EVALUATION OF HETEROGENEITY STATISTICS FOR HYDROLOGICAL REGIONAL FREQUENCY ANALYSIS

by

Michael J. Wright
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Civil and Infrastructure Engineering

Committee:

_____	Dr. Mark Houck, Dissertation Co-Director
_____	Dr. Celso Ferreira, Dissertation Co-Director
_____	Dr. Shanjiang Zhu, Committee Member
_____	Dr. Stephen Nash, Committee Member
_____	Dr. Jason Giovannettone, Committee Member
_____	Dr. Deborah Goodings, Department Chair
_____	Dr. Kenneth S. Ball, Dean, Volgenau School of Engineering
Date: _____	Spring Semester 2014 George Mason University Fairfax, VA

Evaluation of Heterogeneity Statistics for Hydrological Regional Frequency Analysis

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Michael J. Wright
Master of Science
George Mason University, 2011

Directors: Mark Houck and Celso Ferreira
Sid and Reva Dewberry Department of Civil and Infrastructure Engineering

Spring Semester 2014
George Mason University
Fairfax, VA

DEDICATION

This dissertation is dedicated to my family.

ACKNOWLEDGEMENTS

I wish to thank my family for their constant support, my co-workers at the Army Corps of Engineers including Jason Giovannettone for introducing me to the subject of this dissertation, and my advisors and all the professors at George Mason University for their mentorship and tutelage. I am greatly thankful to George Mason University for the provision of financial support through the Presidential Fellowship Award.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

ABSTRACT

EVALUATION OF HETEROGENEITY STATISTICS FOR HYDROLOGICAL REGIONAL FREQUENCY ANALYSIS

Michael J. Wright, Ph.D.

George Mason University, 2014

Dissertation Directors: Drs. Mark Houck and Celso Ferreira

When an engineering firm designs a dam, a government assesses its readiness for flood or famine, or a municipal water supply company invests in new infrastructure, lives and major financial investments are at stake. To design a sturdy dam, prepare a sufficient emergency response, size a water treatment plant, or answer many other important questions, estimates of the magnitude and frequency of future precipitation events are often needed. Because the societal and financial costs of erroneous estimates can be extremely high, mechanisms exist to quantify and minimize the error of estimation. Precipitation gauge records are often pooled to reduce the error of quantile estimates by increasing sample size. This introduces a new error component proportional to heterogeneity, the degree to which the constituent gauges diverge from the regional average. Precipitation frequency analysts, especially in data-poor regions, use heterogeneity statistics to evaluate whether a candidate regionalization adds more error

through heterogeneity than it reduces through increased sample size. This dissertation assesses the relationship between error in quantile estimates and five heterogeneity statistics proposed in the literature, offering precipitation analysts quantifications of these statistics' efficacy and making recommendations for their use. All five-or-more-site regionalizations of a twelve-gauge Minnesota dataset are enumerated and Monte Carlo simulation is used to estimate quantile error and the heterogeneity statistics. Linear relationships found between heterogeneity estimators and quantile error are compared to those found in simulation experiments isolating the heterogeneity-related component of quantile error. Two statistics have highly linear relationships to error in both the simulation and enumeration studies. The less linear statistic is more robust to deviation from the hypothesis that regional coefficient of variation and skewness ranges increase in tandem as heterogeneity rises. Novel heterogeneity thresholds are defined for this statistic. This research offers context and validation for a family of popular heterogeneity statistics whose relative utility has previously been unclear. Precipitation analysts using the regional frequency analysis framework to answer questions about precipitation magnitude now have a full reckoning of the utility of these statistics, increasing the level of confidence they can have in the accuracy of their results.

1: INTRODUCTION

The global water cycle operates on a superhuman scale, forming part of the fundamental context within which all of our priorities, goals, hopes and fears must be evaluated. Precipitation extremes often leave destruction and misery in their wake; this is beyond the power of engineering or science to prevent. However, applying the tools of probability theory to the fundamental workings of the water cycle allows us to plan, in some limited fashion, for these potentially catastrophic events and to mitigate their effects through good design and diligent preparation. To achieve these goals it is vital to quantify and minimize the error of statistical estimates.

A quantile is the value below which a given fraction or percentage of a dataset lies. Decision-makers in fields such as agriculture and engineering use quantile estimates of precipitation for long-term planning. How likely is monthly precipitation to fall below a drought threshold? How often does it rain more than ten inches in a day? Estimating magnitude as a function of probability is perhaps most important, as in these examples, at the extremes. However, extreme events by definition rarely appear in the data records, and an event that was equally likely to impact multiple locations will appear overly likely at one location but go unrecorded at all others.

As sample size increases, the data become a better representation of underlying probability. Increasing record length can only be accomplished by waiting, but multiple

gauges can be pooled to create a regional dataset with many times the sample size of its constituents. However, each gauge's reduced error due to increased sample size must be balanced against the error added by assuming homogeneity – identically shaped probability distributions – among the region's constituent gauges.

If a region of gauges is proposed, some statistical test must be used to validate the appropriateness of the grouping. Simulation studies in which a homogeneous region and equivalent heterogeneous regions with a known degree of variation from the regional mean are compared have been used to isolate the error of quantile estimation added due to heterogeneity. In this study the relationship between heterogeneity statistics and the error of quantile estimation for real data is examined for simulated data and for real precipitation data at durations ranging from one day to thirty days in length. Quantifications of the heterogeneity-error relationship are offered for the consideration of future precipitation analysts, as they provide a statistical basis on which to defend the precipitation gauge regions formed in their studies.

1.1: Regional frequency analysis

The pooling of a region of gauges, or sites, into a regional dataset is typically preceded by normalization; each gauge's record is divided by that gauge's mean or median, or some other value, referred to as the "index flood". The normalized data can be used to generate normalized regional quantile estimates which when multiplied by a site's index flood are transformed into at-site quantile estimates. This allows gauges with different magnitudes but similarly shaped probability distributions to benefit from the increase in sample size provided by a region with low heterogeneity. It also allows the

index flood, which if chosen as a statistic like the mean or median requires less sample size to achieve a given accuracy threshold than extreme quantile estimates, to be determined by at-site data. The procedure of multiplying regional estimates with a more stable at-site statistic means that uncertainties associated with regionalization have less effect at quantiles close to the index flood. However, to achieve the promise of improved extreme quantile estimates the balance between sample size and regional heterogeneity must be struck.

1.2: Linear moments

A statistical summary of the shape of a gauge and a region's probability distribution is necessary for quantification of the concepts used in regional frequency analysis. After data have been re-ordered from low to high their empirical distribution can be quantified and transformed into parameters for a frequency distribution. Regional frequency analysts have long used the conventional moments of data, such as mean, coefficient of variation, skewness, and kurtosis, to summarize the 'shape' of the empirical probability distribution (Dalrymple, 1960). Equations exist to create a probability distribution when given these values as parameters. However, research has shown that hydrological data are better modeled through the use of linear moments, which approximate the behavior of the conventional moments with weighted polynomials (Hosking et al., 1985).

Because higher-order conventional moments require increasingly large exponents to be applied to the difference between each data point and the record's mean, estimates are unstable and outliers have a large effect on the calculations. When these moments are

used to parameterize a distribution and the distribution in turn is used to generate new data, the synthetic dataset often poorly reflects the characteristics of the original data. Also, sample estimates of conventional moments often have unfavorable algebraic bounds, most notably for skewness, whose value is dependent on sample size – a major problem for analysts seeking to find quantile estimates for short gauge records. Linear (or L-) moments are constructed so that the difference between each data point and the mean is not exponentiated but instead is multiplied by a polynomial. This reduces the effect of outliers and can be modeled by sample estimates with more favorable algebraic bounds.

