

Evaluating statistical goodness-of-fit estimators for skewed hydrologic data

A Dissertation submitted by
Caitline A. Barber

in partial fulfilment of the requirements for the degree of
Master of Science
in

Civil and Environmental Engineering

TUFTS UNIVERSITY

February 2020

© 2019, Caitline A. Barber

Advisor: Jonathan R. Lamontagne

ABSTRACT

Model evaluation is necessary in order to determine how well a model performs and therefore how much the model output can be trusted to be an accurate representation of the modeled system's response to imposed conditions. In hydrology, the common goodness-of-fit metrics that are used to evaluate model performance include Pearson's correlation estimator and the Nash-Sutcliffe efficiency (NSE) estimator. These goodness-of-fit metrics provide an accurate assessment of model performance under specific conditions, such as when the data arise from a normal distribution. High frequency hydrologic data, however, is often skewed and more closely resembles a log normal distribution. Alternative goodness-of-fit metrics of correlation and efficiency are developed specifically for skewed hydrologic data. Monte Carlo analyses are then employed to evaluate the performance of these newly developed estimators against other estimators commonly used for model evaluation in hydrology. The results of the Monte Carlo analyses demonstrate that the correlation estimator developed for skewed hydrologic data is a significant improvement over Pearson's correlation estimator. Similarly, the newly proposed efficiency estimators perform much better than NSE, particularly for large sample sizes.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation for the support and guidance provided by my advisory committee: Jonathan Lamontagne, Richard Vogel, and Peter Weiskel. I could not have anticipated all that I would learn in my short time at Tufts, and I am forever grateful for the patience and encouragement they provided through the challenges that we faced.

I would also like to thank those that I have shared an office with: Flannery Dolan, Ghazal Shabestanipou, LiChen Wang, Ashkan Akhlaghi, Azin Mehrjoo, and Nasim Partovi. Thank you for making our office the best one could ask for! This experience would not have been the same without you.

And finally, thank you to my family. Without you, I would not have accomplished this milestone.

TABLE OF CONTENTS

ABSTRACT	i
ACKNOWLEDGEMENTS	II
TABLE OF CONTENTS	III
LIST OF TABLES	V
LIST OF FIGURES	VI
INTRODUCTION	1
CHAPTER 1	4
LITERATURE REVIEW	4
1 INTRODUCTION	4
2 CORRELATION AND PEARSON'S ESTIMATOR	5
2.1 <i>Limitations of Pearson's estimator</i>	6
2.2 <i>A proposed modification to Pearson's estimator</i>	8
2.3 <i>Alternative methods of estimating correlation</i>	9
3 EFFICIENCY AND NSE	13
3.1 <i>Limitations of NSE</i>	13
3.2 <i>Alternative efficiency estimators</i>	15
REFERENCES	18
CHAPTER 2	22
IMPROVED ESTIMATORS OF CORRELATION AND R^2 FOR SKEWED HYDROLOGIC DATA	22
ABSTRACT	23
1 INTRODUCTION	24
1.1 <i>R^2 and correlation coefficient</i>	25
1.2 <i>Application of R^2 and Pearson correlation coefficient ρ in hydrology</i>	26
1.3 <i>Performance of R^2 and r under bivariate normality</i>	28
1.4 <i>Sampling properties of R^2 and r under bivariate non-normality</i>	28
2 STUDY ASSUMPTIONS: BIVARIATE LOGNORMAL MODEL OF HYDROLOGIC DATA	32
3 ALTERNATIVE ESTIMATORS OF R^2 AND ρ FOR NON-NORMAL BIVARIATE HYDROLOGIC DATA	34
3.1 <i>The copula and the dependence structure of hydrologic variables</i>	35
3.2 <i>Stedinger's (1981) lognormal estimator, r_1</i>	37
3.3 <i>Modified Spearman rank correlation estimator, r_2</i>	39
3.4 <i>A modified rank inverse normal correlation estimator, r_3</i>	41
4 MONTE CARLO EXPERIMENTS	42
5 EVALUATIONS USING ACTUAL BIVARIATE STREAMFLOW OBSERVATIONS	45
5.1 <i>Bivariate PRMS streamflow simulations and observations</i>	45
5.2 <i>Goodness-of-fit of bivariate lognormal model to bivariate observations</i>	46
5.3 <i>Comparisons among four correlation estimators</i>	52
5.4 <i>Impact of skewness on the sampling properties of Pearson's r</i>	55
6 CONCLUSIONS	57
REFERENCES	60
APPENDIX. GENERATION OF BIVARIATE LOGNORMAL STREAMFLOW SERIES	64

IMPROVED ESTIMATORS OF EFFICIENCY FOR GOODNESS-OF-FIT OF SKEWED HYDROLOGIC DATA 66

ABSTRACT	67
1 INTRODUCTION	68
2 THEORETICAL DEVELOPMENT OF EFFICIENCY E	74
2.1 <i>Theoretical Efficiency Based on Nash-Sutcliffe</i>	74
2.2 <i>Another Definition of Theoretical Efficiency E' Based on Kling-Gupta</i>	76
3 STUDY ASSUMPTIONS: DAILY STREAMFLOW SIMULATIONS AND OBSERVATIONS	78
3.1 <i>Bivariate Three Parameter Lognormal Model (BLN3)</i>	80
3.2 <i>A BLN3 Monthly Mixture Model for Reproduction of the Probability Distribution of Daily Streamflow</i>	81
4 SAMPLE ESTIMATORS OF EFFICIENCY	84
4.1 <i>Nash-Sutcliffe Efficiency (NSE)</i>	84
4.2 <i>Natural Log Nash-Sutcliffe Efficiency (LNSE)</i>	84
4.3 <i>Kling-Gupta Efficiency (KGE')</i>	84
4.4 <i>Non-Parametric Efficiency (PVSE')</i>	85
4.5 <i>BLN3 Estimators of Efficiency (LBE and LBE')</i>	86
4.6 <i>BLN3 Mixture Estimators of Efficiency (LBE_m and LBE'_m)</i>	88
5 EXPERIMENTAL RESULTS	90
5.1 <i>Monte-Carlo Experiments</i>	91
5.2 <i>Monte Carlo Experiment Results</i>	92
5.3 <i>Comparison of Sample Estimates of E and E'</i>	96
5.4 <i>The Causes of Variability in Efficiency Estimates</i>	101
6 DISCUSSION	103
6.1 <i>Handling Zero Streamflows</i>	104
6.2 <i>Improved Mixture Model</i>	104
7 CONCLUSIONS	105
REFERENCES	109
FUTURE WORK	113
1 GRAND CHALLENGES REMAINING	113
2 SPECIFIC EXPERIMENTS	114
2.1 <i>Further improvements for correlation and efficiency estimators</i>	114
2.2 <i>Robustness analysis of correlation and efficiency estimators</i>	115
2.3 <i>Impacts of sample size and parameters on efficiency estimators</i>	117
2.4 <i>Analysis of efficiency estimators for sub-daily hydrologic data</i>	118

LIST OF TABLES

Table 2.1 Statistics of Streamflow Records of 905 sites across the continental U.S. (see Farmer and Vogel, 2016ab).....	46
Table 3.1 Values of n , C_0 , C_s , α , Δ , ρ , LBE_m , NSE , LBE'_m and KGE Corresponding to Daily Streamflow Observations and Simulations at 447 USGS Gaged Watersheds.....	79
Table 3.2 Estimators of Efficiency and Various Components of Efficiency Based on BLN3 Monthly Mixture Model for 8 selected sites.....	98

LIST OF FIGURES

Figure 2.1 Results of the Monte Carlo experiments are illustrated using boxplots of the four estimators of correlation ρ considered: Pearson r , Stedinger r_1 , modified Spearman r_2 and modified RIN r_3 . Boxplots summarize the sampling distribution of estimators of ρ for synthetic streamflows generated from a bivariate lognormal model with $\rho = 0.7$ (left panels) and $\rho = 0.9$ (right panels), for three different values of coefficient of variation $Co=C_s$ 44

Figure 2.2 Goodness-of-fit evaluation of bivariate normality hypothesis with 50th and 95th confidence interval ellipses drawn. Upper left panel shows synthetic bivariate normal data and the remaining 5 panels are representative sites from the dataset. 48

Figure 2.3 L-moment diagrams of observed and simulated average daily streamflows at the 905 sites summarized in Table 1. 50

Figure 2.4 Evaluation of the ability of the four correlation estimators applied to the 905 sites summarized in Table 1, to reproduce the theoretical dependence structure associated with a bivariate LN3 process given by Equation 5 52

Figure 2.5 Comparison of Pearson’s correlation estimator with the three alternative correlation estimators (Left) using observations and simulations at the 905 sites summarized in Table 1 and (Right) using synthetic bivariate lognormal series generated to reproduce the characteristics of the 905 sites summarized in Table 1. 54

Figure 2.6 The expected difference between Pearson’s estimator r , and Stedinger’s estimator r_1 , as a function of the coefficient of variation of the observations. The values of $E[r - r_1]$ are computed from 500 synthetic bivariate lognormal traces generated to reproduce the characteristics of the bivariate observations and simulations at the 905 sites summarized in Table 1. 56

Figure 3.1 Ratio of Efficiency E Based on NSE to the Efficiency E' Based on KGE’ as a Function of α and ρ , or an unbiased model. 77

Figure 3.2 The Square of the Probability Plot Correlation Coefficients for the BLN3 monthly mixture model $PPCC_m^2$ and a single BLN3 model $PPCC_{LN3}^2$ 83

Figure 3.3 Boxplots of Estimates of Efficiency E Resulting From 1,000 Monte Carlo Experiments Performed at Each of the 447 Sites Summarized in Table 3.1 93

Figure 3.4 Boxplots of Estimates of Efficiency E' Resulting From 1,000 Monte Carlo Experiments Performed at Each of the 447 Sites Summarized in Table 3.1 94

Figure 3.5 Scatter plot of values of $u_i = \ln[o_i - \hat{t}_o]$ versus $v_i = \ln[s_i - \hat{t}_s]$ for 8 sites summarized in Table 3.2 along with results for synthetic data in the lower right plot 97

Figure 3.6 Scatterplots of estimates of E obtained from various estimators, with results from 8 sites in Table 3.2 shown using dark black circles. 100

Figure 3.7 Scatterplots of estimates of E' obtained from various estimators, with results from 8 sites in Table 3.2 shown using dark black circles. 101

Figure 3.8 Root mean squared error of NSE and KGE vs C_0 , ρ , Δ , and α . RMSE of NSE and KGE calculated based on 1,000 Monte Carlo replicates of sample length $n=10,950$ (30 years) for each of the 447 sites.....103

Introduction

Models are used to advance scientific understanding of systems and are increasingly used as a tool to provide essential information for decision making. They assist decision makers by demonstrating how a system will respond under certain conditions. Given the increasing reliance on models, it is necessary to assess how well a model reproduces the observed system response in order to understand how well the model will perform under conditions not observed historically, and therefore how confident one can be in model predictions. To determine how well a model performs, a variety of goodness-of-fit metrics can be computed. Each of these metrics, however, have advantages, disadvantages, and make underlying assumptions about the data. If these underlying assumptions are not met, the metrics may perform poorly. Therefore, in order to accurately understand how well a model performs, it is critical to have a firm understanding of which goodness-of-fit metrics can appropriately be applied.

In hydrology it is common to evaluate model performance using either Pearson's correlation or an efficiency estimator such as the Nash-Sutcliffe Efficiency (NSE) or Kling-Gupta Efficiency (KGE). The statistical properties of these estimators, such as unbiasedness, under certain conditions have resulted in their widespread use in the hydrology community. It is important, however, to assess under which conditions these attractive properties break down in order to avoid over confidently trusting a model due to its reported goodness-of-fit metric. In this dissertation, various correlation and efficiency metrics are evaluated based on their performance when applied to daily historical observations and model simulations generated from a distributed-parameter, precipitation-runoff model. A new estimator of correlation and efficiency are proposed that have been developed specifically for skewed hydrologic data.

This dissertation begins with a thorough literature review of the relevant research that has been conducted on Pearson's correlation estimator and NSE. Limitations of these estimators are discussed and alternative estimators that have been suggested in the literature are reviewed.

The second chapter examines four estimators of correlation, including the frequently used Pearson's correlation. The performance of the estimators under bivariate lognormal conditions is determined for varying sample sizes, correlation, and coefficient of variation through a Monte Carlo analysis. Following the Monte Carlo analysis, the degree to which the assumption of bivariate lognormally distributed data holds is analyzed. Given the widespread use of correlation as a goodness-of-fit metric for high frequency hydrologic data, conclusions regarding the bias and variance of various estimators under these conditions provides tremendous insight as to which estimators are most appropriate under these conditions.

Expanding on the second chapter, the third chapter investigates the performance of efficiency estimators for high frequency hydrologic data. Improving upon the bivariate lognormal model used in the second chapter, a bivariate lognormal mixture model is employed to more accurately capture the relationship between the model observations and simulations. Monte Carlo analysis is conducted to provide a rigorous examination of the performance of the various estimators. Similar to the second chapter, the findings presented increase the understanding of the limitations of commonly used efficiency estimators and provide an evaluation of a newly proposed estimator.

The final section of this dissertation outlines the potential areas for future work that would expand this field of research. While improved estimators of correlation and efficiency have been developed for skewed hydrologic data, there is much more research needed to understand the how

these estimators perform when underlying assumptions about the data are violated. Also, improvements upon commonly used estimators may be possible.

Due to the reliance on models to aid both scientific understanding and decision making, it is crucial to understand the statistical estimators that are used to evaluate the performance of these models. Without a rigorous, nuanced understanding of these goodness-of-fit statistics the conclusions drawn from models may be misleading under certain conditions. This dissertation focuses on evaluating the performance of commonly used goodness-of-fit statistics for skewed hydrologic data to more accurately capture model performance.

Chapter 1

Literature Review

1 Introduction

Statistical models are used to help us understand and predict the response of a system to changes in model inputs. For each model developed, goodness-of-fit metrics are applied in order to understand how well the simulated model output matches the measured data used during the calibration process. Goodness-of-fit metrics are critical to evaluating how well the designed model performs and are frequently used as indicators of how much the model can be trusted.

In the field of hydrology, it is common to model streamflow in order to help inform decisions that affect water resources. To assess how well the model performs, it is common to report either the Pearson correlation coefficient or the Nash-Sutcliffe Efficiency (NSE) as a goodness-of-fit metric between the model simulations and observations. Common estimators for these statistics are parametric and carry implicit assumptions about the underlying distribution of the data. The purpose of this research is to evaluate whether these estimators of goodness-of-fit are appropriate for the hydrologic data that they are often applied to.

High frequency hydrologic data is made readily available by agencies such as the United States Geological Survey (USGS). Daily, hourly, and even sub-hourly data can be used to generate model simulations. In order to appropriately use goodness-of-fit metrics, it is important to understand the underlying distribution of the model simulations and observations. Studies that have evaluated the appropriateness of various distributions to high-frequency hydrologic data have found that the two-parameter and three-parameter lognormal distributions provide a good first approximation to daily streamflow observations (Blum, Archfield, & Vogel, 2017; Limbrunner, Vogel, & Brown, 2000).

This literature review provides distinctions between the theoretical statistics of correlation and efficiency and their commonly used sample estimators of Pearson's correlation coefficient and the Nash-Sutcliffe efficiency estimator. A review of the limitations of these sample estimators is discussed, followed by a brief discussion of alternative estimators that have been proposed to overcome their limitations. The relevant literature discussed provides a basis for the chapters that follow.

2 Correlation and Pearson's estimator

Correlation is a measure of the degree of dependence between two variables. It is a theoretical statistic with probabilistic properties. The theoretical definition of correlation, denoted by ρ , between two variables O and S is given by:

$$\rho = \frac{\text{cov}(O, S)}{\sqrt{\text{var}(O)\text{var}(S)}} = \frac{E[(O - \mu_O)(S - \mu_S)]}{\sigma_O \sigma_S} \quad (1.1)$$

Because the true population mean μ and standard deviation σ of the data are never known, estimates of these moments must be made in order to calculate the correlation between two variables. The decision of which estimators of mean and standard deviation to use impose assumptions regarding the underlying distribution of the data.

One commonly used estimator of correlation is Pearson's correlation estimator which employs product moment estimators of mean μ and standard deviation σ . Pearson's correlation estimator, denoted by r , is given by:

$$r = \frac{\frac{1}{n} \sum_{i=1}^n (s_i - \bar{s})(o_i - \bar{o})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (o_i - \bar{o})^2 \frac{1}{n} \sum_{i=1}^n (s_i - \bar{s})^2}} \quad (1.2)$$

Pearson's correlation is based upon a number of assumptions which have been thoroughly discussed in the statistical literature by Havlicek and Peterson (1976), Horowitz (1974), Guilford and Fruchter (1973), Bradley (1968), Siegel (1956) and others. The assumption that is central to this dissertation is the assumption that the data are bivariate normally distributed.

2.1 Limitations of Pearson's estimator

As Pearson's correlation estimator became more widespread, increasing focus was put on understanding its limitations for hydrologic applications and, subsequently, improving upon it. Some studies have criticized Pearson's correlation estimator and sought to improve it without referring to the original theoretical definition of correlation (D. N. Moriasi et al., 2007; Krause, Boyle, & Bäse, 2005; Legates & Davis, 1997; Legates & McCabe, 1999; McCuen & Snyder, 1975; Willmott, 1981; Willmott et al., 1985). As a result, these studies have broadly criticized the theoretical correlation and in some cases have developed new estimators which may have no valid mathematical relationship with the theoretical statistic.

Since the late 1920s, researchers have investigated the robustness of the distribution of Pearson's correlation estimator when its underlying assumptions are violated. Kowalski (1972) reviewed these studies and found that there was a "dichotomy of opinion". Many previous studies had showed that the distribution of Pearson's estimator was in fact robust to violations of the normality assumption (Cheriyann, 1945; Dunlap, 1931; Hey, 1938; Nair, 1941; Pearson, 1929; Rider, 1932), however many others showed the opposite (Baker, 1930; Chesire, Oldis, & Pearson, 1932; Gayen, 1951; Haldane, 1949). In response, Kowalski performed a Monte Carlo analysis to determine the robustness of the distribution of Pearson's r to non-normal data. He found that the distribution of Pearson's estimator is sensitive to non-normally distributed data, and therefore its use should be restricted to cases for which the assumption of normality is not violated. Although

Kowalski wrote that he hoped the previous decades of conflicting research regarding the widespread and indiscriminate use of Pearson's estimator would conclude and a consensus regarding the need for alternative goodness-of-fit measures would be reached as a result of his research, the debate carried on and the use of Pearson's estimator continued.

An example of this is came just four years following Kowalski's research. In response to conflicting advice in the literature regarding the necessity of meeting the underlying assumptions of Pearson's r to obtain accurate correlation measures between two variables, Havlicek and Peterson (1976) performed Monte Carlo analyses to determine the impact on Pearson's r when the basic assumptions of normality are violated. They concluded that Pearson's r is not sensitive to the shape of the distributions of the variables being correlated. Notably, however, they did not consider the situation in which both variables arise from negatively skewed data, as is often the case for high frequency hydrologic data.

Similar to Havlicek and Peterson (1976), nearly a decade later Willmott et al. (1985) discussed the conflicting recommendations made from the model evaluation literature. Based on their analysis, they recommended using a root-mean-square error approach to assessing the goodness-of-fit of a model.

Numerous researchers have documented the significant impact that large outliers can have on correlation including Moore (1991) and Legates and Davis (1997). As discussed by Legates and McCabe (1999), the considerable influence that outliers have over correlation-based measures can lead to inflated measures of correlation. The finding that outliers can lead to an overestimation of Pearson's r is certainly an important limitation of the estimator, however the studies cited here fail to distinguish between Pearson's r and the theoretical correlation statistic. The theoretical statistic cannot be influenced by outliers because it is a population value and does not depend on

data. Only when data is introduced, as with Pearson's r , can the estimator of correlation exhibit over sensitivity to outliers.

Furthering the discussion on the appropriateness of applying Pearson's estimator of correlation to non-normal data, and recognizing the distinction between a sample estimate of correlation and the population correlation, Hutchinson (1997) suggested that Pearson's estimator is a poor estimator of the population correlation.

2.2 A proposed modification to Pearson's estimator

McCuen and Snyder (1975) found that Pearson's correlation estimator is only an accurate measure of the goodness-of-fit between two hydrographs when the hydrographs have similar means and variances. If these moments are not similar, the hydrographs may have different sizes but maintain a similar shape. In this situation, Pearson's estimator simply indicates the similarity of the shape of the hydrographs and does not reflect differences in their size. As a measure of goodness-of-fit between two models, simply measuring the similarity of the shape of their hydrographs is not adequate. As a result, McCuen and Snyder sought to improve upon Pearson's estimator so that the estimator reflects differences in both the shape and the size of the hydrographs.

To start, McCuen and Snyder set desired characteristics for their new estimator. Similar to Pearson's estimator, they sought to create an estimator which has limits on its magnitude and approaches a null value as dissimilarities in either size or shape, or both, increase. To accomplish their goal, McCuen and Snyder created an estimator which adjusts the Pearson's correlation coefficient downwards when the two hydrographs in question differ in size, and the scale of this adjustment is proportional to the difference in size. Their adjustment factor is based on the sum of

squares of the two hydrographs X and Y , such that the modified correlation, denoted r_m , is proportional to Pearson's estimator, r , through the following relationship:

$$r_{m\ xy} = \frac{A}{B} r_{xy} \quad (1.3)$$

where A and B are the rescaling factor for hydrographs X and Y , respectively, based on the sum of squares.

Of significance to the modification of Pearson's correlation estimator proposed by McCuen and Snyder is the fact that at no point did they verify that their modified estimator of correlation is in fact still an estimator of the theoretical correlation ρ given in (1.1). Their proposed estimator r_m is therefore not guaranteed to estimate the correlation between X and Y at all. Without a thorough analysis to ensure that a proposed modification maintains the consistency between the theoretical statistic and its estimator, the modified statistic may not simply be a modified estimator, but be estimating something else entirely.

2.3 Alternative methods of estimating correlation

The limitations of Pearson's correlation estimator discussed in the previous sections arise from a violation of the assumption that the underlying data is bivariate normally distributed. Nonparametric estimators avoid this limitation because they do not make any assumptions regarding the underlying distribution of the data. Both Spearman's and Kendall's estimators of correlation are reviewed. Another method to address the potential violation of the assumption of normality is to first apply a nonlinear transformation to the data so that it can then be approximated by a normal distribution. The rank-inverse and Box-Cox transformations are discussed.

2.3.1 Nonparametric estimators

Spearman's correlation estimator simply applies Pearson's estimator to the ranks of the data (Spearman, 1904). This nonparametric approach is attractive to use when the underlying data

cannot confidently be assumed to arise from a normal distribution. Concerns, however, have been raised regarding Spearman's estimator. Spearman's estimator has been recognized as an inherently biased statistic (Daniels, 1950; Hoeffding, 1948). It has also been noted that the expected value of Spearman's estimator generally does not reflect the population value of correlation, and that this bias is partly attributable to sample size (Arndt, Turvey, & Andreasen, 1999; Cliff, 1996; M. Kendall & Gibbons, 1990). Of significant importance to correctly understanding and interpreting Spearman's estimator is the discussion provided by Astivia and Zumbo (2017). As they discuss, a common misinterpretation of Spearman's estimator exists in the literature in which Spearman's estimator is assumed to be a substitute estimator for the population correlation defined in Eq. 1, when in fact it is an estimate of a different population correlation entirely. They show that necessary transformations, such as a Gaussian copula transformation, is needed in order to directly compare Spearman's estimator with Pearson's estimator of correlation.

Kendall's rank correlation coefficient, often referred to as Kendall's τ (M. Kendall 1938), is similar to Spearman's estimator in that it estimates correlation based on the ranks of the data, however important distinctions exist between the two estimators. The calculation of Kendall's τ does not depend on the ranks themselves (as with Spearman's estimator), but on the number of concordant and discordant pairs (Conover, 1999). If two observations (a,b) and (c,d) are compared, they are discordant if $a < c$ and $b > d$, or conversely if $a > c$ and $b < d$. The observations are concordant if $a > c$ and $b > d$, or if $a < c$ and $b < d$. Kendall's τ is given by the following equation for the situation in which no ties exist:

$$\tau = \frac{N_c - N_d}{n(n - 1)/2} \quad (1.4)$$

where N_c and N_d are the total number of concordant and discordant pairs, respectively. Kendall's τ can be interpreted as the probability of observing concordant and discordant pairs (Conover,

1999), or in other words, the probability that as the independent variable increases the dependent variable will also increase (Arndt et al., 1999). It therefore measures the monotonic relationship between the two variables in question (Helsel & Hirsch, 1992). Generally, Kendall's τ yields smaller correlation estimates than Spearman's estimator (Conover, 1999; Helsel & Hirsch, 1992), however the two estimators of correlation differ in their interpretation and metric and therefore cannot be compared directly (Arndt et al., 1999). As a result of their rank-based procedures, both Kendall's τ and Spearman's estimator are resistant to effects from outliers and can capture non-linear correlations (Helsel & Hirsch, 1992; Maidment, 1993). Unlike Spearman's estimator, Kendall's τ does not take into account the magnitude of the differences observed between data pairs (Helsel & Hirsch, 1992).

2.3.2 Nonlinear data transformations

Nonlinear data transformations can be an important tool in order to ultimately provide a more accurate estimate of the correlation between two variables. These transformations can decrease the influence of outliers while also producing greater marginal normality and linearity by altering the shapes of variable distributions (Bishara & Hittner, 2012). After applying a nonlinear data transformation, Pearson's estimator may be suitable to use because the data no longer violate the key underlying assumption of normality.

The rank-based inverse normal (RIN) transformation has also been shown to provide an improved estimate of correlation for non-normal data (Bishara & Hittner, 2015; Puth, Neuhäuser, & Ruxton, 2014). First introduced by Fisher and Yates (1963), this transformation can be very useful because it can approximately normalize any distribution shape (Beasley, Erickson, & Allison, 2009; Bishara & Hittner, 2012). Generally speaking, the RIN method converts the data to ranks, then converts the ranks to probabilities, and finally converts the probabilities into an

approximately normal shape by using the inverse cumulative normal function (Bishara & Hittner, 2012). As Beasley, Erickson, and Allison (2009) describe, there are many variants of this general methodology that can be implemented. All methods involve first converting a variable to ranks. Following this preliminary step, RIN transformations can be subdivided into ones that involve a stochastic element and ones that are deterministic. Deterministic RIN transformations can yet again be subdivided into those that use expected normal scores and those that use sample quantiles to approximate the expected normal scores. Bishara and Hittner (2012) note that the RIN transformation may be a useful and widely applicable method for assessing bivariate associations such as correlation.

Another nonlinear transformation that can be applied is the Box-Cox transformation (Box & Cox, 1964). This transformation consists of a family of power transformations that are particularly useful for transforming skewed data into a more normal distribution (Bishara & Hittner, 2012). The Box-Cox transformation is defined as

$$g(x, \lambda) = \begin{cases} \ln(x) & \text{if } \lambda = 0 \\ \frac{x^\lambda - 1}{\lambda} & \text{else} \end{cases} \quad (1.5)$$

This type of transformation depends on the value of the parameter λ . A linear transformation results from a value of $\lambda = 1$, while a convex and concave function result from $\lambda > 1$ and $\lambda < 1$, respectively (Bishara & Hittner, 2012). Undefined values can result for $x < 0$, so often times a constant is added to the data prior to applying the transformation (Bishara & Hittner, 2012). The Box-Cox transformation may provide an improved estimate of correlation, however it can be significantly influenced by the addition of an arbitrary constant (Dougherty, Thomas, Brown, Chrabaszcz, & Tidwell, 2015).

