, only LSTM and S4D-FT include results for both setups; the remaining models (Sac-SMA, MC-LSTM, Transformers, Modified Transformers, and basic S4D)

Table 3
Statistical Performance Comparison of Models Using Various Metrics for the Global Setup and the PUB Setup

Model	NSE (↑)	KGE (↑)	Pearson-r (↑)	FHV (%) (→0)	FLV (%) (→0)	PBias (%) (→0)
Global setup						
Sac-SMA[a]	0.65 (±0.004)	0.66 (±0.006)	0.82 (±0.001)	−21.36 (±0.47)	38.46 (±2.31)	**2.53 (±0.38)**
MC-LSTM[a]	0.72	0.72	0.86	−18.72	−30.84	5.02
Transformers[b]	N/A	0.71 (±0.007)	N/A	−26.66 (±2.83)	3.31 (±2.34)	N/A
Modified Transformers[c]	N/A	0.74 (±0.007)	N/A	−18.00 (±2.94)	**2.28 (±4.24)**	N/A
LSTM	0.72 (±0.005)	0.74 (±0.007)	0.86 (±0.002)	−17.51 (±1.17)	10.63 (±6.18)	5.42 (±1.34)
S4D	0.72 (±0.004)	0.72 (±0.03)	0.86 (±0.001)	−18.07 (±4.13)	16.18 (±23.50)	5.92 (±8.83)
S4D-FT	**0.74 (±0.002)**	**0.75 (±0.019)**	**0.87 (±0.001)**	**−16.98 (±2.26)**	20.17 (±20.77)	5.87 (±3.31)
PUB setup						
LSTM	0.62 (±0.006)	0.61 (±0.009)	0.83 (±0.004)	**−20.82 (±0.95)**	**8.20 (±5.02)**	**6.74 (±0.96)**
S4D-FT	**0.66 (±0.003)**	**0.63 (±0.014)**	**0.85 (±0.001)**	−21.14 (±1.44)	30.22 (±7.90)	9.08 (±2.19)

Note. Each metric value represents the median across 531 watersheds, with standard deviations calculated from different ensemble members shown in parentheses. Values without parentheses and/or "N/A" denotes unavailability. Metrics marked with "↑" indicate that higher values are preferred. "→0" means that values closer to zero are optimal. Best-performing metrics (medians closest to ideal) are in bold and highlighted in red. [a]Adopted from Frame et al. (2022). [b]Adopted from Liu et al. (2024). [c]Adopted from Liu et al. (2024).

report results solely for the global setup. For each model, Median values of evaluation metrics are presented for all study watersheds, with standard deviations across ensemble members shown in parentheses where available. The results for MC-LSTM, Transformers, Modified Transformers and Sac-SMA are directly adopted from previous studies. The results for S4D-FT, the basic S4D, and LSTM are produced by our own simulation experiments. Notably, our LSTM results align with benchmarks reported in the literature (i.e., Frame et al., 2022; Kratzert, Klotz, Herrnegger, et al., 2019; Kratzert, Klotz, Shalev, et al., 2019; Liu et al., 2024).

According to Table 3, under the global setup, S4D-FT demonstrates the best median NSE, KGE, Pearson-r, and FHV among all models. Moreover, S4D-FT shows the lowest standard deviation in NSE and Pearson-r, indicating greater consistency across different random seed initializations. In terms of FLV, the Transformers (including the basic Transformers and the Modified Transformers) outperform Sac-SMA, LSTM-type models (i.e., LSTM and MC-LSTM), and SSMs (i.e., S4D and S4D-FT). Regarding the overall bias (PBias), Sac-SMA still achieves the best performance, which is closest to zero, followed by LSTM-type models, and then SSMs.

Comparing the basic S4D with S4D-FT, the basic S4D achieves only slightly better accuracy than the basic Transformers but still underperforms the LSTM and Modified Transformers. However, frequency tuning (i.e., S4D-FT) notably enhances S4D's performance, improving all metrics except for FLV and establishing S4D-FT as the overall best-performing model among all existing benchmarks.

Under the PUB setup, S4D-FT achieves higher median NSE, KGE, and Pearson-r than LSTM. However, S4D-FT underperforms LSTM in bias-related metrics (PBias, FHV, and FLV) for both median values and standard deviation across random seeds. Compared with the global setup, the PUB setup leads to larger performance gains for S4D-FT over LSTM in NSE and KGE but also reveals more pronounced performance decline in bias-related metrics.

3.2. Regional Performance Comparison of S4D-FT and LSTM

Since LSTM is recognized as the leading model for CONUS-wide rainfall-runoff simulations, we focus on comparing S4D-FT and LSTM in regional performance, under both the global and PUB setups. Under the global setup, Figures 4a and 4b shows NSE and KGE skill scores that demonstrate the relative improvement of S4D-FT over the baseline LSTM. Positive skill scores (red) indicate better performance by S4D-FT, while negative scores (blue) indicate LSTM outperforms S4D-FT. A skill score of 1 reflects theoretical best performance of S4D-FT (NSE or KGE = 1), and a skill score of 0 denotes equal performance between LSTM and S4D-FT. Darker colors represent greater differences.

