



Improved estimators of correlation and R^2 for skewed hydrologic data

Caitline Barber, Jonathan R. Lamontagne & Richard M. Vogel

To cite this article: Caitline Barber, Jonathan R. Lamontagne & Richard M. Vogel (2019): Improved estimators of correlation and R^2 for skewed hydrologic data, Hydrological Sciences Journal, DOI: [10.1080/02626667.2019.1686639](https://doi.org/10.1080/02626667.2019.1686639)

To link to this article: <https://doi.org/10.1080/02626667.2019.1686639>



Accepted author version posted online: 31 Oct 2019.
Published online: 12 Nov 2019.



Submit your article to this journal [↗](#)



Article views: 10




View related articles [↗](#)



View Crossmark data [↗](#)

Improved estimators of correlation and R^2 for skewed hydrologic data

Caitline Barber, Jonathan R. Lamontagne and Richard M. Vogel 

Department of Civil and Environmental Engineering, Tufts University, Medford, Massachusetts, USA

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.

ARTICLE HISTORY

Received 28 May 2019
Accepted 26 September 2019

EDITOR

S. Archfield

ASSOCIATE EDITOR

E. Volpi

KEYWORDS

correlation coefficient; goodness of fit; coefficient of determination; bivariate lognormal; calibration; validation; coefficient of variation; skewness; sampling; bias; copula; Gaussian; Spearman; Pearson; trend

1 Introduction and problem setting

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 :

$$S = H(X|\Omega) + \varepsilon \quad (1)$$

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:

$$O = S + \varepsilon \quad (2)$$

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 the 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 Equations (1) and (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:

$$\rho = R = \frac{\text{cov}(O, S)}{\sqrt{\text{var}(O)\text{var}(S)}} = \frac{E[(O - \mu_O)(S - \mu_S)]}{\sigma_O \sigma_S} \quad (3)$$

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

$$r = \frac{\frac{1}{n} \sum_{i=1}^n (s_i - \bar{s})(o_i - \bar{o})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (o_i - \bar{o})^2 \frac{1}{n} \sum_{i=1}^n (s_i - \bar{s})^2}} \quad (4)$$

Pearson (1896) introduced both the theoretical statistic $R = \rho$ in Equation (3) as well as its sample estimator r , in (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 Equation (3), it is the theory of statistics which governs the sampling properties of the estimator r .

1.2 Application of R^2 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 Equation (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 $\text{NSE} = \rho^2 = R^2$ for any unbiased model which exhibits serially independent residuals ε in Equation (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 Equation (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 Equation (3) and one estimator of that statistic, r , given in Equation (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 Equation (3). Similarly, Legates and McCabe (1999) and many others have criticized the use of the estimator r in (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 Equation (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 Equation (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, 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 20th 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 the 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 the 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 Equation (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 an 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:

$$\text{Var}[r] = \left[(1 - \rho^2)^2 / n \right] \left[(1 + \rho_s \rho_o) / (1 - \rho_s \rho_o) \right] \quad (5)$$

where ρ_s and ρ_o are the lag-one serial correlation coefficients for the s and o series, respectively. The second quantity on the right-hand side of Equation (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 Equations (1) and (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, Fig. 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 Fig. 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