3 Efficiency and NSE

Efficiency is a standardized form of the mean squared error (*MSE*) and, similar to correlation, it is a theoretical statistic governed by probability theory. Efficiency is a preferred goodness-of-fit metric because it captures both the bias and variance associated with the data. Theoretical efficiency, referred to here as E , is given by:

$$E = 1 - \frac{MSE}{E[(O - \mu_o)^2]} = 1 - \frac{MSE}{\sigma_o^2} \quad (1.6a)$$

where

$$MSE = E[(S - O)^2] \quad (1.6b)$$

Analogous to correlation, in order to compute an estimate of this theoretical statistic based on data one must choose estimators of *MSE* and variance.

According to Todini and Biondi (2017), the most widely used estimator of efficiency in hydrology is the Nash-Sutcliffe Efficiency (*NSE*), given by:

$$NSE = 1 - \frac{\overline{MSE}}{s_o^2} = 1 - \frac{\frac{1}{n} \sum_{i=1}^n (s_i - o_i)^2}{\frac{1}{n-1} \sum_{i=1}^n (o_i - \bar{o})^2} \quad (1.7a)$$

where

$$\overline{MSE} = \frac{1}{n} \sum_{i=1}^n (s_i - o_i)^2 \quad (1.7b)$$

Again, as with Pearson's correlation estimator, the calculation of *NSE* implies certain assumptions about the data such as the data being independent and identically distributed.

3.1 Limitations of NSE

As the use of the Nash-Sutcliffe efficiency estimator has become a standard goodness-of-fit metric to evaluate hydrologic models, research has increasingly focused on evaluating its use and

identifying its limitations. Even as research continues to show that *NSE* is not an appropriate goodness-of-fit statistic in certain situations, it maintains its prominence as a standard metric from which to evaluate the accuracy of models.

Martinec and Rango (1989) demonstrated the enormous sensitivity of *NSE* to methods of calculating moments such as the average streamflow. They found that when seasons or years with extreme high or low flow data are being evaluated, it is important to calculate \bar{o} in (1.7a) based on the year or period in question, not on a long-term average over the entire period of record. Failure to calculate \bar{o} from the period in question, and instead calculate it as a long-term average, may result in significantly overestimating the performance of the model in the specific year or season in question. Based on their results, they find that for an unbiased model ($E = \rho^2$), the arithmetic mean of *NSE* calculated in each period of interest (using \bar{o} from that period of interest) gives a more realistic evaluation of model performance than when *NSE* is calculated as a single metric from the entire period of record (using \bar{o} from the entire period of record). Their work essentially employs a mixture estimator of *NSE* and finds that the mixture estimator provides a more accurate measure of model performance.

McCuen, Knight, and Cutter (2006) noted that due to the influence of many factors, a high value of *NSE* may be calculated even when the fit of a model is relatively poor. They also discuss how the value of *NSE* can depend greatly on the sample size. As the sample size increases, *NSE* becomes a better estimator of the population value of efficiency. They propose a hypothesis test in order to systematically assess efficiency values and ultimately avoid the potential for subjective conclusions regarding the goodness-of-fit of a model. Based on the sensitivity of *NSE* to factors such as sample size, outliers, and bias, they ultimately conclude that *NSE* can be a useful index,

but careful attention must be given in order to avoid misapplying *NSE* to situations that it is not suited for.

Criss and Winston (2008) emphasized that *NSE* does not measure how well a model performs in absolute terms, but rather is in reference to a model that uses the mean value of the observations as its prediction. This is the result of the fact that *NSE* is a normalized measure of mean square error which results in a value of 1.0 if there is perfect model performance, a value of zero if the model on average performs as good as simply using the mean of the observations, and less than zero if the model performs worse than simply assuming a predicted value equivalent to the mean of the observations. They demonstrate that *NSE* can lead to misleading interpretations depending on the frequency and strength of seasonality in the data. Based on their findings, Criss and Winston (2008) recommend establishing benchmark models that would provide an appropriate baseline measure of *NSE* for different types of studies.

Jain and Sudheer (2008) illustrate some of the limitations of *NSE* through the consideration of four case studies. They show that high values of *NSE* can be obtained even when visual analysis of model simulations and observations reveals that the model does a poor job of matching the observations. Additionally, they find that even if a model appears to provide a near perfect fit based on the *NSE* value, analysis of the residuals may reveal that the model does a poor job of capturing the structure of the data.

3.2 *Alternative efficiency estimators*

An increasingly popular goodness-of-fit metric for hydrologic models is the Kling-Gupta efficiency (*KGE*) estimator (Gupta, Kling, Yilmaz, & Martinez, 2009). In their decomposition of *NSE*, they find that *NSE* can be represented in terms of just three components: correlation, bias, and variability of flow. They also find that *NSE* does not weight these components equally. This

unequal weighting can lead to unintended consequences if optimizing a model based on *NSE*. To remedy this situation, *KGE* was developed specifically to provide an equal weighting of the correlation, bias, and variability measures. Gupta et al. (2009) introduce *KGE* as the following:

$$KGE = 1 - \sqrt{(r - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \quad (1.8)$$

where r is Pearson's correlation estimator, $\alpha = \sigma_S/\sigma_O$, and $\beta = \mu_S/\mu_O$. Product-moment estimators of mean and standard deviation are utilized for estimation of *KGE*. Santos, Thirel, and Perrin (2018) have shown that numerical flaws are introduced if *KGE* is applied to the logarithmic transformations of the data and it is therefore suggested that this transformation should not be applied if using *KGE* as a goodness-of-fit metric.

To improve upon the available estimators of efficiency, Pool, Vis, and Seibert (2018) proposed an estimator of efficiency that is based on non-parametric components. They note that *KGE* uses estimators of correlation, α , and β that implicitly assume that the data are normally distributed. Therefore, *KGE* may not be suitable for skewed data such as high frequency hydrologic data. The estimator proposed by Pool, Vis, and Seibert (2018) builds off of the estimator introduced by Gupta et al. (2009) by utilizing nonparametric estimators of correlation and α . Their estimator is introduced as:

$$PVSE = 1 - \sqrt{(r_S - 1)^2 + (\alpha_{NP} - 1)^2 + (\beta - 1)^2} \quad (1.9)$$

where r_S is the Spearman rank correlation, $\alpha_{NP} = 1 - \frac{1}{2} \sum_{k=1}^n \left| \frac{S(k)}{n\bar{s}} - \frac{o(k)}{n\bar{o}} \right|$, and $\beta = \mu_S/\mu_O$ where the means are product moment estimators. The nonparametric form of α used in this estimator is based on a normalized flow-duration curve. Based on their initial analysis of *PVSE*, Pool, Vis, and Seibert (2018) conclude that their nonparametric modification to *KGE* generally performs better than *KGE* except when evaluating the magnitude and timing of high flows.

To address the sensitivity of *NSE* to extreme values, some have calculated *NSE* can be calculated based on the natural log of the observations and simulations. This logarithmic transformation results in a flattening of the peak flows while maintaining relatively the same level of the low flows. This causes the influence of the low flows to increase in comparison to the high flows (Krause et al., 2005). Log-transformed discharge is often used in order to provide a goodness-of-fit metric that specifically focuses on low-flows (Santos et al., 2018).

References

- Arndt, S., Turvey, C., & Andreasen, N. C. (1999). Correlating and predicting psychiatric symptom ratings: Spearman's ρ versus Kendall's tau correlation. *Journal of Psychiatric Research*, 33(2), 97–104. [https://doi.org/10.1016/S0022-3956\(98\)90046-2](https://doi.org/10.1016/S0022-3956(98)90046-2)
- Astivia, O. L. O., & Zumbo, B. D. (2017). Population models and simulation methods: The case of the Spearman rank correlation. *British Journal of Mathematical and Statistical Psychology*, 70(3), 347–367. <https://doi.org/10.1111/bmsp.12085>
- Baker, G. A. (1930). The Significance of the Product-Moment Coefficient of Correlation with Special Reference to the Character of the Marginal Distributions. *Journal of the American Statistical Association*, 25(172), 387–396. Retrieved from <http://www.tandfonline.com/doi/abs/10.1080/01621459.1930.10502211>
- Beasley, T. M., Erickson, S., & Allison, D. B. (2009). Rank-based inverse normal transformations are increasingly used, but are they merited? *Behavior Genetics*, 39(5), 580–595. <https://doi.org/10.1007/s10519-009-9281-0>
- Bishara, A. J., & Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological Methods*. <https://doi.org/10.1037/a0028087>
- Bishara, A. J., & Hittner, J. B. (2015). Reducing Bias and Error in the Correlation Coefficient Due to Nonnormality. *Educational and Psychological Measurement*, 75(5), 785–804. <https://doi.org/10.1177/0013164414557639>
- Blum, A. G., Archfield, S. A., & Vogel, R. M. (2017). On the probability distribution of daily streamflow in the United States. *Hydrology and Earth System Sciences*, 21, 3093–3103.
- Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2), 211–243. <https://doi.org/10.1111/j.2517-6161.1964.tb00553.x>
- Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, N.J.: Prentice-Hall, Inc.
- Cheriyān, K. C. (1945). Distributions of Certain Frequency Constants in Samples from Non-Normal Populations. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 7(2), 159–166.
- Chesire, L., Oldis, E., & Pearson, E. S. (1932). Further Experiments on the Sampling Distribution of the Correlation Coefficient. *Journal of the American Statistical Association*, 27(178), 121–128. Retrieved from <http://www.tandfonline.com/doi/abs/10.1080/01621459.1932.10502593>
- Cliff, N. (1996). *Ordinal methods for behavioral data analysis*. Mahwah, N.J.: Erlbaum.
- Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). New York: Wiley.
- 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>

- D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, & T. L. Veith. (2007). Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Transactions of the ASABE*. <https://doi.org/10.13031/2013.23153>
- Daniels, H. E. (1950). *Rank Correlation and Population Models*. Source: *Journal of the Royal Statistical Society. Series B (Methodological)* (Vol. 12).
- Dougherty, M. R., Thomas, R. P., Brown, R. P., Chrabaszcz, J. S., & Tidwell, J. W. (2015). AN INTRODUCTION TO THE GENERAL MONOTONE MODEL WITH APPLICATION TO TWO PROBLEMATIC DATA SETS. *Sociological Methodology*, 45, 223–271.
- Dunlap, H. F. (1931). An Empirical Determination of the Distribution of Means, Standard Deviations and Correlation Coefficients Drawn from Rectangular Populations. *The Annals of Mathematical Statistics*, 2(1), 66–81.
- Fisher, R. A., & Yates, F. (1963). *Statistical tables for biological, agricultural, and medical research*. Hafner Pub. Co. Retrieved from <http://hdl.handle.net/2027/mdp.39015020570803>
- Gayen, A. K. (1951). The frequency distribution of the product-moment correlation coefficient in random samples of any size drawn from non-normal universes. *Biometrika*, 38(1–2), 219–247. <https://doi.org/10.1093/biomet/38.1-2.219>
- Guilford, J. P., & Fruchter, B. (1973). *Fundamental statistics in psychology and education*. New York: McGraw-Hill.
- Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*. <https://doi.org/10.1016/j.jhydrol.2009.08.003>
- Haldane, J. B. S. (1949). A note on non-normal correlation. *Biometrika*, 36(3–4), 467–468. <https://doi.org/10.1093/biomet/36.3-4.467>
- Havlicek, L. L., & Peterson, N. L. (1976). Robustness of the Pearson Correlation against Violations of Assumptions. *Perceptual and Motor Skills*, 43(3_suppl), 1319–1334. <https://doi.org/10.2466/pms.1976.43.3f.1319>
- Helsel, D. R., & Hirsch, R. M. (Eds.). (1992). Correlation. In *Statistical Methods in Water Resources* (Vol. 49, pp. 209–220). Elsevier. [https://doi.org/https://doi.org/10.1016/S0166-1116\(08\)71107-5](https://doi.org/https://doi.org/10.1016/S0166-1116(08)71107-5)
- Hey, G. B. (1938). A New Method of Experimental Sampling Illustrated on Certain Non-Normal Populations. *Biometrika*, 30(1/2), 68–80.
- Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. *Ann. Math. Statist.*, 19(3), 293–325. <https://doi.org/10.1214/aoms/1177730196>
- Horowitz, L. M. (1974). *Elements of statistics for psychology and education*. New York: McGraw-Hill.
- Hutchinson, T. P. (1997). A Comment on Correlation in Skewed Distributions. *The Journal of General Psychology*, 124(2), 211–215. <https://doi.org/10.1080/00221309709595518>
- Jain, S. K., & Sudheer, K. P. (2008). Fitting of hydrologic models: A close look at the nash-

- sutcliffe index. *Journal of Hydrologic Engineering*. [https://doi.org/10.1061/\(ASCE\)1084-0699\(2008\)13:10\(981\)](https://doi.org/10.1061/(ASCE)1084-0699(2008)13:10(981))
- Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1–2), 81–93. <https://doi.org/10.1093/biomet/30.1-2.81>
- Kendall, M., & Gibbons, J. (1990). *Rank Correlation Methods* (5th ed.). New York: Oxford University Press.
- Kowalski, C. J. (1972). On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 21(1), 1–12. <https://doi.org/10.2307/2346598>
- Krause, P., Boyle, D. P., & Bäse, F. (2005). *Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences* (Vol. 5).
- Legates, D. R., & Davis, R. E. (1997). The continuing search for an anthropogenic climate change signal: Limitations of correlation-based approaches. *Geophysical Research Letters*, 24(18), 2319–2322. <https://doi.org/10.1029/97GL02207>
- Legates, D. R., & McCabe, G. J. (1999). Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. *Water Resources Research*. <https://doi.org/10.1029/1998WR900018>
- Limbrunner, J. F., Vogel, R. M., & Brown, L. C. (2000). Estimation of Harmonic Mean of a Lognormal Variable. *Journal of Hydrologic Engineering*, 2(January), 59–66.
- Maidment, D. R. (1993). *Handbook of hydrology*. New York: McGraw-Hill.
- Martinec, J., & Rango, A. (1989). MERITS OF STATISTICAL CRITERIA FOR THE PERFORMANCE OF HYDROLOGICAL MODELS. *Journal of the American Water Resources Association*, 25(2), 421–432. <https://doi.org/10.1111/j.1752-1688.1989.tb03079.x>
- McCuen, R. H., Knight, Z., & Cutter, A. G. (2006). Evaluation of the Nash-Sutcliffe efficiency index. *Journal of Hydrologic Engineering*. [https://doi.org/10.1061/\(ASCE\)1084-0699\(2006\)11:6\(597\)](https://doi.org/10.1061/(ASCE)1084-0699(2006)11:6(597))
- McCuen, R. H., & Snyder, W. M. (1975). A proposed index for comparing hydrographs. *Water Resources Research*, 11(6), 1021–1024. <https://doi.org/10.1029/WR011i006p01021>
- Moore, D. S. (1991). *Statistics : concepts and controversies* (3rd ed.). book, New York: W.H. Freeman.
- Nair, A. N. K. (1941). Distribution of Students “t” and the Correlation Coefficient in Samples from Non-Normal Populations. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 5(4), 383–400.
- Pearson, E. S. (1929). Some Notes on Sampling Tests with Two Variables. *Biometrika*, 21(1/4), 337–360. <https://doi.org/10.2307/2332565>
- Pool, S., Vis, M., & Seibert, J. (2018). Evaluating model performance : towards a non-parametric variant of the Kling-Gupta efficiency Kling-Gupta efficiency. *Hydrological*

- Sciences Journal*, 63(13–14), 1941–1953. <https://doi.org/10.1080/02626667.2018.1552002>
- Puth, M. T., Neuhäuser, M., & Ruxton, G. D. (2014). Effective use of Pearson's product-moment correlation coefficient. *Animal Behaviour*. Academic Press.
<https://doi.org/10.1016/j.anbehav.2014.05.003>
- Rider, P. R. (1932). On the Distribution of the Correlation Coefficient in Small Samples. *Biometrika*, 24(3/4), 382–403.
- Santos, L., Thirel, G., & Perrin, C. (2018). Technical note: Pitfalls in using log-transformed flows within the KGE criterion. *Hydrology and Earth System Sciences*.
<https://doi.org/10.5194/hess-22-4583-2018>
- Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York.
- Spearman, C. (1904). *The Proof and Measurement of Association between Two Things*. Source: *The American Journal of Psychology* (Vol. 15).
- Todini, E., & Biondi, D. (2017). Calibration, parameter estimation, uncertainty, data assimilation, sensitivity analysis, and validation. In V. Singh (Ed.), *Handbook of Applied Hydrology*. New York: McGraw-Hill.
- Willmott, C. J. (1981). ON THE VALIDATION OF MODELS. *Physical Geography*, 2(2), 184–194. <https://doi.org/10.1080/02723646.1981.10642213>
- Willmott, C. J., Ackleson, S. G., Davis, R. E., Feddema, J. J., Klink, K. M., Legates, D. R., ... Rowe, C. M. (1985). Statistics for the evaluation and comparison of models. *Journal of Geophysical Research*, 90(C5), 8995. <https://doi.org/10.1029/jc090ic05p08995>

Chapter 2

Improved Estimators of Correlation and R^2 For Skewed Hydrologic Data

with Jonathan R. Lamontagne & Richard M. Vogel

Publication citation: Barber, C., J. Lamontagne and R.M. Vogel (2019), Improved estimators of correlation and R^2 for skewed hydrologic data, *Hydrological Sciences Journal*, DOI: 10.1080/02626667.2019.1686639

Abstract

The coefficient of determination R^2 and Pearson correlation coefficient $\rho = R$ are standard metrics in hydrology for the evaluation of the goodness-of-fit between model simulations and observations, and as measures of the degree of dependence of one variable upon another. We show that the standard product moment estimator of ρ , termed r , while well behaved for bivariate normal data, is upward biased and highly variable for bivariate non-normal data. We introduce three alternative estimators of ρ which are nearly unbiased and exhibit much less variability than r for non-normal data. We also document remarkable upward bias and tremendous increases in variability associated with r using both synthetic data and daily streamflow simulations from 905 calibrated rainfall–runoff models. We show that estimators of $\rho = R$ accounting for skewness are needed for daily streamflow series because they exhibit high variability and skewness compared to, for example, monthly/annual series, where r should perform well.

1 Introduction

Consider the problem of evaluating the goodness-of-fit of hydrologic model output to observations. For the sake of illustration and without loss of generality, assume an additive error model. Every model has both a deterministic and stochastic element, so that a simulated response S is obtained from the sum of the deterministic model $H(X|\Omega)$ and a stochastic model error component ε :

$$S = H(X|\Omega) + \varepsilon \quad (2.1)$$

where X denotes some set of model input variables and Ω denotes the set of deterministic model parameters. Once a deterministic model is calibrated to observations, hydrologists usually compare the observations O to the simulations S , which are normally computed without adding model error, so that $S = H(X|\Omega)$. Thus, during the calibration period:

$$O = S + \varepsilon \quad (2.2)$$

Streamflow processes and hydrologic model output are unique in part due to the very high degree of variability, skewness, kurtosis and overall non-normality associated with the values of O , S and ε , causing tremendous estimation challenges associated with evaluations of goodness-of-fit. In fact, one could argue that in hydrologic modeling, non-normality is the norm, rather than an exception. In hydrology it has long been known that estimators of goodness-of-fit such as correlation are highly impacted by non-normality, nonlinearity and outliers. This is in part why there are now many well developed nonparametric alternative estimators of correlation which are in common use such as the Spearman and Kendall correlations (see Helsel and Hirsch, 2002, Helsel *et al.* 2019).

As discussed below, there is an extensive literature in hydrology on the advantages and disadvantages of various goodness-of-fit statistics, and it is not our goal to enter into that debate. Instead, we have noticed that nearly all previous studies which have sought to evaluate and compare goodness-of-fit statistics in hydrology have failed to distinguish between the probabilistic properties and behavior of the theoretical statistics, and the rather different sampling (statistical) properties of estimators of those statistics when computed from data. It is this distinction between the theoretical or population statistic and the sampling properties of its various possible estimators which sets our work apart from any previous work on goodness-of-fit statistics in hydrology.

The primary purpose of this paper is to evaluate and compare a number of common estimators of the degree of correlation between the observations O and simulations S with the ultimate goal of developing improved estimators suited for use in (i) evaluating the goodness-of-fit of hydrologic models and (ii) as a measure of the degree of dependence of one variable upon another. Our analysis ignores the model error component ε in (2.1) and (2.2), and we refer the reader to Farmer and Vogel (2016a) and Vogel (2017) for further information on the implications of ignoring model error on goodness-of-fit evaluations and, more importantly, on the use of such models in water resources planning and management.

1.1 R^2 and correlation coefficient

Various metrics have been advanced for quantifying the goodness-of-fit of the simulations S to the observations O and as a measure of the degree of dependence of one variable upon another. In this initial study we focus on the commonly used goodness-of-fit metric known as R^2 , which is simply the square of the Pearson (1896) correlation coefficient ρ between O and S . The theoretical (or probabilistic) definition of $R = \rho$ is given by:

$$\rho = R = \frac{\text{cov}(O, S)}{\sqrt{\text{var}(O)\text{var}(S)}} = \frac{E[(O - \mu_O)(S - \mu_S)]}{\sigma_O \sigma_S} \quad (2.3)$$

and the most common estimator of $R = \rho$ is known as the Pearson product moment correlation coefficient given by:

$$r = \frac{\frac{1}{n} \sum_{i=1}^n (s_i - \bar{s})(o_i - \bar{o})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (o_i - \bar{o})^2 \frac{1}{n} \sum_{i=1}^n (s_i - \bar{s})^2}} \quad (2.4)$$

Pearson (1896) introduced both the theoretical statistic $R = \rho$ in (2.3) as well as its sample estimator r , in (2.4). It is common practice to use uppercase O and lowercase o to denote the theoretical values and their sample realizations, respectively. Similarly, it is common practice to use Greek characters for the theoretical mean, variance and correlation coefficient, μ_o , σ_o^2 and ρ and to use \bar{o} , s_o^2 and r to denote sample estimates based on sample realizations. While the theory of probability governs the behavior and properties of $R = \rho$ in (2.3), it is the theory of statistics which governs the sampling properties of the estimator r .

1.2 Application of R² and Pearson correlation coefficient ρ in hydrology

Numerous hydrologic studies have reviewed the use of the common estimator of the Pearson correlation coefficient r in (2.4) for use in evaluating the goodness-of-fit of hydrologic models (McCuen and Snyder, 1975; Willmott, 1981; Willmott *et al.* 1985; Legates and Davis, 1997; Legates and McCabe, 1999; Krause *et al.*, 2005; Moriasi *et al.* 2007). In each of those studies numerous concerns were raised about the value of using estimates of ρ or R^2 as a goodness-of-fit metric. The primary drawback of the use of ρ or R^2 as a goodness-of-fit metric is that they do not account for model bias. This is in contrast with the more general and useful goodness-of-fit statistic known as the Nash Sutcliffe efficiency (NSE), which is a standardized mean square error (MSE).

The advantage of any MSE type criterion over ρ or R^2 is that it includes both bias and variance aspects of goodness-of-fit. Since $NSE = \rho^2 = R^2$ for any unbiased model which exhibits serially independent residuals ε in Eq. 2.1, the results of this study pertain directly to our follow-up study on the theoretical behavior and sampling properties of an improved estimator of the theoretical statistic which NSE attempts to mimic. Another drawback of the theoretical correlation metric ρ or R^2 is that it is only a measure of linear association or dependency, which is why a host of other nonparametric correlation metrics have been advanced. Again, it is not our goal to evaluate which theoretical goodness-of-fit metric is best for a given application, rather, given the widespread usage (and misuse) of the statistic r in (2.4), it is our goal to obtain improved estimators of ρ or R^2 suited specifically for skewed hydrologic data.

Remarkably, all of the hydrologic studies cited above suffer from the error of not having distinguished between the theoretical statistic ρ given in (2.3) and one estimator of that statistic, r , given in (2.4). This is remarkable because most of the previously cited hydrologic studies criticize the performance of the estimator r , not realizing that it is only one of an infinite number of ways to estimate ρ and that it is possible to come up with improved estimators of ρ for hydrologic applications. For example, McCuen and Snyder (1975) suggested modifications to the estimator r without ever resorting to a theoretical analysis to ensure the modification is consistent with the definition of ρ in (2.3). Similarly, Legates and McCabe (1999) and many others have criticized the use of the estimator r in (2.4) due to its sensitivity to outliers, not realizing that the estimator r is only one of many possible estimators of ρ , some of which considered here are NOT unusually sensitive to outliers. Thus, in effect all of the hydrologic studies cited above have criticized the performance of r , and because they never presented or considered the theoretical definition of ρ in

(2.3) they have, by default, also dismissed and criticized the behavior of ρ . This is illogical and would be analogous to rejecting the theoretical statistic $E[x] = \mu$ just because one of its estimators, the sample mean \bar{x} , is heavily influenced by outliers.

1.3 Performance of R^2 and r under bivariate normality

The statistical properties of the estimator r have been understood for over a century under the condition of bivariate normality. For example, Fisher (1915) derived the exact sampling distribution of r for samples from a bivariate normal distribution. The estimator r in (2.4) is known to yield approximately unbiased estimates of ρ when observations and simulations arise from a bivariate normal process. When data follow a bivariate normal distribution, the sample estimator r of ρ is very well behaved, in the sense that it is a maximum likelihood estimator and thus provides an asymptotically unbiased estimator of the true value, because $E[r] = \rho[1 - (1 - \rho^2)/2n + O(n^{-2})] \rightarrow \rho$ as $n \rightarrow \infty$ (see Balakrishnan and Lai, 2009; and Xu *et al.* 2013). Note that the bias in r is only slight and disappears for $n > 20$ under bivariate normal sampling. Unbiasedness is a very important property for a statistic like r which is so widely used across disciplines and applications. Xu *et al.* (2013) also summarize the variance of r under bivariate normal sampling as $\text{var}[r] \cong (1 - \rho^2)^2/(n - 1)$.

1.4 Sampling properties of R^2 and r under bivariate non-normality

Unfortunately, in hydrologic applications, bivariate *non-normality* is more the norm than is bivariate normality. It has long been known by statisticians that the behavior of r can be quite sensitive to non-normality and that use of r should be limited to situations in which both S and O are normally distributed or nearly so (Kowalski, 1972). Kowalski (1972) provides a detailed historical survey of studies dating back to the early twentieth century which evaluated the impact

of non-normality on the distribution of r . He concluded that “*everyone seems to agree that the distribution of r is quite robust to non-normality when $\rho = 0$, but there is good evidence that this becomes less stable with increasing values of $|\rho|$, especially when kurtosis is in evidence. It is the variance of r which is most vulnerable to the effects of non-normality and this variance may be either larger or smaller than the normal-theory value, depending on the type of non-normality under consideration.*” Embrechts *et al.* (2002) uses theoretical arguments and Habib *et al.* (2001) use simulation results to document some of the challenges in estimation of ρ from bivariate (and multivariate) non-normal processes.

Another problem with r is that it is very sensitive to sample outliers and other features of datasets which create departures from bivariate normality. For example, Xu *et al.* (2013) argued that r “*is notoriously sensitive to the non-Gaussianity caused by impulsive contamination in the data. Even a single outlier can severely distort the value of r and hence result in misleading inference in practice.*” In addition to concerns over the impact of outliers, Xu *et al.* (2013) also report that r performs poorly under monotone nonlinearity, and it is for this reason that alternative measures of dependence have been developed and compared by Devlin *et al.* (1975), Serinaldi (2008), Xu *et al.* (2013), Bishara and Hittner (2015, 2017), and many others.

