Discriminatory Power of Heterogeneity Statistics with Respect to Error of Precipitation Quantile Estimation

Michael J. Wright, Ph.D.¹; Mark. H. Houck, Ph.D., P.E., F.ASCE²; and Celso M. Ferreira, Ph.D., M.ASCE³

Abstract: At low sample size, sampling error may be reduced by pooling multiple gauge records. This creates an error component due to heterogeneity, the degree to which the pooled regional data’s quantile estimates are different from the true at-site quantiles. Heterogeneity statistics attempt to quantify the degree to which error is added due to regional heterogeneity. They are justified through elucidation of a so-called reasonable proxy relationship with error caused by heterogeneity and through the ability of heterogeneity thresholds to detect heterogeneous regions. In this paper, previous findings regarding three heterogeneity statistics \( H_1 - H_3 \) are revisited; a previous finding that \( H_1 \) is superior to \( H_2 \) and \( H_1 \) is amended based on simulation experiments and upon enumeration of all possible regionalizations of a small gauge dataset across time scales from daily to monthly. Thresholds defined based on \( H_1 \) are shown to be \( 4 \times \) too high for application to \( H_2 \) and new thresholds are derived for \( H_3 \). Two nonparametric heterogeneity statistics are tested and found to achieve only the unsatisfactory performance level of \( H_3 \). DOI: 10.1061/(ASCE)HE.1943-5584.0001172. © 2015 American Society of Civil Engineers.

Introduction

A large sample size is required for accurate estimation of the likelihood of extreme precipitation events. However, long precipitation data records are often unavailable, especially in developing nations and sparsely populated locales. Pooling the sample size offered by multiple precipitation gauges can reduce the error of quantile estimation associated with sample size but only at the cost of introducing a new component of error. The degree to which a group or region of gauges or sites violates the hypothesis of identical underlying probability distributions, or homogeneity, is termed heterogeneity. Quantile error associated with heterogeneity must be balanced against the reduction in quantile error that is induced by the pooling of multiple gauge records.

Heterogeneity statistics have been established in the literature based on their ability to detect heterogeneous regions (Viglione et al. 2007) and the degree to which the value of the heterogeneity statistic acts as a so-called reasonable proxy for the magnitude of quantile error added due to heterogeneity (Hosking and Wallis 1997). These studies have established a number of heterogeneity statistics and offered tests of statistical power based on the performance of thresholds at distinguishing heterogeneous from homogeneous regions, but results for the reasonable proxy analyses of several statistics have not been presented. Because the relationship between a heterogeneity statistic and quantile error added due to heterogeneity was used by Hosking and Wallis (1997) to define thresholds, this relationship can also be characterized for other statistics to determine the reasonableness of thresholds that are assigned.

This regionalization method allows analysis of point-based measurements of rainfall to take spatial relationships into account. Other methods of modeling spatial extremes exist in the literature, including kriging and Theissen polygons. Max-stable process models incorporate spatial and temporal dependency relationships to represent areal rainfall as the summation of many independent processes, i.e., storms (Padoan et al. 2009).

In the research reported in this paper, a simulation method previously used in the literature to quantify the reasonable proxy relationship for one heterogeneity statistic is extended to others. Additionally, a novel enumeration method is applied to a small Minnesota daily precipitation gauge dataset. All possible regionalizations of the gauges for data aggregated at time steps from daily to monthly are evaluated through the method of linear moments. Estimates of quantile error are compared to heterogeneity statistics in analogous fashion to the simulation experiment. Because heterogeneity-error relationships are linear in both cases, the Pearson’s \( r \) statistic is a reasonable measure of the utility of a heterogeneity statistic as a proxy for quantile error. Guidance is presented for precipitation frequency analysts regarding the relative utility of the statistics considered in this paper and heterogeneity thresholds are estimated for statistics with strong linear relationships to error.

Data

Twelve daily precipitation gauges with more than 1,000 days of record and no missing months were selected from a Minnesota high-density rain gauge network. The Minnesota State Climatology Office maintains hundreds of long-record volunteer daily precipitation gauges in a statewide high-density precipitation gauge network. A quality-controlled subset of these gauges was included in the dataset for Perica et al. (2013). Maps displaying results for all states currently covered by Perica et al. (2013), including Minnesota, can be accessed at http://hdsc.nws.noaa.gov/hdsc/pfds/pfds_maps.html.

The Minnesota State Climatology Office prepared data from 341 gauges for submission to the National Oceanic and
Atmospheric Administration (NOAA). Of these, 57 gauges which did not skip 1 full month between the start and end of the record were further considered and 12 clustered in the Minneapolis-St.Paul region (i.e., the Twin Cities region) were selected (Fig. 1). The longest record length of the 12 sites is 3,751 days of nonzero precipitation and the shortest is 1,134 days. The wet-day (nonzero) precipitation records of this group are used as the basis for the enumeration experiment.