10.1029/2025WR039888

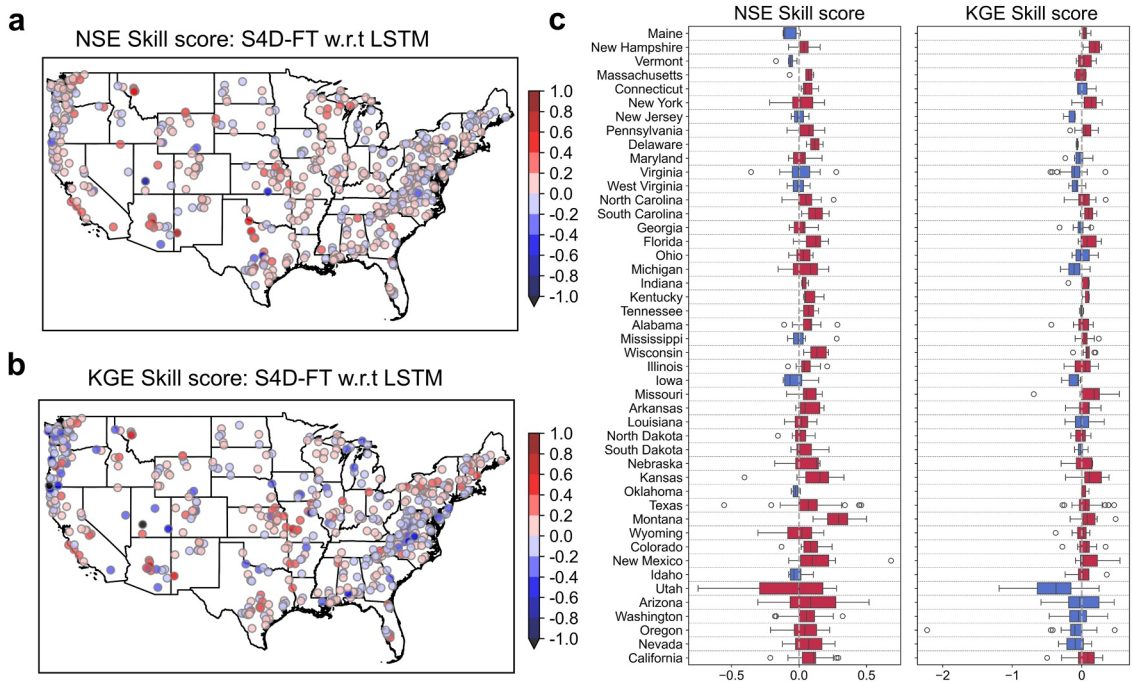

Figure 4. Simulation performance of S4D-FT relative to the LSTM model across study watersheds in CONUS under the global setup. Panel (a) Spatial distribution of NSE skill scores, with red dots (positive NSE skill score) indicating S4D-FT outperformance and blue dots (negative NSE skill score) indicating LSTM outperformance. Panel (b) Spatial distribution of KGE skill scores, following the same color scheme as Panel (a). Panel (c) Boxplots of NSE and KGE skill scores by state, ordered east to west, with red boxes for positive median scores and blue for negative medians. Rhode Island and Minnesota are excluded due to no study watersheds.

According to Figures 4a and 4b, S4D-FT presents better NSE and KGE prevalently across much of the CONUS, as shown by the widespread distribution of red dots. The better performance of S4D-FT is particularly pronounced in the Pacific Southwest and Mid-South (including Kansas, Missouri, Arkansas, and Texas). However, S4D-FT underperforms LSTM in several regions with different spatial patterns between NSE and KGE skill scores. Specifically, NSE skill scores show clusters of negative values (blue dots) along the East Coast (e.g., Maine and Virginia) and scattered negative values in parts of the Midwest. In contrast, KGE skill scores display more frequent negative values, particularly along the East Coast (e.g., Virginia), Great Lakes, Midwest (e.g., Utah), and Pacific Northwest (Washington and Oregon).

A more detailed breakdown of NSE and KGE skill scores by U.S. state is provided in Figure 4c, where red boxes represent states with a positive median skill score for NSE or KGE, and blue boxes indicate a negative median. According to Figure 4c, S4D-FT outperforms LSTM in most states, confirming the observations from Figures 4a and 4b. However, S4D-FT underperforms LSTM along the East Coast (specifically New Jersey, Virginia, and West Virginia) and in the Great Lakes region (specifically Minnesota and Iowa). For KGE skill scores alone, S4D-FT also shows weaker performance in the western U.S., particularly in Utah, Arizona, Washington, Oregon, and Nevada.

The performance of S4D-FT and LSTM under the PUB setup is shown in Figure 5, following the same format as Figure 4. Compared to the global setup, S4D-FT's advantage over LSTM in both NSE and KGE is more pronounced, as indicated by the more prevalence and intensity of red dots in Figures 5a and 5b, particularly across the Appalachia, Great Lakes, Rockies, and Pacific Northwest regions. Such improvements are further quantified in Figure 5c, which shows the state-level breakdown of skill scores. According to Figure 5c, S4D-FT achieves higher NSE in all but four states, namely Maine, Vermont, Iowa, and South Dakota. For KGE, S4D-FT reverses the underperformance observed in the western U.S. under the global setup, and demonstrates outperformance in Utah, Arizona, Oregon, and Nevada.

10.1029/2025WR039888

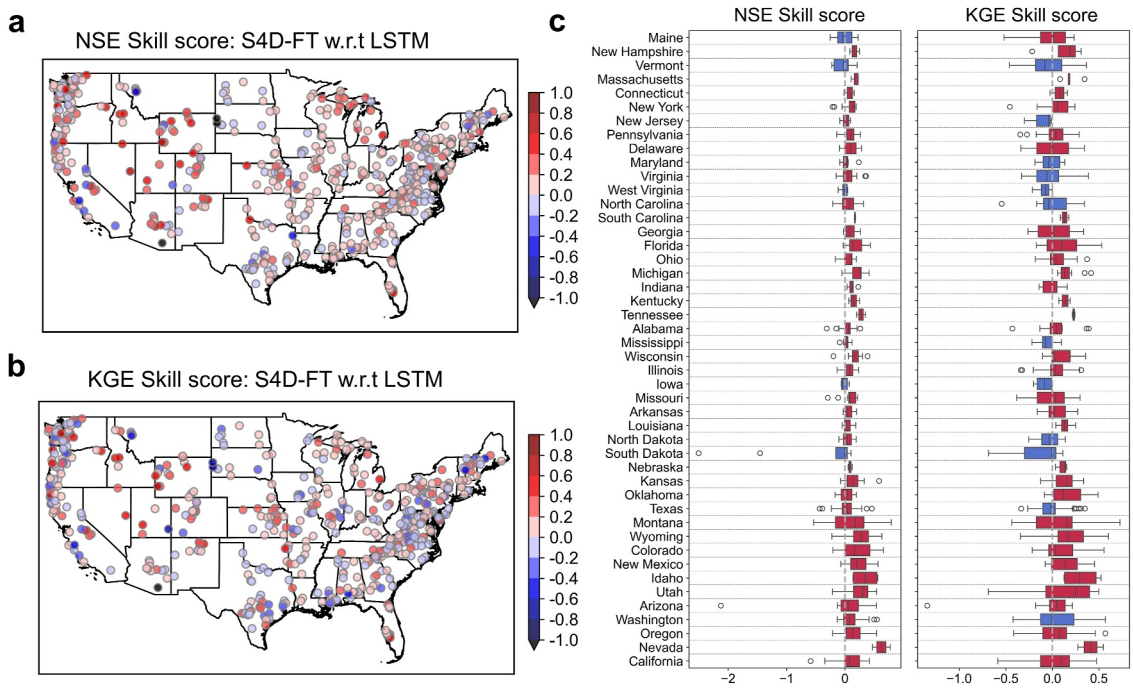

Figure 5. Simulation performance of S4D-FT relative to the LSTM model across study watersheds in CONUS under the PUB setup. Panel (a) Spatial distribution of NSE skill scores, with red dots (positive NSE skill score) indicating S4D-FT outperformance and blue dots (negative NSE skill score) indicating LSTM outperformance. Panel (b) Spatial distribution of KGE skill scores, following the same color scheme as Panel (a). Panel (c) Boxplots of NSE and KGE skill scores by state, ordered east to west, with red boxes for positive median scores and blue for negative medians. Rhode Island and Minnesota are excluded due to no study watersheds.

