Water Resources Research

RESEARCH ARTICLE
10.1029/2020WR027101

Improved Estimators of Model Performance Efficiency for Skewed Hydrologic Data

Jonathan R. Lamontagne[1], **Caitline A. Barber**[1], **and Richard M. Vogel**[1]

[1]Department of Civil and Environmental Engineering, Tufts University, Medford, MA, USA

Key Points:

- This is the first study to perform controlled experiments concerning the widely used Nash-Sutcliffe (*NSE*) and Kling-Gupta (*KGE*) efficiencies
- NSE is unbiased but not suited for use with daily streamflows due to its enormous variability from one sample to another
- Considerably improved estimators of efficiency are introduced that are based on a bivariate lognormal monthly mixture model

Correspondence to:
R. M. Vogel,
richard.vogel@tufts.edu

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

Received 9 JAN 2020
Accepted 27 AUG 2020
Accepted article online 2 SEP 2020

Abstract The Nash-Sutcliffe efficiency (*NSE*) and the Kling-Gupta efficiency (*KGE*) are now the most widely used indices in hydrology for evaluation of the goodness of fit between model simulations *S* and observations *O*. We introduce two theoretical (probabilistic) definitions of efficiency, *E* and *E′*, based on the estimators *NSE* and *KGE*, respectively, which enable controlled Monte Carlo experiments at 447 watersheds to evaluate their performance. Although *NSE* is generally unbiased, it exhibits enormous variability from one sample to another, due to the remarkable skewness and periodicity of daily streamflow data. However, use of *NSE* with logarithms of daily streamflow leads to estimates of *E* with almost no variability from one sample to the next, though with high upward bias. We introduce improved estimators of *E* and *E′* based on a bivariate lognormal monthly mixture model that are shown to yield considerable improvements over *NSE* and slight improvements over *KGE* in controlled Monte Carlo experiments. Our new estimators of *E* should avoid most previous criticisms of *NSE* implied by the literature. Improved estimators of *E* that account for skewness and periodicity are needed for daily and subdaily streamflow series because *NSE* is not suited to such applications.

Plain Language Summary Reliable metrics are needed, which summarize the degree to which simulation model output reproduces the observations. Two of the most widely used metrics are Nash-Sutcliffe efficiency (*NSE*) and the Kling-Gupta efficiency (*KGE*). Remarkably, this is the first study to provide a theoretical definition and treatment of these indices enabling controlled Monte Carlo experiments to evaluate their performance. Controlled experiments at 447 U.S. watersheds enable us to report the degree of bias and variability associated with these indices when applied to daily data. As expected, *NSE* is on average, equal to its theoretical value; thus, it provides an unbiased estimate of its theoretical value. However, *NSE* exhibits enormous variability from one sample to another due to the enormous skewness and periodicity of daily streamflows. Improved estimators are introduced, which account for skewness and periodicity of daily streamflow observations. Our improved estimators yield considerable improvements over *NSE* and slight improvements over *KGE* and are shown to avoid most previous criticisms of *NSE* implied by the literature. Simulation models are increasingly being used to mimic high frequency observations, which exhibit highly skewed and periodic behavior. In such instances, improved estimators of efficiency are needed because *NSE* is no longer suited to such applications.

1. Introduction

The Nash-Sutcliffe efficiency (*NSE*) is a widely used sample statistic that, until this study, did not have a theoretical definition. Numerous concerns have been raised about the sensitivity of *NSE* to outliers, seasonality, and many other issues. The theoretical or probabilistic statistic *E*, introduced here for the first time and upon which *NSE* is based, is, in contrast to *NSE*, entirely independent of the properties of data used in its estimation. This is a central focus of this paper.

1.1. Model Simulation and Calibration

Consider the problem of evaluating the goodness of fit of watershed simulation model output *S*, to observations *O*. Let S_t and O_t represent the simulated and observed daily streamflow on day *t*, $t = 1, \ldots, n$, at the outlet of a watershed. A conceptual simulation model $H[X, \Omega]$ is envisioned, which converts a suite of model parameters Ω and inputs X_t, such as rainfall, potential evapotranspiration, and temperature into simulated watershed responses. The observations O_t, can be interpreted as random realizations from an unknown probability distribution (pd) $f[\ ]$ so that

$$O_t \xrightarrow{d} f[H[X_t, \Omega], \varepsilon] \quad (1)$$

where $\xrightarrow{d}$ denotes the convergence to a pd and ε denotes the errors introduced by model uncertainty as well as by measurement errors in both the inputs X and the observations O.

Model calibration attempts to adjust the model parameter set Ω to obtain model parameter estimates $\widehat{\Omega}$ which ensure that the fitted simulation output

$$S_t = H\left[X_t, \widehat{\Omega}\right] \quad (2)$$

resembles important features of the observations O_t Todini and Biondi (2017) point out that the goodness of fit obtained from such calibrations will generally be better than the goodness of fit associated with the true parameter set Ω if it exists, due to structural model errors and interactions between the model parameters, predictions and errors.

Once a deterministic model is fit to data, hydrologists often use scatterplots to compare the observations O to the simulations S so that for the calibration sequence

$$O = S + \varepsilon \quad (3)$$

such scatterplots of O versus S are a graphical illustration of a realization of the joint pd $f(O,S)$. Two important conditional probability density functions (pds) arise from this joint distribution, namely, (1) the distribution of model predictions given by $f(O|\,S) = f(O,\ S))/f(S)$ and (2) the distribution of model simulations given by $f(S|\ O) = f(O,\ S))/f(O)$. While a representation of prediction uncertainty via $f(O|\,S)$ is central to Bayesian decision approaches efforts, a representation of simulation uncertainty, given by $f(S|\,O)$ is paramount to efforts to improve model performance by improving model parameter estimates and model structure. See Todini (2011, 2017) for a review of these issues.

1.2. Literature Review on NSE

Numerous statistics have been introduced for calibration, hypothesis testing, and goodness-of-fit evaluations of hydrologic models. It is now commonplace and perhaps essential (Reusser et al., 2009) to use multiple goodness of fit measures to calibrate simulation models as evidenced in reviews by Moriasi et al. (2007) and Efstratiadis and Koutsoyiannis (2010). Koppa et al. (2019) review studies, which document improvements resulting from the multiobjective calibration of rainfall-runoff models. Although multiple measures of goodness of fit are generally recommended and applied in practice, we only concentrate on the *NSE* (Nash & Sutcliffe, 1970) and the Kling-Gupta efficiency (*KGE′*) (Gupta et al., 2009) indices. Evidence of the widespread usage of *NSE* is provided by over 19,680 Google Scholar citations (19 August 2020) to Nash and Sutcliffe (1970), as well as recent discussions by Moriasi et al. (2007), Gupta et al. (2009), Ewen (2011), Guinot et al. (2011), Pushpalatha et al. (2012), Todini and Biondi (2017), and many others. For example, Todini and Biondi (2017) report that *NSE* "is by far the most utilized index in hydrological applications." In an effort to provide overall recommendations for model evaluation techniques both ASCE (1993) and Moriasi et al. (2007) recommended the use of *NSE* over numerous other alternative goodness-of-fit metrics.

Over the years, numerous authors have evaluated the behavior of *NSE* (see, e.g., Bardsley, 2013; Gupta & Kling, 2011; Gupta et al., 2009; Jain & Sudheer, 2008; Legates & McCabe, 1999; Liu et al., 2018; Martinec & Rango, 1989; McCuen et al., 2006; Pool et al., 2018; Schaefli & Gupta, 2007, and many others). Ritter and Munoz-Carpena (2013) provide a very thorough review of literature on *NSE*. Several modifications of *NSE* have been proposed, such as a bounded version (Mathevet et al., 2006), a nonparametric estimator introduced by Pool et al. (2018), a version for volumetric efficiency (Criss & Winston, 2008), an index related to both *NSE* and ρ (Bardsley, 2013), a slight variant of *NSE* termed the coefficient of gain (Martinec & Rango, 1989; World Meteorological Organization [WMO], 1986), and others (Krause et al., 2005). Pushpalatha et al. (2012) suggest improved evaluation of goodness of fit associated with low streamflow by computing *NSE* between $1/O$ to $1/S$. See Ritter and Munoz-Carpena (2013) for a review of other transformations that have been advanced to improve the performance of *NSE*. Clark et al. (2008, Figure 4) and

Newman et al. (2015, Figures 10 and 11) document the considerable impact that a small percentage of the observations can exert over the total contribution to mean square error (MSE), a component of *NSE*.

1.3. Confusion Between Probability, Statistics, and NSE

Unlike previous research on *NSE*, we distinguish between the theoretical efficiency, which we term *E,* and one estimator of that statistic; *NSE*. The theoretical statistic *E*, is a standardized form of the *MSE*:

$$MSE = E\left[(S-O)^2\right] \tag{4a}$$

and

$$E = 1 - \frac{MSE}{E\left[(O-\mu_o)^2\right]} = 1 - \frac{MSE}{\sigma_o^2} \tag{4b}$$

where $E[\ ]$ denotes the expectation operator, S and O represent the simulated and observed time series, respectively, and μ_o and σ_o^2 denote the true mean and variance of the observations. Both *MSE* and *E* are defined by the expectation operators $E[\ ]$ in (4a) and (4b), which are grounded in the theory of probability as distinguished from the theory of statistics that would involve developing formulas to estimate *E* from data, the topic of this paper.

It is only after data are introduced, s_i and o_i, $i = 1, \ldots n$, that one needs to replace the expectation operator with estimators of *MSE* and *E* in (4a) and (4b) so that

$$\overline{MSE} = \frac{1}{n}\sum_{i=1}^{n}(s_i - o_i)^2 \tag{5a}$$

$$NSE = 1 - \frac{\overline{MSE}}{s_o^2} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(s_i - o_i)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(o_i - \overline{o})^2} \tag{5b}$$

where *NSE* is the popular estimator introduced by Nash and Sutcliffe (1970). It is common practice to use upper and lower case values ("*O*" vs. "*o*") to denote the theoretical values and their realizations, respectively. It is also common practice to use Greek for the theoretical, population, or probabilistic mean μ_o and variance σ_o^2 and to use either hats over the Greek values $\widehat{\mu}_o$ and $\widehat{\sigma}_o^2$, or symbols such as $\overline{o}$ and s_o^2, to denote sample estimates of those same statistics based on data.

Since previous literature failed to distinguish between the theoretical statistic *E* and its sample estimator *NSE*, previous criticisms of *NSE* have been misinterpreted as a drawback of the population *E*. This is analogous to criticizing the true mean μ_x because the sample mean $\overline{x}$ is sensitive to outliers. This confusion has important consequences and forms the basis of our contribution, because it has led some investigators to suggest that *E* has flaws, when in fact it is only the particular estimator *NSE* that raises concerns. A theoretical treatment of *MSE* and *E* should not depend on actual data because the theoretical properties of the metrics *MSE* and *E* defined in (4a) and (4b) do not depend on data or its properties. Since previous studies have failed to distinguish between the theoretical statistic *E* and one sample estimator *NSE*, those studies have confused the subjects of probability and statistics.

There is a growing literature which has sought to develop methods for constructing hypothesis tests and confidence intervals concerning the true value of efficiency *E*. Examples of such studies include those by McCuen et al. (2006), Ritter and Munoz-Carpena (2013), Libera et al. (2018), and many others reviewed by Liu et al. (2018), all of which have sought to improve our understanding of the sampling properties of *NSE*. This is a very exciting and promising line of research, which could benefit significantly from our results for two reasons. First, it is difficult to develop and evaluate a hypothesis test or confidence interval for the true value of efficiency, without a theoretical definition of *E*. Thus, previous studies that introduced confidence intervals or hypothesis tests concerning the true value of *E* were unable to rigorously evaluate the likelihood of random confidence intervals covering the true value of *E* or the likelihood of Type I or II errors, because those studies did not have knowledge of the true value of *E*. Second, we introduce a new

estimator of E that has considerable advantages over NSE for highly skewed periodic bivariate daily hydrologic series.

1.4. The Influence of Periodicity, Variability, and Skewness

Daily streamflows exhibit deterministic periodic behavior which confounds our ability to estimate reliable summary statistics such as E, μ_o, or σ_o in (4b). Such deterministic seasonal behavior implies that the daily streamflows are not identically distributed but instead exhibit statistics which vary in a deterministic fashion from one season to another. We introduce a monthly mixture model to account for periodic behavior and demonstrate that this approach leads to considerable improvements in our ability to estimate theoretical efficiency E in (4b).

Ever since the legendary paper "Just a Moment" (Wallis et al., 1974), we know that nonnormality and skewness of observations induces both bias and increased variability in product moment estimators of statistics such as standard deviation σ_o, coefficient of variation $C_o = \sigma_o/\mu_o$, and skewness γ_o. Daily streamflows exhibit a very high level of positive skewness γ_o, yet ironically, moment estimators of γ_o only exhibit low bias and variance when the random variable exhibits no skewness, as shown by Wallis et al. (1974). Vogel and Fennessey (1993) further show that sample estimates of C_o and γ_o are highly downward biased and variable even when computed from tens of thousands of daily streamflow observations, with that bias increasing as the skewness of the observations increases. Similarly, Barber et al. (2019) document considerable downward bias and increased variability in estimates of Pearson correlations, due to skewness of the observations.

The coefficient of variation of the observations, $C_o = \sigma_o/\mu_o$, can be used as a surrogate of γ_o for positively skewed observations, because, for example, for Gamma and LN2 variables, γ_o is related to C_o via the relations $\gamma_o = 2C_o$ and $\gamma_o = C_o^3 + 3C_o$, respectively. In our development of suitable estimators of E, μ_o, or σ_o in (4b) we develop improved estimators of C_o that account for both periodicity and skewness and thus are not subject to the bias and variability associated with product moment estimators. In the remainder of this study we use estimates of C_o as a surrogate of skewness.