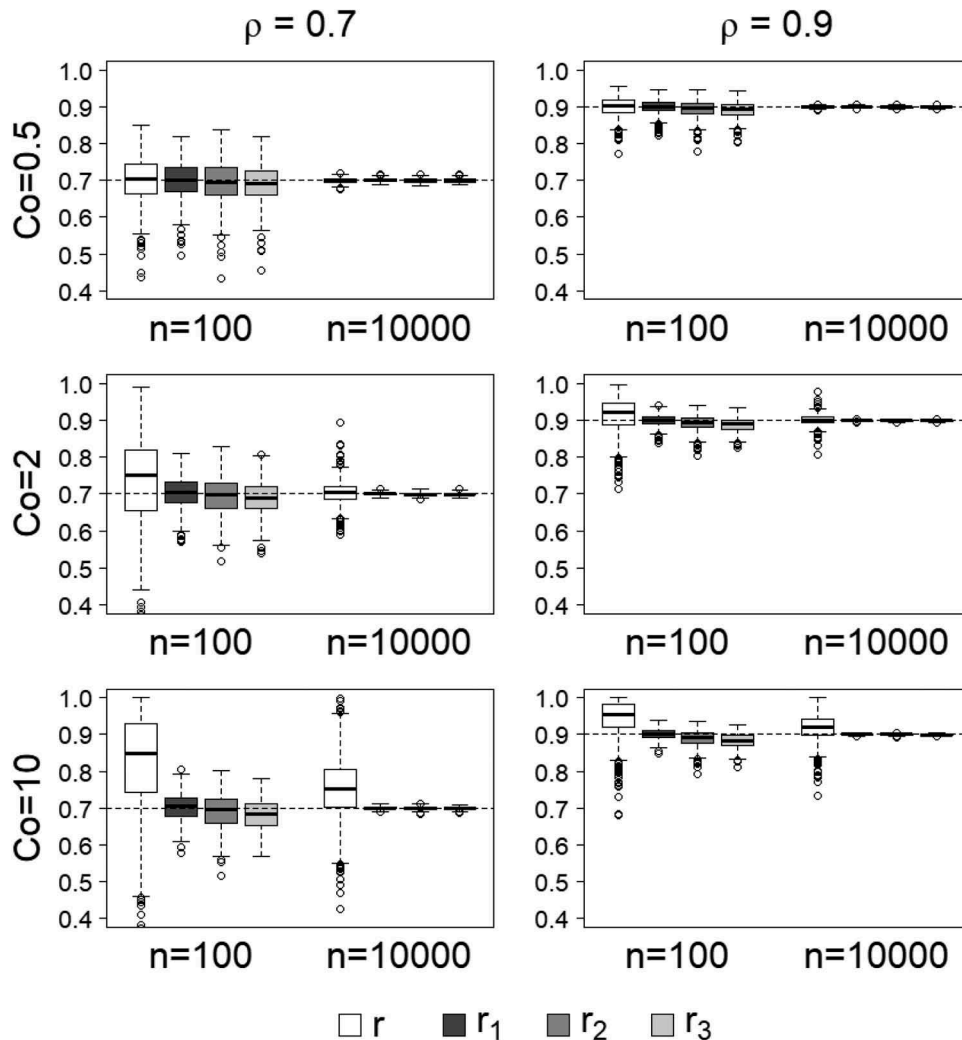


Figure 1. Results of the Monte Carlo experiments are illustrated using boxplots of the four estimators of correlation ρ considered: Pearson r , Stedinger r_1 , modified Spearman r_2 and modified RIN r_3 . Boxplots summarize the sampling distribution of estimators of ρ for serially independent synthetic streamflows generated from a bivariate lognormal model with $\rho = 0.7$ (left) and $\rho = 0.9$ (right), for three different values of coefficient of variation $C_O = C_S$.

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:

$$\rho = \frac{\exp(\rho_{UV}\sigma_U\sigma_V) - 1}{\sqrt{\exp(\sigma_U^2) - 1}\sqrt{\exp(\sigma_V^2) - 1}} \quad (6)$$

(see Mostafa and Mahmoud 1964; Equation 5 in Stedinger 1981 and eq. 11.71 in, Balakrishnan and Lai 2009). Thus, Equation (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 Equation (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 Equation (6) yields $\rho_{UV} = 0.952$. Generally, the coefficient of variation of the observations and simulations will not be equal (see Fig. 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 Equation (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 Equation (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