3.3. Investigating Factors Behind Regional Variability in S4D-FT Performance

To further investigate the varying performance of the S4D-FT over LSTM, we divide all study watersheds into two groups based on the relative performance of S4D-FT over LSTM (described in Section 2.4). Figure 6a

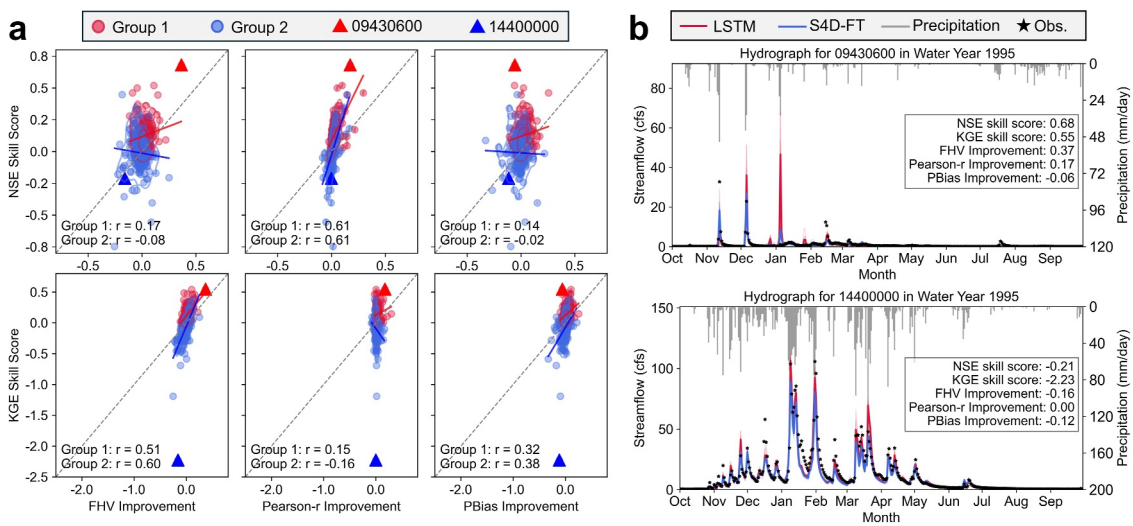

Figure 6. Analysis of S4D-FT's performance relative to LSTM considering multiple evaluation statistics. Panel (a) Scatter plots overlaid with contour density plots of NSE and KGE skill scores against improvements in FHV, Pearson correlation, and PBias for Group 1 (red) and Group 2 (blue) watersheds. Solid red and blue lines represent regression lines. Correlation coefficients are displayed at the bottom left of each plot. The red and blue triangles highlight two specific example watersheds (i.e., 09430600 and 14400000) from Group 1 and Group 2, respectively. Panel (b) Simulated 8-member ensemble hydrographs for LSTM (red) and S4D-FT (blue), along with observed streamflow (black stars), for the highlighted watersheds (i.e., 09430600 and 14400000). Thin and thick lines indicate individual ensemble members and ensemble medians, respectively. Precipitation is shown on the secondary *y*-axis (gray bars). The NSE and KGE skill scores, along with improvements in FHV, Pearson-r, and PBias calculated from the testing results are displayed for each watershed.

presents scatter plots overlaid with contour density of S4D-FT's NSE or KGE skill scores against its improvements or deterioration in FHV, Pearson-r, and PBias for the two groups of watersheds (Group 1: $n = 244$, red; Group 2: $n = 287$, blue). According to Figure 6a, NSE skill scores strongly correlate with Pearson-r improvements ($r = 0.61$) across both groups but have limited correlation with improvements in FHV and PBias, especially for Group 2 watersheds. In contrast, for KGE skill scores, both groups of watersheds exhibit the strongest positive correlation with FHV improvements ($r = 0.51$ and $r = 0.60$ for Groups 1 and 2, respectively), followed by PBias improvements ($r = 0.32$ and $r = 0.38$ for Groups 1 and 2, respectively), but weak correlations with Pearson-r improvements.

Figure 6b displays the simulated and observed hydrographs for two representative watersheds: one with the highest KGE skill score (USGS station 09430600, red triangle in Figure 6a) and one with the lowest KGE skill score (USGS station 14400000, blue triangle in Figure 6a). Note that the hydrographs shown in Figure 6b are only for one water year (WY1995), while all statistical values in Figure 6b are computed over the entire testing period.

According to Figure 6b, at watershed 09430600, S4D-FT improves FHV (by 0.37) and Pearson-r (by 0.17) as compared to LSTM, while showing no improvement in PBias (−0.06). The improved FHV and Pearson-r could be validated by the hygrograph in WY 1995, where S4D-FT alleviates LSTM's significant overestimation in December and January by correctly identifying that the observed precipitation impulses during this period did not translate into streamflow. Conversely, at watershed 14400000, S4D-FT present decreased FHV (by −0.16) and PBias (by −0.12). Specifically in the WY 1995 example, while both LSTM and S4D-FT show good simulation alignments with observation, LSTM captures the flow spikes in January, February, and mid-March more closely than S4D-FT. Pearson-r at watershed 14400000 remains unchanged (improvement by 0) between LSTM and S4D-FT, which aligns with the weak relationship observed in Figure 6a between Pearson-r improvement and KGE skill scores. We provide additional hydrographs in Figure S3 in Supporting Information S1 for two more Group 1 watersheds and two more Group 2 watersheds. These hydrographs reinforce the observations that while S4D-FT mitigates LSTM's overestimations, it also tends to underestimate high-flow events.

In addition, Figure 7 presents the relationship between S4D-FT's performance and watershed streamflow, climate, soil, and vegetation characteristics, as detailed in Section 2.4. Note that although the CAMELS data set includes many watershed characteristics, for parsimony we analyze and report only those that (a) show a percentage difference >10% between the two watershed groups, or (b) have an absolute correlation $|r| \geq 0.2$ with NSE or KGE skill scores. In each panel in Figure 7, the leftmost column shows the percentage differences in watershed characteristics between Group 1 (where S4D-FT outperforms LSTM) and Group 2 (where S4D-FT does not outperform LSTM). Blue and orange indicate higher and lower values in Group 1, respectively. The remaining four columns show Pearson correlation coefficients between each attribute and the NSE and KGE skill scores for each group. Positive correlations are in red and negative correlations are in blue, with darker shades indicating stronger correlations.

