https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

Improving low flow prediction from hydrologic models using alternative model calibration and post-processing techniques

Tong Wan[1], Charles Kroll[2], and Richard Vogel[3]

[1]Department of Environmental Science, College of Environmental Science and Forestry, State University of New York, Syracuse, New York, USA
[2]Department of Environmental Resources Engineering, College of Environmental Science and Forestry, State University of New York, Syracuse, New York, USA
[3]Department of Civil and Environmental Engineering, Tufts University, Medford, Massachusetts, USA

Correspondence: Tong Wan (twan03@esf.edu)

Abstract. Accurate prediction of low flow series and statistics remains a major challenge in hydrologic modeling. This study evaluates the effectiveness of combining model calibration strategies and post-processing approaches to improve low flow simulation from hydrologic models. WRF-Hydro, a fully distributed deterministic watershed model, is calibrated, post-processed, and evaluated using alternative methods that only require observed and simulated streamflows. Model calibration is performed using alternative objective functions that target different flow magnitudes. This study applies two post-processing approaches, quantile mapping bias correction and stochastic ensemble generation using log-streamflow ratios, to three unregulated watersheds in New York State. The skill of the model simulations and post-processing techniques is evaluated by assessing prediction of low flow series and statistics. Calibration alone could not address conditional bias or reduce the variability of low streamflow estimators. While quantile mapping removes conditional bias, estimators of low flow series and design statistics still exhibited large variability. In contrast, ensemble-based methods led to considerable reductions in both bias and variability of low flow series and design statistic estimators. The ensemble methods performed better when statistics were obtained from an average single streamflow trace than as the average of the statistic across all ensembles. In addition, during a forecasting simulation, resampling of errors from the calibration period was shown to improve low flow estimators during forecast periods when observed streamflows are unknown. These findings suggest that improving low flow simulations requires shifting emphasis from calibration and bias correction methods, toward the development of ensemble-based post-processing approaches.

1 Introduction

Hydrologic drought is a natural hazard related to prolonged periods of less than typical surface and subsurface water supplies, resulting in low streamflow (low flow) conditions that could cause public-safety risks. On a global scale, of all the 20th century natural hazards, droughts had the greatest detrimental impact, due in part to the highly interrelated nature of regional water availability, agriculture, food supply, energy and ecosystems (Bruce, 1994; Obasi, 1994). Even relatively humid regions such as the Northeastern United States have experienced recent drought conditions. For instance, in 2020, below average

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

May–September rainfall produced an extreme hydrologic drought across much of New England with record low flow and groundwater, and in 2022, a similar event prompted water use restrictions and disaster declarations in parts of Rhode Island and Connecticut (Lombard et al., 2021; McCarthy et al., 2023). By late September 2025, about 53% of the Northeast was in drought, with areas of extreme drought (D3; < 5% probability) reported by the U.S. Drought Monitor (USDM) in Maine, New Hampshire, and Vermont (U.S. Drought Monitor, 2025). Improving predictions of low flow is essential for sustainable water resources management, hazard mitigation, and policy decision-making. However, it remains a challenge to provide reliable predictions of low flow series and statistics because they are influenced by multiple factors including variations in meteorological drivers, watershed-based heterogeneity, increasing anthropogenic pressures, and challenges in hydrologic modeling (World Meteorological Organization, 2008). These complexities not only amplify prediction errors but also expose fundamental limitations in how hydrological processes are represented, particularly during hydrologic extremes. As a result, improving the simulation and prediction of low flow has become a central focus of some hydrological modeling research (Tallaksen and Lanen, 2023). Vogel and Kroll (2021) argue that improved guidelines and methods are needed for the prediction of low flows as an analog to existing and well-established national guidelines and methods for flood frequency analysis. They further argue that improvements in low flow prediction should result in improved water resources design, planning, operations, and management under low flow conditions.

Deterministic watershed models (DWMs) are widely used to simulate streamflow processes. Representative process-based DWMs, such as SAC SMA, PRMS, VIC, HBV, and WRF-Hydro, transform meteorological forcings into runoff and enable predictions of hydrological variables at multiple scales (Gochis et al., 2020; Singh, 2012). These models simulate physical processes through mathematical formulations that enforce conservation of mass, energy, and sometimes momentum thereby allowing for spatially and temporally continuous predictions of hydrologic variables (Clark et al., 2017). Their physics-based structure enables interpretability, transferability to ungauged basins, and integration with coupled land atmosphere systems (Johnson et al., 2023).

Unfortunately, DWMs are inherently imperfect representations of reality. Uncertainties arise from multiple sources, including (1) measurement and interpolation errors in meteorological inputs, (2) simplified or incomplete representations of physical processes, and (3) limitations of model calibration and parameter estimation. For example, precipitation estimates at coarse spatial scales may fail to capture localized convective storms, while snowmelt or evapotranspiration modules may rely on empirical parameters that poorly reflect heterogeneous land-surface conditions (Chang et al., 2018; Puma et al., 2016). These uncertainties are embedded within the modeling system, leading to systematic biases and reduced predictive skill, particularly for hydrological extremes (Gupta et al., 2009). Such uncertainty cannot simply be ignored and leads to discrepancies between simulated and observed streamflow values. The strong skewness and heavy tails of streamflow distributions further complicate model calibration by adversely impacting common error measures such as the Nash Sutcliffe Efficiency (NSE) (Clark et al., 2021; Lamontagne et al., 2020). Moreover, systematic evaluation of low flow processes remains limited, and available studies indicate only modest skill in the ability of DWMs to reproduce low flow series (Davison and van der Kamp, 2008).

Some DWM studies have attempted to improve low flow predictions by applying model calibration metrics that put more emphasis on smaller streamflows (Nicolle et al., 2014; Pushpalatha et al., 2012). Pushpalatha et al. (2012) investigated the value

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

of various transformations by computing the Nash Sutcliffe efficiency (NSE) on logarithms (LNSE), square-root (SNSE) or inverse-flow transformations. However, many spatially-distributed models are generally more suitable for flood prediction than low flow prediction, similar to the Weather Research and Forecasting Hydrological Modeling System (WRF-Hydro) employed in this study. This is due to their oversimplified representation of groundwater processes and large computational demands, so that model calibration for low flows can be challenging and produce biased and highly variable streamflow estimators (Bosompemaa et al., 2025). Furthermore, traditional model calibration has limited ability to address streamflow bias and variability for two main reasons: (1) models are typically calibrated over the entire hydrograph rather than representative events, and (2) calibration commonly relies on a single performance metric (e.g., NSE, SNSE, LNSE), which is insufficient to capture the full spectrum of model behavior.

To mitigate these limitations, statistical post-processing has emerged as a critical component of deterministic watershed modeling (Li et al., 2017). The integration of stochastic post-processing with DWMs offers a way to enhance predictions of extreme events such as floods and droughts and effectively bridge the gap between process-based modeling and real-world observations. Post-processing refers to methods applied after model calibration to correct systematic biases, align statistical moments of model predictions with observations, and quantify predictive uncertainty (Li et al., 2017). One form of systematic bias in streamflow predictions, known as conditional bias, arises due to the application of DWM's without accounting for model error (i.e. through post-processing). Such bias is termed conditional bias, because the streamflow output from any DWM represents a conditional mean value of streamflow on a given day and thus will generally have lower variance (and all other upper moments) than the streamflow observations used to calibrate the model. For example, Figure 3 in Farmer and Vogel (2016) documents the systematic bias in all streamflow observations which results from not performing post-processing. Li et al. (2017) review dozens of post-processing methods mostly drawn from the meteorology literature. In contrast, Shabestanipour et al. (2023) note that the literature on post-processing methods for use with long-range simulation, which are the focus of this study, is relatively sparse compared with the extensive applications in areas of meteorology and short-term flood forecasting reviewed by Li et al. (2017).

Farmer and Vogel (2016) document that the need for post-processing becomes most apparent when DWMs are used to estimate extreme streamflows. Biases that appear small at moderate flows can result in substantial relative errors under high or low flow conditions (Farmer and Vogel, 2016; Krzysztofowicz, 2014). Consequently, bias correction techniques such as quantile mapping (QM) have been widely adopted to restore statistical consistency and reduce bias across flow regimes, including both high and low flows (Hashino et al., 2007; Li et al., 2017). QM typically maps the simulated cumulative distribution function (CDF) for streamflows to the observed CDF, leading to reproduction of the marginal distributions of the simulations and observations over the modeling period (Bosompemaa et al., 2025; Liu et al., 2022). For example, Liu et al. (2022) applied several QM variants to WRF-Hydro simulations. Their results indicated that post-processing methods improved simulated streamflow relative to the raw model output, even though the QM approaches exhibited comparatively larger uncertainties and greater sensitivity to model calibration. Similarly, Bosompemaa et al. (2025) evaluated bias-correction techniques for improving regional streamflow predictions from a national-scale hydrologic model. Their study applied a QM method termed the Flow Duration Curve (FDC) method which maps the FDC of the simulated streamflows to the FDC of the

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

observed streamflows. By systematically adjusting the streamflow magnitudes across the full range of exceedance probabilities, the FDC approach substantially reduces biases, particularly during low flow conditions.

