AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2022WR032201

Key Points:

- Deterministic watershed models mischaracterize the extremes that are most relevant to hydrologic risk management
- Stochastic watershed model (SWM) developed by post-processing deterministic model output can capture hydrologic extremes
- Verification and validation are critical to ensure that SWMs are properly parameterized and fit for purpose

Supporting Information:

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

Correspondence to:

J. R. Lamontagne,
jonathan.lamontagne@tufts.edu

Citation:

Shabestanipour, G., Brodeur, Z., Farmer, W. H., Steinschneider, S., Vogel, R. M., & Lamontagne, J. R. (2023). Stochastic watershed model ensembles for long-range planning: Verification and validation. *Water Resources Research*, *59*, e2022WR032201. https://doi.org/10.1029/2022WR032201

Received 14 FEB 2022
Accepted 18 JAN 2023

Author Contributions:

Conceptualization: Ghazal Shabestanipour, Zachary Brodeur, William H. Farmer, Scott Steinschneider, Richard M. Vogel, Jonathan R. Lamontagne
Data curation: Ghazal Shabestanipour
Formal analysis: Ghazal Shabestanipour
Funding acquisition: Richard M. Vogel, Jonathan R. Lamontagne
Methodology: Ghazal Shabestanipour, Zachary Brodeur, William H. Farmer, Scott Steinschneider, Richard M. Vogel, Jonathan R. Lamontagne
Software: Ghazal Shabestanipour
Supervision: Jonathan R. Lamontagne
Validation: Ghazal Shabestanipour, Zachary Brodeur
Visualization: Ghazal Shabestanipour

Stochastic Watershed Model Ensembles for Long-Range Planning: Verification and Validation

Ghazal Shabestanipour[1], Zachary Brodeur[2], William H. Farmer[3], Scott Steinschneider[2], Richard M. Vogel[1], and Jonathan R. Lamontagne[1]

[1]Department of Civil & Environmental Engineering, Tufts University, Medford, MA, USA, [2]Department of Biological & Environmental Engineering, Cornell University, Ithaca, NY, USA, [3]Water Resources Mission Area, U.S. Geological Survey, Denver, CO, USA

Abstract Deterministic watershed models (DWMs) are used in nearly all hydrologic planning, design, and management activities, yet they cannot generate streamflow ensembles needed for hydrologic risk management (HRM). The stochastic component of DWMs is often ignored in practice, leading to a systematic bias in extreme events. Since traditional stochastic streamflow models used in HRM struggle to account for anthropogenic change, there is a need to convert DWMs into stochastic watershed models (SWMs) to generate ensembles for use in HRM. A DWM can be converted to an SWM using a post-processing (pp) approach to add error to the DWM predictions. Many pp methods advanced in the area of flood forecasting are useful in HRM and for correcting extreme event biases. Selecting a suitable error model for pp is challenging due to nonnormality, skewness, heteroscedasticity, and autocorrelation. We develop a parsimonious pp method based on an autoregressive (AR) model of the logarithm of the ratio of the observations and simulations, which leads to AR model residuals that are approximately symmetric and independent. We document the value of pp for improving flood and low flow frequency analysis and we reintroduce the concepts of verification and validation of stochastic streamflow ensembles to ensure that the SWM can reproduce both statistics it was and was not designed to reproduce, respectively. These concepts are illustrated on a Massachusetts basin using the USGS Precipitation Runoff Modeling System, with an additional analysis indicating the approach may be applicable to 1,225 other sites across the United States.

Plain Language Summary Deterministic watershed models (DWMs) are used in hydrologic design and water management, yet DWMs systematically misrepresent extremes and do not generate streamflow ensembles needed for hydrologic risk management (HRM). Since traditional stochastic streamflow models struggle to account for climate and land use change, there is a need for approaches to convert DWMs to stochastic watershed models (SWMs) that can provide both unbiased estimates of extremes and streamflow ensembles reflecting both historical and potential future hydrologic conditions. We make two contributions: First, we reintroduce the concepts of verification and validation of stochastic streamflow ensembles, which are essential to ensure an SWM is performing correctly and that it is fit for its intended purpose, respectively. We describe how these concepts are related to, but distinct from the more common verification and validation of hydrologic models. Our second contribution is a parsimonious post-processing approach to convert a DWM to an SWM by adding error to the DWM's predictions. We demonstrate that key modeling assumptions are met (verification) and that streamflow ensembles reproduce important decision-relevant metrics related to hydrologic extremes (validation) for a test basin. We show that our approach is likely to be applicable in more than 1,200 basins across the United States.

1. Introduction

The introduction of stochastic streamflow models by Fiering (1967), Maass et al. (1962), and others led to a revolution in water resources planning, design, and management. These models enabled hydrologists to generate representative streamflow ensembles over future planning horizons, needed to explore the consequences of future hydrologic conditions not experienced historically, and formally characterize the reliability, vulnerability, and resilience of water resource systems (Hashimoto et al., 1982; Loucks & van Beek, 2017). Traditional stochastic streamflow models are typically statistical models rather than mechanistically driven hydrologic models. Such stochastic streamflow models may be adjusted to reflect changes in seasonality or other statistical properties of flow (e.g., Quinn et al., 2018), but tying statistical hydrologic changes to climate and land use change is

Writing – original draft: Ghazal Shabestanipour, Zachary Brodeur, Richard M. Vogel, Jonathan R. Lamontagne
Writing – review & editing: Ghazal Shabestanipour, Zachary Brodeur, William H. Farmer, Scott Steinschneider, Richard M. Vogel, Jonathan R. Lamontagne

not trivial without some mechanistic modeling of the hydrologic system. This increasingly renders stochastic streamflow models as inadequate for long-range planning applications. Instead, stochastic watershed models (SWMs)—defined as stochastic versions of deterministic watershed models (DWMs)—provide a viable alternative for streamflow ensemble generation because they can account for the complex coupling between climate, human, and watershed systems (Montanari & Koutsoyiannis, 2012; Vogel, 2017). This study focuses on advancing the use of SWMs for the purpose of generating stochastic streamflow ensembles, particularly in the context of long-range water resource planning and design. We contribute both methodological innovations for SWM development and strategies to verify the SWM assumptions and validate that the resulting streamflow ensembles are fit for purpose.

1.1. On the Need for Streamflow Ensembles

The purpose of generating streamflow ensembles is to represent the uncertainty associated with the dynamic watershed system by generating multiple sets of streamflow predictions over future planning horizons. We follow Koutsoyiannis and Montanari (2022) and use the term "prediction" to encompass simulation, prediction, and forecasting activities associated with DWMs and SWMs. A calibrated DWM produces a single trace of both streamflow output and model residual error. When reasonably constructed, these streamflow predictions are mean values, conditioned upon climatic, parameter, and other inputs to the DWM. Such conditional mean streamflow values will generally exhibit lower variance (and other upper moments) than the observed streamflows upon which the DWM is calibrated, leading to systematic bias, particularly for extreme events (Farmer & Vogel, 2016). The systematic addition of model residuals to simulated streamflow output, using post-processing (pp) methods produces stochastic output in the form of streamflow ensembles, which can better reproduce the upper moments and many other statistics of observed flows, thus addressing one important source of systematic bias. Such pp methods enable the conversion of a DWM to an SWM, resulting in streamflow ensembles useful in a wide range of hydrologic risk management (HRM) activities.

The development and use of SWM streamflow ensembles have grown over the years, mostly with a focus on flood and hydrometeorological forecasting over relatively short (hourly monthly) time horizons (Cloke & Pappenberger, 2009; Li et al., 2017; Troin et al., 2021; Vannitsem et al., 2019, 2021; Zha et al., 2020). Use of streamflow ensembles for flood forecasting is attractive because, in addition to the benefits of uncertainty quantification, probabilistic hydrological ensemble predictions are often more skillful than deterministic predictions (Cloke & Pappenberger, 2009; Roulin, 2007).

However, in contrast with flood forecasting, less attention has been given to pp methods and the value of SWM streamflow ensembles for use in long-range planning activities, which is a central focus of this study. Streamflow ensembles over long planning horizons enable integration of uncertainty into water resource decision-making. They have been in use by numerous US federal agencies for decades (see review in Vogel (2017)), although often developed using stochastic streamflow models (Fiering, 1967) or ensembles of climate traces and DWMs. Such ensembles are the basis for modern Risk-Based Decision Making (RBDM), which enables determination of an appropriate level of investment based on the expected benefits and damages avoided versus the cost of the infrastructure required under integrated climate and hydrologic uncertainty (Brekke, 2009; Stakhiv, 2011).

The generation of stochastic streamflow ensembles representing alternative climate realizations that are an increasingly important component of robustness and adaptation frameworks within water resources planning (Herman et al., 2015, 2020; Steinschneider et al., 2012, 2015; Steinschneider & Lall, 2015). The complicated relationship between climate change, land use change, and hydrologic extremes presents a significant challenge to stochastic streamflow models. For instance, Sharma et al. (2018) show that the impact of more extreme precipitation on flooding depends on many factors including antecedent hydrologic conditions, the size and geometry of basins, and characteristics of storms. One cannot assume that more intense precipitation leads to more intense flooding. SWMs are a promising tool to generate nonstationary streamflow ensembles because they are based on DWMs which capture (though imperfectly) the factors that interact with the changing climate to produce changes in streamflow. Still, the formal use of SWMs to generate these long-range streamflow ensembles remains in its infancy.

1.2. Post-Processing Approaches for Generating Streamflow Ensembles—A Brief Review

Nearly all pp methods are implemented by performing a stochastic analysis of DWM errors (our focus here), notwithstanding some interesting exceptions (Koutsoyiannis & Montanari, 2022). Two general approaches exist for characterizing model error: (a) aggregated approaches that lump all sources of uncertainty into a single model error term (Montanari & Koutsoyiannis, 2012; Schoups & Vrugt, 2010; Tajiki et al., 2020); and (b) decomposition approaches that model each error source separately (Kuczera et al., 2006; Renard et al., 2011). We follow the aggregated approach, which assumes that all sources of uncertainty, whether they arise from input data measurement errors, model parameter errors, and/or model structural errors, are contained within the model calibration/validation residuals. Such aggregated approaches have been shown to yield more reliable prediction intervals than decomposition approaches (e.g., Valdez et al., 2022).

Aggregated approaches to pp and uncertainty analysis may be further divided into those based on likelihood functions and likelihood-free methods. Most methods based on likelihood functions are developed within a Bayesian framework (see Kuczera et al., 2017), where identification of a suitable likelihood function often introduces assumptions that can be difficult to assess, making the overall approach less transparent to end-users.

Aggregation of all sources of uncertainty into model error leads to some very attractive and simple pp approaches, resulting in a relatively transparent and straightforward generation of streamflow ensembles (Evin et al., 2013; Hunter et al., 2021; Koutsoyiannis & Montanari, 2022; Meyer et al., 2020; Sikorska et al., 2015; Zha et al., 2020). The fundamental challenge becomes selection and estimation of a suitable probabilistic model that can characterize the non-normality, heteroscedasticity, and very high stochastic persistence associated with DWM errors (see Hunter et al. (2021) and McInerney et al. (2017)), a choice which can have a tremendous impact on both DWM model parameter estimation and SWM prediction intervals (Evin et al., 2013).

Due to the widespread use of pp methods in flood forecasting, studies have explored a wide range of methods for handling autocorrelation and heteroscedasticity of model residuals (see Table 1 in Zha et al. (2020)). In this work, we contribute two important methodological insights to this literature. First, we employ an autoregressive (AR) model of a logarithmic transformation of the model residuals to account for heteroscedasticity, serial correlation, skewness, and heavy tails, and couple this transformation with a k-nearest neighbor (k-NN) bootstrap resampling approach to generate streamflow ensembles. While others have employed AR models of logarithmically transformed residuals to develop SWMs (see reviews in McInerney et al. (2017) and Hunter et al. (2021)), we propagate parametric uncertainty in the AR model into the stochastic ensembles, which is often ignored. In addition, our use of k-NN resampling is unique and requires fewer assumptions to support stochastic generation. Importantly, bootstrapping is known to breakdown under heteroscedasticity, serial correlation, skewness, and heavy tails (see discussion in section 5.1 of Clark et al. (2021) and chapter 9 of Chernick (2008)), and so the coupling of k-NN resampling with AR models of log-transformed residuals (which removes these features) is an important advance.