Still most literature evaluating the behavior of r have focused on bivariate normal and other symmetric bivariate distributions (see Chapter 32 in Johnson *et al.* (1995), for a review), whereas our interest focuses on estimation of ρ for skewed bivariate hydrologic data. Serinaldi (2008) provides a good review of challenges and approaches to estimation of the correlation coefficient for skewed hydrologic data and recommends the use of alternative nonparametric measures of correlation including Kendall’s rank correlation, an upper tail dependence coefficient, as well as

several copula approaches. Here we focus on estimation of the most commonly used correlation metric ρ due to its widespread historical use as a measure of the degree of dependence of one variable upon another and as a goodness-of-fit metric.

Numerous authors reviewed by Johnson *et al.* (1995) and Lai *et al.* (1999) have derived expressions for the first four moments of r in terms of the cumulants and cross-cumulants of the parent non-normal population. Despite this attention given to r , the magnitude of the bias and the variance of r are still relatively poorly understood for general bivariate non-normal populations. Although several non-normal populations have been investigated, there is no uniform guidance or understanding of the robustness of r against non-normality (see Johnson *et al.*, 1995, p. 580).

Lai *et al.* (1999) examined the bias and variance in r under bivariate lognormal sampling using both Monte Carlo simulation experiments and analytical derivations. Their experiments revealed tremendous upward bias associated with the estimator r in (2.4) for bivariate lognormal samples. Importantly, Lai *et al.* (1999) concluded that the upward bias in the estimator r for bivariate lognormal samples is only reduced (approximately removed) with sample sizes in the range of 3–4 million observations. The example introduced by Lai *et al.* (1999) has received little attention in the literature, in spite of the fact that bivariate hydrologic samples tend to be much better approximated by a bivariate lognormal model than a bivariate normal model. The only study in hydrology we could locate which noted the upward bias associated with r under non-normal sampling is Habib *et al.* (2001) which dealt with interstation correlation of rainfall series. Following Shimizu (1993), Habib *et al.* (2001) documented a method to correct for the upward bias associated with the estimator r under a bivariate discrete-continuous (mixed) lognormal model. This work is distinctly different from our work because we make use of a bivariate continuous lognormal model. More recently, in reaction to the phenomenon observed by Lai *et al.*

(1999), Zhang and Chen (2015) developed generalized confidence intervals and hypothesis tests for the value of r computed from bivariate LN2 samples. Persistence in each of the bivariate series under consideration is also known to increase the sampling variance of the Pearson correlation estimator, when compared to independent series. For example, Arbabshirani, et al (2014) derived the variance of r when both series, x and y , arise from a lag-one autoregressive (AR(1)) model resulting in:

$$\text{var}[r] = [(1 - \rho^2)^2/n][(1 + \rho_s\rho_o)/(1 - \rho_s\rho_o)] \quad (2.5)$$

where ρ_s and ρ_o are the lag-one serial correlation coefficients for the s and o series, respectively. The second quantity on righthand side of (2.5) represents the inflation in the variance due to autocorrelation which can be quite large for daily flow series which exhibit a very high degree of persistence.

Although numerous authors have recently evaluated the behavior of r under departures from bivariate normality (see for example Bishara and Hittner; 2015, 2017), we are unaware of any literature which has derived expressions for the bias and variance of r under alternatives to bivariate normality. As documented by Bishara and Hittner (2015, 2017) and others, departures to bivariate normality affect not only estimates of ρ but also inflate the probability of type I and II errors when using r to perform hypothesis tests regarding the true value ρ . On the basis of Monte Carlo experiments which generated bivariate non-normal data with known values of ρ , Bishara and Hittner (2015) compared the performance of several alternative estimators of ρ under a wide variety of bivariate distribution shapes, sample sizes and true values of ρ . In some sense, this study can be viewed as a follow-up study to Bishara and Hittner (2015) but suited to the unique features of skewed hydrologic data instead of the type of educational and psychological data in their study.

2 Study assumptions: bivariate lognormal model of hydrologic data

More and more, high frequency hydrologic model simulations are employed at daily, hourly and even sub-hourly time scales to enable increasingly sophisticated water resource management applications. Daily and hourly streamflows are known to exhibit extremely high values of coefficient of variation and skewness, so that typical values of S and O in (2.1) and (2.2) are much more closely approximated by a bivariate lognormal model than a bivariate normal model. Blum *et al.* (2017) and Limbrunner *et al.* (2000, Figure 6) showed that two-parameter and three-parameter lognormal distributions (LN2 and LN3, respectively) provide a very good first approximation to the distribution of daily streamflow observations for hundreds of stations across the conterminous United States. Therefore, we make the reasonable assumption that observations O and simulations S of daily streamflow follow a bivariate lognormal (LN2) distribution. The derivations of our improved estimators of ρ rely on this assumption which not only allows for analytical (closed-form) derivations, but it is also rather general and well-suited for hydrologic variables considered in this study.

The appendix summarizes a simple algorithm for generating synthetic streamflows from a bivariate lognormal model that is equivalent to many other approaches including the more general meta-Gaussian method (see e.g. Papalexiou, 2018, Tsoukalas *et al.* 2018, and references therein). Moreover, we also perform an empirical analysis fitting a bivariate LN3 model to actual streamflow observations.

Another critical feature of our study is that we consider the impact of the extraordinary variability and skewness associated with high frequency (i.e. daily, hourly and sub-hourly) streamflow observations. Vogel *et al.* (2003, see Figure 1) summarize the behavior of estimates of the coefficient of variation, $C_o = \sigma_o/\mu_o$ of daily streamflow series across the conterminous United

States. Their estimates of C_O were obtained using L-moment estimators for an LN3 distribution whose lower bound avoids undefined logarithm values due to zero daily streamflows enabling characterization of rivers with intermittent regimes. Vogel *et al.* (2003) report values of C_O across the USA which range from 0.5 to 10,000, with a median value of 10 and an interquartile range from 3 to 33. The larger values occurred in arid and semi-arid regions of the western U.S. and the extremely large estimates of C_O correspond to sites which had a very large fraction of zero flow values. When zero observations are a concern, an alternative to fitting an LN3 model would be to consider a mixed lognormal distribution of daily streamflow as advocated by Guo *et al.* (2016).

The bivariate LN2 and LN3 models involve two assumptions: (1) the marginal distributions of the two variables O and S are LN2 or LN3, and (2) a linear dependence structure exists between $U = \ln[O]$ and $V = \ln[S]$ for the LN2 case, and between $U = \ln[O - \tau_o]$ and $V = \ln[S - \tau_s]$ for the LN3 case, where τ_o and τ_s are the lower bounds of the LN3 distributions of O and S , respectively. In order to develop suitable alternative estimators of ρ which would perform well under bivariate LN3 sampling, it is necessary to exploit the theoretical relationship between the correlation between O and S and the correlation between their natural logarithms $U = \ln[O - \tau_o]$ and $V = \ln[S - \tau_s]$. The relationship between the log space correlation between U and V , denoted ρ_{UV} and the real space correlation between O and S , denoted as ρ is given by:

$$\rho = \frac{\exp(\rho_{UV}\sigma_U\sigma_V) - 1}{\sqrt{\exp(\sigma_U^2) - 1} \sqrt{\exp(\sigma_V^2) - 1}} \quad (2.6)$$

(see Mostafa and Mahmoud, 1964; equation 5 in Stedinger 1981 and eq. 11.71 in Balakrishnan and Lai, 2009). Thus, (2.6) represents the relationship between the population correlations in real space, ρ , and log space, ρ_{UV} , corresponding to a bivariate LN3 model. In general, $\rho_{UV} > \rho$ (Embrechts *et al.* 2002). For the LN2 case (i.e. $\tau_o = \tau_s = 0$), setting $\sigma_U = \sigma_V$ in (2.6), we have

$\lim_{C_O C_S \rightarrow 0} \rho_{UV} = \rho$ and $\lim_{C_O C_S \rightarrow \infty} \rho_{UV} = 1$. Typically, the difference between ρ and ρ_{UV} increases as both σ_U and σ_V increase, regardless of whether $\sigma_U = \sigma_V$. For example, when the coefficient of variation of the observations, $C_O = \sigma_O/\mu_O$, and simulations, $C_S = \sigma_S/\mu_S$, is $C_O = C_S = 10$ and $\rho = 0.8$, solving Eq. (2.6) yields $\rho_{UV} = 0.952$. Generally, the coefficient of variation of the observations and simulations will not be equal (see Figure 2 in Farmer and Vogel, 2016a); however, this appears to have little impact on the difference between ρ and ρ_{UV} . For example, when $C_O = 10$, $C_S = 6$ and $\rho = 0.8$, from Eq. (2.6), we obtain the almost identical result of $\rho_{UV} = 0.953$. Note that for an LN2 model $C_O = \sqrt{\exp(\sigma_U^2) - 1}$ and $C_S = \sqrt{\exp(\sigma_V^2) - 1}$.

The relationship in Eq. (2.6) and the assumption that O and S follow an LN3 distribution are the two primary assumptions implicit in our work. Both of these assumptions are verified in Section 5.2 using 905 bivariate samples of actual daily streamflow derived from calibrated distributed hydrologic rainfall–runoff models.

3 Alternative estimators of R^2 and ρ for non-normal bivariate hydrologic data

Consider the problem of evaluating the goodness-of-fit metric ρ when O and S are observations and simulations of daily, hourly or sub-hourly streamflow. Daily streamflow typically varies over 4 to 5 orders of magnitude in a single year, resulting in extremely high values of C_O and skewness. Bivariate non-normality is arguably the norm in hydrologic practice; thus, this section considers three alternative estimators of ρ suited to such conditions. In this initial study, we derive estimators based on the assumption of bivariate LN3 streamflows because, as is shown later, this is a good first approximation for modeling bivariate streamflow series considered in this study and many others.

Our work is a departure from previous work on alternative estimators of correlation because our focus is exclusively on the development and evaluation of alternative estimators of the theoretical value of ρ . This distinction is extremely important because (a) most previous hydrologic studies dealing with the behavior of the Pearson correlation coefficient r never distinguished between the estimator r and its theoretical value ρ ; and (b) this is one of the only studies we are aware of to introduce a suite of alternative estimators of ρ based on several widely used nonparametric correlation estimators such as the Spearman correlation coefficient and the rank inverse normal transformation correlation estimator. To better understand this distinction it is necessary to understand the role of both assumptions inherent in our work: (i) the assumed marginal lognormal distribution of O and S ; and (ii) the linear dependence structure between U and V , as well as the highly nonlinear dependence structure between O and S . To better understand the role of the dependence structure between O and S , and between U and V , we briefly review the role of the copula. After that, we introduce three additional estimators of ρ which are all shown to be improvements over r , for skewed bivariate hydrologic and synthetic data.

3.1 The copula and the dependence structure of hydrologic variables

We introduce the copula because all of the improved estimators introduced in this paper are based on a Gaussian copula and because we wish to emphasize that our initial work could and probably should be extended to other copula families, as well as other marginal probability distributions associated with the observations and simulations. Since one goal of our work is to educate hydrologists who use the Pearson correlation estimator r as a performance measure in cases with skewed observations and simulations, our discussion of the copula focuses on concepts without resorting to extensive mathematics.

The copula contains all the information about the dependence between the random variables O and S , as well as between variables resulting from monotonic transformations, such as $U = \ln[O - \tau_O]$ and $V = \ln[S - \tau_S]$. Importantly, the copula enables one to model the dependence structure between the variables separately from their marginal probability distributions. The copula describes the relationship between the exceedance probabilities of each of those variables. Suppose Z and W represent the exceedance probability associated with the variables O and S . According to the probability integral transformation theorem, the exceedance probabilities Z and W always follow a uniform distribution regardless of the original marginal distributions of O and S . For example, suppose Z and W are the exceedance probabilities of two normally distributed variables U and V , then a bivariate normal model can be seen as the combination of a bivariate Gaussian copula describing the linear dependence between the exceedance probabilities Z and W of two normally distributed variables along with the assumption of Gaussian marginal distributions associated with U and V . Since Z and W are uniform and can result from any distribution, a bivariate lognormal model can be seen as a combination of a bivariate Gaussian copula describing the linear dependence between the exceedance probabilities Z and W of two lognormal variables O and S , and lognormal marginal distributions for O and S , which would be referred to as the target process of interest. In this context, our Eq. (2.6) is a special case (for a bivariate lognormal model) of Eq. (8) in Papalexiou (2018), which links the correlation coefficient ρ_{UV} of a parent bivariate Gaussian process associated with U and V , with the correlation coefficient of a target bivariate process with arbitrary marginal distributions (also see Xiao, 2014; and Tsoukalas *et al.*, 2018). Copulas can be extremely useful for modeling and understanding bivariate and multivariate relationships because given a single copula, one can obtain many different bivariate or multivariate distributions by simply selecting different marginal distributions to work with. Genest and

Chebana (2017) provide an illustration for how to select a suitable copula for characterizing the dependence structure between the ranked streamflow values.

Salvadori *et al.* (2007), Salvadori and De Michele (2013) and Genest and Chebana (2017) provide a good overview of the advantages and uses of copulas in hydrology. Embrechts *et al.* (2002) provides a detailed overview of the need for separately understanding the dependence structure and the marginal distributions of the bivariate or multivariate process. We stress here that copulas provide an excellent framework for understanding the derivations of our estimators below, and for extending our work based on a bivariate lognormal process to include other bivariate models in terms of both the dependence structure between the variables and their marginal distributions. Here we assume a linear dependence structure between U and V and a highly nonlinear power law relationship between O and S , with Gaussian marginals in log space and lognormal marginals in real space. Although our contributions are restricted to these assumptions, we highlight that future work could extend our results to other bivariate processes by resorting to copulas.

3.2 *Stedinger's (1981) lognormal estimator, r_1*

For situations in which O and S arise from a bivariate LN2 model, Stedinger (1981) recommended an improved estimator of the correlation coefficient ρ based on the theoretical relationship between ρ and ρ_{UV} given in Eq. (2.6). Here we consider a slight adaptation of Stedinger's (1981) estimator for use with bivariate LN3 samples given by:

$$r_1 = \frac{\exp[\hat{\sigma}_{uv}^2] - 1}{\sqrt{(\exp[\hat{\sigma}_u^2] - 1)(\exp[\hat{\sigma}_v^2] - 1)}} \quad (2.7)$$

where $u_i = \ln[o_i - \hat{t}_O]$ and $v_i = \ln[s_i - \hat{t}_S]$, with

$$\hat{\sigma}_{uv}^2 = \frac{1}{n} \sum_{i=1}^n (u_i - \bar{u})(v_i - \bar{v}) \quad (2.8a)$$

$$\hat{\sigma}_u^2 = \frac{1}{n} \sum_{i=1}^n (u_i - \bar{u})^2 \quad (2.8b)$$

and

$$\hat{\sigma}_v^2 = \frac{1}{n} \sum_{i=1}^n (v_i - \bar{v})^2 \quad (2.8c)$$

A very attractive and efficient estimator of the lower bounds τ_O and τ_S for use in Eq. (2.7) is given in another paper by Stedinger (1980) as:

$$\hat{\tau}_O = \frac{o_{(1)}o_{(n)} - (o_{0.5})^2}{o_{(1)} + o_{(n)} - 2o_{0.5}} \quad (2.9a)$$

and

$$\hat{\tau}_S = \frac{s_{(1)}s_{(n)} - (s_{0.5})^2}{s_{(1)} + s_{(n)} - 2s_{0.5}} \quad (2.9b)$$

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 . The condition $o_{(1)} + o_{(n)} - 2o_{0.5} > 0$ must be satisfied to obtain a reliable estimate of $\hat{\tau}_O$ in Eq. (2.9a). Analogous definitions exist for estimation of $\hat{\tau}_S$ based on the simulations s .

Equation (2.7) is based on the relationship in Eq. (2.6) which is an analytical version of the linkage between the correlation between O and S given by ρ and the correlation between the values of U and V resulting from the parent bivariate Gaussian process. Other estimators analogous to Eq. (2.7) could be derived based on other bivariate processes with different copulas and other marginal distributions of O and S .

Using Monte Carlo experiments based on synthetic bivariate LN2 samples, Stedinger (1981) documents that r_1 is generally preferred over r ; however, his experiments only considered bivariate LN2 samples with coefficient of variations $C_O \leq 1$, typical of series of annual maximum floods and drought. Daily and hourly streamflow are known to exhibit extremely high skewness corresponding to much higher values of C_O than considered by Stedinger (1981), and high values of skewness lead to considerable degradation in the performance of r , and thus to considerable advantages of r_1 over r , as is shown below.

3.3 Modified Spearman rank correlation estimator, r_2

Nonparametric methods are now widely used in hydrology and described in detail by Helsel and Hirsch (2002) and Helsel *et al.* (2019). Most nonparametric methods work with the ranks of the data instead of the data itself, and Spearman's correlation estimator is simply the Pearson correlation estimator r , applied to the ranks of O and S . Here we derive a modified version of Spearman's nonparametric estimator of correlation introduced by Spearman (1904) which is suited for use under the assumptions of our study. The population value of Spearman's correlation is denoted here as ρ_s and its sample estimator is denoted as r_s . The only situation in which the Pearson and Spearman correlations are equal (i.e. $\rho = \rho_s$) would be for bivariate uniform data because the ranks of data are always uniformly distributed, regardless of their underlying distribution. Thus, just as r is an estimator of ρ , the estimator r_s provides an estimate of the correlation of the ranks of the values of O and S . It would not make sense to compare the performance of r and r_s under bivariate models with non-uniform marginal distributions, as has been done in numerous previous studies (see for example Bishara and Hittner, 2015, 2017 and references cited therein), because as Astivia and Zumbo (2017) show so clearly, these are estimates of different population correlation statistics. Similarly within the context of developing an

improved estimator of the NSE, Pool et al. (2018) incorrectly equated the properties of the Spearman and Pearson correlation coefficients. Instead, we must employ the necessary transformations to ensure that the resulting non-parametric correlation estimator is an estimate of the population value of ρ . Here two important transformations are needed, (a) to account for the relationship between Pearsons ρ and Spearmans ρ for bivariate normal data, and (b) to account for the known relationship in Eq, (2.6) between ρ and ρ_{UV} .

When Spearman's estimator r_s is applied to any ranked data (expressed as positive integers 1, 2, 3, ...), with no ties, the estimator can be simplified to:

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (2.10)$$

where d denotes the differences between the ranks and n is the sample size (Astivia and Zumbo, 2017; and Xu et al. 2013). Under a bivariate normal model between U and V , we have $E[r_s] = \rho_S = (6/\pi)\sin^{-1}(\rho_{UV}/2)$ (Kruskal, 1958) which can be inverted to yield:

$$\rho_{UV} = 2\sin\left(\frac{\pi\rho_S}{6}\right) \quad (2.11)$$

Now replacing ρ_S with r_s in Eq. (2.11), and combining Eq. (2.11) with Eq. (2.6), yields our modified Spearman correlation estimator r_2 which is designed to reproduce Pearson's ρ for the case of bivariate LN3 processes:

$$r_2 = \frac{\exp\left[2\sin\left(\frac{\pi r_s}{6}\right)\hat{\sigma}_U\hat{\sigma}_V\right] - 1}{\sqrt{\exp(\hat{\sigma}_U^2) - 1}\sqrt{\exp(\hat{\sigma}_V^2) - 1}} \quad (2.12)$$

where r_s is given in Eq. (2.10), and the estimators $\hat{\sigma}_u^2$ and $\hat{\sigma}_v^2$ are given in Eqs. (2.8a) and (2.8b) with $u_i = \ln[o_i - \hat{t}_o]$ and $v_i = \ln[s_i - \hat{t}_s]$ and the estimators \hat{t}_o and \hat{t}_s are given by Stedinger's (1980) lower bound estimator in Eqs. (2.9a) and (2.9b).

Helsel and Hirsch (2002) and Helsel *et al.* (2019) provide background on the computation of r_s as well as associated hypothesis tests and confidence intervals. Note that, since the ranks of the data are expected to follow uniform distributions, the impact of the highly non-normal populations of the observations and simulations on estimation of correlations is reduced considerably.

3.4 A modified rank inverse normal correlation estimator, r_3

The Spearman correlation coefficient applies the Pearson correlation coefficient to a transformation of the data pairs (o_i, s_i) into their associated ranks. An attractive and related estimator is the more general rank inverse normal (RIN) correlation estimator recommended by Bishara and Hittner (2015, 2017) and many others. Beasley and Erickson (2009) provide a review of applications, advantages and caveats associated with the RIN approach which apparently is increasingly widely used in a variety of fields.

The RIN method consists of four steps. First, each pair (o_i, s_i) is converted to their ranks (j, k) where j and k denote the ranks associated with the observations and simulations, respectively. Next, the ranks are transformed to probabilities by using a plotting position such as the Weibull plotting position to yield the new pairs $(j/(n+1), k/(n+1))$ where j and k take on integer values between 1 and n . The Weibull plotting position is attractive because it yields unbiased estimates of the cumulative probabilities associated with observations and simulations regardless of their underlying marginal probability distributions. The third step involves an inverse normal

transformation from the cumulative probabilities into the standard normal variates so that each pair now becomes $(\Phi^{-1}(j/(n+1)), \Phi^{-1}(k/(n+1)))$, where $\Phi^{-1}(p)$ denotes the inverse of a standard normal variate with cumulative probability equal to p . The RIN estimator is obtained by simply applying the Pearson correlation estimator r in Eq. (2.4) to the inverse normal pairs resulting in the estimator we denote as r_{RIN} . The problem with r_{RIN} is that it is an estimate of the correlation in log space, ρ_{UV} and not the correlation in real space ρ , which we desire; hence, one needs to transform its value into real space through the transformation expression given in Eq. (2.6) resulting in the following corrected RIN estimator:

$$r_3 = \frac{\exp(r_{\text{RIN}}\hat{\sigma}_U\hat{\sigma}_V) - 1}{\sqrt{\exp(\hat{\sigma}_U^2) - 1}\sqrt{\exp(\hat{\sigma}_V^2) - 1}} \quad (2.13)$$

where again the estimators $\hat{\sigma}_u^2$ and $\hat{\sigma}_v^2$ are given in Eqs. (2.8a) and (2.8b) with $u_i = \ln[o_i - \hat{t}_o]$ and $v_i = \ln[s_i - \hat{t}_s]$ and the estimators \hat{t}_o and \hat{t}_s are given by Stedinger's (1980) lower bound estimator in Eqs. (2.9a) and (2.9b).

4 Monte Carlo experiments

We begin our evaluation of the four estimators of ρ by generating synthetic bivariate lognormal streamflow data with a range of coefficients of variation, sample sizes and ρ similar to those observed in practice. After those evaluations, the remainder of the paper evaluates the four estimators of ρ using actual bivariate streamflow observations from hundreds of watersheds across the USA. In our Monte Carlo experiments, $m=500$ bivariate LN2 samples of length $n=100$ and $n=10,000$ and coefficients of variation $C_o = C_s = 0.5, 2.0$ and 10.0 are generated for $\rho = 0.7$ and 0.9 using the methodology outlined in the Appendix. Each of those $m = 500$ experiments leads to estimates of r, r_1, r_2 and r_3 , based on the estimators given in Eqs. (2.4), (2.7), (2.12) and (2.13),

respectively. Boxplots of the resulting values of the four estimators are illustrated in Figure 2.1. Under all the conditions considered, the estimators r_1 , r_2 and r_3 are relatively unbiased and exhibit variability which decreases significantly as sample size increases, as expected. In contrast, the Pearson correlation r exhibits significant upward bias and much more variability than r_1 , r_2 and r_3 , with neither its bias nor its variance disappearing even for very large sample sizes. Importantly, we note from Figure 2.1 that the upward bias and very large variability of the estimator r increases as the coefficient of variation of the observations and simulations increases. Our results in Figure 2.1 are consistent with those of the previous study by Lai et al. (1999), who found that the bias and the inflation in the variance of r does not seem to disappear until samples sizes in the millions are obtained.

It must be highlighted that the bivariate lognormal generation algorithms used in this study and the study by Lai et al. (1999), result in serially independent traces, whereas actual daily streamflow observations are known to exhibit an extremely high level of persistence. The primary effect of serial correlation on estimation of correlations is that it creates an overlap in the information contained in each datapoint which effectively reduces the sample size of the dataset. This usually results in increases in both the bias and variance of correlation estimates when compared with independent samples of the same sample size n , which is a well-known phenomenon. Consider the case when a sample of n simulations and observations each arise from an AR(1) process with lag one correlations ρ_s and ρ_o both equal to 0.9, a typical value for daily streamflow series. Then using the result for $\text{var}(r)$ from Arbabshirani et al. (2014) presented in Section 1.4 (see Eq. [5]) along with the definition of information content introduced by Matalas and Langbein (1962), we obtain the very approximate information content of a daily streamflow series as:

$$I = n / \left[\frac{(1 + \rho_s \rho_o)}{(1 - \rho_s \rho_o)} \right] = 0.10n \quad (2.14)$$

which indicates the gross reduction in information resulting from serial correlation. Thus, the results given in Figure 2.1 for independent streamflow series of length $n = 100$ and $n = 10,000$ correspond very roughly to actual serially correlated daily streamflow series of lengths equal to $n = 1000$ and $n = 100,000$, respectively.

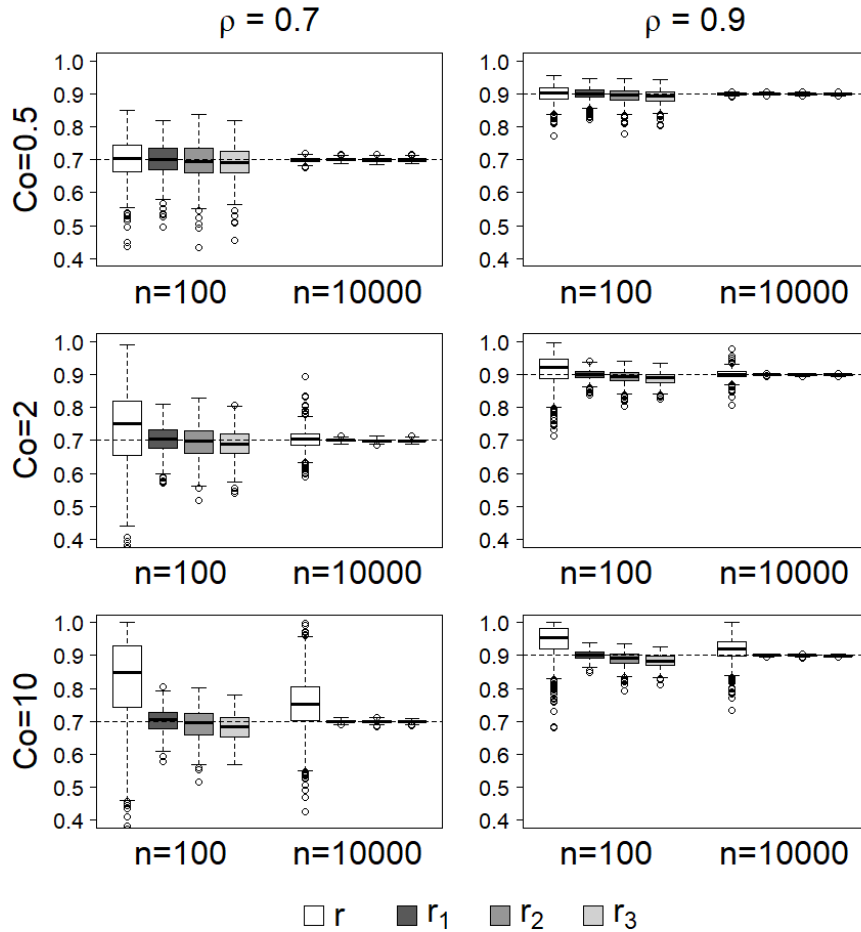


Figure 2.1 Results of the Monte Carlo experiments are illustrated using boxplots of the four estimators of correlation ρ considered: Pearson r , Stedinger r_1 , modified Spearman r_2 and modified RIN r_3 . Boxplots summarize the sampling distribution of estimators of ρ for synthetic streamflows generated from a bivariate lognormal model with $\rho = 0.7$ (left panels) and $\rho = 0.9$ (right panels), for three different values of coefficient of variation $Co=C_s$

5 Evaluations using actual bivariate streamflow observations

The results in Section 4, along with analogous results by Lai et al. (1999), provide evidence of the relatively large upward bias and inflation in variance associated with the estimator r , under bivariate lognormal sampling, particularly for large values of C_o and C_s . Two questions which remain are (a) to what extent are actual bivariate streamflow observations approximated by a bivariate lognormal process; and (b) to what extent is the behavior of the four estimators documented in Section 4 under bivariate lognormal sampling similar to that which could be expected when used with actual daily streamflow observations. The compelling challenge which plagues us in such evaluations is that we will never know the true correlation ρ when working with actual bivariate streamflows; however, we can examine whether or not the general behavior of the four estimators is similar between actual bivariate sequences and synthetic bivariate lognormal sequences, which is the subject of this section.