With recent developments in machine learning techniques, recurrent neural networks, for example, long short-term memory (LSTM) networks have been used as post-processors to improve streamflow predictions from physical hydrologic models (Cho and Kim, 2022; Xiang et al., 2020; Xiao et al., 2025). In a recent application to WRF-Hydro simulations over the Sierra Nevada Mountains, an LSTM post-processor trained on historical monthly streamflow corrected systematic WRF-Hydro errors and improved seasonal forecast skill by modeling temporal error dependence (Xiao et al., 2025). LSTMs effectively corrected systematic errors while maintaining physical consistency with the underlying WRF-Hydro simulations, notably improving low flow representation during the dry season and reducing RMSE in April–July seasonal streamflow forecasts compared to QM bias corrections. However, improvements in historical flow reconstruction did not always transfer directly to improved ensemble forecast performance, indicating that additional uncertainties in meteorological forcing, initial hydrologic states, and ensemble generation procedures limit the post-processing gains. This discrepancy highlights an important limitation of deterministic or mean-corrected forecasts for low flow applications. Although LSTM substantially improves bias and average error metrics, correcting the average flow alone does not ensure a reliable representation of the distributional tails that contribute to the extreme low flow characteristics. In addition, the improvements were achieved when the LSTM was implemented as a physics-informed post-processor that incorporated routed streamflow together with hydrological state variables from WRF-Hydro. When trained using streamflow information alone, without incorporating meteorological forcings or additional hydrologic state variables, the pure data-driven LSTM did not outperform the calibrated raw WRF-Hydro simulations (Xiao et al., 2025). LSTMs reliance on substantial training data, input and output sequences, and careful temporal representation may restrict application at sites with insufficient data.

For drought monitoring, environmental flow management, and water-supply operations, accurately characterizing low flow risk requires more than just bias reduction through post-processing. It also necessitates generating statistically consistent streamflow ensembles that reproduce both the magnitude and variability of low flow extremes. Ensemble streamflow prediction provides such a framework by quantifying uncertainty arising from model structure, residual error dynamics, and forcing variability (Li et al., 2017; Schaake et al., 2007). An attractive and relatively simple post-processing approach was suggested by Shabestanipour et al. (2023), who transformed a DWM into a stochastic watershed model by adding error to the model outputs to generate streamflow ensembles. Their results showed that their log streamflow ratio stochastic post-processing framework substantially improved the representation of predictive uncertainty and reduced systematic bias across flow regimes. It also enhanced the original deterministic model's performance in reproducing both low flow and high-flow statistics, while providing the capability to generate streamflow predictions under future scenarios.

This study aims to explore an advanced methodological framework for simulating low flow series and statistics using DWMs to (1) test the impact of common calibration methods on estimating low flow series, (2) evaluate the extent to which stochastic post-processing methods can improve deterministic hydrological simulations of streamflow extremes during both calibration and forecasting periods, and (3) provide a comparative assessment of their applicability at three watersheds in New York State. We apply two techniques for model post-processing: bias correction using FDC quantile mapping and ensemble

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

generation using the Shabestanipour et al. (2023) log streamflow ratio approach. These techniques are applied to the spatially-distributed and computationally intensive WRF-Hydro DWM. Due to the long processing times of WRF-Hydro, generating streamflow ensembles from climatic input traces becomes computationally intensive. Statistical post-processing therefore may provide a computationally efficient alternative for improving model skill, removing conditional bias, and generally improving reproduction of low flow events.

2 Experimental Design

2.1 Study Sites

This study focuses on three watersheds shown in Figure 1 with U.S. Geological Survey (USGS) stream gaging stations located within the Catskill/Delaware Watersheds of New York State: Schoharie Creek at Prattsville (USGS 01350000), Esopus Creek at Allaben (USGS 01362200), and the West Branch Delaware River at Walton (USGS 01423000). These stations are all within the New York City (NYC) water supply system, which provides nearly one billion gallons of unfiltered drinking water daily to over nine million people (New York City Department of Environmental Protection, 2025). The Schoharie Creek gage at Prattsville, NY, with a drainage area of 237 mi², is located upstream of the Schoharie Reservoir, which is the northernmost reservoir in the Catskill system. The Esopus Creek gage at Allaben, NY, with a drainage area of 63.7 mi², monitors the primary headwaters that flow into the Ashokan Reservoir, a major storage facility supplying the Catskill and ultimately the NYC distribution network. The West Branch Delaware River gage at Walton, NY, with a drainage area of 332 mi², represents the largest unregulated headwater basins contributing to the Cannonsville Reservoir in the Delaware system, supplying both NYC and downstream compact obligations to New Jersey, Pennsylvania, and Delaware. Together, these three sites form the upstream boundary of the principal source reservoirs that sustain NYC's water supply. These sites are located within steep, forested headwater catchments of the New York City water supply system, characterized by thin glacial soils overlying fractured bedrock typical of the Appalachian Plateau region (Wolock, 2003). These hydrogeologic settings lead to rapid infiltration and shallow subsurface storage, with streamflow during non-storm periods dominated by lateral groundwater flow through soil–bedrock interfaces and near-stream zones rather than deep regional aquifers (Dunne and Black, 1970).

As a result, baseflow dynamics are relatively smooth and predictable, and are commonly approximated using linear or near-linear reservoir representations at daily timescales (Brutsaert and Nieber, 1977; Mohan and Hunt, 2026; Vogel and Kroll, 1992). Sustained shallow groundwater discharge leads to perennial flow conditions throughout the year, and low flow variability is primarily governed by variations in groundwater storage and recession behavior.

2.2 Data and Model

While this experiment could be applied to any DWM, we use NCAR's Weather Research and Forecasting Hydrological Modeling System (WRF-Hydro v5.2). WRF-Hydro is a spatially distributed open-source modeling system and has been incorporated into past versions of NOAA's National Water Model to provide operational, high spatial resolution streamflow

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

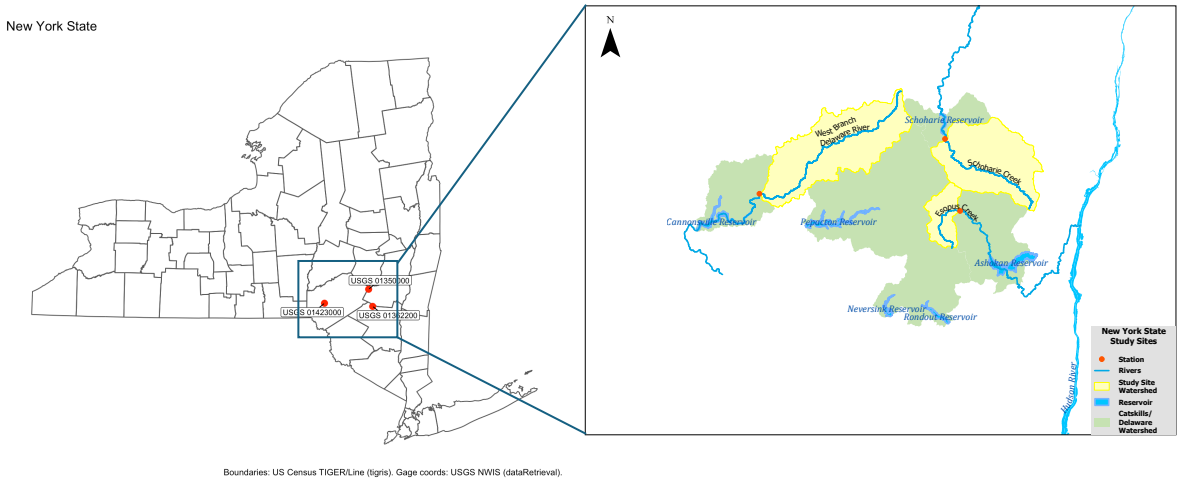

Figure 1. Study sites in New York State, USA.

forecasts (Cosgrove et al., 2024; Gochis et al., 2020; Maidment, 2017). The default groundwater module which we employ in WRF-Hydro is a simple bucket-baseflow module where groundwater discharges to the channel (baseflow) as a linear function of storage (e.g., an exponential discharge model during periods with no groundwater recharge) (Gochis et al., 2020; Niu et al., 2011). Such a simplistic groundwater model should perform well in New England where groundwater dynamics are often described by a linear reservoir model particularly as watershed area increases (Mohan and Hunt, 2026; Vogel and Kroll, 1992). While WRF-Hydro is run at an hourly time-step, here we calibrate WRF-Hydro using hourly streamflows aggregated to a daily time-step. The WRF-Hydro simulation is executed with a gridded configuration with a 1 km horizontal resolution and 1 km river routing resolution.