**Analysis**

Regional frequency analysis, in which multiple gauges’ data are normalized by the at-site mean or median and a unitless regional quantile function is parameterized by the pooled data, has been used in hydrology since Dalrymple (1960). At that time the conventional moments of the dataset were used to parameterize probability distributions, which in turn output quantile estimates. However, a set of polynomial statistics imitating the conventional moments called linear or L-moments have been shown to possess superior properties for the analysis of hydrological data. Their sample estimates have lower bias, are more robust to outliers than conventional sample moments, and are not bounded by sample size (Hosking et al. 1985; Lettenmaier et al. 1987; Hosking 1990; Vogel and Fennessey 1993). The Hosking and Wallis (1997) regional frequency analysis framework, referred to in this paper as regional frequency analysis using linear moments (RFA-LM), uses probability distributions parameterized by the L-moments of real and simulated data to generate synthetic data through Monte Carlo simulation. This method is applied to generate regional quantile estimates and to approximate their error as well as for the calculation of heterogeneity statistics.

The statistical language R (R Core Team 2012) was used to run the calculations in the research reported in this paper. For RFA-LM calculations and calculation of the Hosking-Wallis heterogeneity statistics $H_1$, $H_2$, and $H_3$, the lmomRFA package was used. lmomRFA documentation is available online at http://cran.r-project.org/web/packages/lmomRFA/lmomRFA.pdf (Hosking 2014). The package applies FORTRAN code available at http://lib.stat.cmu.edu/general/lmoments. For the nonparametric bootstrap Anderson-Darling (AD) and Durbin-Knot (DK) test statistics the package termed homtest was used; homtest, which contains functions for the calculation homogeneity tests described in Viglione et al. (2007), was used to calculate nonparametric heterogeneity statistics. Documentation for the homtest package is available at http://cran.r-project.org/web/packages/homtest/homtest.pdf.

**Linear Moments**

Linear moments are derived from the probability-weighted moments (PWMs) of Greenwood et al. (1979). Hosking and Wallis (1997) define the $(r + 1)$th L-moment $\lambda$ of a quantile function $x(u)$ as the product of a polynomial $p_{r,k}^{*}$ and the probability-weighted moment $\beta_k$ in Eqs. (1)–(3)

$$\lambda_{r+1} = \sum_{k=0}^{r} p_{r,k}^{*}\beta_k$$  \hspace{1cm} (1)

$$p_{r,k}^{*} = \frac{(-1)^{-k}(r+k)!}{(k!)^2(r-k)!}$$  \hspace{1cm} (2)

$$\beta_k = \int_0^1 x(u)u^k\,du$$  \hspace{1cm} (3)

The sample estimate of $\beta_k$ is $b_r$, estimated using Eq. (4), where $n$ is the number of data points in the record, and $x_j$ is the $j$th data point ordered from smallest to largest

$$b_r = n^{-1} \sum_{j=1}^{n} \frac{(j-1)(j-2)\ldots(j-r)}{(n-1)(n-2)\ldots(n-r)} x_j$$  \hspace{1cm} (4)

The value of $\lambda_1$ is calculated identically to the arithmetic mean, $\lambda_2$ measures variance, $\lambda_3$ measures skewness, and $\lambda_4$ measures kurtosis. L-moment ratios are often used to represent the second through fourth moments. The coefficient of L-variation (L-CV) is $\tau = \lambda_2/\lambda_1$, while L-skewness $\tau_3$ and L-kurtosis $\tau_4$ are calculated using Eq. (5)

$$\tau_r = \lambda_r/\lambda_2$$  \hspace{1cm} (5)

The sample estimate of $\lambda_r$ is $l_r$, the sample estimate of the L-CV is $t_r$, and the sample estimate of $\tau_r$ is $t_r$. These are calculated using Eqs. (6) and (7)

$$l_{r+1} = \sum_{k=0}^{r} p_{r,k}^{*}b_k$$  \hspace{1cm} (6)

$$t_r = l_r/l_2$$  \hspace{1cm} (7)

**Homogeneity Statistics**

Because the L-moments analytical framework offers parameterizations for probability distributions, Monte Carlo statistics can be

---

Fig. 1. Map of Minnesota gauge network: (a) full state map; (b) Twin Cities region

© ASCE 04015011-2 J. Hydrol. Eng.
formulated that quantify the concepts of heterogeneity and quantile error. In addition, statistics can be formulated from a nonparametric standpoint in which nothing is assumed about the shape of the dataset. Like the L-moment statistics the process begins with the sorting of the data record from low to high, but this information is used not to parameterize a probability distribution from which simulated data will be drawn but to make estimates directly, using the original data.

Nonparametric heterogeneity statistics compare the empirical distribution of each gauge’s data record to the distribution of the pooled dataset, while Monte Carlo heterogeneity statistics generate many simulated regions from a flexible probability distribution and make comparisons between observed and simulated data. Nonparametric statistics avoid assumptions about the data at the cost of assuming the data record contains all possible future events, while Monte Carlo statistics make distributional assumptions in return for sample size limited not by the data record but by computational resources. In addition, if the distributional assumption has validity, the simulated regions are likely to contain rare events that did not appear in the original dataset but have relevance to quantile estimates for future events.

Homogeneity statistics quantify the degree to which the homogeneity assumption is violated for a candidate regionalization of gauge data. If all sites have identical probability distributions (after normalization) the regional pooling step serves to increase sample size with no quantile error added. If only minor differences exist between gauges, the increase in sample size can still decrease quantile error due to the pooled dataset’s inability to represent faithfully each individual gauge’s probability distribution. The numerical value of a heterogeneity statistic should impart information on the magnitude of quantile error due to heterogeneity, which in turn allows thresholds to be assigned for heterogeneity statistics with reference to the equivalent heterogeneity-associated error.