Another methodological contribution of this study is that we derive a bias correction factor to account for retransformation bias. Use of any model error transformation approach introduces retransformation bias, when converting transformed model errors back to real space to generate streamflow ensembles. Others have dealt with retransformation bias in an indirect and/or empirical fashion, for example, Hunter et al. (2021) and others cited therein introduced empirical time-series models that relate the mean of the transformed errors to streamflow over time to ensure unbiased simulation. However, to our knowledge, we are the first to derive a retransformation bias correction based on statistical properties of the model error transformation.

Finally, others have criticized the use of a logarithmic transformation due to its inability to handle zero streamflows. Although we only consider a perennial river in this study, we provide recommendations in Section 4.2 and the Supplement for several approaches to handle zero streamflows when using a logarithmic transformation.

1.3. Verification and Validation of Streamflow Ensembles: Are They Fit for Purpose?

An enormous literature exists on methods for the verification and validation of streamflow ensembles for use in meteorological and hydrologic forecast applications, where concerns over forecast skill are paramount (e.g., Alfieri et al. (2014), Bradley et al. (2004), Laio and Tamea (2007), and Wilks (2019)). Beven (2019) argues that a model should be "fit for purpose," which is why forecast skill is so important to the hydrometeorological

forecasting community. Similarly, an enormous literature exists on methods for the verification and validation of a calibrated DWM with most traditional approaches being a variation on the split-sample technique (see discussions in Klemeš (1986) and Vogel and Sankarasubramanian (2003)). Verification and validation exercises often concentrate on an evaluation of the goodness-of-fit between the streamflow observations and the streamflow simulations obtained from a DWM calibrated to those historical observations (i.e., see Clark et al., 2021). In contrast to both the verification and validation of a DWM, and of streamflow ensembles for use in hydrometeorological forecast applications, much less attention has been given to the verification and validation of the stochastic ensembles generated by an SWM for use in long-range planning, a central focus of this study. While on the one hand, we differentiate our work in part by its focus on the verification and validation of the streamflow ensembles, rather than the DWM itself, on the other hand, we are not the first to address these issues (see e.g., Evin et al. (2013), Hunter et al. (2021), and Schoups and Vrugt (2010), among many others). In this study, we demonstrate approaches to verify our modeling assumptions are met and validation procedures to ensure the resulting ensembles are "fit for purpose," which in this case means they are useful in long-range planning.

Many studies have employed pp methods to develop uncertainty intervals for streamflow simulations, where the notion of "fit for purpose" is usually evaluated using coverage probabilities. However, Vogel (2017) argues that uncertainty intervals, while interesting and useful, are of little value in approaches like RBDM, because such RBDM approaches require the complete set of streamflow ensembles for their implementation, so that the uncertainty intervals would not be adequate to apply widely accepted RBDM approaches. In the remainder of this study, we assume that SWMs will find use in short-range, medium-range, and long-range water resource planning, operations, and management applications, where the notion of a model being "fit for purpose" would include a myriad of concerns outlined in this study, in addition to forecast skill and coverage probabilities.

One focus of this study is the verification and validation of the streamflow ensembles generated by an SWM (not the SWM itself as is common practice), particularly in the context of long-range planning, using the principles advanced for evaluating streamflow ensembles generated from stochastic streamflow models (Salas et al., 1980; Stedinger & Taylor, 1982a). One can think of this aspect of our study as an instructive example of how to apply generally established practices in the development of an SWM for generating streamflow ensembles. Before the application of an SWM to simulate the impact of hydrologic change, one must ensure that the model is credible. Such an evaluation of an SWM would follow the basic guidelines associated with the construction, verification, and validation of any stochastic simulation model, as summarized by Salas et al. (1980) and Stedinger and Taylor (1982a) and now common practice in the much larger field of simulation modeling. As paraphrased from Stedinger and Taylor (1982a): Stochastic streamflow model verification should demonstrate that a conceptual model has been implemented correctly; (stochastic streamflow) model validation is then an additional and more difficult task which compares simulation results (i.e., streamflow ensembles) with real-system data to demonstrate that the model is an adequate description of the real world for the intended investigation.

Model verification of SWM streamflow ensembles would evaluate those properties and assumptions inherent in their generation, which in our case leads to an evaluation of the stochastic behavior of a logarithmic transformation of the model innovation ratios (model simulations divided by the observation) to ensure they are approximately normal, and that an AR model fit to those log transformed innovation ratios are approximately serially independent and symmetric (no skewness). The pp method outlined in this study makes use of those three assumptions, thus their verification should ensure plausible streamflow ensembles.

Model validation should evaluate whether or not the SWM is capable of generating ensembles that are "fit for purpose." This amounts to ensuring that the SWM can generate streamflow ensembles that reproduce important hydrologic and water resource system properties that are related to the actual RBDM activities the model is intended to address. For example, if one's interest is in water supply planning, reproduction of various drought, storage, and water deficit statistics would be a priority, whereas if the focus is on flooding, reproduction of key design flood characteristics would be critical. We note that this definition of validation differs from the common practice of evaluating a calibrated model against data not used in the calibration, though it does not necessarily preclude it.

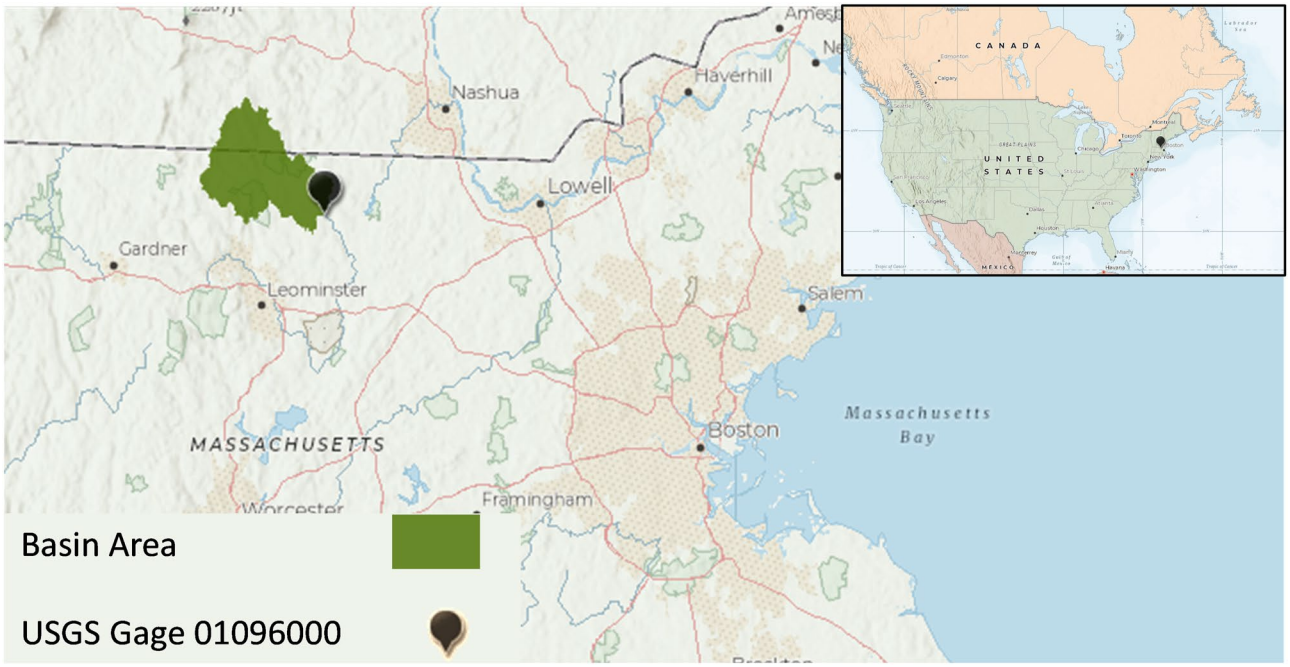

Figure 1. Squannacook watershed area and USGS streamgage location.

1.4. Study Goals

The above introduction has sought to provide perspective on the development of the central goals of this study, which are:

1. Development of a generalized, transparent, and aggregated pp approach to modeling DWM model errors to convert a DWM to an SWM, resulting in streamflow ensembles that are useful in a wide range of HRM activities, including long-range planning.
2. (Re)introduction of ensemble verification procedures to ensure that the aggregated pp approach used to generate streamflow ensembles mimics important statistical properties of the calibration model error upon which the SWM is based (e.g., nonnormality, heteroscedasticity, and autocorrelation).
3. (Re)introduction of ensemble validation approaches to ensure that streamflow ensembles are "fit for purpose," particularly in long-range planning activities, that is, ensuring that streamflow ensembles reproduce key properties needed for most RBDM applications, including flow duration curves (FDCs), storage-yield curves, and the distribution of extreme high and low flow design statistics in addition to reproduction of coverage probabilities associated with uncertainty intervals.

2. Stochastic Watershed Modeling Methodology

In this section, we introduce a pp approach to convert a DWM to an SWM, along with approaches to verify our ensemble modeling assumptions and ensemble validation procedures to ensure the ensembles are "fit for purpose," which implies they are useful in long-range planning. We begin by introducing the basin and DWM used in this study to demonstrate the proposed pp approach.

2.1. Study Basin and the DWM

We adopt the Squannacook River (USGS streamgage 01096000, Figure 1) in northeastern Massachusetts as a demonstration basin for this study. The Squannacook basin has a drainage area of 173.8 km^2 and is moderately impacted by human influences. Less than 8% of the basin surface area is impervious and it contains five dams. This basin topography ranges from a hilly upland plateau in the north and west to flat coastal plain in the south and east.

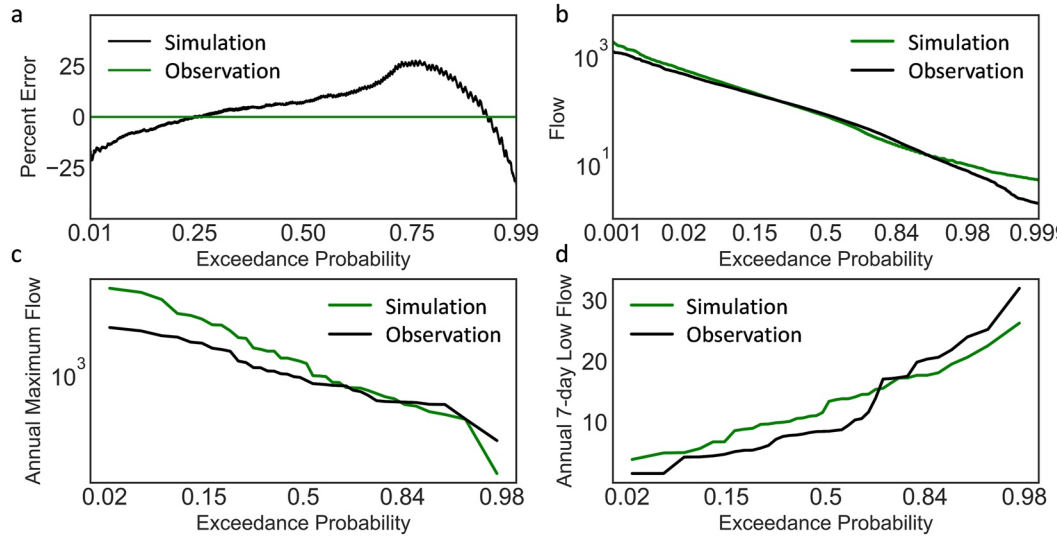

Figure 2. Deterministic watershed model (DWM) performance (USGS PRMS model) for the Squannacook River by exceedance probability, based on (a) percent error, (b) flow duration curves, (c) the distribution of annual maximum streamflow, and (d) the distribution of 7-day low flows.