Our forcing data for this experiment is the North American Land Data Assimilation System Phase 2 (NLDAS-2) dataset, which has been described in Xia et al. (2012). This dataset represents information on essential meteorological variables including wind, air temperature, specific humidity, surface pressure, surface radiation and precipitation (National Aeronautics and Space Administration, 2023). The NLDAS-2 dataset provides a spatial resolution of 1/8th degree (approximately 12.5 km) and a temporal resolution of one hour, making it a valuable tool for detailed analysis of climatic variables (Xia et al., 2012). NLDAS-2 is the common forcing data for WRF-Hydro in the US, and thus is used in this analysis (Lahmers et al., 2019; Lin et al., 2018). At each watershed, during the periods for the model calibration and validation, we initially examined the relationships between daily average precipitation from NLDAS-2 and the observed streamflow to make sure that there was a reasonable relationship between these two series (e.g., streamflow rises after a large precipitation event). We also examined the relationships between daily average precipitation across each site from NLDAS-2 and the recorded daily average precipitation at the nearest monitoring station to confirm general consistency. Daily streamflow recorded by the USGS are used as the observed streamflow data in this study.

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

2.3 Methods

2.3.1 Model Set Up

The primary objective of this experiment is to develop an advanced model calibration and post-processing framework designed for improving the assessment of low flow conditions. The WRF-Hydro model was configured separately for each watershed using a ten-year spin-up period with default parameter settings to establish stable initial hydrologic states. Here a water year is defined as extending from April 1st of the current year to March 31st of the following year; the annual low flow season typically occurs in late summer to early fall in New York State, so the water year captures the typical annual low flow season. This definition differs from the high flow water year used by most agencies in the US, which runs from October 1st to September 30th; using this water year may divide the annual low flow period across two water years. The spin-up period for all three sites extends from April 1, 2000 to March 31, 2010.

2.3.2 Model Calibration

Following a ten-year spin-up period, the model calibration was conducted at each study site using three different calibration metrics. All model calibrations are performed using the dynamically dimensioned search (DDS) method (Tolson and Shoemaker, 2007). Parameters to be calibrated were selected based on the 18 parameters (Tables 1 and 2) chosen by National Center for Atmospheric Research (NCAR) for calibrating the National Water Model (NWM) (Dugger et al., 2017). Initially the model is calibrated by maximizing the commonly employed Nash-Sutcliffe efficiency (NSE) of the flows:

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

where $Q_{obs,i}$ and $Q_{sim,i}$ are the observed and simulated streamflows at day i, respectively, $\overline{Q_{obs}}$ is the average of $Q_{obs,i}$, and n is the total number of observations. NSE is a scaled, or nondimensional real-space sum of squared errors metric, and thus should produce the same result as any squared error metric (e.g., mean squared error).

Initially, 150 iterations of the DDS algorithm are applied in the calibration with NSE to determine the starting values for the 18 model parameters. Of the 18 parameters initially calibrated, 8 parameters, specifically those in categories vegetation, snow, and channel (Table 2) are set as constants. The remaining 10 parameters, which should have a larger impact on surface and subsurface runoff, groundwater, and channel processes (Table 1), are further calibrated in this experiment with their starting values set to the values obtained from the initial calibration. The model is then calibrated for an additional 100 DDS iterations using three different metrics: NSE, log-space NSE (LNSE), and square root of NSE (SNSE).

LNSE is defined as:

$$LNSE = 1 - \frac{\sum_{i=1}^{n} \left(\ln(Q_{obs,i}) - \ln(Q_{sim,i})\right)^2}{\sum_{i=1}^{n} \left(\ln(Q_{obs,i}) - \overline{\ln(Q_{obs})}\right)^2} \quad (2)$$

where $\ln(Q_{obs,i})$and $\ln(Q_{sim,i})$ are the log-transformed observed and simulated streamflows at day i, respectively, $\overline{\ln(Q_{obs})}$ is the mean of $\ln(Q_{obs,i})$, and n is the total number of observations. The LNSE is commonly used to calibrate models focused on

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

Table 1. Ten primary model parameters fully calibrated in this study.

Parameter	Description	Category
BEXP	Pore size distribution index	Soil parameters
SMCMAX	Saturated soil moisture content, i.e., porosity	Soil parameters
DKSAT	Saturated hydraulic conductivity	Soil parameters
REFKDT	Surface runoff parameter. REFKDT is a tunable parameter that significantly impacts surface infiltration and hence the partitioning of total runoff into surface and subsurface runoff. Increasing REFKDT decreases surface runoff.	Runoff parameters
SLOPE	Linear scaling of "openness" of the bottom drainage boundary	Runoff parameters
RETDEPRTFAC	Multiplier on retention depth limit	Runoff parameters
LKSATFAC	Multiplier on lateral hydraulic conductivity, which controls anisotropy between vertical and lateral conductivity	Runoff parameters
Zmax	Maximum groundwater bucket depth	Groundwater parameters
Expon	Exponent controlling rate of bucket drainage as a function of depth	Groundwater parameters
Coeff	Coefficient of the bucket model	Groundwater parameters

low flow, as these transformations increase the weight assigned to smaller observations as opposed to larger observations by reducing the scale and skewness in streamflow observations (Huang et al., 2013; Koch et al., 2018; Pushpalatha et al., 2012). In addition, a logarithmic transformation reduces the variability of the NSE, which has been shown to be an unstable estimator of model efficiency (Lamontagne et al., 2020). Since none of the watersheds considered in this experiment have intermittent streamflow, the log transformation does not require additional handling of zero streamflows.

Pushpalatha et al. (2012) explored many different metrics for evaluating low flow series, including the SNSE of daily flows across all sites. This metric is a balance between NSE and LNSE, with less weight than NSE on the largest observations and less weight than LNSE on the smallest observations, and has been shown to fit flow over a wide range of streamflow conditions (Oudin et al., 2006; Pushpalatha et al., 2012). SNSE is defined as:

$$SNSE = 1 - \frac{\sum_{i=1}^{n} \left(\sqrt{Q_{obs,i}} - \sqrt{Q_{sim,i}}\right)^2}{\sum_{i=1}^{n} \left(\sqrt{Q_{obs,i}} - \overline{\sqrt{Q_{obs}}}\right)^2} \quad (3)$$

where $\overline{\sqrt{Q_{obs}}}$ is the average of $\sqrt{Q_{obs,i}}$. While we can't be certain that the DDS algorithm converges to a global optimal solution for every calibration metric, we ensured consistency across the calibration metrics by verifying that it yields the best value for each calibration metric tested.

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

Table 2. Eight secondary parameters calibrated only during the initial calibration stage in this study.

Parameter	Description	Category
CWPVT	Canopy wind parameter	Vegetation parameters
VCMX25	Maximum carboxylation at 25 °C	Vegetation parameters
MP	Slope of Ball–Berry conductance relationship	Vegetation parameters
HVT	Canopy top height	Vegetation parameters
BtwWdth	Bottom width of channel	Channel parameters
ChSlp	Channel side slope	Channel parameters
N	Manning's N	Channel parameters
MFSNO	Melt factor for snow depletion curve	Snow parameters

2.3.3 Post-processing on Simulated Variables: Bias Correction and Ensemble Stremflow Generation

Hydrological model uncertainty arises from multiple sources including input data, model structural formulations, streamflow measurement error and parameter estimation, and these sources of uncertainty can lead to systematic errors in simulated flows. Statistical post-processing techniques could be applied to reduce bias and potentially improve the predictive skill of the simulation outputs (Brown and Seo, 2010; Li et al., 2017; Moges et al., 2021). In this study, we assess two post-processing techniques: FDC bias correction using quantile mapping (QM), and ensemble generation (EG) of streamflow realizations using Shabestanipour et al.'s (2023) log streamflow ratio approach.

