NOTES AND CORRESPONDENCE

Estimation of Rainfall Interstation Correlation

EMAD HABIB* AND WITOLD F. KRAJEWSKI

IIHR—Hydroscience & Engineering, The University of Iowa, Iowa City, Iowa

GRZEGORZ J. CIACH

Environmental Verification and Analysis Center, The University of Oklahoma, Norman, Oklahoma

8 May 2001 and 27 August 2001

ABSTRACT

This study discusses questions of estimating correlation coefficient of point rainfall as observed at two measuring stations. The focus is on issues such as sensitivity to sample size, extreme rain events, and distribution of rainfall. The authors perform extensive analysis based on a two-point data-driven rainfall model that simulates the intermittence and extreme variability of rainfall using a bivariate mixed-lognormal distribution. The study examines the commonly used product–moment estimator along with an alternative transformation-based estimator. The results show a high level of bias and variance of the traditional correlation estimator, which are caused mostly by significant skewness levels that characterize rainfall observations. Application using data from a high-density cluster indicated the advantages of using the alternative estimator. The overall aim of the study is to draw the attention of researchers working with rainfall to some commonly disregarded issues when they seek accurate and reliable correlation information.

1. Introduction

In this study, we bring to the attention of the hydrometeorological community a little known fact that estimation of the correlation coefficient from a bivariate sample of skewed data results in biased estimates of correlation coefficient. Our focus is on rainfall data, in particular at smaller temporal and spatial scales on which both rainfall rate and rainfall accumulation exhibit considerable skewness. The statisticians have known about this bias problem for years, but numerous works on rainfall analysis neglected the finding. This neglect prompted us to report on it. The issue is important because the bias may be as high as 0.1 or more, painting an overoptimistic picture of model and algorithm performances.

Correlation coefficient is a measure of linear dependency between a pair of random variables. Correlation often is used as a measure of agreement between rainfall estimates based on different sensors; radar–rain gauge comparison studies are a prime example. Correlation of rainfall at two or more gauging stations is used in data quality-control applications, rainfall prediction, and climate characterizations. Correlation has been widely used to characterize the complex spatial structure of rainfall patterns (e.g., Huff and Shipp 1969; Huff 1970; Sharon 1972; Zawadzki 1973). In addition, spatial correlation is a significant component of hydrological design, in particular, design of rainfall networks (e.g., Bras and Rodriguez-Iturbe 1993). Evaluation of the errors of area-averaged rainfall accumulations is largely determined by information provided by the spatial correlation function in the rain fields (Morrissey et al. 1995; Bras and Rodriguez-Iturbe 1993).

For the problem discussed herein, the variables are concurrent rainfall quantities at two locations. In the past, the sample Pearson product–moment correlation coefficient r has been used commonly to estimate the population coefficient ρ [see Eqs. (1) and (2) in section 3]. Many researchers argued that estimation of the correlation coefficient suffers from several shortcomings. For example, in his evaluation of climate model performance, Willmott (1984) illustrated that r is insensitive to additive and proportional differences that may exist between the two variables, which causes a problem if r is used as a performance measure. Like many other statistical measures, r is much more influenced by out-

* Current affiliation: Department of Civil and Environmental Engineering, Tennessee Technological University, Cookeville, Tennessee.

Corresponding author address: Witold Krajewski, IIHR—Hydroscience & Engineering, The University of Iowa, Iowa City, IA 52242. E-mail: witold-krajewski@uiowa.edu

liers than by observations near the mean (Moore 1991). Legates and Davis (1997) and Legates and McCabe (1999) discussed the effect of extreme values on the evaluation of goodness-of-fit measures of hydrologic and hydroclimatic models. They illustrated that the existence of extreme events will lead to an inflated estimate of r that may obscure the true relationship over most of the range of observations. Based on an idealized analysis of radar and rain gauge comparisons, Kessler and Neas (1994) showed that the magnitude of the correlation coefficient tends to increase with the range over which the variables are measured.

Researchers also noted other problems with estimating correlation. Cressie (1993) argued that, as compared with other spatial association measures such as the variogram, the correlation function is not preferred given that bias and trend contamination errors heavily affect its estimates. A similar limitation of the use of the correlation coefficient arises when temporal records of observations are used to estimate the spatial correlation structure. Gunst (1995) illustrated that, with meteorological data, correlations calculated from time series overestimate spatial correlations when temporal trends and autocorrelations exist in the observations.

Statistical literature provides evidence that the source of the problems is the underlying distribution of the variables under consideration. In the case of bivariate normal distribution, the distribution of r is known well (e.g., chapter 32 of Johnson et al. 1995; chapter 6 of Stuart and Ord 1987) and can be used to form confidence intervals or to perform significance tests. However, in the case of nonnormal bivariate populations, little is known about the distribution of r. Kowalski (1972) illustrated that the distribution of r may be very sensitive to nonnormality in the data and suggested that normal correlation analysis should be limited to situations in which the variables are nearly normal. Hutchinson (1997) suggested that, for very skewed distributions, the Pearson sample correlation coefficient is a poor estimator of the population value.

Although several studies have investigated the robustness of the correlation coefficient estimation in case of nonnormal data, statistical literature does not provide a practical methodology to determine the distribution of the sample correlation. Contradicting results are often reported about the robustness of the Pearson sample correlation estimator; see Johnson et al. (1995, chapter 32) for an extensive review. Tedious numerical analyses such as cross-cumulant and asymptotic expansions are required (Nakagawa and Niki 1992; Lai et al. 1999) to address the issue of uncertainties in the correlation estimation.

In the hydrological literature, few studies addressed the correlation estimation problems. Stedinger (1981) presented an extensive investigation of alternative estimators of the correlation coefficients of streamflows. He used a simulation-based analysis to show that, in the case of skewed data such as streamflows, better estimates are possible if logarithmically transformed variables are used to estimate the true correlations. Such arguments indicate the need for a similar practice in the case of rainfall data, which are characterized by even higher skewness levels. Our review of the rainfall literature indicated that, despite the extensive use of correlation-based analyses, it is unclear to what extent researchers are aware of these limitations. Correlation estimates for rainfall patterns observed via a variety of sensors are often reported without discussing relevant issues such as the distribution of the analyzed data, sample size effects, and uncertainty bounds of the computed coefficients (e.g., Moszkowicz 2000; May and Julien 1998). However, in recent studies, Young et al. (2000) and Krajewski et al. (2000) discussed the significant scatter in the computed coefficients and the inability of the normal theory (Stuart and Ord 1987) to explain them; these authors indicated the need to analyze and to explain the sources of such scatter.

