https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

Quantification of Data and Model Uncertainty for Deep Learning-based Streamflow Prediction

Kaidi Peng[1], Daniel B. Wright[1], Lei Yan[1], G. Aaron Alexander[1], Yagmur Derin[2]

[1]Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI, USA.
[2]Department of Civil and Environmental Engineering, University of Iowa, Iowa City, Iowa, USA.

Correspondence to: Kaidi Peng (kaidi.peng@wisc.edu)

Abstract. Accurate and timely flood warnings are essential for reducing flooding risks. Achieving this objective requires both accurate data and careful model parameterization. The absence of high-resolution ground-based precipitation observations in many locations leaves relatively low-accuracy satellite-based precipitation products as one of the few alternatives. At the same time, advancements in deep learning-based flood prediction have led to improvements in both accuracy and computational efficiency. Despite the progress made by integrating satellite-based precipitation data and a deep learning-based hydrologic model, the quantification of uncertainty—stemming from both data and model—remains largely unexamined. This gap results in less reliable predictions, as overfitting and limited explainability are common concerns in deep learning. In this paper, uncertainties arising from satellite-based precipitation inputs, streamflow observations, training data variability, and model parameters are jointly addressed. The basin-averaged precipitation uncertainty is represented by a parametric probabilistic distribution function, whose parameters can be inferred from the deep learning-based hydrologic model, without the need for higher-accuracy precipitation "ground truth". Streamflow observation uncertainty is characterized by a multiplicative Gaussian noise. Training data variability is quantified by using a mixture density network, while uncertainty in model parameters is captured through variational inference. Our results demonstrate that this integrated approach to uncertainty quantification enhances both the prediction accuracy and the explainability of ensemble streamflow predictions. It provides a foundation for "uncertainty-aware" deep learning-based streamflow prediction.

1 Introduction

The frequency of flood-related disasters has more than doubled over the past two decades due to climate change (Payton, 2023). Disaster early warning systems are one effective way to mitigate flood risks, reducing flood-related fatalities and economic costs (Najafi et al., 2024). The effectiveness of flood warning systems heavily relies on the performance of hydrologic models, which are often plagued by uncertainties stemming from meteorological inputs, particularly precipitation, and hydrological model parameterization and structure (Gupta et al., 2023). These uncertainties compromise the accuracy of streamflow predictions.

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

Deep learning-based hydrologic models have been developed at various spatiotemporal scales and in different parts of the world (e.g., Gauch et al., 2021; Kratzert et al., 2018; Nearing et al., 2024). Compared to traditional physical hydrologic models, deep learning-based model studies have reported higher accuracy, less simulation time, and better generalization across different watersheds. However, deep learning, like its physically based counterparts, is not immune to uncertainty. Errors in precipitation propagate through the model, interact with model parameters, and are further contaminated by errors in streamflow observations (Zhang et al., 2023). Streamflow observation uncertainty is often overlooked in hydrologic modelling, even though many studies have recognized its importance (Horner et al., 2018; Velásquez & Krajewski, 2024). This uncertainty propagation and interaction exacerbate the model overfitting and limit model explainability. Addressing this issue requires uncertainty quantification for both data and the model.

Uncertainty Quantification (UQ), which enables expressions of confidence (or lack thereof) in model predictions, has been developed for deep learning methods. Many UQ approaches introduce inherent regularization through priors, encouraging more robust learning and reducing the risk of memorizing training noise. By providing a confidence interval or ensemble spread, UQ indicates which predictions are reliable and which are speculative, making the downstream decision process more transparent. Studies have attempted to disentangle and characterize different uncertainty components (Gawlikowski et al., 2023; Hüllermeier et al., 2021; Kendall et al., 2017; Ovadia et al., 2019). Two general types can be classified. One is aleatoric uncertainty, which measures irreducible random variability of training data and new input data. This uncertainty usually includes observation noise in the training data and inherent data variability when characterizing the physical world. Mixture Density Networks (MDNs; Christopher, 1994) have been proven to be well-suited for capturing aleatoric uncertainty. It incorporates extra head layers to model the conditional probability distribution of the streamflow prediction given the input. Another uncertainty component is epistemic uncertainty, which arises from a lack of knowledge of neural network parameters and structures. This term is specified as model parameter uncertainty in this study. Epistemic uncertainty is usually considered by assigning probability distributions over the model parameters. Bayesian Variational Inference (BVI; Graves, 2011) is one of the main frameworks. BVI seeks a variational distribution that is closest to the true posterior distribution of parameters through optimization. It provides a more theoretically rigorous approximation of epistemic uncertainty and is more computationally feasible than other approaches, such as Monte Carlo Dropout (MCD; Gal et al., 2016) and deep ensembles (Lakshminarayanan et al., 2017).

Despite growing interest in UQ, such studies in the context of deep learning-based water prediction remain limited. Existing efforts have primarily focused on either data uncertainty (e.g., Zhang et al., 2023; Liu et al., 2023) or model parameter uncertainty (e.g., Li et al., 2021, 2022; Quilty et al., 2023; Zheng et al., 2024). For example, Klotz et al. (2022) implemented MCD and MDN separately to demonstrate that Long Short-Term Memory (LSTM)-based hydrologic models can produce statistically reliable uncertainty estimates. They concluded that it is very important to achieve holistic benchmarking of estimation uncertainty. Other studies—such as Tran et al. (2024) in reservoir inflow-outflow modelling and Fang et al. (2020) in soil moisture prediction— have applied Gaussian MDN-based data uncertainty and MCD-based model uncertainty. These two studies, however, did not have explicit treatment of observation noise in the training data. In a

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

different application context, van Leeuwen et al. (2025) described uncertainties from training data, parameters, model structure, and test data systematically and combined them to predict the rate of raindrop formation in clouds. They demonstrated that sufficient consideration of different uncertainty sources can reduce the model sensitivity to out-of-distribution input. They advocate for a stronger scientific foundation in UQ for deep learning applications in geosciences. Up to now, no study has focused on comprehensive uncertainty quantification of deep learning-based hydrologic prediction, with joint consideration of uncertainties arising from both data observational noise, data inherent variability, and model parameters.

One particular challenge in hydrologic modelling is data quality control. Satellite-based precipitation products have become the most accessible source of observed precipitation, offering larger spatial coverage than ground-based rain gauges and radar networks, and shorter latency compared to global land-ocean-atmosphere reanalysis data. Nevertheless, the accuracy of satellite-based precipitation products is constrained by deficiencies in both sensors and retrieval algorithms (Lettenmaier et al., 2015; Peters-Lidard et al., 2018). Different methods to quantify and reduce satellite-based precipitation uncertainty have been explored (Guilloteau et al., 2022, 2025; Hartke et al., 2022; Maggioni et al., 2014; Tang et al., 2022; Wright et al., 2017), though most rely heavily on ground-based precipitation measurements. In contrast, Li et al. (2023) and Peng, et al., (2025) proposed to use space-based radar precipitation measurements as a reference, albeit subject to challenges and limitations. As an alternative approach for quantifying precipitation uncertainty, inverse hydrologic modelling has gained some attention (Kirchner, 2009). Studies have tried to recover precipitation uncertainty from hydrologic modelling by Bayesian inference (e.g., Grundmann et al., 2019; Henn et al., 2015, 2018; Liao et al., 2022; Renard et al., 2011; Vrugt et al., 2008; Wu et al., 2021; Zoccatelli et al., 2024), but retrieval accuracy is seriously limited by interactions between precipitation errors and model parameter errors, while model uncertainty is usually overlooked. Furthermore, inverse approaches have not been explored using deep learning-based hydrologic models. Likewise, even if the effects of precipitation uncertainty on traditional physically-based hydrologic models are reasonably well-documented (Hartke et al., 2023; Maggioni et al., 2013; Peng, Wright, Derin, Alexander, et al., 2025; Tang et al., 2023), few studies have incorporated precipitation uncertainty in deep learning-based hydrologic models.

This study proposes a comprehensive UQ framework for deep learning-based streamflow predictions, integrating uncertainties from satellite-based precipitation inputs, streamflow observations, training data variability, and model parameters. By evaluating the impacts of these uncertainty components and how they interact with one another, this work aims to enhance the accuracy of streamflow predictions and mitigate the overfitting of deep learning models. It also examines the extent to which deep learning models can inherently learn and represent precipitation uncertainty, thereby enhancing the model's explainability. This approach promotes uncertainty-aware deep learning-based hydrologic modelling, especially in ungauged watersheds.

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

2 Data and Study Area

Our simulations were conducted in 17 watersheds. Key characteristics of these watersheds, including drainage area, aridity index, annual potential evapotranspiration (PET), annual precipitation, snow fraction, and class, are summarized in Table S1. Drainage areas range from 510 km^2 to 8,236 km^2. The aridity index, defined as the ratio of annual PET to annual precipitation, ranges from 0.5 to 6.3, with higher values indicating greater aridity. The snow fraction is calculated as the proportion of hourly precipitation depths with concurrent air temperatures lower than 0 °C. This fraction ranges from 0% to 21%, with higher values indicating a greater proportion of snow. Watersheds were classified as either natural or disturbed by human activities, following the GAGES-II classification scheme (Falcone, 2011). Of the 17 watersheds, 8 are natural and 9 are disturbed. Figure 1a shows the locations of 17 watersheds across CONUS. Figure 1b visualizes the watersheds with different drainage areas, aridity indices, and snow fractions. The placement of 17 watersheds on the Budyko curve is also presented in Fig. S1. Throughout the paper, individual watersheds are identified by their USGS gauge identifying number.

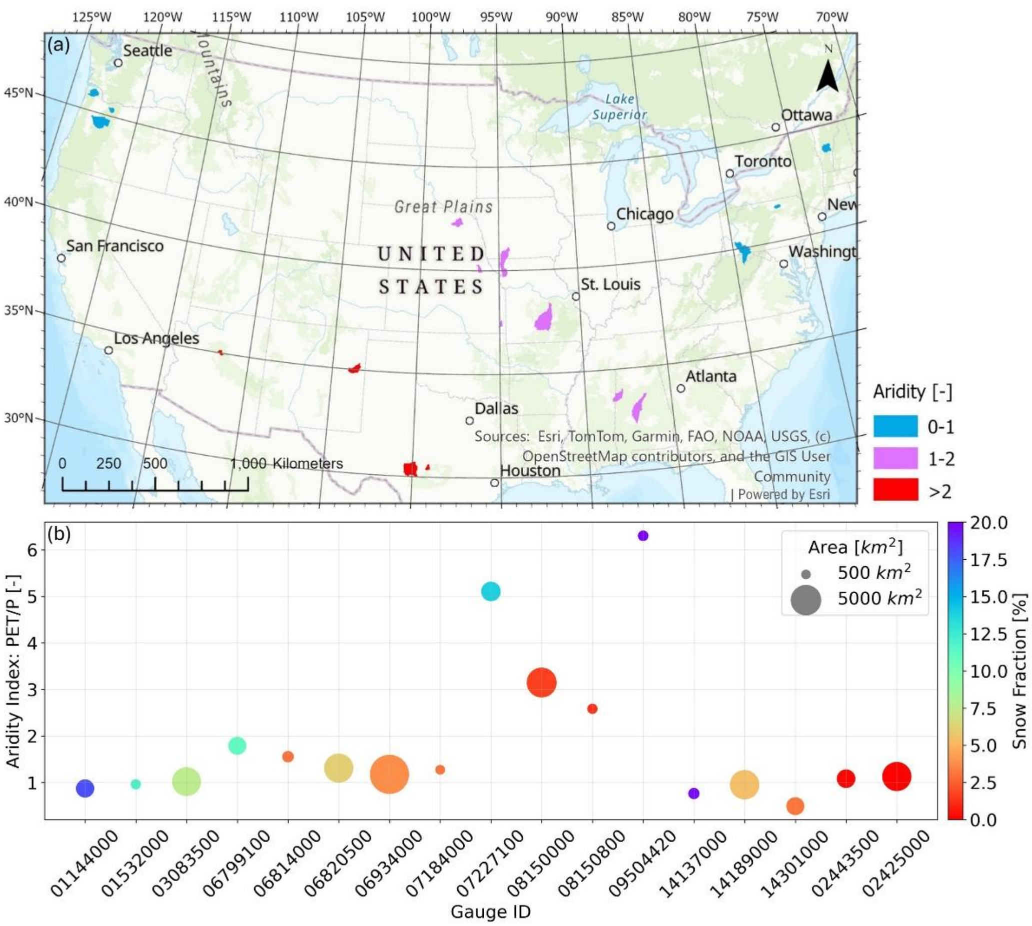


Figure 1. (a) Locations of 17 watersheds across CONUS. Blue watersheds have an aridity index less than 1, pink watersheds have an aridity index between 1 and 2, and red watersheds have an aridity index greater than 2. (b) Watersheds with different drainage areas, aridity indices, and snow fractions.

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

This study is tested on a gridded satellite-based precipitation product—NASA Integrated Multi-satellitE Retrievals for Global Precipitation Measurement (IMERG; Huffman et al., 2023). To evaluate the proposed method, we involved a high-quality ground-based precipitation product —NOAA Analysis of Record for Calibration Dataset (AORC; Fall et al., 2023) as an independent reference. IMERG provides 30-minute, 0.1° gridded precipitation estimates over the entire Earth's surface since 1998 by integrating satellite-based passive microwave retrievals and microwave-calibrated infrared estimates. IMERG has been widely used in precipitation-related disaster modelling (e.g., Gowan et al., 2020; Kirschbaum et al., 2020; Liu, 2023). IMERG offers three products: Early (4-hour latency), Late (14-hour latency), and Final (3.5-month latency), with the latter two incorporating additional data sources and bias corrections. For this study, the "precipitation" field from the most recent IMERG Early V07B dataset was used to leverage its near-realtime capability.