QM is widely implemented for hydrologic model outputs and it adjusts the statistical distribution of simulated variables so that their empirical cumulative distribution function (cdf) aligns with that of observed data (Li et al., 2010; Wood et al., 2004). This method establishes a transfer function between the simulated and observed quantiles and applies it to correct the simulated time series. By modifying model outputs rather than parameters, QM can remove systematic biases remaining after calibration (Bum Kim et al., 2021; Cannon et al., 2015). Previous studies show that QM is effective at correcting bias and improving the representation of extreme events such as droughts and floods (Bosompemaa et al., 2025; Maraun, 2013).

Here streamflow quantiles are mapped based on a Flow Duration Curve (FDC) to adjust daily simulated streamflows at each gaging site to match the corresponding observed flow distributions (Bosompemaa et al., 2025; Farmer et al., 2018). This mapping function preserves the rank order of the simulated flows while adjusting their magnitudes to conform to the observed flow regime. The resulting adjusted streamflow values were used as the bias-corrected simulated series. This method (QM) typically employs the entire FDC. Bosompemaa et al. (2025) found this FDC-based correction performed particularly well for low flows. To ensure that the bias in simulated streamflow was effectively corrected across all flow magnitudes, the FDC mapping technique was also applied separately to each quartile within each water year, referred to here as QMQ. In this approach, daily streamflows within a water year are first divided into four groups based on observed flow quartiles: Q1 (0–25%); Q2 (25–50%); Q3 (50–75%) and Q4 (75–100%). The FDC mapping is then performed independently within each

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

quartile group. By allowing separate bias-correction relationships for different flow regimes, the QMQ approach captures seasonal variations in streamflow prediction more effectively, resulting in less bias in both high-flow and low flow regimes.

In this study we generated streamflow ensembles from the calibrated watershed model to represent residual uncertainty in simulated flows. The core assumption is that the deterministic model produces a single "best estimate" of the conditional mean of the daily streamflow on a given day, but that there remains unrepresented model error variability which should not be ignored (Farmer and Vogel, 2016). Following Shabestanipour et al. (2023), we assume that the log ratio of observed to simulated streamflow contains a stochastic component that can be modelled, and that by adding this error component back to the deterministic simulation we can produce ensemble traces of the streamflow series. We first compute the log of the innovation ratio:

$$I_t = ln(Q_{obs,t}/Q_{sim,\,t}) \tag{4}$$

for each day t over the 6-year calibration period and then fit an autoregressive model (AR) to this series. Various lag AR models were explored to identify a model where the resulting residuals have no remaining significant serial correlation; for these watersheds, a lag-8 AR model (AR(8)) with parameters denoted as c and φ was found to be most appropriate:

$$I_t = c + \varphi_1 I_{t-1} + \varphi_2 I_{t-2} + \varphi_3 I_{t-3} + \varphi_4 I_{t-4} + \varphi_5 I_{t-5} + \varphi_6 I_{t-6} + \varphi_7 I_{t-7} + \varphi_8 I_{t-8} + \varepsilon_t \tag{5}$$

Similar to Shabestanipour et al. (2023), we then partitioned residuals from this model into months, and then randomly resampled these monthly residuals, ε_t, to preserve the seasonal structure of model errors to obtain random sequences of I_t. Starting on the 9th day of the simulation, we then apply this I_t back to the deterministic simulation to generate an ensemble streamflow trace:

$$Q_{ens,i}(t) = \frac{Q_{sim}(t)}{\exp(I_t)} * BCF \tag{6}$$

where BCF is a transformation bias correction factor that is needed to address the bias introduced during the transformation of I_t from log space to real space (Shabestanipour et al., 2023):

$$BCF = \exp(\mu_I - \frac{\sigma_I^2}{2}) \tag{7}$$

where μ_I and σ_I^2 here represent the mean and variance of the I_t series. In this experiment, 1000 ensemble traces were generated.

The ensemble technique requires both the simulated and observed flow series. For the validation (forecast) period, one would not have observed streamflows. To address this, two different approaches were employed to generate ensemble streamflows. In the first approach, observations were assumed to be available, and ensemble streamflows were generated using an AR(8) model derived from validation-period simulated and observed flows. This approach assesses the consistency and robustness of the post-processing framework across different time periods without introducing additional predictive assumptions. In the second approach, observations were assumed to be unavailable and ensemble streamflows were generated using the estimated calibration-period AR model. Residuals were resampled from the calibration-period simulations and observations, thereby mimicking a true forecasting scenario. This method evaluates the ability of this modeling framework to forecast streamflow under conditions where no contemporaneous observations are available.

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

2.3.4 Model Evaluation

The model skill was evaluated by comparing deterministic simulations before and after post-processing to observed streamflow series and statistics, with a particular emphasis on the low flow regime. During the calibration period, model performance was assessed by comparing original deterministic simulations, QM bias-corrected simulations, and the ensemble streamflow traces to evaluate the effectiveness of the calibration and post-processing framework in improving model prediction skill. During calibration, the ensemble traces were evaluated in two ways: (1) computing the mean of the statistic across all ensemble traces (ME), and (2) from the statistics derived from a single mean ensemble trace (EM). In the validation period, two scenarios were considered. The first mirrors the calibration setup, with ensemble traces evaluated in EM and ME. The second is analogous to a forecasting period, ensemble traces were evaluated using the single mean of the statistic across all ensemble forecasts (MEF) as well from statistics derived from a single mean ensemble mean forecast (EMF).

Low flow model skill was assessed using the entire FDC, the lowest quartile (below the 25th percentile) of daily streamflows in each year, annual FDC quantiles with an exceedance probability of 90% (Q_{90}), 95% (Q_{95}), and 99% (Q_{99}), and annual minimum 3-day, 7-day, 14-day and 30-day average flows. These streamflow series and associated statistics describe the frequency, magnitude, and duration of hydrologic drought events, and both FDC quantiles and d-day annual minimums were chosen because they are common low flow design statistics (Wiley, 2006; World Meteorological Organization, 2008).

The same low flow series and statistics that were used to assess the model during the calibration period were also used during the validation period. Through this evaluation, the model's ability to reproduce and forecast critical low flow conditions was assessed, demonstrating its potential utility for drought forecasting and water resource management.

3 Results and Discussion

3.1 Bias Duration Curve

Farmer and Vogel (2016, Figure 3) introduced a Bias Duration Curve (BDC), which is an evaluation tool developed to diagnose and visualize the magnitude of model bias across the full spectrum of streamflow conditions. To create a BDC, at each quantile of the FDC, the simulated streamflow quantile is subtracted from the observed streamflow quantile, and that quantity is then divided by the observed streamflow quantile, resulting in a curve of percent bias versus exceedance probability.

Figure 2 presents BDCs for the Schoharie Creek at Prattsville, NY study site, with each figure corresponding to one of the three calibration metrics: NSE, LNSE, and SNSE. Different colors on the BDC correspond to the deterministic streamflow (Qsim) and the three post-processing methods: QM, ME and EM. Both ME and EM are based on 1000 ensembles. Results for the other two watersheds were similar and thus are not shown here. Results for QM, which performs bias correction across the entire FDC, and QMQ, which performs bias correction across each quarter of a year separately, were nearly indistinguishable, thus only QM is shown here.

As expected, the original calibrated simulations (Qsim) systematically overestimated low flows and underestimated high flows with levels of conditional bias that vary across calibration metrics and flow regimes. This is most apparent for the model

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

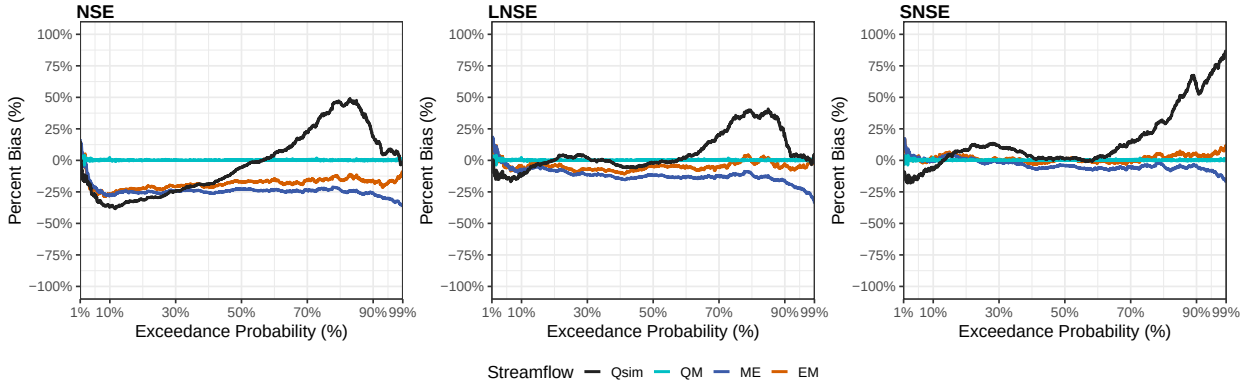

Figure 2. Bias duration curves at Schoharie Creek at Prattsville, NY, for the original simulation (Qsim), quantile mapping (QM), mean across all ensembles (ME), and ensemble mean (EM) for models calibrated by NSE, LNSE, and SNSE.

calibrated with NSE, though similar conditional bias occurred when the model was calibrated with LNSE or SNSE. Applying FDC-based bias correction (QM) eliminates nearly all bias in the BDC regardless of the calibration metric, while the two ensemble post-processing methods assessed (ME and EM) reduce but do not completely remove conditional bias. Evaluating the BDC alone is not sufficient to assess model skill at low flow conditions, as low flow behavior is also sensitive to variability in the lower tail of the distribution.