5.1 *Bivariate PRMS streamflow simulations and observations*

Here, as in Farmer and Vogel (2016a), a moderately complex, distributed-parameter, precipitation–runoff model is used to generate bivariate daily streamflow traces from daily streamflow observations at 1225 river locations across the continental United States. The distributed-parameter model, in this case, the Precipitation–Runoff Modeling System (PRMS; Markstrom *et al.*, 2015), was calibrated at each of 1225 perennial river basins across the conterminous United States. Details and availability of the datasets are described by Farmer and Vogel (2016b). The particulars of the model and the calibration scheme are not relevant to our experiments. Our focus is not on the development and calibration of this model, but rather on the behavior of estimates of ρ derived from observed and modeled daily streamflow, thus further

details of the model are not provided here. The same general conclusions can be expected to result from the use of any hydrologic model used to simulate daily streamflow.

An experienced hydrologist would never resort only to quantitative goodness-of-fit metrics, but would instead perform graphical evaluations to ensure consistent and sensible behavior between the observations, o and the simulations s . To mimic the work of a hydrologist, we examined every scatterplot of $v = \ln[s - \hat{\tau}_s]$ versus $u = \ln[o - \hat{\tau}_o]$ to ensure that they mimic the type of behavior expected from such analyses. Our experience indicates that one expects an approximately ellipsoidal relationship between u and v , which would be consistent with the assumption of a bivariate lognormal relationship between o and s . Removing those sites which led to spurious and non-ellipsoidal relationships between $v = \ln[s - \hat{\tau}_s]$ and $u = \ln[o - \hat{\tau}_o]$ left us with a total of 905 sites which are used in the following analyses. Table 2.1 summarizes the values of sample size n along with values of the coefficient of variation of the observations C_o and simulations C_s across the 905 samples.

Table 2.1 Statistics of Streamflow Records of 905 sites across the continental U.S. (see Farmer and Vogel, 2016ab)

Property	Average	Median	IQR (25 th , 75 th)	Range (min, max)
n	9827	10957	(9862, 10957)	(1262, 11322)
Co	2.3	1.9	(1.4, 2.7)	(0.3, 15.2)
Cs	2.0	1.4	(1.0, 1.9)	(0.2, 142.9)

5.2 *Goodness-of-fit of bivariate lognormal model to bivariate observations*

In this section, we assess the study assumptions summarized in Section 2 using the daily streamflow observations summarized in Table 2.1.

5.2.1 Assessment of bivariate lognormal approximation using probability ellipses

Our overall assumption that O and S arise from a bivariate LN3 process is equivalent to an assumption that the quantities $U = \ln[O - \tau_o]$ and $V = \ln[S - \tau_s]$ arise from a bivariate normal process. Numerous hypothesis tests of multivariate normality (MVN) exist; however, all such tests are based on data series which are serially independent. Given the extremely high degree of serial dependence, seasonality and other periodicities inherent in daily streamflow series, such hypothesis tests would not exhibit their reported type I or II error probabilities. In a review of MVN tests, Meklin and Mundfrom (2004) suggest that there is no clear favorite test; however, the most widely used tests in practice are the Mardia skewness and kurtosis tests (Mardia 1970). Using the p values of these two test statistics, we constructed scatterplots of the bivariate relationship between $v = \ln[s - \hat{\tau}_s]$ and $u = \ln[o - \hat{\tau}_o]$ for five watersheds which capture the complete range of goodness-of-fit, as shown in Figure 2.2. The p values associated with each of the Mardia test statistics are reported above each plot in Figure 2.2. If the observations were independent, the p values would reflect the probability of rejecting the MVN null hypothesis when it is true; thus, one would reject the null hypothesis of MVN for p values of less than 0.05, or so. However, since daily streamflow observations and simulations exhibit a very high degree of serial correlation, we avoid any conclusions concerning the likelihood of type I or II errors and only use the p values to evaluate goodness-of-fit. In general, goodness-of-fit improves as the p value increases.

To each scatterplot, we added two-dimensional confidence intervals, known as ‘probability ellipses’, using the method outlined in Example 10.1 of Wilks (2006) for a bivariate normal process. Figure 2.2 illustrates probability ellipses for the values of u and v corresponding to the five of the 905 watersheds, which are drawn to enclose 50% and 90% of the values of U and V . The probability ellipses were generated using the `dataEllipse` function from the “car” package in

R. For comparison, we include in the upper left panel of Figure 2.2 an example scatterplot and probability ellipses for synthetic MVN data. We conclude from Figure 2.2 that, even though we cannot formally accept or reject the MVN hypothesis, that hypothesis appears to provide a very good first approximation to the bivariate relationship between U and V .

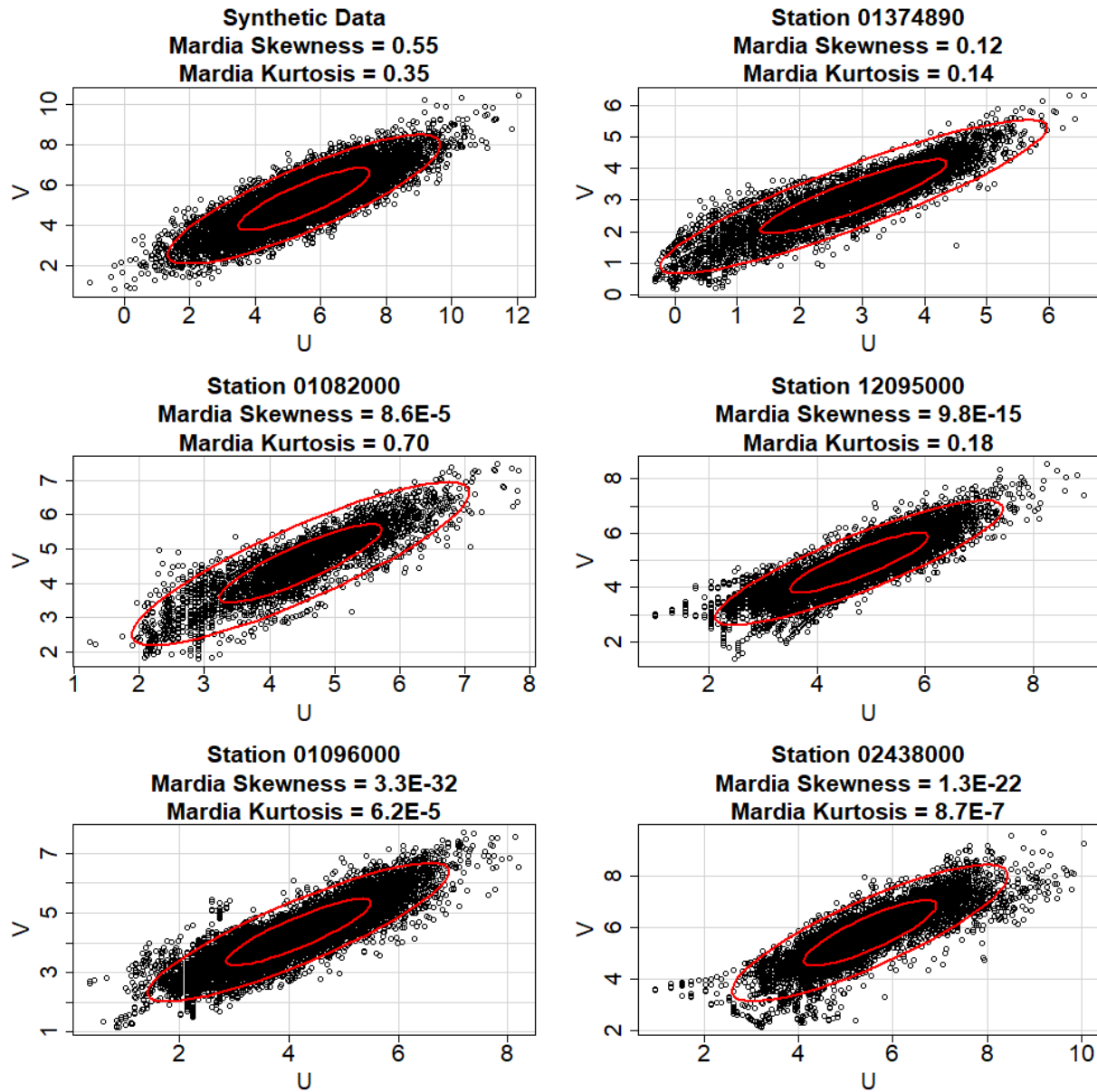


Figure 2.2 Goodness-of-fit evaluation of bivariate normality hypothesis with 50th and 95th confidence interval ellipses drawn. Upper left panel shows synthetic bivariate normal data and the remaining 5 panels are representative sites from the dataset.

5.2.2 Lognormal marginal distributions

In Figure 2.3 we employ L-moment diagrams to assess the goodness-of-fit of an LN2 and LN3 distribution to the actual o and s series. Hosking and Wallis (1997) and Vogel and Fennessey (1993) review the use of L-moment diagrams for use in assessing the goodness-of-fit of various probability distributions to observations. The top plots in Figure 2.3 contrast the theoretical relationship between L-Cv and L-skewness for an LN2 variate, shown using a solid curve, with estimates of those L-moment ratios at each of the 905 sites. Similarly, the bottom plots in Figure 2.3 contrast the theoretical relationship between L-kurtosis and L-skewness for an LN3 variate, shown using a solid curve, with estimates of those L-moment ratios at each of the 905 sites. What we observe from Figure 2.3 is that the observations and simulations are generally consistent with both the LN2 and LN3 hypotheses and, as expected, an LN3 model provides a better fit than the LN2 model because the lower plots exhibit less scatter about the theoretical relationship than the upper plots. These results are consistent with those of both Blum *et al.* (2017) and Limbrunner *et al.* (2000, Figure 6) who considered a larger set of sites across the USA and performed more detailed evaluations, including an analysis of the sampling variability to be expected from L-moment ratios computed from long daily streamflow series. On the basis of our results in Figure 2.3, combined with the results of Blum *et al.* (2017) and Limbrunner *et al.* (2000), we assume the marginal distribution of O and S may be roughly approximated by an LN3 distribution.

It is important to emphasize that we are not claiming that daily streamflow observations arise from an LN3 model. Blum *et al.* (2017) contrast L-moment diagrams computed from daily streamflow observations with L-moment diagrams arising from synthetic series in their Figure 2.2. On the basis of those experiments, they recommend use of a four-parameter kappa (KAP) distribution over an LN3 distribution for daily streamflow series, yet even a KAP distribution can

only provide a rough approximation to the complex distribution of daily streamflows. We are only claiming that the LN3 model provides a good first approximation to the general probabilistic behavior of both O and S , and is thus useful in documenting the behavior of estimates of the correlation coefficient when computed from actual streamflow observations. A natural extension to this study would be to explore the use of a bivariate kappa model, based on a Gaussian copula, for the purpose of evaluating and developing improved correlation estimators.

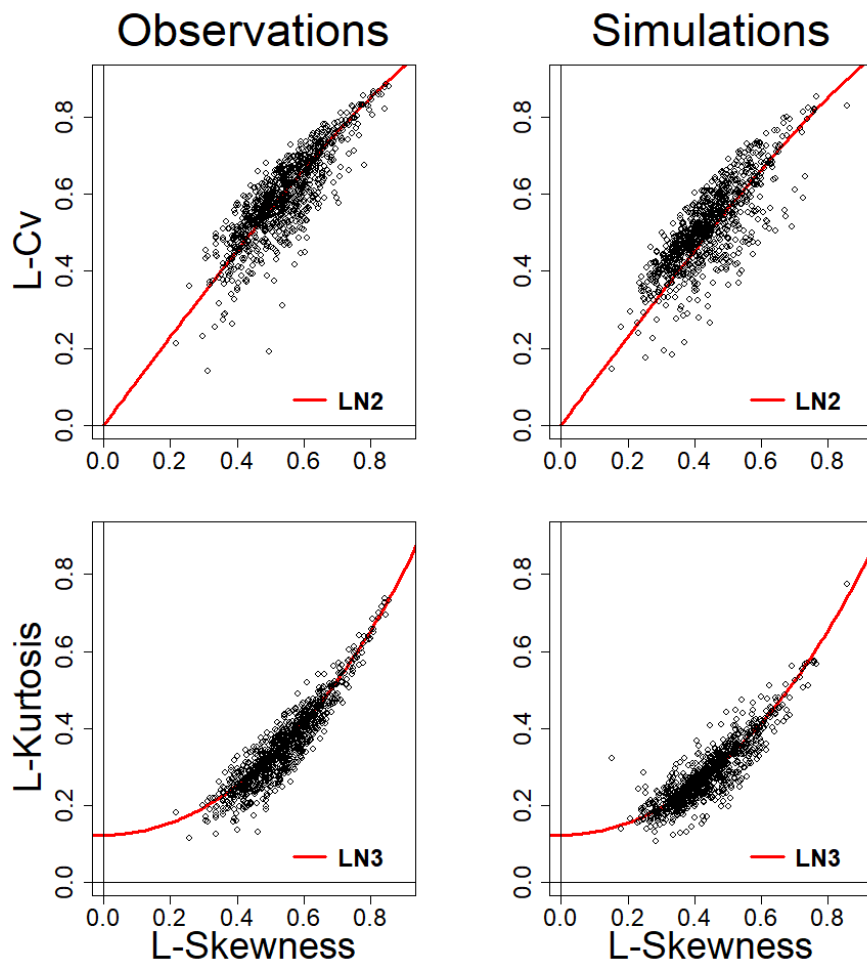


Figure 2.3 L-moment diagrams of observed and simulated average daily streamflows at the 905 sites summarized in Table 1.

5.2.3 Assessment of dependence structure

The bivariate LN3 model assumes a particular theoretical dependence structure given by Eq. (2.6). Here, we assess whether the correlation structure of the observations of O and S at the 905 sites summarized in Table 1, reproduce the theoretical dependence structure in Eq. (2.6) which relates the linear correlation between $U = \ln[O - \hat{\tau}_o]$ and $V = \ln[S - \hat{\tau}_s]$, termed ρ_{UV} , to the nonlinear relationship between O and S , termed ρ . In Figure 2.4, we assess the degree to which the theoretical relationship between ρ and ρ_{UV} in Eq. (2.6) is reproduced by the observations. Figure 2.4 illustrates scatterplots of the four estimators of ρ versus the corresponding estimates of ρ , which would be obtained by the application of Eq. (2.6). Equation (2.6) is the theoretical version of its equivalent sample estimator which is the Stedinger estimator r_1 given in Eq. (2.7). In other words, Stedinger's estimator r_1 is designed to reproduce, exactly, the theoretical dependence structure in Eq. (2.6). The important result in Figure 2.4 is that using actual streamflow observations, Pearson's estimator r does a very poor job of reproducing the theoretical dependence structure associated with the bivariate LN3 model, whereas the three other estimators, r_1 , r_2 and r_3 nicely reproduce that theoretical relationship. We conclude on the basis of Figure 2.2, Figure 2.3, and Figure 2.4 that the bivariate LN3 model approximately reproduces both the marginal distributions of O and S as well as their complex nonlinear dependence structure given in Eq. (2.6). We emphasize that future research is needed to explore more complex marginal distributions and nonlinear dependence structures, to enable derivation of more realistic estimators of correlation than introduced here.

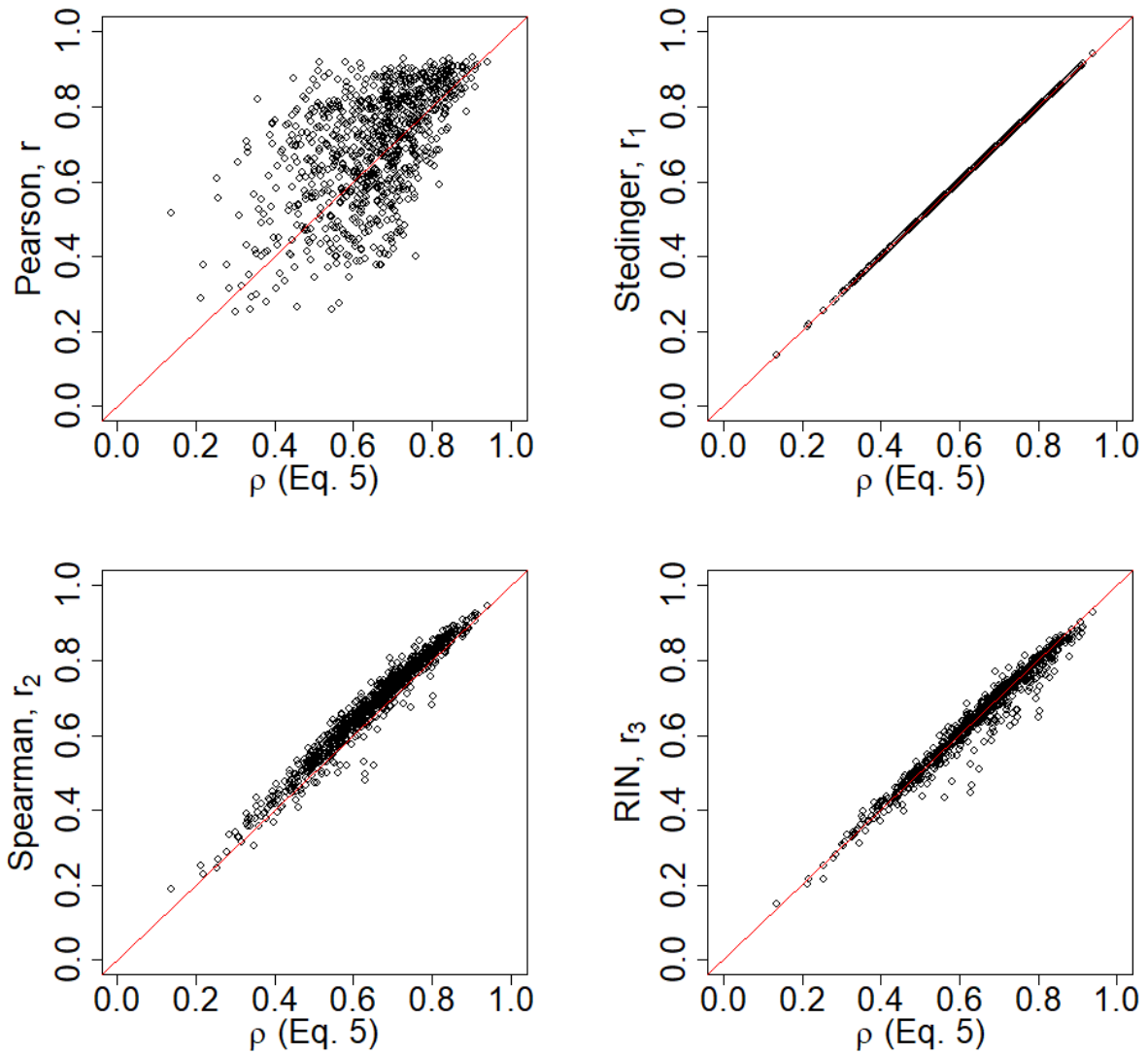


Figure 2.4 Evaluation of the ability of the four correlation estimators applied to the 905 sites summarized in Table 2.1, to reproduce the theoretical dependence structure associated with a bivariate LN3 process given by Equation 2.5

5.3 Comparisons among four correlation estimators

The left panels of Figure 2.5 compare the magnitude of the Pearson r given in Eq. (2.4) with the three competing correlation estimators r_1 , r_2 and r_3 given in equations (2.7), (2.12) and (2.13), respectively, using the actual streamflow simulations and observations at 905 sites across the USA with sample sizes n ranging from 1,262 to 11,322. Analogous comparisons are provided

in the right panels of Figure 2.5 based on synthetic bivariate LN3 samples generated to reproduce the sample sizes and sample moments of U and V associated with each of the O and S series at the 905 sites. The left panels of Figure 2.5 illustrate enormous variability associated with the estimator r compared with all three competing estimators r_1 , r_2 and r_3 . This result, based on actual daily streamflow observations and simulations, is to be expected on the basis of our previous Monte Carlo experiments reported in Figure 2.1, which demonstrated that the estimator r exhibits considerably more variability than any of the other estimators considered, over the wide range of conditions considered, even for very large sample sizes. Interestingly, the left panel of Figure 2.5 indicates that Pearson's r exhibits even greater variability when used with actual streamflow observations than when applied to synthetic bivariate LN3 data in the right panels of Figure 2.5. This result further illustrates that the theoretical bivariate LN3 model can only provide a rough approximation to the behavior of actual bivariate daily sequences of O and S . In other words, the correspondence between the left and right panels in Figure 2.5 provides the ultimate evaluation of the adequacy of the theoretical bivariate LN3 model for its ability to reproduce the sampling properties of the various correlation estimators. We recommend that future studies attempt to improve upon the results in Figure 2.5 by considering more representative marginal distributions for O and S , such as the KAP distribution and by using copulas which are more representative of the observed dependence structures than the Gaussian copula which is implied by the bivariate LN3 model in Eq. (2.6).

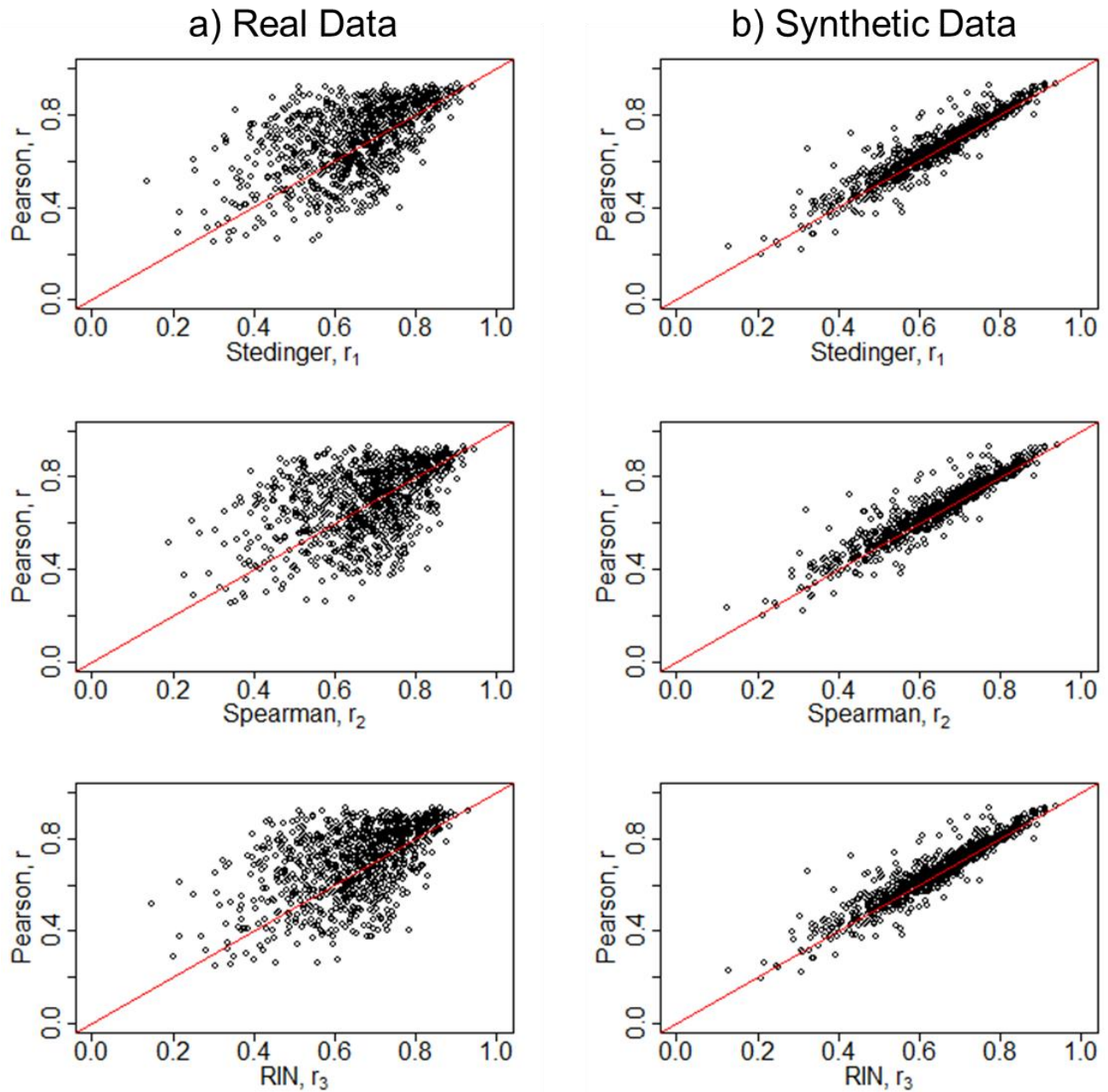


Figure 2.5 Comparison of Pearson's correlation estimator with the three alternative correlation estimators (Left) using observations and simulations at the 905 sites summarized in Table 2.1 and (Right) using synthetic bivariate lognormal series generated to reproduce the characteristics of the 905 sites summarized in Table 2.1.

5.4 *Impact of skewness on the sampling properties of Pearson's r*

Section 1.4 reviewed the sparse literature which summarizes the sampling properties of Pearson's r under non-normal conditions. Of critical importance, and an issue which does not appear to be addressed in any previous literature, is the tremendous sensitivity of Pearson's r to increases in skewness, even for very large sample sizes. This is analogous to, and highly related to, the tremendous sensitivity of all product moment ratio estimators to high values of skewness, reported by Vogel and Fennessey (1993). To document this issue, Figure 2.6 illustrates the expected difference between Pearson's r and Stedinger's r_1 , denoted $E[r - r_1]$, for synthetic bivariate LN3 samples generated to reproduce the sample moments of U and V associated with each of the O and S series at the 905 sites. Figure 2.6 reports the value of $E[r - r_1]$ versus the coefficient of variation of the observations computed using the LN2 estimator $C_o = \sqrt{\exp(\hat{\sigma}_u^2) - 1}$ where $u_i = \ln[o_i]$ and $\hat{\sigma}_u^2 = \frac{1}{n} \sum_{i=1}^n (u_i - \bar{u})^2$.