AORC was developed to calibrate land-surface and hydrologic models for improving flood forecasting. It spans the period from 1979 to the near-present for the Continental US with a spatial resolution of 30 arc seconds (~800 m) and a temporal resolution of one hour. It integrates diverse precipitation sources depending on the time period and region. These sources are primarily rain gauges and ground-based weather radar, while satellite data and reanalysis products are used when no ground alternative is available. AORC also provides other meteorological forcing variables: incoming surface shortwave and longwave radiation flux, surface pressure, 2-meter above ground specific humidity, 2-meter above ground temperature, and 10-meter above ground zonal and meridional wind speed. All forcing data were aggregated into hourly watershed-averaged time series.

Streamflow observations from the US Geological Survey were used. 2008 to 2017 (10 years) were used to train the deep learning-based hydrologic model, June 2001 to December 2007 (6.5 years) were used to validate hyperparameters in the model, and 2018 to 2023 (6 years) were used to test the model. Only test period is presented in the results section. In addition, the Moderate Resolution Imaging Spectroradiometer (MODIS) MOD16A2GF Version 6.1 Evapotranspiration product (Running et al., 2021) was used to provide PET for each watershed.

3 Methods

3.1 Long Short-Term Memory-Based Streamflow Prediction

Long Short-Term Memory (LSTM; Hochreiter et al., 1997) is a type of recurrent neural network designed to model long-term dependencies between input and output data. An LSTM manages its internal states through a series of activation functions that regulate the input–state and state–output relationships, making them resistant to vanishing gradients during training. A more detailed description of LSTM and its application in hydrologic models can be found in Kratzert et al. (2018). This study adopts the multi-timescale LSTM framework (MTS-LSTM; Gauch et al., 2021) as a lumped hydrologic model to generate hourly streamflow predictions. MTS-LSTM processes one-dimensional input time series at two scales: a one-year-long daily scale (365 data points) and a fourteen-day-long hourly scale (336 data points). Separate LSTM networks with the same structure are used to handle daily and hourly scales due to their distinct behaviours and information content.

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

Daily states from fourteen days before the prediction time (when hourly input begins) are passed through a linear state transfer layer to initialize hourly states. The hourly LSTM network then processes the last fourteen days to produce hourly streamflow predictions. This method significantly reduces the computational expense compared to relying entirely on hourly time series. Our MTS-LSTM-based streamflow model was built using the NeuralHydrology library in Python (Kratzert et al., 2022).

3.2 Uncertainty Quantification in LSTM

Equation (1) presents the combination of uncertainty sources from precipitation input observations, streamflow observations, training data variability, and model parameters in the LSTM-based streamflow prediction model. The total uncertainty is denoted as $\mathrm{P}(q|x, X, Q)$:

$$\mathrm{P}(q|x, X, Q) = \int\int\int\int \mathrm{P}(q|\mathbf{w}, x^g)\mathrm{P}(x^g|x)\mathrm{P}(X^g|X)\mathrm{P}(Q^g|Q)\mathrm{P}(\mathbf{w}|X^g, Q^g)d\mathbf{w}dx^g dX^g dQ^g. \quad (1)$$

q is the new streamflow prediction. X and X^g are training input precipitation observations from IMERG and precipitation ground truth approximated by AORC, respectively. x and x^g are the new input precipitation observations from IMERG and precipitation ground truth approximated by AORC, respectively. Q and Q^g are training target streamflow observations and streamflow ground truth, respectively. X^g, x^g and Q^g are unknown in our framework. $\mathrm{P}(X^g|X)$ and $\mathrm{P}(x^g|x)$ are denoted as uncertainty distribution for training input and new input precipitation observations from IMERG, respectively. $\mathrm{P}(Q^g|Q)$ denotes the uncertainty distribution for the training target streamflow observations. $\mathrm{P}(\mathbf{w}|X^g, Q^g)$ is uncertainty quantification for LSTM parameters given training precipitation and streamflow data. $\mathrm{P}(q|\mathbf{w}, x^g)$ is the uncertainty in the new streamflow prediction q conditioned on one model parameter ensemble member $\mathbf{w}$ and new precipitation input x^g. This uncertainty arises from the inherent variability of the training samples, reflecting real-world stochasticity, which prevents the model from reaching a definitive prediction. A detailed derivation of Eq. (1) can be seen in Appendix A. Each component in Eq. (1) will be described in the following sub-sections.

3.2.1 Satellite-based Precipitation Uncertainty

Precipitation uncertainty is the primary source of input data uncertainty. In this study, $\mathrm{P}(X^g|X)$ and $\mathrm{P}(x^g|x)$ are modelled using a censored shifted gamma distribution (CSGD) satellite-based precipitation error modelling framework (Li et al., 2023; Wright et al., 2017). The CSGD error model estimates the conditional probability distribution of what the actual precipitation might have been given the observed satellite-based precipitation at each time step t. CSGD modifies the conventional two-parameter gamma distribution with left-censoring, capturing the probability of precipitation, central tendency (i.e., median or mean), and uncertainty using three parameters (Scheuerer et al., 2015): the mean μ_t, standard deviation σ_t, and shift δ_t. Precipitation observation uncertainty distributions are expressed as follows:

$$\mu_t = \frac{\mu}{\alpha_1}\log\left\{(e^{\alpha_1} - 1)\left[\alpha_2 + \alpha_3\frac{X_t}{\bar{X}}\right] + 1\right\} \quad (2)$$

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

$$\sigma_t = \alpha_4 \sigma \sqrt{\frac{\mu_t}{\mu}} \tag{3}$$

$$\delta_t = \delta, \tag{4}$$

$$\mathrm{P}\left(X_t^g \middle| X_t\right) = \begin{cases} f(X_t; \mu_t, \sigma_t, \delta_t), 0 \le X_t \le +\infty \\ 0, \delta_t \le X_t < 0 \end{cases}, \quad (\mu_t > 0, \sigma_t > 0, \delta_t < 0) \tag{5}$$

where $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are regression parameters that, in this study, are estimated through streamflow modelling (see Sect. 3.3). $f(\cdot)$ is the probability density function (PDF) of the gamma distribution. X_t and $\bar{X}$ are IMERG precipitation at time t and its climatological mean, respectively. μ, σ, and δ are derived from a climatological CSGD fitted to IMERG precipitation time series. μ_t, σ_t and δ_t are functions of X_t and identify a distinct distribution at time t. X_t in Eqs. (2)-(5) can be replaced by x_t to represent $\mathrm{P}\left(x_t^g | x_t\right)$. More details about the CSGD parameterization can be seen in Wright et al. (2017) and Scheuerer et al. (2015).

Once all the parameters in Eqs. (2)-(4) are determined, a time-dependent distribution representing precipitation uncertainty at the hourly time scale is obtained. Five quantiles of this distribution (i.e., 10%, 25%, 50%, 75%, and 90%) are then used as precipitation forcings for LSTM to generate streamflow. Daily scale precipitation is obtained by aggregating hourly scale precipitation forcings.

3.2.2 Streamflow Observation Uncertainty

Streamflow ground truth Q^g is modeled as the observation Q corrupted by multiplicative random error following a Gaussian distribution with mean zero and standard deviation ε (i.e., $N(0,\varepsilon)$). We assume that only random errors exist in streamflow observations. ε is assumed to be a constant for each watershed, and it is usually unknown. This streamflow observation uncertainty is usually from stage measurements and rating curve interpolation and extrapolation. Errors in Q increase as Q increases since river stage measurement error causes a larger error in discharge for a river with a larger cross-section. $\mathrm{P}(Q^g|Q)$ can be described as

$$Q^g = Q\left(1 + N(0, \varepsilon)\right) \tag{6}$$

$$\mathbf{P}(\boldsymbol{Q^g}|\boldsymbol{Q}) = \boldsymbol{N}(\boldsymbol{Q}, \boldsymbol{Q\varepsilon}) \tag{7}$$

ε is estimated through streamflow modelling, together with precipitation uncertainty parameters (see Sect. 3.3).

3.2.3 Training Data Variability

Uncertainty due to training data variability is modelled using a type of MDN—Countable Mixture of Asymmetric Laplacians (CMAL) distribution, which was introduced in Klotz et al. (2022). The CMAL distribution combines multiple asymmetric Laplacian components with different weights, allowing it to approximate heavy-tailed or skewed distributions. CMAL is enabled by replacing the standard fully connected regression layer that links intermediate output to the streamflow prediction

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

in the original LSTM architecture with a head layer, which maps the intermediate output to the CMAL parameters. CMAL models the conditional probability distribution of the streamflow prediction given the model input X^g, capturing the prediction uncertainty due to real-world stochasticity. The parameters in the CMAL head layer are estimated together with other parameters in LSTM by maximizing the likelihood of streamflow ground truth Q^g. CAML's formula is given in Klotz et al. (2022):

$$f(q|\eta,s,\tau) = \frac{\tau(1-\tau)}{s} \times \begin{cases} exp[-(q-\eta)(\tau-1)/s], & if\ q < \eta \\ exp[-(q-\eta)\tau/s], & if\ q \geq \eta \end{cases}, \tag{8}$$

$$\mathrm{P}(q|\mathbf{w},x^g) = \sum_{k=1}^{K} \varphi_k^{\mathbf{w}}(x^g) f\left(q|\eta_k^{\mathbf{w}}(x^g), s_k^{\mathbf{w}}(x^g), \tau_k^{\mathbf{w}}(x^g)\right). \tag{9}$$

K is the total number of mixture components in the distribution. It is a hyperparameter to be identified during model validation. K=3 is selected in this study. $\eta_k^{\mathbf{w}}(x^g), s_k^{\mathbf{w}}(x^g), \tau_k^{\mathbf{w}}(x^g)$ and $\varphi_k^{\mathbf{w}}(x^g)$ are mean, scale, asymmetry, and mixture weight of CMAL with component k. They are dependent on input x^g and model parameter **w.** CMAL can be identified once x^g and **w** are estimated.

3.2.4 Model Parameter Uncertainty

The model parameter distribution given the training precipitation data X^g and streamflow data Q^g is estimated by Bayesian variational inference (BVI). The true posterior $P_{max}(\mathbf{w}|X^g, Q^g)$ is often intractable, due to the high dimensionality of the parameter space and infeasible computation needs. Thus, a variational distribution $P_v(\mathbf{w}|\boldsymbol{\theta}, X^g, Q^g)$, where $\boldsymbol{\theta}$ are variational parameters, is used to approximate the true posterior. The objective is to optimize $\boldsymbol{\theta}$ by maximizing the Evidence Lower Bound (ELBO; Graves, 2011):

$$\mathrm{ELBO} = \log\mathrm{P}(Q^g|X^g) - \mathrm{KL}(P_v(\mathbf{w}|\boldsymbol{\theta}, X^g, Q^g)\ ||\ \mathrm{P}(\mathbf{w}|X^g, Q^g)) \tag{10}$$

The first term in ELBO is the log-likelihood of the model prediction, while the second term is the Kullback-Leibler (KL) divergence between the variational distribution $P_v(\mathbf{w}|\boldsymbol{\theta}, X^g, Q^g)$ and the evolving posterior $\mathrm{P}(\mathbf{w}|X^g, Q^g)$. Maximizing the ELBO simultaneously encourages the variational distribution to fit the observed data (via the likelihood term) while remaining close to the evolving posterior of parameters (via the KL term). The Gaussian distribution is the most widely used variational distribution in practice due to its flexibility and computational efficiency. Parameter **w** is quantified by a Gaussian distribution with mean **u** and standard deviation **v**.

$$\mathrm{P}(\mathbf{w}|X^g, Q^g) \simeq P_v(\mathbf{w}|\boldsymbol{\theta}, X^g, Q^g) = N(\mathbf{w}; \mathbf{u}, \mathbf{v}|X^g, Q^g) \tag{11}$$

More details about variational inference can be seen in Ranganath et al. (2014). In each forward run, different model parameter ensembles are obtained by sampling from the variational distribution and used to generate ensemble streamflow predictions.

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

MCD is also an alternative empirical approach for estimating parameter uncertainty. As illustrated in Fig. S2, a comparison between BVI and MCD across our study sites reveals that BVI outperforms MCD in most watersheds. Therefore, in this study, BVI was selected to estimate parameter uncertainty.

3.3 Experimental Setup

Figure 2 illustrates the full workflow for uncertainty quantification in LSTM-based hydrologic modelling. The process begins with estimating precipitation and streamflow observation uncertainty, as detailed in Sect. 3.2.1 and Sect. 3.2.2. The estimated parameters denoted as $\boldsymbol{\psi}=\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \varepsilon\}$ are identified using an expectation-maximization (E-M) algorithm, which is well-suited for obtaining posterior estimates of $\boldsymbol{\psi}$ when the model relies on unknown LSTM parameters.