**Hosking-Wallis Statistics**

The statistics proposed by Hosking and Wallis (1997), as part of the RFA-LM methodology, have achieved wide use in hydrology. These heterogeneity statistics are referred to in this paper as $H_1$, $H_2$, and $H_3$ individually, and collectively as the Hosking-Wallis statistics.

Quantiles of annual maximum precipitation for events of a known duration (e.g., every 5 min, daily, or weekly) have been estimated using RFA-LM methods, including the use of the Hosking-Wallis heterogeneity statistics (Huff and Angel 1992; Adamowski et al. 1996; Bradley 1998; Allia 1999; Smithers and Schulze 2001; Kysely et al. 2007; Kysely and Pícek 2007; Norbiato et al. 2007; Szolgay et al. 2009; Um et al. 2010; Yang et al. 2010; Nggongondo et al. 2011; Gabriele and Chiaramonti 2013). The RFA-LM has also been used for quantile estimates of monthly (Núñez et al. 2011) and annual (Guttman et al. 1993; Lin and Chen 2006; Parida and Moulafli 2008; Modarres and Sarhadi 2011; Dikbas et al. 2012) precipitation totals. The RFA-LM has also been widely applied to stream flow data, particularly for quantile estimation of annual maximum floods (Zrini and Burn 1994; Burn and Goel 2000; Kjeldsen et al. 2002; Jingyi and Hall 2004; Abida and Ellouze 2006; Atem and Harmançioğlu 2006; Rao and Srinivas 2006; Srinivas et al. 2008; Noto and La Loggia 2009; Gaume et al. 2010; Guse 2010; Saf 2010; Hussain 2011; Kar et al. 2011; Rianna et al. 2012; Seckin et al. 2013).

A wide variety of hydrological variables have been subjected to RFA-LM analysis in an attempt to reduce quantile error of estimation due to low sample size. The Hosking-Wallis statistics have been used to defend regionalizations of partial duration series (Pham et al. 2013) and annual minimum flow over a weekly duration (Modarres 2008; Dodangeh et al. 2013) for stream flow data as well as quantile estimates for maximum daily precipitation immediately after a drought (Feng et al. 2013). The standardized precipitation index (SPI; Santos et al. 2011) and gridded precipitation data (Marx and Kinter 2007; Satyanarayana and Srinivas 2009) have also been subjected to RFA-LM analysis.

Hosking and Wallis (1997) estimate heterogeneity by using the candidate region’s data to parameterize the more flexible four-parameter Kappa distribution, from which simulated regions are drawn using Monte Carlo sampling. For the real region and all simulated regions a statistic $V$ can be calculated as the sum of the squared deviations from the mean across all sites for a given L-moment ratio or ratios. The mean and standard deviation of $V$ across all simulated regions is calculated and compared with $V$ for the true data to calculate $H$, a heterogeneity statistic.

Three formulations of $V$ are proposed by Hosking and Wallis (1997), as follows: (1) using only the L-CV ($t_1$), (2) using the L-CV and L-skewness ($t_3$), and (3) using the L-skewness and L-kurtosis ($t_4$). These statistics are denoted as $V_1$, $V_2$, and $V_3$, respectively, in Eqs. (8)–(10). The $R$ superscript indicates the regional average, and $i$ is the index of the $N$ sites in the region

$$V_1 = \sqrt{\frac{\sum_{i=1}^{N} n_i (t_i^R - \mu_R)^2}{\sum_{i=1}^{N} n_i}}$$

$$V_2 = \sqrt{\frac{\sum_{i=1}^{N} n_i (t_i^R - \mu_R)^2 + (t_3^R - \mu_3)^2}{\sum_{i=1}^{N} n_i}}$$

$$V_3 = \sqrt{\frac{\sum_{i=1}^{N} n_i (t_i^R - \mu_R)^2 + (t_3^R - \mu_3)^2 + (t_4^R - \mu_4)^2}{\sum_{i=1}^{N} n_i}}$$

In Eq. (11), $\mu_R = \text{mean}$; and $\sigma_R = \text{SD}$ of $V$ across all simulated regions. $H_1$ is calculated using $V_1$, $H_2$ using $V_2$, and $H_3$ using $V_3$

$$H = \frac{V - \mu_R}{\sigma_R}$$

**Nonparametric Rank-Order Statistics**

Sorting a dataset from lowest to highest value is the first step of both the moment-based methods previously described and of nonparametric methods in which only observed data are used. Viglione et al. (2007) expresses the Anderson-Darling and Durbin-Knot tests as estimators of regional heterogeneity. The degree of similarity between the empirical distributions of the pooled regional data and each constituent site is established and averaged into a regional estimate of heterogeneity.