In the current study, we investigate and illustrate the relevance of such problems to the correlation estimation in rainfall data. We follow the approach presented by Stedinger (1981) and examine it using rainfall data, which, in contrast to streamflows, has a mixed-type distribution because of the intermittent behavior of the precipitation process. We have organized this paper as follows. We begin with a brief description of the experimental rainfall datasets used throughout the study to illustrate our points. We then briefly discuss the simulation-based analysis with which we investigated the robustness of the correlation coefficient. After that, we examine the performance of the transformation-based estimator and apply it to the experimental data. We close with discussion and conclusions.

2. Data description

To illustrate our investigation, we use rainfall data from a high-density network (cluster) of rain gauges. The cluster was deployed near Melbourne, Florida, as a part of the Texas and Florida Underflights field campaign (TEFLUN-B), in August and September of 1998, in support of the ground validation of the Tropical Rainfall Measuring Mission (TRMM; Simpson et al. 1996). The network consisted of 14 8-in. tipping-bucket (TB) gauges. The separation distances between gauges ranged from a few hundred meters up to about 8 km, covering spatial scales typical for resolution of radar-rainfall products used for hydrologic applications. Figure 1 shows a layout of this experimental network. Within this network, another cluster of rain gauges, operated by The University of Iowa, was also deployed in which three dual-gauge platforms (Krajewski et al. 1998) were set with gauge-separation distances of 1, 10, and 100 m. All the gauges in the cluster have high sampling resolution in time (5 or 10 s), and a bucket size that corresponds to 0.25 mm. Herein we focus on the data collected during the 2-month period from August through September of 1998. Despite this relatively short

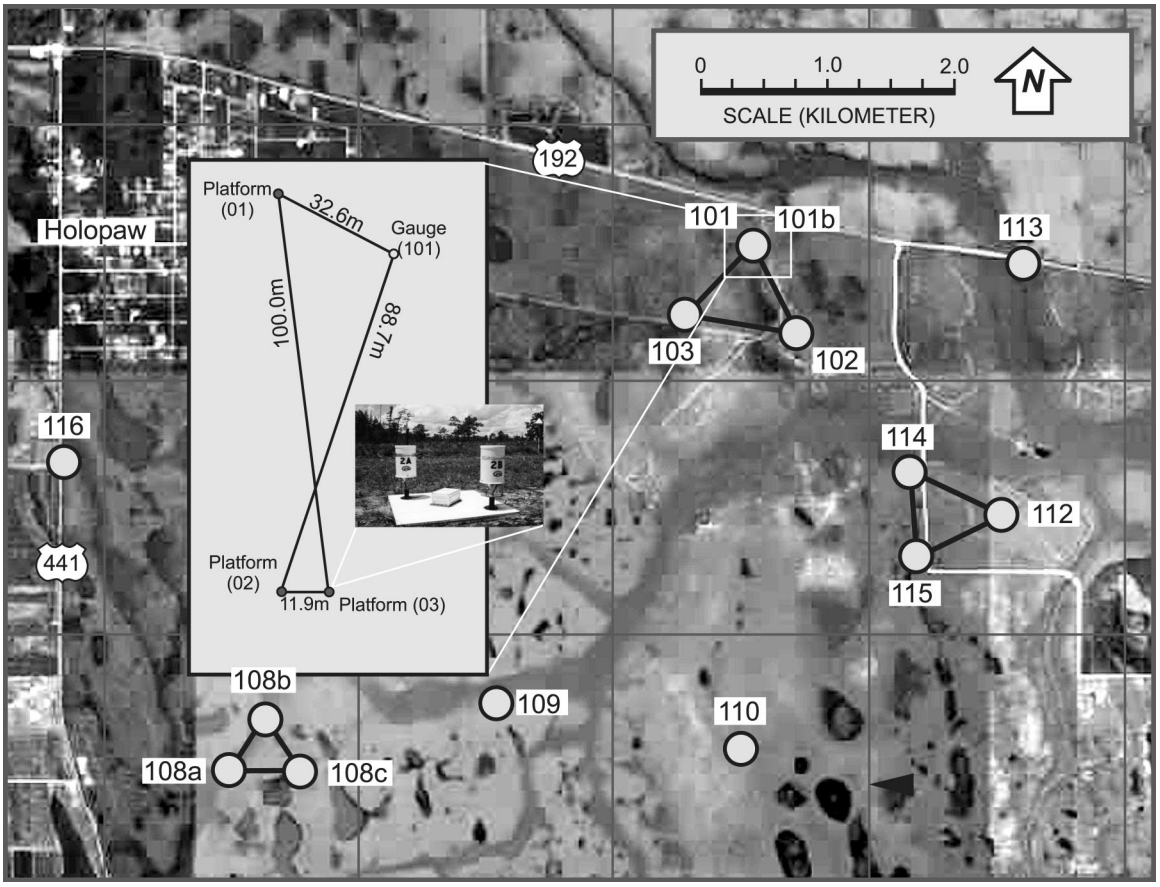

FIG. 1. Aerial photo of the National Aeronautics and Space Administration (NASA) dense rain gauge network deployed during the TEFLUN-B field campaign, Melbourne, FL, Aug–Sep 1998. The rectangular grid spacing is 2 km. The circles show gauge locations with the numbering system that corresponds to the NASA-maintained TRMM database. The insert shows the dual-gauge platforms configuration.

time, the total amount of accumulated rainfall was very high, ranging from 350 to 450 mm, depending on the rain gauge. Using a linear interpolation scheme (Habib et al. 2001), we converted the raw rain gauge data into estimates of 1-min rain rates. These fine-resolution estimates can be used to construct rain-rate averages at any desired time interval for further analysis. For discussion of accuracy of TB-based point rainfall estimates, please see Humphrey et al. (1997) and Habib et al. (2001).