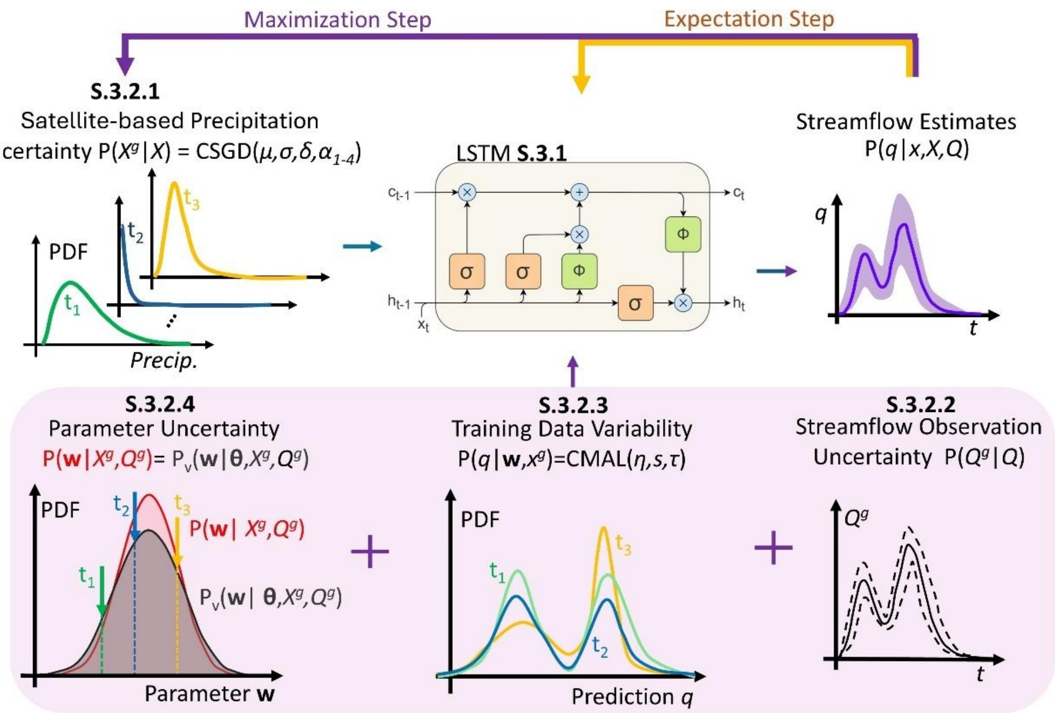


Figure 2. Full workflow for LSTM uncertainty quantification.

In the expectation step of E-M algorithm, deterministic LSTM parameters are optimized under the current estimate of $\boldsymbol{\psi}$ by maximizing the likelihood of streamflow ground truth. One hundred epochs are required to optimize LSTM parameters. In the maximization step, $\boldsymbol{\psi}$ is optimized by maximizing their posterior given streamflow predictions from LSTM. The maximization step employs an iterative ensemble smoother (IES; White, 2018), which has been integrated into the Parameter ESTimation (PEST++; Welter et al., 2012) software suite, to accelerate convergence and reduce computation burden. The prior distribution of $\boldsymbol{\psi}$ is Gaussian, with its bounds set to approximately the 95% confidence interval (i.e., two standard deviations on both sides). The initial centre values and ranges of parameters used in IES are provided in Table 1. One

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

hundred $\boldsymbol{\psi}$ ensemble members are initialized in IES to ensure a large parameter search space. With updated $\boldsymbol{\psi}$, a new set of LSTM parameters can be identified in the expectation step. The streamflow simulations from the optimized LSTM are then used to update $\boldsymbol{\psi}$ for the next maximization step. Typically, four iterations are required for the convergence of $\boldsymbol{\psi}$. This iterative procedure separates precipitation uncertainty from LSTM parameter identification, alleviating the mutual compensation between them. The ensemble median of the estimated CSGD parameters at the fourth iteration is selected as the final estimate of precipitation uncertainty. This method to retrieve CSGD parameters is called CSGD-R.

To evaluate the accuracy of CSGD-R, conventional satellite precipitation error modelling based on CSGD is implemented following the original method in Wright et al. (2017), where AORC serves as precipitation ground truth. This single set of parameter estimates is considered as IMERG's uncertainty ground truth (i.e., CSGD-GT). Comparing CSGD-GT with CSGD-R allows us to evaluate whether deep learning-based hydrologic inversion can accurately recover precipitation uncertainty. Streamflow predictions from LSTMs driven by AORC, IMERG, and CSGD-R are also compared to evaluate the impact of precipitation uncertainty on hydrologic modelling.

Table 1 Initial parameter setting in IES

Name	α_1	α_2	α_3	α_4	ε
Center	0.1	0.2	1.5	0.5	0.3
Range	[0.0001,1]	[0.0001,1]	[0.0001,5]	[0.0001,5]	[0.0001,1]

The estimation of uncertainty from parameter and training data variability is independent from precipitation uncertainty estimation to avoid compensation among different sources of uncertainty. The roles of parameter uncertainty and training data variability are investigated by comparing four different scenarios: 1) deterministic prediction; 2) BVI only to include parameter uncertainty; 3) CMAL only to include training data variability, 4) CMAL+BVI to include both uncertainty components. The accuracy and reliability of streamflow prediction from these four scenarios will be evaluated.

In estimating both model parameter uncertainty and training data variability, only parameters in LSTM's CMAL head layer are subjected to uncertainty quantification, while all other parameters are held deterministic to accelerate convergence in the variational inference. Each CMAL head parameter is assigned to a Gaussian distribution, and a sampling from this distribution generates one parameter ensemble member, yielding a unique CMAL distribution. A total of 20 parameter ensemble members (J = 20 in Eq. (12)) were sampled, resulting in 20 distinct CMALs. These distributions are equally weighted and combined to represent the total prediction uncertainty. Further increasing the LSTM parameter ensemble size does not yield obvious improvements but increases computational cost. 400 independent autocorrelated samplings across time steps were drawn from the combined CMAL to generate streamflow ensemble members that approximately preserve the temporal autocorrelation of streamflow ground truth. This temporal autocorrelation function is estimated nonparametrically using a Fourier transform (Nerini et al., 2017). The computation is as follows:

$$\bar{\eta} = \frac{1}{J}\sum_{j=1}^{J}\sum_{k=1}^{K}\varphi_{jk}^{\mathbf{w}}(x^g)\eta_{jk}^{\mathbf{w}}(x^g) \tag{12}$$

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

$$\Omega_\eta \simeq \bar{\eta} \otimes \Omega = \mathrm{FT}^{-1}\big(\mathrm{FT}(\bar{\eta}) \odot \mathrm{FT}(\Omega)\big) \tag{13}$$

$$\boldsymbol{q} = \mathbf{CDF}^{-1}_{\mathbf{CMALs}}\left(\mathbf{CDF}_{\Omega}(\boldsymbol{\Omega}_{\eta})\right). \tag{14}$$

An autocorrelated Gaussian noise Ω_η is generated by convolving (i.e., $\otimes$) Ω with $\bar{\eta}$, which is equivalent to a point-wise multiplication (i.e., $\odot$) of the Fourier transforms of Ω (i.e., $\mathrm{FT}(\Omega)$) and $\bar{\eta}$ (i.e., $\mathrm{FT}(\bar{\eta})$), followed by an inverse Fourier transform (i.e., FT^{-1}). In this way, the power spectrum of $\bar{\eta}$ is transferred to Ω_η. Finally, CDF mapping is performed between Ω_η and the combined CMAL to produce autocorrelated streamflow samples q. This procedure assumes that the temporal autocorrelation of the prediction mean $\bar{\eta}$ closely approximates that of streamflow ground truth.

In estimating model parameter uncertainty only, the parameters in LSTM's regression head are subjected to uncertainty quantification, with a total of 400 parameter samplings. When only considering training data variability, 400 autocorrelated samplings are drawn from a single CMAL distribution. For deterministic prediction, standard LSTM optimization that minimizes the mean square error (MSE) was conducted.

3.4 Evaluation Metrics

The evaluation of precipitation is against AORC as ground truth. The evaluation of streamflow prediction is against streamflow observation, which is equivalent to the expectation of streamflow ground truth based on our streamflow observation uncertainty assumption.

The continuous ranked probability score (CRPS) and containing ratio (CR) were used to evaluate the probabilistic predictability of ensembles. CRPS measures the divergence between deterministic observations and the ensemble distribution (Thorarinsdottir et al., 2013). CRPS is equivalent to mean absolute error (MAE) when the ensemble size is one. Lower CRPS (or MAE) indicates better performance, with values ranging from zero to positive infinity. CR measures the proportion of instances where the ensemble spread encompasses the observation. CR ranges from 0 to 1, with 1 being optimal. The equations for CRPS and CR are given in Text S1.

Shannon's entropy was used to quantify the uncertainty magnitude. Higher entropy means higher uncertainty. Over- or underestimated uncertainty reduces the reliability of ensemble prediction. The entropy for precipitation uncertainty has closed form solution based on a gamma distribution. The entropy for ensemble streamflow prediction was calculated using the following expression

$$\mathbf{S}(\boldsymbol{q}) = -\int \mathbf{P}(\boldsymbol{q})\mathbf{log}\big(\mathbf{P}(\boldsymbol{q})\big)\boldsymbol{dq} \qquad , \tag{15}$$

where $\mathrm{P}(q)$ is estimated through Gaussian kernel density estimation based on ensembles.

Shapley values (Rozemberczki et al., 2022) quantify the contribution of each input feature to the model's predictions. A positive Shapley value indicates that the feature increases the prediction outcome, while a negative value suggests a decreasing effect. A greater Shapley value reflects a stronger contribution. When evaluating the Shapley value of CSGD-R,

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

the Shapley value of all five quantiles of CSGD-R is summed up. Shapley values were computed using the Python package SHapley Additive exPlanations (SHAP; Lundberg & Lee, 2017).

4 Results

4.1 Retrieved Precipitation Uncertainty Parameters and Distributions

To demonstrate the effectiveness of precipitation uncertainty retrieval in our framework, we visualized CSGD-R error model parameters and compared them with CSGD-GT parameters. 100 sets of CSGD-R parameters were retrieved by IES, while CSGD-GT is a single set of parameters. A set of synthetic study was first conducted with CSGD-GT-driven simulated streamflow with four levels of added uncertainty (i.e., ε={0,0.01,0.03,0.05}) used as streamflow "ground truth." This setting can eliminate other systematic errors from models and evaluate parameter identifiability. The evolution of five parameters (i.e., $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ and ε) for the synthetic study of watershed 03083500 across four IES iterations are shown in Fig. 3a for ε=0 and Fig. S3a,c&e for ε={0.01,0.03,0.05}. To further examine the differences between CSGD-R and CSGD-GT, Fig. 3b and Fig. S3b,d&f present the CDFs of CSGD conditioned on IMERG precipitation values of 0.1, 1, 5, and 10 mm/h with all ensemble members from the fourth iteration of CSGD-R. Results for all the study watersheds in standard retrieval are presented in Fig. S10.

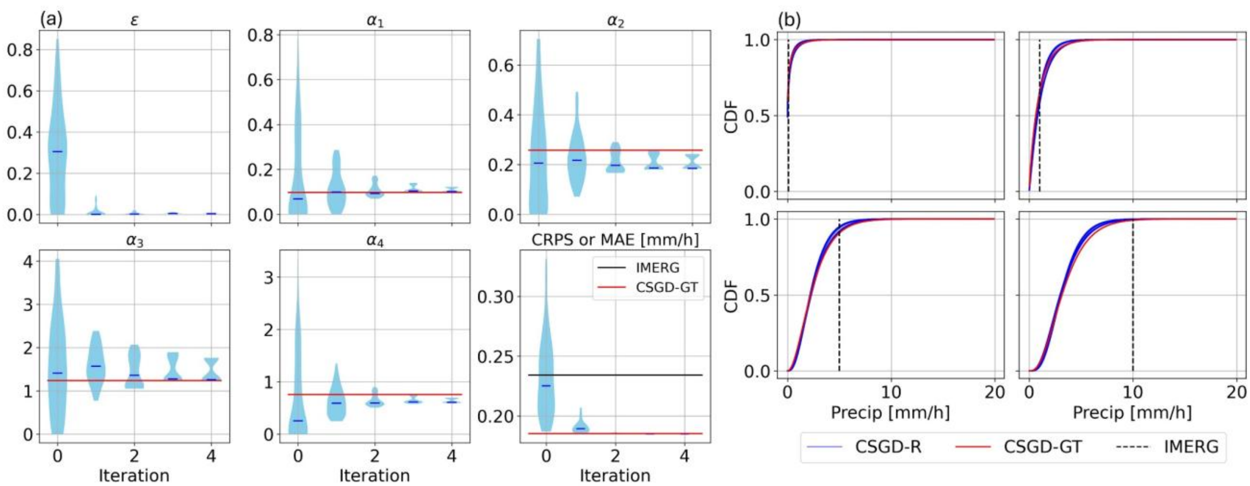


Figure 3. (a) Retrieved CSGD-R parameters' evolution in four iterations. (b) CDFs of CSGD-R and CSGD-GT conditioned on four different IMERG values (i.e., 0.1, 1, 5, and 10 mm/h). All ensemble members in the fourth iteration are shown. The results are for the synthetic study of watershed 03083500 with ε=0.

In synthetic study with ε=0, the retrieved ε effectively converges to zero, compared to a value of around 0.03 in the standard retrieval for 03083500 (Fig. S10e). Meanwhile, both CSGD-R parameters and CDFs in ε=0 case align more closely with the ground truth than those from the standard retrieval and the synthetic study with ε={0.01,0.03,0.05}. While the retrieved ε remains relatively consistent with the actual ε for ε=0.01 and 0.03 in Fig. S3a&c, the retrieval of both

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

precipitation and streamflow uncertainty parameter becomes problematic when ε=0.05 in Fig. S3e&f. This synthetical study reflects the confounding influence of modelling error and streamflow error on precipitation uncertainty retrieval. Our retrieval framework may fail when applied to highly erroneous streamflow data.