Our approach to SWM construction begins with a calibrated DWM. We use the USGS National Hydrologic Model Precipitation Runoff Modeling System (NHM-PRMS; Markstrom et al., 2015; Regan et al., 2019) segment for the Squannacook River. The calibrated model was extracted directly from the NHM-PRMS framework. Regan et al. (2019) describe the NHM-PRMS as a medium-complexity continuous watershed simulation model that was calibrated for the entire continental United States. For the NHM-PRMS, calibration was accomplished through a normalized squared error on streamflow along several calibration steps (e.g., high flows, low flows, monthly flows, and daily flows) across similarly behaved basins; a full description is provided by Regan et al. (2019).

Once extracted from the NHM-PRMS, further adjustments were conducted to account for local conditions. Adjustments were mainly performed on the climate input data; all monthly evapotranspiration coefficients were increased by 20% and the groundwater coefficient was decreased by 10% to improve model fit and month-to-month consistency. Over the period of record (October 1980 to September 2017), the model produced simulations with a Nash-Sutcliffe Model Efficiency (NSE) of 0.64 (Nash & Sutcliffe, 1970) and a Kling-Gupta Efficiency (KGE) of 0.68 (Gupta et al., 2009), and the logarithms of daily streamflow produced Nash-Sutcliffe Model Efficiency (LNSE) of 0.71. There is, of course, significant uncertainty associated with NSE and KGE, and although LNSE is a biased estimator of real space efficiency, it has much lower uncertainty and is generally preferred over either NSE or KGE, as described by Clark et al. (2021) and Lamontagne et al. (2020).

Figure 2 compares the properties of the observed and DWM-simulated daily streamflows for the study basin. The DWM underestimates both high and low daily streamflows (see Figures 2a and 2b). The percent underestimation increases for streamflow extremes, exceeding 25% error for the 1% and 99% quantiles of daily flows. The DWM also underestimates annual flow statistics used in long-range planning including the annual maximum flood (Figure 2c) and the 7-day low flow (Figure 2d), which could result in under design or overdesign of infrastructure, respectively. These patterns of bias are characteristic of those documented by Farmer and Vogel (2016), who showed across 1,400 modeled basins in the contiguous United States that the underestimation of extremes is in part due to the general underestimation of variance and other upper moments when DWMs are applied without pp methods. This can be understood through analogy to linear regression, where an unbiased regression model will only reproduce the variance of the observations if $R^2 = 1$, which is unlikely in practice. Generally, as goodness-of-fit drops, the downward bias in the variance of model simulations increases, leading to corresponding downward/upward bias in the floods/drought streamflows. Farmer and Vogel (2016) document that such bias in extreme events can be ameliorated through pp which effectively adds variability to the streamflow ensembles, a central focus of this study.

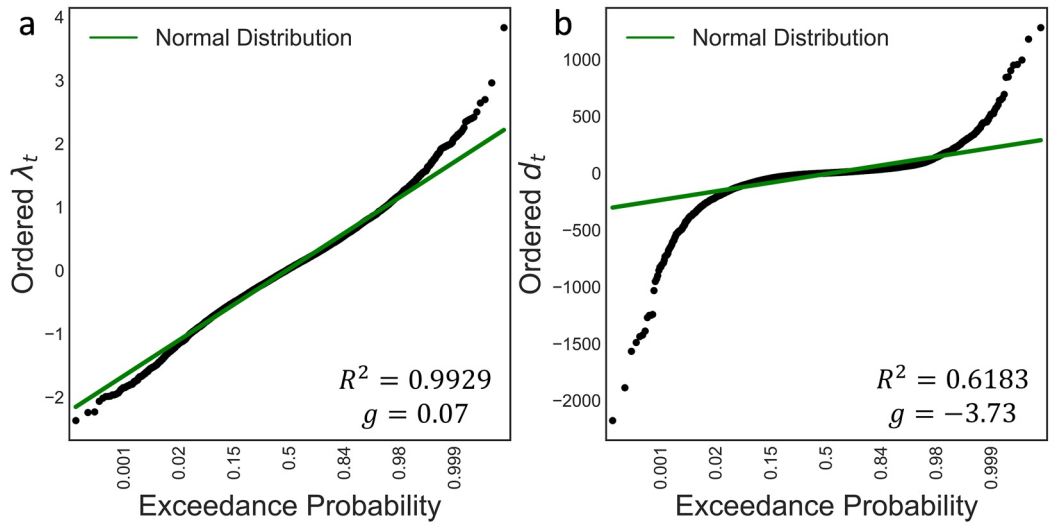

Figure 3. Normal probability plots of (a) log-ratio errors λ_t and (b) differenced residuals $d = S_t - O_t$. R^2 is square of the correlation coefficient between the theoretical quantiles and ordered residuals, and g is the sample skew coefficient of the residuals.

2.2. A Log-Ratio Post-Processing Approach to Stochastic Watershed Modeling

Our pp approach constructs an SWM by adding variability to the DWM's predictions to better reproduce the variance and other higher moments of the observed streamflows. We assume that errors from all sources (except for streamflow measurement errors) are contained in the model residuals over the historical period (as in Valdez et al. (2022)), so our pp approach aims to generate random variability that mimics important properties of the observed residuals, as described below and summarized in Figure 6. One challenge is that the differenced residuals (i.e., $d = Simulation - Observation = S - O$) are known to exhibit significant asymmetry, heavy tails, heteroscedasticity, and serial correlation which can confound standard statistical approaches to stochastic simulation. Thus, our approach begins with a transformation of the residuals to address those issues.

Though most previous attempts to characterize model error involve the differenced residual, d, recent work suggests residual transformations may reduce heteroscedasticity, serial correlation, and non-normality. McInerney et al. (2017) evaluated eight different model error formulations and concluded that although no single one was preferred in all cases, the Box-Cox transformation with transformation parameter between 0 (i.e., the log transformation) and 0.2 usually performed best. Based on the recommendations of Farmer et al. (2021), McInerney et al. (2017), Meyer et al. (2020), and Morawietz et al. (2011), as well as our own analysis in Figures 3 and 7 below, our pp approach characterizes residuals using a log ratio model:

$$\lambda_t = \ln\left(\frac{S_t}{O_t}\right) = \ln(S_t) - \ln(O_t) \tag{1}$$

where S_t and O_t are the simulated and observed streamflows in time t, respectively. Figure 3 uses normal probability plots to compare the distribution of the differenced residuals d and λ_t residuals for the Squannacook Basin. While the difference residuals are skewed with heavy tails, the λ_t are approximately normally distributed with 0 mean (which is shown to be a rather general result for the conterminous United States in Section 2.4).

Figure 4a reports the empirical autocorrelation function (ACF) of the λ_t series for the Squannacook basin. The ACF in Figure 4a is a plot of the sample estimates of the autocorrelation $r(k)$ in a time series versus the lag k in days and is denoted using solid circles. Note that the autocorrelation of λ_t dies off very slowly and only approaches 0 after about 100 lags (see the observed autocorrelation line in Figure 9).

We use an AR(p) model to represent the very slow decay associated with the ACF of the λ_t shown in Figure 4a

$$\lambda_t + \varphi_0 + \sum_{i=1}^{p} \varphi_i \lambda_{t-i} + \varepsilon_t \tag{2}$$

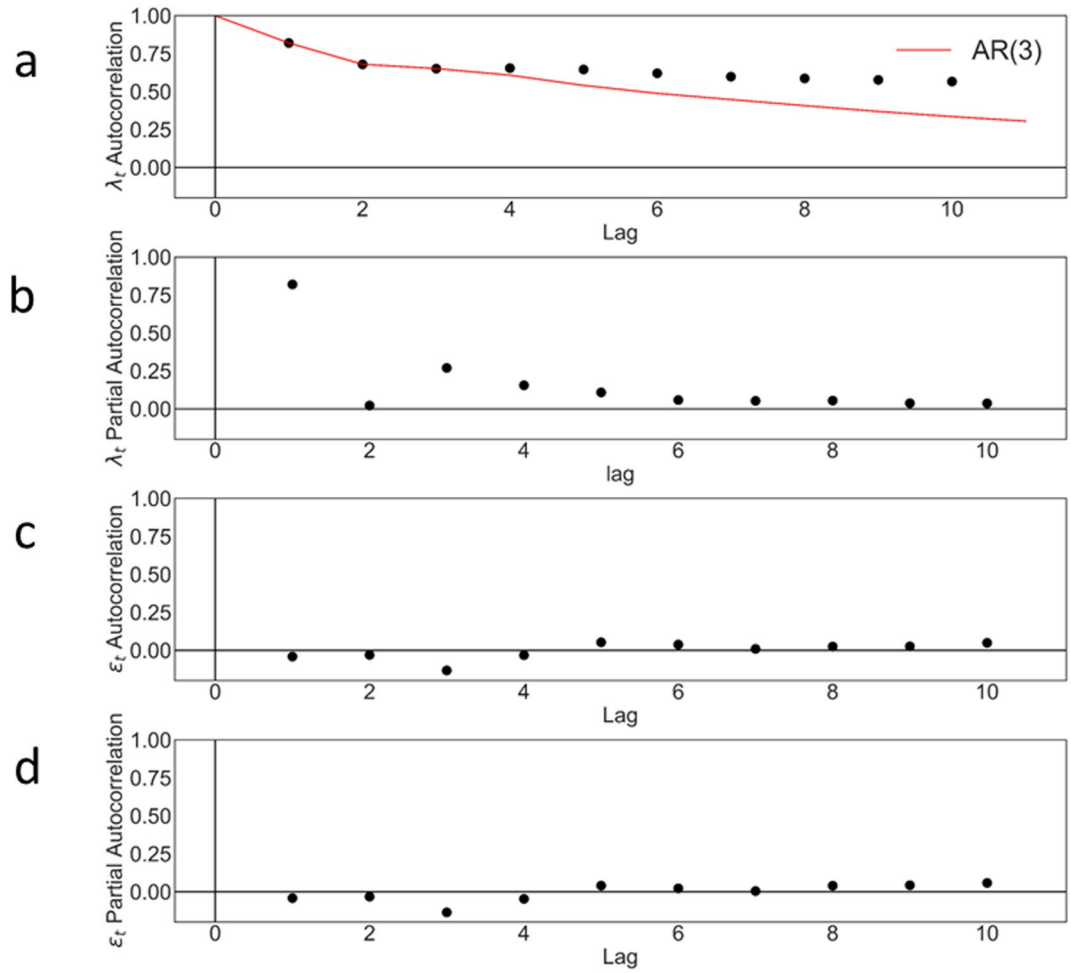

Figure 4. (a) Empirical autocorrelation function (ACF) for λ_t series, (b) empirical partial autocorrelation function (PACF) for λ_t series, (c) empirical ACF of residuals ε_t series, and (d) empirical PACF of residuals ε_t series.

where φ_i and ε_t are the coefficients and residuals of the AR model, respectively. We select an AR(3) model in this application, which minimized the Akaike Information Criteria (AIC; Teegavarapu et al., 2019).

Figure 4c reports the ACF of the ε_t time series corresponding to the fitted AR(3) model which indicates that nearly all autocorrelation has been removed. In addition, the residuals ε_t exhibit little skew (L-coefficient of skewness = 0.02). Thus we may treat ε_t as approximately symmetric and independent. Our pp approach to an SWM then involves generating random ε_t, denoted $\tilde{\varepsilon}_t$, to generate random λ_t, denoted $\tilde{\lambda}_t$, and ultimately random S_t, denoted $\tilde{S}_t$.

Since the residuals ε_t in Equation 2 exhibit neither skewness nor serial correlation, we employ bootstrap resampling to generate random sequences of $\tilde{\varepsilon}_t$ from the historical observations of ε_t. This avoids the need for distributional assumptions in modeling ε_t. In Figure 5, we observe some evidence that ε_t are not identically distributed as a function of simulated streamflow, S_t, and across months. To account for this, we tested two bootstrap approaches: a k-NN bootstrap (Lall & Sharma, 1996; Prairie et al., 2006) based on simulated flow, and a monthly k-NN bootstrap wherein the ε_t are segregated by month before k-NN bootstrapping. For brevity, we focus on the results from k-NN bootstrap approach, with the monthly k-NN results provided in the Supporting Information, and a brief discussion contrasting the two approaches in Section 3. A range of "k" values were tested to determine which bootstrap resulted in the best verification and validation results, and a value of $k = 700$ was selected. We emphasize that the level of heteroscedasticity associated with the ε_t values is significantly lower than corresponding levels of heteroscedasticity exhibited by the differenced residuals d_t, so that it would be much more difficult to implement a plausible bootstrap with the d_t values than for the ε_t values.