If $k$ sites, whose index is $i$, with $n_i$ data points each, are pooled into the ordered sample $Z_1 < \ldots < Z_N$, where $N$ is the sum of $n_i$ at all sites $i$, the $k$-sample Anderson-Darling test statistic, can be calculated using Eq. (12)

$$AD = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i-1} \frac{(NM_{ij} - jn_i)^2}{jn(N-j)}$$

© ASCE 04015011-3 J. Hydrol. Eng.
where \( M_{ij} \) = number of data points at site \( i \) that are less than or equal to \( Z_j \). A nonparametric bootstrap approach determines percentage points for the test. Sampling with replacement from the pooled sample to create \( k \) artificial sites with \( n_i \) data points each, normalizing each site by its mean or median, and calculating AD for the synthetic region gives \( N_{\text{syn}} \) values of AD which can be ranked low to high. A threshold equivalent to a 5% probability of heterogeneity is extrapolated from the 95th percentile of the empirical distribution of AD values, which serves as an approximation of the distribution of AD under the null hypothesis of homogeneity. The true region is accepted as homogeneous if it has a lower AD score than at least 95% of a population of synthetic regions formed by sampling with replacement from the pooled regional data.

The Durbin-Knott test quantifies the heterogeneity amongst a candidate region’s dispersion or variance, analogously to \( H_1 \), but it is a rank test. A measure \( D_i \) can be calculated at each of \( k \) sites in which the empirical distribution function of the pooled regional data \( H_N(x) \) is evaluated at each data point \( j \) of \( n_i \), the length of the site’s record. The equation for \( D_i \) is shown in Eq. (13)

\[
D_i = \sum_{j=1}^{n_i} \cos[2\pi H_N(X_j)]
\]

Because \( D_i \) is normal under the hypothesis of homogeneity, a statistic \( DK \) defined by Eq. (14) has a chi-square distribution with \( k-1 \) degrees of freedom, allowing a threshold value to be determined for the 5% significance level. If the region’s DK is above the threshold, the region is considered heterogeneous. No simulations are required in the calculation of DK.

\[
DK = \sum_{i=1}^{k} D_i^2
\]

**RMS Error of Regional Quantile Estimates**

As part of their explication of the RFA-LM method, Hosking and Wallis (1997) introduced a simulation procedure for quantifying the error of quantile estimation. While the simulation experiment defending the \( H \) statistic sampled from a distribution parameterized by known L-moment ratios to obtain simulated quantile estimates, the L-moment ratios of real data represent sample estimates and incorporate sampling error. If the simulated regions were generated from the observed sample L-moment ratios they would receive a double dose of sampling error, so the range of L-moment ratio variation within the candidate region must be shrunk so that the simulated regions will have a similar range of L-moment variation to that seen in the original data.

**Table 1. Correlation between 12 Sites in Dataset for the 1-Day Time Step**

<table>
<thead>
<tr>
<th>Site</th>
<th>11</th>
<th>35</th>
<th>39</th>
<th>46</th>
<th>78</th>
<th>104</th>
<th>149</th>
<th>150</th>
<th>266</th>
<th>268</th>
<th>272</th>
<th>328</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0.7</td>
<td>0.5</td>
<td>0.7</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>35</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>0.8</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>39</td>
<td>-</td>
<td>1</td>
<td>0.5</td>
<td>0.7</td>
<td>0.7</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>46</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0.6</td>
<td>0.7</td>
<td>0.7</td>
<td>0.8</td>
<td>0.7</td>
<td>0.8</td>
<td>0.7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>78</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>104</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
<td>0.7</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>149</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0.9</td>
<td>0.7</td>
<td>0.7</td>
<td>0.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>150</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0.8</td>
<td>0.8</td>
<td>0.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>266</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0.7</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>268</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0.7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>272</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>328</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table 2. Simulation Study**

<table>
<thead>
<tr>
<th>Region type</th>
<th>Average</th>
<th>Range</th>
<th>Number of sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hom ( n = 30 )</td>
<td>0.2</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Hom ( n = 60 )</td>
<td>0.2</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Het ( 30% \ n = 30 )</td>
<td>0.2</td>
<td>0.06</td>
<td>11</td>
</tr>
<tr>
<td>Het ( 50% \ n = 30 )</td>
<td>0.2</td>
<td>0.1</td>
<td>6</td>
</tr>
<tr>
<td>Het ( 30% \ n = 60 )</td>
<td>0.2</td>
<td>0.06</td>
<td>6</td>
</tr>
<tr>
<td>Het ( 50% \ n = 60 )</td>
<td>0.2</td>
<td>0.1</td>
<td>11</td>
</tr>
<tr>
<td>Het ( 30%, \ Region A )</td>
<td>0.2</td>
<td>0.06</td>
<td>21</td>
</tr>
<tr>
<td>Het ( 30%, \ Region B )</td>
<td>0.2</td>
<td>0.06</td>
<td>21</td>
</tr>
<tr>
<td>Het ( 30%, \ Region C )</td>
<td>0.2</td>
<td>0.06</td>
<td>21</td>
</tr>
<tr>
<td>Het ( 30%, \ Region D )</td>
<td>0.2</td>
<td>0.06</td>
<td>21</td>
</tr>
<tr>
<td>Hom ( n = 30 )</td>
<td>0.2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Bimodal ( 20% \ n = 30 )</td>
<td>0.2</td>
<td>0.04</td>
<td>2</td>
</tr>
<tr>
<td>Bimodal ( 30% \ n = 30 )</td>
<td>0.2</td>
<td>0.06</td>
<td>2</td>
</tr>
<tr>
<td>Bimodal ( 50% \ n = 30 )</td>
<td>0.2</td>
<td>0.1</td>
<td>2</td>
</tr>
</tbody>
</table>

**Shrinkage Estimators**

If a dataset with a given range of variation in L-moment ratios like L-CV, L-skewness, and L-kurtosis is used to parameterize a probability distribution, simulated data outputted by the distribution will generally have a wider range than the original data. This is due to the randomizing effect of sampling error. Therefore, shrinking the degree of variation in these L-moment ratios toward the regional mean to some appropriate level will result in the generation of data which due to sampling error regains the range of variation seen in the original data. Theoretical support in the general multivariate case can be found in Stein (1956), while Bayesian logic supporting shrinkage estimators is offered by Lindley and Smith (1972).