For study watersheds in Fig. S10, ε always diminishes as the precipitation retrieval error reduces with increasing iterations. It initially compensates for precipitation errors in the early stages and converges to a low value once CSGD parameters stabilize. The initial values of CSGD-R parameters are broad to ensure sufficient search space. After two iterations, they converged toward true values from CSGD-GT. The CRPS value of CSGD-R closely aligns with CSGD-GT and is lower than the MAE of IMERG after convergence. The CDFs from CSGD-R generally match those of CSGD-GT, while performance varies across watersheds. The dynamic groundwater recharge in 08150000 may cause trouble in convergence, resulting in large uncertainty in parameter identification. The large errors in 09504420 and 07227100 are primarily due to their high aridity level greater than 5. Heavily snow-fed watersheds also show slight biases. For example, 01144000 experiences significant snowmelt in early spring, while 14137000 is affected by rain-on-snow events that dominate winter discharge. Watersheds with larger areas (>4,000 km^2) also tend to exhibit wider ensemble spreads after convergence, such as 06934000 and 03083500, though the ensemble of CDF can generally capture the CDF ground truth. Overall, the best precipitation retrieval performance is achieved in the natural watersheds with small-to-moderate drainage areas (≤2,000 km^2), low aridity (≤1.5), and low-to-moderate snow influence (snow fraction ≤15%).

This conclusion highlights that precipitation retrieval from streamflow is susceptible to degradation when streamflow generation mechanisms are complicated by groundwater recharge, snow processes, and human activities. Landscape heterogeneity and rainfall spatiotemporal variability also challenge retrieval of lumped precipitation uncertainty, as evidenced by relatively degraded performance in large watersheds, since those factors weaken the precipitation-streamflow relationship.

4.2 Performance of Precipitation Retrieval

To further investigate the performance of precipitation retrieval across watersheds, we visualized the effects of drainage area, aridity index, and snow fraction. The CRPS of CSGD-R normalized by mean precipitation for different precipitation quantiles is shown in Fig. 4a. Lower normalized CRPS means a better performance. Normalized CRPS decreases with increasing precipitation quantiles. When averaging all flows, smaller or more arid watersheds exhibit a higher normalized CRPS in Fig. 4b.

To evaluate the benefits of CSGD-R relative to IMERG, a comparison of CRPS of CSGD-R to the MAE of IMERG for different precipitation quantiles is shown in Fig. 5a. A y-axis below 1 indicates that CSGD-R outperforms IMERG. The degradation in precipitation retrieval accuracy at higher precipitation quantiles is more pronounced in arid watersheds than in humid ones, whereas humid watersheds show a more consistent performance across quantiles. For arid watersheds, the reduction in CRPS of CSGD-R relative to MAE of IMERG is greater than humid ones at low precipitation quantiles, while an increase in CRPS of CSGD-R compared with MAE of IMERG occurs at high quantiles. Except for high precipitation

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

quantiles in arid watersheds, the CRPS of CSGD-R remains lower than the MAE of IMERG. This is likely to reflect that limited high-flow data in arid watersheds deteriorates precipitation retrieval accuracy.

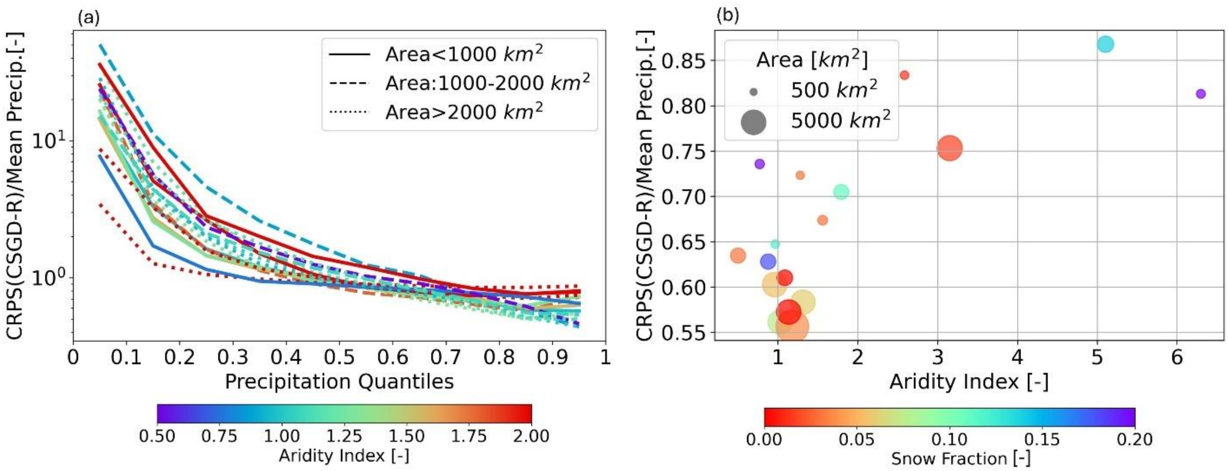
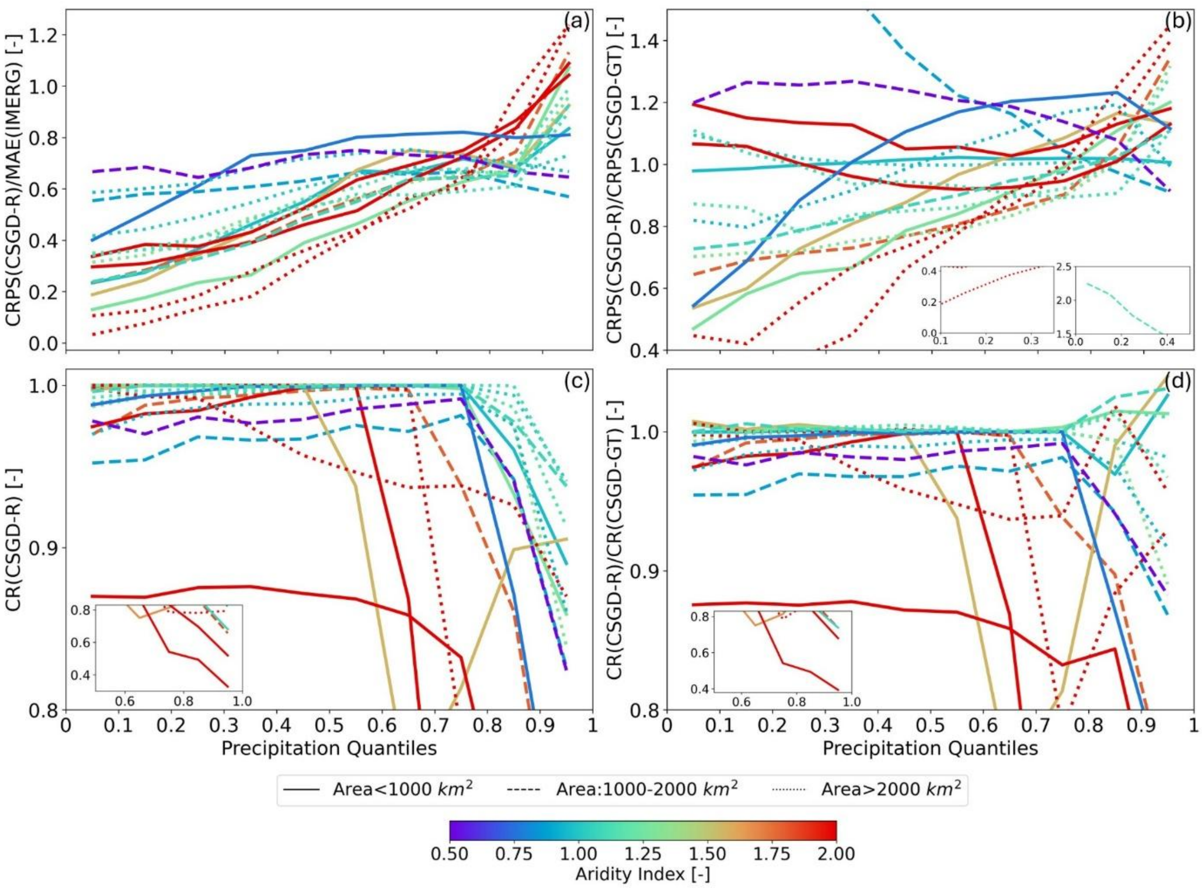


Figure 4. (a) CRPS of CSGD-R normalized by mean precipitation for different precipitation quantiles. (b) is the average of (a) over all precipitation.



https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

Figure 5. CRPS of CSGD-R normalized by (a) MAE of IMERG and (b) CRPS of CSGD-GT for different precipitation quantiles. (c) CR of CSGD-R, and (d) CR of CSGD-R normalized by CR of CSGD-GT for different precipitation quantiles. Truncated portions of each panel are shown in the insets.

When compared to CSGD-GT in Fig. 5b, CSGD-GT generally outperforms CSGD-R (i.e., the y-axis larger than 1) at higher precipitation quantiles across most watersheds. However, CSGD-R occasionally achieves better accuracy than CSGD-GT at low-to-moderate precipitation quantiles. This difference likely arises from the distinct optimization objectives between CSGD-R and CSGD-GT. CSGD-GT is calibrated to minimize the average precipitation CRPS, which is dominated by moderate-to-high precipitation events. In contrast, CSGD-R is primarily optimized using low-to-moderate flows, which can be more readily simulated by the hydrologic model, whereas the large modelling errors associated with high-flow events may make it difficult for CSGD-R to capture high precipitation dynamics. Figure 5c&d show the CR of CSGD-R and its comparison with that of CSGD-GT, respectively. Arid watersheds exhibit more pronounced decreases in CR with increasing precipitation quantiles compared to humid ones. CR of CSGD-R is mostly lower than that of CSGD-GT (i.e., y-axis lower than 1), particularly at high precipitation quantiles.

Figure 6 summarizes the average performance of CSGD-R versus CSGD-GT and IMERG. Values below 1 on the y-axis indicate that CSGD-R outperforms IMERG or CSGD-GT. It reveals a clear trend of superior performance in humid watersheds. For watersheds with similar size and aridity, the ones with larger snow fractions tend to have larger errors. Overall, the performance of CSGD-R is generally better than that of IMERG but worse than that of CSGD-GT.

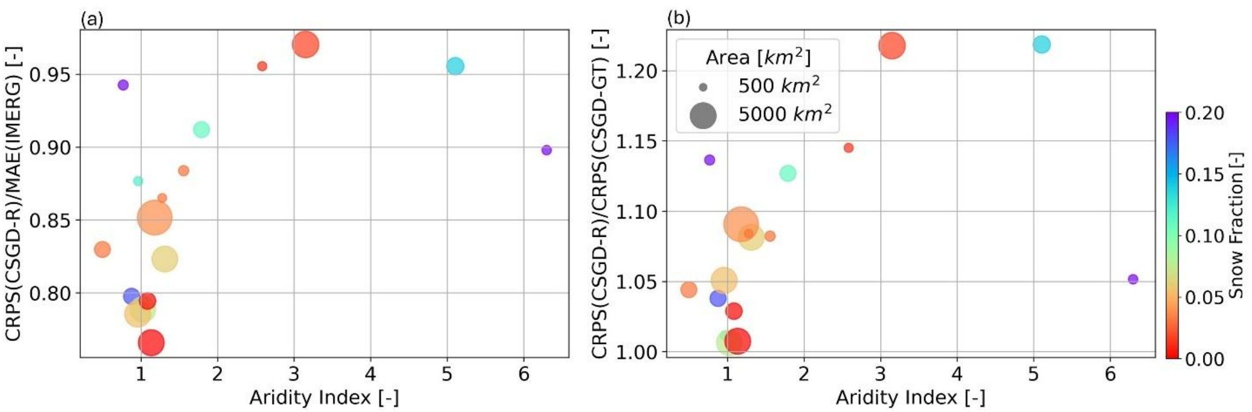


Figure 6. Average of Fig. 5a&b over all precipitation.