1.2.1: Fitting probability distributions using L-moments

After L-moment ratios have been calculated for a dataset (whether at a single gauge or after regional pooling) they can be used as parameters in equations that calculate quantile, cumulative distribution and probability density functions for probability distributions defined by statisticians. Parameterizations using both conventional and linear moments are available for the most commonly used distributions. The most famous probability distribution is the Normal, which for centuries has been recognized as a reasonable model for many types of data. Another family of distributions including the Fréchet, Weibull and Gumbel distributions is derived from extreme-value theory and fits many types of data well, especially at extreme quantiles.

The Normal is a two-parameter distribution, taking only the mean and coefficient of variance of data and assuming a given skewness and kurtosis value. A three-parameter distribution additionally takes the skewness as an input so that only the kurtosis is assumed; except for the first parameter, which is the mean or first moment, distributions

are parameterized by moment ratios. “Generalized” three-parameter probability distributions have been defined which bring together multiple two-parameter distributions, such as the Generalized extreme-value distribution. The four-parameter Kappa distribution unifies several three-parameter distributions and the five-parameter Wakeby acts as a “parent distribution” to an even larger set of distributions. Equations for these and other probability distributions can be found in Appendix A of Hosking and Wallis (1997).

Distributions with too many parameters run the risk of overfitting the data. Higher-order L-moments require more sample size for accurate estimation; at low sample sizes, sample estimates of higher-order L-moments are likely to be significantly different from the properties of the underlying distribution. This risks enshrining random variation due to insufficient sample size as a characteristic of the data, resulting in inaccurate quantile estimates. However, distributions with too few parameters enforce distributional assumptions regarding higher-order moment ratios that may be at odds with the data. For example, precipitation data is typically highly skewed and the Normal distribution models unskewed data; extreme quantile estimates outputted from the Normal for most precipitation data would be wildly inaccurate. A useful distribution must possess distributional assumptions for higher-order moments that have been shown to be broadly applicable to the type of data in question while retaining the ability to represent the idiosyncrasies of the analyzed data record.

Hydrologists typically approach this problem by using the mean, the coefficient of variation and the skewness of data as parameters while using the kurtosis of the data to

determine which of a set of three-parameter distributions is most applicable to the data. In this way the issue of overfitting is ameliorated but the kurtosis data excluded from the parameter set is used for guidance as the distribution is chosen. Three-parameter distributions which have been found to provide close fits to hydrologic data include the Pearson Type III, the Lognormal, and the Generalized extreme-value distributions. The process requires an assumption that events of a given magnitude occur at a fixed probability over time, with the shape of the underlying probability distribution remaining unchanged between past and future data.

1.3: Estimating heterogeneity

After data have been sorted from lowest to highest, L-moment ratios can be calculated for the data or nonparametric statistics based only on the empirical distribution of the observed data can be formulated. In both cases a measure of the difference between the 'shape' of the dataset at each site and in the pooled regional dataset can be averaged across all sites in the region to find a regional heterogeneity value. The squared difference between at-site and regional L-moment ratios averaged across all sites in the region offers a summary of heterogeneity with respect to variation, skewness, or kurtosis. When the fraction of a site's dataset below a reference threshold is compared to the fraction of the total pooled data under that threshold (after correcting for normalization) the difference between the two can be averaged across all member sites of a region to create a nonparametric measure of regional heterogeneity. Equations for evaluated statistics are presented in the manuscripts that have been incorporated into this dissertation below.

Because the L-moment ratios of a dataset can also be used to parameterize a probability distribution, data generated from that distribution can be assessed for heterogeneity across a large number of Monte Carlo simulations. Monte Carlo methods can also be applied to nonparametric statistics through sampling with replacement from the observed data. In the first case a distributional assumption is incorporated into the method; the second requires the assumption that the data record represents an unbiased sampling from the underlying probability distribution.

1.4: Estimating quantile error

Reducing error of quantile estimation, especially at extreme quantiles, is ultimately the purpose of regional frequency analysis. For simulated data where 'true' L-moment ratios are assigned at each site, the difference between at-site and pooled regional quantile estimates (after multiplying the normalized regional estimate by the at-site index flood) can be squared and averaged across all sites in the region to obtain an estimate of the error of quantile estimation. Through application of the assumption of an underlying probability distribution Monte Carlo simulation can be applied to real data; averaging simulated regions' error measures across all simulations can quantify the degree to which regional quantile estimation adds error. The simulated regions used in this analysis can be generated with intersite correlations taken into account.

Because true at-site L-moment ratios are unknown for real data and the estimates which are available are sample statistics, the range of variation in L-moment ratios across the region and therefore the magnitude of a heterogeneity estimate is increased by sampling error. If simulated regions are generated from a distribution fitted with L-

moment ratios estimated from real data, another round of sampling error will further increase the range of variation observed between the sites' L-moment ratios. The at-site L-moment ratios used as the basis for generation of simulated regions must therefore be shrunk toward the regional mean. The degree of shrinkage must be calibrated so as to counteract the effects of sampling error, resulting in the generation of simulated regions with estimated heterogeneity nearly equal to the original region. Equations for this method are presented in the manuscripts included in this dissertation, below.

The four subsequent chapters of this dissertation comprise three manuscripts submitted to peer-reviewed journals and a Conclusion section. The first paper has been accepted at the *Journal of Hydrology*, the second has been submitted to *Water Resources Research*, and the third has been submitted to the *Journal of Hydrologic Engineering*. Drs. Mark Houck and Celso Ferreira are co-authors on all papers, while Dr. Jason Giovannettone is a co-author as well on the second. These authors assisted in creating the experimental framework of the research. They also offered constructive commentary at every stage of the research and offering edits to the manuscripts of which they are co-authors. The papers present methods for evaluating heterogeneity statistics and offer conclusions regarding the effectiveness of various heterogeneity statistics used in the literature. The Conclusion section summarizes the findings of the three papers and offers future directions for research. References are listed after the Conclusion section.

The citations for the papers are as follows:

Wright, M.J., Ferreira, C.M., Houck, M.H., 2014. Evaluation of heterogeneity statistics as reasonable proxies of precipitation quantile estimation in the Minneapolis-St. Paul region. Journal of Hydrology, doi: 10.1016/j.jhydrol.2014.03.056.

Wright, M.J., Ferreira, C.M., Houck, M.H., Giovannettone, J.P., 2014. The relationship between Monte Carlo estimators of heterogeneity and error for daily to monthly time steps in a small Minnesota precipitation gauge network. Water Resources Research (submitted).

Wright, M.J., Ferreira, C.M., Houck, M.H., 2014. Discriminatory power of heterogeneity statistics with respect to error of precipitation quantile estimation. Journal of Hydrologic Engineering (submitted).