Shrinkage estimators are applied to the error estimation routine by running preliminary simulations in which various fractional multipliers are applied to the range of variation about the regional average in L-CV, L-skewness, and L-kurtosis. The multiplier resulting in an \( H_1 \) that is closest to the \( H_1 \) of the original observed data is...
taken as the appropriate degree of shrinkage so that added sampling error restores the appropriate range of variation. This shrinkage multiplier is used to create simulated data for the error estimation routine.

**Goodness of Fit for a Probability Distribution**

The Hosking and Wallis (1997) Z statistic is a Monte Carlo statistic measuring goodness of fit for a 3-parameter distribution. Monte Carlo simulation using a given distribution parameterized by the mean, L-CV, and L-skewness of the regional data generates the mean, L-CV, and L-skewness of the regional data. The bias $B_4$ and the SD $\sigma_4$ of L-kurtosis, used in calculating Z, are calculated in Eqs. (15) and (16). For a 3-parameter distribution (DIST) with a fixed L-kurtosis after parameterization, $Z_{DIST}$ is calculated using the at-site L-kurtosis of site $m$, $t_4^R(m)$, and regional average L-kurtosis $t_4^R$ in Eq. (17).

$$B_4 = N_{sim}^{-1} \sum_{m=1}^{N_{sim}} (t_4^R(m) - t_4^R)$$

$$\sigma_4 = \sqrt{(N_{sim} - 1)^{-1} \sum_{m=1}^{N_{sim}} (t_4^R(m) - t_4^R)^2 - N_{sim} B_4^2}$$

$$Z_{DIST} = \frac{t_4^DIST - t_4^R + B_4}{\sigma_4}$$

Hosking and Wallis (1997) list formulas providing $t_4^DIST$ for five 3-parameter distributions, the generalized logistic, generalized Pareto, generalized extreme-value, Pearson Type 3, and lognormal, in Table A.3.

**Monte Carlo Estimation of Quantile Error**

The Monte Carlo process by which simulated data are generated, described previously for the Hosking-Wallis Monte Carlo heterogeneity statistics, can be modified to incorporate intersite correlations, allowing for simulations with greater fidelity to the original data at the cost of increased computational time. Data fitting a multivariate normal distribution are generated. Next a shrinkage multiplier is applied to the sample L-moment ratios of the real data to counteract the effects of sampling error. Finally these shrunken L-moment ratios are used to parameterize a distribution, which transforms the correlated normal at-site data so that the simulated region has the so-called shape of the appropriate probability distribution while maintaining correlation between sites. Many Monte Carlo iterations of this simulation are performed, quantile estimates are calculated for each iteration, and the results are averaged to obtain a Monte Carlo estimate of relative RMS error.

The first step in the process requires a matrix containing information on intersite correlations, which is calculated from the sample data as described next. Intersite correlation $r_{ij}$ between each pair of sites $i$ and $j$ in the dataset is calculated according to Hosking and Wallis (1997) using Eq. (18).

$$r_{ij} = \frac{\sum_k (Q_{ik} - \bar{Q}_i)(Q_{jk} - \bar{Q}_j)}{\sqrt{\sum_k (Q_{ik} - \bar{Q}_i)^2 \sum_k (Q_{jk} - \bar{Q}_j)^2}}$$

where the index for time points at which both gauges $i$ and $j$ have nonzero data is $k$; $n_{ij} = $ number of such time points for the pairing of sites $i$ and $j$; and $Q_{ik} =$ value at site $i$ for the shared time point $k$. A mean value for gauge $i$ across all $k$, $\bar{Q}_i$, is calculated in Eq. (19).

**Fig. 2.** At-site L-skewness and L-kurtosis for different time steps at starting point of one plotted against curves representing L-moment ratios of data outputted by five 3-parameter distributions
The index \( k \) and therefore all of these statistics must be calculated uniquely for each pair of sites

\[
\tilde{Q}_i = n_{ij}^{-1} \sum_k Q_{ik}
\]  

(19)

Eq. (20) calculates the average intersite correlation for the region, \( \bar{r} \), from \( r_{ij} \)

\[
\bar{r} = \left[ \frac{1}{2} N (N - 1) \right]^{-1} \sum_{i \leq N} \sum_{j \neq N} r_{ij}
\]  

(20)

Table 1 shows the correlation at the 1-day time step between all 12 sites. Sites with correlations of zero share no time points with common data; the start date of one is after the end date of the other.

In the lnomRFA package for R, the function regsimq uses \( \bar{r} \) to create a correlation matrix with as many columns and rows as there are sites in the region. The regional average correlation, not the specific intersite correlation \( r_{ij} \), is used as the correlation between all pairs of sites.