To the extent possible, we performed careful quality control of the gauge records to detect instances of mechanical or electronic failures of the kind that often affect the TB-type gauges. This check was facilitated by the closely located gauges. Thus, we believe that the behavior of the analyses reported in subsequent sections of this paper is not contaminated by the occurrence of outliers in the dataset.

3. Empirical properties of correlation coefficient in rainfall

Let (X, Y) denote a pair of rainfall processes observed at two gauge locations. The population correlation coefficient can be defined as

$$\rho(X, Y) = \frac{E(XY) - E(X)E(Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}. \quad (1)$$

The following Pearson product–moment sample coefficient r, derived from N observations of the pair (X, Y), is usually used as an estimate of $\rho(X, Y)$:

$$r(X, Y) = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sqrt{(\overline{X^2} - \overline{X}^2)(\overline{Y^2} - \overline{Y}^2)}}, \quad (2)$$

where the overbars indicate average values over the sample of size N. In this section, we elaborate on some of the problems associated with the estimation of correlation from rainfall data. Effect of sample size was addressed by Berndtsson (1987). Using daily rainfall data, he showed that scatter in correlation variation with distance increases significantly when the sample size is reduced. We performed similar investigations using the TEFLUN-B gauges and computed the correlation coefficients for the entire sample and for smaller-size subsamples at a timescale of 15 min. We observed that when the sample size is reduced, the scatter becomes more pronounced and the estimated values become more biased upward. Application of the commonly used ''Fisher

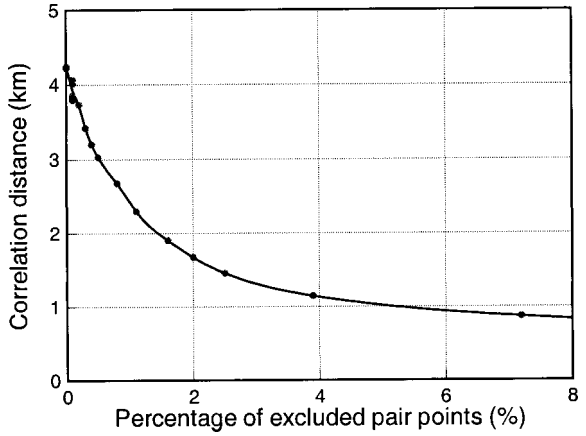

FIG. 2. Variation of correlation distance with the percentage of highest rainfalls excluded from the correlation coefficient estimation.

z transformation'' (Fisher 1921; Stuart and Ord 1987) does not explain the observed scatter. This fact indicates that the correlation estimation in rainfall is highly sensitive to the sample size and cannot be treated with the normal-based traditional statistical methods. To investigate this issue empirically, extremely long records of homogeneous rainfall observations are needed. The question, ''How long do the datasets need to be to obtain stable statistics?,'' cannot be answered with the usual setups of rainfall-measuring networks. Thus, we will investigate this issue using simulation.

Another problem with the correlation analysis comes from the fact that storms are characterized by short intense rainfall periods often associated with localized convective cells. To examine the sensitivity of the correlation coefficient to the existence of such high values, we again considered the correlation structure estimated for the TEFLUN-B network at 15-min timescale. We used a least squares method to fit the variation of the correlation coefficient r with distance d, assuming the following isotropic exponential model:

$$r(d) = r_0 \exp[(-d/d_0)^{\mu}], \tag{3}$$

where r_0, d_0, and μ are parameters. The parameter d_0 is usually referred to as the correlation distance and characterizes the decay of correlation. We recomputed the correlation coefficients for gauge pairs after cutting off the tail of their rainfall distributions at various levels. This approach corresponds to excluding certain portions of the highest rainfalls. In Fig. 2, the correlation distance d_0 is plotted as a function of the percentage of the excluded points. It shows that the correlation distance decreases rapidly when extreme rain rates are excluded from the correlation calculations. Using all the observations, d_0 attains a value of about 4.2 km, but after we cut off as little as 1% of the highest rain rates, it decreases to 2.7 km and to about 1 km after rejecting 5% of points. This demonstrates that the correlation is highly sensitive to the extreme rainfall values. Because extreme rainfall events are short-lived and infrequent, the sample product–moment coefficient [Eq. (2)] is not a robust estimate of the population parameter: adding just a few more extreme values to the sample significantly changes its estimated values. The above analysis also implies that for highly skewed joint distributions the correlation may be overestimated.

4. Monte Carlo simulations

To investigate the performance of the sample correlation coefficient with respect to the underlying distribution of the rainfall data, we performed a simulation-based analysis. Rain rate has a mixed-type distribution, because there is a positive probability of no rain caused by the intermittent behavior of the rainfall process. One accordingly, has to decide on the continuous-distribution component of rain rate. Kedem et al. (1990) presented a brief review of several models that have been suggested in the literature and indicated that distributions such as gamma, hyperbolic, and lognormal are all possible candidates. Houze and Cheng (1977), Kedem and Chiu (1987), Shimizu (1993), and Ciach and Krajewski (1999) all adopted the lognormal distribution. In addition, researchers often use logarithmic transformation of rainfall data (e.g., McDonald 1960; Kitchen and Blackall 1992; Smith and Krajewski 1991; Anagnostou et al. 1999). Considering that there is no general agreement within the research community on a specific form of the distribution (Kedem et al. 1990), we chose to use the lognormal model as the underlying distribution of rain rate in the simulation analysis. In the application described in next sections, we will evaluate the agreement of actual rainfall data with the lognormal distribution.

a. Bivariate intermittent lognormal distribution

A realistic distribution that takes into account the intermittent nature of rainfall was developed by Shimizu (1993) as a part of a two-point rainfall model and serves as a useful tool for our simulations. The distribution is explained in detail in Shimizu's paper, and we only present a brief description here. The distribution allows representation of situations with zero rainfall at either of the two observing stations. Four cases have to be considered: zero rain at both stations (0, 0), zero at one station and rain at the other (x^*, 0) and (0, y^*), and rain at both stations (x, y). For a sample size of N, the four cases will have sample sizes of n_0, n_1, n_2, and n_3, where $n_0 + n_1 + n_2 + n_3 = N$. The distribution accordingly can be described fully by the following parameters: δ_0, δ_1, δ_2, δ_3, μ_1^*, σ_1^*, μ_2^*, σ_2^*, μ_1, μ_2, σ_1, σ_2, and ρ_N. Parameters δ_i ($i = 0, 1, 2, 3$) are the probabilities of rainfall occurrence for each case. Parameters μ_i^* and σ_i^* ($i = 1, 2$) are conditional means and standard deviations for cases n_1 and n_2. Parameters μ_i and σ_i ($i = 1, 2$), and ρ_N are, respectively, conditional means, stan-