1.5. Study Goals

This is one of the first studies to draw a distinction between the theoretical efficiency E in (4b) and its common sample estimator NSE in (5b). Previous studies used actual streamflow observations and hydrologic model simulations to evaluate the performance of NSE and its variants, yet such studies can never report definitively on the performance of NSE because the true value of NSE, which we denote as E, is always unknown in such situations. Our approach of distinguishing between E and NSE enables implementation of controlled Monte Carlo experiments to rigorously evaluate the performance of various alternative estimators of efficiency E. Finally, we test our findings and recommendations using the output of hundreds of calibrated U.S. Geological Survey (USGS) Precipitation Runoff Modeling System (PRMS) rainfall-runoff models (Markstrom et al., 2015) analogous to the recent work of Farmer and Vogel (2016a) and Barber et al. (2019).

2. Theoretical Development of Efficiency E

Here we introduce theoretical expressions for efficiency that are based on the widely used Nash-Sutcliffe and Kling-Gupta definitions of efficiency. We also perform probabilistic analysis of these two statistics, which enable us to arrive at numerous conclusions without resorting to the use of any data, or sample statistics, whatsoever. Again, our approach is unique because all previous analyses, discussions, and criticisms of these two statistics in the literature were made using data and estimators without resorting to a probabilistic analysis as is performed here.

2.1. Theoretical Efficiency Based on Nash-Sutcliffe

The definition of MSE in (4a) is based on the bivariate relationship between S and O, which can also be expressed in terms of the univariate model residual ε defined in (1) so that

$$E\left[(S-O)^2\right] = E\left[\varepsilon^2\right] = MSE[\varepsilon] \tag{6}$$

where $MSE[\varepsilon]$ is referred to as the MSE of the model residuals ε. It is easily shown that $MSE[\varepsilon]$ is the sum of the bias squared and variance

$$MSE[\varepsilon] = E[\varepsilon - E(\varepsilon)]^2 + E\left[(\varepsilon - E[\varepsilon])^2\right] = Bias(\varepsilon)^2 + Var(\varepsilon) \tag{7}$$

so that degradation in the goodness of fit of a model results from any increase in bias and/or variance of the error term. Both E and MSE are impacted by bias and variance in ε, and it is that unique feature that distinguishes them from some other metrics such as the correlation coefficient ρ, which is not influenced by bias in ε.

Using the theory of probability, one can expand the expectation in (6) to obtain

$$MSE[\varepsilon] = (\mu_o - \mu_s)^2 + \sigma_o^2 + \sigma_s^2 - 2\sigma_s\sigma_o\rho \tag{8}$$

where μ_0 and $\mu_s$denote the means of O and S, respectively, σ_0^2 and σ_S^2 denote the variances of O and S, respectively, and ρdenotes the Pearson correlation between O and S, respectively. The expansion in (8) is also given by Murphy (1988, see Equation 10) and Gupta et al. (2009) using sample estimators of the various terms, instead of their population values.

Our central goal is to develop improved estimators of E, which are generally preferred (for skewed and periodic hydrologic data), to the commonly used NSE estimator given in (5b) as well as the Kling-Gupta efficiency estimator (KGE') introduced by Gupta et al. (2009) and the nonparametric estimator of E introduced by Pool et al. (2018). Analogous to Gupta et al. (2009), we rewrite $MSE[\varepsilon]$ in (8) as

$$MSE[\varepsilon] = \Delta^2\mu_o^2 + \sigma_o^2\left[1 + \alpha^2 - 2\alpha\rho\right] \tag{9}$$

where $\Delta = \dfrac{\mu_o - \mu_s}{\mu_o}$, and $\alpha = \dfrac{\sigma_s}{\sigma_o}$.

Here Δ is bias as a fraction of the mean of the observations, and α is the ratio of the standard deviation of the simulated response to the standard deviation of the observations. The primary difference between our treatment in (9) and Gupta et al. (2009), Pool et al. (2018), and others is that they employ sample estimates of the various terms in (9) without referring to their true values.

We have chosen to introduce the bias as a fraction of the mean observations Δ in (9) because this form of standardized bias is so easy to compare and contrast across models or watersheds and is consistent with the traditional (statistical) definition of bias given in (7), unlike the nonstandard bias term $\beta = \mu_s/\mu_o$ introduced by Gupta et al. (2009). We note that Δ is related to $\beta = \mu_s/\mu_o$ so that $\Delta = 1 - \beta$.

Combining (9) and (4b) leads to

$$E = 2\alpha\rho - \alpha^2 - \frac{\Delta^2}{C_o^2} \tag{10}$$

where $C_0 = \sigma_o/\mu_o$ is the coefficient of variation of the observations and again $\Delta = (\mu_o - \mu_s)/\mu_o$ and $\alpha = \sigma_s/\sigma_o$.

2.2. Another Definition of Theoretical Efficiency E' Based on Kling-Gupta

Although they did not distinguish between the theoretical value of efficiency and its sample estimator, we infer that Gupta et al. (2009) introduced a new definition of efficiency, which is based on a different (nonequivalent) form of the expression for E given in (4b) and (10). Their implied definition of theoretical efficiency E' is

$$E' = 1 - \sqrt{w_\beta(\beta - 1)^2 + w_\alpha(\alpha - 1)^2 + w_\rho(\rho - 1)^2} \tag{11}$$

where $\beta = 1 - \Delta = 1 - [(\mu_o - \mu_s)/\mu_o]$ with ρ and $\alpha = \sigma_s/\sigma_o$ defined previously and the weights $w_\beta = w_\alpha = w_\rho = 1$. The statistic E' was formulated as a measure of the Euclidian distance between the three-dimensional Pareto frontier and an ideal solution corresponding to $\alpha = 1$, $\rho = 1$ and zero bias ($\beta = 1$, $\Delta = 0$).

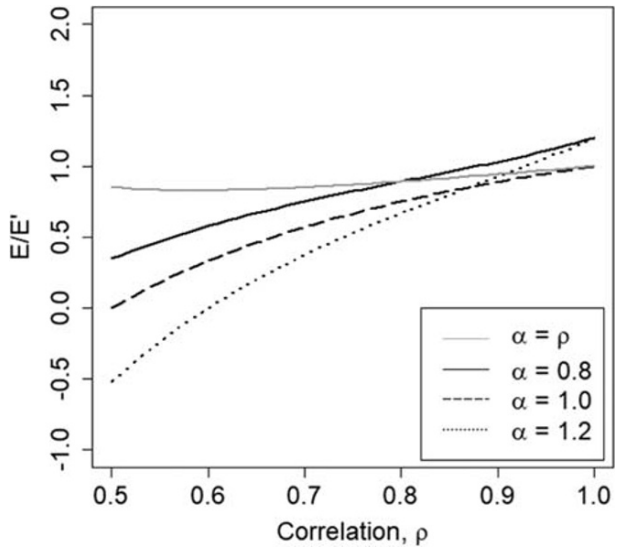

Figure 1. Ratio of efficiency E based on *NSE* to the efficiency E' based on *KGE'* as a function of α and ρ for an unbiased model.

The theoretical efficiency E' in (11) introduced in its empirical form by Gupta et al. (2009) could have several advantages over the definition of E in (10) as is shown later on and discussed by Gupta et al. (2009); Knoben et al. (2019) and others. Here, similar to the work of Knoben et al. (2019), we document the markedly different behavior of E and E'. The distinction between our approach and the approach taken by Knoben et al. (2019) is that we derive general analytical expressions to distinguish between E and E', whereas they employed Monte Carlo experiments to distinguish between the properties of *NSE* and *KGE* for particular data sets.

To better understand the differences in the behavior of E and E', in (10) and (11) respectively, one can derive their ratio for an unbiased model ($\Delta = 0$) as

$$\frac{E}{E'} = \frac{\alpha^2 - 2\rho\alpha}{\sqrt{(\rho-1)^2 + (\alpha-1)^2} - 1} \tag{12}$$

Figure 1 compares the ratio of E and E', as a function of both α and ρ. In general, the two theoretical statistics E and E' are only very roughly equal when $\alpha = \rho$, which would be the case for simple linear regression or when the simulations are independent of the model residuals. Otherwise, for more realistic and complicated models with $\alpha \neq \rho$, the values of E and E' can be expected to differ and quite significantly so. Surely, in any rigorous and objective evaluation of the behavior of sample estimators of E and E' (such as the evaluations of *NSE* and *KGE* reported later on), one must account for the important and marked differences in their theoretical values reported in Figure 1.

The definition of efficiency E' in (11) is not a measure of standardized *MSE* as is the case for E. One can show that $\sqrt{1-E} = 1 - E'$ when the weights in (11) are defined as $w_\beta = 1/C_o^2$, $w_\alpha = 1$, and $w_\beta = 2\alpha/(1-\rho)$. Thus, it is only under the very unlikely circumstances that $C_o = 1$ and $2\alpha = 1 - \rho$ in which case E and E' can be expected to yield similar behavior.

2.3. Limiting Behavior of E and E' for Unbiased Models

In this section we make no additional assumptions, other than that the simulations are generated without adding error as in (2) and that during calibration (3) holds, so that $O = S + \varepsilon$ and $\rho > 0$. Under those conditions $\sigma_o^2 = \sigma_s^2 + \sigma_\varepsilon^2 + 2\rho_{s,\varepsilon}\sigma_s\sigma_\varepsilon$ and since $\varepsilon = O - S$, $\sigma_\varepsilon^2 = \sigma_o^2 + \sigma_s^2 + 2\rho\sigma_o\sigma_s$. Combining these expressions leads to an expression for the correlation between S and ε resulting in $\rho_{s,\varepsilon} = (\rho - \alpha)/\sqrt{1 + \alpha^2 - 2\rho\alpha}$. Thus, if the calibration is performed in such a manner as to ensure an unbiased model with errors ε, which are independent of the simulations S, (so that) then $\alpha = \rho$ in which case the efficiency in (10) reduces to $E = \rho^2$ and the efficiency in (11) reduces to $E' = 1 + \sqrt{2} - \rho\sqrt{2} \approx 2.41 - 1.41\rho$.

The statement that $E = \rho^2$ for an unbiased model with model residuals, which are independent of the simulations is a more general and correct statement than the conclusion reached by McCuen et al. (2006, Equation 3), which stated that $E = \rho^2$ for a linear model. It is possible for a linear model without an intercept to exhibit bias, and it is also possible for any linear model to exhibit nonzero $\rho_{s,\varepsilon}$; thus, it is possible for a linear model, with or without an intercept, to exhibit an $E \neq \rho^2$. This is also a different and more general interpretation of the conditions under which $E = \rho^2$ than is given by either Gupta et al. (2009) or Gupta and Kling (2011). Note also that this is also a different result from the incorrect result $E = 1 - (1/\rho^2)$ given by Bardsley (2013) for an unbiased model.

3. Study Assumptions: Daily Streamflow Simulations and Observations

A fundamental challenge in stochastic hydrology is that we do not know the true distribution from which our data arises. Without this knowledge it is difficult to make general recommendations concerning estimators of hydrologic statistics such as *NSE*, *KGE*, or the 100-year flood. Comparisons based on actual hydrologic

data alone can only provide anecdotal evidence because we do not know the correct answer, and the historical record is only one realization of a random process that will never be repeated. Conversely, considering only the properties of theoretical pds, either analytically or with Monte Carlo analysis, is likely to be of little interest to hydrologists because the assumed distributions may not reflect the properties of our data. Thus, a proper analysis of a hydrologic statistic must be grounded in a theoretical model of the natural process with known properties but should also validate that model's ability to approximate important properties of the (true) natural process. Through such an analysis, it is possible to make defensible claims about a hydrologic statistic, which are useful to hydrologists.

Since daily streamflows are neither normally nor identically distributed, it is necessary to introduce a theoretical model, which can accommodate the high degree of skewness and periodicity inherent in observed streamflow sequences. In the following sections we document how a bivariate lognormal monthly mixture model can provide a good approximation to daily streamflow observations and simulations at hundreds of watersheds in the United States. That model is used in subsequent sections to perform controlled Monte Carlo experiments, and as the foundation of improved efficiency statistics.

3.1. A Simple Model of Streamflow Observations and Simulations

Daily, hourly, and subhourly streamflow are known to exhibit extremely high values of skewness, so that typical observations O, and simulations S, are much more closely approximated by a bivariate three-parameter lognormal (*BLN3*) model, than a bivariate normal model as was shown by Barber et al. (2019) and others. Barber et al. (2019), Blum et al. (2017), and Limbrunner et al. (2000, Figure 6) used L-moment diagrams to illustrate that two- and three-parameter lognormal distributions (*LN2* and *LN3*, respectively) provide a good approximation to the pd of daily streamflow observations for hundreds of stations across the conterminous United States. Barber et al. (2019) also document that the *BLN3* model is equivalent to a Gaussian copula with a LN3 marginal pd.

However, in comparisons of the behavior of the Pearson correlation coefficient estimator r, from synthetic *BLN3* samples versus actual daily flow series, Barber et al. (2019, Figure 5) found that synthetic *BLN3* series could not reproduce the behavior of r estimated from actual O and S series. To address this issue, we introduce a *BLN3* monthly mixture model, a necessary and major innovation over Barber et al. (2019) that accounts for the skewness and periodicity of the streamflow series.

3.2. A BLN3 Monthly Mixture (BLN3-MM) Model of the PD of Daily Streamflow

Baldwin and Lall (1999) and many others have shown that annual and intra-annual seasonal variations in streamflow can lead to complex bimodal pds. An LN3 monthly mixture model is needed to account for the strong deterministic periodicity within the daily flow series which can lead to bimodal and other far more complex pd shapes than a single LN3 model could mimic. We introduce a bivariate *LN3* monthly mixture model (denoted *BLN3-MM*) which involves fitting a separate *BLN3* model to the daily streamflows in each month. We employ a $12 \times 4 = 48$ parameter *BLN3-MM* model to generate synthetic streamflow series, which better mimic the marginal distribution of the observations and simulations than a single four-parameter *BLN3* model.