2: EVALUATION OF HETEROGENEITY STATISTICS AS REASONABLE PROXIES OF THE ERROR OF PRECIPITATION QUANTILE ESTIMATION IN THE MINNEAPOLIS-ST. PAUL REGION

Abstract

Estimating precipitation frequency is important in engineering, agriculture, land use planning, and many other disciplines. The index flood method alleviates small sample size issues due to short record length by calculating normalized quantile estimates for averaged data from a "region" of gauges. For a perfectly homogeneous region this adds no error; heterogeneity statistics seek to quantify a real-world region's deviation from this assumption. Hosking and Wallis (1997) introduced a Monte Carlo heterogeneity statistic called here H_1 and used a simulation study to assess its utility while rejecting two similar statistics called here H_2 and H_3. A nearly linear relationship was found between H_1 and the percentage root mean square error (RMSE) increase due to heterogeneity, establishing H_1 as a "reasonable proxy" of quantile error. The H_1-percent RMSE added relationship found in the simulation experiment was used to find equivalent RMSEs for heterogeneity thresholds against which all three H statistics were tested. In this study the "reasonable proxy" relationship is evaluated across a highly skewed daily precipitation dataset in Minnesota for H_1, H_2, and H_3. Simulated regions used in quantile error estimation are generated using at-site L-moment ratios scaled toward the regional mean with a shrinkage

multiplier. A linear relationship is found between Monte Carlo estimates of quantile RMSE and both H_1 and H_2 across all possible regionalizations of twelve gauges. H_2's relationship is less linear than H_1's as quantified by Pearson's r. A synthetic study is also undertaken using the same sample sizes, regional L-moment averages, and between-site variations as the Hosking and Wallis (1997) simulation. The H_2-percent RMSE added relationship is found to be nearly as linear as for H_1, complementing the enumeration study's findings. Because H_2's linear relationship with percent RMSE added has approximately one-fourth the slope of the H_1-RMSE relationship, heterogeneity thresholds calculated with reference to H_1 should not be applied to H_2. H_2 thresholds can be derived from the H_2-percent RMSE added relationship in analogous fashion to the method used in Hosking and Wallis (1997) for H_1. The resulting thresholds are one-fourth the magnitude of the H_1 thresholds.

2.1: Introduction

Rare or extreme precipitation events, which include events classified as natural disasters, have major ecological, economic, and public safety significance. Sample size is often a limiting factor in the estimation of extreme hydrological events; one rule of thumb for flood frequency estimation is that for reliable estimates of a return period T the record length in station-years must exceed 5T (Robson and Reed, 1999). Many statistical hydrologists have followed the index flood method of Dalrymple (1960), in which sample size is increased by grouping gauges, or "sites", into regions, calculating a regional "growth curve" normalized by an index such as the mean or median of the at-

gauge data, and estimating at-site quantiles by multiplying the index flood and the regional growth curve.

Linear moments, analogues of conventional central moments like skewness and kurtosis based on probability weighted moments (Greenwood et al., 1979), are often used in this context. L-moment estimators have lower bias than other common methods of estimation at small sample size (Hosking et al., 1985; Lettenmaier et al, 1987). They are less biased than conventional moment estimators, are not bounded by sample size, and are more robust to outliers. L-moment ratios can be more reliably predicted from a subsample than conventional moment ratios. L-moment ratios, the second through fourth of which are denoted the coefficient of L-variance (L-CV), L-skewness, and L-kurtosis (the first L-moment ratio does not exist), provide greater insight into the underlying distribution of high-skew data than conventional moment ratios. For example, L-moment ratio diagrams are used as decision aides for identifying the underlying distribution of regional data (Hosking, 1990; Vogel and Fennessey, 1993; Hosking and Wallis, 1997; Zafirakou-Koulouris et al., 1998).

L-moment analysis of regions formed according to hydrological characteristics has been conducted in recent decades on streamflow (Vogel et al., 1993a; Ouarda et al., 2008; Noto and Loggia, 2009) and precipitation data (Guttman et al., 1993; Werick et al., 1994; Adamowski et al., 1996; Alila, 1999; Smithers and Schulze, 2001; Kyselý et al., 2007; Modarres and Sarhadi, 2011). L-moment ratios for daily data series using "wet-day" (non-zero only) and full datasets have been evaluated across the United States (Hansong and Vogel, 2008). Regional frequency analysis models using fuzzy regions

(Jingyi and Hall, 2004; Rao and Srinivas, 2006) and fractional-membership regions of influence (Burn, 1990; Zrinji and Burn, 1994; Gaál et al., 2008) represent alternatives to the strict regional membership model.

The regional pooling mechanism of the index flood method involves the assumption of homogeneity across the sites in a candidate region - at-site differences in L-moment ratios are assumed to be due solely to sampling variability. The degree to which the homogeneity assumption is violated is therefore likely to be related to quantile error. Statistics quantifying the heterogeneity of a region based on Monte Carlo simulation have been proposed which sample from a Generalized extreme-value distribution (Lu and Stedinger, 1992; Alila, 1999).

Hosking and Wallis (1997) define three statistics based on the between-site variation of L-moment ratios, H_1, H_2, and H_3. All three H statistics fit the Kappa distribution with the average L-moment ratios of the region in question and use Monte Carlo simulation to generate simulated regions from the Kappa. For the real region and for each simulated region a statistic called V_1, V_2, or V_3 is calculated using the sum of the squared difference between each site's L-moment ratio values and the regional average. V_1 uses only the L-CV, V_2 incorporates L-CV and L-skewness, and V_3 incorporates L-skewness and L-kurtosis. H_1 is calculated when V_1 for the real region minus the mean of V_1 for simulated regions is divided by the standard deviation of simulated regions' V_1; H_2 and H_3 are calculated analogously (see Equations 6-9).

A simulation study is used in Hosking and Wallis (1997) to reject H_2 and H_3 and to accept H_1. Heterogeneous regions' RMSEs are divided by their equivalent

homogeneous region's RMSE. This isolates the RMSE increase due to heterogeneity. H_1 is shown to have a linear relationship with percent RMSE added due to heterogeneity. Results for H_2 and H_3 are not reported.

Hosking and Wallis (1997) define thresholds below which regions can be considered “possibly” and “definitely” heterogeneous with reference to a range of percent RMSE added values implied by the H_1-percent RMSE added relationship. $H_1 = 1$ is found to indicate a 20-40% increase in RMSE, while $H_1 = 2$ is associated with 40-80% increases. These thresholds are also applied to H_2 and H_3, which are found to rarely exceed them.

Viglione et al. (2007) investigate H_1 and H_2 as well as two nonparametric heterogeneity statistics by measuring the fraction of simulated regions that are correctly and incorrectly identified as heterogeneous. The threshold of $H = 2$ is used for both H_1 and H_2. They confirm the utility of H_1 for simulated data with L-skewness below 0.23 and reject H_2. The bootstrap Anderson-Darling test is found to be more powerful than either statistic for data with higher skewness.

Two approaches are used in this study to quantify the power of H_1, H_2, and H_3 as proxies of error due to heterogeneity. The original Hosking and Wallis (1997) simulation study is recapitulated and results for H_2, and H_3 are presented alongside those for H_1. Thresholds for H_2 are found using its linear relationship to quantile error, not H_1's. An enumeration study is also conducted, estimating H_1, H_2, H_3, and quantile error for all possible regionalizations of a small daily precipitation gauge dataset. Components of error unrelated to heterogeneity are preserved in this study, allowing the heterogeneity

statistics' relationships with total estimated quantile error to be compared to the ideal case represented in the simulation experiment.

The remainder of this paper is structured as follows. Daily precipitation gauge data from which all possible regions are to be enumerated are presented in the following section, Section 2. Section 3 introduces the equations and methods used to calculate linear moments of the data, estimate regional heterogeneity, assign a regional distribution, estimate the RMSE of regional quantile estimates, and perform a simulation experiment analogous to that presented in Hosking and Wallis (1997). Section 4 presents the results first of the simulation experiment, then of the enumeration experiment, which compares heterogeneity and error estimates across all possible regions formed from the selected precipitation gauges. Section 5 discusses the results; section 6 summarizes the paper and offers conclusions and potential avenues of future research.

2.2: Data

Mean annual precipitation in Minnesota ranges from the low teens to above 30 inches, with the mean annual precipitation generally increasing from the northwest to the southeast. Moist air carried from the Gulf of Mexico is an important source of precipitation in Minnesota. Almost half of yearly precipitation occurs in June, July, and August (Baker et al., 1967). Rainfall quantiles have been estimated for Minnesota using the Generalized extreme-value distribution on the annual maximum series (Blumenfeld and Skaggs, 2011) and strict-membership regional analysis using L-moments (Werick et al., 1994), as well as a study comparing several methods including regional L-moment

analysis (Huff and Angel, 1992). Updates to the NOAA Precipitation Frequency Atlas of the United States (Perica and coauthors, 2013) for Minnesota are underway.