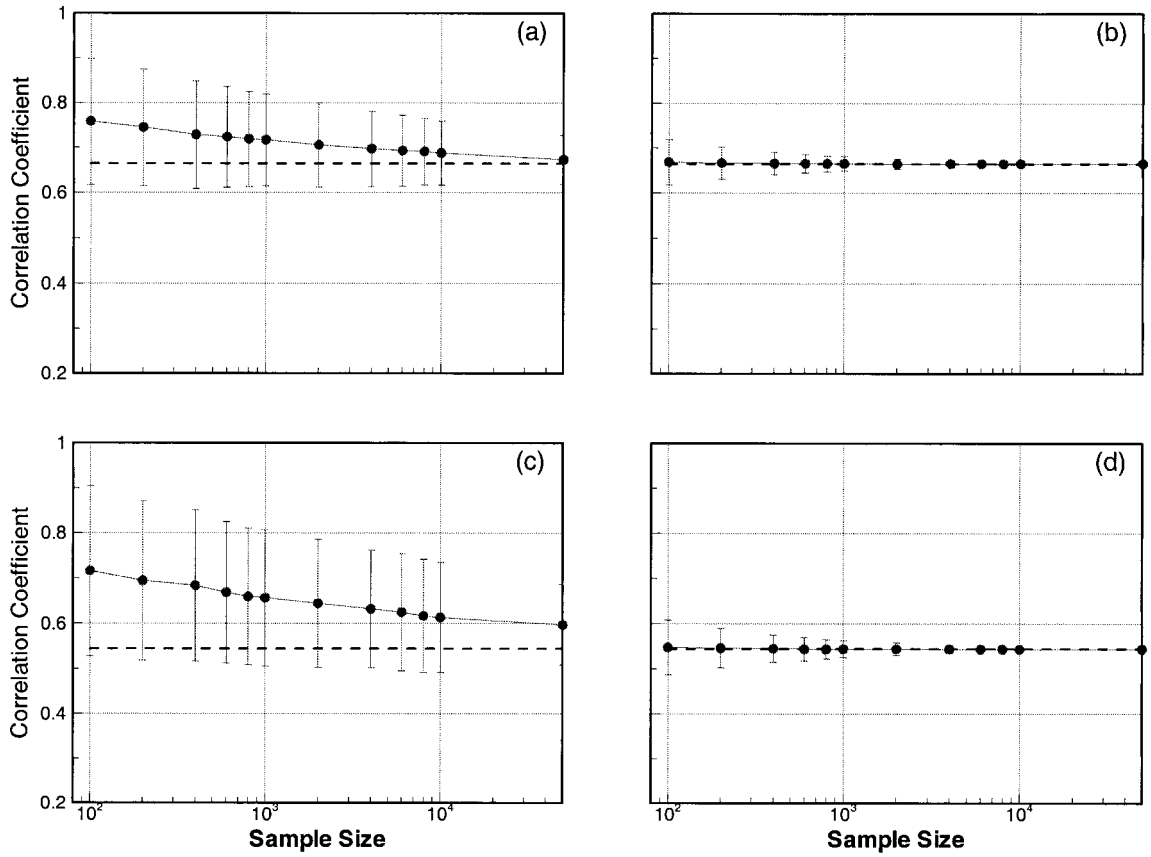

FIG. 3. Averages and standard errors (vertical bars) of the correlation coefficient estimates as a function of sample size. Samples of bivariate mixed-lognormal distribution were simulated with parameters (a), (b) $\sigma_1 = \sigma_2 = 1.6$ and (c), (d) $\sigma_1 = \sigma_2 = 2$. The rest of the parameters had the following values: $\delta_0 = 0$, $\delta_1 = \delta_2 = 0.2$, $\delta_3 = 0.6$, $\mu_1^* = \mu_2^* = -1.5$, $\mu_1 = \mu_2 = 0.25$, $\rho_N = 0.85$, and $\sigma_1^* = \sigma_2^* = 1$. Horizontal dashed lines show population correlation values: (a), (b) 0.66 and (c), (d) 0.54. Two correlation estimators were used: (a), (c) the Pearson formula and (b), (d) the transformation-based estimator.

dard deviations, and correlation coefficient for case n_3. We remind the reader that this set of parameters corresponds to normal distributions of the logarithmic transformation of the variables X and Y. Note also that the set of standard deviations σ_i and σ_i^* of the normal variables determines the skewness of the distribution; that is, higher standard deviations result in a more skewed lognormal distribution. For specific values of these parameters, simulation of such a distribution is possible through assigning the desired probabilities δ_i ($i = 0, 1, 2, 3$), sampling from a normal distribution for both of the cases (x^*, 0) and (0, y^*) and sampling from a bivariate normal distribution for the case (x, y). Appropriate exponential transformations then can be applied to get the desired intermittent bivariate lognormal sample.