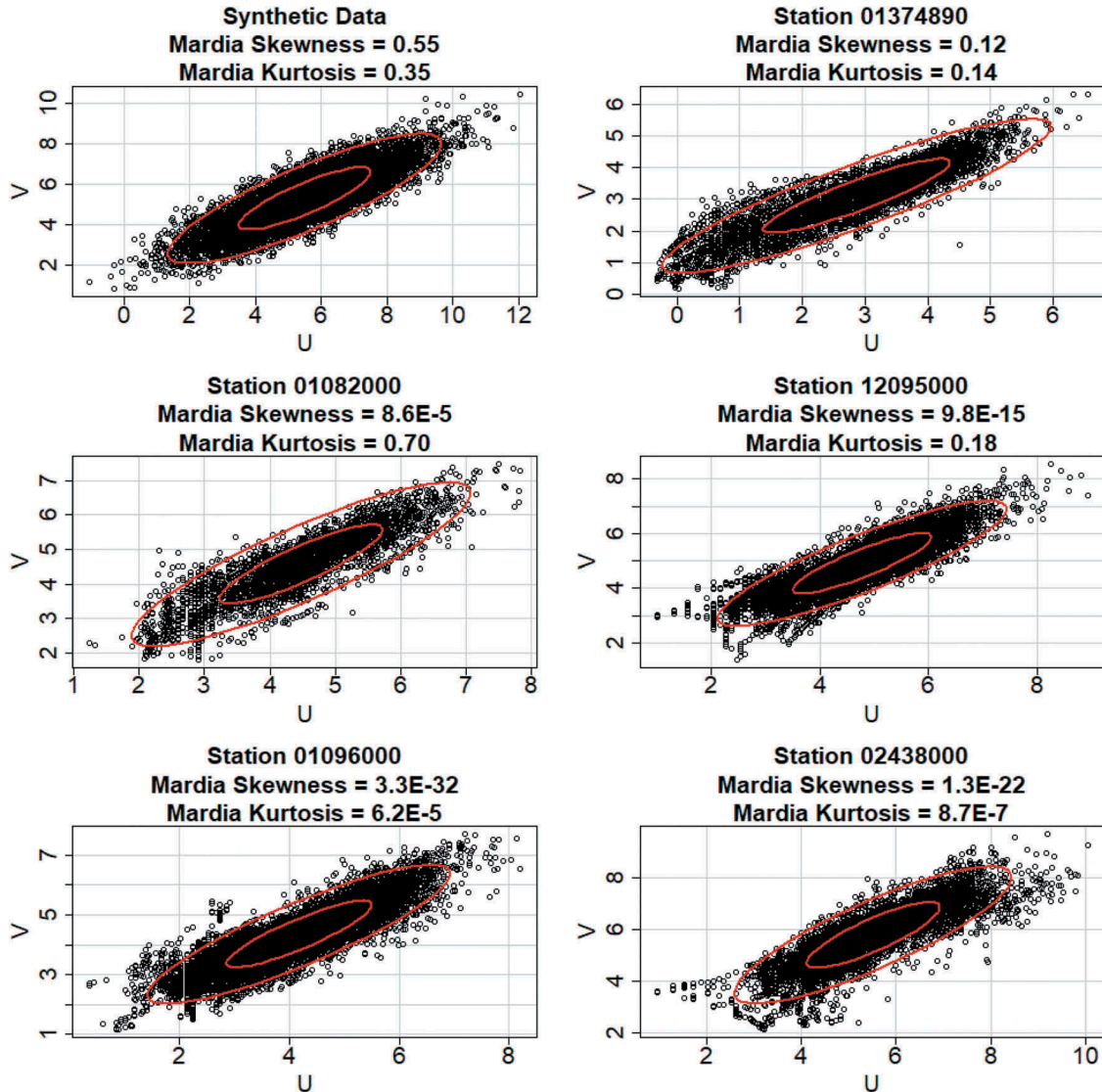


Figure 2. Goodness-of-fit evaluation of bivariate normality hypothesis with 50th and 95th confidence interval ellipses drawn. The upper left panel shows synthetic bivariate normal data and the remaining five panels are representative sites from the national dataset. Also shown are the p values corresponding to the Mardia skewness and kurtosis tests of bivariate normality.

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 non-linear 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 Equation (6) is a special case (for a bivariate lognormal model) of Equation (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, 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 non-linear 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 Equation (6). Here we consider a slight adaptation of Stedinger's (1981) estimator for use with bivariate LN3 samples given by:

$$r_1 = \frac{\exp[\hat{\sigma}_{UV}^2] - 1}{\sqrt{(\exp[\hat{\sigma}_U^2] - 1)(\exp[\hat{\sigma}_V^2] - 1)}} \quad (7)$$

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

$$\hat{\sigma}_{UV}^2 = \frac{1}{n} \sum_{i=1}^n (u_i - \bar{u})(v_i - \bar{v}) \quad (8a)$$

$$\hat{\sigma}_U^2 = \frac{1}{n} \sum_{i=1}^n (u_i - \bar{u})^2 \quad (8b)$$

and

$$\hat{\sigma}_V^2 = \frac{1}{n} \sum_{i=1}^n (v_i - \bar{v})^2 \quad (8c)$$

A very attractive and efficient estimator of the lower bounds τ_O and τ_S for use in Equation (7) is given in another paper by Stedinger (1980) as:

$$\hat{\tau}_O = \frac{o_{(1)}o_{(n)} - (o_{0.5})^2}{o_{(1)} + o_{(n)} - 2o_{0.5}} \quad (9a)$$

and

$$\hat{\tau}_S = \frac{s_{(1)}s_{(n)} - (s_{0.5})^2}{s_{(1)} + s_{(n)} - 2s_{0.5}} \quad (9b)$$

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 Equation (9a). Analogous definitions exist for estimation of $\hat{\tau}_S$ based on the simulations s .

Equation (7) is based on the relationship in Equation (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 Equation (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 Spearman's ρ for bivariate normal data, and (b) to account for the known relationship in Eq. (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:

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (10)$$

where d denotes the differences between the ranks and n is the sample size (Xu *et al.* 2013, Astivia and Zumbo 2017). 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:

$$\rho_{UV} = 2 \sin\left(\frac{\pi \rho_s}{6}\right) \quad (11)$$

Now replacing ρ_s with r_s in Equation (11), and combining Equation (11) with Equation (6), yields our modified Spearman correlation estimator r_2 which is designed to reproduce Pearson's ρ for the case of bivariate LN3 processes:

$$r_2 = \frac{\exp\left[2 \sin\left(\frac{\pi r_s}{6}\right) \hat{\sigma}_U \hat{\sigma}_V\right] - 1}{\sqrt{\exp(\hat{\sigma}_U^2) - 1} \sqrt{\exp(\hat{\sigma}_V^2) - 1}} \quad (12)$$

where r_s is given in Equation (10), and the estimators $\hat{\sigma}_U^2$ and $\hat{\sigma}_V^2$ are given in Equation (8) with $u_i = \ln[o_i - \hat{\tau}_O]$ and $v_i = \ln[s_i - \hat{\tau}_S]$ and the estimators $\hat{\tau}_O$ and $\hat{\tau}_S$ are given by Stedinger's (1980) lower bound estimator in Equation (9).