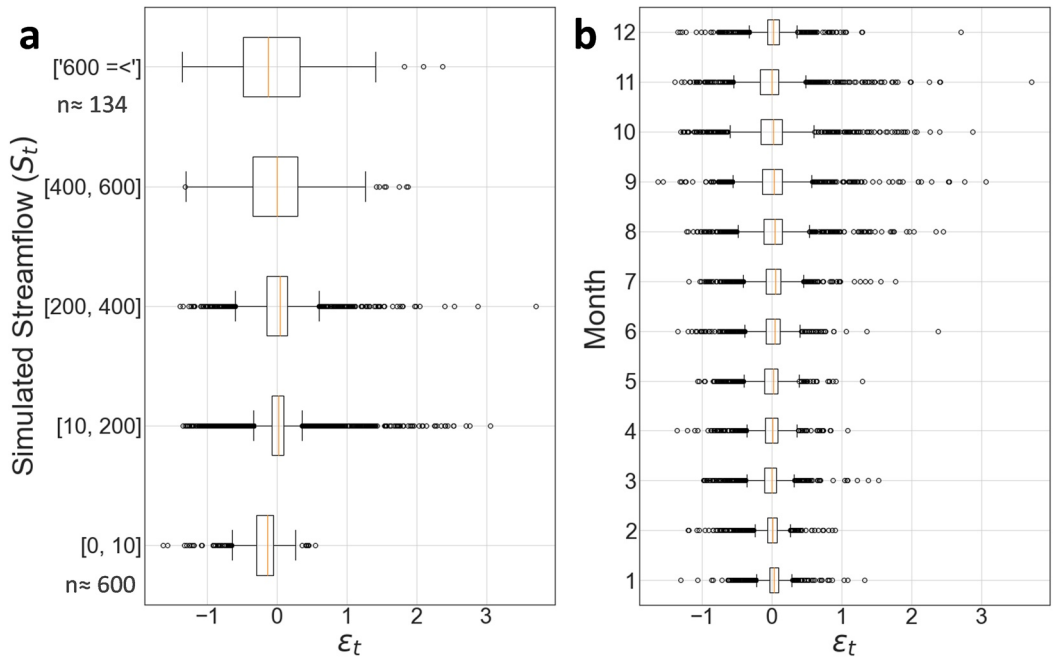

Figure 5. Distribution of ε_t in Equation 2 as a function of (a) simulated streamflow S_t and (b) month, demonstrating some evidence of heteroscedasticity.

One factor that is often overlooked when developing an SWM or stochastic streamflow generator is the uncertainty introduced by estimating the stochastic model parameters (Stedinger & Taylor, 1982b). Here, we account for error model parameter uncertainty in the estimated coefficients φ_i of the AR model by randomly generating these parameters during ensemble generation. Samples of $\tilde{\varphi}_i$ are drawn from a multivariate normal distribution with mean vector equal to the maximum likelihood coefficient estimates (φ_i) and covariance matrix inferred from the Fisher information matrix, based on standard asymptotic properties of the MLE (Stedinger & Taylor, 1982b). For each sample $\tilde{\varphi}_i$, we convert random sequences of $\tilde{\varepsilon}_t$ from bootstrap resampling to random sequences of $\tilde{\lambda}_t$ using Equation 2.

The synthetic series of log-ratios $\tilde{\lambda}_t$ are then converted to a series of synthetic streamflow using

$$\tilde{Q}_t = \frac{S_t}{e^{\tilde{\lambda}_t}} \mathrm{BCF} \tag{3}$$

where BCF is a transformation bias correction factor needed to account for the bias introduced by having to retransform the values of $\tilde{\lambda}_t$ back into real space. The bias correction factor BCF (Equation 4) is derived in the appendix under the assumption that the distribution of λ_t is approximately normal as was shown in Figure 3a, where μ_λ and σ_λ represent the mean and standard deviation of the $\tilde{\lambda}_t$ series:

$$\mathrm{BCF} = \exp\left(\mu_\lambda - \frac{\sigma_\lambda^2}{2}\right) \tag{4}$$

The full procedure for implementing the SWM is summarized in Figure 6. This procedure begins with the transformation in Equation 1, followed by removal of autocorrelation using the AR(3) model in Equation 2. Random AR(3) residuals $\tilde{\varepsilon}_t$, are generated using either a k-NN or monthly k-NN bootstrap, and then transformed

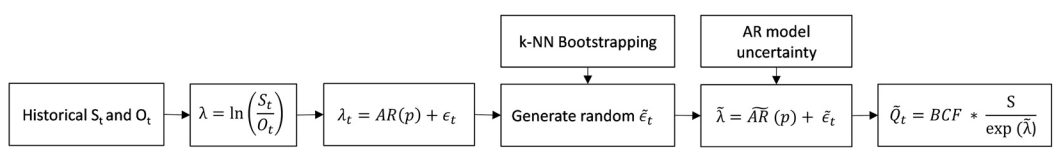

Figure 6. Stochastic watershed model post-processing method for generation of streamflow ensembles.

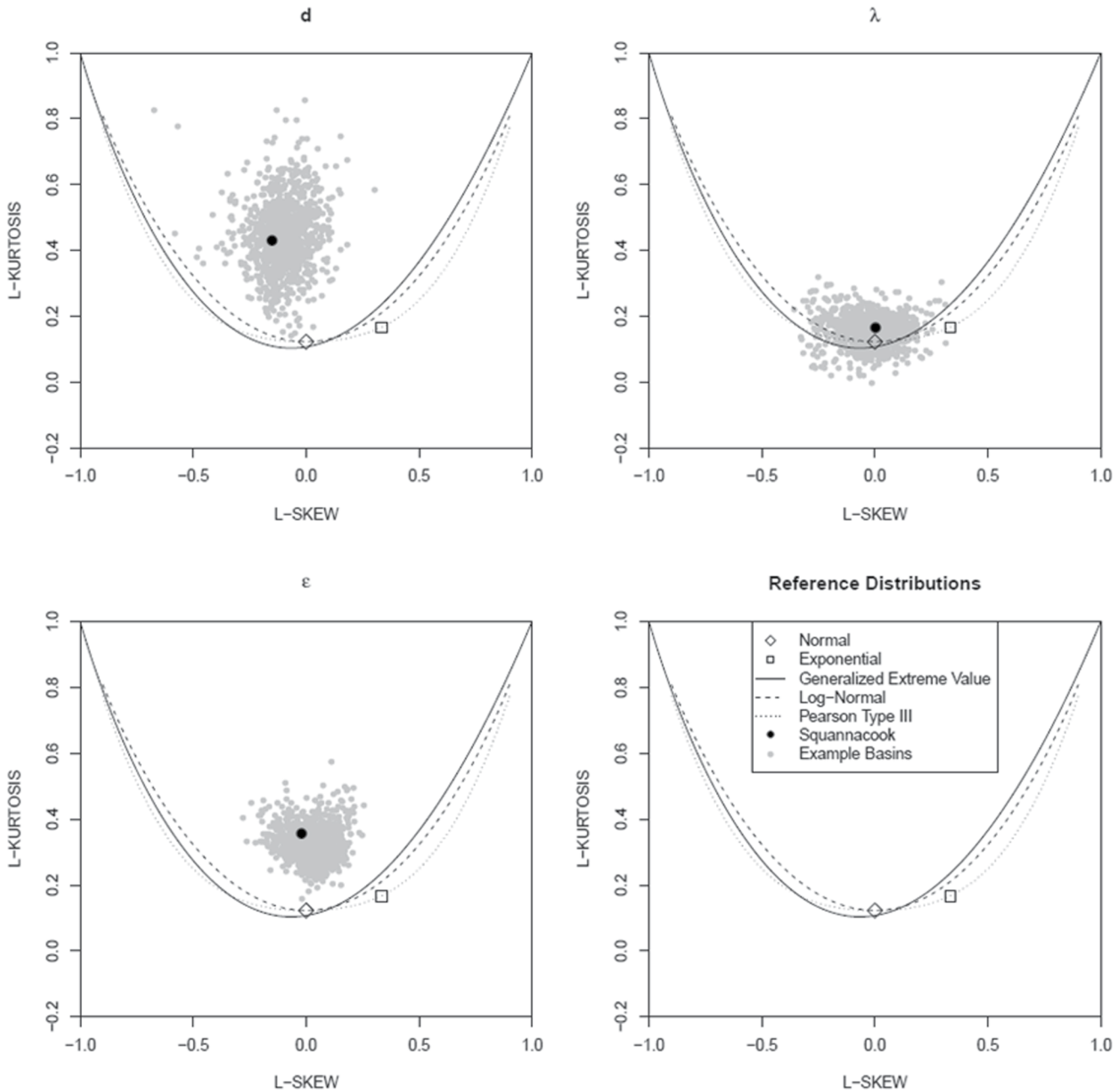

Figure 7. L-moment diagrams of (a) differenced residuals $d = S - O$, (b) log ratio residuals λ_t, and (c) AR(3) model residuals ε_t for the 1,225 USGS sites considered by Farmer and Vogel (2016). (d) Provides a legend of the reference distributions.

back to $\tilde{\lambda}_t$ using Equation 2. Finally, the real-space SWM streamflow ensembles are obtained by taking the back-transformation in Equation 3 with the transformation bias correction factor in Equation 4.

The unique features of our proposed pp approach to the development of an SWM, in comparison to past studies, is that in addition to the parsimonious accounting for the complex stochastic dependence and heteroscedastic residual structure, it includes a necessary correction for retransformation bias, a nonparametric k-NN bootstrap approach to residual resampling, and an accounting for the additional uncertainty introduced by having to estimate error model parameters. Importantly, the proposed SWM relies on three assumptions: the λ_t are normally distributed, the ε_t are symmetric, and that the ε_t are independent. These assumptions are more easily met using log-ratio residuals λ rather than differenced residuals d. We assess the general applicability of these assumptions for perennial watersheds below.

2.3. On the General Applicability of a Log Ratio Approach to Post-Processing

Recall that Figure 3 used normal probability plots to document for the case study that the differenced residuals, d, are slightly skewed with heavy tails yet the log ratio residuals λ_t in Equation 1 are approximately normally distributed. In this section, we examine the generality of these findings with extension to the distributional behavior of the AR(3) model residuals ε_t in Equation 2. Figure 7 depicts L-Moment diagrams for the differenced residuals $d = S - O$, log ratio residuals λ_t and for AR(3) model residuals ε_t corresponding to the 1,225 perennial watersheds across the contemporaneous United States considered by Farmer and Vogel (2016). What we observe in Figure 7 is that the differenced residuals generally exhibit negative skewness and very large values of L-Kurtosis, and thus exhibit extremely heavy tails, compared to the λ_t and to a lesser degree the AR(3) residuals ε_t. Our findings in Figure 7 may be significant for pp approaches like ours that involve a bootstrap approach to generate errors a method which is purported to fail under heavy tails (see chapter 9 of Chernick (2008)), because the L-Kurtosis is generally lower for ε_t than for d_t. Unfortunately, the literature is unclear on how heavy tails must be for the bootstrap to break down, thus future research is needed to address this issue and we do not consider it further. Farmer et al. (2021) also show that both λ_t and ε_t are approximately homoscedastic, whereas the differenced residuals exhibit enormous heteroscedasticity, creating another tremendous challenge for pp approaches based on the differenced residuals.