According to Figure 7a, Group 1 watersheds tend to have lower streamflow volumes (q_mean, q5, q95) and less flashy flow regimes (slope_fdc) compared to Group 2 watersheds, as indicated by the negative percentage differences in the first column of the heatmaps. On the other hand, Group 1 watersheds exhibit higher frequency and longer duration of both high and low flow events (high_q_freq, high_q_dur, low_q_freq, low_q_dur, and zero_q_freq), as shown by the positive percentage differences. The observed inter-group differences in hydrologic signatures align with intra-group correlations. Specifically, within Group 1, lower NSE and KGE skill scores are associated with higher streamflow volumes and flashier streamflow behavior, as suggested by the negative correlation with q_mean, q_95, and slope_fdc. In contrast, higher skill scores are associated with more frequent and sustained high- and low-flow conditions as suggested by positive correlation with high_q_freq, high_q_dur, low_q_freq, low_q_dur, zero_q_freq. However, within Group 2, little correlation is shown between skill scores and hydrologic signatures, with the exception of moderate negative correlations between KGE skill score and q_mean and slope_fdc.

Figure 7b provides additional insight into the relationship between watershed climate conditions and S4D-FT's performance. Compared to Group 2, Group 1 watersheds receive less annual precipitation but exhibit a higher fraction of snowfall, pointing to potentially snow-dominated climates. Group 1 watersheds also exhibit greater aridity, reflected by a higher aridity index and longer durations of dry periods (i.e., higher low_prec_dur). Within Group 1, higher skill scores are associated with more arid conditions (i.e., higher aridity index). Interestingly, the fraction of snow does not correlate with S4D-FT's performance within Group 1 watersheds while it serves as a

10.1029/2025WR039888

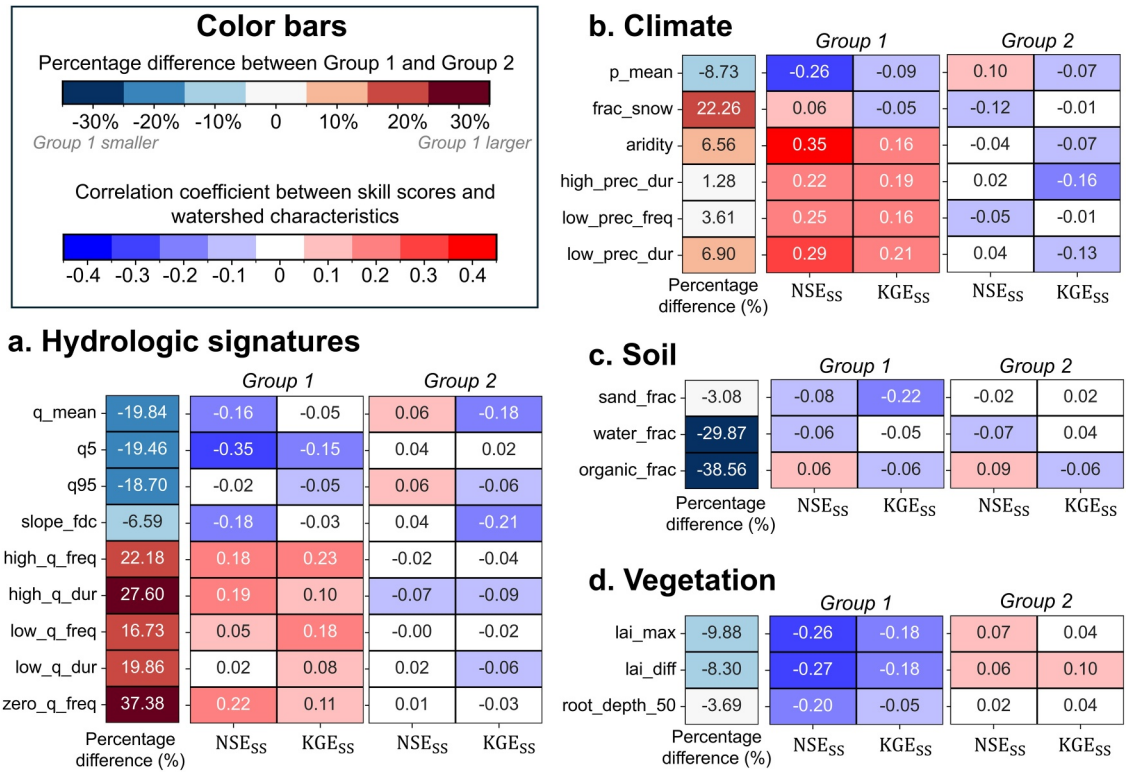

Figure 7. Relationship between watershed characteristics and the performance of S4D-FT relative to LSTM. Panel (a) Heatmap of percentage differences in hydrologic signatures between Group 2 and Group 1 watersheds (left column, blue-orange color scale), alongside correlation coefficients between each characteristic and the NSE/KGE skill scores (i.e., NSE_{SS} and KGE_{SS}) within each group (right four columns, blue-red color scale). Panels (b–d) same information as Panel (a), but for climate, soil, and vegetation characteristics, respectively. All values are labeled within the heatmap cells. The color bars above the panels indicate shared color scales used across all subplots.

distinguishing characteristic between Group 1 and Group 2. Instead, stronger correlations are observed between better skill scores and the temporal patterns of precipitation (i.e., high_prec_freq, high_prec_dur, low_prec_freq, and low_prec_dur), suggesting the better performance of S4D-FT in watersheds with more frequent and sustained high- and low-precipitation events.

Examining the soil properties (Figure 7c), Group 1 watersheds contain significantly lower fractions of water and organic material as shown by large negative percentage differences in water_frac and organic_frac, whereas the sand content (sand_frac) is comparable across both groups. Within Group 1, S4D-FT tends to perform worse in sandy soils. However, the fractions of water and organic matter themselves do not show meaningful correlations with skill scores. In Group 2, none of the soil characteristics display notable relationships with model performance.

Lastly, vegetation properties (Figure 7d) indicate that Group 1 watersheds generally have less vegetation, smaller seasonal vegetation variations, and shallower root systems, as indicated by negative percentage differences in lai_max, lai_diff, and root_depth_50. Within Group 1, S4D-FT tends to perform better in areas with sparser or less seasonal vegetation cover, as suggested by negative correlations between the skill scores and all three vegetation characteristics. In addition, little correlation is observed for Group 2 watersheds in terms of model performance and vegetation conditions.