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 the 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 Equation (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 Equation (6) resulting in the following corrected RIN estimator:

$$r_3 = \frac{\exp(r_{\text{RIN}} \hat{\sigma}_U \hat{\sigma}_V) - 1}{\sqrt{\exp(\hat{\sigma}_U^2) - 1} \sqrt{\exp(\hat{\sigma}_V^2) - 1}} \quad (13)$$

where again the estimators $\hat{\sigma}_U^2$ and $\hat{\sigma}_V^2$ are given in Equation (8) with $u_i = \ln[o_i - \hat{\tau}_O]$ and $v_i = \ln[s_i - \hat{\tau}_S]$ and the estimators $\hat{\tau}_O$ and $\hat{\tau}_S$ are given by Stedinger's (1980) lower bound estimator in Equation (9).

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 Equations (4), (7), (12) and (13), respectively. Boxplots of the resulting values of the four estimators are illustrated in Fig. 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 Fig. 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 Fig. 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 the 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 Equation [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:

$$I = n / \left[\frac{(1 + \rho_s \rho_o)}{(1 - \rho_s \rho_o)} \right] = 0.10n \quad (14)$$

which indicates the gross reduction in information resulting from serial correlation. Thus, the results given in Fig. 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.

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 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 1. Statistics of streamflow records of 905 sites across the continental United States (also see Farmer and Vogel 2016a, 2016b).

Property	Average	Median	IQR (25th, 75th)	Range (min, max)
n	9827	10,957	(9862, 10,957)	(1262, 11,322)
C_O	2.3	1.9	(1.4, 2.7)	(0.5, 15.2)
C_S	2.0	1.4	(1.0, 1.9)	(0.2, 142.9)

5.2 Goodness of fit of the 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 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 Fig. 2. The p values associated with each of the Mardia test statistics are reported above each plot in Fig. 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 the 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 illustrates probability ellipses for the values of u and v corresponding to 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 Fig. 2 an example scatterplot and probability ellipses for synthetic MVN data. We conclude from Fig. 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 .

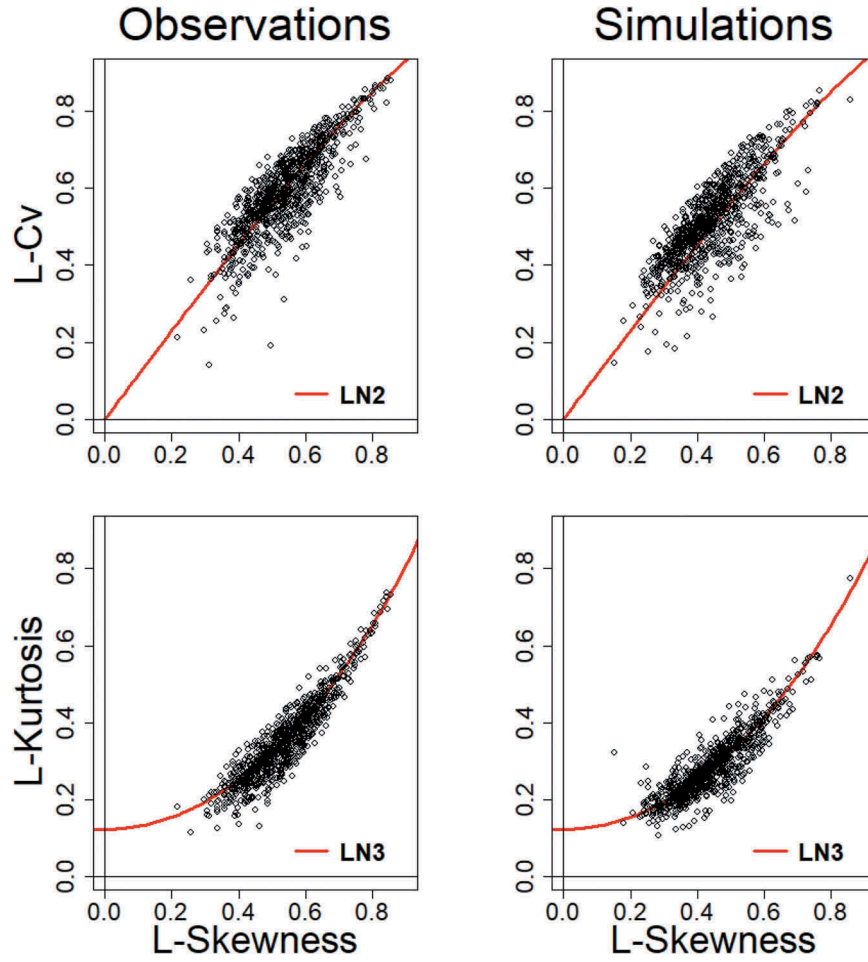


Figure 3. L-moment diagrams of observed and simulated average daily streamflows at the 905 sites summarized in Table 1.