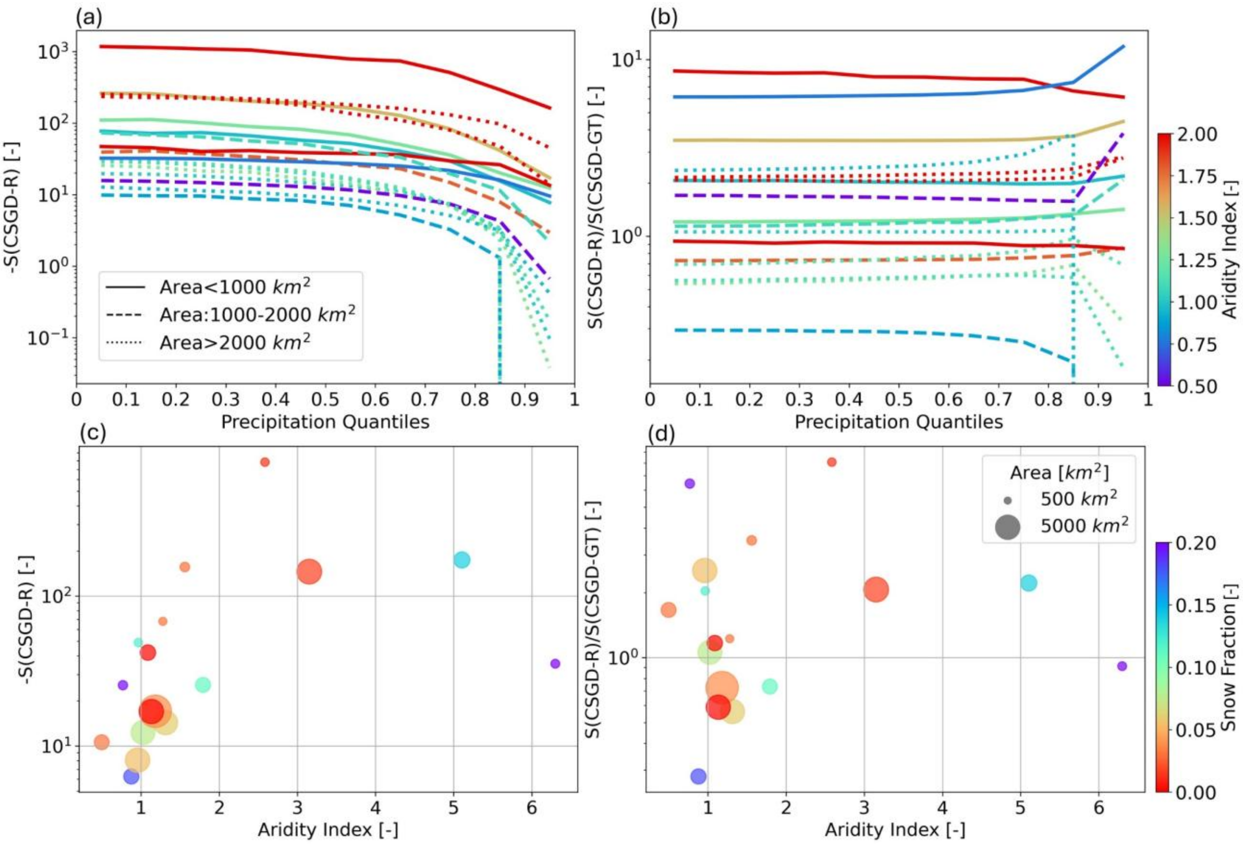
4.3 Uncertainty of Precipitation Retrieval

The entropy of CSGD is used to assess the reliability of precipitation uncertainty retrievals, and can identify over- or under-dispersion. Since the CSGD distribution yields negative entropy in this setting, Fig. 7a presents the negative entropy of CSGD-R on a logarithmic scale, with higher values indicating lower uncertainty. Across all watersheds, uncertainty increases with precipitation intensity. Comparison with CSGD-GT in Fig. 7b shows that the entropy differences remain relatively consistent across precipitation levels, suggesting that CSGD-R captures the scaling of uncertainty with precipitation intensity but introduces systematic bias. Y-axis values below 1 indicate that CSGD-R exhibits higher entropy

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

(i.e., uncertainty) than CSGD-GT, and vice versa. Averaged over all precipitation values (Fig. 7c), more arid watersheds exhibit lower uncertainty. Among watersheds with similar aridity, larger basins tend to have higher uncertainty. The relatively low uncertainty in the severe arid watershed (aridity index >6) is likely attributable to the precipitation retrieval errors.

395 Comparing the average entropy of CSGD-R and CSGD-GT in Fig. 7d further reveals a scale-dependent bias. Larger watersheds are more likely to overestimate uncertainty, while smaller watersheds typically underestimate uncertainty. This pattern highlights again the increased difficulty in parameter identification in larger basins. In contrast, for smaller basins, LSTM can reproduce streamflow dynamics using a relatively narrow range of precipitation realizations, potentially because of a more direct streamflow response to precipitation.



400

Figure 7. (a) negative entropy of CSGD-R and (b) entropy of CSGD-R normalized by entropy of CSGD-GT, for different precipitation quantiles. (c) and (d) are the average of (a) and (b) over all precipitation, respectively.

4.4 Performance of Ensemble Streamflow Prediction

The normalized CRPS of ensemble streamflow prediction driven by CSGD-R and incorporating both training data variability 405 by CMAL and model parameter uncertainty by BVI is shown in Fig. 8. The scenarios driven by IMERG and AORC are shown in Fig. S4. The normalized CRPS exhibits a relatively consistent performance with different streamflow quantiles for

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

large and humid watersheds. In contrast, model performance is more variable for small or arid watersheds, likely due to the highly skewed streamflow distributions used in model training. Averaging all flow regimes, in Fig. 8b, larger and more humid watersheds generally exhibit lower normalized CRPS. When comparing the accuracy of ensemble streamflow driven by CSGD-R, IMERG, and AORC in Fig. 9, AORC shows the best performance for most watersheds, while predictions driven by CSGD-R tend to have better performance than those driven by IMERG in smaller, or more snow-influenced watersheds. This is largely because small basins are more sensitive to precipitation variability, and satellite-based precipitation products often struggle in snow-fed landscapes.

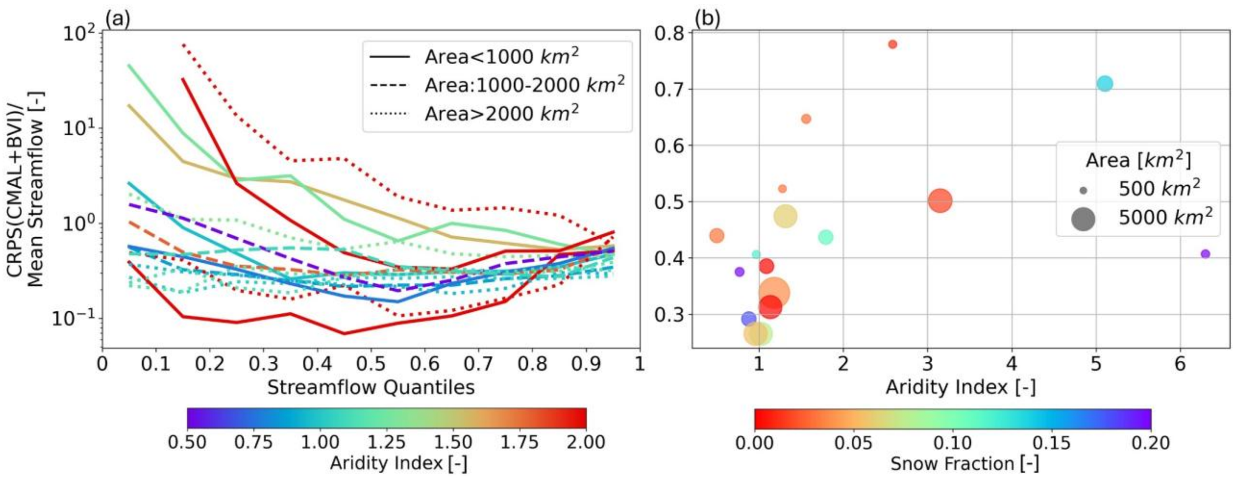


Figure 8. (a) CRPS of ensemble streamflow driven by CSGD-R with both BVI and CMAL, normalized by mean streamflow for different streamflow quantiles. (b) is the average of (a) over all flows.

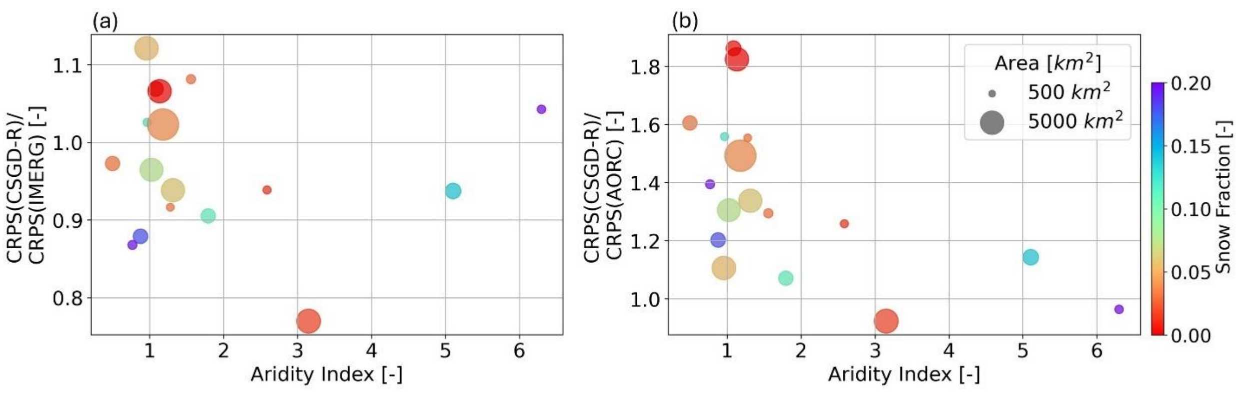


Figure 9. CRPS of ensemble streamflow driven by CSGD-R normalized by CRPS of ensemble streamflow driven by (a) IMERG and (b) AORC. Both BVI and CMAL uncertainty components are included.

To assess the individual and combined contributions of training data variability by CMAL and model parameter uncertainty by BVI relative to deterministic streamflow prediction, Fig. 10a,c&e show CRPS of ensemble streamflow driven by CSGD-R with different uncertainty components, normalized by MAE of deterministic streamflow prediction. Values greater than one indicate worse performance than the deterministic benchmark. The combination of CMAL and BVI

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

outperforms both the deterministic prediction and the ensemble predictions using only CMAL or only BVI. Similarly, CR is higher for the CMAL+BVI version than for the single-component versions. The decrease of CR with BVI is more pronounced than other versions, whereas CMAL results in consistently low CR values (around 0.7) across many watersheds. This pattern suggests that BVI struggles to capture high flow dynamics, while CMAL cannot resolve systematic bias likely from parameterization. Corresponding results for predictions driven by AORC and IMERG are provided in Fig. S5 and S6, and similar conclusions can be drawn.

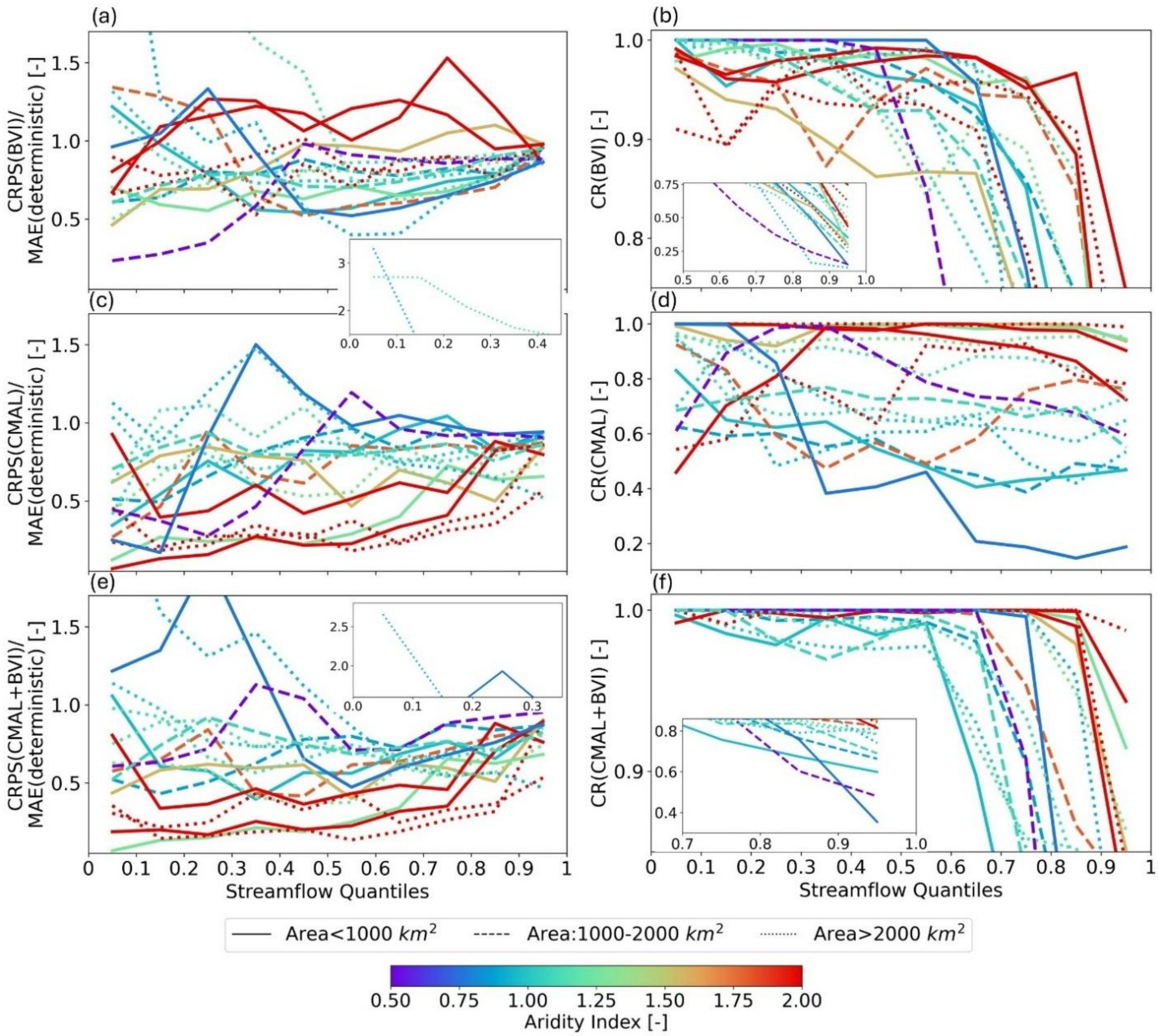


Figure 10. CRPS of ensemble streamflow driven by CSGD-R with (a) only BVI, (b) only CMAL, and (c) both CMAL and BV, normalized by the MAE of deterministic streamflow prediction, for different precipitation quantiles. Truncated portions of each panel are shown in the insets.