4. Discussion

4.1. Performance of S4D-FT Across CONUS

The growing use of DL models in hydrology highlights the need for a standardized evaluation framework to better understand their strengths and limitations. Current studies often differ in training and testing data set, study periods, geospatial regions, and/or the number of watersheds analyzed, making direct comparisons difficult and

potentially misleading. To facilitate a clearer cross-model comparison, we adopt a configuration that has been used by several prior studies (e.g., Frame et al., 2023; Liu et al., 2024) as a point of reference. Specifically, we use 531 CAMELS watersheds with NLDAS forcing training from 10/1/1999 to 9/30/2008 and testing from 10/1/1989 to 9/30/1999. In prior work using this training and testing configuration, the decades-old LSTM has been the prevailing benchmark in rainfall-runoff simulations, outperforming newer architectures like Transformers. This has led to speculation that LSTM may have already reached, or is approaching, the predictive limit for rainfall-runoff simulations (Liu et al., 2024).

In an effort to push the boundaries of simulation accuracy, we revisit that benchmark by introducing the latest SSMs, specifically the S4D-FT, to the hydrology community. We explore S4D-FT's capability in advancing rainfall-runoff simulations, and comparing its performance to the existing prevailing DL benchmark for both in-sample ("global setup") and out-of-sample ("PUB setup") conditions. Our results show a favorable performance of S4D-FT as compared to the LSTM. Specifically, under the global setup, the overall median NSE value and KGE value increase from 0.72 to 0.74, and from 0.74 to 0.75 over a total of 531 watersheds across CONUS, respectively. The improvements are more pronounced under the PUB setup, where the median NSE increases from 0.62 to 0.66 and the median KGE from 0.61 to 0.63. Moreover, the outperformance of S4D-FT is generally consistent across different regions in CONUS, with exceptions in a few areas such as the eastern U.S. (e.g., Vermont and Virginia), and the Pacific Northwest considering both NSE and KGE (Figures 4 and 5).

Interestingly, S4D-FT shows greater improvements in NSE than KGE across more regions (Figures 4 and 5). Given that both NSE and KGE are composite metrics, such a discrepancy leads to two additional scientific questions: (a) Why does S4D-FT underperform on KGE more frequently than on NSE? (b) And does this suggest specific strengths or limitations of S4D-FT in rainfall-runoff simulation compared to LSTM? Our results in Figure 6 reveal that S4D-FT's KGE performance is strongly correlated with high-flow regime bias (FHV) but shows only a minimal correlation with temporal correspondence (Pearson-r correlation). In contrast, NSE is less affected by high-flow regime bias and is primarily driven by temporal consistency with observations. Although S4D-FT outperforms LSTM in both FHV and Pearson-r in terms of the overall statistics (Table 3), our further analysis indicates that S4D-FT results in a higher proportion of watersheds with improved Pearson-r compared to FHV (75% and 50% of the study watersheds, respectively according to Table S5 in Supporting Information S1). As a result, the FHV-sensitive nature of KGE leads to fewer improvements compared to the correlation-sensitive NSE. Given such information, we suspect that S4D-FT's primary strength over LSTM lies in its capability of capturing temporal correspondence. However, S4D-FT may offer limited improvement in simulating high-flow regimes compared to LSTM.

The relationship between hydrologic signatures and S4D-FT's NSE or KGE skill scores (Figure 7a) further clarifies the scenarios when S4D-FT exhibits performance advantages over LSTM. Specifically, S4D-FT tends to favor watersheds with smaller flow magnitudes, less flashy streamflow regimes, but more frequent and prolonged high-flow, low-flow, and zero-flow events. These patterns point to streamflow regimes where variability is more structured (e.g., seasonal, long dry spell, etc.) rather than dominated by frequent and short-lived flashy events. As further reinforced by the accompanying analysis of watershed climate characteristics (Figure 7b), S4D-FT tends to perform better in snow-dominated or arid regions, consistent with the structured streamflow regimes identified in the hydrologic signature analysis. In addition, S4D-FT's improved performance in landscapes with less open water, lower organic content, and sparser vegetation (Figures 7c and 7d) suggests its performance may be affected by complex hydrologic processes associated with these aforementioned hydrologic characteristics (Autio et al., 2020; Jones et al., 2019; Wu and Lane, 2017).

4.2. From Accuracy to Hydrologic Process Understanding

The discussed hydrologic, climate, and land surface characteristics associated with S4D-FT's improved performance could be found in several regions across CONUS. For instance, the Rocky Mountains and Sierra Nevada exhibit classic snowmelt-driven regimes (Addor et al., 2017; Brunner et al., 2020; Pham et al., 2021; Yang & Olivera, 2023; Yue et al., 2025), where streamflow follows a seasonal pattern with low flows during snow accumulation and high flows during snowmelt. Additionally, the Great Plains (from North Dakota to Texas, as well as Missouri and Arkansas) feature intermittent streamflow regimes characterized by prolonged drought periods, driven by high aridity and convective precipitation (e.g., thunderstorms or frontal systems) (Addor et al., 2017; Brunner et al., 2020; Yang & Olivera, 2023; Zhang, Yang, et al., 2025).

Conversely, we also identify a subset of watersheds where S4D-FT underperforms LSTM in KGE, notably clustered along the East Coast and West Coast, particularly in the Appalachia and the Pacific Northwest (Figures 2 and 3). These regions are characterized by pluvio-nival streamflow regimes (i.e., a combination of rainfall and snowmelt) that feature infrequent but intense high-flow events (Addor et al., 2017; Yang & Olivera, 2023). Although snow processes are also important in these regions, the interaction between snowmelt and precipitation leads to extreme high-flow volumes and peak events that deviate from the smaller-flow conditions and strong seasonal patterns. As a result, the humid Appalachia and Pacific Northwest emerge as regions where S4D-FT underperforms. This compromised performance not only supports our earlier suspicion that S4D-FT is less capable of representing high-flow volumes (i.e., relatively less improved FHV) but also agrees with our attribution analysis that shows its difficulty in capturing high-magnitude and flashy flow dynamics in humid regions.

The regions where S4D-FT outperforms LSTM coincide with the regions identified by Fang and Shen (2017) as being strongly influenced by water storage, such as snow pack and groundwater in streamflow generation. In contrast, the Appalachian Plateau, where S4D-FT underperforms, is also noted by Fang and Shen (2017) as water storage-limited due to the thin layer of soils. Considering the correspondence, we suspect that from a hydrologic process perspective, S4D-FT might be more capable of simulating storage-dependent streamflow regimes, but less effective if the storage impact is weak.