Precipitation gauge data were provided by the Minnesota State Climatology Office from a high-density rain gauge network made up of hundreds of long-record volunteer gauges. Reliable observers' records were collected and quality controlled for inclusion in the updated version of NOAA's Precipitation Frequency Atlas of the United States (Perica and coauthors, 2013), which incorporates the L-moment, regionalization, and Monte Carlo error estimation procedures of Hosking and Wallis (1997). Estimates for Volume 8, "Midwestern States" are available online at http://hdsc.nws.noaa.gov/hdsc/pfds/pfds_map_cont.html.

From a database of daily precipitation totals at 341 Minnesota gauges, 57 which did not skip a full month between the start and end of the record were selected for further analysis. Twelve of these gauges were clustered in the Minneapolis-St. Paul region, mostly in Ramsey County which contains St. Paul (Figure 1, Table 1). The longest record length of the twelve sites is 3751 days of non-zero precipitation and the shortest is 1134 days. The wet-day precipitation records of this group are subjected to further analysis below.

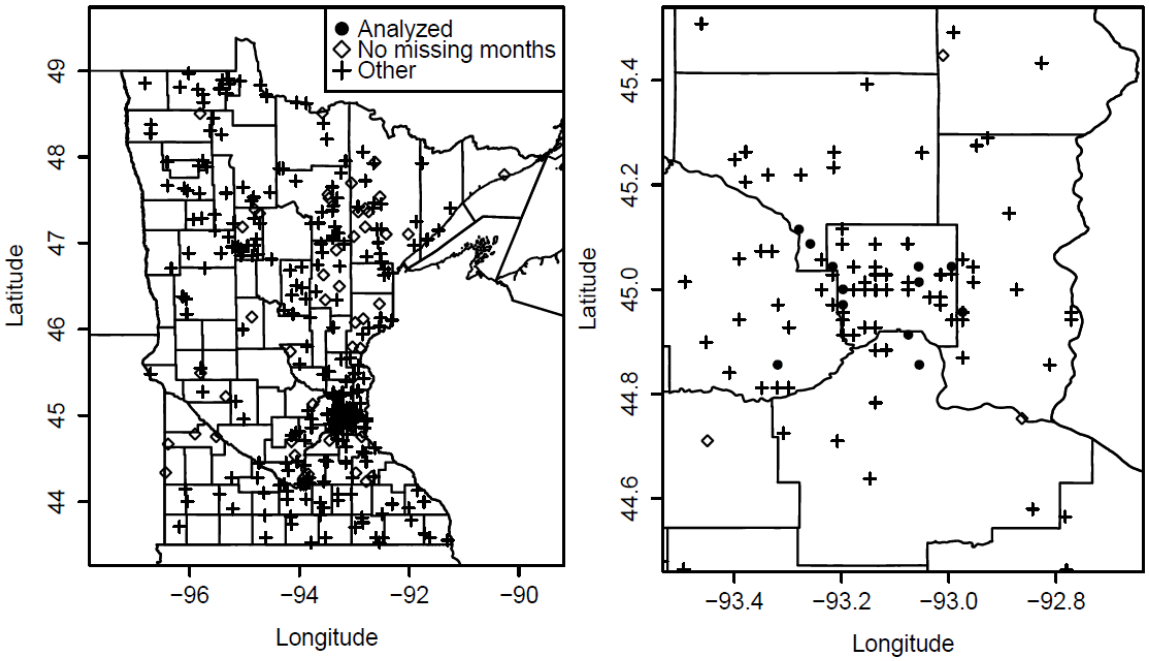

Figure 1 Location of gauges in degrees latitude and longitude

Table 1 Selected Minnesota rain gauges; ID, sample size, first and last month of data record, and location in degrees latitude and longitude

Gauge ID #	# days	Start date	End date	Latitude	Longitude
11	1714	Feb 1982	Feb 1996	44.86	-93.32
35	1191	Jan 1999	Jan 2009	45.01	-93.06
39	2419	Jan 1978	Aug 1998	45.04	-92.99
46	2765	Dec 1980	Nov 2008	45.12	-93.28
78	1518	Dec 1975	Apr 1990	44.97	-93.20
104	3751	Jan 1976	Nov 2008	45.09	-93.26
149	1488	Aug 1993	Nov 2008	44.86	-93.05
150	1787	Jul 1993	Nov 2008	44.91	-93.08
266	1859	May 1992	Oct 2008	45.00	-93.20
268	1438	Sep 1990	Nov 2004	45.04	-93.06
272	1134	Sep 1998	Nov 2008	45.04	-93.22
328	3572	Sep 1979	Dec 2007	44.96	-92.97

The “lmomRFA” package for the statistical language R (R Core Team, 2012) was used to calculate L-moments, distribution parameters, and both the tested heterogeneity statistics and the error statistics used to compare them. “lmomRFA” documentation is available online at http://cran.r-project.org/web/packages/lmomRFA/lmomRFA.pdf. The package applies FORTRAN code available at http://lib.stat.cmu.edu/general/lmoments. Scientific computing resources were provided through the Extreme Science and Engineering Discovery Environment (XSEDE) network on the Steele computing cluster at Purdue University.

2.3: Analysis

2.3.1: Linear moments

The Hosking (1990) approach begins from the probability weighted moments (PWM) of Greenwood et al. (1979), defined in Equation (1):

Equation 1

$$\beta_r = \mathrm{E}\{\mathrm{X}[\mathrm{F}(\mathrm{X})]^r\}$$

Where F(X) is the cumulative distribution function (cdf) of X, X(F) is the inverse cdf or quantile function of X for probability F, and β_r is the rth-order PWM (β_0 is equal to the mean $\mu = E(X)$). Hosking (1990) defines the L-moments λ_{r+1} in Equation (2):

Equation 2

$$\lambda_{r+1} = \sum_{k=0}^{r} p^*_{r,k}\beta_k$$

With $p^*_{r,k}$ calculated according to Equation (3):

Equation 3

$$\mathrm{p}^*_{r,k} = \frac{(-1)^{r-k}(r+k)!}{(k!)^2(r-k)!}$$

λ_1 is the mean of a data record, λ_2 measures its dispersion, λ_3 measures its skewness, and λ_4 measures its kurtosis. From these L-moments, L-moment ratios are

constructed; the L-CV or coefficient of L-variation is $\tau = \lambda_2 / \lambda_1$ and the ratios of L-skewness and L-kurtosis are calculated using r = 3 and 4 respectively in Equation (4):

Equation 4

$$\tau_r = \lambda_r / \lambda_2$$

The sample estimate of β_r is b_r, estimated using Equation (5):

Equation 5

$$b_r = n^{-1} \sum_{j=r+1}^{n} \frac{(j-1)(j-2)\dots(j-r)}{(n-1)(n-2)\dots(n-r)} x_{j:n}$$

Where n is the number of data points at the precipitation gauge and $x_{j:n}$ is the jth-smallest of the data points.

The sample estimate of λ_r is l_r and the sample estimate of τ_r is t_r with the sample estimate of the L-CV denoted as t; these are calculated using Equations (2-4), substituting b_r for β_r. L-moment ratio values for selected gauges are presented in Table 2.