Consider a monthly mixture distribution which consists of a separate LN3 pd $f(o;\mu_i,\sigma_i,\tau_i)$ in each of $i = 1, \ldots 12$, months where o denotes the daily streamflow observations within month i, τ_i denotes the lower bound of the fitted *LN3* distribution in month i and μ_i and σ_i denote the mean and standard deviation of the transformed streamflows, $u = \ln(o - \tau_i)$, in month i. The resulting mixture pd of all the daily streamflows is given by

$$f(o;\mu_1,\ldots,\mu_{12},\sigma_1,\ldots,\sigma_{12},\tau_1,\ldots,\tau_{12}) = \sum_{i=1}^{12} w_i f_i(o;\mu_i,\sigma_i,\tau_i) \tag{13}$$

where $f()$ denotes the overall pd of the observations and $f_i()$ denotes the pd of the observations in month i, and $\sum_{i=1}^{12} w_i = 1$. Assuming each month has the same number of days, and that daily observations exist on every day, we employ the fixed mixture weights, $w = w_i = 1/12$; however, future work may benefit from

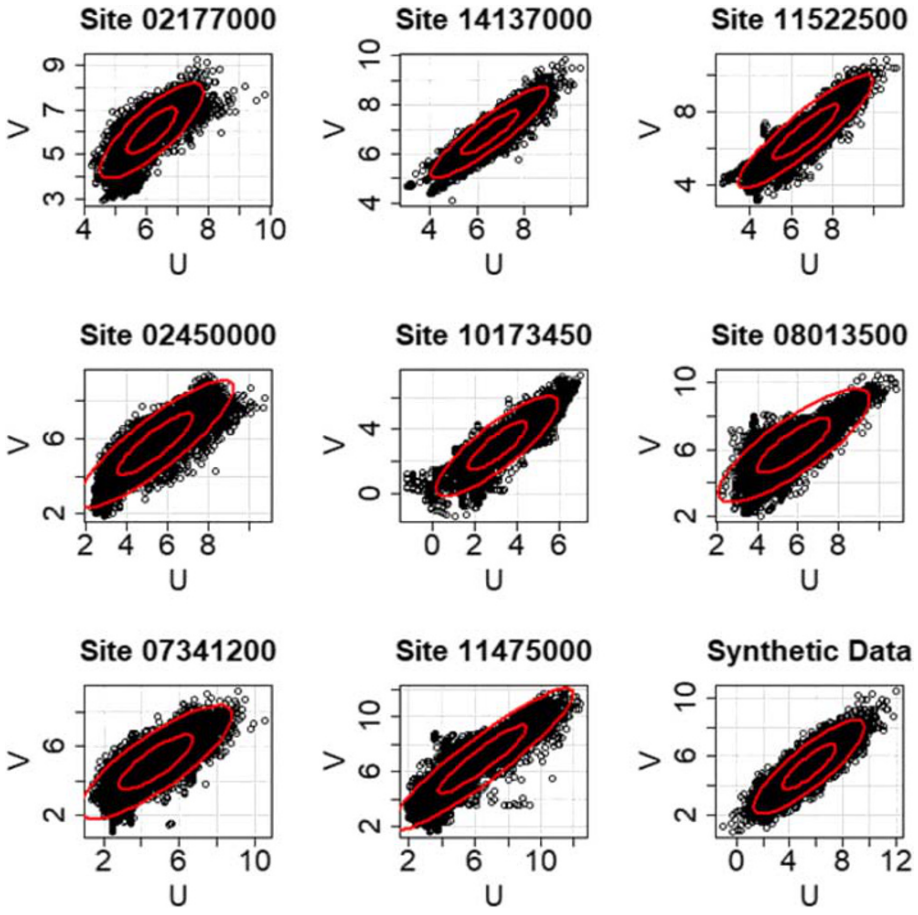

Figure 2. Scatter plot of values of $u_i = \ln[o_i - \widehat{\tau}_O]$ versus $v_i = \ln[s_i - \widehat{\tau}_S]$ for eight sites summarized in Table 3.2 of Barber (2020) along with results for synthetic data on the lower right.

use of unequal weights based on maximum likelihood or Bayesian estimators of the mixture weights as recommended by McLachlan et al. (2019).

3.3. Daily Streamflow Observations and Simulations

Here, as in Farmer and Vogel (2016a) and Barber et al. (2019), a moderately complex, distributed-parameter, precipitation-runoff model is used to generate bivariate daily streamflow traces from daily streamflow observations at 1,225 river locations across the continental United States. The distributed model, in this case PRMS (Markstrom et al., 2015), was calibrated at each of 1,225 perennial river basins across the conterminous United States. Details and availability of the data sets are described by Farmer and Vogel (2016b). The particulars of the model and the calibration scheme are not relevant to our experiments.

In addition to numerous assessments relating to reproduction of the water balance and various aspects of hydrograph behavior, an experienced hydrologist would normally examine scatterplots of the observations, o versus the simulations s. We examined scatterplots of the logs of the transformed simulations, $v = \ln(s - \widehat{\tau}_s)$, versus the logs of the transformed observations, $u = \ln(o - \widehat{\tau}_o)$, at each site to ensure that they mimic the behavior of reasonable models. The parameters $\widehat{\tau}_o$ and $\widehat{\tau}_s$ reflect that the lower bound for observed and simulated flows at many sites is greater than 0 and are estimated using Stedinger's (1980) lower bound estimator of an LN3 distribution (see Equations 18a and 18b). One expects an approximately ellipsoidal relationship between u and v, which would be consistent with the assumption of a *BLN3* relationship between o and s.

Figure 2 plots u vs v for eight watersheds, which represent a range of the type of results we observed. Figure 2 includes a few sites for which there are not nice elliptical relationships between U and V, meaning the *BLN3*

Table 1
Values of n, Co, Cs, α, Δ, ρ, LBE_m, LBE'_m, NSE, and KGE Corresponding to Daily Streamflow Observations and Simulations at 447 USGS Gaged Watersheds

			IQR	Range
Property	Average	Median	(25th, 75th)	(min, max)
n	10,944	10,957	(10,957, 10,957)	(10,014, 11,322)
Co	1.54	1.42	(1.32, 1.73)	(0.51, 6.30)
Cs	1.39	1.24	(1.04, 1.54)	(0.45, 6.56)
α	0.94	0.92	(0.79, 1.09)	(0.50, 1.50)
Δ	−0.03	−0.03	(−0.08, 0.02)	(−0.31, 0.30)
ρ	0.7	0.7	(0.64, 0.77)	(0.50, 0.91)
LBE_m	0.4	0.42	(0.31, 0.53)	(−0.74, 0.80)
NSE	0.52	0.52	(0.39, 0.68)	(−0.09, 0.86)
LBE'_m	0.63	0.64	(0.55, 0.71)	(0.24, 0.87)
KGE	0.62	0.65	(0.51, 0.76)	(−0.02, 0.90)

assumption is more tenuous, because those cases illustrate what can happen in some unusual situations. Also shown in Figure 2 are two-dimensional confidence intervals, known as "probability ellipses," drawn to enclose 50% and 90% of the values of U and V, if they arose from a *BLN2* model (see Barber et al., 2019). In the lower right-hand corner of Figure 2 we include for comparison, a scatterplot of synthetic series generated from the *BLN3-MM* model. We expect these probability ellipses to give only a very rough approximation to the relationship between U and V because a *BLN3-MM* model will NOT yield *LN3* marginal distributions for S and O as does a *BLN3* model.

Removing those sites that led to spurious and highly nonellipsoidal relationships between u and v left a total of 905 sites. To ensure enough streamflow data to reliably estimate sample statistics to inform our Monte Carlo experiments, we also dropped sites with record lengths less than 10,000 days leaving 673 watersheds. Finally, to ensure that we only consider plausible simulation results, we dropped sites that led to estimates of bias $\Delta = (\mu_o - \mu_s)/\mu_o$ and $\alpha = \sigma_s/\sigma_o$ outside the ranges of [−0.33, 0.33] and [0.5, 1.5], respectively, leading to a total of 447 sites used in the following analyses. Table 1 summarizes the range, interquartile range and median values of sample size n as well as estimates of the coefficient of variation of the observations C_o and simulations C_s, Δ, ρ, α, across the 447 sites. Barber (2020, Table 3.2) reports those statistics for the eight highlighted sites. Although all of these statistics were estimated from observations, we do not use hats to denote estimated values, because all of these values are considered to be the true values in our subsequent Monte Carlo experiments. Also shown in Table 1 are some of the estimators of efficiency described below. Estimators of all the statistics in Table 1 are based on the *BLN3-MM*, which provides more reliable estimates of all of these statistics than alternative methods, as shown below.

3.4. Evaluation of the BLN3 Monthly Mixture (*BLN3-MM*) Model

Figure 2 provides evidence of the goodness of fit of the *BLN3-MM* model at eight sites. To evaluate the goodness of fit of the *BLN3-MM* model across all 447 sites, we use the well-known probability plot correlation coefficient (PPCC) statistic. We employ probability-probability (pp) probability plots, which involve plotting the empirical cumulative probability of the observations versus an estimate of those cumulative probabilities for the fitted mixture model. The mixture cumulative distribution function is obtained by integration of (13) which leads to

$$F(o;\mu_1,...,\mu_{12},\sigma_1,...,\sigma_{12},\tau_1,...,\tau_{12}) = \frac{1}{12}\sum_{i=1}^{12} F_i(o;\mu_i,\sigma_i,\tau_i) \tag{14}$$

where $F()$ denotes the overall cumulative pd of the observations o and $F_i()$ denotes the cumulative pd of the observations in month i.

Suppose we have a total of $j = 1,2,\ldots,n$ observations denoted $o_{i,j}$ where i denotes which month each flow occurs in. The transformed observations are denoted as $u_{i,j} = \ln(o_{i,j} - \widehat{\tau}_i)$ for $i = 1,2,...,12$ and $j = 1,2,\ldots,n$. For the *BLN3-MM* model a pp probability plot is constructed by first ranking all the transformed observations denoted $u_{i,(j)}$ where the subscript parenthesis is standard notation for ranked variables. Under the *BLN3* hypothesis, O follows an *LN3* distribution and $U_{i,j} = \ln(O_{i,k} - \tau_i)$ follows a normal distribution, in each month i, and a pp probability plot is obtained by plotting a Weibull plotting position estimate of the cumulative probabilities $p_j = j/(n+1)$ versus an estimate of the cumulative probability of the fitted *BLN3-MM* distribution obtained from

$$\widehat{F}(u_{i,(j)}) = \frac{1}{12}\sum_{i=1}^{12} \Phi\left(\frac{u_{i,(j)} - \overline{u}_i}{s_{u,i}}\right) \tag{15}$$

where $\Phi()$ denotes the cumulative distribution function of a normal variable and $\overline{u}_i$ and $s_{u,i}$ are the mean and standard deviation of the transformed flows $u_{i,(j)} = \ln(o_{i,(j)} - \widehat{\tau}_i)$ in month i. A Weibull plotting position is suitable here, because it yields an unbiased estimate of the cumulative probability associated with

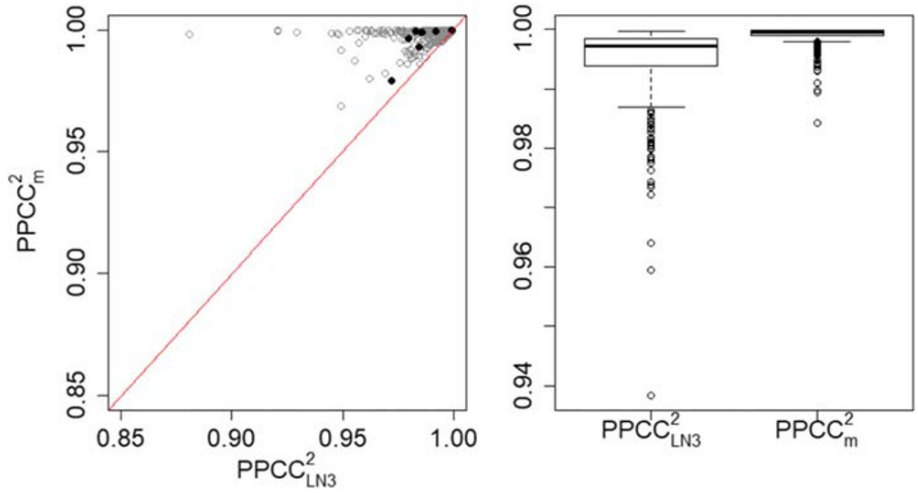

Figure 3. The square of the probability plot correlation coefficients for the BLN3 monthly mixture model $PPCC^2_m$ and a single LN3 model $PPCC^2_{LN3}$. The solid circles denote the eight sites summarized in Figure 2 and in Figure 3.2 of Barber (2020).

the observations, regardless of their pd. The *PPCC* is then obtained by computing the correlation between the n values of the plotting positions p_j and $\widehat{F}(u_{i,(j)})$ obtained from (15). Figure 3 uses boxplots and a scatterplot to summarize the square of the *PPCC* values associated with the *BLN3-MM* model (denoted $PPCC^2_m$) versus the $PPCC^2$ value of fitting a single *LN3* model (denoted $PPCC_{LN3}$) to the entire n day series at the 447 sites. Figure 3 documents the considerable improvement in the goodness of fit of the 48-parameter *BLN3-MM* model over the four-parameter *BLN3* model used by Barber et al. (2019), which is expected given the 44 additional parameters associated with the *BLN3-MM* model.

4. Sample Estimators of Efficiency

In this section we summarize estimators of E and E', which have been introduced by others and improved estimators derived by us. In section 5, these estimators are compared and evaluated using Monte Carlo experiments.

4.1. NSE

This estimator of theoretical efficiency E defined in (4b) and (10) was first introduced by Nash and Sutcliffe (1970) and is given in Equation 5b.

4.2. Kling-Gupta Efficiency (*KGE′*)

Gupta et al. (2009) developed an estimator of E' defined in (11), which is based on a different (nonequivalent) form of the expression for E given in (4b) and (10). Using standard statistical notation, where hats over variables denote estimates of that variable, their estimator now widely referred to as the *Kling-Gupta* estimator, takes the form:

$$KGE' = 1 - \sqrt{(\hat{\beta}-1)^2 + (\hat{\alpha}-1)^2 + (\hat{\rho}-1)^2} \tag{16}$$

where $\widehat{\rho} = r = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(s_i-\bar{s})(o_i-\bar{o})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i-\bar{o})^2\frac{1}{n}\sum_{i=1}^{n}(s_i-\bar{s})^2}}$, $\hat{\alpha} = \dfrac{\hat{\sigma}_s}{\hat{\sigma}_o} = \dfrac{s_s}{s_o}$ and $\hat{\beta} = 1 - \hat{\Delta} = 1 - \dfrac{\hat{\mu}_o - \hat{\mu}_s}{\hat{\mu}_o} = \dfrac{\bar{s}}{\bar{o}}$.

with $\bar{o} = \frac{1}{n}\sum_{i=1}^{n} o_i$, $\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s_i$. $s_o = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(o_i-\bar{o})^2}$ and $s_s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(s_i-\bar{s})^2}$.