We suspect the underlying reason can be traced to the mathematical structure of S4D-FT. Specifically, S4D-FT parameterizes temporal dynamics through a state-space kernel as a sum of damped sinusoids and applies the kernel as a convolution across the entire input sequence. This global kernel representation naturally (a) supports long-range memory that integrates information over weeks to months, and (b) represents smooth trends (from the exponential decay) and periodic signals (from the sinusoidal terms) effectively (Gu, Goel, & Ré, 2021; Gu et al., 2022). In hydrologic contexts, S4D-FT may more faithfully capture delayed runoff contributions from snowmelt or groundwater release, as well as intermittent streamflow to carry depletion information across extended dry spells. However, limitations could also arise from the S4D-FT kernel formulation. Since each S4D-FT's kernel operates on the sequence globally, and is computed in the frequency domain that naturally emphasizes low-frequency (smooth) components (Gu, Goel, & Ré, 2021; Gu et al., 2022), S4D-FT tends to attenuate sharp, high-frequency (rapid) components such as short-lived flood peaks or high-flow extremes. Consequently, S4D-FT suffers from fidelity in simulating local, noisy, or rapidly varying hydrologic events.

By contrast, LSTM relies on recursive hidden states updated sequentially at each time step. Inherently, LSTM suffers from memory decay over long steps, limiting its ability to capture sustained storage-release processes. Nevertheless, the recursive structure could excel at modeling local or event-scale dynamics. As a result, LSTM could be more effective than S4D-FT at simulating sudden rainfall-runoff responses, presenting advantages in flashy, high-magnitude regimes where S4D-FT struggles.

4.3. Limitation and Future Work

While the improved accuracy of S4D-FT represents advancements in applying DL models to hydrologic simulations, its black-box nature remains a limitation. As an attempt, we provide a joint analysis that links S4D-FT's model mechanics with hydrologic regime characteristics. However, we acknowledge that our explanation, though plausible, remains superficial, as it does not fully uncover the specific processes that S4D-FT excels at or struggles with in rainfall-runoff simulations. We believe this challenge is not unique to S4D-FT; rather, it is a recurring question for all data-driven DL models in hydrology: *What specific hydrologic processes are these models effectively simulating, and where do they fall short?*

This challenge arises in part from the differing priorities between the computer science (CS) and hydrology communities. While the CS community prioritizes achieving higher predictive accuracy, the hydrology community focuses equally, if not more, on model interpretability. For hydrologists, a model's value is not only determined by its statistical performance but also by its ability to provide insight into the mechanisms driving certain hydrologic phenomena. Although it is reported that the examination of physical-meaningful hidden states of DL could also provide the underlying physical insights (Lees et al., 2021; Wang et al., 2025b), we note such an approach may not be applied universally as the identification of the physical-meaningful hidden node(s) is not guaranteed.

Alternatively, we believe further advancing physics-aware DL models is a promising direction. By integrating physical principles into data-driven model structures, physics-aware DL models could strike a balance between purely physics-based and purely data-driven approaches to achieve both interpretability and high simulation accuracy. A noteworthy example is the Mass Conserving (MC) LSTM (Hoedt et al., 2021), which imposes a strict mass conservation constraint on the standard LSTM structure. It is reported that MC-LSTM achieves significantly higher accuracy than physically-based models while providing greater interpretability than conventional LSTM by linking the model states and real-world hydrologic components (Frame et al., 2022). More recently, Wang et al. (2025b) proposed an alternative, more realistic mass-conservation constraint for LSTM, and showed improved performance over the original MC-LSTM.

We therefore believe that transferring the idea of integrating mass conservation constraints to S4D-FT could be a compelling future direction. However, we note that applying such constraints is non-trivial due to the fundamental architectural differences of S4D-FT compared to recurrent LSTM models. Unlike recurrent models that explicitly carry and update internal states (e.g., the cell states in LSTM) at each time step, S4D-FT implicitly encodes the full sequence of system dynamics into a single global convolutional kernel ($\overline{\boldsymbol{K}}$ in Equation 10) for efficient parallel sequence modeling. As a result, the intermediate hidden state trajectory in S4D-FT is obscured, making it difficult to enforce mass conservation constraints where storage changes need to equal the difference between incoming and outgoing mass at each timestep. We believe this challenge is not limited to mass conservation but applies to any physical constraint that depends on temporal continuity (e.g., such as energy conservation, momentum balance) due to the lack of access to intermediate states of S4D-FT.

A practical first step may be enforcing a sequence-level constraint (i.e., between the initial and final hidden states) instead of at individual time steps in S4D-FT. For mass conservation, this could mean requiring that the net difference between cumulative inputs and outputs across the full sequence matches the net change in storage between the initial and final hidden states. We hypothesize that properly embedding mass conservation constraints could enhance S4D-FT's ability to simulate both long-term storage dynamics and short-term hydrologic extremes that current S4D-FT does not capture well (e.g., Hortonian runoff and rain-on-snow events) by ensuring physically consistent alignment between cumulative inputs, outputs, and long-term storage change. Moreover, comparative analyses of S4D-FT with and without mass conservation constraints may offer diagnostic insights into how such constraints influence sequence-level water balance fidelity (Wang et al., 2025a, 2025b). It is important to note that, despite the theoretical advancements, it is possible that physics-aware DL comes at the cost of some predictive accuracy (Frame et al., 2022). Nevertheless, we argue that the trade-off between accuracy and interpretability shall be considered worthwhile if it enables hydrologists to trace model predictions back to specific physical processes, deepening our understanding of hydrologic systems and providing insights beyond pure statistical analysis. Looking ahead, we advocate for DL models being tailored to the needs of the hydrology community to achieve a better balance between performance and interpretability following significant progress made thus far (Feng et al., 2022; Hou et al., 2024; Ji et al., 2024; Shen et al., 2023; Tsai et al., 2021; Wang et al., 2025b).

Last but not least, while this work establishes S4D-FT as a strong CONUS-wide benchmark under a widely used evaluation setup, it should be viewed as an initial step. In contrast to LSTM whose performance has been tested across many regions, time scales, and tasks, SSMs (including S4D-FT) are still relatively new to hydrology. Therefore, broader validation is needed to assess generalizability, especially in settings beyond our experiments (e.g., global domains, forecasting, and extremes; see Nearing et al., 2024). We encourage community-wide benchmarking of S4D-FT (and other models) at regional and global scales using common, transparent protocols (shared training and testing splits, consistent metrics, and reported computational cost), to helps ensure that observed differences reflect model behavior rather than experimental setup.

5. Conclusions

This study proposes adopting a first-of-kind S4D-FT model for rainfall-runoff simulations of a total of 531 watersheds in CONUS. Through a comprehensive evaluation of statistical metrics and spatial performance, we demonstrate that S4D-FT outperforms the current leading model, that is, the decades-old LSTM model, in large-scale rainfall-runoff simulations for both in-sample and out-of-sample (i.e., PUB) setups.