2.4. Verification and Validation of Streamflow Ensembles

A thorough evaluation of an SWM involves both verification and validation of the DWM upon which it is based, as well as verification and validation of the resulting stochastic streamflow ensembles. Here we focus on the latter. Verification of the proposed SWM's ensembles is implemented by evaluating whether the residuals of the fitted AR model ε_t in Equation 2 are symmetric and independent (to justify the application of a bootstrap), and by evaluating whether these properties are maintained in the simulated residuals $\tilde{\varepsilon}_t$.

Validation of the SWM is implemented by confirming that the model can reproduce important characteristics of the streamflow observations that may be impactful for long-range water resources planning. SWM ensemble validation procedures include several steps that are specific to a given application. In our case, we first analyze how well the ensemble spread of synthetic stochastic streamflow trajectories capture the spread of observed streamflow values using coverage probabilities. We then evaluate the ability of the SWM to reproduce some of the curves, distributions, and statistics commonly used in long-range water resources planning and HRM, including: the FDC of daily streamflow, the storage yield curve, the distribution of annual minimum 7-day streamflow (7-day low flow), and the distribution of annual maximum daily streamflow. These distributions and curves broadly characterize both aggregate and extreme behavior of daily streamflow relevant to a broad range of SWM applications. We also consider common statistics used in the design of water resources infrastructure, including the 7Q10 (7-day low flow with 10-year return period) and various design flood events with return periods between 2 and 500 years.

The "observed" design flood events and their uncertainty are estimated using a log-Pearson Type III (LP3) distribution fit to the observed annual maximum series following the recommendation of England et al. (2019). Sampling uncertainty in the LP3 design quantiles due to the limited record lengths is quantified through the quantile confidence interval (CI) procedure proposed by Chowdhury and Stedinger (1991). The fitted LP3 quantiles and their uncertainty are compared to the distribution of design events across ensembles obtained by fitting an LP3 distribution to each ensemble member. This enables us to assess whether the SWM ensembles are able to reproduce the underlying uncertainty in important design statistics useful for long-term planning.

3. Results

In this section, we verify and validate the proposed pp approach to develop an SWM for the Squannacook River (see Figure 1). We present results for the k-NN bootstrap resampling approach, with results for the monthly k-NN presented in the Supporting Information.

3.1. Verification of Streamflow Ensembles

The pp approach described in Section 2.2 was used to generate an ensemble of 10,000 realizations each consisting of 38 years of daily streamflows. For verification of the ensembles, we evaluate the symmetry and independence

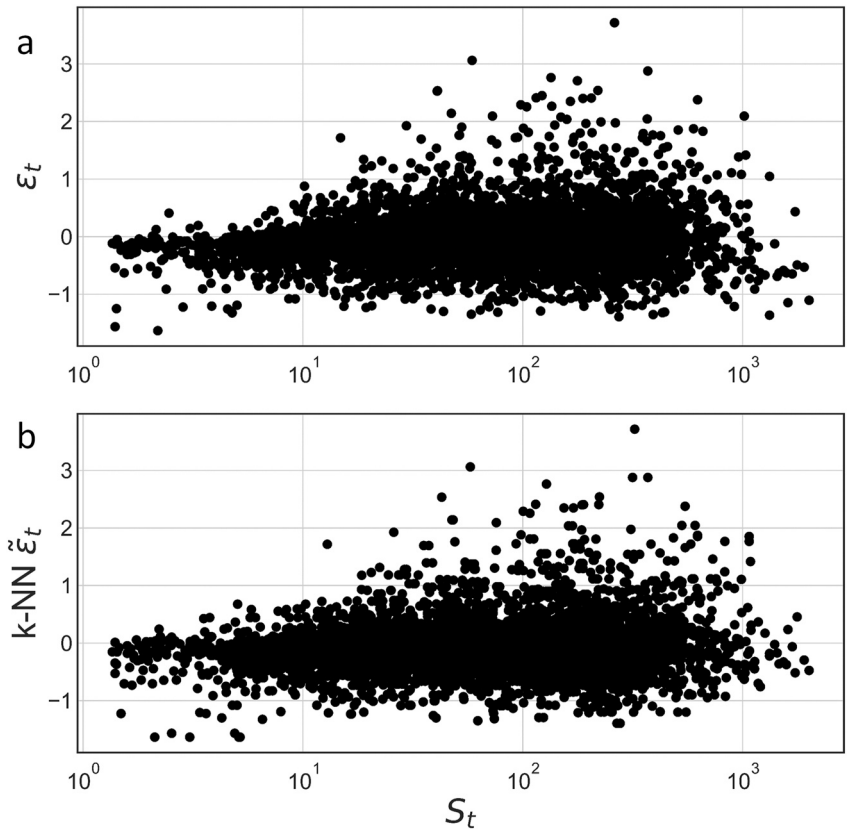

Figure 8. (a) Fitted AR model residuals ε_t in Equation 2 versus simulated flow S_t. (b) k-NN bootstrap residuals $\tilde{\varepsilon}_t$ versus simulated flow, S_t.

of the fitted AR(3) residuals ε_t in Equation 2. The distribution of ε_t for our pilot basin is centered at zero and is approximately symmetric (L-coefficient of skewness equal to 0.02).

In Figure 8, we evaluate whether the synthetic $\tilde{\varepsilon}_t$ generated using the k-NN bootstrap resemble those associated with the fitted residuals computed from the AR(3) model in Equation 2. Figure 8a plots the residuals ε_t in Equation 2 versus simulated streamflow S_t and illustrates some heteroscedasticity as evidenced by an increase in the variability of ε_t as S_t increases. We also find that the mean of ε_t varies, because days with $S_t \leq 10$ CFS have a lower mean ($\mu = -0.20$) than days with $S_t \geq 10$ CFS ($\mu = -0.01$), see also Figure 5. Figure 8b illustrates a sample realization of k-NN bootstrap residuals $\tilde{\varepsilon}_t$. The k-NN bootstrap residual distribution in Figure 8b exhibits good qualitative agreement with the fitted AR model residuals in Figure 8a especially in capturing the asymmetric heteroscedastic structure near the upper and lower tails of S_t.

As shown previously in Figures 4c and 4d, the AR model residuals ε_t are independent, justifying use of a bootstrap approach. In Figure 9, we examine the ability of the k-NN bootstrap of the errors ε_t to reproduce the ACF

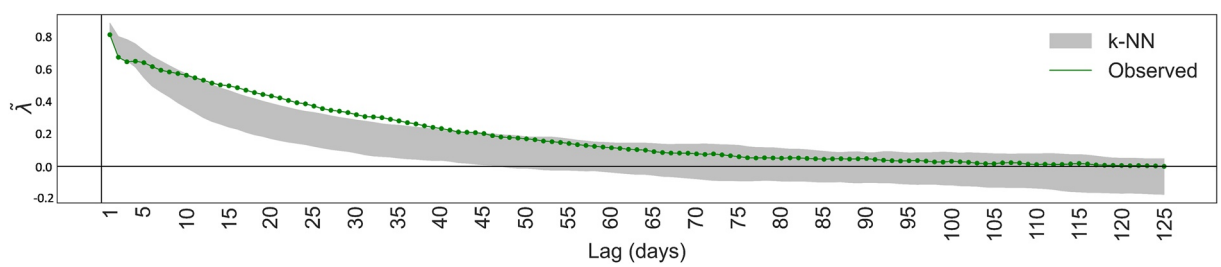

Figure 9. Comparison of autocorrelation functions for the observed λ_t in Equation 2 and those generated by k-nearest neighbor (k-NN) bootstrap. The green points denote the empirical autocorrelation function (ACF) based on the observed, λ_t series. The gray bands are the ACFs of $\tilde{\lambda}_t$ of the stochastic watershed model ensemble.

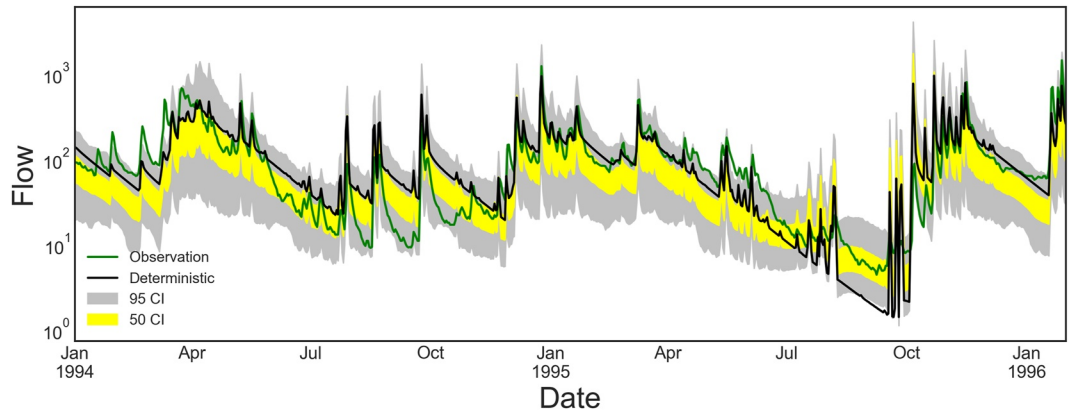

Figure 10. Partial validation of daily streamflow ensembles from the stochastic watershed model (SWM), using coverage probability. Hydrograph of model output between January 1994 and January 1996 with and 50%/95% confidence interval (CI) from the SWM ensemble.

of the λ_t series associated with the observations. As shown in Figure 9, $\tilde{\lambda}_t$ generated by the AR(3) model with the k-NN bootstrap reproduce the observed autocorrelation structure of λ out to lags of as much as 125 days. This is because the conditional k-NN bootstrap accounts for the slight heteroscedasticity which results from conditional bias in the residuals of the AR model by bootstrapping from streamflows that are similar to those on the day of interest.

3.2. Validation of Streamflow Ensembles

To date, perhaps the most common approach to validation of stochastic streamflow ensembles is to evaluate their coverage probabilities (Montanari & Brath, 2004; Montanari & Grossi, 2008; Montanari & Koutsoyiannis, 2012; Sikorska et al., 2015). Here, the coverage probabilities for our SWM are 89% for the 95% CI and 35% for the 50% CI. Figure 10 provides the hydrograph for a 2-year period in which the basin experienced an extreme drought, with particularly low flows in August and September of 1995. The figure suggests that the SWM performs well, even during low flow events when the deterministic model struggles. This is thanks, in part to the k-NN bootstrap which accounts for the change in distribution of ϵ_t as simulated flow S_t varies (see Figure 8). If one's only concern is coverage probability, as is often the case in flood forecasting applications, this result may constitute adequate validation (e.g., demonstration that the method is fit for purpose). However, water resources planners often have broader concerns, including the method's ability to capture the timing, magnitude, and distribution of extreme events and/or the usefulness in other water resources applications including estimating the flow-duration curve and the storage-yield curve. We consider this wider range of validation criteria below.

Figure 11 evaluates the ability of the SWM to generate ensembles that can reproduce various streamflow curves and statistics that are important to water resources planning and design. Figure 11a documents that the storage yield curve based on the sequent peak algorithm computed from observations (green) compares favorably with those based on the stochastic ensemble (gray). A comparison of the FDCs in Figure 11b illustrates that the stochastic ensemble generally reproduces the FDC of the observations, largely correcting the underestimation of the low flows exhibited by the DWM (black). This point is supported by considering the distribution of the 7-day low flows in Figure 11c, a common concern in drought, low-flow, and water quality planning applications. While the DWM (black) consistently underestimates the observed 7-day low flows (green), the stochastic ensemble (gray) neatly encloses the two. Figure 11d reports the distribution of the annual maximum flows, and shows that the SWM ensemble nicely reproduces the distribution of the observed annual maxima. Of particular importance is that while the DWM underestimates the magnitude of the events with the lowest annual exceedance probability (left-hand side of Figure 11d), the stochastic ensemble nicely captures the observed annual maxima within its uncertainty bounds for these low exceedance probabilities.