Following the method of Shimizu (1993), population means and variances of the lognormal variables are expressed as follows:

$$E(X) = \delta_1 \exp[\mu_1^* + (\sigma_1^{*2}/2)] + \delta_3 \exp[\mu_1 + (\sigma_1^2/2)],$$

$$E(Y) = \delta_2 \exp[\mu_2^* + (\sigma_2^{*2}/2)] + \delta_3 \exp[\mu_2 + (\sigma_2^2/2)],$$

$$E(XY) = \delta_3 \exp[\mu_1 + \mu_2 + (\sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2\rho_N)/2], \quad (4a)$$

$$\mathrm{Var}(X) = \delta_1 \exp(2\mu_1^* + 2\sigma_1^{*2}) + \delta_3 \exp(2\mu_1 + 2\sigma_1^2) - \{\delta_1 \exp[\mu_1^* + (\sigma_1^{*2}/2)] + \delta_3 \exp[\mu_1 + (\sigma_1^2/2)]\}^2, \quad \text{and}$$

$$\mathrm{Var}(Y) = \delta_2 \exp(2\mu_2^* + 2\sigma_2^{*2}) + \delta_3 \exp(2\mu_2 + 2\sigma_2^2) - \{\delta_2 \exp[\mu_2^* + (\sigma_2^{*2}/2)] + \delta_3 \exp[\mu_2 + (\sigma_2^2/2)]\}^2. \quad (4b)$$

After substituting these expressions into Eq. (1), an expression of the population correlation coefficient for the mixed-lognormal variables can be obtained.

b. Simulation results

In the following data-driven simulations, we examine the performance of the correlation estimation using the bivariate intermittent lognormal distribution. We evaluate the effects of sample size N, degree of skewness (measured by σ), and levels of population correlation. We varied the simulated sample sizes from 100 to 50 000

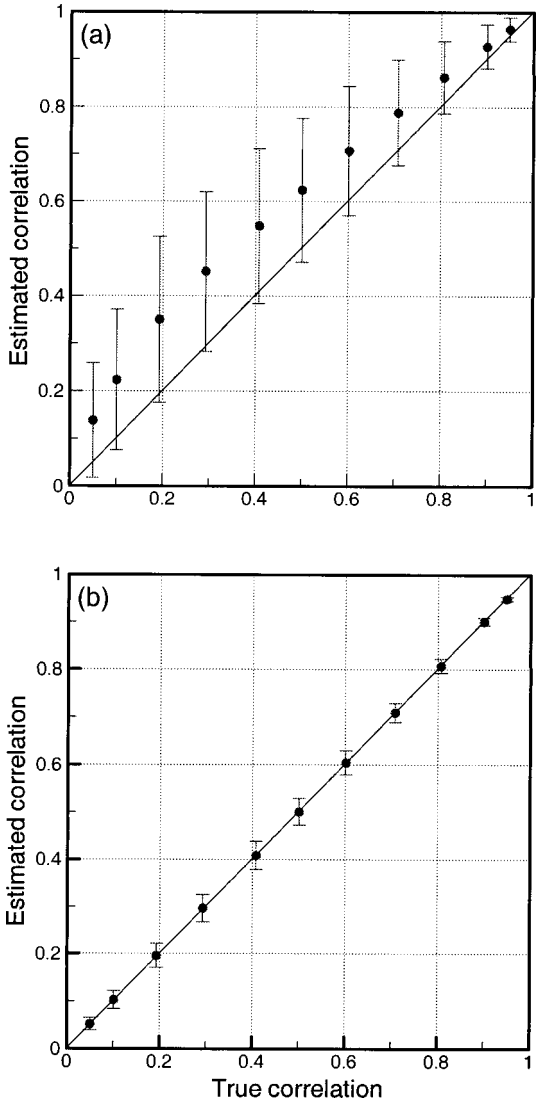

FIG. 4. Comparison of averages and standard deviations (vertical bars) of correlation coefficient estimates for (a) Pearson formula and (b) the transformation-based estimator at different population correlation levels. Samples of size 500 were simulated from the bivariate intermittent lognormal distribution with parameters $\delta_0 = 0$, $\delta_1 = \delta_2 = 0.2$, $\delta_3 = 0.6$, $\mu_1^* = \mu_2^* = -1.5$, $\mu_1 = \mu_2 = 0.25$, $\sigma_1^* = \sigma_2^* = 1$, and $\sigma_1 = \sigma_2 = 2$.

and fixed the rest of the parameters at $\delta_0 = 0$, $\delta_1 = \delta_2 = 0.2$, $\delta_3 = 0.6$, $\mu_1^* = \mu_2^* = -1.5$, $\mu_1 = \mu_2 = 0.25$, $\sigma_1^* = \sigma_2^* = 1$,$\sigma_1 = \sigma_2 = 1.6$, and $\rho_N = 0.85$; this set of values corresponds to a true population correlation coefficient ρ of 0.66. These particular values of the parameters are realistic because we used the actual rainfall data (the TEFLUN-B sample described in section 2) to obtain them.

For the case of a bivariate continuous lognormal distribution, Lai et al. (1999) showed that the Pearson estimator in Eq. (2) suffers from significant bias and scatter levels. Stedinger (1981) alternatively suggested estimating the correlation coefficient after performing logarithmic transformation of the considered variables. We chose to compute the sample correlation coefficient by estimating $E(X)$, $E(Y)$, $E(X, Y)$, $\mathrm{Var}(X)$, and $\mathrm{Var}(Y)$ in two ways: first by applying the Pearson estimator [Eq. (2)] to the untransformed sample and second by using the logarithmic-transformed pairs to get sample estimates of the parameters μ^*, μ, σ^*, σ, and ρ_N, which were in turn plugged into Eq. (4) to get the desired statistics. The simulation results—that is, averages and standard errors of the estimated correlations—are plotted versus the sample size in the upper panels of Fig. 3. We used a logarithmic scale for clarity. To examine the effect of increased skewness, we carried another set of simulations in which σ_1 and σ_2 were increased to 2 that correspond to a population correlation coefficient of 0.54. For this case, the results for both correlation estimation methods are shown in the lower panels of Fig. 3. The plots indicate that significant bias and error characterize the sample correlation coefficient when estimated in the untransformed space. In addition, the bias and variance do not decay fast enough with the increase of the sample size, and they notably increase, as expected, with the increase in the skewness of the sample. On the other hand, the alternative transformation-based estimator shows negligible bias and much lower variance values that drop fast with the increasing sample size.

Last, we evaluated the two estimators at various levels of correlations. We performed the simulations using a fixed set of parameters except for ρ_N, which was varied to obtain several levels of the population correlation ρ. The results are shown in Fig. 4 and confirm the lack of bias and low error of the alternative estimator. The results prove that the transformation-based computation of sample correlations should be favored whenever the data appear to follow the probability distribution described above.