Our analysis highlights that S4D-FT excels in watersheds where high- and low-flow events are frequent and prolonged, but with smaller high- and low-flow magnitudes among all study watersheds. These watersheds are often associated with snowmelt-driven regimes that feature strong seasonal cycles, such as those in the Rocky Mountains, and intermittent flow-dominated regions that feature long dry spells, like parts of the Great Plains. However, S4D-FT tends to be less effective in regions with high daily mean and peak streamflow values, such as the pluvio-nival watersheds in the Appalachia and Pacific Northwest. The limited performance of S4D-FT might be associated with less accurate simulations of flashy and high-flow regimes. Such performance strengths and weaknesses likely stem from S4D-FT's architectural design, which models state dynamics through global convolutional kernels in the frequency domain. These kernels are well-suited for capturing low-frequency, smooth dynamics, such as seasonal patterns, slow groundwater release, or long dry spell, but tend to suppress high-frequency, localized signals in flashy hydrologic responses, such as Hortonian runoff or rain-on-snow floods.

To conclude, our findings show that S4D-FT transcends the predictive limits of LSTM in rainfall-runoff simulations over a large number of study cases in CONUS. Nonetheless, to fully realize the potential of S4D-FT in hydrology, broader validation is still needed across regions, timescales, and hydrologic applications. Additionally, extending the model by incorporating physical constraints such as enforcing mass conservation across input-output boundaries offers a promising path toward developing physics-aware variants of S4D-FT with improved physical fidelity and interpretability.

Conflict of Interest

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

Data Availability Statement

The training and testing data (hydrometeorological variables, static catchment attributes, and USGS gauge streamflow) is from the Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) data set (Newman et al., 2014). The Sacramento Soil Moisture Accounting (Sac-SMA) with SNOW-17 simulations are available on HydroShare (Frame, 2022). The Python codes to reproduce the results of this paper is available at https://github.com/Pandas-Paws/S4D_rainfall_runoff_simulations and on Zenodo (Wang, 2025).

Acknowledgments

The University of Oklahoma (OU) team acknowledges support of the National Science Foundation (NSF) CAREER Award (No. 2236926). The OU team would also like to thank the Department of Defense, Army Corps of Engineers (DOD-COR) Engineering With Nature (EWN) Program (Award No. W912HZ-21-2-0038), the U.S. Bureau of Reclamation (USBR) Project No. R24AC00032, and the National Oceanic and Atmospheric Administration (NOAA)'s Climate Program Office, CVP and MAPP programs (Award Number: NA23OAR4310459). AY would like to thank the SciAI Center, funded by the Office of Naval Research under Grant N00014-23-1-2729. NBE would like to acknowledge partial support from the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Scientific Discovery through Advanced Computing (SciDAC) program, under Contract Number DE-AC02-05CH11231. We would also like to thank anonymous reviewers for their valuable feedback and constructive suggestions on this work.

References

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

Agarwal, N., Suo, D., Chen, X., & Hazan, E. (2023). Spectral state space models. *arXiv preprint arXiv:2312.06837*.

Anderson, M., & McDonnell, J. (2005). Sacramento soil moisture accounting model (SAC-SMA). *Encyclopedia of Hydrological Sciences*.

Autio, A., Ala-Aho, P., Ronkanen, A. K., Rossi, P. M., & Kløve, B. (2020). Implications of peat soil conceptualization for groundwater exfiltration in numerical modeling: A study on a hypothetical peatland hillslope. *Water Resources Research*, *56*(8), e2019WR026203. https://doi.org/10.1029/2019wr026203

Beven, K. (1989). Changing ideas in hydrology—The case of physically-based models. *Journal of Hydrology*, *105*(1–2), 157–172. https://doi.org/10.1016/0022-1694(89)90101-7

Beven, K. J. (1996). *Distributed hydrological modelling* (pp. 255–278). Springer.

Beven, K. J. (2012). *Rainfall-runoff modelling: The primer*. John Wiley & Sons.

Beven, K. J., & Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrological Sciences Journal*, *24*(1), 43–69. https://doi.org/10.1080/02626667909491834

Brunner, M. I., Melsen, L. A., Newman, A. J., Wood, A. W., & Clark, M. P. (2020). Future streamflow regime changes in the United States: Assessment using functional classification. *Hydrology and Earth System Sciences*, *24*(8), 3951–3966. https://doi.org/10.5194/hess-24-3951-2020

Clark, M. P., Bierkens, M. F., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., et al. (2017). The evolution of process-based hydrologic models: Historical challenges and the collective quest for physical realism. *Hydrology and Earth System Sciences*, *21*(7), 3427–3440. https://doi.org/10.5194/hess-21-3427-2017

Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D. (2017). *Language modeling with gated convolutional networks* (pp. 933–941). PMLR.

Erichson, N. B., Azencot, O., Queiruga, A., Hodgkinson, L., & Mahoney, M. W. (2020). Lipschitz recurrent neural networks. *arXiv preprint arXiv:2006.12070*.

Erichson, N. B., Lim, S. H., & Mahoney, M. W. (2022). Gated recurrent neural networks with weighted time-delay feedback. *arXiv preprint arXiv:2212.00228*.

Fang, K., & Shen, C. (2017). Full-flow-regime storage-streamflow correlation patterns provide insights into hydrologic functioning over the continental US. *Water Resources Research*, *53*(9), 8064–8083. https://doi.org/10.1002/2016wr020283

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

Frame, J. M. (2022). MC-LSTM papers, model runs [Dataset]. *HydroShare*. https://doi.org/10.4211/hs.d750278db868447dbd252a8c5431affd

Frame, J. M., Kratzert, F., Gupta, H. V., Ullrich, P., & Nearing, G. S. (2023). On strictly enforced mass conservation constraints for modelling the rainfall-runoff process. *Hydrological Processes*, *37*(3), e14847. https://doi.org/10.1002/hyp.14847

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

Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A., Goel, K., Gupta, A., & Ré, C. (2022). On the parameterization and initialization of diagonal state space models. In *Advances in neural information processing systems* (Vol. 35, pp. 35971–35983).

Gu, A., Goel, K., & Ré, C. (2021). Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., & Ré, C. (2021). Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in Neural Information Processing Systems*, *34*, 572–585.

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hasani, R., Lechner, M., Wang, T.-H., Chahine, M., Amini, A., & Rus, D. (2022). Liquid structural state-space models. *arXiv preprint arXiv: 2209.12951*.

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (GELUS). *arXiv preprint arXiv:1606.08415*.

Hochreiter, S. (1997). *Long short-term memory*. Neural Computation MIT-Press.

Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G. S., et al. (2021). *MC-LSTM: Mass-conserving LSTM* (pp. 4275–4286). PMLR.

Hou, Q., Li, Y., Singh, V. P., & Sun, Z. (2024). Physics-informed neural network for diffusive wave model. *Journal of Hydrology*, *637*, 131261. https://doi.org/10.1016/j.jhydrol.2024.131261

Ji, Y., Zha, Y., Yeh, T.-C. J., Shi, L., & Wang, Y. (2024). Groundwater inverse modeling: Physics-informed neural network with disentangled constraints and errors. *Journal of Hydrology*, *640*, 131703. https://doi.org/10.1016/j.jhydrol.2024.131703

Jones, C. N., Ameli, A., Neff, B. P., Evenson, G. R., McLaughlin, D. L., Golden, H. E., & Lane, C. R. (2019). Modeling connectivity of non-floodplain wetlands: Insights, approaches, and recommendations. *JAWRA Journal of the American Water Resources Association*, *55*(3), 559–577. https://doi.org/10.1111/1752-1688.12735

Knapp, H. V., Durgunoğlu, A., & Ortel, T. W. (1991). A review of rainfall-runoff modeling for stormwater management ISWS Contract Report CR 516.

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

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., et al. (2021). Hydrological concept formation inside long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences Discussions*, *2021*, 1–37.

Liu, J., Bian, Y., Lawson, K., & Shen, C. (2024). Probing the limit of hydrologic predictability with the transformer network. *Journal of Hydrology*, *637*, 131389. https://doi.org/10.1016/j.jhydrol.2024.131389

Moradkhani, H., & Sorooshian, S. (2008). *General review of rainfall-runoff modeling: Model calibration, data assimilation, and uncertainty analysis*. Springer.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., et al. (2024). Global prediction of extreme floods in ungauged watersheds. *Nature*, *627*(8004), 559–563. https://doi.org/10.1038/s41586-024-07145-1

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. https://doi.org/10.5194/hess-19-209-2015

Newman, A. J., Sampson, K., Clark, M. P., Bock, A., Viger, R. J., & Blodgett, D. (2014). A large-sample watershed-scale hydrometeorological dataset for the contiguous USA [Dataset]. *UCAR/NCAR*. https://doi.org/10.5065/D6MW2F4D

Parnichkun, R. N., Massaroli, S., Moro, A., Smith, J. T., Hasani, R., Lechner, M., et al. (2024). State-free inference of state-space models: The transfer function approach. *arXiv preprint arXiv:2405.06147*.

Patro, B. N., & Agneeswaran, V. S. (2024). Mamba-360: Survey of state space models as transformer alternative for long sequence modelling: Methods, applications, and challenges. *arXiv preprint arXiv:2404.16112*.

Pham, L. T., Luo, L., & Finley, A. (2021). Evaluation of random forests for short-term daily streamflow forecasting in rainfall-and snowmelt-driven watersheds. *Hydrology and Earth System Sciences*, *25*(6), 2997–3015. https://doi.org/10.5194/hess-25-2997-2021

Rusch, T. K., & Mishra, S. (2021). *UNICORNN: A recurrent model for learning very long time dependencies* (pp. 9168–9178). PMLR.

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., et al. (2023). Differentiable modelling to unify machine learning and physical models for geosciences. *Nature Reviews Earth & Environment*, *4*(8), 552–567. https://doi.org/10.1038/s43017-023-00450-9

Sivapalan, M., Takeuchi, K., Franks, S., Gupta, V., Karambiri, H., Lakshmi, V., et al. (2003). IAHS decade on predictions in ungauged basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal*, *48*(6), 857–880. https://doi.org/10.1623/hysj.48.6.857.51421

Smith, J. T., Warrington, A., & Linderman, S. W. (2022). Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*.

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., et al. (2021). From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature Communications*, *12*(1), 5988. https://doi.org/10.1038/s41467-021-26107-z

Vaswani, A. (2017). Attention is all you need. In *Advances in neural information processing systems*.

Wang, Y. (2025). A deep state space model for rainfall-runoff simulations [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.16988503

Wang, Y., Zhang, L., Erichson, N. B., & Yang, T. (2025a). Investigating the streamflow simulation capability of a new mass-conserving long short-term memory (MC-LSTM) model across the contiguous United States. *Journal of Hydrology*, *658*, 133161. https://doi.org/10.1016/j.jhydrol.2025.133161

Wang, Y., Zhang, L., Erichson, N. B., & Yang, T. (2025b). A mass conservation relaxed (MCR) LSTM model for streamflow simulation across CONUS. *Water Resources Research*, *61*(8), e2024WR039131. https://doi.org/10.1029/2024wr039131

Wu, Q., & Lane, C. R. (2017). Delineating wetland catchments and modeling hydrologic connectivity using LiDAR data and aerial imagery. *Hydrology and Earth System Sciences*, *21*(7), 3579–3595. https://doi.org/10.5194/hess-21-3579-2017

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., et al. (2012). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research*, *117*(D3). https://doi.org/10.1029/2011jd016048

Yang, M., & Olivera, F. (2023). Classification of watersheds in the conterminous United States using shape-based time-series clustering and random forests. *Journal of Hydrology*, *620*, 129409. https://doi.org/10.1016/j.jhydrol.2023.129409

Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research*, *44*(9). https://doi.org/10.1029/2007wr006716

Yu, A., Lyu, D., Lim, S. H., Mahoney, M. W., & Erichson, N. B. (2024). Tuning frequency bias of state space models. *arXiv preprint arXiv: 2410.02035*.

Yu, A., Mahoney, M. W., & Erichson, N. B. (2024). There is HOPE to avoid HiPPOs for long-memory state space models. *arXiv preprint arXiv: 2405.13975*.

Yue, H., Wang, Y., Zhang, L., & Yang, T. (2025). A machine learning-based water supply forecasting model to quantify the impact of snow water equivalent on seasonal streamflow variability over the western US. *Journal of Hydrology*, *660*, 133465. https://doi.org/10.1016/j.jhydrol.2025.133465

Zhang, J., Yue, H., Basirifard, M., Cao, J., & Yang, T. (2025). A Mamba-type of deep state space model for reservoir release simulation with a large-scale verification over 441 dams across CONUS. *Journal of Hydrology*, 134145.

Zhang, L., Yang, T., Gao, S., Fan, M., Lu, D., Xu, H., & Xiao, C. (2025). An alternative ensemble streamflow prediction approach using improved subseasonal precipitation forecasts from the north America multi-model ensemble phase II. *Journal of Hydrometeorology*, *26*(3), 309–326. https://doi.org/10.1175/jhm-d-24-0048.1