This last point is examined in more detail in Figure 12 and Table 1, which compare the sampling distribution of various common planning statistics derived from the stochastic ensemble to those estimated directly by the DWM and directly from the observations themselves. Ensemble predictions are often more skillful than deterministic

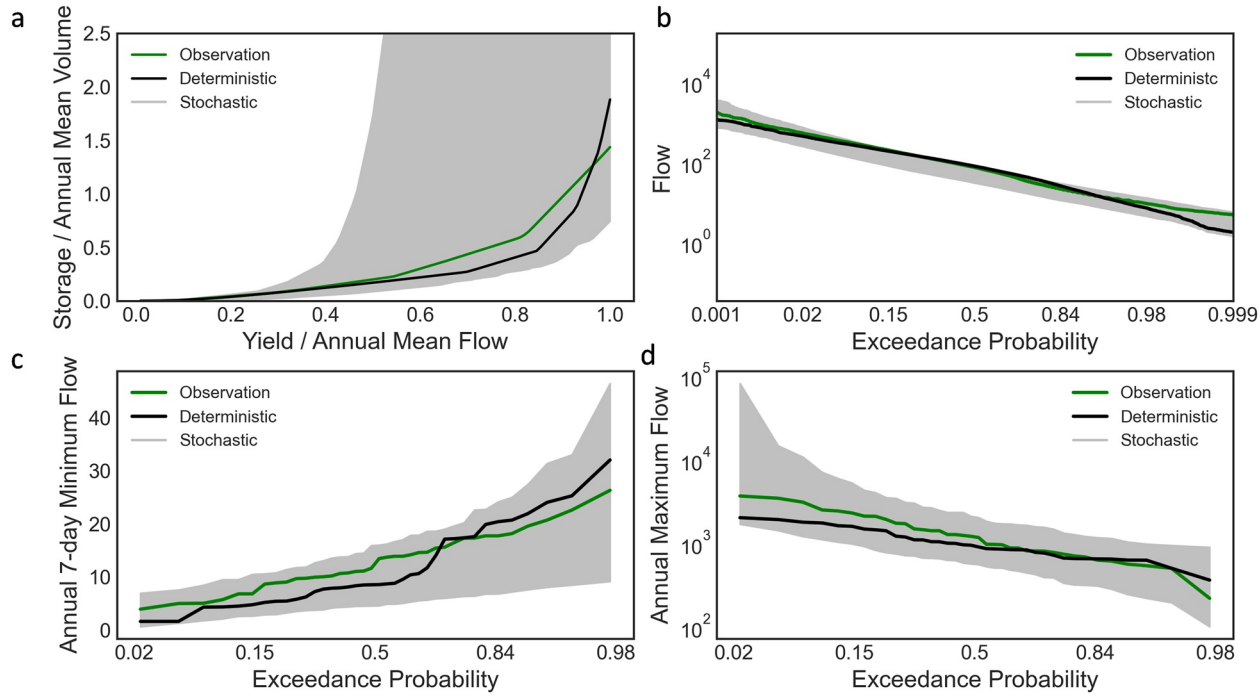

Figure 11. Performance of the stochastic watershed model in reproducing common curves and distributions used in long-range water resources planning and design. In each figure panel, observations are plotted in green, deterministic watershed model in black, and the stochastic ensemble in gray. (a) The storage ratio of a hypothetical reservoir versus yield ratio. (b) Daily flow duration curve, with horizontal axis as the exceedance probability. (c) Empirical cumulative distribution of annual minimum 7-day flow. (d) Empirical cumulative distribution of annual maximum flow.

ones (Cloke & Pappenberger, 2009; Roulin, 2007). Across the flood statistics and drought statistics considered in Figure 12 and Table 1, the mode of the stochastic ensemble is nearly always closer to the observed value than the deterministic model prediction. Furthermore, the LP3 quantile 90% CIs in Figures 12b–12d (green dashed lines) agree closely with the distribution derived by the stochastic ensemble, particularly for more extreme events. We

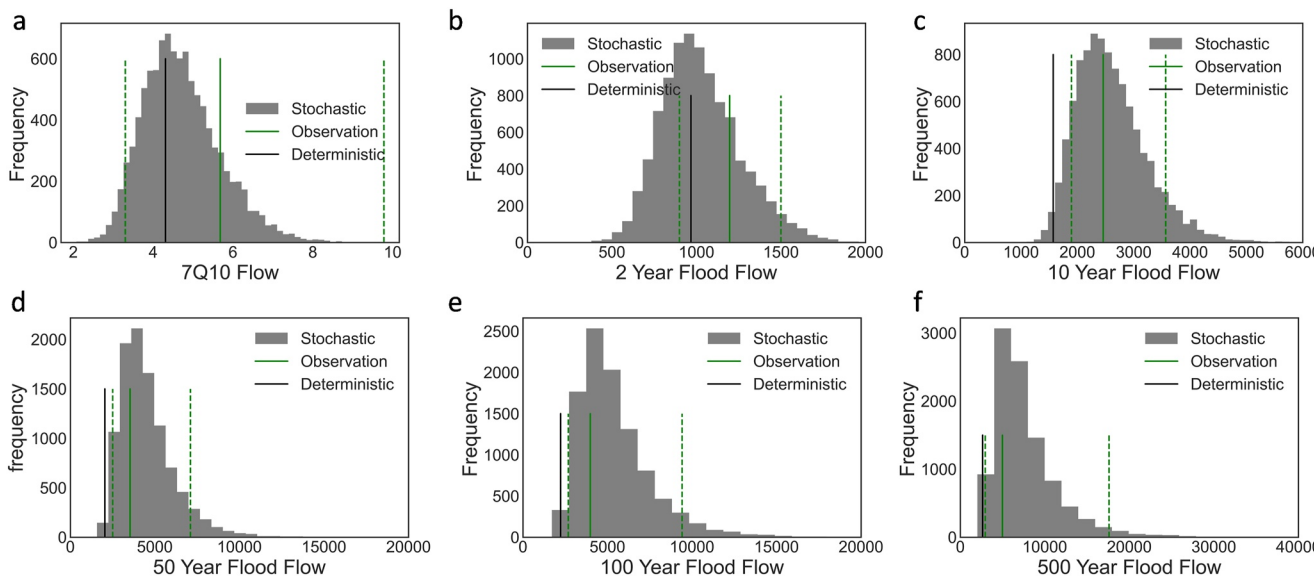

Figure 12. Comparison of the distribution of design flow quantile estimates based on 10,000 simple bootstrap stochastic watershed model ensembles (gray histograms) versus observations (solid green bar) and deterministic model (black bar). The dashed green lines are representative of 90% confidence intervals associated with the observed design flow. Results shown for the (a) 7Q10, (b) 2-year flood, (c) 10-year flood, (d) 50-year flood, (e) 100-year flood, and (f) 500-year flood.

Table 1
Summary of DWM and SWM Performance in Estimating Low-Flow and Flood-Flow Design Statistics

Design quantile	Observed flow (cfs)	Percent error DWM	Percent error SWM mode	SWM coverage of quantile 90% CI
7Q10	5.68	−24	−22.5	95
2-Year flood	1,195	−19	−21.3	64
10-Year flood	2,458	−36	−5.2	81
50-Year flood	3,534	−42	6.8	88
100-Year flood	3,978	−44	0.5	91
500-Year flood	4,987	−47	10.3	96

conclude that the mode of the SWM ensemble corrects for the underestimation of extreme floods associated with the DWM.

Though the mode of the stochastic ensemble of the 7-day low flow (7Q10) is marginally closer to the observed 7Q10, and the majority of the stochastic ensemble falls within the 90% CI of the observed 7Q10 (green dashed lines), the performance is not as compelling as for flood flows. This is in large part attributable to the poor performance of the DWM for very low flows (in the case of the Squannacook, particularly when $S_t < 10\,cfs$). Recall that in Figures 5 and 8, ε_t has a distinctly different distribution for lower flows. Though the k-NN bootstrap based on simulated S_t accounts for this, there is likely a limit to which any pp can be expected to correct deterministic biases from poor data, calibration, or process representation errors in the underlying DWM. Though improving the DWM would be one solution, it is often beyond the scope of the project (as is the case here) or may well be infeasible if applying a pp approach across a wide geographic region, for instance to every watershed in the U.S. National Hydrologic Model. An important conclusion from Figure 12 and Table 1 is that the proposed pp approach improves the estimation of both peak and low flow statistics compared to the DWM alone, but improvements may be limited as DWM biases grow. Further work should address whether this important result can be generalized to other basins and DWM models.

An alternative to the k-NN bootstrap approach presented here is the monthly k-NN bootstrap, because the ε_t from Equation 2 showed heteroscedasticity by both month and simulated flow S_t. The validation results for this approach are presented in the Supporting Information. In general, the results for the monthly k-NN and k-NN SWM approaches are comparable. They yield similar coverage probability, storage-yield curve, FDC, annual maximum distribution, and estimates of extreme flood quantiles (Figures S1–S3 in Supporting Information S1). The monthly k-NN also exhibits a similar ACF to the standard k-NN, which roughly approximates the observed value, a verification criterion (Figure S4 in Supporting Information S1). However, the monthly k-NN SWM fails to replicate the distribution of the 7-day low flows and produces a mode 7Q10 estimate that is worse than the DWM prediction, unlike the standard k-NN SWM. This is, in part because the very low flows are distributed across several months, so that there were an insufficient number of low flow residuals (in the case of the Squannacook $S_t < 10\,cfs$) to ensure good performance of the SWM. A seasonal k-NN may correct this issue, but we decline to pursue this. A promising direction for future work is a k-NN resampling that utilizes additional information about the hydrologic process when bootstrapping ε_t.

4. Discussion

4.1. Model Bias and Fitness for Purpose

In general, three different sources of bias can arise when generating ensembles of streamflow, including (a) deterministic biases, (b) stochastic biases, and (c) transformation biases. Deterministic biases arise from a flaw in the DWM simulation that results from misrepresentation of underlying physical processes or biases in input data and/or streamflow observations. Stochastic biases result from other factors which introduce variability and uncertainty into the properties of the model error and may be influenced by the behavior and properties of streamflow observations, model inputs, model outputs, parameter estimates, and the calibration approach. Finally, transformation bias arises when working with transformed model errors, because in general, the transformed moments will not equal to their untransformed moments so that $(E[f(x)] \neq f(E[x]))$. SWMs are developed to address stochastic biases, and the log transformation bias correction factor derived in the appendix corrects for transformation bias incurred by retransforming the log space errors back to real space when generating streamflow ensembles. However, SWMs can only indirectly address deterministic bias via a statistical correction (e.g., by using the k-NN bootstrap). When deterministic bias is identified, it should be addressed (if possible) by investigating corrections related to the input data, streamflow observations, calibration method, the underlying processes being modeled, or any other source of error determined to root cause of the deterministic bias. This is preferred to a statistical bias correction, which assumes that the deterministic bias is stationary and will continue into the future, a poor assumption in the face of land use and climate change. If deterministic bias correction is infeasible, one potential method to more fully encapsulate the uncertainty in the stochastic bias is to perform the kNN sampling on randomly sampled subsets of the historical data as recommended by Fadhel et al. (2017).

A parsimonious SWM developed in this work addressed stochastic and transformation biases through the use of several modeling choices: log-transformed innovation ratios, an AR(3) model, a k-NN bootstrap, and a transformation bias correction factor. To verify the appropriateness of these choices, we first evaluated the underlying assumption that the AR model residuals were symmetric and independent, and that any heteroscedasticity in the AR model residuals with simulated flow was captured by the k-NN bootstrap. We also verified that these modeling features together were able to reproduce the slowly decaying autocorrelation structure of the SWM errors λ_t (Figure 9). This verification process is fundamental to the development of a useful SWM, because it ensures that the aggregated pp approach used to generate streamflow ensembles mimics important statistical properties of the calibration model error upon which the SWM is based.

The SWM was further evaluated in a validation process of the SWM streamflow ensembles to ensure the model was fit for the purpose it was intended, which in this case is to support long-range water resources planning. The most common validation approach for ensembles is to compute the coverage probability of SWM flow ensembles with observed streamflow data (Laio & Tamea, 2007; Montanari & Koutsoyiannis, 2012; Sikorska et al., 2015). The SWM produced coverage probabilities associated with the 95% and 50% streamflow CIs that were slightly lower than the desired values (89% and 36%, respectively). This result is likely due to reduced variance in its associated AR residuals arising from the k-NN bootstrap, which caused corresponding lower variability of the resulting SWM streamflow ensembles.