From our earlier Monte Carlo experiments summarized in Figure 2.1, we know that Pearson's r is generally upward biased and r_1 is generally unbiased for synthetic bivariate LN2 samples. Figure 2.6 illustrates the general and considerable increase in the upward bias associated with Pearson's r which results as the value of C_o increases. We conclude from Figure 2.6 that one should be very skeptical of estimates of Pearson's r arising from samples of daily, hourly and sub-hourly streamflow data which exhibit high variability, as evidenced by large values of C_o . We remind the reader to use L-moment ratios instead of product moment ratios when computing coefficients of variation, skewness and kurtosis, as recommended by Vogel and Fennessey (1993) and others.

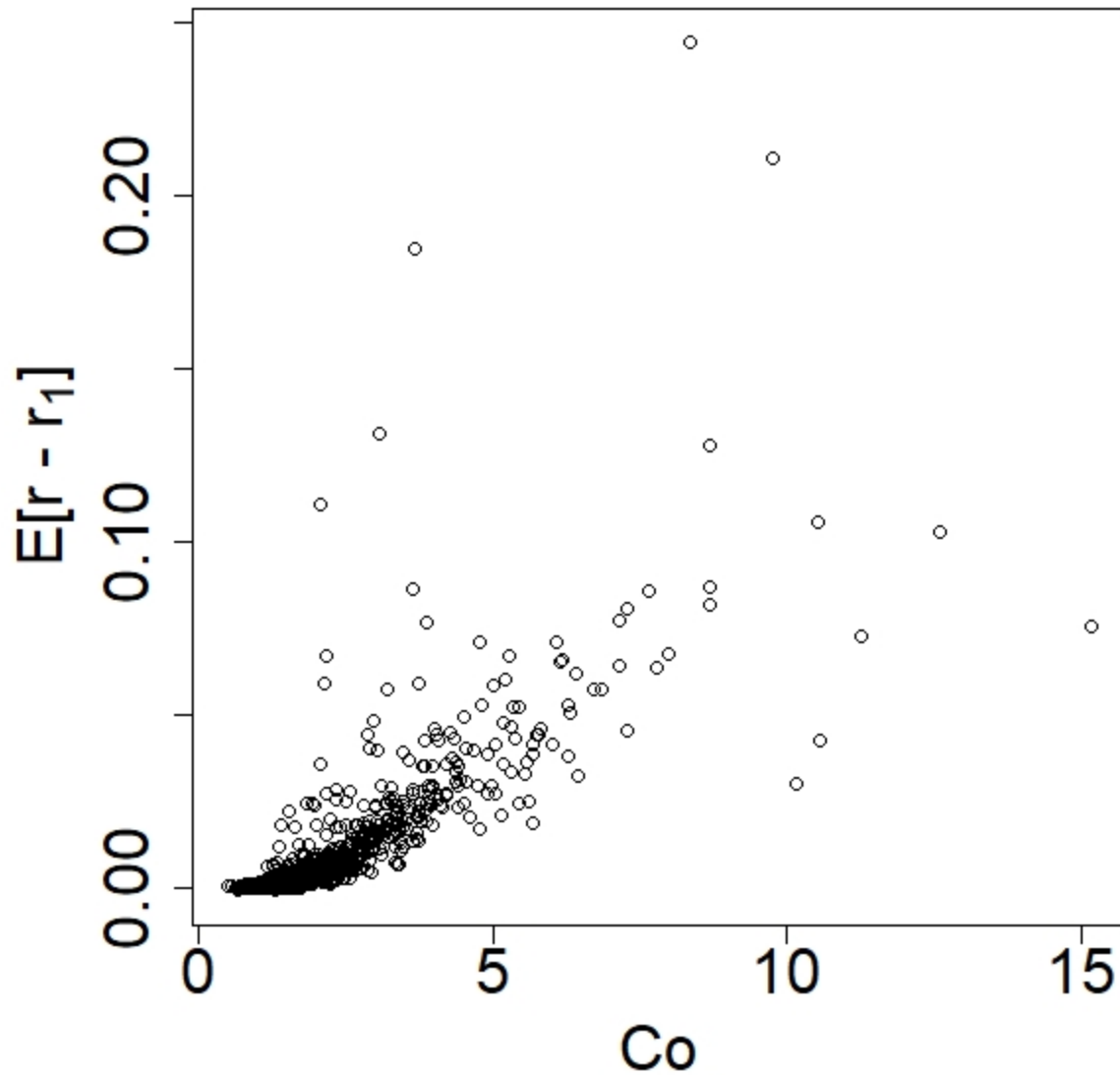


Figure 2.6 The expected difference between Pearson’s estimator r , and Stedinger’s estimator r_1 , as a function of the coefficient of variation of the observations. The values of $E[r - r_1]$ are computed from 500 synthetic bivariate lognormal traces generated to reproduce the characteristics of the bivariate observations and simulations at the 905 sites summarized in Table 2.1.

6 Conclusions

We have sought to evaluate the performance of, and to develop improved estimators for, the Pearson correlation coefficient ρ , which is widely used in the field of hydrology and water resources management: (a) for evaluations of goodness-of-fit of model simulations and observations and (b) in determination of the relationship among hydrologic variables. In our Monte Carlo experiments summarized in Figure 2.1, the widely used estimator r of the Pearson correlation coefficient was shown to exhibit significant upward bias and enormous variability for skewed bivariate lognormal samples and, importantly, that bias and variance does not disappear even for very large sample sizes in the thousands and even tens of thousands. While this result was demonstrated earlier by Lai *et al.* (1999), their message seems to be lost in the literature and, importantly, they did not do enough experiments to document the severe sensitivity of the sampling properties of the Pearson correlation r to increases in the variability and skewness of the bivariate data to be summarized, which is central to hydrologic studies, nor did they develop and evaluate improved estimators of ρ , as we have in this study.

We have discussed many previous hydrology studies which have criticized the behavior of the estimator r for its sensitivity to outliers, nonlinearity and non-normality, yet nearly every study failed to distinguish between the theoretical statistic ρ and the estimator r . Thus, those studies have incorrectly equated their criticisms of r with a critical evaluation of the theoretical statistic ρ . That logic would be like criticizing and dispensing with the expected value $E[X]$ because one of its estimators known as the sample mean \bar{x} is sensitive to large observations.

A central goal of this study was to uncover an important sampling problem associated with the Pearson correlation coefficient estimator r and to provide three estimators that, to first order,

should be improvements for the type of skewed samples encountered in hydrology. Our secondary goal was to provide guidance on when our estimators may be useful, but, more importantly, to provide recommendations for the future derivation of suitable estimators for bivariate samples that are known to exhibit more complex marginal distributions and dependence structures than the bivariate lognormal model assumed here. We have introduced a suite of three alternative estimators of ρ all of which were shown to exhibit less bias and variance than r for the types of skewed samples typically and increasingly encountered in hydrology. While the estimator r may perform reasonably well for annual and monthly hydrologic series, its performance degrades as the time interval decreases to daily, hourly and sub-hourly, thus warranting greater attention to this issue in the future. Our evaluations of the four alternative estimators of correlation were made using synthetic bivariate lognormal samples, as well as using actual bivariate samples of observations and simulations arising from the application of distributed rainfall-runoff model at 905 sites across the USA. Our evaluations led us to conclude that a bivariate lognormal model can only provide a first approximation to the behavior of actual bivariate daily streamflow series, but that it was instrumental in developing improved estimators of ρ which are much better suited to goodness-of-fit evaluations and evaluations of relationships among skewed hydrologic samples. We can only recommend use of the three improved estimators introduced here under conditions when bivariate samples are well approximated by a bivariate lognormal distribution. In practice, the bivariate lognormal model is likely to provide only a first-order approximation, thus we recommend that future research use the theory of copulas to develop improved correlation estimators based on marginal distributions such as the Kappa and Wakeby distributions (see Blum et al. 2017) as well as more accurate nonlinear dependence structures than exhibited by the bivariate lognormal model.

Ongoing work considers the impact of the bias and increased variance of ρ , demonstrated in this paper on NSE, an even more widely used goodness-of-fit metric in hydrology. Those ongoing investigations led us to realize that Pearson's ρ is simply a special case of NSE, because, for an unbiased model with serially independent residuals, $\text{NSE} = \rho^2$; thus, we felt that it would be important to begin our investigations by developing improved estimators for ρ , the subject of this initial study. Since NSE is a function of ρ , we expect to observe similar upward bias and increased variance associated with the commonly used real space estimator of NSE, as well as recent reported improvements in NSE termed the Kling-Gupta efficiency (KGE), introduced by Gupta *et al.* (2009), and the nonparametric efficiency estimator recently introduced by Pool *et al.* (2018). Those reported improvements in estimation of NSE arise from a burgeoning literature which has criticized the behavior of NSE. Interestingly, those criticisms of NSE, analogous to the criticisms of r reported here (Section 1.2), have confused the theoretical efficiency statistic with sample estimators such as NSE and KGE; thus, it would be very unlikely that improvements to estimation of a theoretical statistic could result without understanding the theoretical properties of that statistic. Importantly, every issue addressed and highlighted in this study is relevant to the development of improved estimators of NSE, the subject of an ongoing sequel to this study.

Acknowledgments

The authors are extremely grateful to Francesco Serinaldi for his very detailed and constructive review of two earlier versions of this manuscript, which led to considerable improvements. The authors are also grateful to associate editor Elena Volpi and two anonymous reviewers for their constructive comments which led to considerable improvements.

References

- Arbabshirani, M.R., Damaraju, E., Phlypo, R., Plis, S., Allen, E., Ma, S., Mathalon, D., Preda, A., Vaidya, J.G., Adali, T., Calhoun, V.D., 2014. Impact of Autocorrelation on Functional Connectivity. *NeuroImage*. <https://doi.org/10.1016/j.neuroimage.2014.07.045>.
- Astivia, O.L.O and B.D. Zumbo, 2017. Population models and simulation methods: The case of the Spearman rank correlation. *British Journal of Mathematical and Statistical Psychology*, 70, 347-367.
- Balakrishnan, N. and C.D., Lai, 2009. *Continuous Bivariate Distributions*, Dordrecht, The Netherlands: Springer, Second Edition.
- Beasley, T.M., and S. Erickson, 2009. Rank-Based Inverse Normal Transformations are Increasingly Used, But are They Merited? *Behaviour Genetics*, 39, 580-595.
- Bishara, A. J., & Hittner, J. B. 2015. Reducing bias and error in the correlation coefficient due to nonnormality. *Educational and Psychological Measurement*, 75, 785–804.
- Bishara, A. J., & Hittner, J. B., 2017. Confidence intervals for correlations when data are not normal. *Behaviour Research*, 49, 294-309.
- Blum, A.G., S.A. Archfield, and R.M. Vogel, 2017. The probability distribution of daily streamflow in the United States. *Hydrology and Earth System Sciences*, 21, 3093–3103, <https://doi.org/10.5194/hess-21-3093-2017>.
- Devlin, S.J., Gnanadesikan, R., Kettnering, J.R., 1975. Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62, 531–545, 1975.
- Embrechts P., McNeil A.J., Straumann D., 2002. Correlation and dependence in risk management: properties and pitfalls. In: Dempster MAH (eds) *Risk management: value at risk and beyond*. Cambridge, UK: Cambridge University Press, pp 176–223.
- Farmer, W. H., and R. M. Vogel, 2016a. On the deterministic and stochastic use of hydrologic models. *Water Resources Research*, 52, doi:10.1002/2016WR019129.
- Farmer, W.H., and Vogel, R.M., 2016b. On the Deterministic and Stochastic Use of Hydrologic Models: Data Release: US Geological Survey data release, <https://dx.doi.org/10.5066/F7W37TF4> .
- Fisher, R.A., 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10, 507-521.
- Guo, Y., A. Quader and J.R. Stedinger, 2016. Analytical estimation of geomorphic discharge indices for small intermittent streams. *Journal of Hydrologic Engineering*, 21(7).
- Genest, C. and F. Chebana, 2017. Copula modeling in hydrologic frequency analysis, Chapter 30 in: *Handbook of Applied Hydrology*, V.P. Singh ed.i.Chief, New York, NY: McGraw-Hill Education.

- Gupta, H.V., H. Kling, K.K. Yilmaz and G.F. Martinez, 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modeling. *Journal of Hydrology* 377, 80–91 .
- Habib, E., W.R. Krajewski, and G.J. Ciach, 2001. Estimation of rainfall interstation correlation. *Journal of Hydrometeorology*, 2, 621-629.
- Helsel, D.R. and R. M. Hirsch, 2002. *Statistical Methods in Water Resources Techniques of Water Resources Investigations*, Book 4, chapter A3. US Geological Survey.
- Helsel, D., R.M. Hirsch, K.R. Ryberg, S.A. Archfield, and E. Gilroy, 2019. *Statistical Methods for Water Resources*, 2nd edition, US Geological Survey, in press.
- Hosking, J. R. M. and Wallis, J. R., 1997. *Regional frequency analysis: an approach based on L-moments*, Cambridge, UK: Cambridge University Press.
- Johnson, N.L., S. Kotz and N. Balakrishnan, 1995. *Continuous univariate distributions*, Volume 2. New York, NY: Wiley.
- Kowalski, C.J., 1972. On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 21(1), 1-12.
- Krause P., Boyle D.P., Base F., 2005. Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, 29(5):89–97.
- Kruskal, W. H., 1958. Ordinal measures of association. *Journal of the American Statistical Association*, 53(284), 814-861.
- Legates, D.R. and R.E. Davis, 1997. The continuing search for an anthropogenic climate change signal: Limitations of correlation-based approaches. *Geophysical Research Letters*, 24(18), 2319-2322.
- Legates, D.R. and G.J. McCabe, 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, 35(1) 233-241.
- Limbrunner, J.F., R.M. Vogel, and L.C. Brown, 2000. Estimation of the Harmonic Mean of a Lognormal Variable. *Journal of Hydrologic Engineering*, 5(1), 59-66.
- Lai, C.D., J.C.W. Rayner and T.P. Hutchinson, 1999. Robustness of the sample correlation – The bivariate lognormal case. *Journal of Applied Mathematics & Decision Sciences*, 3(1), 7-19.
- Mardia, K.V., 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57, 519–530.
- Markstrom, S.L., Regan, R.S., Hay, L.E., Viger, R.J., Webb, R.M.T., Payn, R.A., and LaFontaine, J.H., 2015. PRMS-IV, the precipitation-runoff modeling system, version 4: US Geological Survey Techniques and Methods, Book 6, Chap. B7. <http://dx.doi.org/10.3133/tm6B>
- Matalas, N.C. and W.B. Langbein, 1962. Information content of the mean. *Journal of Geophysical Research*, 67(9), 3441-3448.

- McCuen, R.H., and W.M. Snyder, 1975. A proposed index for comparing hydrographs. *Water Resources Research*, 11(6) 1021-1024.
- Meklin, C.J., and D.J. Mundfrom, 2004. An Appraisal and Bibliography of Tests for Multivariate Normality. *International Statistical Review / Revue Internationale de Statistique*, 72,(1) 123-138.
- Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50, 885–900.
- Mostafa, M.D., and M.W. Mahmoud, 1964. On the Problem of Estimation for the Bivariate Lognormal Distribution. *Biometrika*, 51(3/4) 522-527.
- Papalexiou, S.M., 2018. Unified theory for stochastic modelling of hydroclimatic processes: Preserving marginal distributions, correlation structures, and intermittency. *Advances in Water Resources*, 115, 234-252.
- Pearson, K., 1896. Mathematical contributions to the theory of evolution III. Regression, heredity and panmixia. *Philosophical Transactions A*, 373, 253-318.
- Pool, S., M. Vis and J. Seibert, 2018. Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency. *Hydrological Sciences Journal*, 63:13-14, 1941-1953, DOI: 10.1080/02626667.2018.1552002
- Salvadori, G., De Michele, C., Kottegoda, N. T., & Rosso, R., 2007. *Extremes in nature: an approach using copulas* (Vol. 56). Springer Science & Business Media.
- Salvadori, G., & De Michele, C., 2013. Multivariate extreme value methods. In *Extremes in a changing climate* (pp. 115-162). Springer, Dordrecht.
- Serinaldi, F., 2008. Analysis of inter-gauge dependence by Kendall's τ , upper tail dependence coefficient, and 2-copulas with application to rainfall fields. *Stochastic Environmental Research and Risk Assessment*, 22:671-688.
- Shimizu, K., 1993. A bivariate mixed lognormal distribution with an analysis of rainfall data. *Journal of Applied Meteorology*, 32, 161–171.
- Spearman, C., 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15, 72-101.
- Stedinger, J.R., 1980. Fitting lognormal distributions to hydrologic data. *Water Resources Research*, 16(3) 481-490.
- Stedinger, J.R., 1981. Estimating correlations in multivariate streamflow models. *Water Resources Research*, 17(1) 200-208.
- Tsoukalas, I., A. Efstratiadis, A., and C. Makropoulos, 2018. Stochastic Periodic Autoregressive to Anything (SPARTA): Modeling and simulation of cyclostationary processes with arbitrary marginal distributions. *Water Resources Research*, 54, 161–185. <https://doi.org/10.1002/2017WR021394>

- Vogel, R.M. and N.M. Fennessey, 1993. L-Moment Diagrams Should Replace Product-Moment Diagrams. *Water Resources Research*, 29(6) pp 1745-1752.
- Vogel, R.M., J.R. Stedinger and R.P. Hooper, 2003. Discharge Indices for Water Quality Loads. *Water Resources Research*, 39(10), 1273, doi:10.1029/2002WR001872.
- Vogel, R.M., 2017. Stochastic watershed models for hydrologic risk management. *Water Security*, doi: <http://dx.doi.org/10.1016/j.wasec.2017.06.001>.
- Wilks, D.S., 2006. *Statistical Methods in the Atmospheric Sciences*, Second Edition, Academic Press, Burlington, MA.
- Willmott, C.J., 1981. On the validation of models, *Physical Geography*, 2, 184-194.
- Willmott, C. J., S. G. Ackleson, R. E. Davis, J. J. Feddema, K. M. Klink, D. R. Legates, J. O'Donnell, and C. M. Rowe, 1985. Statistics for the evaluation and comparison of models, *Journal of Geophysical Research*, 90, 8995-9005.
- Xiao, Q., 2014. Evaluating correlation coefficient for Nataf transformation. *Probabilistic Engineering Mechanics*, 37, 1-6.
- Xu, W., Hou, Y., Hung, Y. S., & Zou, Y. 2013. A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models. *Signal Processing*, 93, 261–276. <http://dx.doi.org/10.1016/j.sigpro.2012.08.005>
- Zhang, G. and Z. Chen, 2015. Inferences on correlation coefficients of bivariate log-normal distributions. *Journal of Applied Statistics*, 42:3, 603-613, DOI: 10.1080/02664763.2014.980786

Appendix. Generation of bivariate lognormal streamflow series

We describe here a methodology for generating bivariate two- (LN2) and three-parameter (LN3) lognormal series. Balakrishnan and Lai (2009) introduce a bivariate LN2 model and review numerous applications of bivariate lognormal series in a variety of different fields. Without any loss of generality, we assume that the mean of both series equal unity so that $\mu_o = \mu_s = 1$. For assumed values of the coefficient of variation of the observations $C_o = \sigma_o / \mu_o$, and simulations $C_s = \sigma_s / \mu_s$, 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

$$\mu_U = \ln \left[\frac{\mu_o - \tau_o}{\sqrt{1 + \left(\frac{\sigma_o}{\mu_o - \tau_o} \right)^2}} \right] \quad \sigma_U = \sqrt{\ln \left[1 + \left(\frac{\sigma_o}{\mu_o - \tau_o} \right)^2 \right]} \quad (\text{A1a})$$

$$\mu_V = \ln \left[\frac{\mu_s - \tau_s}{\sqrt{1 + \left(\frac{\sigma_s}{\mu_s - \tau_s} \right)^2}} \right] \quad \sigma_V = \sqrt{\ln \left[1 + \left(\frac{\sigma_s}{\mu_s - \tau_s} \right)^2 \right]} \quad (\text{A1b})$$

Note that 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 were inputs to the experiments, and not estimated from data. One approach to generation of bivariate LN3 streamflows, is to first generate the observations O , from the lognormal quantile function

$$O_i = \tau_o + \exp[\mu_U + z(p_i)\sigma_U] \quad (\text{A2})$$

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 LN3 variates is easily implemented by making use of the log space regression so that

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

with errors ε_i generated from a normal distribution with zero mean and variance equal to

$$\sigma_\varepsilon^2 = \sigma_V^2(1 - \rho_{UV}^2).$$

Chapter 3

Improved Estimators of Efficiency For Goodness-of-fit of Skewed Hydrologic Data

with Jonathan R. Lamontagne & Richard M. Vogel

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 different theoretical (probabilistic) definitions of efficiency, E and E' , based on the estimators NSE and KGE , respectively, which enable us to derive improved estimators of E and E' which are shown to yield considerable improvements over NSE and slight improvements over KGE in controlled Monte-Carlo experiments. We document that although NSE is a nearly unbiased estimator, it exhibits enormous variability due to the remarkable skewness and periodicity of daily streamflow data, whereas NSE performs well for bivariate nonperiodic normal data. Dozens of studies have criticized NSE , but all such studies have failed to distinguish between the probabilistic properties of E and the sampling properties of NSE . Thus previous criticisms of NSE are misinterpreted as a drawback of the population E . Our new bivariate lognormal monthly mixture estimator of E should avoid most previous criticisms of E implied by the literature. Improved estimators of E which account for skewness and periodicity are needed for daily and subdaily streamflow series because NSE is not suited to such applications.

1 Introduction

To appreciate the primary contribution of this paper, we begin with a simple example. Consider the expected value of a random variable x , termed $\mu_x = E[x]$ with a very formal probabilistic or theoretical definition, given by $E[x] = \int xf(x)dx$ where $f(x)$ is the probability distribution (pd) of x . Knowledge of this probabilistic definition is absolutely central to casinos, lotteries, and the flood insurance program (FIP), because without that knowledge, they would be unable to use that probabilistic definition to ensure that their revenues exceed their costs over the long term, or on expectation. To evaluate long term profits, they might use the common sample estimator $\bar{x} = \sum_{i=1}^n x_i/n$ to estimate the true value of μ_x . If instead, the casino, lottery or FIP used the median value of x instead of μ_x to estimate their revenues, they would likely go out of business if x arises from a pd which exhibits right hand skew. Some may criticize the statistic \bar{x} as a reliable estimator of μ_x because it is heavily influenced by outliers, which often occur in highly skewed samples. That would be a reasonable criticism because other estimators of μ_x exist which are less subject to outliers [Helsel et al. 2019]. However, it would be unreasonable to criticize the theoretical statistic μ_x simply because one of its sample estimators performs poorly. This is exactly what has happened in the hydrologic literature concerning the well-known Nash-Sutcliffe Efficiency (*NSE*). Numerous concerns have been raised about the sensitivity of the estimator *NSE* to outliers and other issues, when in fact the theoretical or probabilistic statistic, upon which it is based, is entirely independent of the properties of data used in its estimation. This is the central focus of this paper.

Consider the problem of evaluating the goodness-of-fit of model output to observations. Every model has both a deterministic and stochastic element so that a simulated response S is obtained from the sum of the deterministic model $H(X|\Omega)$ and a stochastic model error component ε

$$S = H(X|\Omega) + \varepsilon \quad (3.1)$$

where X denotes some set of model input variables and Ω denotes the set of deterministic model parameters. Estimates of the deterministic model parameters are obtained by fitting the model to a set of observations, O . This step, known as model calibration, is usually accomplished with an optimization algorithm which yields an estimate of the model parameters $\hat{\Omega}$ by minimizing the model error term ε . Once calibrated, model simulations S , are often obtained by dropping the model error term and generating the simulated response using

$$S = H(X|\hat{\Omega}) \quad (3.2)$$

Using model simulations from thousands of deterministic rainfall-runoff models, *Farmer and Vogel* [2016a] show that the common practice of dropping the model error term as in Eq. (3.2), leads to systematic upward/downward bias in the simulation of hydrologic extremes such as droughts/floods. *Vogel* [2017] suggests a generalized approach termed ‘stochastic watershed models’ which involves simulation of both the deterministic and stochastic elements in Eq. (3.1) to avoid the systematic simulation bias elaborated by *Farmer and Vogel* [2016a].

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

$$O = S + \varepsilon \quad (3.3)$$

Numerous statistics have been introduced, over the years, for evaluating the goodness-of-fit of model simulations to observations. Within the field of hydrology, the most widely used goodness-of-fit index, in this context, is the Nash-Sutcliffe Efficiency (*NSE*) (*Nash and Sutcliffe, 1970*) as evidenced by over 18,000 Google Scholar citations (December 10, 2019) to their paper 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.

This is the first study which draws a distinction between the theoretical statistic which we term efficiency *E*, and its common sample estimator *NSE*. The theoretical statistic *E*, is simply a standardized form of the mean square error (*MSE*) so that

$$MSE = E[(S - O)^2] \quad (3.4a)$$

and

$$E = 1 - \frac{MSE}{E[(O - \mu_o)^2]} = 1 - \frac{MSE}{\sigma_o^2} \quad (3.4b)$$

where $E[\]$ denotes the expectation operator, *O* and *S* represent the simulated and observed time series, respectively, and μ_o and σ_o^2 denote the mean and variance of the observations. Both *MSE* and *E* are defined by the expectation operators in Eq. (3.4a) which are grounded in the theory of probability as distinguished from the theory of statistics which 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 , $i=1, \dots, n$, that one needs to replace the expectation operator with estimators of MSE and E in Eq. (3.4) so that

$$\overline{MSE} = \frac{1}{n} \sum_{i=1}^n (s_i - o_i)^2 \quad (3.5a)$$

$$NSE = 1 - \frac{\overline{MSE}}{s_o^2} = 1 - \frac{\frac{1}{n} \sum_{i=1}^n (s_i - o_i)^2}{\frac{1}{n-1} \sum_{i=1}^n (o_i - \bar{o})^2} \quad (3.5b)$$

where NSE is the popular estimator introduced by *Nash and Sutcliffe* [1970]. Note that it is common practice to use upper O, and lower o, case values to denote the theoretical values and their realizations. Similarly it is common practice to use Greek for the theoretical, population or probabilistic mean and variance μ_o and σ_o^2 and to use \bar{o} and s_o^2 to denote sample estimates of those same statistics based on data.

The theoretical efficiency statistics MSE and E in Eq. (3.4a) have advantages over other metrics such as the correlation coefficient between S and O termed ρ , because they are an amalgamation of both bias and variance associated with model simulations S . Unfortunately, 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 .

Over the years, numerous authors have discussed and evaluated the behavior of the estimator NSE [see for example: *Pool et al.* 2018; *Liu et al.* 2018; *Bardsley*, 2013; *Gupta and Kling*, 2011; *Gupta et al.*, 2009; *Jain and Sudheer*, 2008; *Schaepli and Gupta* 2007; *McCuen et al.* 2006; *Legates and McCabe*, 1999; *Martinec and Rango*, 1989, 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 and Winston 2008], an index related to both NSE and ρ [Bardsley, 2013], a slight variant of NSE termed the coefficient of gain (WMO, 1986; and Martinec and Rango, 1989), 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 which have been advanced to improve the performance of NSE.