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

Averaging different flow regimes, the benefit of BVI remains relatively consistent across different watersheds, reducing the MAE of deterministic predictions by up to 20% (Fig. 11). In contrast, CMAL shows a stronger positive influence for most watersheds. Overall, the ensemble incorporating both CMAL and BVI achieves the best accuracy across watersheds. The results driven by AORC and IMERG in Fig. S7 exhibit similar behavior. This indicates that both training data variability and model parameter uncertainty should not be ignored to achieve accurate ensemble streamflow prediction across different flow regimes and watersheds.

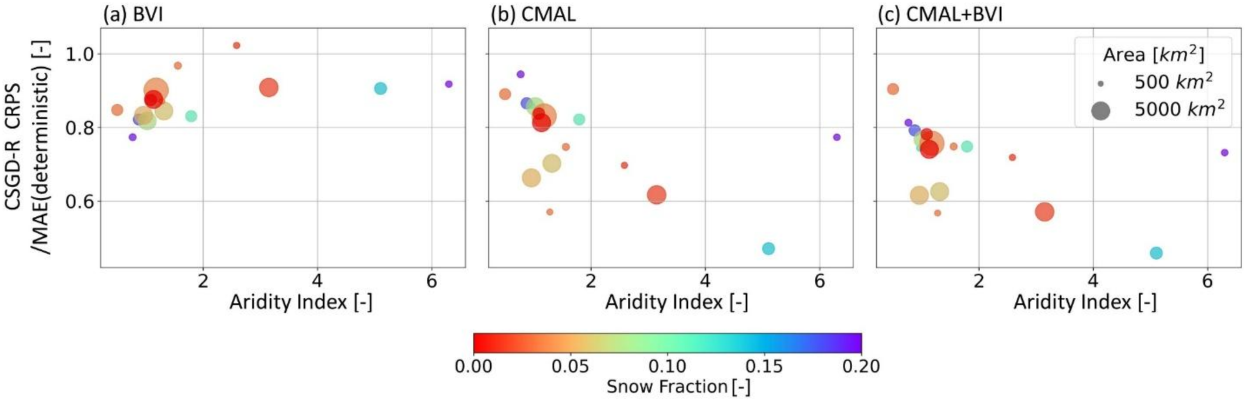


Figure 11. Average of Fig. 10 a,c&e over all flows.

4.5 Uncertainty of Ensemble Streamflow Prediction

To evaluate the contribution of different uncertainty components to the ensemble streamflow predictions, the uncertainty measured by Entropy is presented in Fig. 12a-c for CSGD-R driven scenario and Fig. S8 for the AORC- and IMERG-driven scenarios. The entropy of ensemble streamflow predictions with CMAL+BVI is taken as the reference, given its best performance. The entropy of ensemble streamflow predictions with only CMAL or only BVI is then compared against this reference. Higher y-axis values in Fig. 12a denote greater prediction uncertainty. In Fig. 12b&c, values below 1 show that BVI and CMAL result in lower uncertainty than the CMAL+BVI scenario.

Generally, ensemble streamflow shows a higher entropy with a higher streamflow quantile. Most watersheds show a decrease in entropy from BVI with higher streamflow quantiles. It suggests that a great contribution of uncertainty from BVI for low streamflow quantiles, while a reduced contribution for high streamflow quantiles. This is because the low flow is usually regulated by model parameters to identify soil moisture and evapotranspiration conditions. In contrast, CMAL shows an increasing entropy with higher streamflow quantiles. This aligns with the higher data variability during high-flow events. A similar, stronger contribution of data uncertainty to high-flow events and a stronger contribution of parameter uncertainty to low-flow events was also reported by Peng et al. (2025) for a physics-based hydrologic model. Figure 12d-f and S9 show the average Entropy of ensemble streamflow prediction. BVI tends to overestimate uncertainty, especially for low flow quantiles. The consistently lower level of uncertainty from CMAL also indicates the necessity for parameter uncertainty.

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

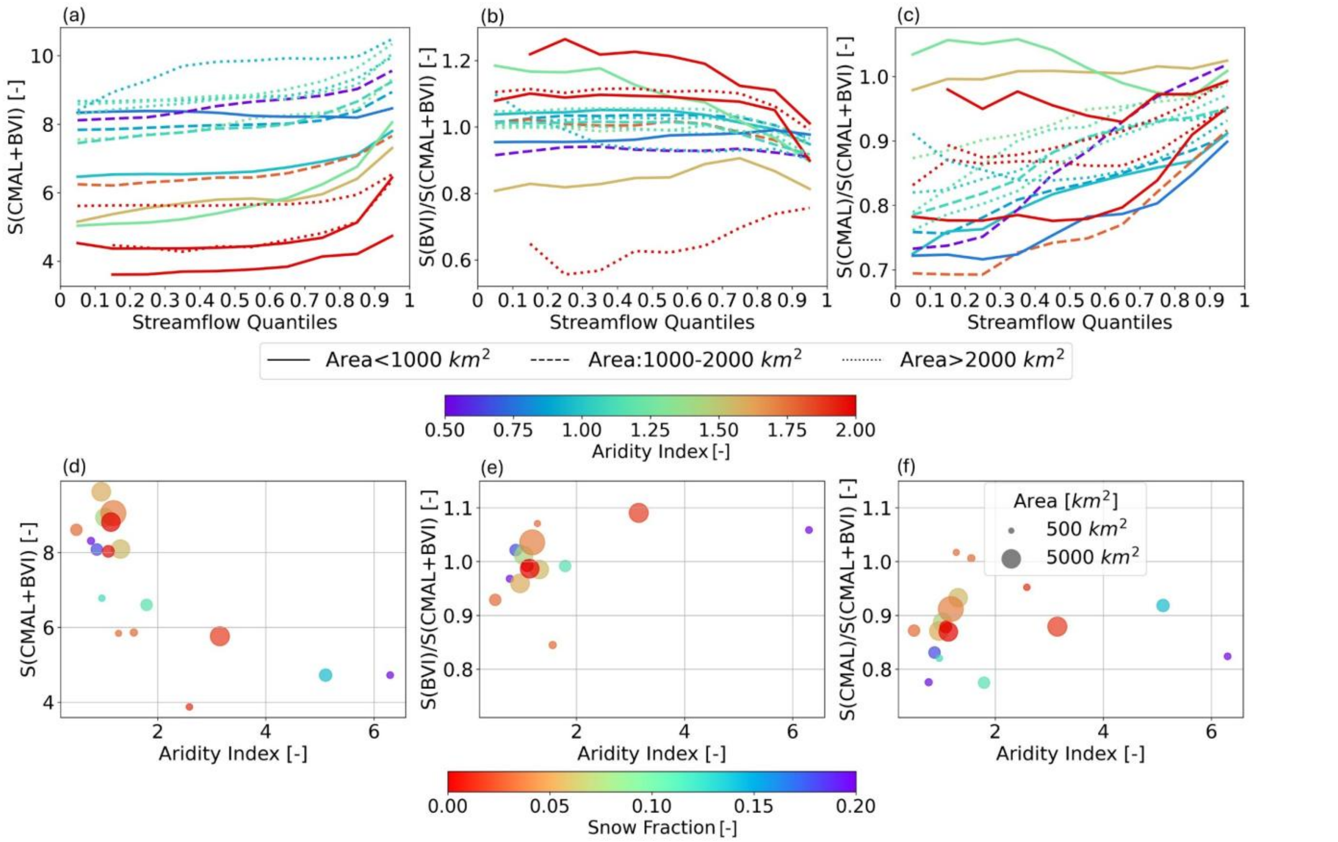


Figure 12. Entropy of ensemble streamflow prediction driven by CSGD-R (a) with both CMAL and BV1, (b) with only BVI and normalized by (a), (c) with only CMAL and normalized by (a), for different streamflow quantiles. (d), (e) and (f) are the average of (a), (b) and (c) over all flows, respectively.

4.6 Precipitation Explainability in Streamflow Prediction

CSGD-R is an uncertainty-quantified version of IMERG that reduces systematic biases and represents random errors through distribution quantiles. We evaluated the extent to which this precipitation uncertainty quantification improves the contribution of precipitation to streamflow prediction. Figure 13a presents the Shapley values of CSGD-R for streamflow prediction across different streamflow quantiles. Higher Shapley values indicate a stronger contribution of precipitation to streamflow. The Shapley value increases with higher streamflow quantiles across all watersheds, since high flows are predominantly driven by precipitation, whereas low flows are more strongly regulated by temperature, humidity, or other forcings. Figures 13b and 13c compare the Shapley values of the CSGD-R scenario with those of the AORC and IMERG scenarios, respectively. Y-axis values greater than 1 indicate a stronger contribution from CSGD-R compared to AORC or IMERG. The differences in Shapley values remain relatively consistent across streamflow quantiles.

Averaging all flow regimes in Fig. 13d-f, humid watersheds exhibit a stronger contribution from precipitation to streamflow prediction than arid watersheds, primarily because they experience more precipitation-driven events. CSGD-R

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

exhibits a higher contribution than AORC in some arid watersheds. This is likely because CSGD-R can better capture precipitation generation stochasticity in arid regions. CSGD-R usually exhibits a lower Shapley value than AORC in humid watersheds, potentially due to biases in the CSGD-R precipitation estimates. When compared with the original IMERG product, CSGD-R generally provides a stronger contribution to streamflow prediction, especially in arid, snow-influenced, and small watersheds. This is consistent with the better performance of CSGD-R than IMERG in driving ensemble streamflow prediction.

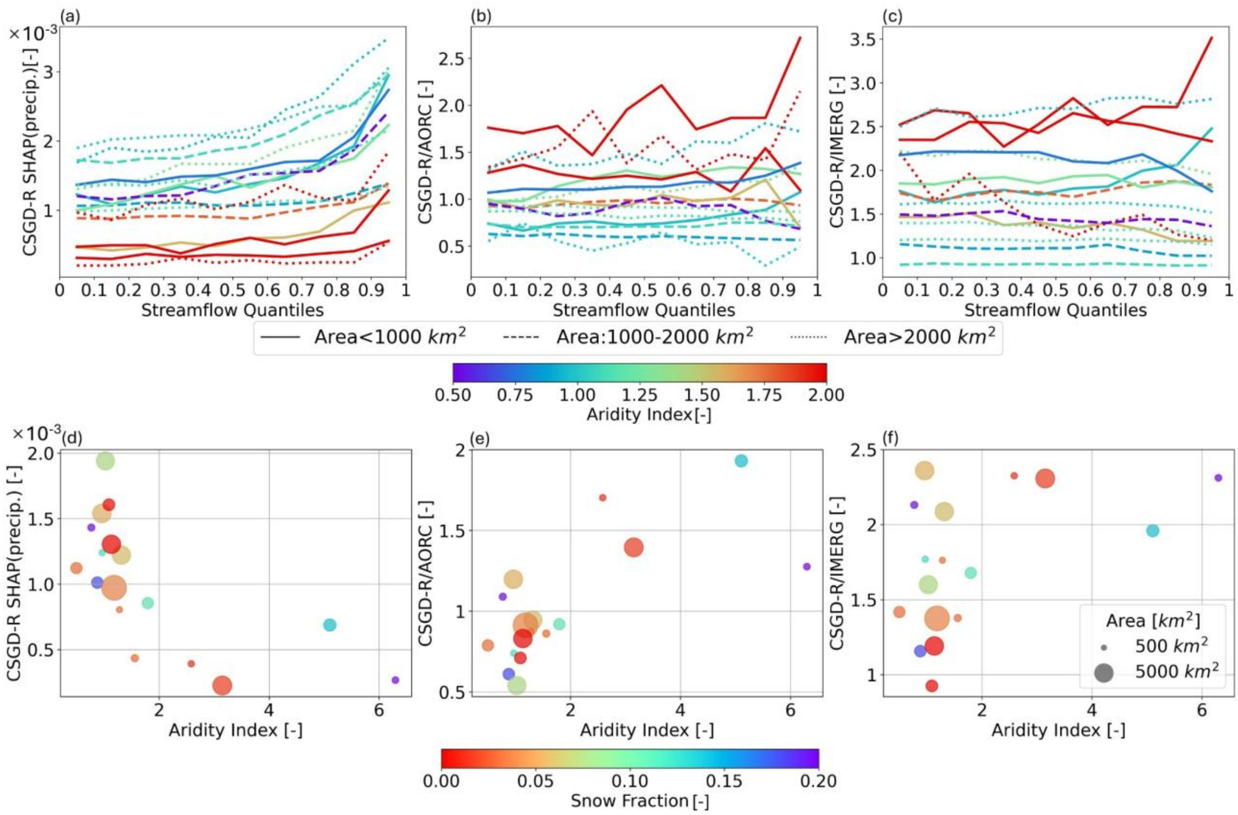


Figure 13. (a) Shapley value of CSGD-R in streamflow prediction for different streamflow quantiles. (b) and (c) are (a) normalized by the Shapley value of AORC and IMERG, respectively. (d), (e) and (f) are the average of (a), (b) and (c) over all flows, respectively.