5.2.2 Lognormal marginal distributions

In Fig. 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 Fig. 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 Fig. 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 Fig. 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, Fig. 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 Fig. 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 Fig. 2. On the basis of those experiments, they recommend the 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.

5.2.3 Assessment of dependence structure

The bivariate LN3 model assumes a particular theoretical dependence structure given by Equation (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

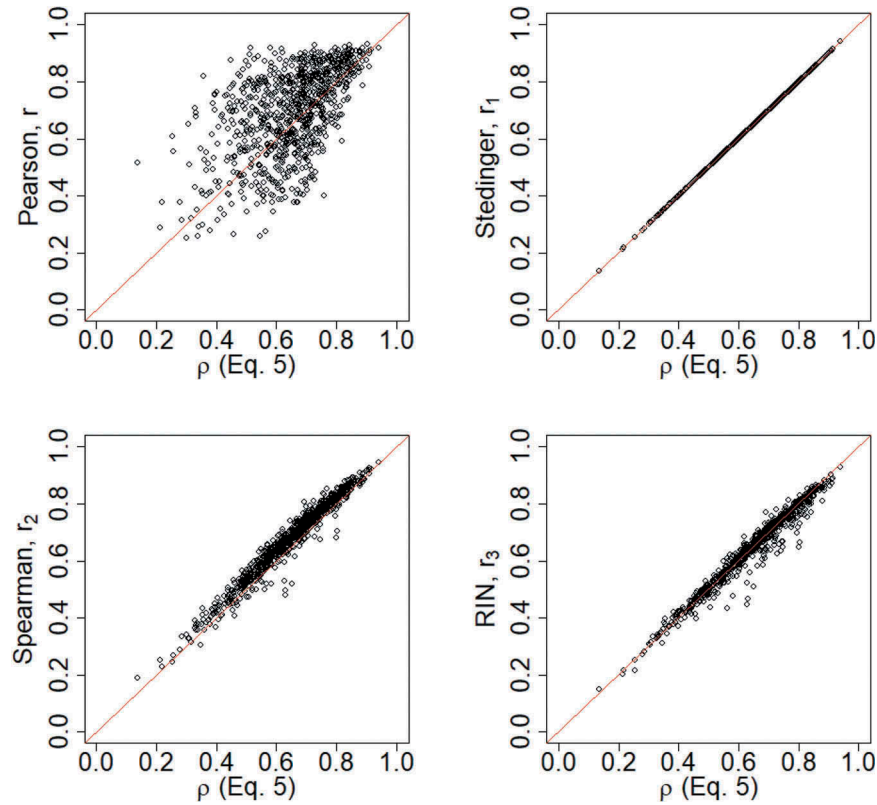


Figure 4. Evaluation of the ability of the four correlation estimators applied to the 905 sites summarized in Table 1, to reproduce the theoretical dependence structure associated with a bivariate LN3 process given by Equation (6).

theoretical dependence structure in (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 Fig. 4, we assess the degree to which the theoretical relationship between ρ and ρ_{UV} in Equation (6) is reproduced by the observations. Figure 4 illustrates scatterplots of the four estimators of ρ versus the corresponding estimates of ρ , which would be obtained by the application of Equation (6). Equation (6) is the theoretical version of its equivalent sample estimator which is the Stedinger estimator r_1 given in Equation (7). In other words, Stedinger's estimator r_1 is designed to reproduce, exactly, the theoretical dependence structure in Equation (6). The important result in Fig. 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 Figs. 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 Equation (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.

5.3 Comparisons among four correlation estimators

The left panels of Fig. 5 compare the magnitude of the Pearson r given in Equation (4) with the three competing correlation

estimators r_1 , r_2 and r_3 given in Equations (7), (12) and (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 Fig. 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 Fig. 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 Fig. 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 Fig. 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 Fig. 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 Fig. 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 Fig. 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