What is common to nearly all of the above cited studies are that they only use 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 is always unknown in such situations. Our approach is unique because we first introduce a theoretical definition of the true value of efficiency in section 2 followed by implementation of controlled Monte-Carlo experiments to rigorously evaluate the performance of various alternative estimators of efficiency E including NSE as well as improved estimators introduced here. Finally, we test out our findings and recommendations using the output of hundreds of calibrated PRMS rainfall-runoff models analogous to the recent work of *Farmer and Vogel* [2016a] and *Barber et al.* [2019].

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] and many others reviewed by *Liu et al.* [2018], all of which have sought to improve our understanding of the sampling properties of the most widely used estimator of efficiency 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 a reliable hypothesis test or confidence interval for the true value of efficiency without a theoretical definition of E and, to our knowledge, this is the first study which has distinguished the theoretical (or true value) of E from the estimator NSE . Thus all previous studies which introduced confidence intervals for the true value of E were unable to rigorously evaluate the ability of random confidence intervals to provide proper coverage of the true value of E because those studies did not have knowledge of the true value of E . Similarly, previous studies which advanced hypothesis tests concerning the true value of E were unable to rigorously evaluate the likelihood of either type I or type II errors because they did not have knowledge of the true value of E . Second, our study introduces a new estimator of E which has considerable advantages over NSE for highly skewed and periodic bivariate daily streamflows which exhibit significant periodicity (seasonality), thus 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.

Unfortunately, nearly all previous studies cited above which discuss the properties of MSE and NSE have failed to distinguish between their theoretical values defined above, and sample estimates of those values derived from actual data and models, discussed below. Thus, they have confused the subjects of probability and statistics, because their theoretical treatment begins with empirical estimators of both MSE and E based on samples of data of length n . 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 of NSE they have used 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 Eq. (3.4a) do not depend on data or sample size.

2 Theoretical Development of Efficiency E

In the following section we introduce theoretical (probabilistic) expressions for efficiency which are based on both 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. All previous analyses, discussions and criticisms of these two statistics were made using data and estimators without resorting to probabilistic analyses as in this section and later on in section 6.

2.1 Theoretical Efficiency Based on Nash-Sutcliffe

The definition of MSE in Eq. (3.4a) 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 Eq. (3.1) so that

$$E[(S - O)^2] = E[\varepsilon^2] = MSE[\varepsilon] \quad (3.6)$$

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

$$MSE[\varepsilon] = E[\varepsilon - E(\varepsilon)]^2 + E[(\varepsilon - E[\varepsilon])^2] = Bias(\varepsilon)^2 + Var(\varepsilon) \quad (3.7)$$

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

Using the theory of probability, one can show by expanding the expectation in Eq. (3.6) that

$$MSE[\varepsilon] = (\mu_o - \mu_s)^2 + \sigma_o^2 + \sigma_s^2 - 2\sigma_s\sigma_o\rho \quad (3.8)$$

where μ_o and μ_s denote the means of O and S , respectively, σ_o^2 and σ_s^2 denote the variances of the O and S , respectively, and ρ denotes the Pearson correlation between O and S , respectively. The expansion in Eq. (3.8) 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.

A 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 Eq. (3.5b) 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 Eq. (3.8) as

$$MSE[\varepsilon] = \Delta^2\mu_o^2 + \sigma_o^2[1 + \alpha^2 - 2\alpha\rho] \quad (3.9)$$

where $\Delta = \frac{\mu_o - \mu_s}{\mu_o}$, and $\alpha = \frac{\sigma_s}{\sigma_o}$. Here Δ is the bias as a fraction of the mean 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 Eq. (3.9) and *Gupta et al.* [2009], *Pool et al.* [2018] and others, is that they employ sample estimates of the various terms in Eq. (3.9) without referring to their true values.

We have chosen to introduce the bias as a fraction of the mean observations Δ in Eq. (3.9) because this form of standardized bias is easy to compare and contrast across models or watersheds, and is consistent with the traditional (statistical) definition of bias given in Eq. (3.7), 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 Eqs. (3.9) and (3.4b) leads to

$$E = 2\alpha\rho - \alpha^2 - \frac{\Delta^2}{C_o^2} \quad (3.10)$$

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

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 Eqs. (3.4b) and (3.10). Using their notation along with our notation, their implied definition of theoretical efficiency E' would be

$$E' = 1 - \sqrt{(\beta - 1)^2 + (\alpha - 1)^2 + (\rho - 1)^2} \quad (3.11)$$

where $\beta = 1 - \Delta = 1 - \frac{\mu_o - \mu_s}{\mu_o}$ with ρ and $\alpha = \frac{\sigma_s}{\sigma_o}$ defined previously.

The theoretical efficiency E' in Eq. (3.11) introduced in its empirical form by *Gupta et al.* [2009] could have several advantages over the definition of E in Eq. (3.10) as is shown later on and discussed by *Gupta et al.* [2009]; *Knoben et al.* [2019] and others. In this section, 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 theoretical 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 datasets.

To better understand the differences in the behavior of E and E' , in Eqs. (3.10) and (3.11) respectively, one can derive their ratio for an unbiased model ($\Delta = 0$) as

$$\frac{E}{E'} = \frac{\alpha^2 - 2\rho\alpha}{\sqrt{(\rho - 1)^2 + (\alpha - 1)^2} - 1} \quad (3.12)$$

Figure 3.1 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 $\alpha = \rho$. Otherwise, for the more common 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 3.1.

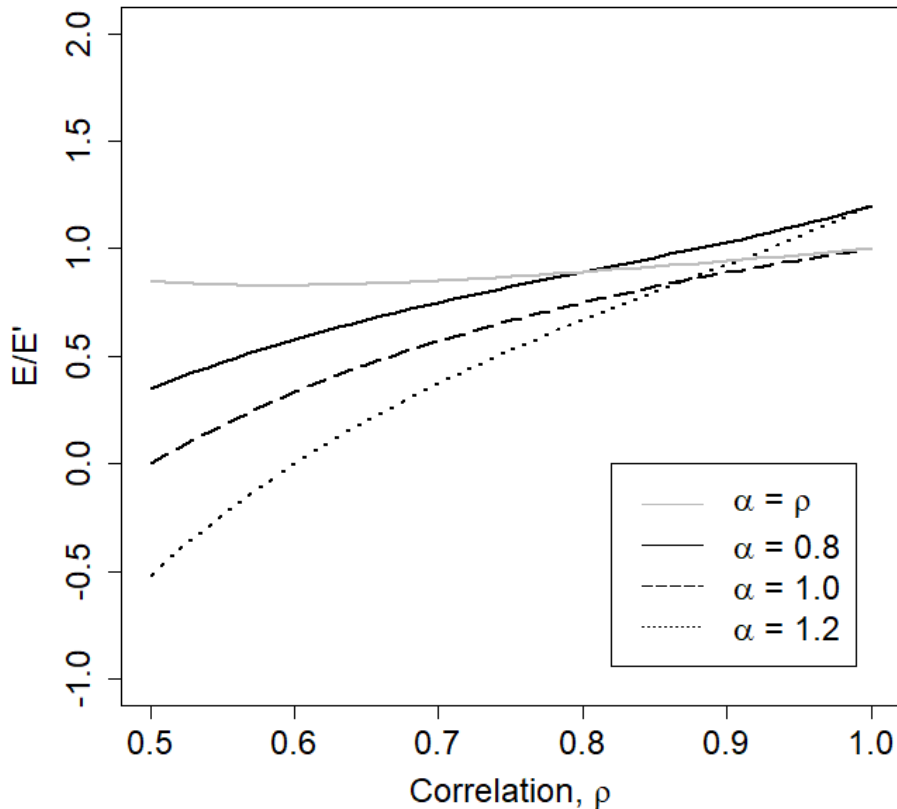


Figure 3.1 Ratio of Efficiency E Based on NSE to the Efficiency E' Based on KGE' as a Function of α and ρ , or an unbiased model.

3 Study Assumptions: Daily Streamflow Simulations and Observations

Here, as in *Farmer and Vogel* (2016a), 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 U.S. The distributed-parameter model, in this case, the Precipitation-Runoff Modeling System [*Markstrom et al.*, 2015], was calibrated at each of 1,225 perennial river basins across the conterminous United States. Details and availability of the datasets are described by *Farmer and Vogel* [2016b]. The particulars of the model and the calibration scheme are not relevant to our experiments.

An experienced hydrologist would not only consider quantitative goodness-of-fit metrics, but would also perform graphical evaluations to ensure consistent and sensible behavior between the observations, o and the simulations s . To mimic the work of a hydrologist, we examined every scatterplot of $v = \ln(s - \hat{\tau}_s)$ versus $u = \ln(o - \hat{\tau}_o)$ to ensure that they mimic the type of behavior expected from such analyses. Here $\hat{\tau}_o$ and $\hat{\tau}_s$ are Stedinger's (1980) lower bound estimator of an LN3 distribution (also see equation 18). Our experience indicates that one expects an approximately ellipsoidal relationship between u and v which would be consistent with the assumption of a bivariate LN3 relationship between o and s . Removing those sites which led to spurious and non-ellipsoidal relationships between $v = \ln(s - \hat{\tau}_s)$ and $u = \ln(o - \hat{\tau}_o)$ left us with a total of 905 sites. To ensure enough streamflow data to inform our Monte-Carlo experiments, we 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 which 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, as well as sites which had a correlation ρ value outside the range of $[0.5, 1]$. We also

restricted the analysis to sites with a C_o value greater than 0.5. As is described in section 3.2, a bivariate three parameter lognormal mixture model is able to capture the periodicity that daily flow series often exhibit. We therefore also restrict our analysis to sites which have a higher probability plot correlation coefficient from the BLN3 mixture model than when single BLN3 model is applied. After applying the described restrictions, a total of 447 sites were found to be suitable for our study and are used in the following analyses. Table 3.1 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. Also shown in Table 3.1 are some of the estimators of efficiency described below. Estimators of all the statistics in Table 3.1 are based on the bivariate lognormal monthly mixture model which provides more reliable estimates of all of these statistics than alternative methods, as is shown later on.

Table 3.1 Values of n , C_o , C_s , α , Δ , ρ , LBE_m , NSE , LBE'_m and KGE Corresponding to Daily Streamflow Observations and Simulations at 447 USGS Gaged Watersheds

Property	Average	Median	IQR (25 th , 75 th)	Range (min, max)
n	10,944	10,957	(10,957, 10,957)	(10,014, 11,322)
C_o	1.54	1.42	(1.32, 1.73)	(0.51, 6.30)
C_s	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)

3.1 *Bivariate Three Parameter Lognormal Model (BLN3)*

Daily, hourly and sub hourly streamflow are known to exhibit extremely high values of skewness, so that typical observations O , and simulations S , in Eqs. (3.1)-(3.3) are much more closely approximated by a bivariate three parameter lognormal model (BLN3), than a bivariate normal model as was recently shown by *Barber et al.* [2019]. For example, *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 very good first approximation to the probability distribution (pd) of daily streamflow observations for hundreds of stations across the conterminous US. A much better approximation than the BLN3 model is introduced in the next section which deals with fact that daily streamflow observations are periodic and hence are not identically distributed.

Barber et al. [2019, Figure 2; Thesis: Figure 2.2] also use scatterplots with bivariate two parameter lognormal BLN2 confidence regions (or probability ellipses) to document the approximate, yet impressive goodness-of-fit of the BLN3 model at 905 sites across the conterminous U.S. They also document that the BLN3 is equivalent to a Gaussian copula with an LN3 marginal pd. *Barber et al.* [2019] also perform independent evaluations which verify that the nonlinear dependence structure implied by the BLN3 model is reproduced by the bivariate series of O and S .

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; Thesis Figure 2.5] found that generation of synthetic streamflow series from a BLN3 model was not entirely adequate to reproduce the behavior of estimates of r derived from the actual streamflow observations and simulations. A major innovation over *Barber et al.* [2019] which we consider in

section 3.2 is the use of a BLN3 monthly mixture model to better represent the highly complex marginal pds which result from the complex seasonal and other periodic behavior embedded within daily streamflow series.

3.2 A BLN3 Monthly Mixture Model for Reproduction of the Probability Distribution 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 probability distributions (pds). A mixture model is needed to account for the strong deterministic signals associated with seasonal and other causes of periodicity within the daily flow series which give rise to such multimodal behavior in the pd. We introduce a monthly mixture model which involves fitting a separate BLN3 model to the daily streamflows in each month. We employ this 36 parameter BLN3 monthly mixture model to generate synthetic streamflow series which better mimic the marginal distribution of the observations and simulations than a single BLN3 model and we also use that model to provide improved estimators of all the statistics reported in this study.

Consider a mixture distribution made up of fitting a separate LN3 distribution $f(o; \mu_i, \sigma_i, \tau_i)$ to the daily streamflows in each of $i=1, \dots, 12$, 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 $u = \ln(o - \tau_i)$. The resulting mixture pd of all the daily streamflows is given by

$$f(o; \mu_1, \dots, \mu_{12}, \sigma_1, \dots, \sigma_{12}, \tau_1, \dots, \tau_{12}) = \sum_{i=1}^{12} w_i f_i(o; \mu_i, \sigma_i, \tau_i) \quad (3.13)$$

where $f()$ denotes the overall pd of the observations and $f_i()$ denotes the pd of the observations in month i , and $\sum_{i=1}^{12} w_i = 12$. Assuming each month has the same number of days, $w = w_i = 1/12$.

To evaluate the goodness-of-fit of the mixture model in Eq. (3.13) we construct probability plots and summarize their goodness-of-fit using the well known probability plot correlation coefficient (PPCC) statistic. We employ 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 Eq. (3.13) which leads to:

$$F(o; \mu_1, \dots, \mu_{12}, \sigma_1, \dots, \sigma_{12}, \tau_1, \dots, \tau_{12}) = \frac{1}{12} \sum_{i=1}^{12} F_i(o; \mu_i, \sigma_i, \tau_i) \quad (3.14)$$

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

For the BLN3 mixture model a pp probability plot is constructed by first ranking all the values of the daily streamflow observations $o_{(j)}$ $j = 1, 2, \dots, n$ where n is the total record length of the daily flow series. Each of those observed daily flows $o_{(j)}$ have the same rank as their transformed flows $u_{(j)} = \ln(o_{(j)} - \hat{\tau}_i)$ for $i = 1, 2, \dots, 12$ and $j = 1, 2, \dots, n$. Under the LN3 hypothesis, O follows an LN3 distribution and $U_{(j)} = \ln(O_{(j)} - \tau_i)$ follows a normal distribution so that 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 mixture distribution obtained from

$$\hat{F}(u_{(j)}) = \frac{1}{12} \sum_{i=1}^{12} \Phi \left(\frac{u_{(j)} - \bar{u}_i}{s_{u,i}} \right) \quad (3.15)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a normal variable and \bar{u}_i and $s_{u,i}$ are the mean and standard deviation of the transformed flows $u_{(j)} = \ln(o_{(j)} - \hat{\tau}_i)$ in month i . A Weibull plotting position is suitable here, because it yields an unbiased estimate of the cumulative probability associated with 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 $\hat{F}(u_{(j)})$ obtained from Eq. (3.15). Figure 3.2 uses boxplots and a scatterplot to summarize the PPCC values associated with the BLN3 monthly mixture model (denoted $PPCC_m^2$) versus the PPCC value of fitting a single BLN3 model (denoted $PPCC_{LN3}^2$) to the entire n day series. Figure 3.2 documents the considerable improvement in the goodness-of-fit of the 36-parameter BLN3 monthly mixture model over the single BLN3 model used by Barber *et al.* [2019] at all of the 447 sites considered in this study.

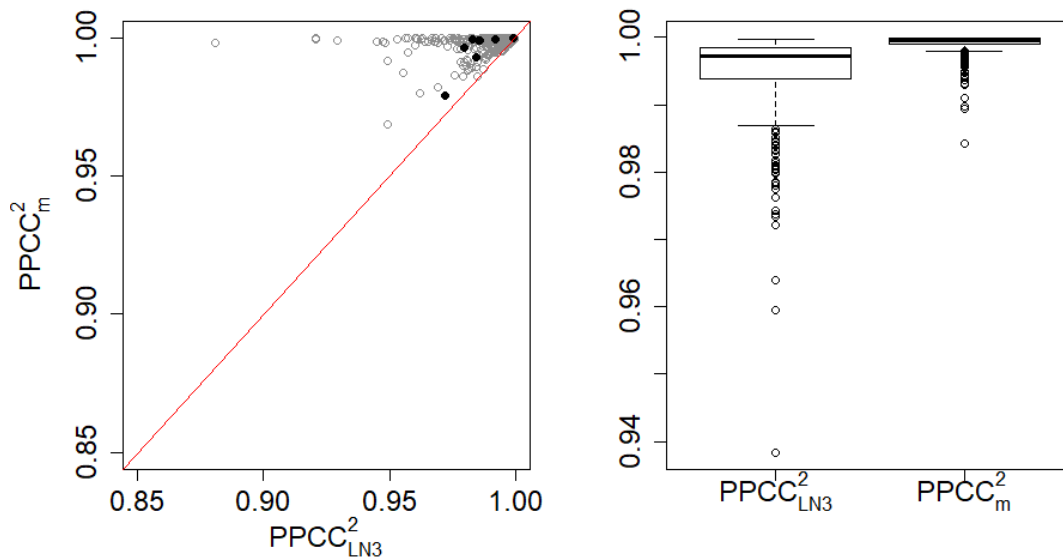


Figure 3.2 The Square of the Probability Plot Correlation Coefficients for the BLN3 monthly mixture model $PPCC_m^2$ and a single BLN3 model $PPCC_{LN3}^2$

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, which will all be compared and evaluated in our subsequent Monte-Carlo experiments in Section 5.

4.1 Nash-Sutcliffe Efficiency (NSE)

This estimator of theoretical efficiency E defined in Eq. (3.4b) and Eq. (3.10) was first introduced by *Nash and Sutcliffe* [1970] and is given in Eq. (3.5b).

4.2 Natural Log Nash-Sutcliffe Efficiency (LNSE)

This estimator requires taking the natural log of the observations o and simulations s prior implementing the Nash-Sutcliffe efficiency given in Eq. (3.5b).

4.3 Kling-Gupta Efficiency (KGE')

Gupta et al. [2009] developed an estimator of E' defined in Eq. (3.11), which is based on a different (nonequivalent) form of the expression for E given in Eqs. (3.4b) and (3.10). Using their notation along with our notation, combined with 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:

$$KGE' = 1 - \sqrt{(\hat{\beta} - 1)^2 + (\hat{\alpha} - 1)^2 + (\hat{\rho} - 1)^2} \quad (3.16)$$

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

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

In equation Eq. (3.16) we employ the notation KGE' instead of the usual notation KGE , to highlight that KGE' is an estimator of E' in Eq. (3.11), and is NOT necessarily a good estimator of E in Eq. (3.10) as was so clearly shown in Figure 3.1.

On the one hand there are several important advantages to the theoretical statistic E' introduced and outlined by *Gupta et al.* [2009] however there are several potential concerns with KGE' , apart from the important fact that it is based on a different theoretical definition of efficiency than that introduced in Eqs. (3.4b) and (3.10). One obvious problem with KGE' is that it is based entirely on product moment estimators for all the components given in Eq. (3.11). 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 ratio 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 Eq. (3.11) 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 normally distributed observations, but was shown by *Barber et al.* [2019] to perform poorly for BLN3 streamflows, because it exhibits considerable upward bias and extreme variability compared to the BLN3 and nonparametric estimators of ρ they introduced.

4.4 Non-Parametric Efficiency (PVSE')

To address the concerns raised above for KGE' , *Pool et al.* [2018] developed an estimator of E' in Eq. (3.11), which we term the Pool-Vis-Siebert estimator ($PVSE'$), which employs nonparametric estimators for each of the three components given by:

$$PVSE' = 1 - \sqrt{(\hat{\beta} - 1)^2 + (\hat{\alpha} - 1)^2 + (\hat{\rho} - 1)^2} \quad (3.17)$$

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 simply applying Pearson product moment estimator $\hat{\rho} = r$ given

in Eq. (3.16) 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 Eqs. (3.10) and (3.11) (see *Barber et al.*, 2019) and their nonparametric estimator $\hat{\alpha}$ is an estimator of a different theoretical statistic than α defined in Eqs. (3.10) and (3.11). Nevertheless, as we show later on $PVSE'$ does have attractive sampling properties.

4.5 *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 Eqs. (3.10) and (3.11), 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 variables of interest (e.g. O and S , or the streamflow values at two stations) follow a BLN3 monthly mixture model, which provides a good approximation to actual observations across the U.S. as shown recently by *Barber et al.* [2019] and further verified in section 5 below. On the one hand, this choice allows for an analytical (closed-form) derivation of improved estimators of E and E' , and

on the other hand, this assumption is 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 Eq. (3.10) of *Stedinger* [1980] as well as an adaptation of the efficient LN2 estimator of the Pearson correlation coefficient ρ introduced by *Stedinger* [1981] and evaluated recently for highly skewed observations by *Barber et al.* [2019]. We do not claim that daily or subdaily streamflow observations follow an LN3 model, rather, the idea here is that the streamflow observations much more closely resemble the behavior of an LN3 model than a normal distribution, thus we choose the best available parameter estimators for an LN3 model. Our estimators of E and E' , which we term the Lamontagne-Barber efficiency estimators LBE and LBE' , respectively, take the form:

$$LBE = 2\hat{\alpha}\hat{\rho} - \hat{\alpha}^2 - \frac{\hat{\Delta}^2}{\hat{C}_o^2} \quad (3.18a)$$

$$LBE' = 1 - \sqrt{(\hat{\beta} - 1)^2 + (\hat{\alpha} - 1)^2 + (\hat{\rho} - 1)^2} \quad (3.18b)$$

where

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

$$\hat{\alpha} = \frac{\hat{\sigma}_s}{\hat{\sigma}_o} = \sqrt{\frac{\exp(2\bar{v} + s_v^2)(\exp(s_v^2) - 1)}{\exp(2\bar{u} + s_u^2)(\exp(s_u^2) - 1)}}$$

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

$$\hat{\rho} = r_1 = \frac{\exp[s_{uv}^2] - 1}{\sqrt{(\exp[s_u^2] - 1)(\exp[s_v^2] - 1)}}$$

where $u_i = \ln[o_i - \hat{\tau}_o]$ and $v_i = \ln[s_i - \hat{\tau}_s]$

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

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

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

A very attractive and efficient estimator of the lower bounds τ_o and τ_s for use in Eq. (3.18) is given by *Stedinger* (1980) as

$$\hat{\tau}_o = \frac{o_{(1)}o_{(n)} - (o_{0.5})^2}{o_{(1)} + o_{(n)} - 2o_{0.5}} \quad \text{and} \quad \hat{\tau}_s = \frac{s_{(1)}s_{(n)} - (s_{0.5})^2}{s_{(1)} + s_{(n)} - 2s_{0.5}}$$

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 . 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 $\hat{\tau}_o$ and $\hat{\tau}_s$ in Eq. (3.18). In situations when that condition cannot be satisfied, we resort to setting $\hat{\tau}_o = 0$ and $\hat{\tau}_s = 0$ which implies an LN2 instead of an LN3 model. Also note that for an LN2 distribution the formula for \hat{C}_o in Eq. (3.18) reduces to the simpler expression $\hat{C}_o = \sqrt{\exp(s_u^2) - 1}$.

4.6 *BLN3 Mixture Estimators of Efficiency (LBE_m and LBE'_m)*

A natural improvement and extension to the estimators *LBE* and *LBE'* introduced in the previous section would be to introduce analogous estimators which exploit properties of the BLN3

mixture mode introduced earlier. For the BLN3 mixture model summarized in section 3.2, for the observations O , (and analogously for the simulations S), we can use the fact that

$E[O^k] = \sum_{i=1}^{12} w_i E[O_i^k]$ follows directly from Eq. (3.13), where each month is assumed to have an

equal number of days so that $w_i = 1/12$. The choice of a monthly mixture model leads to the following two expressions for the mean and variance of the observations for the BLN3 monthly mixture model:

$$\mu_o = \sum_{i=1}^{12} \frac{\mu_i}{12} \quad (3.19a)$$

$$\sigma_o^2 = \sum_{i=1}^{12} \frac{[\sigma_i^2 + \mu_i^2]}{12} - \mu_o^2 \quad (3.19b)$$

Here we employ Eq. (3.19) and analogous expressions for the mean and variance of the simulations to develop improved BLN3 mixture estimators of both E and E' which we term LBE_m and LBE'_m , respectively, given by

$$LBE_m = 2\hat{\alpha}_m r_m - \hat{\alpha}_m^2 - \frac{\hat{\Delta}_m^2}{\hat{C}_{m,o}^2} \quad (3.20a)$$

$$LBE'_m = 1 - \sqrt{\hat{\Delta}_m^2 + (\hat{\alpha}_m - 1)^2 + (r_m - 1)^2} \quad (3.20b)$$

with the mixture estimators (denoted using subscript m) obtained from

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

with $\hat{\mu}_{m,o}$, $\hat{\mu}_{m,s}$, $\hat{\sigma}_{m,o}$ and $\hat{\sigma}_{m,s}$ computed from transformed observations $u_i = \ln(o_i - \hat{\tau}_{o,i})$ and

$v_i = \ln(s_i - \hat{\tau}_{s,i})$ in each month using

$$\hat{\mu}_{m,s} = \sum_{i=1}^{12} \hat{\mu}_{s,i} / 12$$

$$\hat{\mu}_{s,i} = \hat{\tau}_{s,i} + \exp\left(\bar{v}_i + \frac{s_{v,i}^2}{2}\right)$$

$$\hat{\mu}_{m,o} = \sum_{i=1}^{12} \hat{\mu}_{o,i} / 12$$

$$\hat{\mu}_{o,i} = \hat{\tau}_{o,i} + \exp\left(\bar{u}_i + \frac{s_{u,i}^2}{2}\right)$$

$$\hat{\sigma}_{m,o}^2 = \sum_{i=1}^{12} \left[\frac{\hat{\sigma}_{o,i}^2 + \hat{\mu}_{o,i}^2}{12} \right] - \hat{\mu}_{m,o}^2$$

$$\hat{\sigma}_{o,i}^2 = \exp(2\bar{u}_i + s_{u,i}^2) \left(\exp(s_{u,i}^2) - 1 \right)$$

$$\hat{\sigma}_{m,s}^2 = \sum_{i=1}^{12} \left[\frac{\hat{\sigma}_{s,i}^2 + \hat{\mu}_{s,i}^2}{12} \right] - \hat{\mu}_{m,s}^2$$

$$\hat{\sigma}_{s,i}^2 = \exp(2\bar{v}_i + s_{v,i}^2) \left(\exp(s_{v,i}^2) - 1 \right)$$

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

where $r_{1,i}$ is the modified Stedinger [1981] estimator of ρ given in Eq. (3.18) obtained from the transformed observations and simulations in each month i . Here $\hat{\tau}_{o,i}$, \bar{u}_i and $s_{u,i}$ denote *Stedingers* [1980] lower bound, sample mean and sample standard deviation of the values of the transformed observations $u_i = \ln(o_i - \hat{\tau}_{o,i})$ in each month. Similarly, $\hat{\tau}_{s,i}$, \bar{v}_i and $s_{v,i}$ denote *Stedingers* [1980] lower bound, the sample mean and sample standard deviation of the value of the transformed simulations $v_i = \ln(s_i - \hat{\tau}_{s,i})$ in each month.