Table 2 Selected Minnesota rain gauges: ID and selected L-moment statistics

Gauge ID #	Mean (L_1)	L_2	t_3	t_4
11	0.292	0.188	0.516	0.282
35	0.271	0.180	0.538	0.295
39	0.289	0.187	0.531	0.304
46	0.327	0.200	0.475	0.240
78	0.279	0.177	0.502	0.269
104	0.277	0.178	0.509	0.264
149	0.335	0.213	0.513	0.282
150	0.305	0.198	0.514	0.267
266	0.316	0.206	0.522	0.286
268	0.334	0.215	0.520	0.285
272	0.311	0.202	0.514	0.267
328	0.271	0.176	0.520	0.273

2.3.2: Heterogeneity measures

In the heterogeneity statistics of Hosking and Wallis (1997), pooled gauge data are fitted to the Kappa distribution and simulated regions with the same number of sites and record lengths as the real data are generated. A measure V of between-site L-moment ratio variation is calculated for the regional data and for each of many iterations of each simulated region. The mean of V for all simulated regions, μ_V, is subtracted from true regional V and the result is divided by the standard deviation of V for all simulated regions, σ_V, to produce the heterogeneity statistic H (Equation (6)).

Equation 6

$$H = \frac{V - \mu_V}{\sigma_V}$$

Three variants of V and therefore of H are identified. H_1 is calculated from V_1, H_2 from V_2, and H_3 from V_3. V_1 is calculated from the second L-moment ratio, V_2 from the second and third, and V_3 from the third and fourth (Equations 7-9):

Equation 7

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

Equation 8

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

Equation 9

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

The (i) superscript indicates the at-site L-moment ratio value for gauge i; the R superscript indicates the regional average L-moment ratio. N is the number of sites in the region and n_i is the number of data points at site i. The region containing all twelve sites has H_1, H_2 and H_3 values 4.01, 1.44, and 0.77, respectively.

Hosking and Wallis (1997) conduct simulations of regions with varying homogeneity, defined using linear variations in the at-site L-moment ratio values to which the Kappa distribution is fitted, to find a nearly linear relationship between H_1 and the ratio between RMSEs of a heterogeneous region and its corresponding homogeneous

region. This finding is interpreted to show that H_1 is "a reasonable proxy for the likely error in quantile estimates." H_2 and H_3 were found to have little discriminatory power based on similar reasoning. Heterogeneity is found to be indicated by $H_1 > 2$ and homogeneity by $H_1 < 1$, while regions with intermediate values may be improved by redefinition.

2.3.3: Distribution fitting

L-moment ratio diagrams, especially L-skewness/L-kurtosis plots, are useful for choosing a distribution to fit to the data. Power-law approximations of the relationship between t_3 and t_4 for each of several three-parameter distributions are provided in FORTRAN at http://lib.stat.cmu.edu/general/lmoments.

Graphing these curves against the t_3 and t_4 values of the data allows for graphical testing of the hypothesis that simulated data drawn randomly from the distribution will have similar higher-order moment values to the fitted data. The higher-order L-moment ratios of the 57 precipitation gauges selected from the Minnesota dataset are best described by the $t_3 – t_4$ relationship inherent in the Pearson type III distribution (Figure 2); the pattern is comparable to that observed in Hanson and Vogel (2008)'s nationwide wet-day dataset. Hosking and Wallis (1997) provide a mathematical analogue to this graphical line-fitting test.

Equations for deriving the Pearson type III distribution's parameters from L-moment ratios and plotting its cumulative distribution and probability density functions can be found in Appendix A.9 of Hosking and Wallis (1997).

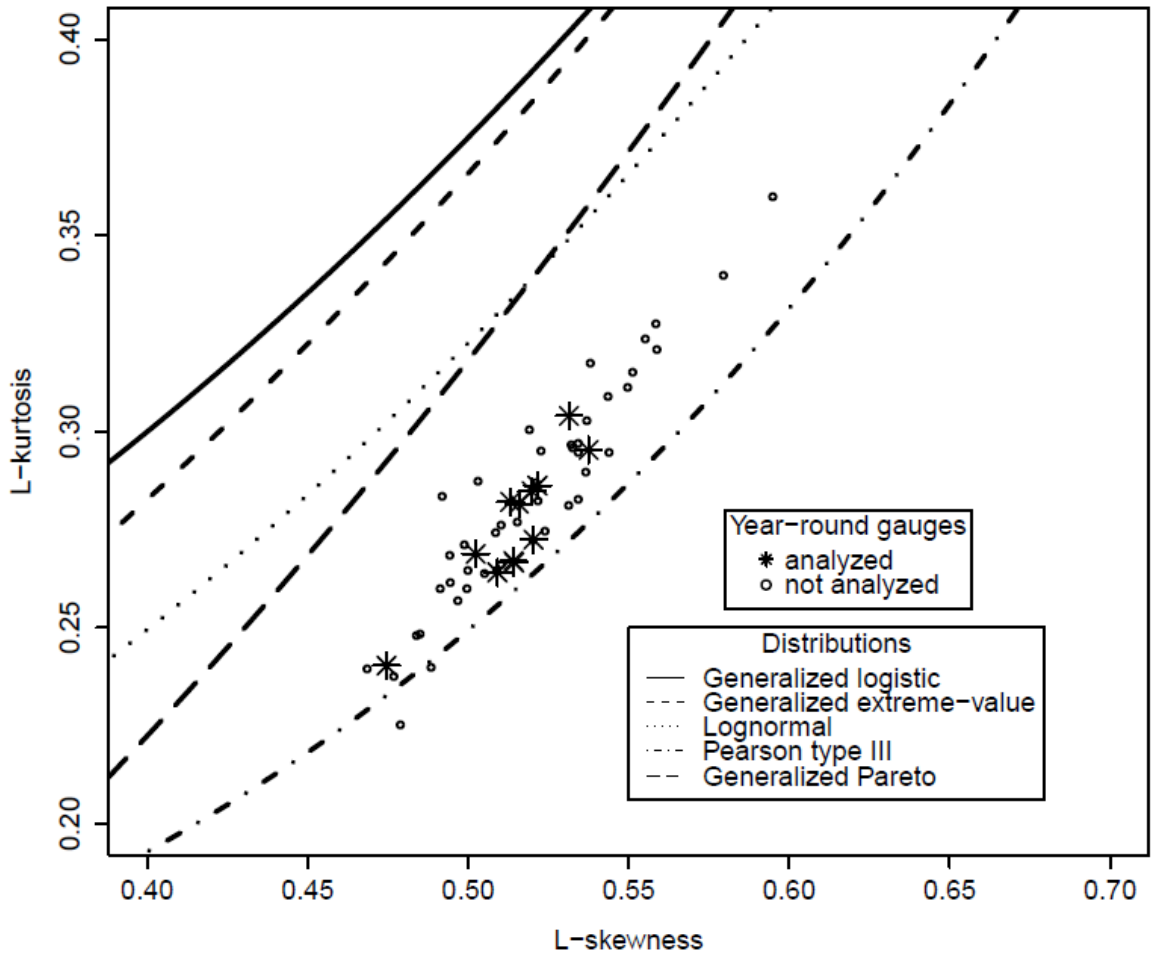

Figure 2 L-skewness/L-kurtosis plot of Minnesota data

2.3.4: Quantile error estimation

Hosking and Wallis (1997) outline a Monte Carlo procedure for estimating the error of quantile estimates taking heterogeneity and inter-site dependence into account (Section 6.4, Table 6.1). This procedure has been implemented in the R package "lmomRFA" as the function "regsimq". "Regsimq" creates simulated regional data on a

site-by-site basis from a list of at-site frequency distributions and parameters; in this study all sites will be fitted to the Pearson type III distribution.