5. Application to TEFLUN-B dataset

In this section, we illustrate the application of the correlation estimation procedure using the rainfall dataset that was collected during August–September of 1998 as part of the TEFLUN-B field experiment (see section 2). We used a timescale of 15 min, which is of interest in radar-rainfall studies. Also, the shorter the timescale is, the more skewed the rainfall data are; thus it is easier to illustrate the problem at hand. First, we test the assumption of lognormality for this dataset. Testing for lognormality is equivalent to testing for normality after performing log-transformation of the rainfall values. We did this for each gauge separately and then for the bivariate sample of every pair of gauges. We applied two statistical tests to the transformed data: the Shapiro–Wilk W test to univariate samples and the Mardia's multivariate skewness and kurtosis-based test to the bivariate sample. Details of these tests are available in most statistics books, and here we discuss only

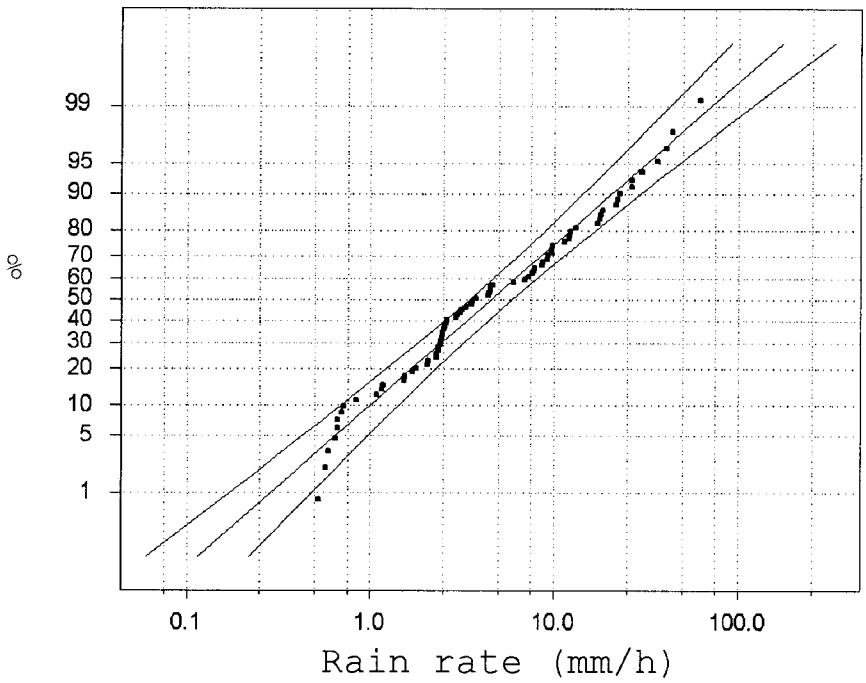

FIG. 5. Probability plot of rain gauge 15-min data fitted using a lognormal distribution; 95% confidence interval bounds are also plotted for illustration.

the results. Using the data collected by the 14-gauge TEFLUN-B cluster, we performed the tests for each pair of rain gauges with the subsets that correspond to cases (x^*, 0), (0, y^*), and (x, y). Both sets of tests accepted the lognormality hypothesis for almost all the cases. For illustration, we show in Fig. 5 an example of the lognormal distribution fitted to the time series of one of the gauges. Despite slight deviations at both tails, the plot indicates a fair agreement between the experimental data and the assumed distribution within 95% confidence bounds.

After the tests, we proceeded with the correlation estimation based on the assumption of mixed bivariate lognormal distribution. Following the procedure described in the previous section, we estimated the parameters δ_0, δ_1, δ_2, δ_3, μ_1^*, μ_2^*, μ_1, μ_2, σ_1^*, σ_2^*, σ_1, σ_2, and ρ_N for each pair of gauges. Then, for each set of these estimates, we generated 1000 samples of the same size from the mixed-lognormal distribution corresponding to the same parameter values. We then used each of these samples to estimate the correlation coefficient using the transformation-based estimator described in the previous section. Standard errors for each set of 1000 estimated correlation coefficients were computed to obtain the uncertainty bounds of the estimates. In Fig. 6a, averages of the 1000 estimated correlation coefficients are plotted together with error bars of 2 standard errors. These error bars are comparable to the standard errors in Fig. 4, except for the cases in which the data sample was atypically small because of the gaps in the observations. For comparison, we also plotted the correlation values obtained directly from the standard Pearson product–moment estimator Eq. (2) as shown in Fig. 6b. In addition, we fitted a functional approximation [Eq. (3)] to the estimated correlation points using the Levenberg–Marquardt algorithm (Press et al. 1988). The algorithm minimizes the sum of squares of differences between data points and the desired model but also takes into account the data uncertainty information which is available only for the transformation-based estimates. As expected, transformation-based correlation estimates are systematically lower than the straightforward Pearson product–moment values. The bias of the latter is on the order of 0.1 over most of the distance range in Fig. 6. We also examined the fitting quality using the sum of squares of differences between the analytical function and the corresponding estimated coefficients. This examination revealed that the scatter in the case of the transformation-based estimator is relatively less than that of the Pearson's estimates. It is obvious now that the estimated uncertainty bounds can explain the amount of scatter associated with the correlation estimates. These effects are significant and, without the bias correction, any further correlation-based analysis, such as rain gauge representativeness errors and the area-point differences, are systematically underestimated.

To complete our illustration, we comment on the possible effect of autocorrelation in the used samples. The mixed bivariate model we used in the simulations ignored the serial correlation that may exist in the data. To address this issue, we computed temporal autocorrelations for different time lags, and we noticed low