3.2 Estimation of the lowest quantile of streamflow

Figure 3 summarizes the average bias and root mean square error (RMSE) of the streamflow smaller than the 25th percentile for the original simulation, quantile mapping, and ensemble generation across three calibration metrics (NSE, LNSE, SNSE) for three watersheds.

Importantly, and what is immediately apparent in Figure 3, is that the ensemble mean streamflow (EM) generally achieves the lowest bias and RMSE across nearly all sites and calibration metrics, demonstrating its capacity to reduce unexplained variability and improve predictive accuracy during low flow periods. EM performed better than the mean of the ensembles (ME), which may have been adversely impacted by some particularly unusual ensembles. Across all watersheds and metrics, the conditional mean daily streamflows from the original deterministic model after calibration (Qsim) exhibit substantial positive bias indicating a systematic tendency to overpredict low flow. This pattern persists regardless of whether the model is calibrated using NSE, LNSE, or SNSE, demonstrating that calibration alone is insufficient to address the issue of conditional bias inherent in the simulations, especially for low flow conditions.

Application of all post-processing methods (QM, ME and EM) generally leads to reductions in bias across all three watersheds compared to Qsim. The FDC-based bias correction on quarterly flows (QMQ) more effectively mitigates conditional bias for the smallest streamflows and consistently provides better results than QM. The bias of the lowest 25% of streamflows averaged across ensembles (ME) and from the single mean ensemble trace (EM) produces the same value for bias (which is to be

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

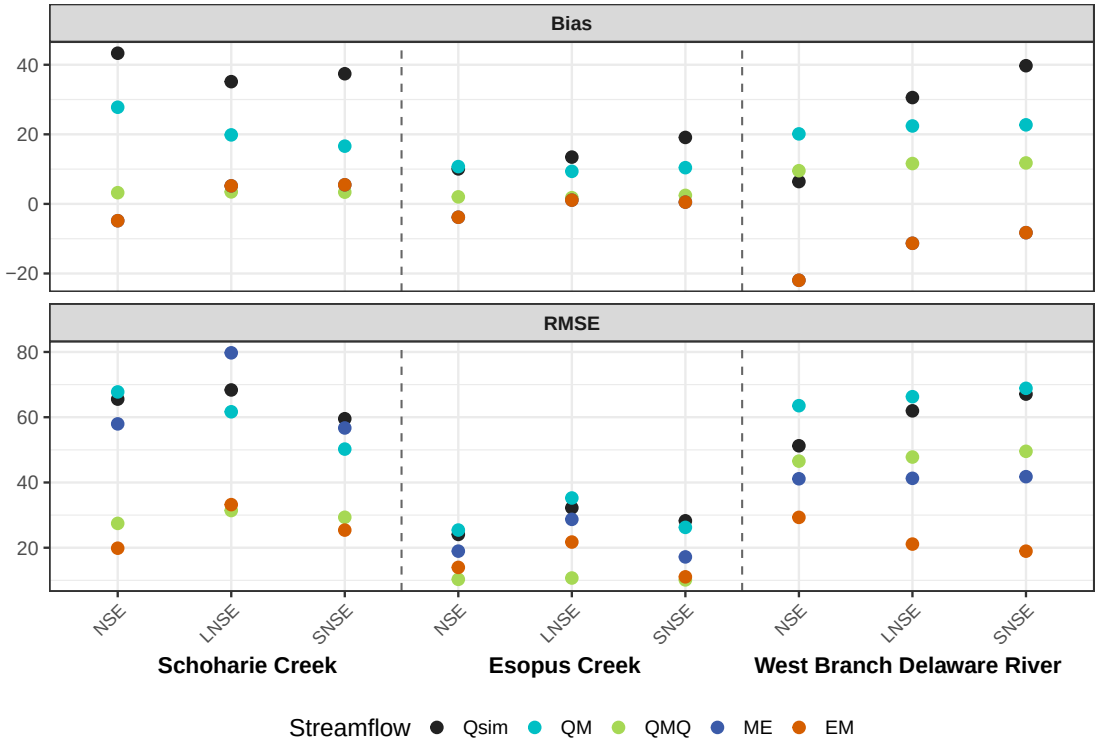

Figure 3. Bias and root mean square error (RMSE) for the original simulation (Qsim), quantile mapping (QM), quantile mapping based on quartered flow (QMQ), mean of ensembles (ME), and ensemble mean (EM) for streamflows below the 25th percentile for models calibrated with NSE, LNSE, and SNSE at the three watersheds.

expected), and in general was similar to that of QMQ, especially at Schoharie Creek and Esopus Creek. The bias of the lowest 25% of the flows for the QMQ, ME and EM results was closest to zero for the model calibrated using SNSE.

Calibrated streamflows without post-processing (Qsim) consistently exhibit a large RMSE when evaluated over the lowest quartile of flows, indicating substantial variability even after site-specific calibration. Bias-corrected simulations (QM) produce limited reductions in RMSE, while QMQ produce noticeable reductions in RMSE, particularly for LNSE- and SNSE-calibrated models, which is consistent with the low flow emphasis of these objective functions. Across all three study sites, for the lowest 25% of streamflow, EM generally has the lowest RMSE, especially when the model is calibrated with SNSE.

3.3 Estimation of the low flow design statistics

Methods were also evaluated using two classes of low flow design statistics: AM_d, the annual minimum d-day average flow commonly used in hydrologic design, where d = 3, 7, 14 and 30 days, and Qp, the annual FDC quantile with an exceedance-probability p where p = 90%, 95%, and 99%. Each statistic was calculated separately at each watershed for each water year and then results were averaged across the three study sites. Figure 4 summarizes the bias and RMSE of annual Qp and AM_d

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

estimators during the calibration period, averaged across all three study sites. Across all calibration metrics (LNSE, NSE, SNSE), Qsim exhibited substantial positive bias and high RMSE values for both AM_d and Qp. Application of the QMQ bias correction, which is possible only during the calibration period because it requires observed flows to derive the bias correction, effectively reduces bias to near zero for both classes of statistics. In contrast, RMSE of QMQ increases compared to Qsim, suggesting that although QMQ is effective at correcting average tendencies, it is not successful in reducing the overall spread and variability of low flow design statistic estimators. Because our previous results indicate that QM performs similarly to or slightly worse than QMQ, only QMQ results are presented in these and subsequent analyses. Because the QMQ method is designed to match the flow duration curve, it performs especially well for statistics that mainly reflect the central tendency of the corrected portion of the distribution, as illustrated by the daily streamflow comparison in Figure 3. However, low-flow design statistics depend strongly on the timing and magnitude of the most extreme low-flow events, which QMQ does not explicitly preserve. As a result, bias can still be reduced while RMSE remains high because event-specific errors and temporal mismatches persist. The most consistent improvements in both bias and RMSE are achieved using the ensemble streamflow methods (ME and EM), which produce bias closer to zero than Qsim while also resulting in the smallest RMSE for both AM_d and Qp. Again, EM was slightly better (bias closer to zero and smaller RMSE) than ME. Models calibrated using SNSE led to improved results compared to models calibrated using NSE and LNSE, those these results were similar and EM and ME post-processing improved the estimators of low flow design statistics for all calibrated models.

Figure 5 illustrates the same comparisons during the validation period as those shown for the calibration period in Figure 4. During the validation period, bias correction (QM and QMQ) cannot be applied because observations are not available. The comparison in Figure 5 therefore focuses on two forms of ensemble streamflow: ME, generated as if observations were known, consistent with the calibration approach, and MEF generated under the assumption that observations are unavailable and the AR model and residuals from the calibration period are employed to forecast the 6-year validation streamflow sequence. In addition to the average statistics across all the ensemble traces, the statistics obtained were also derived from a single averaged mean trace (EM), and EM with the AR model and residuals from the calibration (EMF). Similar to the calibration period results, both ensemble methods (EM and ME) have smaller bias and RMSE than the original simulation (Qsim) when estimating AM_d and Qp, with EM producing the lowest bias and RMSE. Similarly, EMF and MEF also show some improvements compared to the original simulation, indicating the potential to use these techniques for improved model forecasting, but there was some drop in performance compared to EM and ME. While EM performed better than ME during the validation period, EMF and MEF performed similarly here. Among all the calibration metrics, results using a model calibrated with SNSE consistently led to the best performance with the bias near zero and RMSE very small during both calibration and validation periods.