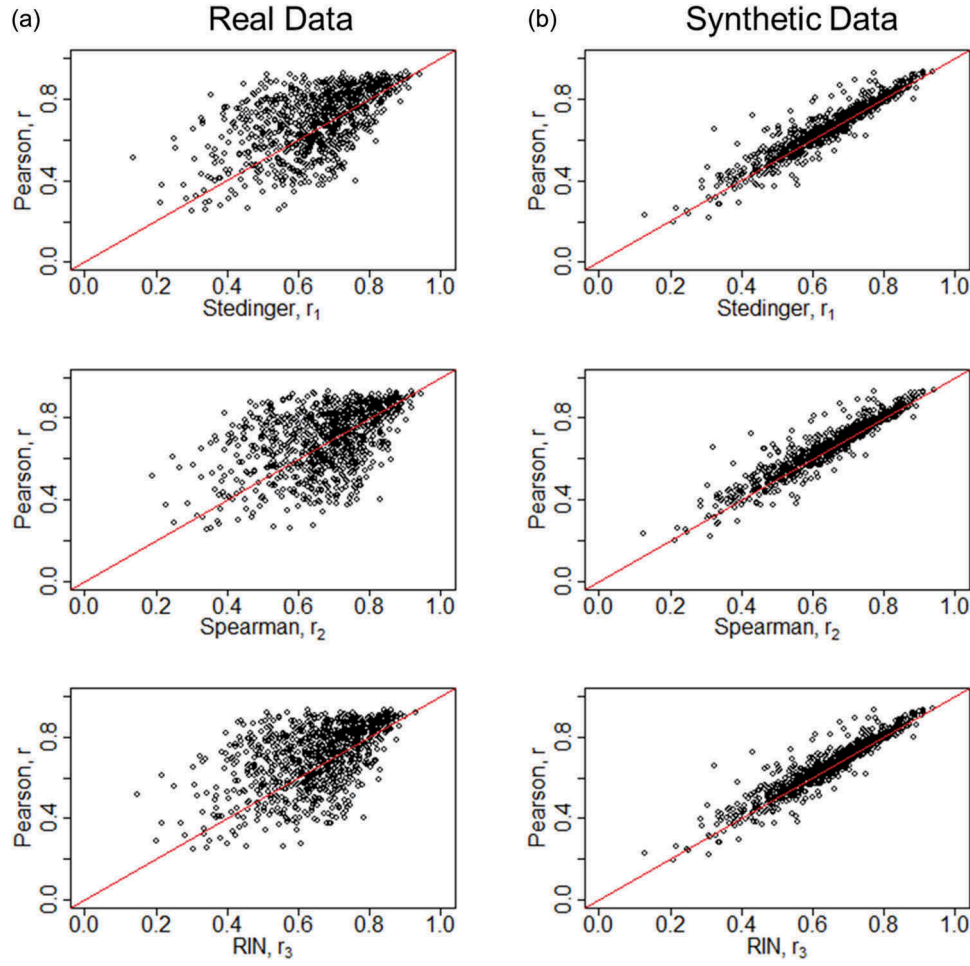


Figure 5. Comparison of Pearson's correlation estimator with the three alternative correlation estimators (left) using observations and simulations at the 905 sites summarized in Table 1 and (right) using synthetic bivariate lognormal series generated to reproduce the characteristics of the 905 sites summarized in Table 1.

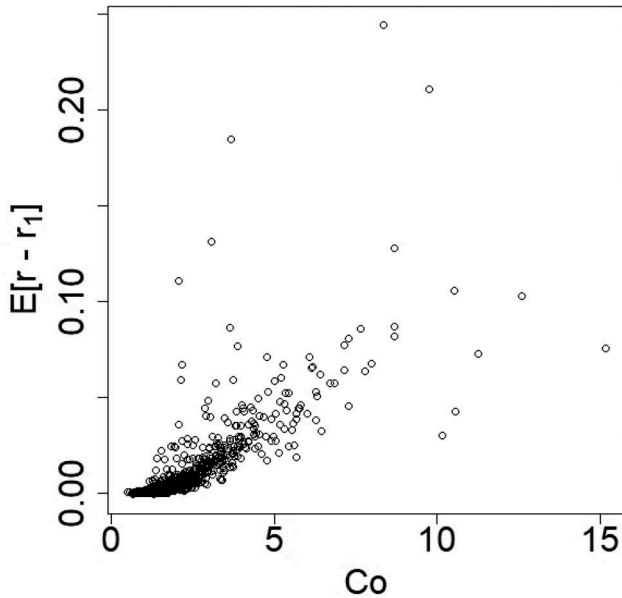


Figure 6. The expected difference between Pearson's estimator, r , and Stedinger's estimator, r_1 , as a function of the coefficient of variation of the observations. The values of $E[r - r_1]$ are computed from 500 synthetic bivariate lognormal traces generated to reproduce the characteristics of the bivariate observations and simulations at the 905 sites summarized in Table 1.

observed dependence structures than the Gaussian copula which is implied by the bivariate LN3 model in Equation (6).

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, Fig. 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 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 Fig. 1, we know that Pearson's r is generally upward biased and r_1 is generally unbiased for synthetic bivariate LN2 samples. Figure 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 Fig. 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.