The shrinkage estimator procedure described previously is applied to each site’s L-moment ratios, which are moved toward or away from the mean by a multiplier determined through a separate series of Monte Carlo simulations. If the between-site variance in the L-moment ratios added through sampling error is of the same magnitude as the shrinkage effected through use of the multiplier, the simulated data will have a similar \( H_1 \) to the original region. An appropriate multiplier is determined iteratively using multiple simulations. The correlated multivariate normal data are then fed into the quantile function of the probability distribution used to model the region, resulting in data having the appropriate so-called shape

---

**Fig. 3.** Estimated RMS error calculated using the Pearson Type 3 distribution plotted against the following for the 1-day time step at 0.999 non-exceedance frequency: (a) \( H_1 \); (b) \( H_2 \); (c) \( H_3 \); (d) AD; (e) DK

© ASCE 04015011-6 J. Hydrol. Eng.
given the distributional parameters while retaining the appropriate intersite correlations.

The lmomRFA package’s function regsimq requires the user to supply a value for \( r \), a probability distribution, and a set of parameters for that distribution at each site. For consistency, the same distribution is used for all regions at a given starting point and time step. At-site L-moment ratios shrunken from the observed data are used to parameterize the probability distribution.

The function generates multivariate normal correlated data at each simulated gauge, transforming the data using the selected distribution. The data are normalized and the simulated region’s average L-moment ratios are found for each artificial site, which are fitted to the chosen distribution and used to output at-site quantile estimates. Eq. (21) estimates the RMS error of the regional normalized quantile estimate \( \bar{Q}_i \) as a predictor of normalized at-site quantile estimate \( Q_i \) for a list of nonexceedance probabilities \( F \). These estimates are averaged across many simulations.

\[
R^H(F) = N^{-1} \sum_{i=1}^{N} \left\{ M^{-1} \sum_{m=1}^{M} \left( \frac{\bar{Q}_i^{(m)}(F) - Q_i(F)}{Q_i(F)} \right)^2 \right\}^{1/2}
\]  

\( (21) \)

**Pearson’s \( r \) and Linearity**

The establishment of \( H \) as a heterogeneity statistic in Hosking and Wallis (1997) was primarily driven by the relationship found between the statistic and quantile error added due to heterogeneity. This relationship is linear in nature and exhibits tight correlation without a great deal of spread. This characteristic was used to establish that certain threshold values of \( H \) were equivalent to relatively narrow ranges of added error. The phrase reasonable proxy was used to describe the relationship of \( H \) to percent RMS error added due to heterogeneity. However, no quantitative method of assessing the reasonableness of the proxy relationship was presented.

Because the relationship between \( H \) and percent RMS error added due to heterogeneity was linear, the degree to which the one is a reasonable proxy of the other can be assessed by quantifying the linearity of the relationship between the two statistics. A common statistical measure of the strength of a linear relationship is Pearson’s \( r \). In this paper, Pearson’s \( r \) is used to quantify the degree to which a heterogeneity statistic is a reasonable proxy for error. Because only linear relationships with positive slopes represent a valid proxy relationship in this case, \( r^2 \) is not used.

**Aggregation of Data for Enumeration Study**

Daily precipitation data from the Twin Cities region of Minnesota were aggregated into larger time steps for all possible starting points. For example, a time step of 2 days in length has two possible starting points, while a weekly time step has 7 and a monthly 30. Starting points are essentially arbitrary and are defined with reference to the first day of record in the oldest of the 12 gauges. Missing days of record are noted and any time step including a missing day is eliminated from the aggregated record. Otherwise, the daily records are summed to create each aggregated data point.

**Simulation Study**

In Hosking and Wallis (1997) a simulation study is presented in which the error of quantile estimates at regions of varying heterogeneity is divided by error at an equivalent homogeneous region to isolate the component of error due to heterogeneity. Known L-moment ratios for each simulated site are used to find a so-called quantile estimate. The RMS error with reference to the true quantile estimate is calculated for quantile estimates drawn from data generated using a generalized extreme-value distribution fitted to the known L-moment ratios.

The RMS error due to heterogeneity is found to scale linearly with a measure based on \( V_1 \) and heterogeneity thresholds are established through reference to this linear relationship; \( H = 1 \) represents 20–40% RMS error added and \( H = 2 \) represents 40–80% RMS error added. \( H \) statistics based on \( V_2 \) and \( V_3 \) were found to rarely exceed these thresholds and were not recommended as heterogeneity estimators.

Results for \( H_2 \) and \( H_3 \) were not presented in the original study. In the research reported in this paper, the experiment is recapitulated and is additionally applied to the nonparametric heterogeneity statistics. The reasonable proxy relationship with regard to heterogeneity-related error is described for all five heterogeneity statistics. Each heterogeneity statistic/RMS error relationship is evaluated using the degree of linearity as quantified using Pearson’s \( r \). For statistics possessing a clear linear relationship with error, the slope of the relationship is used in analogous fashion to Hosking and Wallis (1997) to propose heterogeneity thresholds.

Table 2 presents the L-moment ratio values used in the simulation study, which can also be seen in Table 4.1 in Hosking and Wallis (1997). For a number of homogeneous regions, equivalent regions with varying degrees of heterogeneity are compared and the ratio between each heterogeneous region’s quantile error and the error of its homogeneous equivalent are calculated. The simulated regions designated in the table have a variety of sample sizes and L-moment ratio ranges, allowing for a fairly wide range of error added due to heterogeneity to be investigated.