In Equation 16 we employ the notation *KGE′* instead of the usual notation *KGE*, to highlight that *KGE′* is an estimator of E' in (11) and is NOT necessarily a good estimator of E in (10) as shown in Figure 1.

On the one hand, there are several important advantages of the theoretical statistic E' outlined by Gupta et al. (2009). However, there are several potential concerns with *KGE′*, apart from the fact that it is based on a different theoretical definition of efficiency than the value of E introduced in (4b) and (10). One obvious problem with *KGE′* is that it is based entirely on product moment estimators for all the components given in (11). This would not be a problem if applications were for bivariate normally distributed data; however, for skewed hydrologic data such as daily streamflow, Vogel and Fennessey (1993) show that ratios of product moment estimators exhibit enormous bias, even for extremely large sample sizes in the tens of thousands and should generally be avoided. Thus, the estimators $\hat{\alpha}$, r and $\hat{\beta}$ in (16) will exhibit considerable bias and variability, even for very large samples, because they are ratio estimators based on product moments of skewed observations. The estimator $\hat{\rho} = r$ is the well-known Pearson correlation coefficient (Pearson, 1896), which performs well for bivariate normal observations, but was shown by Barber et al. (2019) to perform poorly for *BLN3* series because it exhibits considerable upward bias and extreme variability compared to the *BLN3* and nonparametric estimators of ρ they introduced.

4.3. Nonparametric Efficiency (*PVSE′*)

To address the concerns raised above for *KGE′*, Pool et al. (2018) developed an estimator of E' in (11), which we term the Pool-Vis-Seibert estimator (*PVSE′*), which employs nonparametric estimators for each of the three components given by

$$PVSE' = 1 - \sqrt{(\hat{\beta} - 1)^2 + (\hat{\alpha} - 1)^2 + (\hat{\rho} - 1)^2} \tag{17}$$

where again $\hat{\beta} = 1 - \hat{\Delta} = \bar{s}/\bar{o}$, $\hat{\rho}$ is estimated using the nonparametric Spearman's correlation coefficient, which is obtained by applying Pearson product moment estimator $\hat{\rho} = r$ given in (16) to the ranks of the observations and simulations and $\hat{\alpha} = 1 - \frac{1}{2}\sum_{k=1}^{n}\left|\frac{s_{(k)}}{n\bar{s}} - \frac{o_{(k)}}{n\bar{o}}\right|$. Here $s_{(k)}$ and $o_{(k)}$ denote the ordered values of the simulations and observations, respectively. While the general idea behind Pool et al. (2018) to employ nonparametric estimators of the components of E' is a good one, we note that Spearman's correlation is an estimator of a different theoretical correlation coefficient than the correlation ρ in (10) and (11) (see Barber et al., 2019) and their nonparametric estimator $\hat{\alpha}$ is an estimator of a different theoretical statistic than α defined in (10) and (11).

4.4. BLN3 Estimators of Efficiency: (LBE and LBE′)

Here we derive improved estimators of E and E' using estimators of each of the components ρ, α, Δ, and C_o of the definitions of E and E' in (10) and (11), respectively, which are suited to highly skewed streamflow observations and simulations. The derivation of our improved estimators of E and E' rely on the assumption that the S and O series follow a *BLN3* model, which was tested by Barber et al. (2019) and section 5 of this study. This choice allows for an analytical (closed-form) derivation of improved estimators of E and E', and this assumption is also rather general and well suited for the skewed hydrologic variables considered in this study.

Given the *BLN3* assumption, we use what has proven to be an extremely effective estimator of the lower bound of an *LN3* model given in Equation 10 of Stedinger (1980) as well as an adaptation of the efficient *LN2* estimator of the Pearson correlation coefficient ρ introduced by Stedinger (1981) and recently evaluated for highly skewed observations by Barber et al. (2019). Our estimators of E and E', which we term the Lamontagne-Barber efficiency estimators *LBE* and *LBE′*, respectively, take the form

$$LBE = 2\hat{\alpha}\hat{\rho} - \hat{\alpha}^2 - \frac{\hat{\Delta}^2}{\hat{C}_o^2} \tag{18a}$$

$$LBE' = 1 - \sqrt{(\hat{\beta} - 1)^2 + (\hat{\alpha} - 1)^2 + (\hat{\rho} - 1)^2} \tag{18b}$$

where

$$\widehat{C}_o = \frac{\sqrt{\exp\left(2\overline{u} + s_u^2\right)\left(\exp\left(s_u^2\right) - 1\right)}}{\widehat{\tau}_o + \exp\left(\overline{u} + \frac{s_u^2}{2}\right)}$$

$$\widehat{\alpha} = \frac{\widehat{\sigma}_s}{\widehat{\sigma}_o} = \sqrt{\frac{\exp\left(2\overline{v} + s_v^2\right)\left(\exp\left(s_v^2\right) - 1\right)}{\exp\left(2\overline{u} + s_u^2\right)\left(\exp\left(s_u^2\right) - 1\right)}}$$

$$\widehat{\Delta} = 1 - \widehat{\beta} = \frac{\widehat{\mu}_o - \widehat{\mu}_s}{\widehat{\mu}_o} = 1 - \frac{\widehat{\tau}_s + \exp\left(\overline{v} + \frac{s_v^2}{2}\right)}{\widehat{\tau}_o + \exp\left(\overline{u} + \frac{s_u^2}{2}\right)}$$

$$\widehat{\rho} = r_s = \frac{\exp\left[s_{uv}^2\right] - 1}{\sqrt{\left(\exp\left[s_u^2\right] - 1\right)\left(\exp\left[s_v^2\right] - 1\right)}}$$

where r_s denotes Stedinger's (1981) estimator and $u_i = \ln[o_i - \widehat{\tau}_O]$ and $v_i = \ln[s_i - \widehat{\tau}_S]$ with

$$\overline{u} = \frac{1}{n}\sum_{i=1}^{n} u_i \quad \text{and} \quad \overline{v} = \frac{1}{n}\sum_{i=1}^{n} v_i$$

$$s_{uv}^2 = \frac{1}{n}\sum_{i=1}^{n}(u_i - \overline{u})(v_i - \overline{v})$$

$$s_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(u_i - \overline{u})^2 \text{ and } s_v^2 = \frac{1}{n-1}\sum_{i=1}^{n}(v_i - \overline{v})^2$$

A very attractive and efficient estimator of the lower bounds τ_O and τ_S for use in (18a) and (18b) is given by Stedinger (1980) as

$$\widehat{\tau}_O = \frac{o_{(1)}o_{(n)} - (o_{0.5})^2}{o_{(1)} + o_{(n)} - 2o_{0.5}} \text{ and } \widehat{\tau}_S = \frac{s_{(1)}s_{(n)} - (s_{0.5})^2}{s_{(1)} + s_{(n)} - 2s_{0.5}}$$

where $o_{(1)}$ and $o_{(n)}$ are the smallest and largest observations, respectively, and $o_{0.5}$ is an estimate of the median observation, o. Analogous definitions exist for estimation of $\widehat{\tau}_S$ based on the simulations s. The conditions $o_{(1)}+o_{(n)} - 2o_{0.5} > 0$ and $s_{(1)}+s_{(n)} - 2s_{0.5} > 0$ must be satisfied to obtain reliable estimates of $\widehat{\tau}_o$ and $\widehat{\tau}_s$ in (18a) and (18b). In situations when that condition cannot be satisfied, we resort to setting either $\widehat{\tau}_o = 0$ and/or $\widehat{\tau}_s = 0$, which implies an LN2 instead of an LN3 model. We advise against allowing $\widehat{\tau}_o < 0$ or $\widehat{\tau}_s < 0$ until such time as a zero-inflated model is introduced to handle the occurrence of zeros and the resulting discontinuity in the pd, which results. Also note that for an LN2 distribution the formula for $\widehat{C}_o$ in (18a) and (18b) reduces to the simpler expression $\widehat{C}_o = \sqrt{\exp\left(s_u^2\right) - 1}$.

4.5. BLN3 Mixture Estimators of Efficiency: (LBE_m and LBE'_m)

A natural improvement to the estimators *LBE* and *LBE′* introduced in the previous section are estimators which exploit the *BLN3-MM* model, which accounts for both the skewness and periodicity of the observations and simulations. Given the *BLN3-MM* model summarized in section 3.2, we can use the fact that $E\left[O^k\right] = \sum_{i=1}^{12} w_i E\left[O_i^k\right]$ for the observations O (and analogously for the simulations S), which follows directly from (13), where each month is assumed to have an equal number of days, with nonzero observations on every day, so that $w_i = 1/12$. That fact leads to the following expressions for the mean and variance of the observations for the *BLN3-MM* model:

$$\mu_o = \sum_{i=1}^{12} \mu_i/12 \tag{19a}$$

$$\sigma_o^2 = \left[\sum_{i=1}^{12}\left(\sigma_i^2 + \mu_i^2\right)/12\right] - \mu_o^2 \tag{19b}$$

We employ (19a) and (19b) and analogous expressions for the mean and variance of the simulations, to develop improved *BLN3-MM* estimators of both E and E', which we term LBE_m and $LBE_m^{'}$ respectively:

$$LBE_m = 2\widehat{\alpha}_m r_m - \widehat{\alpha}_m^2 - \frac{\widehat{\Delta}_m^2}{\widehat{C}_{m,o}^2} \tag{20a}$$

$$LBE_m^{'} = 1 - \sqrt{\widehat{\Delta}_m^2 + (\widehat{\alpha}_m - 1)^2 + (r_m - 1)^2} \tag{20b}$$

with the *BLN3-MM* mixture estimators (denoted using subscript m) obtained from

$$\widehat{\alpha}_m = \frac{\widehat{\sigma}_{m,s}}{\sigma_{m,o}}, \quad \widehat{\Delta}_m = 1 - \frac{\widehat{\mu}_{m,s}}{\widehat{\mu}_{m,o}}, \quad \widehat{C}_{m,o} = \frac{\widehat{\sigma}_{m,o}}{\widehat{\mu}_{m,o}}, \quad r_m = \frac{\widehat{\mu}_{m,so} - \widehat{\mu}_{m,o}\widehat{\mu}_{m,s}}{\widehat{\sigma}_{m,o}\widehat{\sigma}_{m,s}}$$

with $\widehat{\mu}_{m,o}, \widehat{\mu}_{m,s}\widehat{\sigma}_{m,o}, \widehat{\sigma}_{m,o}$, and $\widehat{\mu}_{m,so}$ computed from transformed observations $u_i = \ln\left(o_i - \widehat{\tau}_{o,i}\right)$ and $v_i = \ln\left(s_i - \widehat{\tau}_{s,i}\right)$ in each month using

$$\widehat{\mu}_{m,s} = \sum_{i=1}^{12}\widehat{\mu}_{s,i}/12 \qquad \widehat{\mu}_{s,i} = \widehat{\tau}_{s,i} + \exp\left(\overline{v}_i + \frac{s_{v,i}^2}{2}\right)$$

$$\widehat{\mu}_{m,o} = \sum_{i=1}^{12}\widehat{\mu}_{o,i}/12 \qquad \widehat{\mu}_{o,i} = \widehat{\tau}_{o,i} + \exp\left(\overline{u}_i + \frac{s_{u,i}^2}{2}\right)$$

$$\widehat{\sigma}_{m,o}^2 = \sum_{i=1}^{12}\left[\left[\widehat{\sigma}_{o,i}^2 + \widehat{\mu}_{o,i}^2\right]/12\right] - \widehat{\mu}_{m,o}^2 \quad \widehat{\sigma}_{o,i}^2 = \exp\left(2\overline{u}_i + s_{u,i}^2\right)\left(\exp\left(s_{u,i}^2\right) - 1\right)$$

$$\widehat{\sigma}_{m,s}^2 = \sum_{i=1}^{12}\left[\left[\widehat{\sigma}_{s,i}^2 + \widehat{\mu}_{s,i}^2\right]/12\right] - \widehat{\mu}_{m,s}^2 \quad \widehat{\sigma}_{s,i}^2 = \exp\left(2\overline{v}_i + s_{v,i}^2\right)\left(\exp\left(s_{v,i}^2\right) - 1\right)$$

$$\widehat{\mu}_{m,so} = \frac{1}{12}\sum_{i=1}^{12}\left[\widehat{\mu}_{s,i}\widehat{\mu}_{o,i} + r_{s,i}\widehat{\sigma}_{s,i}\widehat{\sigma}_{o,i}\right]$$

where $r_{s,i}$ is the modified Stedinger (1981) estimator of $\widehat{\rho} = r_s$ given in (18a) and (18b) and applied to the transformed observations and simulations in each month i. Here $\widehat{\tau}_{o,i}$, $\overline{u}_i$, and $s_{u,i}$ denote Stedinger (1980) lower bound, sample mean, and sample standard deviation of the values of the transformed observations $u_i = \ln\left(o_i - \widehat{\tau}_{o,i}\right)$ in each month i. Similarly, $\widehat{\tau}_{s,i}$, $\overline{v}_i$, and $s_{v,i}$ denote Stedinger (1980) lower bound, the sample mean, and sample standard deviation of the value of the transformed simulations $v_i = \ln\left(s_i - \widehat{\tau}_{s,i}\right)$ in each month i. We advise against allowing $\widehat{\tau}_{o,i}<0$ or $\widehat{\tau}_{s,i}<0$ until such time as a zero-inflated model is introduced to handle the occurrence of zeros and the bimodal pd which results.