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 Fig. 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 a 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 the 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, $NSE = \rho^2$; 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 the 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.

Acknowledgements

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.

We are also indebted to William Farmer of the U.S. Geological Survey, for sharing the rainfall runoff simulation data used in this study.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Richard M. Vogel  <http://orcid.org/0000-0001-9759-0024>

References

- Arbabshirani, M.R., et al., 2014. Impact of autocorrelation on functional connectivity. *NeuroImage*, 102, 294–308. doi:10.1016/j.neuroimage.2014.07.045
- Astivia, O.L.O. and Zumbo, B.D., 2017. Population models and simulation methods: the case of the Spearman rank correlation. *British Journal of Mathematical and Statistical Psychology*, 70, 347–367. doi:10.1111/bmsp.12085
- Balakrishnan, N. and Lai, C.D., 2009. *Continuous bivariate distributions*. 2nd ed. Dordrecht, The Netherlands: Springer.
- Beasley, T.M. and Erickson, S., 2009. Rank-based inverse normal transformations are increasingly used, but are they merited? *Behavior Genetics*, 39, 580–595. doi:10.1007/s10519-009-9281-0
- Bishara, A.J. and Hittner, J.B., 2015. Reducing bias and error in the correlation coefficient due to nonnormality. *Educational and Psychological Measurement*, 75, 785–804. doi:10.1177/0013164414557639
- Bishara, A.J. and Hittner, J.B., 2017. Confidence intervals for correlations when data are not normal. *Behavior Research Methods*, 49, 294–309. doi:10.3758/s13428-016-0702-8
- Blum, A.G., Archfield, S.A., and Vogel, R.M., 2017. On the probability distribution of daily streamflow in the United States. *Hydrology and Earth System Sciences*, 21, 3093–3103. doi:10.5194/hess-21-3093-2017
- Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R., 1975. Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62, 531–545. doi:10.1093/biomet/62.3.531
- Embrechts, P., McNeil, A.J., and Straumann, D., 2002. Correlation and dependence in risk management: properties and pitfalls. In: M.A. H. Dempster, ed. *Risk management: value at risk and beyond*. Cambridge, UK: Cambridge University Press, 176–223.
- Farmer, W.H. and Vogel, R.M., 2016a. On the deterministic and stochastic use of hydrologic models. *Water Resources Research*, 52, 5619–5633. 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*. Reston, VA: U.S. Geological Survey. doi: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.
- Genest, C. and Chebana, F., 2017. Copula modeling in hydrologic frequency analysis. Chapter 30. In: V.P. Singh, ed. *Handbook of applied hydrology* (pp. 1–10). New York, NY: McGraw-Hill Education.
- Guo, Y., Quader, A., and Stedinger, J.R., 2016. Analytical estimation of geomorphic discharge indices for small intermittent streams. *Journal of Hydrologic Engineering*, 21 (7), 04016015. doi:10.1061/(ASCE)HE.1943-5584.0001368
- Gupta, H.V., et al., 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *Journal of Hydrology*, 377, 80–91. doi:10.1016/j.jhydrol.2009.08.003
- Habib, E., Krajewski, W.R., and Ciach, G.J., 2001. Estimation of rainfall interstation correlation. *Journal of Hydrometeorology*, 2, 621–629. doi:10.1175/1525-7541(2001)002<0621:EORIC>2.0.CO;2
- Helsel, D., et al., 2019. *Statistical methods for water resources*. 2nd ed. Reston, VA: US Geological Survey.
- Helsel, D.R. and Hirsch, R.M., 2002. *Statistical methods in water resources techniques of water resources investigations*. Book 4, chapter A3. Reston, VA: US Geological Survey.
- 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., Kotz, S., and Balakrishnan, N., 1995. *Continuous univariate distributions*. Vol. 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., and Base, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, 5 (5), 89–97. doi:10.5194/adgeo-5-89-2005
- Kruskal, W.H., 1958. Ordinal measures of association. *Journal of the American Statistical Association*, 53 (284), 814–861. doi:10.1080/01621459.1958.10501481
- Lai, C.D., Rayner, J.C.W., and Hutchinson, T.P., 1999. Robustness of the sample correlation – the bivariate lognormal case. *Journal of Applied Mathematics & Decision Sciences*, 3 (1), 7–19. doi:10.1155/S1173912699000012
- Legates, D.R. and 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. doi:10.1029/97GL02207
- Legates, D.R. and McCabe, G.J., 1999. Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, 35 (1), 233–241. doi:10.1029/1998WR900018
- Limbrunner, J.F., Vogel, R.M., and Brown, L.C., 2000. Estimation of the harmonic mean of a lognormal variable. *Journal of Hydrologic Engineering*, 5 (1), 59–66. doi:10.1061/(ASCE)1084-0699(2000)5:1(59)
- Mardia, K.V., 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57, 519–530. doi:10.1093/biomet/57.3.519
- Markstrom, S.L., et al., 2015. *PRMS-IV, the precipitation-runoff modeling system, version 4: US Geological Survey techniques and methods*. Book 6, Chap. B7. Reston, VA: US Geological Survey. doi:10.3133/tm6B
- Matalas, N.C. and Langbein, W.B., 1962. Information content of the mean. *Journal of Geophysical Research*, 67 (9), 3441–3448. doi:10.1029/JZ067i009p03441
- McCuen, R.H. and Snyder, W.M., 1975. A proposed index for comparing hydrographs. *Water Resources Research*, 11 (6), 1021–1024. doi:10.1029/WR011i006p01021
- Meklin, C.J. and Mundfrom, D.J., 2004. An appraisal and bibliography of tests for multivariate normality. *International Statistical Review/Revue Internationale De Statistique*, 72 (1), 123–138.
- Moriasi, D.N., et al., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50, 885–900. doi:10.13031/2013.23153
- Mostafa, M.D. and Mahmoud, M.W., 1964. On the problem of estimation for the bivariate lognormal distribution. *Biometrika*, 51 (3/4), 522–527. doi:10.1093/biomet/51.3-4.522
- Papalexiou, S.M., 2018. Unified theory for stochastic modelling of hydro-climatic processes: preserving marginal distributions, correlation structures, and intermittency. *Advances in Water Resources*, 115, 234–252. doi:10.1016/j.advwatres.2018.02.013
- Pearson, K., 1896. Mathematical contributions to the theory of evolution III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 187, 253–318. doi:10.1098/rsta.1896.0007
- Pool, S., Vis, M., and Seibert, J., 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. and De Michele, C., 2013. Multivariate extreme value methods. In: A. AghaKouchak et al. eds. *Extremes in a changing climate* (pp. 115–162). Dordrecht, Netherlands: Springer. doi:10.1007/978-94-007-4479-0_5
- Salvadori, G., et al., 2007. *Extremes in nature: an approach using copulas*. Vol. 56. Dordrecht, Netherlands: Springer Science & Business Media.
- 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. doi:10.1007/s00477-007-0176-4