4 Conclusions

This experiment was designed to assess the skill of a fully distributed deterministic hydrologic model, WRF-Hydro, to reproduce observed low flow series and statistics. Alternative calibration metrics as well as the use of two post-processing approaches were evaluated. Both post-processing approaches, the FDC-based quantile mapping for the entire FDC, which

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

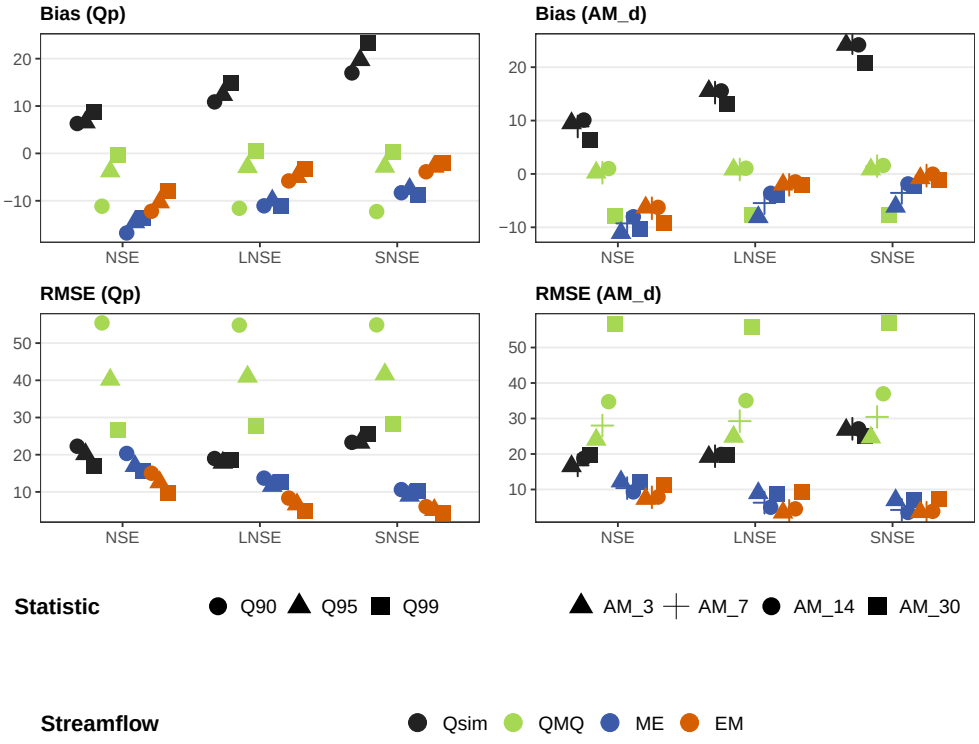

Figure 4. Average bias and RMSE for the original simulation (Qsim), quantile mapping based on quartered flow (QMQ), mean of ensembles (ME), and ensemble mean trace (EM) when estimating annual FDC quantiles Q90, Q95, and Q99 (left) and annual minimum d-day flows AM3, AM7, AM14, and AM30 (right) across all three watersheds during the calibration period.

was evaluated by Bosompemaa et al. (2025), and the log-ratio ensemble post-processing method, which was proposed by Shabestanipour et al. (2023), do not require extensive model inputs and outputs; they only require simulated and observed streamflows. The FDC-based quantile mapping was evaluated over the entire FDC (QM) and the FDC of each quartile (QMQ). This experiment was carried out at three unregulated watersheds in New York State which provide streamflow to New York City's water supply reservoirs. This experiment examined bias across the entire spectrum of streamflows, and the bias and root mean square error (RMSE) of the lowest 25th percentile of streamflow and common low flow design statistics. The experiment was performed across a calibration period, where the observations were assumed to be known, and a validation period, where the observations were assumed to be unknown. The results of these experiments were as follows:

1. Calibration alone, regardless of the calibration metrics used, cannot eliminate conditional bias in simulated streamflow across the watersheds considered. Here we explored the use of NSE, the log-space NSE (LNSE), and the square root NSE (SNSE), all of which are objective functions that assign different weight to observations depending on their magnitude. All three calibration approaches resulted in considerable conditional bias, especially for low and high extreme streamflows.

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

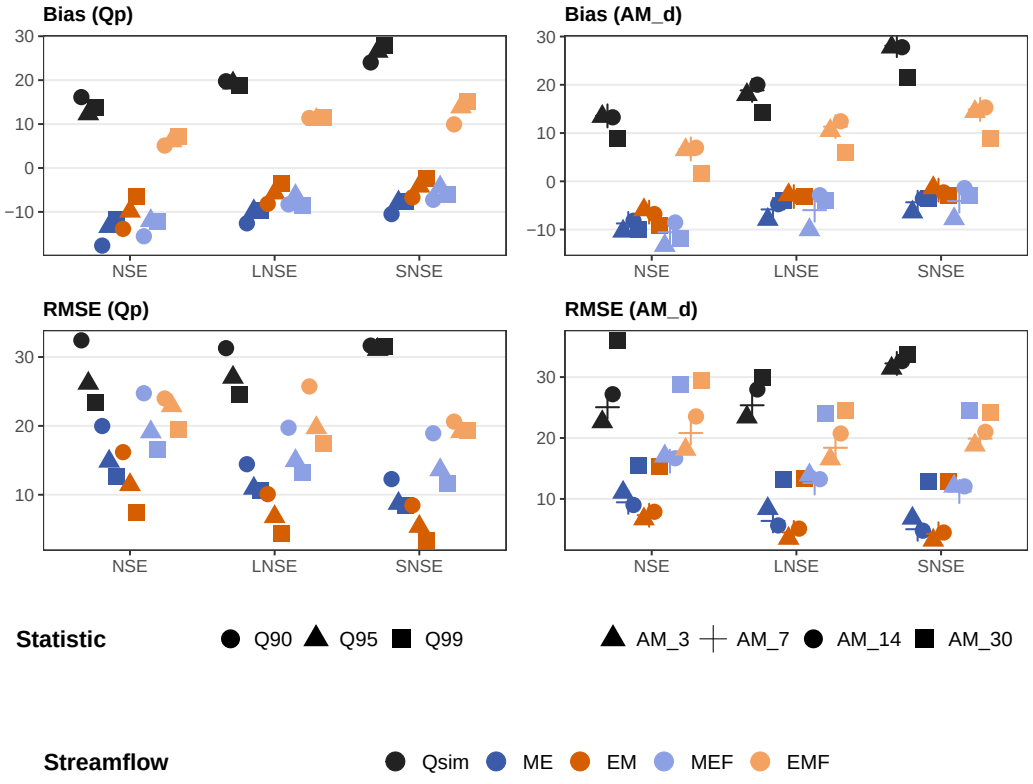

Figure 5. Bias and RMSE for original simulation (Qsim), mean of ensemble (ME), mean of ensemble forecast (MEF), ensemble mean trace (EM), ensemble mean forecast (EMF) when estimating annual FDC quantiles Q90, Q95 and Q99 (left) and annual minimum d-day flows AM3, AM7, AM14 and AM30 (right) across all three watersheds during the validation period.

2. Overall, we found that use of ensemble-based post-processing methods led to considerable reductions in both the bias and RMSE of low flow series and design statistics compared to either the deterministic streamflow simulation (Qsim) or bias correction quantile mapping methods (QM and QMQ) during both calibration (Figures 2, 3 and 4) and validation (Figure 5) periods. This important finding suggests that to improve DWM simulations, greater attention should be given to improvements in ensemble-based post-processing approaches than to the vast literature and attention given to model calibration and bias correction methods.

3. Another important finding was that among all the tested calibration metrics, the SNSE calibration objective combined with log-ratio ensemble post-processing generally provided the best performance across all watersheds and flow statistics, though post-processing techniques improved the performance of all models, regardless of the calibration metric employed. The SNSE objective reduces the influence of large streamflows which is especially important when the flow distribution is highly skewed with heavy tails. As a result, SNSE promotes a more balanced treatment of model errors across the flow range,

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

leading to transformed residuals with more uniform variance and reduced heteroscedasticity. This finding is consistent with the recommendations of Lamontagne et al. (2020) and Clark et al. (2021) who document the enormous impact of outliers on the performance of NSE. While there were no intermittent streamflows in this analysis, unlike LNSE, SNSE can more easily adapt to incorporating zero streamflows into this analysis.