4.6. Log-NSE (*LNSE*)

Numerous investigators have suggested to apply the estimator *NSE* to a logarithmic transformation of the observations and simulations; we term this estimator *LNSE*. For example, *LNSE* is commonly used for both model calibration and validation (Krause et al., 2005; Santos et al., 2018), particularly when simulating low flows is a focus (Pushpalatha et al., 2012). The rationale for using *LNSE* as opposed to *NSE* is that it increases the relative weight assigned to the smallest observations by reducing the asymmetry in streamflow observations. It is also important to realize that *LNSE* will always produce a biased estimate of E because $E[LNSE] \neq E$. In general, one expects $E[LNSE] > E$ because the log-space correlation is generally greater than the real space correlation (see Barber et al., 2019, Equation 6), though the exact relationship is complicated and depends on the log-space moments of the observations and simulations. Importantly, Santos et al. (2018) document that a log transformation should not be applied to *KGE* and its variants, because spurious results can be obtained which depend arbitrarily upon which set of units were used. Thus, we do not consider such estimators here.

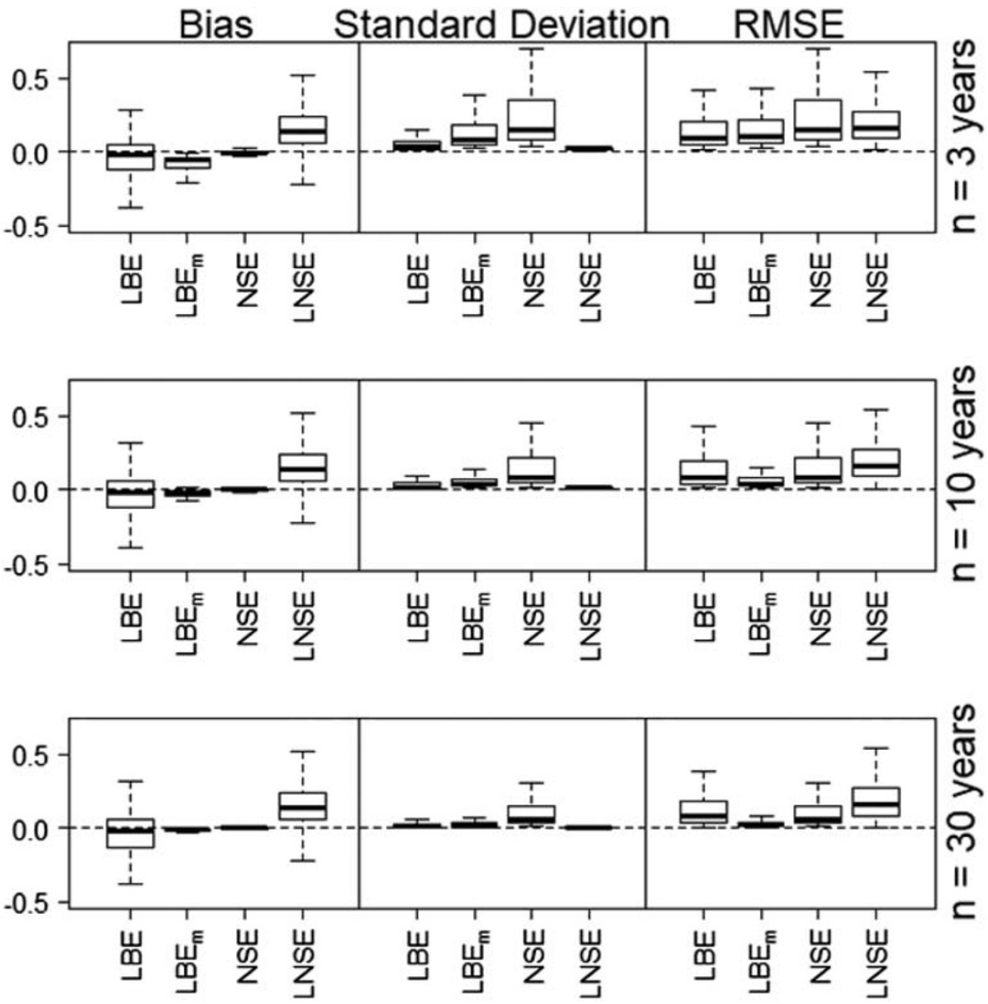

Figure 4. Boxplots of estimates of efficiency E resulting from 1,000 Monte Carlo experiments performed at each of the 447 sites summarized in Table 1.

5. Experimental Results

Monte Carlo experiments enable evaluation of the sampling properties (*bias*, *standard deviation*, and *root-mean-square error [RMSE]*) of the four estimators of E (*LBE; LBE$_m$; NSE; LNSE*) and E' (LBE', LBE_m', KGE', $PVSE'$), summarized in section 4, when applied to synthetic daily streamflow series generated from the *BLN3-MM* model. We also compare the behavior of the eight estimators of efficiency when applied to the PRMS model output summarized in Table 1.

5.1. Monte Carlo Experiments

Our controlled Monte Carlo experiments are unique because to our knowledge, this is the first time that controlled experiments have been performed for evaluating the performance of *NSE* and KGE'. A more attractive estimator is one that yields a "better" estimate of E or E' across all 447 sites considered. Of course, the choice of a "best" estimator will depend on the choice of a metric, or loss function. Among statisticians, the most common performance index is the *MSE* criterion of optimality (see Everitt, 2002, page 128). It is well known that the *MSE*, variance, and bias of an estimator, such as *NSE*, are related via $MSE[NSE] = E[(NSE - E)^2] = (E[NSE] - E)^2 + E[(NSE - E[NSE])^2] = Bias[NSE]^2 + Var[NSE]$ so that the *MSE* of *NSE* is made up of both its bias and variance (also see Equation 8). In the following section, we report all three metrics, because they are all related and important for different reasons described below.

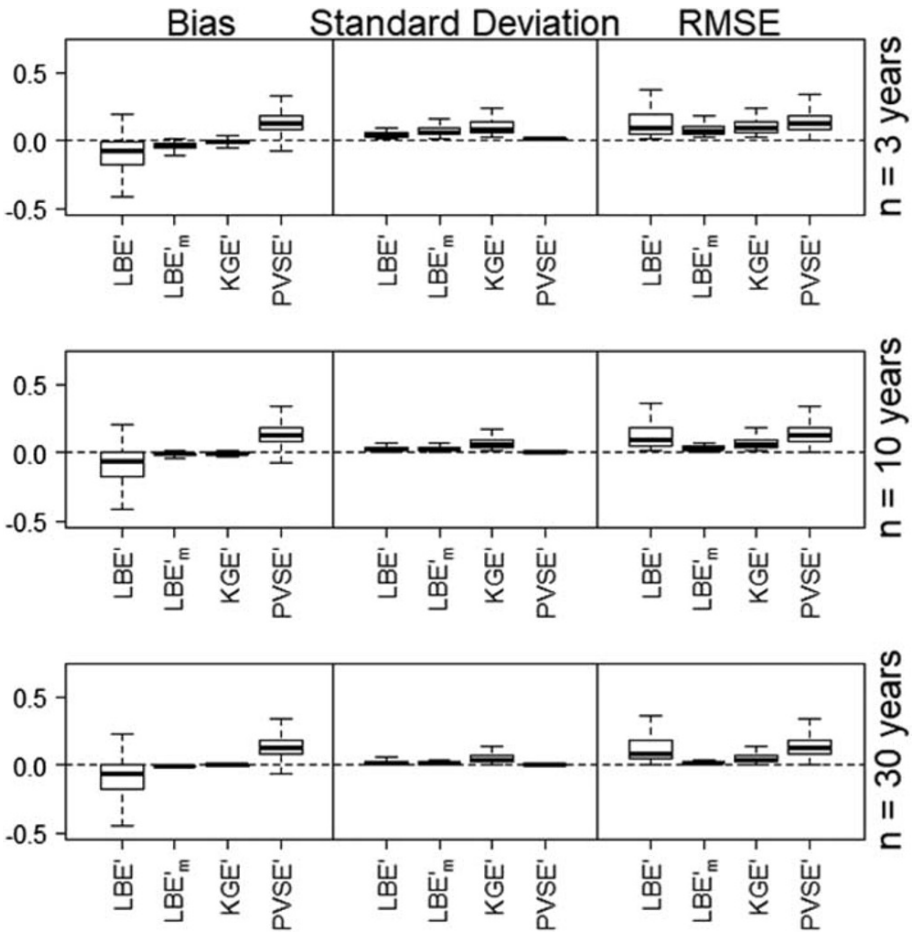

Figure 5. Boxplots of estimates of efficiency E' resulting from 1,000 Monte Carlo experiments performed at each of the 447 sites summarized in Table 1.

Any Monte Carlo experiment requires an assumption of the true value of the statistic (in this case E and E') in advance; otherwise, no definitive insights concerning an estimator's performance could be derived. Put differently, one cannot compute either bias or *MSE* without assuming the true value for the statistic of interest. We assume that the true (population) values of E and E' are equal to the sample values of LBE_m and LBE'_m, computed from the complete period of record of the 447 sites summarized in Table 1. These values are chosen as true values because they are the only estimators of E and E', which account for skewness and periodicity, both central aspects of our contribution.

We generate 1,000 sets of streamflow simulations and observations each of length n = 1,095 days (N = 3 years), n = 3,650 days (N = 10 years), and n = 10,950 days (N = 30 years) to capture the range of conditions typically encountered when calibrating a hydrologic model to observations. Synthetic sequences of daily streamflows are generated at each of the sites summarized in Table 1 by first generating *BLN3* sequences in each month of length n/12, using the algorithm described in the appendix. True values of the required statistics for generating B*LN3* streamflows in each month are assumed equal to the sample statistics based on the *BLN3-MM* model obtained from the full period of record at each site. A complete set of synthetic streamflows is then created by assembling N = 3-, 10-, and 30-year sequences where each year contains synthetic daily streamflows from each of the 12 months.

5.2. Results of Monte Carlo Experiments

Figures 4 and 5 summarize the bias, standard deviation, and *RMSE* associated with the estimators of E and E' respectively, obtained from the Monte Carlo experiments at the 447 sites summarized in Table 1, for sample sizes equal to 3, 10, and 30 years. Statistics for the eight featured sites in Figure 2 are reported in Barber (2020,

Table 3.2). We emphasize that the bias, variability, and *RMSE* in estimators of E and E' illustrated using boxplots in Figures 4 and 5 should be interpreted as occurring across the 447 sites, which reflect diverse hydrologic conditions across the contiguous United States.

Perhaps the most important finding in Figure 4 is the remarkably high variability (evidenced by standard deviation) associated with *NSE*, when compared with the other three estimators of E. Even though *NSE* is consistently unbiased across sites, it exhibits enormous variability from one sample to the next at most sites. The *RMSE* associated with the estimator *NSE* is also generally higher than LBE_m, because *RMSE* is made up of both bias and variance, and variance dominates the behavior of *NSE*. As a result, we cannot recommend use of *NSE* with daily or subdaily streamflows. Instead, to obtain nearly unbiased estimates of E, we would recommend use of LBE_m, which is approximately unbiased at most sites and exhibits much lower *RMSE* than either *NSE* or *LBE*, particularly for the larger sample sizes. Small sample sizes cause increased sampling variability and bias associated with all estimators of efficiency, and particularly LBE_m because the bivariate monthly mixture model requires estimation of 48 parameters.

There are several important conclusions which may be drawn from Figure 5. First, the increasingly widely used estimator KGE' is approximately unbiased across all sites and exhibits much lower standard deviation and *RMSE* than was illustrated for *NSE* in Figure 4. Therefore, KGE' appears to be a more useful and stable statistic than *NSE*. It is important to reemphasize, however, that KGE' is an estimator of E' which, as we have shown, is a very different statistic than the statistic E that *NSE* attempts to estimate. We also note that the estimator $LBE_m^{'}$ generally exhibits much lower standard deviation and *RMSE* than KGE' and only exhibits a very slight downward bias for small samples, thus we generally recommend use of $LBE_m^{'}$ over KGE'.

Another interesting finding from Figures 4 and 5 is the remarkably low standard deviation associated with both *LNSE* and $PVSE'$ when compared with all other estimators of E and E', respectively, for all sample sizes considered. One expects the very high upward bias in *LNSE* at most sites illustrated in Figure 3, in part because the correlation between the natural logarithms of O and S is always greater than the correlation between their real space values (see page 5 in Barber et al., 2019).

Low standard deviation associated with estimates of E and E' is paramount when one's interest is in development of the best possible model for a given watershed. In other words, the very low standard deviation associated with all the estimators considered here (except *NSE* and KGE'), across so many sites, implies that they would all be useful for model calibration and/or for any evaluations of model performance at a single watershed. This is because estimators of E with low variance will tend to give estimates with high precision and thus very little variability from one sample to the next, even if the corresponding estimates are all biased (on average far from the true values). The reason *NSE* and KGE' exhibit such high variability is due to the fact that they are both based on product moment estimators, and all product moment estimators are known to perform poorly for highly skewed daily streamflow samples, even for very large sample sizes (Vogel & Fennessey, 1993).

Unbiasedness of estimators of both E and E' is paramount when comparing the performance of models across watersheds or when developing regional relationships among watershed model parameters and basin characteristics. This is due to the fact that when comparing biased estimators, one will never know if the differences are due to the sampling bias or due to actual differences in model performance, whereas when comparing models using unbiased efficiency estimators, the differences will arise mostly from differences in model performance. It is evident from our experiments that the only nearly unbiased estimators of E and E', which also have acceptably low values of standard deviation and thus *RMSE*, are the estimators LBE_m and $LBE_m^{'}$ when used with sample sizes in excess of roughly 10 years of daily streamflow. In contrast, we note that both *LNSE* and $PVSE'$ exhibit a very large and unpredictable level of mostly upward bias, which would cause severe interpretation problems if these statistics were used to compare goodness of fit across sites.