Despite coverage probabilities that were lower than may be desired, Figure 11 shows that the proposed SWM reproduced the observed storage-yield curve, FDC, the distribution of annual 7-day low flows and annual maximum flows very well. Furthermore, Figure 12 and Table 1 illustrate that the proposed SWM reproduced low flow extremes and small flood flows reasonably well and large flood flow extremes exceptionally well, as compared to the extreme events estimated directly from the observations and from direct use of the DWM. For each study, specific validation criteria should be driven by the purpose of the proposed model. As this study was primarily concerned with quantifying extremes used in long-range hydrologic planning and design, the coverage probabilities were deemed acceptable in light of the good performance of the SWM in capturing extreme hydrologic events and their uncertainty.

4.2. Extension to Intermittent Sites

This study has only advanced an approach to the development of an SWM for perennial sites, yet a major challenge remains to extend our pp approach to intermittent sites with zero observations. Nearly all previous pp efforts and/or DWM calibration efforts at intermittent sites tend to combine the zero and nonzero streamflows, treating them as continuous random variables. This is implicit, for example, when one computes goodness-of-fit metrics such as NSE and/or when the zeros and nonzeros are all combined into a single objective function during model calibration. Although this study does not attempt to deal with zero streamflows, we wish to clarify at the outset, that we believe zero streamflows should be treated separately from nonzero streamflows, because they represent a distinct state of the watershed.

In addition to the SWM advanced here for perennial sites, we argue that a mixture model is needed that can generate sequences of zero streamflows that reproduce the transition probabilities between nonzero and zero flows and vice versa. A major challenge associated with both the calibration of a DWM and generation of streamflow ensembles using an SWM at intermittent sites involves the fact that most DWM baseflow generation algorithms will never generate zero observations. Therefore, it may be necessary to treat the zero observations as censored observations. In the Supporting Information, we describe the use of a ROC curve during model calibration to determine the optimal censoring threshold to distinguish very small simulations S from actual zero observations. Overall, intermittent sites create two challenges: (a) new approaches to DWM calibration which are designed to handle zero observations as separate from the nonzeros and (b) generation of zeros which capture the structure of all relevant transition probabilities as well as the stochastic structure of the intermittent process. Each of these challenges is discussed in the Supporting Information.

5. Conclusion

SWMs are an increasingly important component of water resource planning due to the need to account for climate, land use, and other change, combined with the widespread use of risk-based approaches in modern

decision-making processes, nearly all of which require streamflow ensembles. There is an increased need for SWMs that can be easily deployed at multiple temporal and spatial scales. Most available SWMs for long-range water resource planning applications have relied on complex Bayesian approaches, making their immediate use in operational hydrology challenging. We develop a simpler AR log-ratio SWM bootstrap approach which is shown to have great potential for long-range planning applications. This approach circumvents much of the stochastic complexity of traditional error modeling approaches that rely on the arithmetic difference between simulations and observations to define predictive uncertainty. Instead, our log-ratio AR(p) approach yields AR model errors ε_t which were well approximated by a symmetrically distributed and independent process, enabling a k-NN bootstrap resampling method which makes the modeling process substantially more approachable than alternative methods.

Using calibrated DWM's at hundreds of basins across the United States, Figure 7 in this study and Farmer et al. (2021) have shown that the stochastic AR log ratio error model employed in this study may be quite general. We argue that the relative simplicity of such an SWM makes it easily adaptable to a wide range of uses, including long-range planning applications to which SWMs have not yet been widely implemented. We emphasize that while our pp method is promising, we have not considered the stochastic variation associated with model inputs; precipitation and temperature. A promising approach for further work would involve the use of the generalized COSMOS stochastic modeling package (Papalexiou, 2018) for the stochastic generation of precipitation, temperature, and possibly for the log ratio model errors.

As SWMs become widespread across the field of water resources planning and hydrology, the need for a proper model selection process has increased. We have developed a rigorous model selection process (verification and validation of streamflow ensembles) in addition to model development, and show how commonly used metrics (e.g., coverage probabilities associated with uncertainty intervals) are not sufficient criteria on their own for model acceptance. Any model selection framework should include a variety of metrics that capture model performance across the conditions most likely to expose system vulnerabilities. In general, these should include assessing whether the watershed model error assumptions are maintained across the SWM ensembles (verification), as well as evaluating SWM ensembles for their ability to reproduce important behavioral properties of the historical flows such as the FDC, storage yield curve, and the distribution of low or high flow extremes (validation). For a generic long-range planning problem, we propose the following verification and validation steps for SWM ensembles:

- (Verification of Ensembles) Stochastic error model should reproduce entire stochastic structure of the deterministic model calibration errors across generated ensembles including adequate removal and consideration of retransformation bias, skewness, heteroscedasticity, and serial correlation.
- (Validation of Ensembles) SWM ensembles should be used to construct CIs for: storage yield curve, FDC, and extreme design event quantiles and those CIs should enclose (true) values based on both the DWM and the historical streamflow observations.

An important finding in this study is shown in Figure 12, which demonstrates that the AR(3) k-NN bootstrap pp algorithm led to design floods/droughts based on the mode of the ensemble which was nearly always closer to the design flood/drought based on the observations than to the design events based on the deterministic simulations. This shows the value of the pp algorithm to improve flood and low flow frequency analysis at one site. This finding is consistent with other studies (Cloke & Pappenberger, 2009; Roulin, 2007).

In addition to reproduction of coverage probabilities, we argue that SWM ensembles that pass the above outlined verification-validation steps would be a promising candidate for a variety of long-range planning and decision-making applications in water resources. Nevertheless, we recognize that each application is different and will have its own specific requirements, so we offer these verification-validation steps merely as a starting point.

While the proposed parsimonious SWM captures the uncertainty needed in long-term planning without implementation of much more complex mathematical processes, future work is needed to extend implementation of the approach in a multivariate setting for spatially correlated basins as well as to small ungauged basins (Grimaldi et al., 2022). Future work should also consider alternatives to the kNN approach used here, particularly if the residual resampling is conditioned on multiple predictors to describe the hydrologic state (see e.g., Sharma et al., 2016). It should also be noted that the presented DWM and SWM were only implemented in a relatively wet basin with perennial flows. Ongoing research expands the SWM introduced here to intermittent basins with

observed and modeled zero flows and addresses technical challenges of dealing with log ratio residual of zero flows (see Section 4.2 and Supporting Information for specific recommendations). Ongoing work also considers multisite applications of the proposed SWM, which raises the challenge of accounting for cross-correlation in the DWM errors in adjacent basins.

Appendix A: Transformation Bias Correction Factor Derivation

A bias correction is needed due to the retransformation bias introduced by having to retransform from log space to real space. Stochastic streamflow ensembles are generated using

$$\tilde{S}_t = S_t \exp\left(-\tilde{\lambda}_t\right) \tag{A1}$$

where $\tilde{\lambda}$ is normally distributed with mean μ_λ and standard deviation σ_λ, S_t is simulated streamflow from the DWM, and $\tilde{S}_t$ is the resulting stochastic ensemble streamflow. The problem is that

$$E\left[\tilde{S}_t\right] = E\left[S_t \exp\left(-\tilde{\lambda}_t\right)\right] \neq E[O_t] \tag{A2}$$

where O_t is the observations. Thus, a bias correction is needed to ensure that $E\left[\tilde{S}_t\right] = E[O_t]$ which is obtained by adding a bias correction factor BCF to Equation A1 so that

$$\tilde{S}_t = \frac{S_t \exp\left(-\tilde{\lambda}_t\right)}{BCF} = \frac{S_t \exp\left(-\tilde{\lambda}_t\right)}{E\left[\exp\left(-\tilde{\lambda}_t\right)\right]} \tag{A3}$$

Thus, the challenge is to derive an expression for $BCF = E\left[\exp\left(\tilde{\lambda}_t\right)\right]$ assuming $\tilde{\lambda}_t$ is normally distributed with mean μ_λ and standard deviation σ_λ. Noting from the moment generating function of a lognormal variable that for any normally distributed variable Z with mean μ and standard deviation σ,

$$E\left[e^{tZ}\right] = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) \tag{A4}$$

The result in Equation A4 can be used to derive the BCF in Equation A3 by noting that $\tilde{\lambda}_t$ is normally distributed with mean μ_λ and standard deviation σ_λ. Substitution of those moments along with $t = -1$ into Equation A4 leads to

$$BCF = E\left[\exp\left(-\tilde{\lambda}_t\right)\right] = \exp\left(-\mu_\lambda + \frac{\sigma_\lambda^2}{2}\right). \tag{A5}$$

Data Availability Statement

Computer code and data used in this study are available online at: http://10.5281/zenodo.7510764 (Gshabestani, 2022).

Acknowledgments
The authors would like to thank Viki Zoltay, Scott Olson, and Greg Stewart for their comments on early versions of this work, and the journal reviewers whose constructive comments improved the manuscript. The authors also wish to thank Simon Papalexiou for numerous thoughtful conversations and ideas which led to improvements in our approach. This work was supported by the Massachusetts Executive Office of Energy and Environmental Affairs.

References

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., & Salamon, P. (2014). Evaluation of ensemble streamflow predictions in Europe. *Journal of Hydrology*, *517*, 913–922. https://doi.org/10.1016/j.jhydrol.2014.06.035

Beven, K. (2019). Towards a methodology for testing models as hypotheses in the inexact sciences. *Proceedings of the Royal Society A: Mathematical, Physical & Engineering Sciences*, *475*(2224), 20180862. https://doi.org/10.1098/rspa.2018.0862

Bradley, A. A., Schwartz, S. S., & Hashino, T. (2004). Distributions-oriented verification of ensemble streamflow predictions. *Journal of Hydrometeorology*, *5*(3), 532–545. https://doi.org/10.1175/1525-7541(2004)005<0532:DVOESP>2.0.CO;2

Brekke, L. D. (2009). *Climate change and water resources management: A federal perspective*. U.S. Geological Survey.

Chernick, M. (2008). *Bootstrap methods: A guide for practitioners and researchers* (2nd ed.). John Wiley and Sons.

Chowdhury, J. U., & Stedinger, J. R. (1991). Confidence interval for design floods with estimated skew coefficient. *Journal of Hydraulic Engineering*, *117*(7), 811–831. https://doi.org/10.1061/(ASCE)0733-9429

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., et al. (2021). The abuse of popular performance metrics in hydrologic modeling. *Water Resources Research*, *57*(9), e2020WR029001. https://doi.org/10.1029/2020WR029001

Cloke, H. L., & Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, *375*(3–4), 613–626. https://doi.org/10.1016/j.jhydrol.2009.06.005

England, J., Cohn, T. A., Faber, B. A., Stedinger, J. R., Jr, W. O. T., Veilleux, A. G., et al. (2019). Guidelines for determining flood flow frequency—Bulletin 17C. In *Techniques and methods (No. 4-B5)*. U.S. Geological Survey. https://doi.org/10.3133/tm4B5

Evin, G., Kavetski, D., Thyer, M., & Kuczera, G. (2013). Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration: Technical note. *Water Resources Research*, *49*(7), 4518–4524. https://doi.org/10.1002/wrcr.20284

Fadhel, S., Rico-Ramirez, M. A., & Han, D. (2017). Uncertainty of Intensity–Duration–Frequency (IDF) curves due to varied climate baseline periods. *Journal of Hydrology*, *547*, 600–612. https://doi.org/10.1016/j.jhydrol.2017.02.013

Farmer, W. H., Shabestanipour, G., Lamontagne, J., & Vogel, R. (2021). *Stochastic watershed models using a logarithmic transformation of ratio residuals*. Copernicus Meetings.

Farmer, W. H., & Vogel, R. M. (2016). On the deterministic and stochastic use of hydrologic models. *Water Resources Research*, *52*(7), 5619–5633. https://doi.org/10.1002/2016wr019129

Fiering, M. B. (1967). *Stream flow synthesis*. Harvard University Press.