5 Experimental Results

In this section we compare the eight estimators of E and E' summarized in Section 4 and applied to the USGS PRMS model output and we also perform Monte-Carlo experiments to compare the performance (sampling properties) of those eight estimators of E and E' when the true value of those statistics is known. If one estimator, say *LBE*, is a more attractive estimator than, say *NSE*, then the estimator *LBE* will yield a “better” estimate of the statistic E than *NSE*. The notion of a “best” estimator of a statistic relies upon the choice of a loss function

corresponding to the problem of interest, where a loss function quantifies the relative economic and other losses associated with estimation errors.

Among statisticians, the most common choice of the form of a loss function is quadratic, resulting in the mean squared error criterion of optimality (see *Everitt, 2002*, page 128). It is well known that the *MSE*, variance and bias of an estimator, say *NSE*, are related via

$$MSE[NSE] = E[(NSE - E)^2] = E[(E[NSE] - E)^2] + E[(NSE - E[NSE])^2] = Bias[NSE]^2 + Var[NSE]$$

so that *MSE* of an estimator, say *NSE*, is made up of both its bias and variance (also see Eq. (3.8)).

In the following section, we report all three metrics, because they are all related and important for different reasons as we describe below.

5.1 Monte-Carlo Experiments

In this section we evaluate the sampling properties (*Bias*, *Standard Deviation* and *RMSE*) of the four estimators of E (LBE , LBE_m , NSE , $LNSE$) and the four estimators of E' (LBE' , LBE'_m , KGE' , $PVSE'$), summarized in Section 4, when applied to synthetic daily streamflow series generated from the BLN3 monthly mixture model summarized in Section 3.2. The Monte-Carlo experiments are controlled experiments which ensure that we know the true value of both E and E' in advance. We assume that the true (population) values of E and E' are equal to the values of LBE_m and LBE'_m respectively, computed from the complete period of record of the 447 sites summarized in summarized in Table 3.1. This is the first time, to our knowledge, that such controlled experiments have ever been performed for any of the efficiency statistics discussed here including *NSE* and *KGE*.

We generate 1,000 sets of streamflow simulations and observation series each of length $n=1,095$ ($N=3$ years), $n=3,650$ days ($N=10$ years) and $n=10,950$ days ($N=30$ years) to capture a

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 3.1 by first generating *BLN3* sequences in each month of length $n/12$, using the algorithm described in the Appendix of *Barber et al.* [2019]. True values of the required statistics for generating *BLN3* streamflows in each month, using the approach given in *Barber et al.* [2019, Appendix], are assumed equal to the sample statistics 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 Monte Carlo Experiment Results

Figure 3.3 and Figure 3.4 summarize the bias, variance and *RMSE* associated with the estimators of E and E' respectively, obtained from the Monte-Carlo experiments at the 447 sites summarized in Table 3.1 for sample sizes equal to 3, 10 and 30 years. We emphasize that the bias, variability and *RMSE* in estimators of E and E' illustrated using boxplots in Figure 3.3 and Figure 3.4 should be interpreted as occurring across the 447 sites.

Perhaps the most important finding in Figure 3.3 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, it exhibits enormous variability from one sample to the next, at most sites, and as a result, we cannot recommend its use with daily or subdaily streamflows. Instead, to obtain nearly unbiased estimates of E , we would recommend the use of LBE_m which is approximately unbiased at most sites and exhibits much lower *RMSE* than *NSE* or *LBE*, particularly for the larger sample sizes. Small sample sizes cause increased sampling variability and bias associated with LBE_m because the monthly model requires 36

parameters and evidently a more parsimonious seasonal estimator might be preferred for small samples.

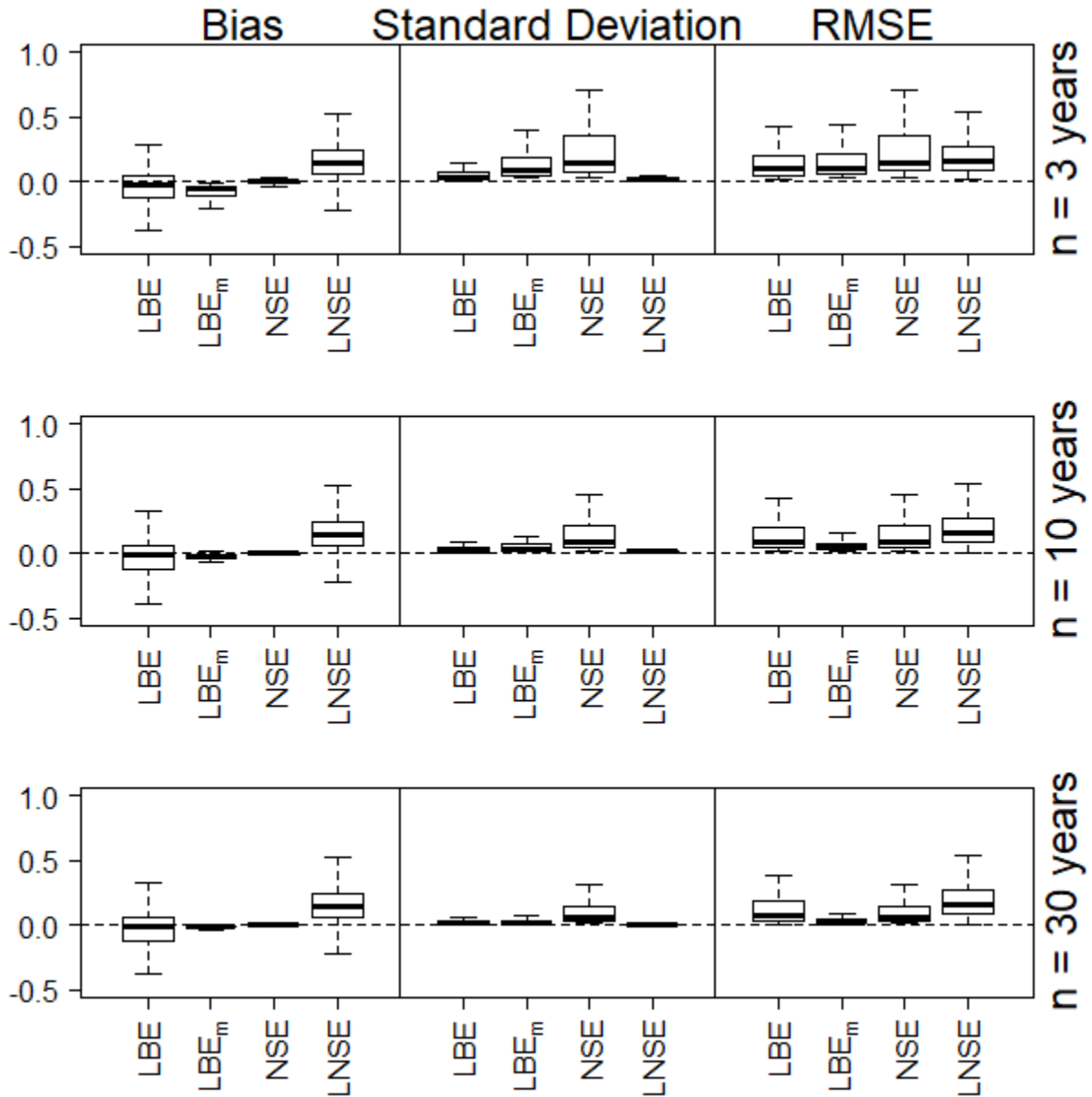


Figure 3.3 Boxplots of Estimates of Efficiency E Resulting From 1,000 Monte Carlo Experiments Performed at Each of the 447 Sites Summarized in Table 3.1

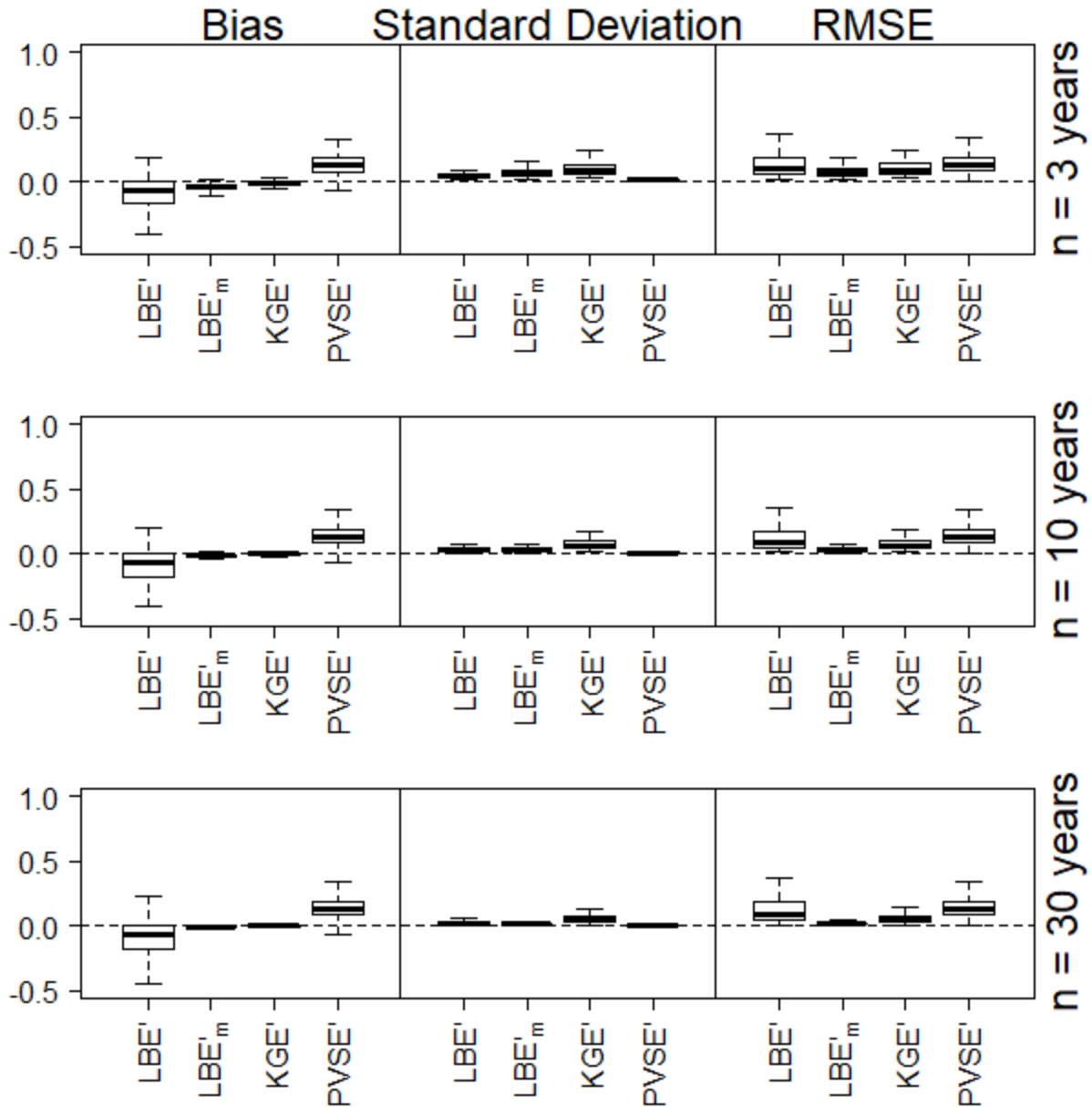


Figure 3.4 Boxplots of Estimates of Efficiency E' Resulting From 1,000 Monte Carlo Experiments Performed at Each of the 447 Sites Summarized in Table 3.1

There are several important conclusions which may be drawn from Figure 3.4. First, the increasingly widely used estimator KGE' is approximately unbiased from one sample to the next, across all sites and exhibits much lower variability and $RMSE$ than was illustrated for NSE in Figure 3.3. Therefore, we are unable to reject KGE' for use with daily and subdaily streamflows.

We also note that the estimator LBE'_m generally exhibits much lower variability 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 Figure 3.3 is the remarkably low standard deviation associated with $LNSE$ compared with the three other estimators of E . One expects the very high upward bias in $LNSE$ at most sites illustrated in Figure 3.3, 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 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, 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 and Fennessey, 1993).

Unbiasedness of estimators of both E and E' is paramount if one's interest is in comparing the performance of models across watersheds, or if one's interest is in 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.

5.3 Comparison of Sample Estimates of E and E'

In this section we compare the various estimators of E and E' summarized in Section 4, when applied to the 447 PRMS watershed simulations summarized in Table 3.1, as well when applied to a single synthetic series of the same length, generated at each watershed from the BLN3 monthly mixture model summarized in Sections 3.2.

We begin our comparisons by choosing a sample of 8 watersheds which represent a wide range of the type of results one can expect from our overall analyses. We include a few cases for which there are not nice elliptical relationships between U and V because those cases illustrate what can happen in some unusual situations. Table 3.2 summarizes values of sample size n , correlation coefficient ρ , coefficient of variation of the observations C_o , fractional bias Δ , and α , as well as the four estimators of E (LBE , LBE_m , NSE , $LNSE$) and the four estimators of E' (LBE' , LBE'_m , KGE' , $PVSE'$), and additionally the values of $PPCC_m$ and $PPCC_{LN3}$ for the 8 chosen sites. Table 3.2 illustrates the gross differences in estimates of E and E' which can result. Figure 3.5 illustrates scatterplots of the bivariate relationship between $v = \ln[s - \hat{\tau}_s]$ and $u = \ln[o - \hat{\tau}_o]$ for the 8 watersheds summarized in Table 3.2. In the lower right hand corner of Figure 3.5 we include for comparison, a scatterplot of synthetic series generated from the BLN3 monthly mixture model. Also shown in Figure 3.5 are two-dimensional confidence intervals, known as “probability ellipses”, which are 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] for further details on their construction. Note that we can only expect these probability ellipses to give a very rough approximation to the relationship between U and V because a BLN3 monthly mixture model will NOT generally yield LN3 marginal distributions for S and O .

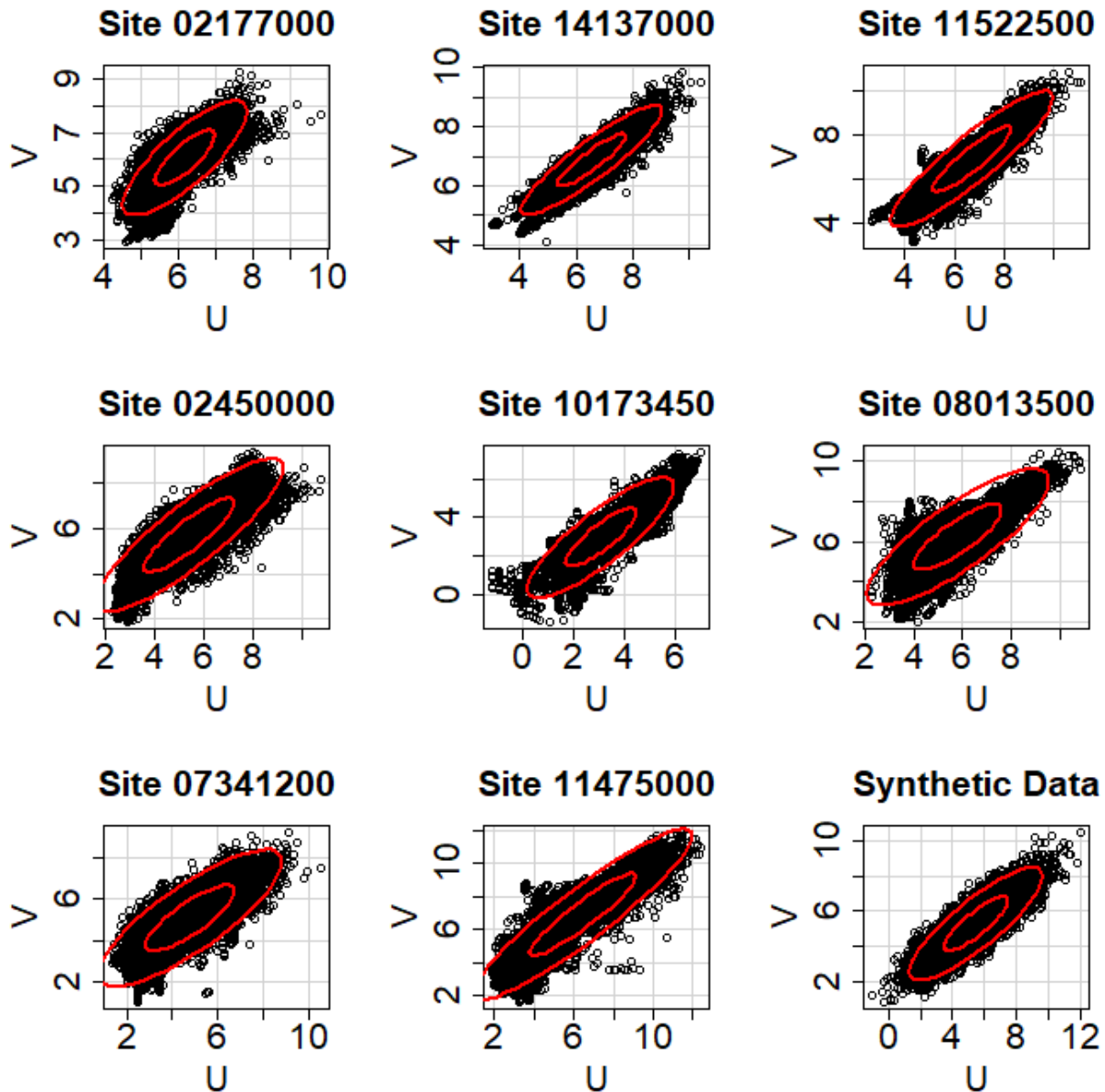


Figure 3.5 Scatter plot of values of $u_i = \ln[o_i - \hat{\tau}_o]$ versus $v_i = \ln[s_i - \hat{\tau}_s]$ for 8 sites summarized in Table 3.2 along with results for synthetic data in the lower right plot

Table 3.2 Estimators of Efficiency and Various Components of Efficiency Based on BLN3 Monthly Mixture Model for 8 selected sites

	Site							
	02177000	14137000	11522500	02450000	10173450	08013500	07341200	11475000
n	10957	11322	10957	10957	10956	10591	10957	10957
ρ	0.73	0.83	0.82	0.69	0.78	0.69	0.61	0.62
Co	0.74	1.01	1.50	2.01	2.49	2.85	3.29	3.95
Δ	-0.03	0.02	-0.10	-0.08	0.01	-0.18	0.14	-0.24
α	1.44	0.80	1.09	0.91	1.05	1.01	0.60	1.48
PPCC_{LN3}	0.9997	0.9929	0.9915	0.9960	0.9898	0.9923	0.9859	0.9928
PPCC_m	0.9998	0.9995	0.9996	0.9997	0.9982	0.9964	0.9893	0.9996
LBE_m	0.02	0.68	0.59	0.42	0.54	0.37	0.37	-0.35
LBE	0.07	0.70	0.65	0.52	0.33	0.34	0.37	0.48
NSE	0.25	0.68	0.61	0.24	0.71	0.75	0.33	0.48
LNSE	0.40	0.80	0.81	0.73	0.73	0.61	0.65	0.83
LBE'_m	0.48	0.74	0.77	0.66	0.77	0.64	0.42	0.35
LBE'	0.48	0.65	0.78	0.65	0.66	0.59	0.38	0.62
KGE'	0.62	0.70	0.77	0.35	0.85	0.75	0.44	0.67
PVSE'	0.83	0.88	0.85	0.82	0.83	0.79	0.79	0.87

Figure 3.6 and Figure 3.7 illustrate scatterplots among the various estimators of E and E' respectively, corresponding to the output from the USGS PRMS model (left column) and corresponding to synthetic sequences generated from the BLN3 monthly mixture model for the 447 watersheds. The dark black circles in Figure 3.6 and Figure 3.7 indicate results for the 8 watersheds shown in Table 3.2. We conclude from Figure 3.6 and Figure 3.7 that the general relationship between the various estimates of both E and E' are similar when using the USGS PRMS model and when using a single synthetic sample generated from the BLN3 monthly mixture model. The enormous variability in the estimates of both E and E' in Figure 3.6 and Figure 3.7 highlight the importance of our earlier 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.

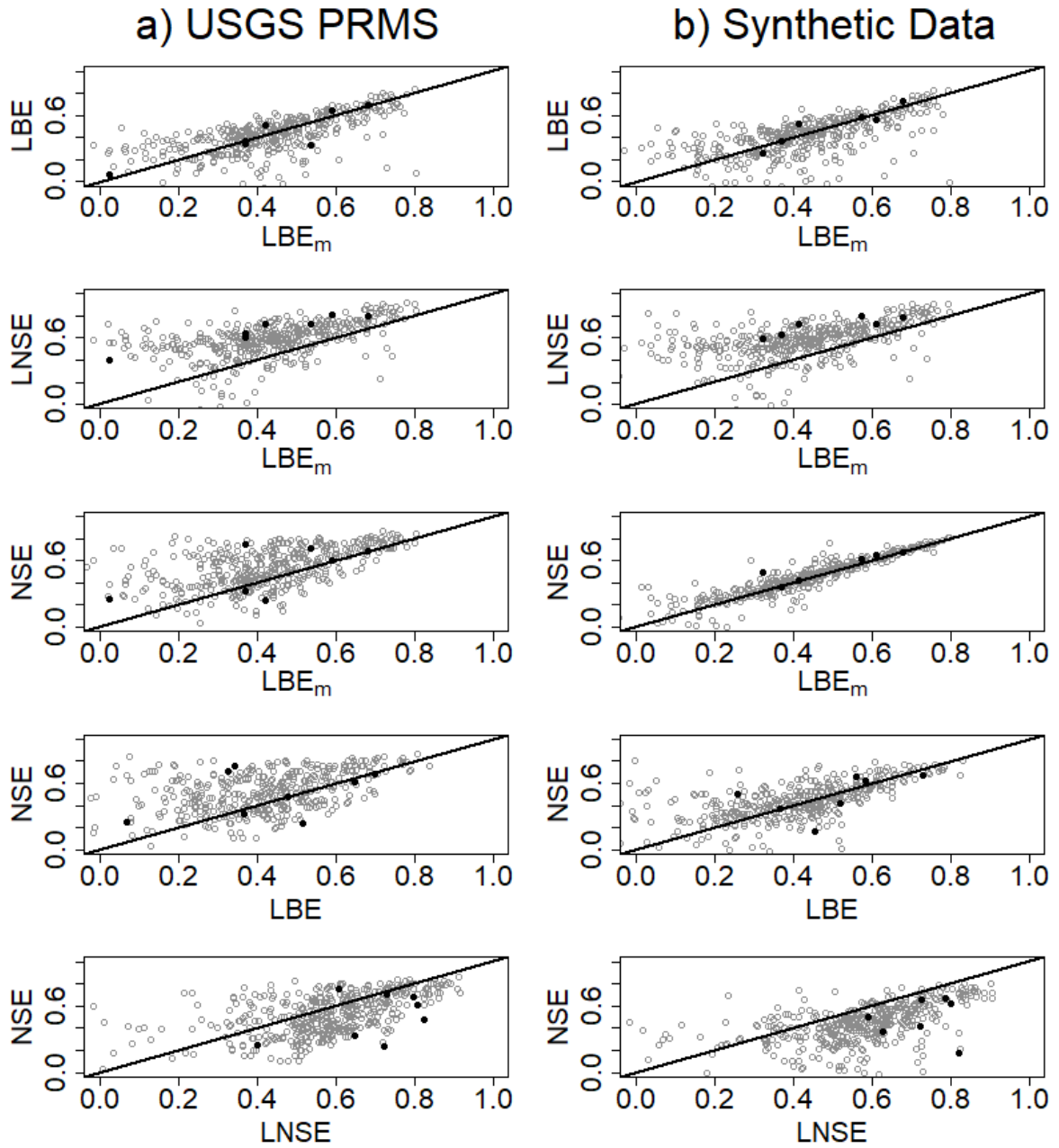


Figure 3.6 Scatterplots of estimates of E obtained from various estimators, with results from 8 sites in Table 3.2 shown using dark black circles.

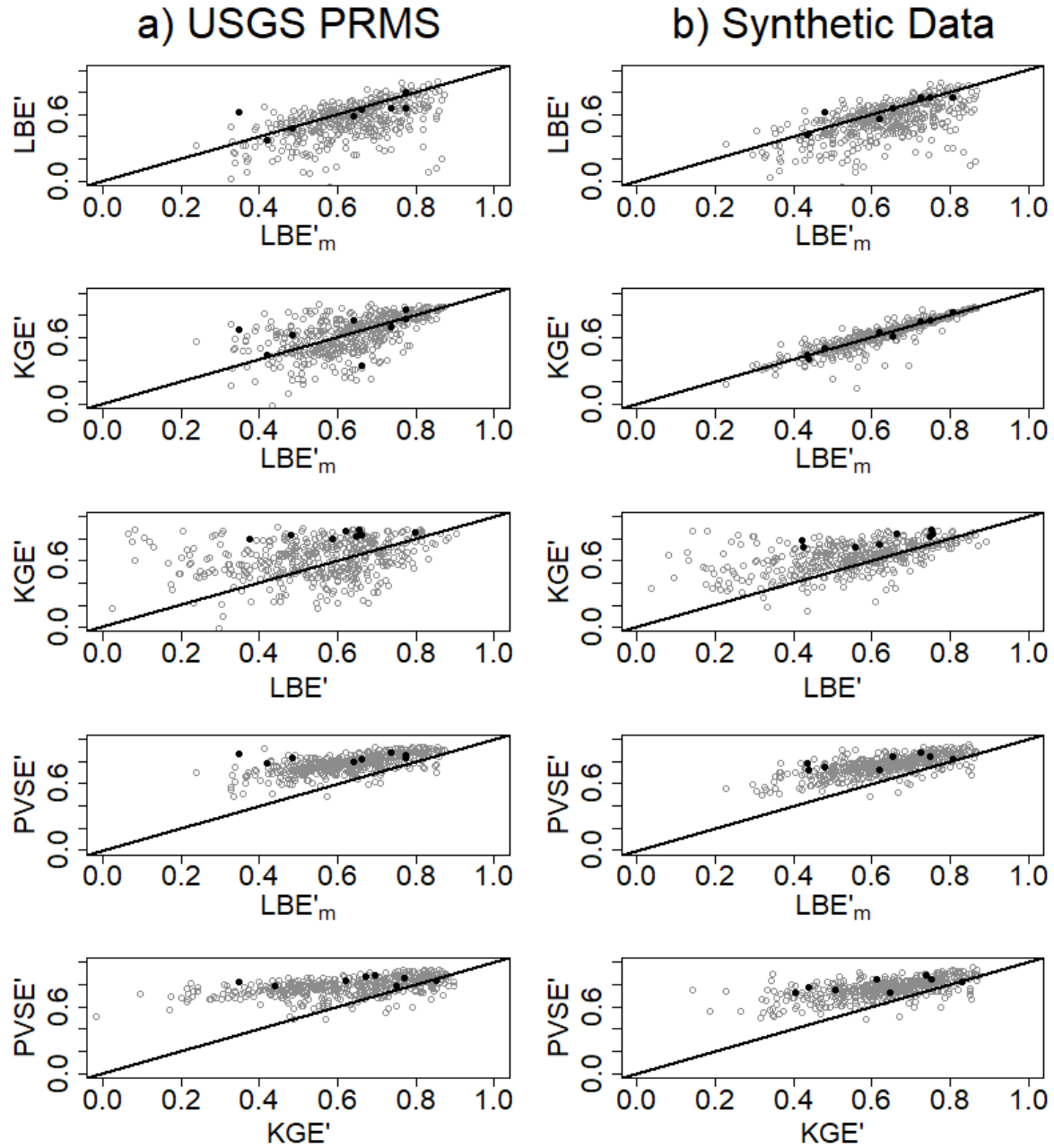


Figure 3.7 Scatterplots of estimates of E' obtained from various estimators, with results from 8 sites in Table 3.2 shown using dark black circles.