4. Ensemble generation was analyzed in two ways: calculating the series or statistic of interest for individual ensembles and then averaging the statistic across all ensembles (ME) and taking the mean of all ensembles to obtain a single mean streamflow trace and calculating the statistic of interest from that single ensemble trace (EM). Both methods generally performed well for both calibration and validation, reducing the bias and RMSE of low flow series and statistics; however, EM consistently shows a smaller RMSE than ME, indicating that using the ensemble mean trace may be more effective than averaging performance statistics across individual ensemble members.

5. For the forecasting/validation experiment, we assess the decrease in skill of the ensemble method when observations are not known and model residuals are sampled from the calibration period to produce forecast ensembles. While there was a decrease in skill of the ensemble method compared to the ensemble method when observations were unknown, both ensemble forecasting methods (EMF and MEF) performed similarly and provided some benefits over the original simulation with reduced bias and RMSE.

6. Quantile mapping was performed in two ways: using the entire flow duration curve (FDC) at once (QM), and quantile mapping using each quarter of the flow duration curve (QMQ). Although QMQ and QM rely on the same FDC bias correction framework, QMQ consistently had a bias and RMSE similar or lower than that of QM across all analyses, indicating that applying the correction separately to each streamflow quarter enhances its effectiveness in producing low flow series and statistics. While FDC-based quantile mapping appears to mitigate the systematic bias of low flow series and statistics, it results in higher RMSE than ensemble-based approaches, highlighting the need for further research into the development of ensemble-based post-processing approaches and stochastic watershed models (Vogel, 2017).

While this experiment showed the potential benefits of post-processing when modeling low flow series and statistics, further analyses are warranted. Only three relatively homogeneous watersheds were examined. Low flow processes can vary widely in different regions, and the proposed methods should be analyzed at more watersheds with varying groundwater and streamflow processes. In addition, the tradeoffs between using the proposed bias correction methods based on quantile mapping as well as the log-ratio ensemble generation approaches should be compared to post-processing methods based on machine learning such as LSTMs. Overall, this experiment provides further evidence that post-processing can improve the estimation of low flow series and statistics compared to the original deterministic model simulation, which represents only the conditional mean streamflow on a given day. In addition, model calibration alone cannot remove conditional bias or reduce RMSE of low flow series and statistics, but model post-processing can greatly improve these estimators, regardless of the calibration metric employed.

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

Code and data availability. The meteorological forcing data used in this study are from the North American Land Data Assimilation System Phase 2 (NLDAS-2) forcing dataset. Streamflow observations used for model evaluation were obtained from the U.S. Geological Survey National Water Information System (USGS NWIS). Hydrologic simulations were conducted using WRF-Hydro version 5.2.0, whose source code is available from the official NCAR WRF-Hydro public repository at https://github.com/NCAR/wrf_hydro_nwm_public/releases/tag/v5.2.0. Model calibration was performed using PyWrfHydroCalib, whose source code is available from the official NCAR PyWrfHydroCalib repository at https://github.com/NCAR/PyWrfHydroCalib. The flow-duration-curve-based bias-correction procedure follows the method described by Bosompemaa et al. (2025), and the stochastic ensemble-generation procedure follows Shabestanipour et al. (2023). The scripts used for calibration, bias correction, ensemble generation, statistical analysis, and figure production, together with the processed data required to reproduce the results, are archived at [will be made available after acceptance].

Author contributions. Charles Kroll and Tong Wan designed the study, Tong Wan conducted the hydrologic simulations, performed model calibration and post-processing analyses, developed the bias-correction and ensemble-generation workflows, produced the figures, interpreted the results, and wrote the original manuscript draft. Charles Kroll and Richard Vogel contributed to the conceptual development of the study, provided guidance on the methodology and interpretation of results, and reviewed and edited the manuscript. All authors discussed the results, contributed to manuscript revision, and approved the final version of the manuscript.

Competing interests. The authors declare no competing interests.

Acknowledgements. The authors would like to acknowledge the support of the National Aeronautics and Space Administration under Grant 80NSSC21K1731 issued through the Science Mission Directorate. Additional support was also obtained from the Department of Environmental Resources Engineering and Division of Environmental Science at the State University of New York College of Environmental Science and Forestry. We would like to acknowledge high-performance computing support from Cheyenne (https://doi.org/10.5065/D6RX99HX) and Derecho (doi:10.5065/qx9a-pg09) provided by the NSF National Center for Atmospheric Research (NCAR), sponsored by the National Science Foundation, and the OrangeGrid high-throughput computing cluster supported by Syracuse University Information Technology and Services (NSF award ACI-1341006). We also thank Priyanka Rajashekar for their comments and suggestions that helped to improve the manuscript. The authors also acknolwedge the use of ChatGPT, developed by OpenAI, to assist with language editing, grammar checking, and improving the clarity and readability of selected manuscript text. The tool was not used to generate original scientific results, data analysis, figures, interpretations, or conclusions. All AI-assisted text was reviewed, edited, and verified by the authors, who take full responsibility for the content of the manuscript.

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

References

Bosompemaa, P., Brookfield, A., Zipper, S., and Hill, M. C.: Using national hydrologic models to obtain regional climate change impacts on streamflow basins with unrepresented processes, Environmental Modelling & Software, 183, 106 234, https://doi.org/10.1016/j.envsoft.2024.106234, 2025.

Bruce, J. P.: Natural Disaster Reduction and Global Change, Bulletin of the American Meteorological Society, 75, 1831–1835, https://doi.org/10.1175/1520-0477(1994)075<1831:NDRAGC>2.0.CO;2, 1994.

Brutsaert, W. and Nieber, J. L.: Regionalized drought flow hydrographs from a mature glaciated plateau, Water Resources Research, 13, 637–643, https://doi.org/10.1029/WR013i003p00637, 1977.

Chang, L., Dwivedi, R., Knowles, J. F., Fang, Y., Niu, G., Pelletier, J. D., Rasmussen, C., Durcik, M., Barron-Gafford, G. A., and Meixner, T.: Why Do Large-Scale Land Surface Models Produce a Low Ratio of Transpiration to Evapotranspiration?, Journal of Geophysical Research: Atmospheres, 123, 9109–9130, https://doi.org/10.1029/2018JD029159, 2018.

Cho, K. and Kim, Y.: Improving streamflow prediction in the WRF-Hydro model with LSTM networks, Journal of Hydrology, 605, 127 297, https://doi.org/10.1016/j.jhydrol.2021.127297, 2022.

Clark, M. P., Bierkens, M. F. P., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., Pauwels, V. R. N., Cai, X., Wood, A. W., and Peters-Lidard, C. D.: The evolution of process-based hydrologic models: historical challenges and the collective quest for physical realism, Hydrology and Earth System Sciences, 21, 3427–3440, https://doi.org/10.5194/hess-21-3427-2017, 2017.

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., Gharari, S., Freer, J. E., Whitfield, P. H., Shook, K. R., and Papalexiou, S. M.: The Abuse of Popular Performance Metrics in Hydrologic Modeling, Water Resources Research, 57, e2020WR029 001, https://doi.org/10.1029/2020WR029001, 2021.

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

Dugger, A. L., Gochis, D. J., Yu, W., Barlage, M., Yang, Y., McCreight, J., Karsten, L., Rafieeinasab, A., and Sampson, K.: Learning from the National Water Model: Regional Improvements in Streamflow Prediction through Experimental Parameter and Physics Updates to the WRF-Hydro Community Model, in: 31st Conference on Hydrology, American Meteorological Society, Seattle, WA, https://ams.confex.com/ams/97Annual/webprogram/Paper314352.html, paper 6A.3, 2017.

Dunne, T. and Black, R. D.: Partial Area Contributions to Storm Runoff in a Small New England Watershed, Water Resources Research, 6, 1296–1311, https://doi.org/10.1029/WR006i005p01296, 1970.

Farmer, W. H. and Vogel, R. M.: On the deterministic and stochastic use of hydrologic models, Water Resources Research, 52, 5619–5633, https://doi.org/10.1002/2016WR019129, 2016.

Gochis, D., Barlage, M., Dugger, A., Fitzgerald, K., Karsten, L., McAllister, M., McCreight, J., Mills, J., Rafieeinasab, A., Read, L., Sampson, K., Yates, D., and Yu, W.: The NCAR WRF-Hydro Modeling System V5 Technical Description, NCAR Technical Note 107, National Center for Atmospheric Research, 2020.