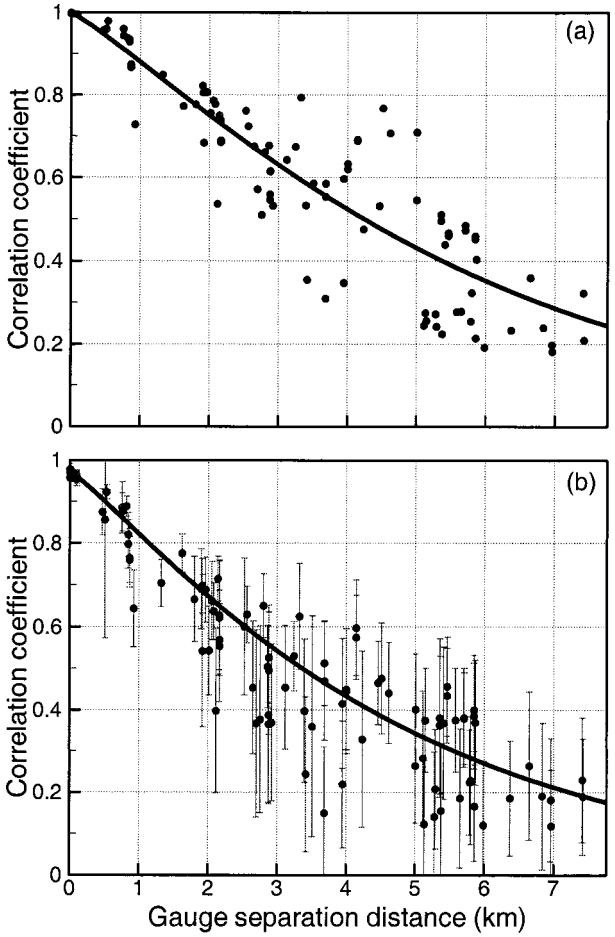

FIG. 6. Scatterplot of estimated correlation coefficients as function of gauge separation distances based on 15-min data from the TEFLUN-B dense network. (a) Correlations computed using the Pearson product–moment formula. (b) Corresponding results using the transformation-based estimator. Two–standard deviation error bar is shown for the transformation-based estimates. Exponential fits for the estimates also are presented.

levels of autocorrelation. The autocorrelation value at one time step of 15-min lag is about 0.5. It becomes negligible (~0.15) for 30-min lag. To address the effect of the serial correlation, we performed a simple experiment in which subsamples were constructed by skipping every other pair of observation in the original time series of the gauges. By doing this, we remove any significant levels of autocorrelation in the samples. We recomputed the correlation coefficients using the transformation-based estimator applied to the constructed subsamples and compared them with the full sample results. The results indicated that the recomputed coefficients show comparable levels to the previously obtained values using the full sample. This result indicates that serial correlation has negligible effect on the reported estimates. We refer the reader to further analysis in Stedinger (1981) on how to deal with this issue in data that have nonnegligible autocorrelation levels.

6. Summary and discussion

In this study, we examined the problems of estimation of correlation coefficients in rainfall data. Researchers need to pay more attention to estimation problems such as effects of sample size, departures from normal distribution, and sensitivity to extreme rain rates. The large deviation of rainfall data from the normal distribution invalidates the usually made assumption of normal distribution of sample correlation. We performed simulation analyses based on bivariate mixed-lognormal distributions to examine the distribution of the sample correlation in conditions closer to reality. The results indicate that the sample correlation coefficient suffers from large bias and estimation variance, which increase with the increase of the skewness of the distribution. We also found that the bias and variance maintain high values, even with large sample sizes not typically available in most archives. We examined the performance of an alternative transformation-based estimator suggested earlier by Stedinger (1981) for streamflows. This estimator provided unbiased and more stable estimates of the correlation coefficient. We illustrated an application of this alternative estimation method to 2-month observations from a high-density rain gauge network deployed in Florida. In the future, it should be worthwhile to examine estimation procedures applied to longer periods of rainfall observation not currently available. Other factors, such as different rainfall regimes and seasonal effects, need to be addressed so that these effects of nonstationarity can be distinguished from the sampling- and distribution-related variability.

There are several other relevant issues with regard to the alternative estimator. First, we consider the procedure we proposed merely as an example of approaches that are needed for improved estimation of correlation over the traditional Pearson's formula. We do not advocate the above procedure as a full remedy of the problem. A major assumption—supported to some degree by the literature and statistical testing—was the underlying bivariate model for two-station rainfall observations. However, other bivariate distributions may offer better fit and/or result in better estimation properties of the sought correlation. For example, the described two-parameter lognormal transformation estimator can be easily formulated to suit a three-parameter case (Stedinger 1981). Another candidate bivariate model (Moran 1969) may be based on the gamma distribution that is known to provide good fit to rainfall observations. Furthermore, testing the validity of the assumed distribution is always required before applying the proposed estimator to a specific rainfall dataset. The robustness of the proposed estimator with respect to the model choice is another important issue.

For some applications it may be appropriate to use nonparametric correlation methods, such as Kendall's τ and Spearman's rank correlations (e.g., Conover 1971; Kendall 1970; Gibbons and Chakraborti 1992). Similar

to other nonparametric statistics, rank correlation is free from distributional assumptions and is less subject to some of the limitations we discussed earlier. However, we did not consider this class of estimators given that our focus is on the population correlation coefficient as defined in Eq. (1).

Acknowledgments. This study was partially supported by NASA Grant NAG5-9664, NOAA Grant NA16GP1013, and Oklahoma NASA EPSCoR NCC 5-171. We are grateful to all the participants of the NASA TEFLUN-B experiment who assisted with the data collection and to Anton Kruger at IIHR—Hydroscience & Engineering for his helpful input at various stages of this study.

REFERENCES

Anagnostou, E. N., W. F. Krajewski, and J. Smith, 1999: Uncertainty quantification of mean-areal radar-rainfall estimates. *J. Atmos. Oceanic Technol.,* **16,** 206–215.

Berndtsson, R., 1987: On the use of the cross-correlation analysis in studies of patterns of rainfall variability. *J. Hydrol.,* **93,** 113–134.

Bras, R. L., and I. Rodriguez-Iturbe, 1993: *Random Functions and Hydrology.* Dover, 559 pp.

Ciach, G. J., and W. F. Krajewski, 1999: Radar–rain gauge comparisons under observational uncertainties. *J. Appl. Meteor.,* **38,** 1519–1525.

Conover, W. J., 1971: *Practical Nonparametric Statistics.* John Wiley and Sons, 462 pp.

Crane, R. K., 1990: Space–time structure of rain rate fields. *J. Geophys. Res.,* **95,** 2011–2020.

Cressie, N. A. C., 1993: *Statistics for Spatial Data.* John Wiley and Sons, 900 pp.

Fisher, R. A., 1921: On the ''probable error'' of a coefficient of correlation deduced from a small sample. *Metron,* **1** (4), 3–32.