A region of simulated sites is formed with similar heterogeneity to the true region for which error is being estimated. These simulated sites are used in "regsimq" as the "true" data against which root mean square error (RMSE) estimates are calculated. Hosking and Wallis (1997) state that "some arbitrariness is inevitable" at this stage, as many different patterns of L-moment ratio variation can produce similar heterogeneity. In section 6.5 of that monograph an example is offered in which L-moment ratios are created for each member gauge of the simulated region by linearly scaling the second L-moment ratio, L-CV, across a fraction of the observed variance, the fraction being varied until simulated H_1 is nearly equal to true H_1. Higher L-moment ratios are set equal to the regional average.

However, for this paper's dataset it is found that such a procedure often outputs physically impossible at-site Pearson type III parameters, i.e. the minimum quantile estimate is negative. Linearly scaling the L-CV and the higher-order L-moment ratios, with or without permutation to vary at-site combinations of L-moment ratios, does not eliminate this problem, which is not observed in the true at-site data. An alternative procedure is used which avoids precipitation estimates below zero.

The observed difference between at-site and regional average L-moment ratios is reduced by shrinking the at-site values toward the regional averages. This counteracts the tendency of sampled data from a distribution to vary across a wider range of L-moment ratio values than seen in the original population. Theoretical justification of shrinkage

estimators can be found originally in Stein (1956), who showed that the multivariate sample mean is inadmissible, and from a Bayesian perspective in Lindley and Smith (1972). Shrinkage is accomplished by applying a multiplier with value between 0 and 1. A multiplier of 0 sets all at-site L-moment ratios equal to the regional average, a multiplier of 1 leaves the true at-site L-moment ratios unchanged, and a multiplier of 0.5 creates a simulated region with at-site L-moment ratios halfway between the regional average and the true at-site values.

Prior to running "regsimq" the average correlation between each pair of gauges in the dataset is calculated according to Hosking and Wallis (1997) using Equation (10):

Equation 10

$$r_{ij} = \frac{\sum_k (Q_{ik} - \bar{Q}_i)(Q_{jk} - \bar{Q}_j)}{\{\sum_k (Q_{ik} - \bar{Q}_i)^2 \sum_k \left(Q_{jk} - \bar{Q}_j\right)^2\}^{1/2}}$$

r_{ij} is the sample correlation between gauges I and j, k is the index and n_{ij} is the number of time points for which both gauges I and j have data Q_{ik}, and $\overline{Q}_i$ is calculated using Equation (11).

Equation 11

$$\bar{Q}_i = n_{ij}^{-1} \sum_k Q_{ik}$$

Table 3 shows the correlation between all twelve sites; a correlation of zero indicates that the sites shared no days of record.

Table 3 Correlation between twelve sites in dataset

Site ID	11	35	39	46	78	104	149	150	266	268	272	328
11	1	0	0.69	0.49	0.67	0.61	0.71	0.77	0.70	0.60	0	0.54
35	-	1	0	0.83	0	0.60	0.80	0.87	0.81	0.88	0.66	0.84
39	-	-	1	0.46	0.69	0.70	0.30	0.36	0.48	0.54	0	0.50
46	-	-	-	1	0.57	0.67	0.73	0.77	0.77	0.82	0.67	0.73
78	-	-	-	-	1	0.84	0	0	0	0	0	0.66
104	-	-	-	-	-	1	0.49	0.55	0.58	0.66	0.87	0.64
149	-	-	-	-	-	-	1	0.86	0.70	0.73	0.55	0.77
150	-	-	-	-	-	-	-	1	0.76	0.79	0.60	0.83
266	-	-	-	-	-	-	-	-	1	0.79	0.62	0.76
268	-	-	-	-	-	-	-	-	-	1	0.66	0.78
272	-	-	-	-	-	-	-	-	-	-	1	0.67
328	-	-	-	-	-	-	-	-	-	-	-	1

The average inter-site correlation $\bar{r}$ is then calculated using Equation (12).

Equation 12

$$\bar{r} = \{\frac{1}{2}N(N-1)\}^{-1} \sum_{1 \le i <} \sum_{j \le N} r_{ij}$$

"Regsimq" uses $\bar{r}$ as an estimate of the correlation ρ between all gauge pairs and creates a correlation matrix of the form in Equation (13):

Equation 13

$$R = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & & \ddots & \\ \rho & \rho & & & 1 \end{bmatrix}$$

"Regsimq" generates multivariate Normal correlated data for each simulated gauge, transforms them using the Pearson type III distribution parameterized as above, normalizes the data and finds the simulated region's average L-moment ratios, which are fitted to the Pearson type III distribution. Normalized quantile estimates, also known as

"growth curves", are calculated for the simulated region for the specified number of iterations. For a list of non-exceedance probabilities F the root mean square error comparing at-site growth curve estimation Q_i to the regional growth curve $\widehat{Q}_i^{(m)}$ is averaged across all iterations m and all sites i, creating a root mean square error estimate for each non-exceedance probability (Equation 14).

Equation 14

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

All possible regions containing five or more sites from the twelve selected gauges are enumerated. Each region's heterogeneity statistics are calculated over 2500 iterations in "regtst", the function in "lmomRFA" for calculating the three heterogeneity statistics. One hundred iterations are performed for each of ten randomly generated, correlated simulated regions with 0.75 of observed L-moment ratio deviation from the regional average. The fraction of observed deviation is decreased by 0.1 if simulated H_1 is greater than true regional H_1 and increased by 0.1 otherwise, and one hundred iterations of ten regions are run again. When simulated H_1 drops below, or rises above, true H_2, the mean of the last two multipliers is tested and the closest fit of the three is used to scale true at-site L-moment ratios toward the regional mean, creating the simulated sites used in a 2500-iteration run of the error estimation routine "regsimq". The maximum multiplier allowed is 0.95 and the minimum is 0.05.

For example, if simulated H_1 was greater than true H_1 for a multiplier of 0.75, but simulated H_1 was less than true H_1 for 0.65, a multiplier of 0.7 would also be evaluated;

the multiplier among 0.65, 0.7, and 0.75 producing the smallest value of |true H_1 – simulated H_1| would be used to create the simulated sites inputted into "regsimq".

2.3.5: Evaluation of heterogeneity statistics across simulated data

A synthetic study is undertaken recapitulating and extending the simulation experiment conducted in Hosking and Wallis (1997) and summarized there in Section 4.3.4, Table 4.1, and Figure 4.2. This study features simulated regions whose sites' L-moment ratio values exhibit 0% to 50% variation around the regional average L-CV and L-skewness values. 100 realizations of each region are generated. Both regional average L-CV and L-skewness and the ranges of their variation between sites are set equal, with two minor exceptions at low L-skewness.

The Generalized extreme-value distribution is used to output quantile estimates at each simulated site for both initial L-moment ratios and data generated using these ratios, from which the RMSE is calculated for each simulated region at any given non-exceedance probability using Equation (14). Heterogeneity estimates are calculated according to Equations 6-9 using 500 Monte Carlo iterations and averaged across each of the 100 realizations of each region. The RMSE ratio between a heterogeneous region and its homogeneous equivalent is then compared to heterogeneity estimates; a linear relationship signifies "reasonable proxy" status for the heterogeneity statistic in question. Pearson's r is used to quantify the linearity of the relationship.

2.4: Results

2.4.1: Results of simulation experiment

The simulation experiment described in the preceding section outputted similar results to those described in Table 4.1 of Hosking and Wallis (1997) for H_1. Averages of the H statistics are taken across 100 simulations for each set of L-moment ratios. In Figure 3, depicting the linear relationship between each H statistic and percent RMSE added due to heterogeneity for a non-exceedance frequency of 0.01, results for H_1 are similar to Figure 4.2 in Hosking and Wallis (1997). Results for H_2 and H_3 are not reported in that monograph but Figure 3 indicates a relationship between H_2 and percent RMSE added due to heterogeneity that is similar to that found for H_1, albeit less linear and with a lower slope. Pearson's r score of heterogeneity statistics plotted against percent RMSE added is found to decrease from H_1 to H_2 to H_3 (Figure 4).