Grimaldi, S., Volpi, E., Langousis, A., Michael Papalexiou, S., Luciano De Luca, D., Piscopia, R., et al. (2022). Continuous hydrologic modelling for small and ungauged basins: A comparison of eight rainfall models for sub-daily runoff simulations. *Journal of Hydrology*, *610*, 127866. https://doi.org/10.1016/j.jhydrol.2022.127866

Gshabestani (2022). Gshabestani/LRM-Squannacook: (SWM). *Zenodo*. https://doi.org/10.5281/ZENODO.6084085

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

Hashimoto, T., Stedinger, J. R., & Loucks, D. P. (1982). Reliability, resiliency, and vulnerability criteria for water resource system performance evaluation. *Water Resources Research*, *18*(1), 14–20. https://doi.org/10.1029/WR018i001p00014

Herman, J. D., Quinn, J. D., Steinschneider, S., Giuliani, M., & Fletcher, S. (2020). Climate adaptation as a control problem: Review and perspectives on dynamic water resources planning under uncertainty. *Water Resources Research*, *56*(2), e24389. https://doi.org/10.1029/2019WR025502

Herman, J. D., Reed, P. M., Zeff, H. B., & Characklis, G. W. (2015). How should robustness be defined for water systems planning under change? *Journal of Water Resources Planning and Management*, *141*(10), 04015012. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000509

Hunter, J., Thyer, M., McInerney, D., & Kavetski, D. (2021). Achieving high-quality probabilistic predictions from hydrological models calibrated with a wide range of objective functions. *Journal of Hydrology*, *603*, 126578. https://doi.org/10.1016/j.jhydrol.2021.126578

Klemeš, V. (1986). Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, *31*(1), 13–24. https://doi.org/10.1080/02626668609491024

Koutsoyiannis, D., & Montanari, A. (2022). Bluecat: A local uncertainty estimator for deterministic simulations and predictions. *Water Resources Research*, *58*(1), e2021WR031215. https://doi.org/10.1029/2021WR031215

Kuczera, G., Kavetski, D., Franks, S., & Thyer, M. (2006). Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. *Journal of Hydrology*, *331*(1–2), 161–177. https://doi.org/10.1016/j.jhydrol.2006.05.010

Kuczera, G., Kavetski, D., Renard, B., & Thyer, M. (2017). Bayesian methods, Chapter 23. In V. P. Singh (Ed.), *Handbook of applied hydrology*. McGraw Hill Book Co.

Laio, F., & Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, *11*(4), 1267–1277. https://doi.org/10.5194/hess-11-1267-2007

Lall, U., & Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research*, *32*(3), 679–693. https://doi.org/10.1029/95WR02966

Lamontagne, J. R., Barber, C. A., & Vogel, R. M. (2020). Improved estimators of model performance efficiency for skewed hydrologic data. *Water Resources Research*, *56*(9), e2020WR027101. https://doi.org/10.1029/2020WR027101

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., & Di, Z. (2017). A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *WIREs Water*, *4*(6), e1246. https://doi.org/10.1002/wat2.1246

Loucks, D. P., & van Beek, E. (2017). *Water resource systems planning and management*. Springer International Publishing. https://doi.org/10.1007/978-3-319-44234-1

Maass, A., Hufschmidt, M. M., Dorfman, R., Thomas, J. H. A., Marglin, S. A., & Fair, G. M. (1962). *Design of water-resource systems*. Harvard University Press.

Markstrom, S. L., Regan, R. S., Hay, L. E., Viger, R. J., Webb, R. M., Payn, R. A., & LaFontaine, J. H. (2015). PRMS-IV, the precipitation-runoff modeling system, version 4. In *Techniques and methods (No. 6-B7)*. U.S. Geological Survey. https://doi.org/10.3133/tm6B7

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., & Kuczera, G. (2017). Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. *Water Resources Research*, *53*(3), 2199–2239. https://doi.org/10.1002/2016WR019168

Meyer, E. S., Sheer, D. P., Rush, P. V., Vogel, R. M., & Billian, H. E. (2020). Need for process based empirical models for water quality management: Salinity management in the Delaware River Basin. *Journal of Water Resources Planning and Management*, *146*(9), 05020018. https://doi.org/10.1061/(asce)wr.1943-5452.0001260

Montanari, A., & Brath, A. (2004). A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. *Water Resources Research*, *40*(1), W01106. https://doi.org/10.1029/2003wr002540

Montanari, A., & Grossi, G. (2008). Estimating the uncertainty of hydrological forecasts: A statistical approach. *Water Resources Research*, *44*(12), W00B08. https://doi.org/10.1029/2008wr006897

Montanari, A., & Koutsoyiannis, D. (2012). A blueprint for process-based modeling of uncertain hydrological systems. *Water Resources Research*, *48*(9), 2011WR011412. https://doi.org/10.1029/2011WR011412

Morawietz, M., Xu, C.-Y., Gottschalk, L., & Tallaksen, L. M. (2011). Systematic evaluation of autoregressive error models as post-processors for a probabilistic streamflow forecast system. *Journal of Hydrology*, *407*(1–4), 58–72. https://doi.org/10.1016/j.jhydrol.2011.07.007

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

Papalexiou, S. M. (2018). Unified theory for stochastic modelling of hydroclimatic processes: Preserving marginal distributions, correlation structures, and intermittency. *Advances in Water Resources*, *115*, 234–252. https://doi.org/10.1016/j.advwatres.2018.02.013

Prairie, J. R., Rajagopalan, B., Fulp, T. J., & Zagona, E. A. (2006). Modified K-NN model for stochastic streamflow simulation. *Journal of Hydrologic Engineering*, *11*(4), 371–378. https://doi.org/10.1061/(ASCE)1084-0699

Quinn, J. D., Reed, P. M., Giuliani, M., Castelletti, A., Oyler, J. W., & Nicholas, R. E. (2018). Exploring how changing monsoonal dynamics and human pressures challenge multireservoir management for flood protection, hydropower production, and agricultural water supply. *Water Resources Research*, *54*(7), 4638–4662. https://doi.org/10.1029/2018WR022743

Regan, R. S., Juracek, K. E., Hay, L. E., Markstrom, S. L., Viger, R. J., Driscoll, J. M., et al. (2019). The U. S. Geological Survey National Hydrologic Model infrastructure: Rationale, description, and application of a watershed-scale model for the conterminous United States. *Environmental Modelling & Software*, *111*, 192–203. https://doi.org/10.1016/j.envsoft.2018.09.023

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), W11516. https://doi.org/10.1029/2011wr010643

Roulin, E. (2007). Skill and relative economic value of medium-range hydrological ensemble predictions. *Hydrology and Earth System Sciences*, *11*(2), 725–737. https://doi.org/10.5194/hess-11-725-2007

Salas, J. D., Delleur, J. W., Yevjevich, V. M., & Lane, W. L. (1980). *Applied modeling of hydrologic time series*. Water Resources Publications.

Schoups, G., & Vrugt, J. A. (2010). A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research*, *46*(10), W10531. https://doi.org/10.1029/2009wr008933

Sharma, A., Mehrotra, R., Li, J., & Jha, S. (2016). A programming tool for nonparametric system prediction using Partial Informational Correlation and Partial Weights. *Environmental Modelling & Software*, *83*, 271–275. https://doi.org/10.1016/j.envsoft.2016.05.021

Sharma, A., Wasko, C., & Lettenmaier, D. P. (2018). If precipitation extremes are increasing, why aren't floods? *Water Resources Research*, *54*(11), 8545–8551. https://doi.org/10.1029/2018WR023749

Sikorska, A. E., Montanari, A., & Koutsoyiannis, D. (2015). Estimating the uncertainty of hydrological predictions through data-driven resampling techniques. *Journal of Hydrologic Engineering*, *20*(1), A4014009. https://doi.org/10.1061/(asce)he.1943-5584.0000926

Stakhiv, E. Z. (2011). Pragmatic approaches for water management under climate change uncertainty 1. *Journal of the American Water Resources Association*, *47*(6), 1183–1196. https://doi.org/10.1111/j.1752-1688.2011.00589.x

Stedinger, J. R., & Taylor, M. R. (1982a). Synthetic streamflow generation: 1. Model verification and validation. *Water Resources Research*, *18*(4), 909–918. https://doi.org/10.1029/wr018i004p00909

Stedinger, J. R., & Taylor, M. R. (1982b). Synthetic streamflow generation: 2. Effect of parameter uncertainty. *Water Resources Research*, *18*(4), 919–924. https://doi.org/10.1029/WR018i004p00919

Steinschneider, S., & Lall, U. (2015). A hierarchical Bayesian regional model for nonstationary precipitation extremes in Northern California conditioned on tropical moisture exports. *Water Resources Research*, *51*(3), 1472–1492. https://doi.org/10.1002/2014WR016664

Steinschneider, S., Polebitski, A., Brown, C., & Letcher, B. H. (2012). Toward a statistical framework to quantify the uncertainties of hydrologic response under climate change. *Water Resources Research*, *48*(11), W11525. https://doi.org/10.1029/2011WR011318

Steinschneider, S., Wi, S., & Brown, C. (2015). The integrated effects of climate and hydrologic uncertainty on future flood risk assessments. *Hydrological Processes*, *29*(12), 2823–2839. https://doi.org/10.1002/hyp.10409

Tajiki, M., Schoups, G., Hendricks Franssen, H. J., Najafinejad, A., & Bahremand, A. (2020). Recursive Bayesian estimation of conceptual rainfall-runoff model errors in real-time prediction of streamflow. *Water Resources Research*, *56*(2), e2019WR025237. https://doi.org/10.1029/2019WR025237

Teegavarapu, R. S. V., Salas, J. D., & Stedinger, J. R. (Eds.), (2019). *Statistical analysis of hydrologic variables: Methods and applications*. American Society of Civil Engineers. https://doi.org/10.1061/9780784415177

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

Valdez, E. S., Anctil, F., & Ramos, M.-H. (2022). Choosing between post-processing precipitation forecasts or chaining several uncertainty quantification tools in hydrological forecasting systems. *Hydrology and Earth System Sciences*, *26*(1), 197–220. https://doi.org/10.5194/hess-26-197-2022

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., et al. (2021). Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, *102*(3), E681–E699. https://doi.org/10.1175/BAMS-D-19-0308.1

Vannitsem, S., Wilks, D. S., & Messner, J. W. (Eds.), (2019). *Statistical postprocessing of ensemble forecasts*. Elsevier. https://doi.org/10.1016/B978-0-12-812372-0.09993-3

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

Vogel, R. M., & Sankarasubramanian, A. (2003). Validation of a watershed model without calibration. *Water Resources Research*, *39*(10), 1292. https://doi.org/10.1029/2002WR001940

Wilks, D. S. (2019). *Statistical methods in the atmospheric sciences* (4th ed.). Elsevier.

Zha, X., Xiong, L., Guo, S., Kim, J.-S., & Liu, D. (2020). AR-GARCH with exogenous variables as a postprocessing model for improving streamflow forecasts. *Journal of Hydrologic Engineering*, *25*(8), 04020036. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001955

References From the Supporting Information

Nowak, K., Prairie, J., Rajagopalan, B., & Lall, U. (2010). A nonparametric stochastic approach for multisite disaggregation of annual to daily streamflow. *Water Resources Research*, *46*(8), W08529. https://doi.org/10.1029/2009WR008530

Oommen, T., Baise, L. G., & Vogel, R. M. (2011). Sampling bias and class imbalance in maximum-likelihood logistic regression. *Mathematical Geosciences*, *43*(1), 99–120. https://doi.org/10.1007/s11004-010-9311-8

Wang, Q. J., & Robertson, D. E. (2011). Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. *Water Resources Research*, *47*(2), W02546. https://doi.org/10.1029/2010WR009333

Ye, L., Gu, X., Wang, D., & Vogel, R. M. (2021). An unbiased estimator of coefficient of variation of streamflow. *Journal of Hydrology*, *594*, 125954. https://doi.org/10.1016/j.jhydrol.2021.125954