5 Discussion

5.1 Retrieving Precipitation from Streamflow

Estimating precipitation uncertainty from streamflow data using hydrologic models is a potentially valuable inverse problem in hydrology for regions where precipitation measurements are unreliable (Kirchner, 2009; Vrugt et al., 2008). Precipitation retrieval accuracy depends on the precipitation uncertainty model, the hydrologic model, the retrieval scheme, and the

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

streamflow generation mechanism in the study region. Our results have demonstrated that this inverse modelling may be less suitable in regions where streamflow is primarily driven by groundwater or snowmelt, as these conditions complicate precipitation effects. The method also faces limitations in larger basins, where precipitation spatial variability and landscape heterogeneity are difficult to capture without additional spatial information, often resulting in over-dispersion. Bayesian frameworks are commonly employed for such inverse problems, and in this study, we use the E-M algorithm coupled with IES as our inference method, which is more computationally feasible than Markov Chain Monte Carlo-based methods (Peng et al., 2025). Though physical hydrologic models have been employed in such a retrieval process, our study demonstrates the feasibility of precipitation retrieval in deep learning-based hydrologic models. It investigates the extent to which an LSTM can internally learn and represent precipitation uncertainty by augmenting the precipitation forcing with a set of trainable parameters.

Beyond the parametric precipitation uncertainty distribution used in this study, deep learning methods—such as diffusion models—have recently been proposed for quantifying precipitation uncertainty. However, these were not pursued here, as streamflow alone cannot provide sufficient physical constraints for multiple interconnected deep learning models. As a result, such an approach may struggle to generate physically realistic precipitation uncertainty. Nevertheless, hybrid approaches that combine deep learning-based precipitation uncertainty quantification with physics-based hydrologic models offer promising alternatives. In such frameworks, the governing equations in the hydrologic model can impose constraints on the precipitation generated by the deep learning model. A related differentiable framework has been proposed by Shen et al. (2023) to handle deep learning-based parameter estimation in a conceptual hydrologic model. Further investigation into the integration of different precipitation uncertainty models with different types of hydrologic models is warranted.

5.2 Joint Consideration of Data and Model Uncertainty

Precipitation observation uncertainty, streamflow observation uncertainty, training data variability, and model parameter uncertainty are explicitly considered in our UQ framework. Precipitation uncertainty provides a stronger contribution to the streamflow prediction than the original satellite-based precipitation input, and it is much closer to contribution observed in the precipitation ground truth-driven streamflow prediction. We also observed marginal improvements in ensemble streamflow driven by CSGD-R compared to those driven by IMERG, mainly for arid, small, or snow-influenced watersheds. These watersheds are more sensitive to precipitation variability, and satellite-based precipitation products often struggle in snow-fed landscapes(Derin & Yilmaz, 2014). CSGD-R also sometimes exhibits benefits over AORC in driving streamflow prediction, as it can better capture precipitation generation stochasticity in arid regions.

Second, we show that accounting solely for input data variability in deep learning-based ensemble streamflow prediction is generally insufficient to produce unbiased predictions, as simulated low flows are usually governed by parameterization of soil moisture and evapotranspiration. Conversely, considering only model parameter uncertainty leads to degraded performance during high-flow conditions, as higher data variability exists during high-flow events. Therefore, incorporating both sources of uncertainty is essential for generating reliable ensemble predictions with improved generalization to different

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

flow regimes and watersheds. Overall, our approach offers a promising path toward more transparent and "uncertainty-aware" deep learning-based hydrologic modelling.

5.3 Towards Ungauged Basins

In this study, streamflow observation uncertainty is modelled as a Gaussian distribution conditioned on streamflow observations. However, in practice, errors that arise from the rating curve used to convert river stage to discharge often exhibit systematic biases and non-Gaussian behaviour (Horner et al., 2018; Velásquez et al., 2024), particularly in regions with poor streamflow gauge quality. The identification of more suitable streamflow observation error models may therefore offer greater utility.

However, the reliance on ground-based streamflow gauges still limits the applicability of our framework in ungauged basins. The recent launch of the Surface Water and Ocean Topography (SWOT) satellite offers an unprecedented opportunity to measure river discharge from space. (Durand et al., 2023). SWOT is anticipated to augment ground-based streamflow measurements, improving hydrologic prediction capabilities in poorly gauged basins. While studies have demonstrated SWOT's potential in advancing traditional hydrologic modelling (Elmer et al., 2020, 2021; Yang et al., 2019), its capability in deep learning-based hydrologic models remains underexplored. Integrating SWOT data into our framework introduces two primary challenges: (1) the uncertainty in SWOT-derived discharge data and (2) the limited temporal coverage due to the satellite's recent launch and its orbital sampling strategy, which yields intermittent rather than continuous observations. Addressing the first issue requires accounting for the uncertainty of SWOT discharge. Multi-model ensemble approaches have been proposed for this purpose (e.g., Frasson et al., 2021, 2023), and align with the uncertainty framework used in this study. Specifically, a physically applicable SWOT observation error model can be involved as the streamflow observation uncertainty. The limited training data from SWOT can potentially be mitigated by the generalization of hydrologic models to collect data from a range of watersheds. The generalization capability of the LSTM has been demonstrated in several streamflow prediction studies (Arsenault et al., 2023; Lees et al., 2021; Nearing et al., 2024). Developing a large-scale computationally feasible uncertainty quantification framework relying on satellite-based data remains a key future research direction.

6 Conclusions

High-quality input data and appropriate model parameterization are critical for delivering timely flood warnings. Although the integration of satellite data and deep learning has enhanced the accuracy and timeliness of flood predictions over the globe, few studies have addressed the quantification of uncertainty in both data and models. This ignorance may reduce confidence in predictions, given well-known challenges of overfitting and limited explainability in deep learning models. In this study, we introduce an uncertainty quantification framework that jointly accounts for uncertainties in satellite-based precipitation input, streamflow observations, training data variability, and model parameters. Precipitation uncertainty is

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

recovered independently of ground truth data and is represented as a parametric distribution. Streamflow observation uncertainty is involved by a multiplicative Gaussian noise. MDN and BVI capture training data variability and model parameter uncertainty effectively. By implementing tests in 17 watersheds across the CONUS, our results demonstrate that the comprehensive treatment of uncertainty improves the accuracy and reliability of ensemble streamflow predictions, enhances model explainability, and strengthens generalization to different flow regimes and watersheds. The coupling of the precipitation uncertainty model and the hydrologic model presents opportunities for further exploration. Developing a large-scale computationally feasible uncertainty quantification framework relying on satellite-based data remains a key future research direction.

Appendix A

$$\begin{aligned}
\mathrm{P}(q|x,X,Q) &= \int \mathrm{P}(q,x^g|x,X,Q)\,dx^g = \int \mathrm{P}(q|x^g,X,Q)\,\mathrm{P}(x^g|x)dx^g \\
&= \int\int\int \mathrm{P}(q,X^g,Q^g|x^g,X,Q)\,\mathrm{P}(x^g|x)dx^g dX^g dQ^g \\
&= \int\int\int \mathrm{P}(q|x^g,X^g,Q^g)\,\mathrm{P}(x^g|x)\mathrm{P}(X^g|X)\mathrm{P}(Q^g|Q)dx^g dX^g dQ^g \\
&= \int\int\int\int \mathrm{P}(q,\mathbf{w}|x^g,X^g,Q^g)\,\mathrm{P}(x^g|x)\mathrm{P}(X^g|X)\mathrm{P}(Q^g|Q)d\mathbf{w}dx^g dX^g dQ^g \\
&= \int\int\int\int \mathrm{P}(q|\mathbf{w},x^g)\mathrm{P}(x^g|x)\mathrm{P}(X^g|X)\mathrm{P}(Q^g|Q)\mathrm{P}(\mathbf{w}|X^g,Q^g)d\mathbf{w}dx^g dX^g dQ^g
\end{aligned}$$

Code and data availability

AORC and IMERG Early are available in Fall et al. (2023) and Huffman et al. (2024) respectively. CSGD codes can be found via Peng (2025). PEST++ IES is available at White (2018). NeuralHydrology is available at Kratzert et al. (2022). The SHAP Python package is available at Lundberg et al. (2017). Sample code is available at Peng, (2026).

Author contributions

Conceptualization: KP, DW; Investigation: KP, DW; Supervision: DW, LY, GA, YD; Visualization: KP; Writing original draft: KP; Writing–review and editing: KP, DW, LY, GA, YD.

Competing interests

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

Acknowledgements

The authors thank the Centre for High Throughput Computing at the University of Wisconsin-Madison for computational resources and support.

Financial support

K. Peng was supported by the NASA Precipitation Measurement Mission (Grant Number 80NSSC22K0600) and the NASA Future Investigators in NASA Earth and Space Science and Technology program (Grant Number 80NSSC24K1689). D.B. Wright and Y. Derin were supported by the NASA Precipitation Measurement Mission (Grant Number 80NSSC22K0600).

References

Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F., & Mai, J. (2023). Continuous streamflow prediction in ungauged basins: Long short-term memory neural networks clearly outperform traditional hydrological models. *Hydrology and Earth System Sciences*, *27*(1), 139–157. https://doi.org/10.5194/hess-27-139-2023

Christopher, B. (1994). Mixture Density Networks. *Neural Computing Research Group Report*.

Derin, Y., & Yilmaz, K. K. (2014). Evaluation of Multiple Satellite-Based Precipitation Products over Complex Topography. *Journal of Hydrometeorology*, *15*(4), 1498–1516. https://doi.org/10.1175/JHM-D-13-0191.1

Durand, M., Gleason, C. J., Pavelsky, T. M., Prata de Moraes Frasson, R., Turmon, M., David, C. H., Altenau, E. H., Tebaldi, N., Larnier, K., Monnier, J., Malaterre, P. O., Oubanas, H., Allen, G. H., Astifan, B., Brinkerhoff, C., Bates, P. D., Bjerklie, D., Coss, S., Dudley, R., … Wang, J. (2023). A Framework for Estimating Global River Discharge From the Surface Water and Ocean Topography Satellite Mission. *Water Resources Research*, *59*(4), e2021WR031614. https://doi.org/10.1029/2021WR031614

Elmer, N. J., Hain, C., Hossain, F., Desroches, D., & Pottier, C. (2020). Generating Proxy SWOT Water Surface Elevations Using WRF-Hydro and the CNES SWOT Hydrology Simulator. *Water Resources Research*, *56*(8), e2020WR027464. https://doi.org/10.1029/2020WR027464

Elmer, N. J., McCreight, J., & Hain, C. (2021). Hydrologic Model Parameter Estimation in Ungauged Basins Using Simulated SWOT Discharge Observations. *Water Resources Research*, *57*(10), e2021WR029655. https://doi.org/10.1029/2021WR029655

Falcone, J. (2011). *GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow* [Report]. USGS Publications Warehouse. https://doi.org/10.3133/70046617

Fall, G., Kitzmiller, D., Pavlovic, S., Zhang, Z., Patrick, N., St. Laurent, M., Trypaluk, C., Wu, W., & Miller, D. (2023). The Office of Water Prediction's Analysis of Record for Calibration, version 1.1: Dataset description and precipitation evaluation. *JAWRA Journal of the American Water Resources Association*, *59*(6), 1246–1272. https://doi.org/10.1111/1752-1688.13143

Fang, K., Kifer, D., Lawson, K., & Shen, C. (2020). Evaluating the Potential and Challenges of an Uncertainty Quantification Method for Long Short-Term Memory Models for Soil Moisture Predictions. *Water Resources Research*, *56*(12), e2020WR028095. https://doi.org/10.1029/2020WR028095

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

Frasson, R. P. de M., Durand, M. T., Larnier, K., Gleason, C., Andreadis, K. M., Hagemann, M., Dudley, R., Bjerklie, D., Oubanas, H., Garambois, P.-A., Malaterre, P.-O., Lin, P., Pavelsky, T. M., Monnier, J., Brinkerhoff, C. B., & David, C. H. (2021). Exploring the Factors Controlling the Error Characteristics of the Surface Water and Ocean Topography Mission Discharge Estimates. *Water Resources Research*, *57*(6), e2020WR028519. https://doi.org/10.1029/2020WR028519

Frasson, R. P. de M., Turmon, M. J., Durand, M. T., & David, C. H. (2023). Estimating the Relative Impact of Measurement, Parameter, and Flow Law Errors on Discharge from the Surface Water and Ocean Topography Mission. *Journal of Hydrometeorology*, *24*(3), 425–443. https://doi.org/10.1175/JHM-D-22-0078.1

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16*, 1050–1059.

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021). Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network. *Hydrology and Earth System Sciences*, *25*(4), 2045–2062. https://doi.org/10.5194/hess-25-2045-2021

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., & Zhu, X. X. (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, *56*(1), 1513–1589. https://doi.org/10.1007/s10462-023-10562-9

Graves, A. (2011). Practical variational inference for neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems, NIPS'11*, 2348–2356.