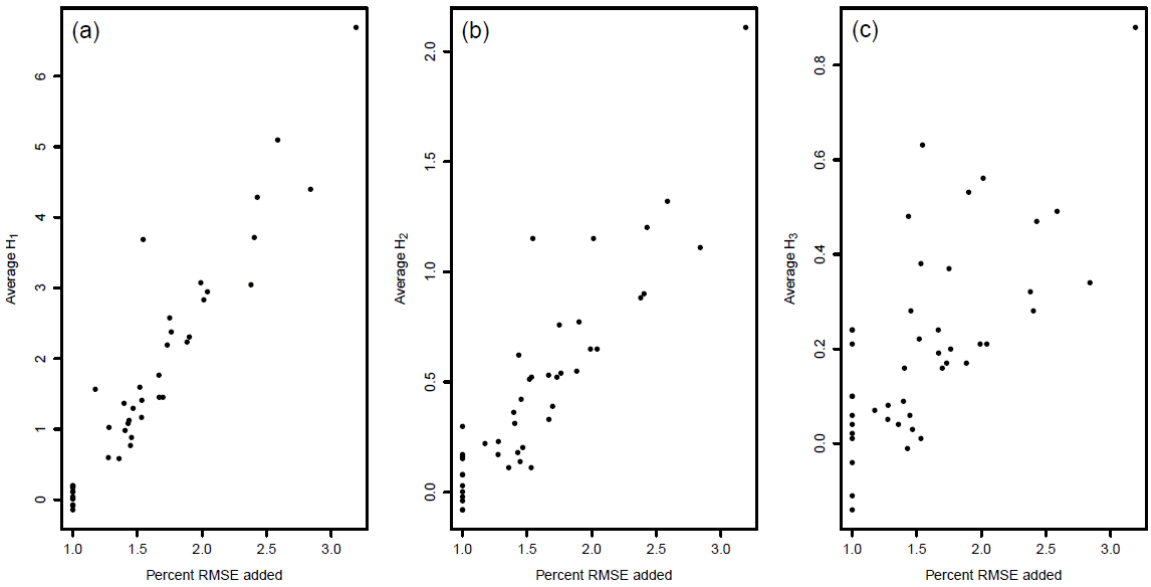

Figure 3 Percent RMSE added due to heterogeneity for simulated regions plotted against (a) H_1, (b) H_2, and (c) H_3 at non-exceedance probability of 0.01.

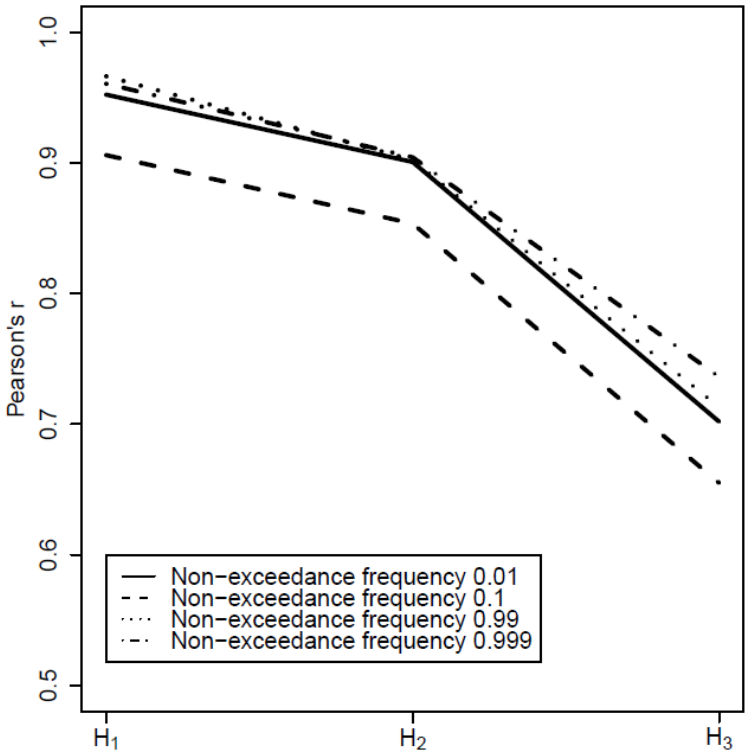

Figure 4 Pearson's r of linear fit between percent RMSE added due to heterogeneity and the H statistics.

2.4.2: Results of enumeration experiment

The scaling procedure for creating simulated datasets produced similar H_1 values to that of the true region outputted; the average discrepancy was less than 0.02, while more than half of the fitted regions had a discrepancy lower than 0.125 and the maximum was 1.065. The ranges of discrepancies were similar for low, medium, and high multipliers. Most regions with positive H_1 were fitted using a multiplier of 0.85 or above, while regions with negative H_1 required multipliers as low as 0.05. Figure 5 shows the number of regions to which a given multiplier was assigned.