- Shimizu, K., 1993. A bivariate mixed lognormal distribution with an analysis of rainfall data. *Journal of Applied Meteorology*, 32, 161–171. doi:10.1175/1520-0450(1993)032<0161:ABMLDW>2.0.CO;2
- Spearman, C., 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15, 72–101. doi:10.2307/1412159
- Stedinger, J.R., 1980. Fitting lognormal distributions to hydrologic data. *Water Resources Research*, 16 (3), 481–490. doi:10.1029/WR016i003p00481
- Stedinger, J.R., 1981. Estimating correlations in multivariate streamflow models. *Water Resources Research*, 17 (1), 200–208. doi:10.1029/WR017i001p00200
- Tsoukalas, I., Efstratiadis, A., and Makropoulos, C., 2018. Stochastic Periodic Autoregressive to Anything (SPARTA): modeling and simulation of cyclostationary processes with arbitrary marginal distributions. *Water Resources Research*, 54, 161–185. doi:10.1002/2017WR021394
- Vogel, R.M., 2017. Stochastic watershed models for hydrologic risk management. *Water Security*, 1, 28–35. doi:10.1016/j.wasec.2017.06.001
- Vogel, R.M. and Fennessey, N.M., 1993. L-moment diagrams should replace product-moment diagrams. *Water Resources Research*, 29 (6), 1745–1752. doi:10.1029/93WR00341
- Vogel, R.M., Stedinger, J.R., and Hooper, R.P., 2003. Discharge indices for water quality loads. *Water Resources Research*, 39 (10), 1273. doi:10.1029/2002WR001872
- Wilks, D.S., 2006. *Statistical methods in the atmospheric sciences*. 2nd ed. Burlington, MA: Academic Press.
- Willmott, C.J., 1981. On the validation of models. *Physical Geography*, 2, 184–194. doi:10.1080/02723646.1981.10642213
- Willmott, C.J., et al., 1985. Statistics for the evaluation and comparison of models. *Journal of Geophysical Research*, 90, 8995–9005. doi:10.1029/JC090iC05p08995
- Xiao, Q., 2014. Evaluating correlation coefficient for Nataf transformation. *Probabilistic Engineering Mechanics*, 37, 1–6. doi:10.1016/j.probengmech.2014.03.010
- Xu, W., et al., 2013. A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models. *Signal Processing*, 93, 261–276. doi:10.1016/j.sigpro.2012.08.005
- Zhang, G. and Chen, Z., 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 the 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 $\sigma_\varepsilon^2 = \sigma_V^2(1 - \rho_{UV}^2)$.