5.3. Comparison of Sample Estimates of E and E' for Real and Synthetic Data

The Monte Carlo experimental results are only interesting to hydrologists to the extent that they approximate real hydrologic conditions. Section 3 demonstrated that the *BLN3-MM* model provides a first-order approximation of hydrologic observations and simulations across the 447 PRMS watersheds. Here we

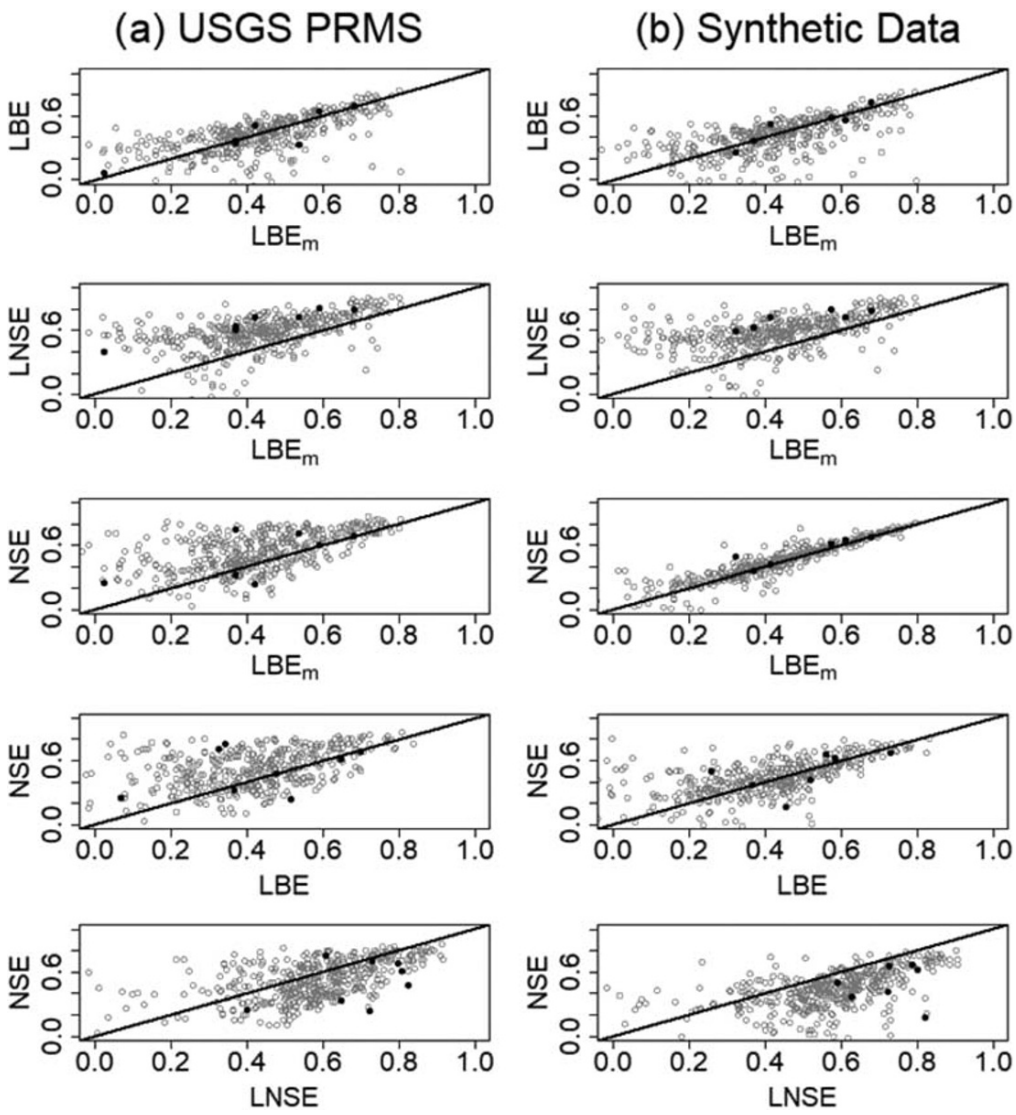

Figure 6. (a and b) Scatterplots of estimates of E obtained from various estimators, with results from eight sites in Figure 2 and Table 3.2 of Barber (2020) shown using dark black circles.

document the similarity in the behavior of estimators of E and E' between synthetic (Monte Carlo) data and actual data.

Figure 6 reports scatterplots among the various estimators of E corresponding to the observations and simulations from the PRMS model (left column) and synthetic sequences generated from the *BLN3-MM* model for the 447 watersheds. The dark black circles in Figure 6 indicate results for the eight watersheds highlighted Figure 2. The similarity in the behavior of estimates of E between the synthetic and real data in Figure 6 provides additional evidence that the *BLN3-MM* model used in our Monte Carlo experiments approximates realistic hydrologic conditions, which are relevant to hydrologists. Analogous plots for E' are reported by Barber (2020, Figure 3.4) The enormous variability in the estimates of both E and E' in Figures 6 and Barber (2020, Figure 3.4) highlight the importance of our controlled Monte Carlo experiments, where we were able to compare the various estimators to their true values, something which cannot be done in these figures.

5.4. The Source of Variability in Efficiency Estimates

What causes the enormous sampling variability associated with the estimator *NSE*, and to a lesser extent *KGE′*, that was observed in our Monte Carlo experiments? Variability in *NSE* and *KGE′* from one sample

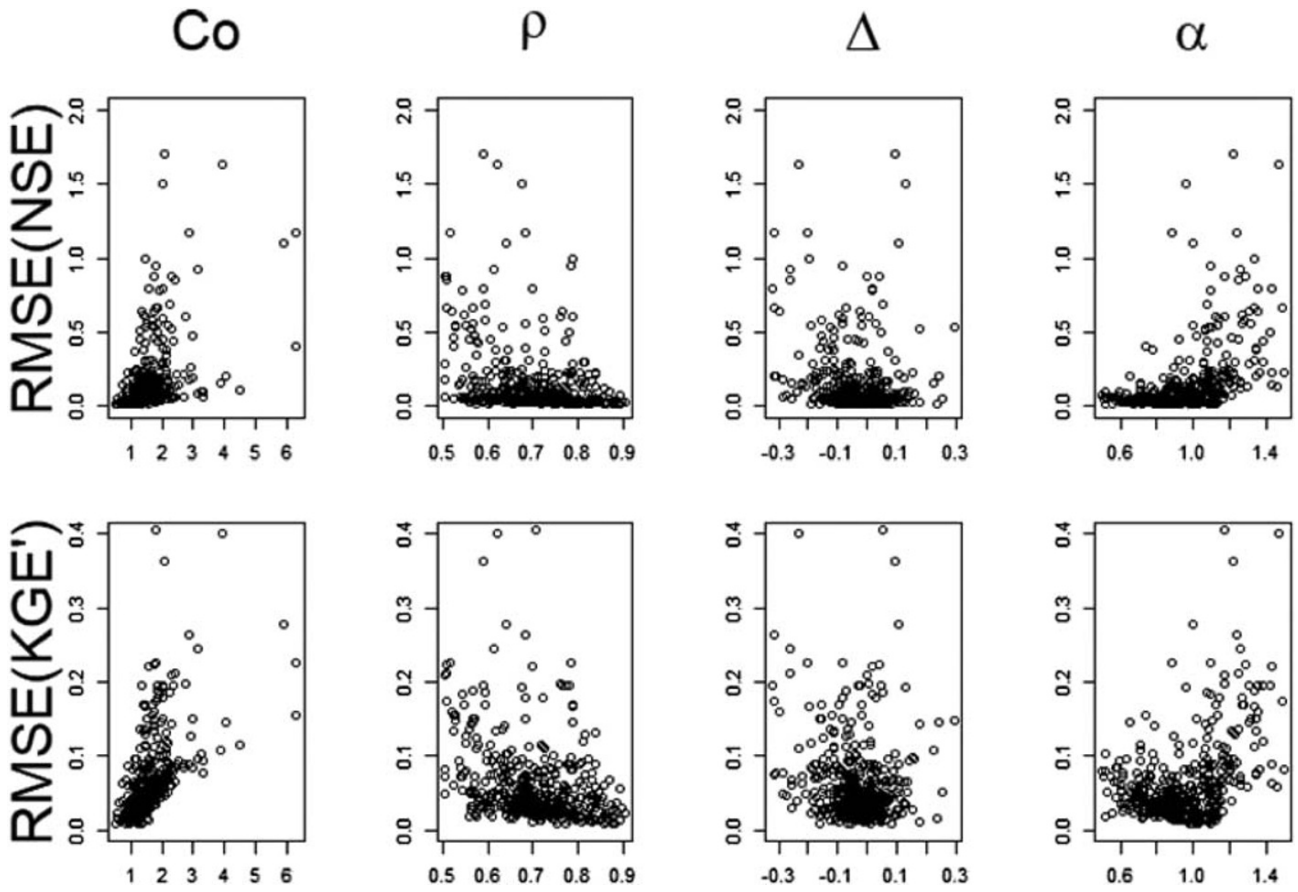

Figure 7. Root-mean-square error of *NSE* and *KGE′* versus *Co*, ρ, Δ, and α. RMSE (*NSE*) and RMSE (*KGE′*) computed from 1,000 Monte Carlo experiments of length $n = 10{,}950$ (30 years) for each of the 447 sites.

to the next arises from numerous sources including the sample size and the degree of cross correlation, variability, skewness, periodicity, and autocorrelation present in the simulations and observations under consideration. Each of these sources is discussed separately below.

5.4.1. Impact of Variability and Skewness of Observations

Figure 7 illustrates the relationship between both *RMSE*(*NSE*) and *RMSE*(*KGE′*) versus the *BLN3-MM* model estimates of C_o, ρ, Δ, and α computed at each of the 447 sites. Here values of *RMSE*(*NSE*) and *RMSE*(*KGE′*) are based on the results of 1,000 Monte Carlo replicates of sample length $n = 10{,}950$ (30 years) for each of the 447 sites. We note that increases in both *RMSE*(*NSE*) and particularly *RMSE*(*KGE′*) result from increases in streamflow variability reflected by C_o and, to a lesser degree, from decreases in ρ. Note that the wide range of values of C_o reported in Figure 7 and Table 1 does not even reflect the enormous variability possible in some regions documented by Vogel et al. (2003) who used *LN3* estimators to report a range in C_o values for daily flow series across the United States from approximately 0.5 to 10,000 with a median value of 10, and an interquartile range from 3 to 33. Thus, the error in estimates of *NSE* and *KGE′* is likely substantially higher in many hydrologic modeling applications not considered here, particularly in arid and semi-arid regions.

5.4.2. Impact of Periodicity of Observations

In addition to skewness, the periodicity of streamflow plays an equally important role in causing variability associated with estimates of *E*. This can be seen by contrasting the results of Figure 6 in this paper, with those of Figure 5 in Barber et al. (2019), which ignored periodicity. Ignoring the periodicity of daily streamflow as was done by Barber et al. (2019) could not reproduce the expected variability in estimates of ρ (and thus *E*) derived from actual streamflow series.

5.4.3. Impact of Goodness of Fit, Sample Size, and Autocorrelation

Here we approximate the information content of daily streamflow series, an extremely complex problem well beyond the scope of this study. Credible efforts to characterize daily streamflows involve characterization of marginal distributions, spatiotemporal correlation structures, and intermittency (Papalexiou & Serinaldi, 2020), as well as climatic indices and epochal variations in predictability (Rajagopalan et al., 2019). Archfield et al. (2013) argue that at least seven fundamental streamflow statistics are needed to fully characterize daily streamflow. Here we approximate the role of streamflow persistence, sample size, and model goodness of fit on the information content of daily streamflow series.

Our Monte Carlo simulations ignore the impacts of the enormous serial correlation exhibited by daily streamflow. Serial correlation results in reductions in the information content or effective sample size of the flow record. Two hypothetical extremes exist: independent series of length n with no loss of information and, conversely, series with lag-one autocorrelation ρ_1 equal to unity resulting in an effective sample size equal to $n = 1$. Daily streamflows exhibit values of ρ_1 near unity, which results in dramatic decreases in the effective record length of the streamflow observations. Recall from section 2.3 that for an unbiased model with residuals that are independent of the simulations $E = \rho^2$ and $E' \approx 2.41 - 1.41\rho$, thus, it is instructive to understand the sampling properties of the Pearson correlation coefficient r in (16) the most common estimator of ρ. Barber et al. (2019) report that for an AR(1) normal process

$$Var[r] = \left[\left(1-\rho^2\right)^2/n\right]\left[\left(1+\rho_{1,S}\rho_{1,O}\right)/\left(1-\rho_{1,S}\rho_{1,O}\right)\right] \tag{21}$$

where $\rho_{1,S}$ and $\rho_{1,O}$ are lag-one correlations of S and O, respectively. Additional variability associated with estimates of r over and above that described by (21) are expected for highly skewed and periodic flow series as discussed in section 1.4 and by Barber et al. (2019).

Equation 21 documents the three critical factors which influence the variability of *NSE* and *KGE* in situations when they are impacted only by correlation ρ between O and S, for the hypothetical AR(1) normal case. Equation 21 highlights the large reductions in the variance of r, and thus *NSE* and *KGE′*, as model goodness of fit and sample size increase. For example, the term $(1-\rho^2)^2$ decreases from 0.26 to 0.036 for ρ equal to 0.7 and 0.9, respectively, an order of magnitude reduction in variance of r. The second quantity in square brackets in (21) represents the inflation in the variance due to autocorrelation. For example, when $\rho_{1,S} = \rho_{1,O} = 0.9$ that factor is equal to $[(1+0.9 \cdot 0.9)/(1 - 0.9 \cdot 0.9)] = 9.53$ or a nearly tenfold increase in variance over independent series.

Ignoring the impact of skewness and periodicity, (21) can be used to approximate the impact of ignoring serial correlation by defining an effective record length $n' = n[(1+\rho_{1,S}\rho_{1,O})/(1 - \rho_{1,S}\rho_{1,O})]^{-1}$ (see Matalas & Langbein, 1962). For example, a 10-year record of correlated $n = 3{,}650$ daily streamflows with $\rho_{1,S} = \rho_{1,O} = 0.9$ is equivalent to only $n' = 383$ independent observations, or a nearly tenfold decrease in information. The primary impact of serial correlation on estimates of summary statistics is to inflate their variance; thus by ignoring the serial correlation of the daily streamflows, we are considerably *understating* the resulting variability associated with the various estimators of E.