5.4 The Causes of Variability in Efficiency Estimates

Figure 8 illustrates the relationship between both $RMSE(NSE)$ and $RMSE(KGE')$ versus the BLN3 monthly mixture model estimates of C_o , ρ , Δ , and α computed at each of the 447

sites. Again, 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. Here we note that increases in both $RMSE(NSE)$ and particularly $RMSE(KGE')$ result from increases in streamflow variability reflected by C_o . Note that the range of values of C_o reported in Figure 8 and Table 1 does not reflect the enormous variability possible as evidenced from the fact that *Vogel et al.* [2003] used lognormal estimators to report a range in the values in C_o for daily flow series across the U.S. from approximately 0.5 to 10,000 with a median value of 10, and an interquartile range from 3 to 33.

For positively skewed data such as daily streamflows, increases in C_o are expected to lead to considerable increases in skewness. For example, for Gamma and LN2 variables, skewness γ is related to coefficient of variation C via the relations $\gamma = 2C$ and $\gamma = C^3 + 3C$, respectively. tely 0.5 to 10,000 with a median value of 10, and an interquartile range from 3 to 33.

We also note that 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]. Such a comparison reveals very clearly that ignoring the periodicity of daily streamflow as was done by *Barber et al.* [2019] cannot reproduce expected variability in estimates of ρ derived from actual streamflow series.

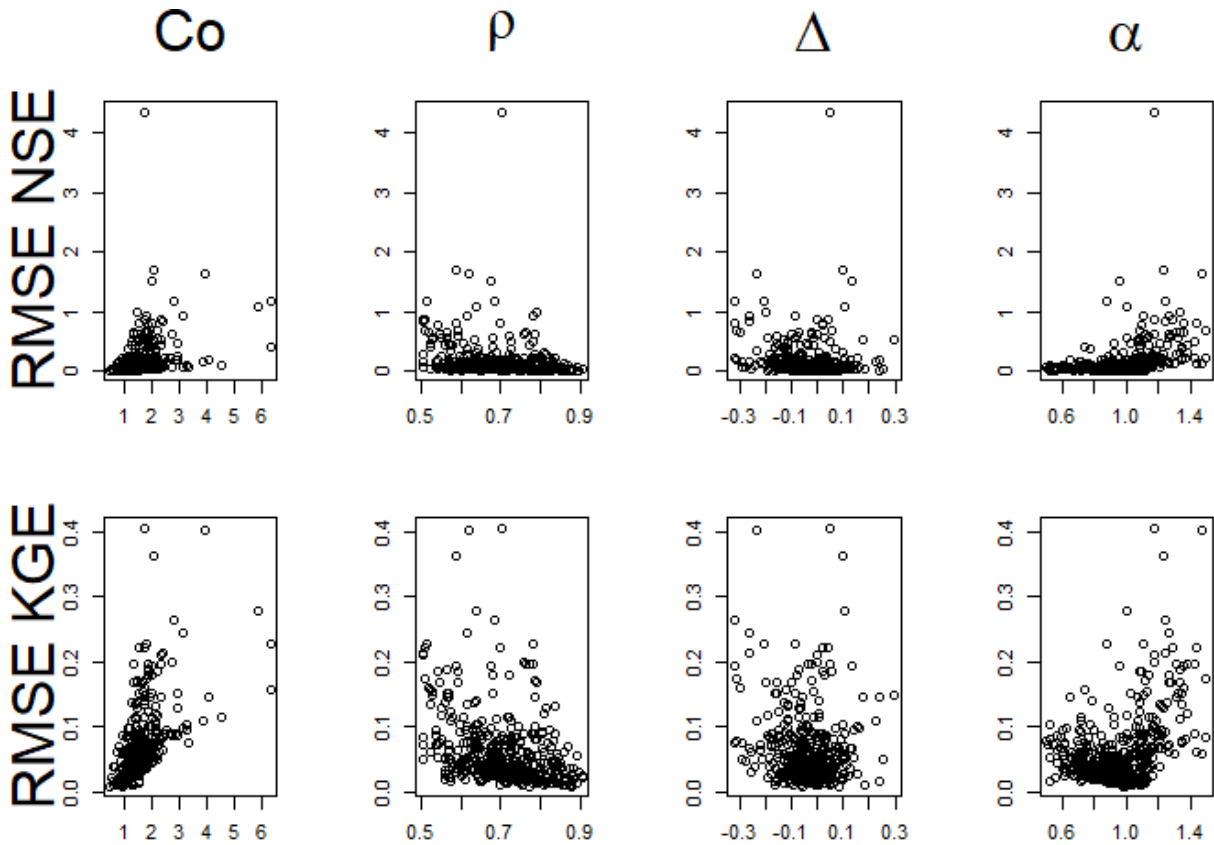


Figure 3.8 Root mean squared error of NSE and KGE vs C_o , ρ , Δ , and α . RMSE of NSE and KGE calculated based on 1,000 Monte Carlo replicates of sample length $n=10,950$ (30 years) for each of the 447 sites.

6 Discussion

We have introduced a BLN3 monthly mixture model which appears to provide an excellent representation of the distribution of daily streamflows across the conterminous U.S., however, several caveats exist and several improvements and extensions are possible as discussed below. We are not claiming that daily streamflows follow an LN3 distribution in a given month, rather, we are simply arguing that a BLN3 monthly mixture model provides a much better first order approximation to daily streamflow observations and simulations than a bivariate normal model which is the assumption behind the use of the existing product moment statistics embedded within the traditional statistics *NSE* and *KGE*.

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 U.S., is approximately 0.01 cfs [Grenato *et al.*, 2017]. Of the 20,438 U.S. Geological Survey (USGS) river gages evaluated by Grenato *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 semi-arid Southwest according to the USGS National Hydrography Dataset. Such streams usually reside in the headwaters or major tributaries of perennial streams in the Southwest. A natural extension to this study would be to develop estimators of E based on a zero-inflated LN3 monthly mixture model analogous to the zero-inflated BLN3 mixture model introduced by Shimizu [1993] for modeling rainfall.

6.2 *Improved Mixture Model*

Another natural extension to this study would be to develop estimators of E based on a bivariate Kappa monthly mixture model, because Blum *et al.* [2017] and others have shown that a four-parameter Kappa model provides a better fit to the distribution of daily streamflows than an LN3 model. Such a bivariate Kappa mixture model could be combined with a Gaussian or other suitable copula to provide an improved representation of both the dependence structure and marginal distributions of the daily streamflow observations.

We have introduced a BLN3 monthly mixture model for use in deriving improved estimators of efficiency. Such a model is not parsimonious because it requires estimation of 36 parameters which could lead to increased sampling bias and variance associated with our recommended estimators for short samples. Alternatively, a seasonal model could be advanced which enables fewer parameters yet still captures the important deterministic periodic seasonal component of streamflows. Future studies are needed to better understand the degree of parsimony needed in the mixture model so as to provide efficient estimates of E while still capturing the critical complexities of the hydrologic process of interest including periodicity, occurrence of zeros and skewness.

7 Conclusions

Our approach differs from all previous research relating to the estimation of efficiencies for evaluation of goodness-of-fit, because our conclusions are based on both theoretical (probabilistic) and empirical (statistical) analyses of the concept of efficiency. 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 indeed surprising that it took this long to advance a theoretical or probabilistic definition of efficiency for use in goodness-of-fit evaluations. Perhaps our most fundamental contribution was to clarify, for the first time, 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] and *Gupta and Kling* [2011]. Below we summarize our 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. In general, the two theoretical statistics E and E' are shown in Figure 3.1 to exhibit completely different behavior. Importantly, the theoretical statistic E has a well known interpretation because it is a standardized form of MSE , whereas the theoretical interpretation of E' is unclear. Measures of MSE and $RMSE$ are perhaps the most widely used metrics of goodness-of-fit across all disciplines. Still, we have shown that the statistic E' has numerous attractive features, however, if E' is to be considered further, attention should be given to its physical interpretation analogous to the definition of E as a measure of standardized MSE . The concepts of MSE and $RMSE$ are fundamental and widely used as metrics across many disciplines. For example, MSE and $RMSE$ are the most common performance metrics in the fields of statistics (Everitt, 2002, page 128) and signal processing (Wang and Bovik, 2009). Before E' is accepted as a standard performance metric in hydrology, its theoretical relationship to $RMSE$, MSE and E should be defined more clearly.

Rejection of NSE But Not KGE' : Even though NSE is consistently unbiased, it exhibits extraordinary variability from one sample to the next, 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 increasingly widely used estimator KGE' was found to be an approximately unbiased estimator or E' , across all sites and exhibits lower variability and $RMSE$ than was illustrated for NSE , thus we are unable to reject KGE' for use with daily and subdaily streamflows.

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 KGE' , particularly for the larger sample sizes. The primary reasons that these estimators are favored is because they address the critical issues of skewness and periodicity which both confound the performance of both NSE and KGE' . Increased concerns over skewness are anticipated in arid and semi-arid regions and at intermittent and ephemeral sites not considered in this study.

Remarkably Low Variability Associated with LNSE: Due to its remarkable and consistently low variance, the estimator $LNSE$ is recommended for calibration of models at an individual site, however, due to its considerable and unpredictable bias, one should not compare values of $LNSE$ across sites.

Synthetic Series are Similar to PRMS Model Output: Comparisons of the behavior of estimates of E and E' corresponding to both the output of USGS PRMS models and to synthetic daily streamflows generated from a BLN3 monthly mixture model indicate very similar behavior, which illustrates that synthetic series must be mimicking, to a great extent, the behavior of the USGS PRMS model output. This is an important innovation and improvement over the results of *Barber et al.* [2019] who used a BLN3 model but did not consider the BLN3 monthly mixture model introduced here.

The Importance of Accounting for Skewness and Periodicity: Daily streamflow series exhibit considerable periodicity and skewness; two properties which were shown to lead to corresponding variability in estimates of efficiency. We have shown that introduction of a bivariate lognormal monthly mixture model addresses both skewness and periodicity and consequently, leads to

considerable improvements in the performance of associated estimators of efficiency when working with daily streamflows. Essentially, daily streamflows are neither normally nor identically distributed, both critical assumptions needed for both NSE and KGE' to perform as expected. We anticipate that studies in arid and semi-arid regions, as well as studies in ephemeral and intermittent streams and studies in which there is marked seasonal behavior of streamflow will lead to even greater variability associated with estimates of NSE and KGE' than was reported here.

Implications of Findings: We have shown that application of NSE to bivariate daily streamflow series can lead to highly spurious and variable results from one sample to another at a single site so that its use in goodness-of-fit evaluations, model calibration and/or regionalization studies would lead to highly spurious results which 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 a suite of estimators based on a BLN3 monthly mixture model which led to much more consistent and reproducible estimates of efficiency at a single site or across sites. Thus we anticipate that application of our improved estimators could lead to associated improvements in simulation model calibration, validation, and in hydrologic regionalization efforts which seek to develop multivariate relationships among model parameters and watershed characteristics.

Acknowledgements

The authors are indebted to Francesco Serinaldi for his insightful comments on a very early draft of this paper.

References

- ASCE (1993), Criteria for evaluation of watershed models, *Journal of Irrigation and Drainage Engineering*, 119(3) 429-422.
- Baldwin, C.K., and U. Lall (1999), Seasonality of streamflow: The upper Mississippi River, *Water Resources Research*, 35(4), 1143-1154.
- Barber, C., J. Lamontagne and R.M. Vogel (2019), Improved estimators of correlation and R^2 for skewed hydrologic data, *Hydrological Sciences Journal*, DOI: 10.1080/02626667.2019.1686639
- Bardsley, W.E. (2013), A goodness of fit measure related to r^2 for model performance assessment, *Hydrological Processes*, 27, 2851-2856.
- Blum, A.G., S.A. Archfield, and R.M. Vogel (2017), The probability distribution of daily streamflow in the United States, *Hydrology and Earth Systems Science*, 21, 3093–3103, <https://doi.org/10.5194/hess-21-3093-2017>.
- Criss R.E. and W.E. Winston (2008), Do Nash values have value? Discussion and alternate proposals, *Hydrological Processes*, 22:2723–2725.
- 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, 178–187.
- Farmer, W. H., and R. M. Vogel (2016a), On the deterministic and stochastic use of hydrologic models, *Water Resources Research*, 52, doi:10.1002/2016WR019129.
- Farmer, W.H., and R.M. Vogel (2016b), On the Deterministic and Stochastic Use of Hydrologic Models: Data Release: U.S. Geological Survey data release, <https://dx.doi.org/10.5066/F7W37TF4>.
- Granato G.E., Ries, K.G., III, and 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., B. Cappelaere, C. Delenne, and D. Ruelland (2011), Towards improved criteria for hydrological model calibration: theoretical analysis of distance- and weak form-based functions, *Journal of Hydrology*, 401, 1–13.
- Gupta H.V., and H. Kling (2011), On typical range, sensitivity and normalization of Mean Squared Error and Nash-Sutcliffe Efficiency type metrics, *Water Resources Research*, 47, doi:10.1029/2011WR010962,.

- Gupta, H.V., H. Kling, K.K. Yilmaz and G.F. Martinez (2009), Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modeling, *Journal of Hydrology*, 377, 80-91.
- Helsel, D., et al. (2019), *Statistical methods for water resources*. 2nd ed. Reston, VA: US Geological Survey.
- Jain S.K., and K.P. Sudheer (2008), Fitting of Hydrologic Models: A close look at the Nash-Sutcliffe Index, *Journal of Hydrologic Engineering*, 13(10) 981-986.
- Knoben, W.J.M., J.E. Freer and R.A. Woods (2019), Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores, *Hydrology and Earth System Sciences*, <https://doi.org/10.5194/hess-2019-327>.
- Krause P., D.P. Boyle, and F. Base (2005), Comparison of different efficiency criteria for hydrological model assessment, *Advances in Geosciences*, 29(5):89–97.
- Legates, D.R. and G.J. McCabe (1999), Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, *Water Resources Research*, 35(1) 233-241.
- Levick, L., J. Fonseca, D. Goodrich, M. Hernandez, D. Semmens, J. Stromberg, R. Leidy, M. Scianni, D. P. Guertin, M. Tluczek, and W. Kepner (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., A. Sankarasubramanian, A. Sharma and B.J. Reich (2018), A non-parametric bootstrapping framework embedded in a toolkit for assessing water quality model performance, *Environmental Modelling and Software*, 107, 25-33.
- Limbrunner, J.F., R.M. Vogel, and L.C. Brown (2000), Estimation of the Harmonic Mean of a Lognormal Variable, *Journal of Hydrologic Engineering*, 5(1), 59-66.
- Liu, D., Guo, S., Wang, Z. et al. (2018), Statistics for sample splitting for the calibration and validation of hydrological models, *Stochastic Environmental Research and Risk Assessment*, 32: 3099. <https://doi.org/10.1007/s00477-018-1539-8>.
- Markstrom, S.L., R.S. Regan, L.E. Hay, R.J. Viger, R.M.T. Webb, R.A. Payn, and J.H. LaFontaine (2015), PRMS-IV, the precipitation-runoff modeling system, version 4: U.S. Geological Survey Techniques and Methods, book 6, chap. B7, 158 p, <http://dx.doi.org/10.3133/tm6B>.
- Martinez, J. and A. Rango (1989), Merits of statistical criteria for the performance of hydrological models, *Water Resources Bulletin*, 25(2). 421-432.

- Mathevet, T., C. Michel, V. Andréassian, and C. Perrin, (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: Results of the Model Parameter Experiment–MOPEX. vol. 307. IAHS Publication, pp. 211–219.
- McCuen, R.H., Z. Knight, and A. G. Cutter (2006), Evaluation of the Nash-Sutcliffe Efficiency Index, *Journal of Hydrologic Engineering*, 11(6), 597.
- Moriasi, D.N., J.G. Arnold, M.W. Van Liew, R.L. Bingner, R.D. Harmel and T.L. Veith (2007), Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations, *Transactions of the ASABE*, 50(3) 885-900.
- Murphy, A.H. (1988), Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient, *Monthly Weather Review*, 116, 2417-2424.
- Nash, J. E., and J.V. Sutcliffe (1970), River flow forecasting through conceptual models. Part 1: A discussion of principles, *Journal of Hydrology*, 10(3), 282–290.
- Pearson, K. (1896), Mathematical contributions to the theory of evolution III. Regression, heredity and panmixia, *Philosophical Transactions A*, 373, 253-318.
- Pool, S., M. Vis and J. Seibert (2018), Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency, *Hydrological Sciences Journal*, 63:13-14, 1941-1953, DOI: 10.1080/02626667.2018.1552002.
- Pushpalatha, R., C. Perrin, N. Le Moine, V. Andreassian (2012), A review of efficiency criteria suitable for evaluating low-flow simulations, *Journal of Hydrology*, 420, 171-182.
- Ritter A, and R. Munoz-Carpena (2013), Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments, *Journal of Hydrology*, 480(3):33–45
- Schaefli B, and H.V. Gupta (2007), Do Nash values have value? *Hydrological Processes*, 21:2075–2080
- Shimizu, K. (1993), A Bivariate Mixed Lognormal Distribution With an Analysis of Rainfall Data, *Journal of Applied Meteorology*, 32, 161-171.
- Stedinger, J.R.,(1980), Fitting lognormal distributions to hydrologic data, *Water Resources Research*, 16(3) 481-490.
- Stedinger, J.R. (1981), Estimating correlations in multivariate streamflow models, *Water Resources Research*, 17(1) 200-208.

- Todini, E. and D. Biondi (2017), Calibration, parameter estimation, uncertainty, data assimilation, sensitivity analysis, and validation, Chapter 22 in “*Handbook of Applied Hydrology*”, V.P. Singh, editor-in-chief, McGraw Hill, New York.
- Vogel, R.M. and N.M. Fennessey (1993), L-Moment Diagrams Should Replace Product-Moment Diagrams, *Water Resources Research*, 29(6), pp 1745-1752.
- Vogel, R.M., J.R. Stedinger and R.P. Hooper (2003), Discharge Indices for Water Quality Loads, *Water Resources Research*, 39(10), 1273, doi:10.1029/2002WR001872.
- Vogel, R.M. (2017), Stochastic watershed models for hydrologic risk management, *Water Security*, <http://dx.doi.org/10.1016/j.wasec.2017.06.001>.
- Wang Z., and A. C. Bovik (2009), Mean squared error: love it or leave it? A new look at signal fidelity measures, *IEEE Signal Processing Magazine*, 26(1) 98–117.
- WMO (1986), Intercomparison of Models of Snowmelt Runoff Operational Hydrology Report No. 23, World Meteorological Organization, Geneva.

Future Work

This dissertation has added to the already numerous studies documenting the limitations of commonly used goodness-of-fit metrics such as Pearson's correlation and Nash-Sutcliffe Efficiency. It has also provided alternative estimators for the theoretical measures of correlation and efficiency which have been shown to be more suitable for skewed, periodic hydrologic data. Although suitable estimators have been developed, improvement upon these estimators is still possible and more detailed guidance should be given as to when the LBE and LBE_m estimators are appropriate to use. To ensure that the improved estimators are utilized by practicing hydrologists, an R package could be developed. There continues to be many opportunities to expand upon this research and it is certainly an exciting and active area that should be pursued further.

1 Grand challenges remaining

The work presented in this thesis has added to the growing literature that demonstrates significant issues with the most commonly used estimator of efficiency in hydrology: NSE . Despite the well documented problems with using NSE as a goodness-of-fit metric for hydrologic data, it continues to be used by researchers and practicing hydrologists alike. Although KGE is increasingly being used, it too has documented limitations and is in fact not estimating the same theoretical efficiency as NSE . Similarly, estimates of correlation have been shown to have severe limitations when the underlying assumptions are violated. Correlation is a goodness-of-fit metric used in many fields, not just hydrology. As a result of the shortcomings of these goodness-of-fit metrics, evaluating model performance is not an entirely objective practice. As more people recognize the weaknesses of these statistics, questions will grow regarding whether statistical models perform as well as the estimators suggest.

Although estimators of correlation and efficiency have been developed here for high frequency hydrologic data and have been shown to provide improvements over other estimators, they are certainly not the best estimators for all data. In order to improve the general field of model assessment, other estimators could be developed which would provide accurate estimates of efficiency based on the specific distribution from which the data appear to arise. These estimators need not be limited to distributions that are commonly observed in hydrologic applications. Any models that are currently evaluated with parametric statistics while violating the underlying assumptions of these parametric statistics can be improved upon. If the distribution of the underlying data can be reasonably approximated, then a parametric estimator developed specifically for the distribution in question could provide the most accurate assessment of model performance. This proposed work could have broad implications on model evaluation techniques in a variety of fields.

2 Specific experiments

2.1 Further improvements for correlation and efficiency estimators

As was shown in the analysis of efficiency estimators, the mixture estimator LBE_m provides significant improvement over commonly used estimators of efficiency, such as NSE and KGE , as well as over LBE , an estimator developed specifically for bivariate lognormal data. This finding suggests that mixture estimators for other common hydrologic statistics may perform better for skewed hydrologic data. Therefore, a mixture estimator of correlation could be developed and analyzed to determine whether similar results are seen. Research into the performance of a mixture estimator of correlation could have a profound impact if it is found that the mixture estimator performs better than Pearson's correlation and Stedinger's correlation estimator. The Pearson correlation coefficient is used as a goodness-of-fit metric in hydrology, but also in numerous other

fields of study. Chapter 2 of this dissertation demonstrated the improved performance of Stedinger's correlation estimator over the commonly used Pearson's estimator, however a mixture estimator of correlation may improve performance even further.

Building upon the mixture estimator of efficiency developed in this dissertation for bivariate lognormal data, a comparison of performance between LBE_m and mixture estimators of NSE and KGE is warranted. This analysis would improve understanding of the potential improvements that could be made to already commonly used estimators, as well as evaluate the benefits and limitations of LBE_m when compared with other mixture estimators. Analogous to the methods used in this dissertation, Monte Carlo experiments could be performed in order to compare the mixture estimators of NSE and KGE with LBE_m . A limitation of the research presented in Chapter 3 is that a mixture estimator, LBE_m , was compared with NSE and KGE , non-mixture estimators. The development and subsequent analysis of the performance of LBE_m compared to mixture estimators of NSE and KGE would provide a more level playing field for comparison and may provide further insights into their performance for skewed, periodic hydrologic data.

2.2 Robustness analysis of correlation and efficiency estimators

The estimators r_l and LBE developed in this research were designed to perform well for skewed hydrologic data. LBE_m was shown to be a further improvement on LBE because it also took advantage of the periodic nature of hydrologic data. To assess the performance of these estimators against their more commonly used goodness-of-fit metrics, Monte Carlo experiments were performed which specifically generated bivariate lognormal data. It has been shown in this research and previous research that a three-parameter lognormal (LN3) distribution provides a good first approximation of daily streamflow data, however the true distribution of daily streamflow can never be known. Therefore, it is essential to understand how the estimators

developed through this research perform when the data come from a distribution other than the one they were developed for (LN3). This would provide insights into how robust the different estimators are when their initial assumptions are violated.

To analyze how well the efficiency estimators LBE and LBE_m perform when the data are not LN3 distributed, a series of Monte Carlo experiments could be performed. Other distributions that approximate the distribution of daily streamflow could be used to generate data. The efficiency estimators could then be calculated from this generated data. This analysis would provide better understanding as to when these estimators can appropriately be applied. If it is found that they do perform well when the data arise from a distribution other than LN3 the recommendation for their widespread use can more confidently be made. If, however, it is found that they do not perform well under conditions other than LN3 distributed data, recommendations for its use must clearly convey that it is suitable only for situations in which the data are LN3 distributed.

Pursuing this analysis further, incorporation of commonly used goodness-of-fit metrics is recommended. This would entail conducting Monte Carlo experiments that analyze the performance of the estimators developed through this research (LBE , LBE_m) with the performance of NSE and KGE under conditions in which the data arise from a distribution other than LN3. This would allow for direct comparison of efficiency estimators under various possible conditions. From this analysis, conclusions regarding the suitability of each estimator under different conditions could be made.

Analogous to the research suggested here for efficiency estimators, similar research into the correlation estimators (namely Stedinger's r_I and Pearson's r) could be conducted. As with the conclusions drawn from the efficiency estimator analysis, this would provide guidance as to how

well these estimators of correlation perform when their underlying assumptions are violated. This would ultimately provide an indication of how robust the estimators are.

2.3 Impacts of sample size and parameters on efficiency estimators

More research is needed in order to understand the impact of sample size on efficiency estimators. Based on the Monte Carlo results presented in Chapter 3 of this dissertation, LBE_m , LBE'_m , and NSE improve significantly as sample size increases. This is largely due to a reduction in the standard deviation of these estimators at a given site as sample size increases. Based on the findings presented, sample sizes of 10 years or greater provide a sufficient amount of data to ensure that LBE_m and LBE'_m yield unbiased estimates of efficiency and have a root mean square error of nearly zero, however for sample sizes of 3 years the performance of these estimators degrades. The performance of NSE is notably unsatisfactory for small sample sizes which raises concerns regarding its use in applications which do not have long record lengths. This may be a result of the sensitivity of NSE to large outliers. More research is needed in order to demonstrate the potential benefit of using LBE_m over NSE in cases which do not have a sufficiently long record length. Specifically, additional experiments for record lengths of 5 and 7 years are needed in order to provide more detailed recommendations as to when NSE is not appropriate to use due to an insufficiently long record length.

In addition to exploring the impact of sample size on the performance of efficiency estimators, future work should also investigate the impact of the parameters ρ , Co , Δ , and α on efficiency estimators. In order to accurately accomplish this, reliable estimators of these parameters for skewed hydrologic data must be employed. Chapter 2 of this dissertation provides evidence that r_I provides accurate estimate of ρ for skewed hydrologic data, however similar formal analyses of estimators of Co , Δ , and α for the data considered here is needed. Chapter 3

includes BLN3 estimators of each of these parameters. Monte Carlo methods could be applied in order to evaluate the accuracy of the BLN3 estimators of these parameters as compared to more commonly used estimators. Following these analyses, thorough research into how efficiency estimators are impacted by the parameters ρ , Co , Δ , and α can be conducted. The results of research focused in this area could provide further guidance as to when different estimators of efficiency can be expected to yield an accurate account of the goodness-of-fit between model simulations and observations.

2.4 Analysis of efficiency estimators for sub-daily hydrologic data

As has been shown through this research, common goodness-of-fit metrics such as Pearson's r and NSE are not adequate measures of model performance for daily hydrologic data. This is in large part a function of the frequency of the data. Daily hydrologic data is highly skewed which presents problems when using Pearson's r and NSE . As data is more readily available at even higher frequency, such as hourly or 15 minute intervals, the degradation of performance may become even more pronounced for Pearson's r and NSE . In order to have confidence in the models developed based on sub-daily data, it is important to understand how accurate different goodness-of-fit metrics are for this data. LBE and LBE' should be investigated in order to understand if they outperform NSE and KGE , respectively, for sub-daily data. Similarly, a comparison of the different correlation estimators presented in Chapter 2 should be performed for sub-daily data. More guidance is needed in order to understand how widely applicable the alternative estimators presented here are, particularly as higher frequency data is used. Such an analysis would provide much needed insight into the confidence modelers, researchers, and decision makers alike can place in their estimates of correlation and efficiency.