Table 2 indicates each region’s average L-CV and the range between the highest and lowest at-site L-CVs. L-skewness is equal to L-CV at all sites, with the exception of the two simulation scenarios where average L-CV is equal to 0.15, its lowest value. The homogeneous, 30% heterogeneous, and 50% heterogeneous regions with average L-CV equal to 0.15 have average L-skewness of 0.1; the

![Fig. 4. Pearson’s \( r \) between estimated RMS error and five heterogeneity statistics at 0.999 nonexceedance frequency for 1–35 day time steps at starting point 1 day](image-url)
30% heterogeneous region has an L-skewness range of 0.09; and the 50% heterogeneous region has a range of 0.15. For regions otherwise unspecified, at-site L-CV and L-skewness varies linearly around the regional mean. Bimodal regions have half of the sites equal to the lower bound of the range and half equal to the upper. Most regions have equal sample size at all sites. The exceptions, which are all composed of 21 sites, are marked as Regions A–D. Sites in Region A have sample sizes of 50, 48, 46, . . ., 10; sites in Region B have sample size 10, 12, 14, . . ., 50; sites in Region C have sample sizes 50, 46, . . ., 14, 10, 14, . . ., 46, 50; and sites in Region D have sample sizes 10, 14, . . ., 46, 50, 46, . . ., 14, 10.

The L-CV and L-skewness of the simulated regions, along with a mean of 1, are inputted as parameters to the generalized extreme-value distribution and one hundred simulated regions are created. The RMS error estimates at the 0.01, 0.1, 0.99, and 0.999 nonexceedance frequencies are calculated, along with the five heterogeneity statistics [i.e., (1) $H_1$, (2) $H_2$, (3) $H_3$, (4) AD, and (5) DK]. The first four statistics are calculated at each of the 100 instantiations of each region using 500 simulations, whereas DK is not a Monte Carlo statistic and so was calculated once for each instantiation. The 100 values of each heterogeneity statistic and RMS error estimate are averaged for each region in Table 2.

**Results**

**Enumeration Study Results**

The relationship between each of five heterogeneity statistics and estimated quantile RMS error is investigated at time steps ranging from daily to monthly. The distribution chosen to model the data is based on the basis of the $Z$ scores for 3-parameter distributions. All regions in simulations with time steps shorter than eleven days record the lowest absolute value of $Z$ score for the Pearson Type 3 distribution. Conversely, all tested time step/starting point combinations for starting points of greater than 18 days in length saw the generalized Pareto record the lowest absolute $Z$ score for all enumerated regions. For time steps of 12–17 days in length some regions fit the generalized Pareto better according to the $Z$ test while others fit the Pearson Type 3 better. For these intermediate time steps the distribution chosen has little effect on the observed

![Fig. 5.](image)

Fig. 5. Percent RMS error added due to heterogeneity for simulated regions plotted against the following at nonexceedance probability of 0.99: (a) $H_1$; (b) $H_2$; (c) $H_3$; (d) AD; (e) DK
relationship between error and heterogeneity statistics, probably because the distributions output very similar L-kurtosis when parameterized with a given L-skewness at these magnitudes. The generalized Pareto is used to model all regions for time steps greater than 13, while all regions for time steps of 13 and less are modeled with the Pearson Type 3 distribution.

Fig. 2 offers a graphical analogue to the results of these Z tests; Fig. 2 also illustrates the similarity between the Pearson Type 3 and generalized Pareto distributions in the sector of the L-skewness/L-kurtosis graph where time steps with regions fitting both distributions are found to plot. The presence of sites from the biweekly, or 14-day, time step near the intersection of the Pearson Type 3 and Generalized Pareto lines offers a graphical rationale for using generalized Pareto at and above the 14-day time step and Pearson Type 3 below the 14-day time step.

While regions whose shrinkage multipliers are below 0.50 do not fall into a linear cluster, perhaps due to the dissimilarity between the heavily shrunken simulation regions and the original data, high-multiplier regions often do, especially for $H_2$ and $H_3$ and to a lesser degree for $H_1$. Examples of H-RMS error plots at a nonexceedance probability of 0.999 are illustrated for the 1-day time step in Fig. 3. Despite the high RMS error, low-heterogeneity points visible in the $H_1$, $H_2$, and $H_3$ plots in Fig. 3, which are found to comprise the majority of regions with shrinkage multipliers below 0.50, a linear relationship with a positive slope between these heterogeneity estimators and RMS error exists for a large population of regions at the 1-day time step. The AD and DK do not exhibit linearity to the degree seen for $H_1$, $H_2$, and $H_3$. Pearson’s $r$ between the five heterogeneity statistics and estimated RMS error at the 0.999 nonexceedance probability is plotted across time steps from daily to 35 days in length at the starting point (1 day) in Fig. 4. Similar patterns are observed for other starting points, while nonexceedance probabilities below 0.95 exhibit low Pearson’s $r$ for all heterogeneity statistics, possibly due to the predominance of nonheterogeneity-related components in the RMS error of quantile estimates. The clustering observed in $H_1$ and $H_2$ plots translates to a higher Pearson’s $r$ than seen for the other heterogeneity statistics due to these statistics’ greater linearity with respect to RMS error. $H_2$ consistently exhibits a less linear relationship with RMS error than $H_1$ or $H_3$, while AD and DK’s nonlinear relationship with RMSE visible in Fig. 3 translates to consistently low Pearson’s $r$ scores.