6. Discussion: Caveats, Improvements, and Extensions

We have introduced a *BLN3-MM* model which appears to provide a very good representation of the pd and various other properties of daily streamflows across the conterminous United States; however, there are several caveats, improvements, and extensions that are possible, as discussed below. We do not claim that daily streamflows follow a *BLN3* distribution in a given month; rather, we argue that a *BLN3-MM* model provides a much better approximation to daily streamflow observations and simulations than a bivariate normal model, which is a required assumption for the product moment statistics embedded within *NSE* and *KGE* to exhibit low bias and variance.

6.1. Handling Zero Streamflows

The occurrence of zero streamflow was not considered here but leads to considerable increases in both C_o and C_s and corresponding increases in the variance of estimators of efficiency and correlation (Barber et al., 2019); thus, it is important to accommodate the occurrence of zeros. Zero streamflows are defined as streamflow

below the measurement threshold which, in the United States is approximately 0.01 cfs (Granato et al., 2017). Of the 20,438 USGS river gages evaluated by Granato et al. (2017), 36% of those gages had at least one occurrence of zero streamflow and 2.6% of those gages had more than 297 days per year (or 81.3%) of zero streamflow. According to Levick et al. (2008), ephemeral and intermittent streams make up approximately 59% of all streams in the United States (excluding Alaska), and over 81% in the arid and semiarid Southwest according to the USGS National Hydrography Dataset. The family of *LBE* estimators introduced here should not be used at sites with zero observations, because the occurrence of zeros introduces a discontinuity in the pd of daily streamflows, which is not captured by either the *BLN3* or the *BLN3-MM* models. A natural extension to this study would be to develop estimators of E based on a zero-inflated *BLN3-MM* model analogous to the zero-inflated *BLN2* mixture model introduced by Shimizu (1993) for modeling rainfall.

6.2. Improvements to Mixture Model and Estimators of E and E'

A natural extension to this study would be to develop improvements to the *BLN3-MM* model, which lead to better reproduction of the pd and stochastic persistence of the observations which in turn, should lead to improved estimators of both E and E'; recommendations are provided below.

6.2.1. Improved Reproduction of Stochastic Persistence

Section 5.4.3 described the general impact of ignoring the serial correlation of S and O. Improved estimators of E and E' over those developed here would result from improvements to our *BLN3-MM* model, which capture the stochastic persistence of O and S, including the within-month and month-to-month serial correlation structure of the daily flows. Such improvements are described in literature dealing with the development of daily stochastic streamflow models (see Papalexiou & Serinaldi, 2020, and recent literature reviews in Vogel, 2017, and Brunner et al., 2019).

6.2.2. An Improved Bivariate Kappa Mixture Model

Blum et al. (2017), Brunner et al. (2019), and others have shown that a Kappa pd provides a better fit to the distribution of daily streamflows than an *LN3* pd. Therefore, a natural improvement would be to (1) generate synthetic streamflow series and (2) develop improved estimators of E and E' based on a bivariate Kappa monthly mixture (*BKAP-MM*) model. Such a *BKAP-MM* model could be combined with a suitable copula to provide an improved representation of both the dependence structure and marginal distributions of the daily streamflow observations and simulations. Generation of synthetic streamflows from a *BKAP-MM* model would also enable a robustness study which would evaluate how well the estimators LBE_m and LBE'_m perform when streamflows arise from a more realistic process than the *BLN3-MM* model.

6.2.3. Improved Estimators of E and E'

Future work may benefit from using maximum likelihood or Bayesian estimators of the parameters of the *BLN3*-MM model as recommended by McLachlan et al. (2019). Another alternative approach to our *BLN3-MM* model would be the use of seasonal reference values of E and E' to account for seasonal and other dynamic variations in goodness of fit as recommended by Schaefli and Gupta (2007) and shown by Reusser et al. (2009) to be necessary for diagnosis of model performance.

6.2.4. A More Parsimonious Mixture Model

The *BLN3-MM* model introduced here is not parsimonious because it requires estimation of 48 parameters, which could lead to increased sampling bias and variance associated with our recommended estimators for short samples. Alternatively, a seasonal model could be considered, which has fewer parameters yet still captures the important deterministic periodic behavior of streamflows. For example, generalized linear models and/or generalized additive models whose parameters depend on sine/cosine functions could account for seasonal behavior using a smaller number of parameters (McCullagh & Nelder, 1989). Future studies are needed to better understand the degree of parsimony needed in the mixture model so as to provide efficient and robust estimators of E while still capturing the critical complexities of the hydrologic process of interest including periodicity, occurrence of zeros, and skewness. The attractive idea of Clarke (2008) of adding the number of model parameters to estimators of efficiency could prove useful in evaluating the impact of parsimony.

7. Conclusions and Recommendations

Our approach differs from past research relating to the estimation of model performance efficiency for evaluation of goodness of fit, because our conclusions are based on both theoretical (probabilistic) and

empirical (statistical) analyses which enabled controlled experiments. Considering the myriad of previous applications and evaluations of *NSE* combined with the fact that Todini and Biondi (2017) report that *NSE* "is by far the most utilized index in hydrological applications," it is surprising that it took this long to advance a theoretical or probabilistic definition of efficiency to enable controlled Monte Carlo experiments which evaluate alternative estimators of efficiency. Perhaps our most fundamental contribution was to clarify and distinguish for the first time, both the theoretical (probabilistic) properties of efficiency and the empirical sampling (statistical) properties of various estimators of efficiency introduced by Nash and Sutcliffe (1970) and later improved upon by Gupta et al. (2009), Gupta and Kling (2011) and others. Below we summarize our major findings:

General comments on E and E′: We have introduced two different probabilistic definitions of efficiency termed E and E', which are consistent with the now widely used empirical estimators known as *NSE* and *KGE′*, respectively. The theoretical statistic E has a well-known interpretation as a standardized form of *MSE*, whereas the interpretation of E' is somewhat less clear because it is only loosely related to *MSE* and *RMSE*. Figure 1 clearly shows the different behavior of the two theoretical statistics E and E'. Measures of *MSE* and *RMSE* are perhaps the most widely used metrics of goodness of fit across all disciplines and E is simply a standardized version of those statistics, whereas E' is only loosely related to *MSE* and *RMSE*. We have shown that the statistic E' has attractive sampling properties (low bias and variance), yet if E' is to be considered further, attention should be given to its interpretation analogous to the definition of E as a measure of standardized *MSE*.

Enormous variability of NSE but not KGE′: Even though *NSE* was shown to be a consistently unbiased estimator of E, it exhibits extraordinary variability from one sample to another, at most sites, and as a result, we cannot recommend its use with daily or subdaily streamflows. The statistic *NSE* is likely to be even more variable at intermittent and ephemeral sites, which were not considered in this study. The estimator *KGE* was shown to be an approximately unbiased estimator of E' across all sites and to exhibit considerably lower variability and *RMSE* than was illustrated for *NSE*. An important finding of our work was that *NSE* was shown to be consistently unbiased; thus, future research is needed, which applies variance reduction methods (Avramidis & Wilson, 1996; Chernick, 2008) to obtain more efficient (lower variance) versions of the *NSE* estimator.

Improved BLN3 monthly mixture model estimators: To obtain nearly unbiased estimates of E and E', we recommend the use of LBE_m and LBE'_m, respectively, which are approximately unbiased at most sites and exhibit much lower *RMSE* than either *NSE* or *PVSE′* and slightly lower *RMSE* than *KGE′*, particularly for the larger sample sizes. The primary reasons that these estimators are favored are because they address the critical issues of skewness and periodicity, which both confound the performance of *NSE* and *KGE′*. We highlight that the LBE_m and LBE'_m estimators introduced here are only expected to be improvements over *NSE* under the relatively restricted watershed conditions of nonzero streamflows and observations which are well approximated by a *BLN3* model in all months.

Extremely low variability associated with LNSE and PVSE′: The estimators *LNSE* and *PVSE′* are generally highly upward biased estimators of E and E', respectively; however, both estimators exhibit extremely low variance from one sample to the next and thus are both recommended over use of *NSE* in calibration and goodness-of-fit evaluations at a single site. However, due to their considerable and unpredictable bias, *LNSE* and *PVSE′* should not be used in regional or other studies, which attempt to compare model goodness of fit across sites, nor should they be used to draw comparisons with other unbiased estimators of E or E' respectively. The *RMSE* associated with the estimators *LNSE* and *PVSE′* was also generally higher than other estimators, because *RMSE* is made up of both bias and variance, and in this case the bias dominated our comparisons.

Synthetic series are similar to PRMS model output: Comparisons of the behavior of estimates of E and E' corresponding to the output of PRMS models and to synthetic streamflows from the *BLN3-MM* model in Figure 6 indicate very similar behavior, which illustrates that the synthetic series must be mimicking, to a great extent, the behavior of the PRMS model output.

The importance of accounting for skewness and periodicity: Daily streamflows exhibit considerable periodicity and skewness, two properties that lead to increased variability in estimates of efficiency. Our *BLN3-MM*

model addresses both skewness and periodicity and consequently; when that model provided a good approximation to the streamflow observations and simulations, it led to considerable improvements in the performance of estimators of efficiency based on that model. Daily streamflows are neither normally (zero skew) nor identically distributed, both critical assumptions needed for *NSE* and *KGE′* to exhibit low bias and variance. Evidence of the importance of accounting for periodicity is provided by the marked reduction in *RMSE* of efficiency estimators obtained from the *BLN3-MM* model so that the *RMSE* associated with both seasonal estimators LBE_m and LBE'_m are markedly lower than the *RMSE* associated with either of the nonseasonal estimators *LBE* and *LBE′*, which is in part due to the addition of 44 parameters that capture seasonal variations. Studies in arid, semiarid, ephemeral, and intermittent streams with marked seasonal behavior and very high values of skewness are likely to exhibit even greater variability associated with the estimators *NSE* and *KGE′* then was reported here.

Implications of findings: We have shown that the application of *NSE* to bivariate monthly mixtures of *LN3* samples of daily streamflow can lead to highly variable results from one sample to another, at a single site. These findings indicate that its use in goodness-of-fit evaluations, model calibration, model hypothesis testing, and/or regionalization studies could lead to highly variable results, which would depend arbitrarily upon characteristics of the watershed(s) of interest. Remarkably, this is true even with streamflow record lengths in the thousands. To address these issues, we have introduced initial estimators based on a *BLN3-MM* model, which led to much more consistent and reproducible estimates of efficiency at a single site and across sites. We anticipate that application of our improved estimators will lead to improvements in simulation model calibration, validation, and hypothesis testing and in hydrologic regionalization efforts, which seek to develop multivariate relationships among model parameters and watershed characteristics. Perhaps the most important contribution of our work is to clarify the previous confusion in the literature between theoretical efficiency *E* introduced here, and properties of the widely used sample estimator *NSE*. This confusion has profound consequences because it has led some investigators to suggest that *E* has flaws, when in fact it is only the particular estimator *NSE* that raises concerns.

Recommendations: The *LBE* family of estimators was only evaluated in a set of controlled yet limited experiments; thus, these estimators are not ready for general usage. This is because they were only tested on perennial rivers with daily streamflow observations and simulations, which are well approximated by an *BLN3-MM* model. Future extensions are needed to accommodate zeros and to evaluate the robustness of our estimators to departures from the *BLN3-MM* model. In section 6, we have outlined numerous caveats, improvements, and extensions to the *BLN3-MM* model and associated efficiency estimators. Since this is the first study to perform controlled experiments concerning the performance of model efficiency estimators, our experiments were not nearly exhaustive enough to provide the type of general guidance needed to determine the necessary sample size needed to provide stable, reliable, and unbiased estimates of both *E* and *E′*. We have shown that such guidelines will depend critically upon the sample size as well as the degree of seasonality, skewness, and serial correlation associated with the observations, and it is our hope that future studies will perform additional controlled experiments to arrive at more general guidelines than provided here. In addition to the numerous improvements suggested in section 6, a natural extension to this study would be to develop hypothesis tests and confidence intervals for the true value of *E* using the improved estimators introduced here.

Appendix A: Generation of Streamflow Series From Bivariate LN3 Monthly Mixture Model

We describe a methodology for generating daily streamflows and observations from a bivariate three-parameter lognormal monthly mixture (*BLN3-MM*) model. Balakrishnan and Lai (2009) introduce a bivariate LN2 model and review numerous applications of bivariate lognormal series in a variety of different fields.

The following approach is used to generate bivariate sequences of daily streamflow observations *o*, and simulations *s*, in each month. For example, suppose we wish to generate 10 years of daily streamflows, then $n = 10(365) = 3{,}650$ days, and $n/12 = 304$ daily streamflows are generated from the *BLN3*

monthly model in each month using the procedure below. Here all statistics correspond to the assumed true values of those statistics for a given month described in section 3.3. For assumed values of the coefficient of variation of the observations $C_O = \sigma_O/\mu_O$, and simulations $C_S = \sigma_S/\mu_S$, in a given month, the moments of the natural logarithms of the observations and simulations, $U = \ln[O - \tau_o]$ and $V = \ln[S - \tau_s]$ are given by

$$\mu_U = \ln\left[\frac{\mu_O - \tau_O}{\sqrt{1 + \left(\frac{\sigma_O}{\mu_O - \tau_O}\right)^2}}\right], \quad \sigma_U = \sqrt{\ln\left[1 + \left(\frac{\sigma_O}{\mu_O - \tau_O}\right)^2\right]} \tag{A1a}$$

$$\mu_V = \ln\left[\frac{\mu_S - \tau_S}{\sqrt{1 + \left(\frac{\sigma_S}{\mu_S - \tau_S}\right)^2}}\right], \quad \sigma_V = \sqrt{\ln\left[1 + \left(\frac{\sigma_S}{\mu_S - \tau_S}\right)^2\right]} \tag{A1b}$$

We do not advocate estimation of coefficients of variation from sample data, due to the findings of Vogel and Fennessey (1993); instead, we simply report how we generated artificial data in this section, in which case the values of C_O and C_S are assumed inputs to the experiments, and not estimated from data. Our approach to generation of *BLN3-MM* daily streamflows in a given month is to first generate the observations O, from the lognormal quantile function:

$$O_i = \tau_O + \exp[\mu_U + z(p_i)\sigma_U] \tag{A2}$$

where p_i is a uniform random variate over the interval (0,1) and $z[p_i]$ is the standard normal quantile function evaluated at p_i. Generation of *BLN3* variates is easily implemented by making use of the log space regression so that

$$S_i = \tau_S + \exp\left[\mu_V + \rho_{UV}\frac{\sigma_V}{\sigma_U}(\ln(O_i - \tau_O) - \mu_U) + \kappa_i\right] \tag{A3}$$

with errors κ_i generated from a normal distribution with zero mean and variance equal to $\sigma_\kappa^2 = \sigma_V^2(1 - \rho_{UV}^2)$.