Grundmann, J., Hörning, S., & Bárdossy, A. (2019). Stochastic reconstruction of spatio-temporal rainfall patterns by inverse hydrologic modelling. *Hydrology and Earth System Sciences*, *23*(1), 225–237. https://doi.org/10.5194/hess-23-225-2019

Guilloteau, C., Foufoula-Georgiou, E., Kirstetter, P., Tan, J., & Huffman, G. J. (2022). How Well Do Multisatellite Products Capture the Space–Time Dynamics of Precipitation? Part II: Building an Error Model through Spectral System Identification. *Journal of Hydrometeorology*, *23*(9), 1383–1399. https://doi.org/10.1175/JHM-D-22-0041.1

Guilloteau, C., Kerrigan, G., Nelson, K., Migliorini, G., Smyth, P., Li, R., & Foufoula-Georgiou, E. (2025). A Generative Diffusion Model for Probabilistic Ensembles of Precipitation Maps Conditioned on Multisensor Satellite Observations. *IEEE Transactions on Geoscience and Remote Sensing*, *63*, 1–15. https://doi.org/10.1109/TGRS.2025.3548518

Hartke, S. H., Wright, D. B., Li, Z., Maggioni, V., Kirschbaum, D. B., & Khan, S. (2022). Ensemble Representation of Satellite Precipitation Uncertainty Using a Nonstationary, Anisotropic Autocorrelation Model. *Water Resources Research*, *58*(8), e2021WR031650. https://doi.org/10.1029/2021WR031650

Hartke, S. H., Wright, D. B., Quintero, F., & Falck, A. S. (2023). Incorporating IMERG satellite precipitation uncertainty into seasonal and peak streamflow predictions using the Hillslope Link hydrological model. *Journal of Hydrology X*, *18*, 100148. https://doi.org/10.1016/j.hydroa.2023.100148

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

Henn, B., Clark, M. P., Kavetski, D., & Lundquist, J. D. (2015). Estimating mountain basin-mean precipitation from streamflow using Bayesian inference. *Water Resources Research*, *51*(10), 8012–8033. https://doi.org/10.1002/2014WR016736

Henn, B., Clark, M. P., Kavetski, D., Newman, A. J., Hughes, M., McGurk, B., & Lundquist, J. D. (2018). Spatiotemporal patterns of precipitation inferred from streamflow observations across the Sierra Nevada mountain range. *Journal of Hydrology*, *556*, 993–1012. https://doi.org/10.1016/j.jhydrol.2016.08.009

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

Horner, I., Renard, B., Le Coz, J., Branger, F., McMillan, H. K., & Pierrefeu, G. (2018). Impact of Stage Measurement Errors on Streamflow Uncertainty. *Water Resources Research*, *54*(3), 1952–1976. https://doi.org/10.1002/2017WR022039

Huffman, G., Bolvin, D. T., Joyce, R., Nelkin, E. J., Tan, J., Hsu, K., Kelley, O. A., Nguyen, P., Watters, D. C., West, B. J., & Xie, P. (2023). NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG) Version 07. *National Aeronautics and Space Administration*.

Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J., & Tan, J. (2024). *GPM IMERG Early Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V07* [Dataset]. Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/GPM/IMERG/3B-HH-E/07

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*(3), 457–506. https://doi.org/10.1007/s10994-021-05946-3

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, 5580–5590.

Kirchner, J. W. (2009). Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward. *Water Resources Research*, *45*(2). https://doi.org/10.1029/2008wr006912

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., & Nearing, G. (2022). Uncertainty estimation with deep learning for rainfall–runoff modeling. *Hydrology and Earth System Sciences*, *26*(6), 1673–1693. https://doi.org/10.5194/hess-26-1673-2022

Kratzert, F., Gauch, M., Nearing, G., & Klotz, D. (2022). NeuralHydrology—A Python library for Deep Learning research in hydrology. *Journal of Open Source Software*, *7*, 4050. https://doi.org/10.21105/joss.04050

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

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

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

Li, D., Marshall, L., Liang, Z., & Sharma, A. (2022). Hydrologic multi-model ensemble predictions using variational Bayesian deep learning. *Journal of Hydrology*, *604*, 127221. https://doi.org/10.1016/j.jhydrol.2021.127221

Li, D., Marshall, L., Liang, Z., Sharma, A., & Zhou, Y. (2021). Bayesian LSTM With Stochastic Variational Inference for Estimating Model Uncertainty in Process-Based Hydrological Models. *Water Resources Research*, *57*(9), e2021WR029772. https://doi.org/10.1029/2021WR029772

Li, Z., Wright, D. B., Hartke, S. H., Kirschbaum, D. B., Khan, S., Maggioni, V., & Kirstetter, P.-E. (2023). Toward a Globally-Applicable Uncertainty Quantification Framework for Satellite Multisensor Precipitation Products Based on GPM DPR. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–15. https://doi.org/10.1109/TGRS.2023.3235270

Liao, M., & Barros, A. P. (2022). Toward optimal rainfall – Hydrologic QPE correction in headwater basins. *Remote Sensing of Environment*, *279*, 113107. https://doi.org/10.1016/j.rse.2022.113107

Liu, S., Lu, D., Painter, S. L., Griffiths, N. A., & Pierce, E. M. (2023). Uncertainty quantification of machine learning models to improve streamflow prediction under changing climate and environmental conditions. *Frontiers in Water*, *Volume 5-2023*. https://doi.org/10.3389/frwa.2023.1150126

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Maggioni, V., Sapiano, M. R. P., Adler, R. F., Tian, Y., & Huffman, G. J. (2014). An Error Model for Uncertainty Quantification in High-Time-Resolution Precipitation Products. *Journal of Hydrometeorology*, *15*(3), 1274–1292. https://doi.org/10.1175/JHM-D-13-0112.1

Maggioni, V., Vergara, H. J., Anagnostou, E. N., Gourley, J. J., Hong, Y., & Stampoulis, D. (2013). Investigating the Applicability of Error Correction Ensembles of Satellite Rainfall Products in River Flow Simulations. *Journal of Hydrometeorology*, *14*(4), 1194–1211. https://doi.org/10.1175/JHM-D-12-074.1

Najafi, H., Shrestha, P. K., Rakovec, O., Apel, H., Vorogushyn, S., Kumar, R., Thober, S., Merz, B., & Samaniego, L. (2024). High-resolution impact-based early warning system for riverine flooding. *Nature Communications*, *15*(1), 3726. https://doi.org/10.1038/s41467-024-48065-y

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T. Y., Weitzner, D., & Matias, Y. (2024). Global prediction of extreme floods in ungauged watersheds. *Nature*, *627*(8004), 559–563. https://doi.org/10.1038/s41586-024-07145-1

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

Nerini, D., Besic, N., Sideris, I., Germann, U., & Foresti, L. (2017). A non-stationary stochastic ensemble generator for radar rainfall fields based on the short-space Fourier transform. *Hydrol. Earth Syst. Sci.*, *21*(6), 2777–2797. https://doi.org/10.5194/hess-21-2777-2017

Ngoc Tran, V., Ivanov, V. Y., Tien Nguyen, G., Ngoc Anh, T., Huy Nguyen, P., Kim, D.-H., & Kim, J. (2024). A deep learning modeling framework with uncertainty quantification for inflow-outflow predictions for cascade reservoirs. *Journal of Hydrology*, *629*, 130608. https://doi.org/10.1016/j.jhydrol.2024.130608

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc.

Payton, E. A. (2023). Fifth National Climate Assessment. *Washington, DC, USA: U.S. Global Change Research Program*.

Peng, K. (2025). *KaidiWisc/CSGD_error_model: CSGD error model for Satellite precipitation uncertainty quantification* (Version csgd) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.15485071

Peng, K. (2026). *KaidiWisc/LSTM_UQ: LSTMUQ* (Version Pub) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.19582312

Peng, K., Wright, D. B., Derin, Y., Alexander, G. A., Pradhan, A., Zoccatelli, D., Hartke, S. H., Li, Z., & Tan, J. (2025). Advancing Ensemble Streamflow Prediction Through Satellite-Based Precipitation Product and Model Parameter Uncertainty Quantification. *Journal of Advances in Modeling Earth Systems*, *17*(10), e2025MS004953. https://doi.org/10.1029/2025MS004953

Peng, K., Wright, D. B., Derin, Y., Hartke, S. H., Li, Z., & Tan, J. (2025). STREAM-Sat: A Novel Near-Realtime Quasi-Global Satellite-Only Ensemble Precipitation Dataset. *Water Resources Research*, *61*(3), e2023WR036756. https://doi.org/10.1029/2023WR036756

Quilty, J., Jahangir, M. S., You, J., Hughes, H., Hah, D., & Tzoganakis, I. (2023). Bayesian extreme learning machines for hydrological prediction uncertainty. *Journal of Hydrology*, *626*, 130138. https://doi.org/10.1016/j.jhydrol.2023.130138

Rajesh Ranganath, Sean Gerrish, & David Blei. (2014). Black Box Variational Inference. In Samuel Kaski & Jukka Corander (Eds.), *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* (Vol. 33, pp. 814–822). PMLR. https://proceedings.mlr.press/v33/ranganath14.html

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., & Franks, S. W. (2011). Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. *Water Resources Research*, *47*(11), 2011WR010643. https://doi.org/10.1029/2011WR010643

Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., & Sarkar, R. (2022). The Shapley Value in Machine Learning. In L. D. Raedt (Ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22* (pp. 5572–5579). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2022/778

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

Running, S., Mu, Q., Zhao, M., & Moreno, A. (2021). *MODIS/Terra Net Evapotranspiration Gap-Filled 8-Day L4 Global 500m SIN Grid V061* [Dataset]. NASA Land Processes Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MOD16A2GF.061

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, F., Kifer, D., Li, L., Liu, X., Ren, W., Zheng, Y., Harman, C. J., Clark, M., Farthing, M., Feng, D., Kumar, P., Aboelyazeed, D., … Lawson, K. (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

Tang, G., Clark, M. P., Knoben, W. J. M., Liu, H., Gharari, S., Arnal, L., Beck, H. E., Wood, A. W., Newman, A. J., & Papalexiou, S. M. (2023). The Impact of Meteorological Forcing Uncertainty on Hydrological Modeling: A Global Analysis of Cryosphere Basins. *Water Resources Research*, *59*(6), e2022WR033767. https://doi.org/10.1029/2022WR033767

Tang, G., Clark, M. P., & Papalexiou, S. M. (2022). EM-Earth: The Ensemble Meteorological Dataset for Planet Earth. *Bulletin of the American Meteorological Society*, *103*(4), E996–E1018. https://doi.org/10.1175/BAMS-D-21-0106.1

Thorarinsdottir, T. L., Gneiting, T., & Gissibl, N. (2013). Using Proper Divergence Functions to Evaluate Climate Models. *SIAM/ASA Journal on Uncertainty Quantification*, *1*(1), 522–534. https://doi.org/10.1137/130907550

van Leeuwen, P. J., Chiu, J.-Y. C., & Yang, C.-K. K. (2025). Uncertainty quantification for deep learning. *Environmental Data Science*, *4*, e53. Cambridge Core. https://doi.org/10.1017/eds.2025.10027

Velásquez, N., & Krajewski, W. F. (2024). Effect of streamflow measurement error on flood frequency estimation. *Stochastic Environmental Research and Risk Assessment*, *38*(7), 2903–2910. https://doi.org/10.1007/s00477-024-02707-1

Vrugt, J. A., Ter Braak, C. J. F., Clark, M. P., Hyman, J. M., & Robinson, B. A. (2008). Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research*, *44*(12). https://doi.org/10.1029/2007wr006720

Welter, D. E., Doherty, J. E., Hunt, R. J., Muffels, C. T., Tonkin, M. J., & Schreüder, W. A. (2012). *Approaches in Highly Parameterized Inversion—PEST++, a Parameter ESTimation Code Optimized for Large Environmental Models: Book 7, Section C5*. U.S. Geological Survey. https://pubs.usgs.gov/tm/tm7c5

White, J. T. (2018). A model-independent iterative ensemble smoother for efficient history-matching and uncertainty quantification in very high dimensions. *Environmental Modelling & Software*, *109*, 191–201. https://doi.org/10.1016/j.envsoft.2018.06.009

Wright, D. B., Kirschbaum, D. B., & Yatheendradas, S. (2017). Satellite Precipitation Characterization, Error Modeling, and Error Correction Using Censored Shifted Gamma Distributions. *Journal of Hydrometeorology*, *18*(10), 2801–2815. https://doi.org/10.1175/JHM-D-17-0060.1

Wu, X., Marshall, L., & Sharma, A. (2021). Quantifying input error in hydrologic modeling using the Bayesian error analysis with reordering (BEAR) approach. *Journal of Hydrology*, *598*, 126202. https://doi.org/10.1016/j.jhydrol.2021.126202

https://doi.org/10.5194/egusphere-2026-3466
Preprint. Discussion started: 22 July 2026

Yang, Y., Lin, P., Fisher, C. K., Turmon, M., Hobbs, J., Emery, C. M., Reager, J. T., David, C. H., Lu, H., Yang, K., Hong, Y., Wood, E. F., & Pan, M. (2019). Enhancing SWOT discharge assimilation through spatiotemporal correlations. *Remote Sensing of Environment*, *234*, 111450. https://doi.org/10.1016/j.rse.2019.111450

Zhang, Y., Ye, A., Analui, B., Nguyen, P., Sorooshian, S., Hsu, K., & Wang, Y. (2023). Comparing quantile regression forest and mixture density long short-term memory models for probabilistic post-processing of satellite precipitation-driven streamflow simulations. *Hydrology and Earth System Sciences*, *27*(24), 4529–4550. https://doi.org/10.5194/hess-27-4529-2023

Zheng, F., Yin, H., Zhang, J., Duan, H.-F., & Gupta, H. V. (2024). A Bayesian deep learning approach for video-based estimation and uncertainty quantification of urban rainfall intensity. *Journal of Hydrology*, *640*, 131706. https://doi.org/10.1016/j.jhydrol.2024.131706

Zoccatelli, D., Wright, D. B., White, J. T., Fienen, M. N., & Yu, G. (2024). Precipitation uncertainty estimation and rainfall-runoff model calibration using iterative ensemble smoothers. *Advances in Water Resources*, *186*, 104658. https://doi.org/10.1016/j.advwatres.2024.104658