**Simulation Study Results**

Results confirmed the analysis of Hosking and Wallis (1997) with regard to the linearity of $H_1$ and the slope of its relationship with percent RMS error added due to heterogeneity. This relationship is depicted for the five heterogeneity statistics considered in this paper in Fig. 5 for the nonexceedance frequency 0.99. At nonexceedance frequencies of 0.01, 0.1, 0.99, and 0.999 the heterogeneity statistic/RMS error added relationship was evaluated and Pearson’s $r$ of the relationship was taken to provide a numerical basis for cross comparison between the statistics. $H_1$ and $H_2$ had the most and second-most linear relationships with RMS error added at all four nonexceedance frequencies, as seen in Fig. 6. $H_3$ consistently has approximately one-fourth the magnitude of $H_1$ across all simulations (Fig. 7).

**Conclusions**

It is found that $H_1$ and $H_2$ outperform $H_3$ and the nonparametric statistics AD and DK in both methodologies described previously. Additionally, threshold values used in the literature for $H_1$, $H_2$, and $H_3$ were applicable only to $H_1$. $H_1$ thresholds have been compared to equivalent ranges of RMS error added due to heterogeneity using the $H_1$-RMS error relationship; analogous thresholds can be constructed for $H_2$ using the $H_2$-RMS error relationship. Because a roughly 4:1 relationship between values of $H_1$ and $H_2$ was noted, proposed $H_2$ thresholds are one-fourth the magnitude of $H_1$ thresholds. These findings offer precipitation frequency analysts a basis for utilizing $H_1$, $H_2$, or both in combination to quantify heterogeneity across their dataset. The performance of $H_2$ at detecting heterogeneity should improve with application of the lower thresholds proposed in this paper.
The Hosking-Wallis heterogeneity statistics $H_1$ and $H_2$ have superior performance to $H_3$ and the nonparametric statistics in both experiments. Because the error estimation methods used in this study are L-moment based and utilize Monte Carlo simulation, it is possible that the superior performance of the Hosking-Wallis statistics can be partly explained by the Monte Carlo and L-moment structure they share with the error estimation routine. A nonparametric measure of error such as bootstrapped standard error could be used in order to provide an alternative error measure which may be biased in favor of the nonparametric statistics.

$H_1$ emerges from the research reported in this paper as the most effective proxy of error among the heterogeneity statistics. This applies for total error and for simulation experiments in which error due to heterogeneity is isolated. However, $H_2$ is nearly as effective, and because its values are generally one-fourth the magnitude of those for $H_1$ the thresholds derived with reference to the $H_1$-RMS error relationship severely understate the efficacy of $H_2$ at identifying heterogeneous regions.

The similarity between the $H_1$-RMS error and $H_2$-RMS error relationships, and the presence in the literature of arguments defending threshold values for $H_1$ based on the $H_1$-RMS error relationship, allows for the analogous characterization of threshold values for $H_2$. They are one-fourth the magnitude of the equivalent thresholds for $H_1$; $H_2 < 0.25$ would indicate a homogeneous region, while $H_2 > 0.5$ would indicate heterogeneity. Due to the slightly greater degree of scatter in an $H_2$-RMS error plot as compared to an $H_1$-RMS error plot the range of added RMS error that is considered equivalent to a given $H_2$ value should be wider than the RMS error ranges described as equivalent to $H_1$ in Hosking and Wallis (1997). If $H_1 = 1$ is equivalent to 20–40% added RMS error, $H_2 = 0.25$ can be considered equivalent to 10–50% added RMS error; if $H_1 = 2$ is equivalent to 40–80% added RMS error, $H_2 = 0.50$ can be considered equivalent to 30–90% added RMS error. For both $H_1$ and $H_2$, equivalent RMS errors are approximations drawn from the empirical relationship between each $H$ statistic and heterogeneity-associated RMS error in simulated experiments. Analysts could use these $H_2$ thresholds to determine whether candidate regions are acceptably homogeneous, perhaps in combination with $H_1$.

Because $H_2$ is measured using two L-moment ratios while $H_1$ uses only one, $H_2$ is more robust to violations of the assumption that regional heterogeneity will be expressed relatively evenly across the L-moment ratios. At low sample sizes, the most important use case for regional frequency analysis, there exists the potential for this assumption to be violated purely due to the randomizing effects of sampling error. This potential for robustness must be balanced against effectiveness as a proxy for error; for this reason, $H_3$ is not likely to serve as an effective heterogeneity statistic despite its relatively robust two-L-moment-ratio formulation.

These findings are based on simulation results and data from a small geographical area. Considerations regarding study areas featuring larger spatial and climatic differences, as well as study areas with higher or lower mean precipitation values or markedly different distributions of precipitation across the seasons, could be addressed through the methods defined previously using data collected from such regions. Simulated data with different statistical properties could also be examined, again using the methods described previously.

Acknowledgments

Data was provided by the State Climatology Office, Minnesota Department of Natural Resources, Division of Ecological and Water Resources. The writers thank Jason Giovannettone for assistance. Financial support provided by the George Mason University Presidential Scholarship is gratefully acknowledged.

References