Data Availability Statement

Computer code (both R and Python) implementing the estimators *LBE*, *LBE′*, LBE_m, and LBE_m' are available online (at https://doi.org/10.5281/zenodo.3813836). The streamflow observations and simulations used in this study are available from Farmer and Vogel (2016b).

Acknowledgments
The authors are especially indebted to Francesco Serinaldi for acting as Associate Editor and for his insightful and extremely constructive comments on two early versions of this paper. The authors are also indebted to George Kuczera, Charles N. Kroll, Jose Luis Salinas, A. Sankarasubramanian, and Hoshin Gupta for their insightful comments on an early version of this manuscript. We are also indebted to Ezio Todini, John England, and another anonymous reviewer for their detailed and constructive review of our manuscript.

References

Archfield, S. A., Kennen, J. G., Carlisle, D. M., & Wolock, D. M. (2013). An objective and parsimonious approach for classifying natural flow regimes at a continental scale. *River Research and Applications*, *30*(9), 1166–1183. https://doi.org/10.1002/rra.2710

ASCE (1993). Criteria for evaluation of watershed models. *Journal of Irrigation and Drainage Engineering*, *119*(3), 429–422.

Avramidis, A. N., & Wilson, J. R. (1996). Integrated variance reduction strategies for simulation. *Operations Research*, *44*(2), 327–346. https://doi.org/10.1287/opre.44.2.327

Balakrishnan, N., & Lai, C. D. (2009). *Continuous bivariate distributions* (Second ed., p. 684). Dordrecht, Heidelberg, London, New York: Springer.

Baldwin, C. K., & Lall, U. (1999). Seasonality of streamflow: The upper Mississippi River. *Water Resources Research*, *35*(4), 1143–1154. https://doi.org/10.1029/1998WR900070

Barber, C., Lamontagne, J., & Vogel, R. M. (2019). Improved estimators of correlation and R^2 for skewed hydrologic data. *Hydrological Sciences Journal*, *65*(1), 87–101. https://doi.org/10.1080/02626667.2019.1686639

Barber, C. A. (2020). Evaluating statistical goodness-of-fit estimators for skewed hydrologic data, MS Thesis, Tufts University, Medford, MA.

Bardsley, W. E. (2013). A goodness of fit measure related to r^2 for model performance assessment. *Hydrological Processes*, *27*(19), 2851–2856. https://doi.org/10.1002/hyp.9914

Blum, A. G., Archfield, S. A., & Vogel, R. M. (2017). The probability distribution of daily streamflow in the United States. *Hydrology and Earth Systems Science*, *21*(6), 3093–3103. https://doi.org/10.5194/hess-21-3093-2017

Brunner, M. I., Bárdossy, A., & Furrer, R. (2019). Stochastic simulation of streamflow time series using phase randomization. *Hydrology and Earth System Sciences*, *23*(8), 3175–3187. https://doi.org/10.5194/hess-23-3175-2019

Chernick, M. R. (2008). Chapter 7 efficient and effective simulation. In *Bootstrap methods: A guide for practitioners and researchers* (Second ed., pp. 128–138). Hoboken, N.J: John Wiley and Sons.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., et al. (2008). Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, *44*, W00B02. https://doi.org/10.1029/2007WR006735

Clarke, R. T. (2008). Issues of experimental design for comparing the performance of hydrologic models. *Water Resources Research*, *44*, W01409. https://doi.org/10.1029/2007WR005927

Criss, R. E., & Winston, W. E. (2008). Do Nash values have value? Discussion and alternate proposals. *Hydrological Processes*, *22*(14), 2723–2725. https://doi.org/10.1002/hyp.7072

Efstratiadis, A., & Koutsoyiannis, D. (2010). One decade of multiobjective calibration approaches in hydrological modelling: A review. *Hydrological Sciences Journal – Journal des Sciences Hydrologiques*, *55*(1), 58–78. https://doi.org/10.1080/02626660903526292

Everitt, B. S. (2002). *The Cambridge dictionary of statistics* (2nd ed.). New York: Cambridge University Press. ISBN: 0-521-81099-X.

Ewen, J. (2011). Hydrograph matching method for measuring model performance. *Journal of Hydrology*, *408*(1–2), 178–187. https://doi.org/10.1016/j.jhydrol.2011.07.038

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

Farmer, W. H., & Vogel, R. M. (2016b). on the deterministic and stochastic use of hydrologic models: Data release: U.S. Geological Survey data release, https://doi.org/10.5066/F7W37TF4

Granato, G. E., Ries, K. G., III, & Steeves, P. A. (2017). Compilation of streamflow statistics calculated from daily mean streamflow data collected during water years 1901–2015 for selected U.S. Geological Survey streamgages: U.S. Geological Survey Open-File Report 2017–1108, 17 p., https://doi.org/10.3133/ofr20171108

Guinot, V., Cappelaere, B., Delenne, C., & Ruelland, D. (2011). Towards improved criteria for hydrological model calibration: Theoretical analysis of distance- and weak form-based functions. *Journal of Hydrology*, *401*(1–2), 1–13. https://doi.org/10.1016/j.jhydrol.2011.02.004

Gupta, H. V., & Kling, H. (2011). On typical range, sensitivity and normalization of mean squared error and Nash-Sutcliffe efficiency type metrics. *Water Resources Research*, *47*, W10601. https://doi.org/10.1029/2011WR010962

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 modeling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Jain, S. K., & Sudheer, K. P. (2008). Fitting of hydrologic models: A close look at the Nash-Sutcliffe index. *Journal of Hydrologic Engineering*, *13*(10), 981–986. https://doi.org/10.1061/(ASCE)1084-0699(2008)13:10(981)

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. https://doi.org/10.5194/hess-2019-327

Koppa, A., Gebremichael, M., & Yeh, W. W. G. (2019). Multivariate calibration of large scale hydrologic models: The necessity and value of a Pareto optimal approach. *Advances in Water Resources*, *130*, 129–146. https://doi.org/10.1016/j.advwatres.2019.06.005

Krause, P., Boyle, D. P., & Base, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, *29*(5), 89–97.

Legates, D. R., & McCabe, G. J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, *35*(1), 233–241. https://doi.org/10.1029/1998WR900018

Levick, L., Fonseca, J., Goodrich, D., Hernandez, M., Semmens, D., Stromberg, J., et al. (2008). The ecological and hydrological significance of ephemeral and intermittent streams in the arid and semi-arid American Southwest. U.S. Environmental Protection Agency and USDA/ARS Southwest Watershed Research Center, EPA/600/R-08/134, ARS/233046, 116 pp.

Libera, D. A., Sankarasubramanian, A., Sharma, A., & Reich, B. J. (2018). A non-parametric bootstrapping framework embedded in a toolkit for assessing water quality model performance. *Environmental Modelling and Software*, *107*, 25–33. https://doi.org/10.1016/j.envsoft.2018.05.013

Limbrunner, J. F., Vogel, R. M., & Brown, L. C. (2000). Estimation of the harmonic mean of a lognormal variable. *Journal of Hydrologic Engineering*, *5*(1), 59–66. https://doi.org/10.1061/(ASCE)1084-0699(2000)5:1(59)

Liu, D., Guo, S., Wang, Z., Liu, P., Yu, X., Zhao, Q., & Zou, H. (2018). Statistics for sample splitting for the calibration and validation of hydrological models. *Stochastic Environmental Research and Risk Assessment*, *32*(11), 3099–3116. https://doi.org/10.1007/s00477-018-1539-8

Markstrom, S. L., Regan, R. S., Hay, L. E., Viger, R. J., Webb, R. M. T., Payn, R. A., & LaFontaine, J. H. (2015). PRMS-IV, the precipitation-runoff modeling system. In *version 4: U.S. Geological Survey techniques and methods, book 6* (Chap. B7, p. 158). Reston, VA: U.S. Geological Survey. https://doi.org/10.3133/tm6B

Martinec, J., & Rango, A. (1989). Merits of statistical criteria for the performance of hydrological models. *Water Resources Bulletin*, *25*(2), 421–432. https://doi.org/10.1111/j.1752-1688.1989.tb03079.x

Matalas, N. C., & Langbein, W. B. (1962). Information content of the mean. *Journal of Geophysical Research*, *67*(9), 3441–3448. https://doi.org/10.1029/JZ067i009p03441

Mathevet, T., Michel, C., Andréassian, V., & Perrin, C. (2006). A bounded version of the Nash-Sutcliffe criterion for better model assessment on large sets of basins, large sample basin experiments for hydrological model parameterization. In *Results of the model parameter experiment–MOPEX* (Vol. 307, pp. 211–219). Wallingford, Oxfordshire, UK: IAHS Publication.

McCuen, R. H., Knight, Z., & Cutter, A. G. (2006). Evaluation of the Nash-Sutcliffe efficiency index. *Journal of Hydrologic Engineering*, *11*(6), 597–602. https://doi.org/10.1061/(ASCE)1084-0699(2006)11:6(597)

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed., p. 532). Boca Raton, FL: Chapman and Hall/CRC Press.

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and Its Application*, *6*(1), 355–378. https://doi.org/10.1146/annurev-statistics-031017-100325

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), 885–900. https://doi.org/10.13031/2013.23153

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, *116*(12), 2417–2424. https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models. Part 1: A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

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

Papalexiou, S. M., & Serinaldi, F. (2020). Random fields simplified: Preserving marginal distributions, correlations, and intermittency, with applications from rainfall to humidity. *Water Resources Research*, *56*, e2019WR026331. https://doi.org/10.1029/2019WR026331

Pearson, K. (1896). Mathematical contributions to the theory of evolution III. Regression, heredity and panmixia. *Philosophical Transactions A*, *373*, 253–318.

Pool, S., Vis, M., & Seibert, J. (2018). Evaluating model performance: Towards a non-parametric variant of the Kling-Gupta efficiency. *Hydrological Sciences Journal*, *63*(13–14), 1941–1953. https://doi.org/10.1080/02626667.2018.1552002

Pushpalatha, R., Perrin, C., Le Moine, N., & Andreassian, V. (2012). A review of efficiency criteria suitable for evaluating low-flow simulations. *Journal of Hydrology*, *420*, 171–182.

Rajagopalan, B., Erkyihun, S. T., Lall, U., Zagona, E., & Nowak, K. (2019). A nonlinear dynamical systems-based modeling approach for stochastic simulation of streamflow and understanding predictability. *Water Resources Research*, *55*, 6268–6284. https://doi.org/10.1029/2018WR023650

Reusser, D. E., Blume, T., Schaefli, B., & Zehe, E. (2009). Analysing the temporal dynamics of model performance for hydrological models. *Hydrology and Earth System Sciences*, *13*(7), 999–1018. https://doi.org/10.5194/hess-13-999-2009

Ritter, A., & Munoz-Carpena, R. (2013). Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. *Journal of Hydrology*, *480*(3), 33–45. https://doi.org/10.1016/j.jhydrol.2012.12.004

Santos, L., Thirel, G., & Perrin, C. (2018). Technical note: Pitfalls in using log-transformed flows within the KGE criterion. *Hydrology and Earth System Sciences*, *22*(8), 4583–4591. https://doi.org/10.5194/hess-22-4583-2018

Schaefli, B., & Gupta, H. V. (2007). Do Nash values have value? *Hydrological Processes*, *21*(15), 2075–2080. https://doi.org/10.1002/hyp.6825

Shimizu, K. (1993). A bivariate mixed lognormal distribution with an analysis of rainfall data. *Journal of Applied Meteorology*, *32*(2), 161–171. https://doi.org/10.1175/1520-0450(1993)032<0161:ABMLDW>2.0.CO;2

Stedinger, J. R. (1980). Fitting lognormal distributions to hydrologic data. *Water Resources Research*, *16*(3), 481–490. https://doi.org/10.1029/WR016i003p00481

Stedinger, J. R. (1981). Estimating correlations in multivariate streamflow models. *Water Resources Research*, *17*(1), 200–208. https://doi.org/10.1029/WR017i001p00200

Todini, E. (2011). History and perspectives of hydrological catchment modeling. *Hydrology Research*, *42*(2–3), 73–85. https://doi.org/10.2166/nh.2011.096

Todini, E. (2017). Predictive uncertainty assessment and decision making. In V. P. Singh (Ed.), *Chap 26 in Handbook of applied hydrology* (pp. 26-1–26-16). New York: McGraw Hill.

Todini, E. & Biondi, D. (2017). Calibration, parameter estimation, uncertainty, data assimilation, sensitivity analysis, and validation. In V. P. Singh (Ed.), *Chap 22 in Handbook of applied hydrology* (pp. 22-1–22-19). New York: McGraw Hill.

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., & Fennessey, N. M. (1993). L-moment diagrams should replace product-moment diagrams. *Water Resources Research*, *29*(6), 1745–1752. https://doi.org/10.1029/93WR00341

Vogel, R. M., Stedinger, J. R., & Hooper, R. P. (2003). Discharge indices for water quality loads. *Water Resources Research*, *39*(10), 1273. https://doi.org/10.1029/2002WR001872

Wallis, J. R., Matalas, N. C., & Slack, J. R. (1974). Just a moment! *Water Resources Research*, *10*(2), 211–219. https://doi.org/10.1029/WR010i002p00211

World Meteorological Organization (WMO) (1986). *Intercomparison of models of snowmelt runoff operational hydrology report No. 23*. Geneva: World Meteorological Organization.