Gibbons, J. D., and S. Chakraborti, 1992: *Nonparametric Statistical Inference.* 3d ed. Marcel Dekker, 544 pp.

Gunst, R., 1995: Estimating spatial correlation from spatial–temporal meteorological data. *J. Climate,* **8,** 2454–2470.

Habib, E., W. F. Krajewski, and A. Kruger, 2001: Sampling errors of tipping-bucket rain gauge measurements. *J. Hydrol. Eng.,* **6,** 159–166.

Houze, R. A., and C.-P. Cheng, 1977: Radar characteristics of tropical convection observed during GATE: Mean properties and trends over the summer season. *Mon. Wea. Rev.,* **105,** 964–980.

Huff, F. A., 1970: Spatial distribution of rainfall rates. *Water Resour. Res.,* **6,** 255–259.

——, and W. L. Shipp, 1969: Spatial correlations of storm, monthly, and seasonal precipitation. *J. Appl. Meteor.,* **8,** 542–550.

Humphrey, M. D., J. D. Istok, J. Y. Lee, J. A. Hevesi, and A. L. Flint, 1997: A new method for automated dynamic calibration of tipping-bucket rain gauges. *J. Atmos. Oceanic Technol.,* **14,** 1513–1519.

Hutchinson, T. P., 1997: A comment on correlation in skewed distributions. *J. Gen. Psychol.,* **124,** 211–215.

Johnson, N. L., S. Kotz, and N. Blakrishnan, 1995: *Continuous Univariate Distributions.* 2d ed. *Probability and Mathematical Statistics—Applied Probability and Statistics,* Vol. 2, John Wiley and Sons, 752 pp.

Kedem, B., and L. S. Chiu, 1987: On the lognormality of rain rate. *Proc. Natl. Acad. Sci.,* **84,** 901–905.

——, ——, and G. R. North, 1990: Estimation of mean rain rate: Application to satellite observations. *J. Geophys. Res.,* **95,** 1965–1972.

Kendall, M., 1970: *Rank Correlation Methods.* 4th ed. Charles Griffin, 202 pp.

Kessler, E., and B. Neas, 1994: On correlation, with applications to the radar and raingage measurement of rainfall. *Atmos. Res.,* **34,** 217–229.

Kitchen, M., and R. M. Blackall, 1992: Representativeness errors in comparisons between radar and gauge measurements of rainfall. *J. Hydrol.,* **134,** 13–33.

Kowalski, C. J., 1972: On the effect of non-normality on the distribution of the sample product-moment correlation coefficient. *Appl. Stat.,* **27,** 1–12.

Krajewski, W. F., A. Kruger, and V. Nespor, 1998: Experimental and numerical studies of small-scale rainfall measurements and variability. *Water Sci. Technol.,* **37,** 131–138.

——, G. J. Ciach, J. R. McCollum, and C. Bacotiu, 2000: Initial validation of the Global Precipitation Climatology Project over the United States. *J. Appl. Meteor.,* **39,** 1071–1086.

Lai, C. D., J. C. W. Rayner, and T. P. Hutchinson, 1999: Robustness of the sample correlation—the bivariate lognormal case. *J. Appl. Math. Decis. Sci.,* **3,** 7–19.

Legates, D. R., and R. E. Davis, 1997: The continuing search for an anthropogenic climate change signal: Limitations of correlation-based approaches. *Geophys. Res. Lett.,* **24,** 2319–2322.

——, and G. J. McCabe Jr., 1999: Evaluating the use of ''goodness-of-fit'' measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.,* **35,** 233–241.

May, D. R., and Julien, P. Y. 1998: Eulerian and Lagrangian correlation structures of convective rainstorms. *Water Resour. Res.,* **34,** 2671–2683.

McDonald, J. E., 1960: Remarks on correlation methods in geophysics. *Tellus,* **12,** 176–183.

Moore, D. S., 1991: *Statistics: Concepts and Controversies.* 3d ed. Freeman, 439 pp.

Moran, P. A. P., 1969: Statistical inference with bivariate gamma distributions. *Biometrika,* **56,** 627–634.

Morrissey, M. L., J. A. Maliekal, J. S. Greene, and J. Wang, 1995: The uncertainty in simple spatial averages using raingage networks. *Water Resour. Res.,* **31,** 2011–2017.

Moszkowicz, S., 2000: Small-scale structure of rain field—preliminary results basing on a digital gauge network and on MRL-5 Legionowo Radar. *Phys. Chem. Earth,* **25B,** 933–938.

Nakagawa, S., and N. Niki, 1992: Distribution of the sample correlation coefficient for nonnormal populations. *J. Japan Soc. Comput. Stat.,* **5,** 1–19.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, 1988: *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, 818 pp.

Sharon, D., 1972: The spottiness of rainfall in a desert area. *J. Hydrol.,* **17,** 161–175.

Shimizu, K., 1993: A bivariate mixed lognormal distribution with an analysis of rainfall data. *J. Appl. Meteor.,* **32,** 161–171.

Simpson, J., C. Kummerow, W.-K. Tao, and R. F. Adler, 1996: On the Tropical Rainfall Measuring Mission (TRMM). *Meteor. Atmos. Phys.,* **60,** 19–36.

Smith, J. A., and W. F. Krajewski, 1991: Estimation of the mean field bias of radar rainfall estimates. *J. Appl. Meteor.,* **30,** 397–412.

Stedinger, J. R., 1981: Estimating correlations in multivariate streamflow models. *Water Resour. Res.,* **17,** 200–208.

Stuart, A., and J. K. Ord, 1987: *Kendall's Advanced Theory of Statistics.* Vol. 1. 5 ed. Charles Griffin, 676 pp.

Willmott, C. J., 1984: On the evaluation of model performance in physical geography. *Spatial Statistics and Models,* G. L. Gaile and C. J. Willmott, Eds., D. Reidel, 443–460.

Young, C. B., A. A. Bradley, W. F. Krajewski, A. Kruger, and M. L. Morrissey, 2000: An evaluation of NEXRAD multisensor precipitation estimates for operational hydrologic forecasting. *J. Hydrometeor.,* **1,** 241–254.

Zawadzki, I., 1973: Statistical properties of precipitation patterns. *J. Appl. Meteor.,* **12,** 459–472.