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

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

Hashino, T., Bradley, A. A., and Schwartz, S. S.: Evaluation of bias-correction methods for ensemble streamflow volume forecasts, Hydrology and Earth System Sciences, 11, 939–950, https://doi.org/10.5194/hess-11-939-2007, 2007.

Johnson, J. M., Fang, S., Sankarasubramanian, A., Rad, A. M., Kindl Da Cunha, L., Jennings, K. S., Clarke, K. C., Mazrooei, A., and Yeghiazarian, L.: Comprehensive Analysis of the NOAA National Water Model: A Call for Heterogeneous Formulations and Diagnostic Model Selection, Journal of Geophysical Research: Atmospheres, 128, e2023JD038 534, https://doi.org/10.1029/2023JD038534, 2023.

Krzysztofowicz, R.: Corrigendum to "Integrator of uncertainties for probabilistic river stage forecasting: Precipitation-dependent model" [J. Hydrol. 249 (1–4) (2001) 69–85], Journal of Hydrology, 515, 345, https://doi.org/10.1016/j.jhydrol.2014.04.053, 2014.

Lamontagne, J. R., Barber, C. A., and Vogel, R. M.: Improved Estimators of Model Performance Efficiency for Skewed Hydrologic Data, Water Resources Research, 56, e2020WR027 101, https://doi.org/10.1029/2020WR027101, 2020.

Li, H., Sheffield, J., and Wood, E. F.: Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching, Journal of Geophysical Research: Atmospheres, 115, 2009JD012 882, https://doi.org/10.1029/2009JD012882, 2010.

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., and Di, Z.: A review on statistical postprocessing methods for hydrometeorological ensemble forecasting, WIREs Water, 4, e1246, https://doi.org/10.1002/wat2.1246, 2017.

Liu, S., Wang, J., Wang, H., and Wu, Y.: Post-processing of hydrological model simulations using the convolutional neural network and support vector regression, Hydrology Research, 53, 605–621, https://doi.org/10.2166/nh.2022.004, 2022.

Maidment, D. R.: Conceptual Framework for the National Flood Interoperability Experiment, JAWRA Journal of the American Water Resources Association, 53, 245–257, https://doi.org/10.1111/1752-1688.12474, 2017.

Maraun, D.: Bias Correction, Quantile Mapping, and Downscaling: Revisiting the Inflation Issue, Journal of Climate, 26, 2137–2143, https://doi.org/10.1175/JCLI-D-12-00821.1, 2013.

Mohan, S. and Hunt, A. G.: Linear Reservoir Behaviour Across Large Spatial Scales, Hydrological Processes, 40, e70 441, https://doi.org/10.1002/hyp.70441, 2026.

National Aeronautics and Space Administration: NLDAS-2 Forcing Dataset Information, https://ldas.gsfc.nasa.gov/nldas/v2/forcing, accessed: 2026-05-06, 2023.

New York City Department of Environmental Protection: Water Supply, https://www.nyc.gov/site/dep/water/water-supply.page, accessed: 2025-12-17, 2025.

Nicolle, P., Pushpalatha, R., Perrin, C., François, D., Thiéry, D., Mathevet, T., Le Lay, M., Besson, F., Soubeyroux, J.-M., Viel, C., Regimbeau, F., Andréassian, V., Maugis, P., Augeard, B., and Morice, E.: Benchmarking hydrological models for low-flow simulation and forecasting on French catchments, Hydrology and Earth System Sciences, 18, 2829–2857, https://doi.org/10.5194/hess-18-2829-2014, 2014.

Obasi, G. O. P.: WMO's Role in the International Decade for Natural Disaster Reduction, Bulletin of the American Meteorological Society, 75, 1655–1661, https://doi.org/10.1175/1520-0477(1994)075<1655:WRITID>2.0.CO;2, 1994.

Oudin, L., Andréassian, V., Mathevet, T., Perrin, C., and Michel, C.: Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations, Water Resources Research, 42, 2005WR004 636, https://doi.org/10.1029/2005WR004636, 2006.

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

Puma, M. J., Celia, M. A., Rodriguez-Iturbe, I., Nordbotten, J. M., Guswa, A. J., and Kavetski, D.: Effects of Spatial Heterogeneity in Rainfall and Vegetation Type on Soil Moisture and Evapotranspiration, https://doi.org/10.48550/ARXIV.1606.05256, version Number: 1, 2016.

Pushpalatha, R., Perrin, C., Moine, N. L., and Andréassian, V.: A review of efficiency criteria suitable for evaluating low-flow simulations, Journal of Hydrology, 420-421, 171–182, https://doi.org/10.1016/j.jhydrol.2011.11.055, 2012.

Schaake, J. C., Hamill, T. M., Buizza, R., and Clark, M.: HEPEX: The Hydrological Ensemble Prediction Experiment, Bulletin of the American Meteorological Society, 88, 1541–1548, https://doi.org/10.1175/BAMS-88-10-1541, 2007.

Shabestanipour, G., Brodeur, Z., Farmer, W. H., Steinschneider, S., Vogel, R. M., and Lamontagne, J. R.: Stochastic Watershed Model Ensembles for Long-Range Planning: Verification and Validation, Water Resources Research, 59, e2022WR032 201, https://doi.org/10.1029/2022WR032201, 2023.

Singh, V. P., ed.: Computer models of watershed hydrology, Water Resources Publications, Highlands Ranch, Colorado, revised edition edn., ISBN 978-1-887201-74-2, 2012.

Tallaksen, L. M. and Lanen, H. A. J.: Hydrological drought: processes and estimation methods for streamflow and groundwater, Elsevier, Amsterdam, 2nd ed edn., ISBN 978-0-12-819082-1, 2023.

Tolson, B. A. and Shoemaker, C. A.: Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, Water Resources Research, 43, 2005WR004 723, https://doi.org/10.1029/2005WR004723, 2007.

U.S. Drought Monitor: U.S. Drought Monitor: Northeast Drought Conditions, September 2025, https://droughtmonitor.unl.edu/CurrentMap/StateDroughtMonitor.aspx?Northeast=, accessed: 2026-05-07, 2025.

Vogel, R. M.: Stochastic watershed models for hydrologic risk management, Water Security, 1, 28–35, https://doi.org/10.1016/j.wasec.2017.06.001, 2017.

Vogel, R. M. and Kroll, C. N.: Regional geohydrologic-geomorphic relationships for the estimation of low-flow statistics, Water Resources Research, 28, 2451–2458, https://doi.org/10.1029/92WR01007, 1992.

Wiley, J. B.: Low-flow analysis and selected flow statistics representative of 1930-2002 for streamflow-gaging stations in or near West Virginia, Scientific Investigations Report 5002, ISSN 2328-0328, https://doi.org/https://doi.org/10.3133/sir20065002, series: Scientific Investigations Report, 2006.

Wolock, D. M.: Hydrologic landscape regions of the United States, Open-File Report 145, U.S. Geological Survey, ISSN 2331-1258, https://doi.org/https://doi.org/10.3133/ofr03145, series: Open-File Report, 2003.

Wood, A. W., Leung, L. R., Sridhar, V., and Lettenmaier, D. P.: Hydrologic Implications of Dynamical and Statistical Approaches to Downscaling Climate Model Outputs, Climatic Change, 62, 189–216, https://doi.org/10.1023/B:CLIM.0000013685.99609.9e, 2004.

World Meteorological Organization: Manual on Low-flow Estimation and Prediction, Tech. Rep. Operational Hydrology Report No. 50, WMO-No. 1029, World Meteorological Organization, Geneva, Switzerland, 2008.

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Duan, Q., and Lohmann, D.: Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow, Journal of Geophysical Research: Atmospheres, 117, 2011JD016 051, https://doi.org/10.1029/2011JD016051, 2012.

Xiang, Z., Yan, J., and Demir, I.: A Rainfall-Runoff Model With LSTM-Based Sequence-to-Sequence Learning, Water Resources Research, 56, e2019WR025 326, https://doi.org/10.1029/2019WR025326, 2020.

https://doi.org/10.5194/egusphere-2026-3023
Preprint. Discussion started: 12 June 2026

565 Xiao, M., Pan, M., Yang, Y., Cao, Q., Dixon, T., Lewis, G., Su, L., Hartman, R., DeFlorio, M. J., Kalansky, J. F.,
566 Monache, L. D., and Ralph, F. M.: LSTM-based Post-Processing Improves Streamflow Prediction in the Sierra Nevada,
567 https://doi.org/10.22541/essoar.174534330.09680378/v1, 2025.