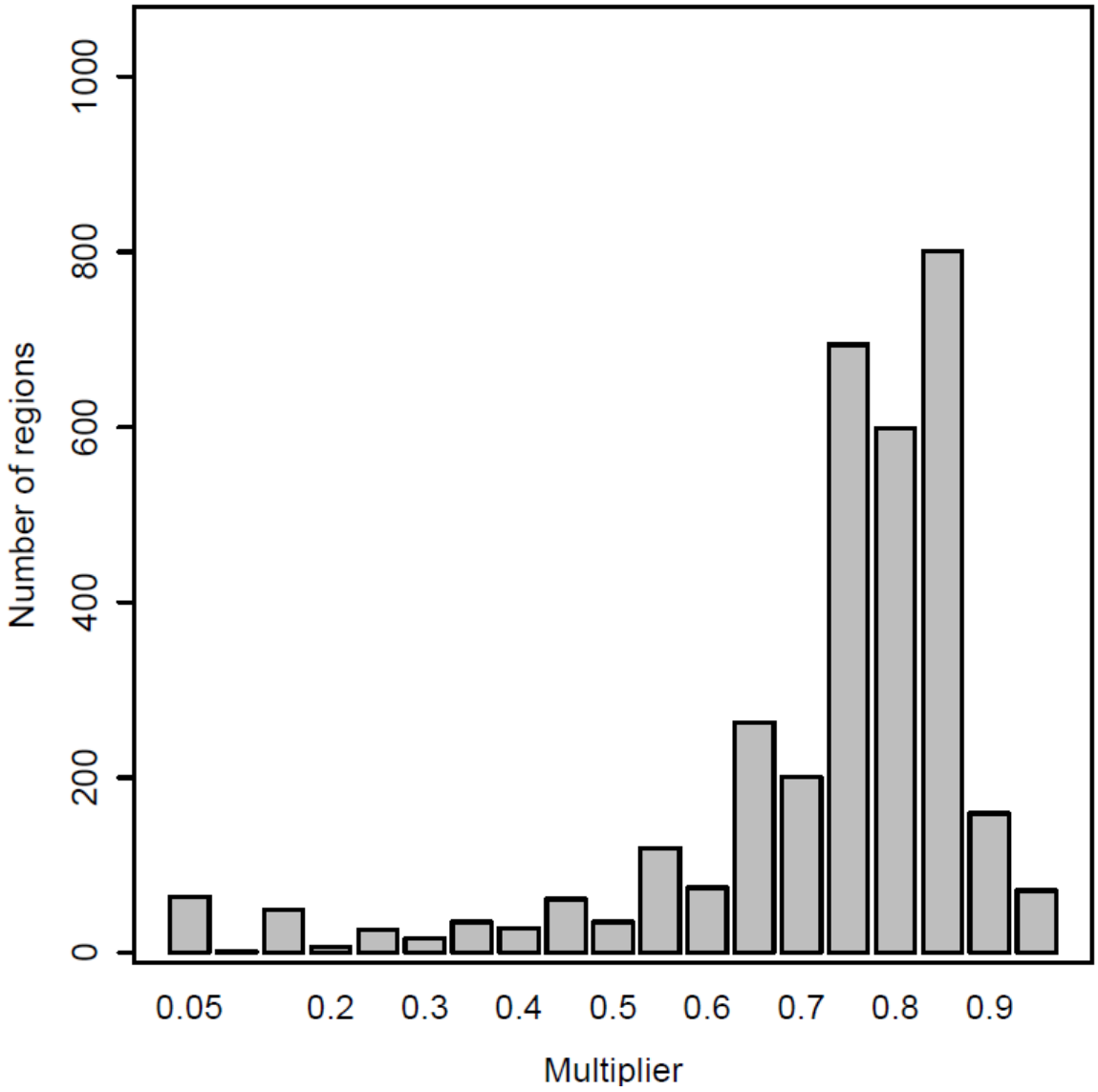

Figure 5 Bar plot of multipliers used to create simulated data for 3,302 enumerated regions

For all enumerated regions, estimates of the relative RMSE for quantile estimates were found using "regsimq" for non-exceedance probabilities 0.001, 0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 0.99, 0.999, and 0.9999. Estimates at non-exceedance probabilities 0.001, 0.01, 0.02, … 0.98, 0.99, and 0.999 are presented for a representative region in Figure 1006. RMSE estimates are lowest for non-extreme probabilities above 0.5, and are larger at

extreme precipitation events (i.e. non-exceedance probabilities closest to 0.0 or to 1.0), with low non-exceedance probabilities registering significantly higher RMSE estimates than high non-exceedance probabilities. All regions generally follow this pattern.

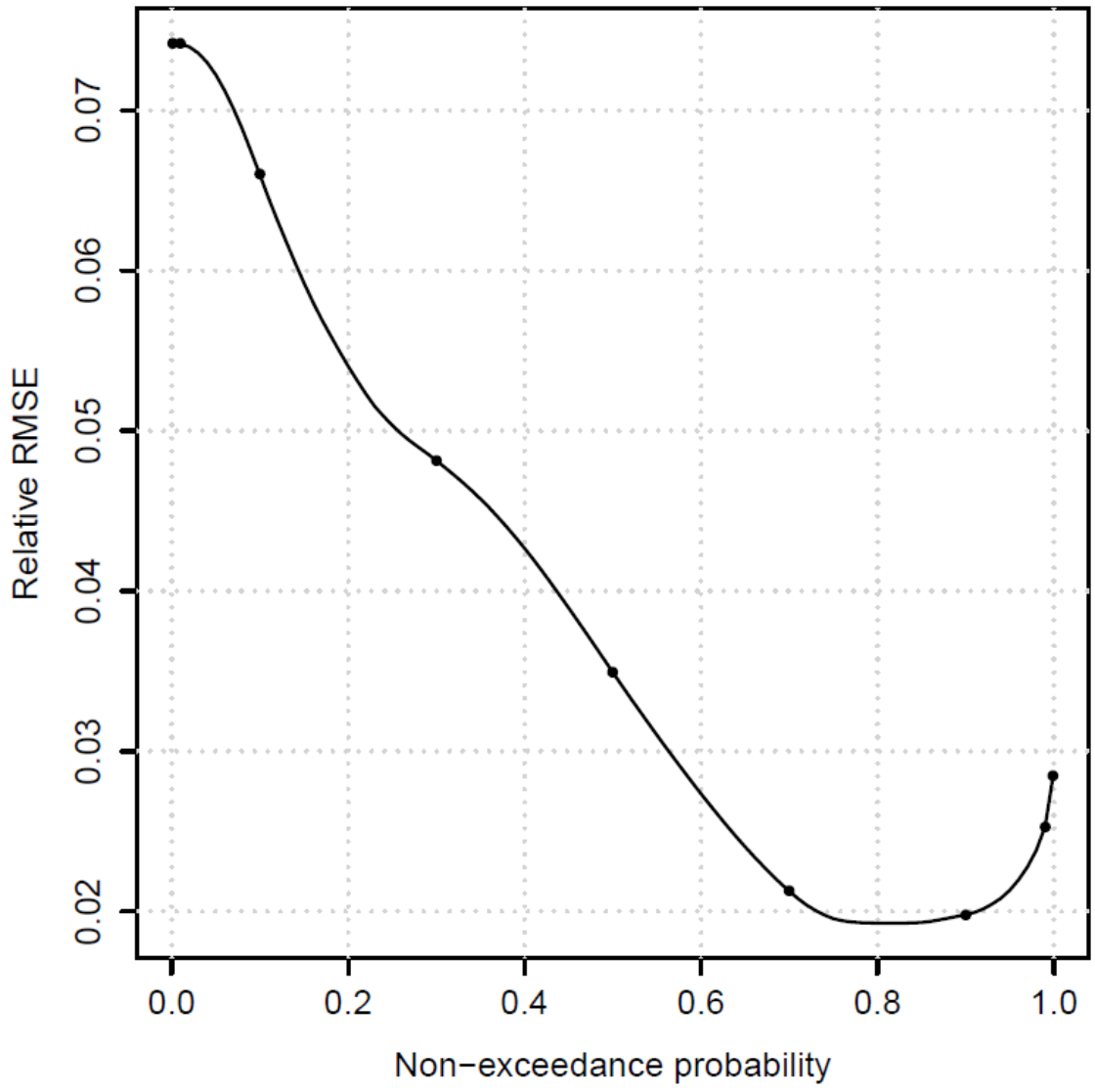

Figure 6 Relative RMSE plotted against non-exceedance probability for 1000 iterations of a representative region. Probabilities sampled in the analysis are marked with dots

Plotting heterogeneity estimates against an estimate for quantile error over all possible regions formed from a gauge network reveals several discrete populations of regions. These populations exhibit different relationships between heterogeneity and error and different patterns of variation in that relationship across non-exceedance frequency and the heterogeneity statistics H_1 to H_2 to H_3.

Regions with multipliers below 0.65 are a subset of regions with negative H_1 for which an unusually small range of at-site L-moment ratio variation was required to match the observed H_1. This procedure, which is a part of the Hosking and Wallis (1997) methodology for finding the error of quantile estimates, obfuscates the relationship between heterogeneity and error for a subset of low-heterogeneity regions, resulting in a diffuse cloud rather than linear correlation. Low absolute values of root mean square error estimates also obfuscate any heterogeneity-error relationship that may be present, but through a different mechanism; as error approaches zero, the relationship between heterogeneity and error grows more linear, but the slope becomes perfectly horizontal and heterogeneity cannot function as a predictor of error. The Pearson's r correlation measure approaches zero in both of these cases and can in fact become negative, for which reason the r^2 statistic was not used. Taking these shortcomings of the method into account, Pearson's r offers a good statistical summary of the degree to which heterogeneity is a “reasonable proxy” of error.

The definitions of these populations, which are based on gauge constituency and the multiplier used to scale at-site L-moment ratios toward the regional mean, may reflect variation in the relationship between heterogeneity and quantile error across regions with

different characteristics. Plots of H_1, H_2, and H_3 against RMSE colored by site- and multiplier-based filters are presented at representative non-exceedance probabilities (Figure7), and Pearson's r for the relationship between the three H statistics and RMSE across representative non-exceedance probabilities is presented in Figure 8 for each of Figure 7's clusters.

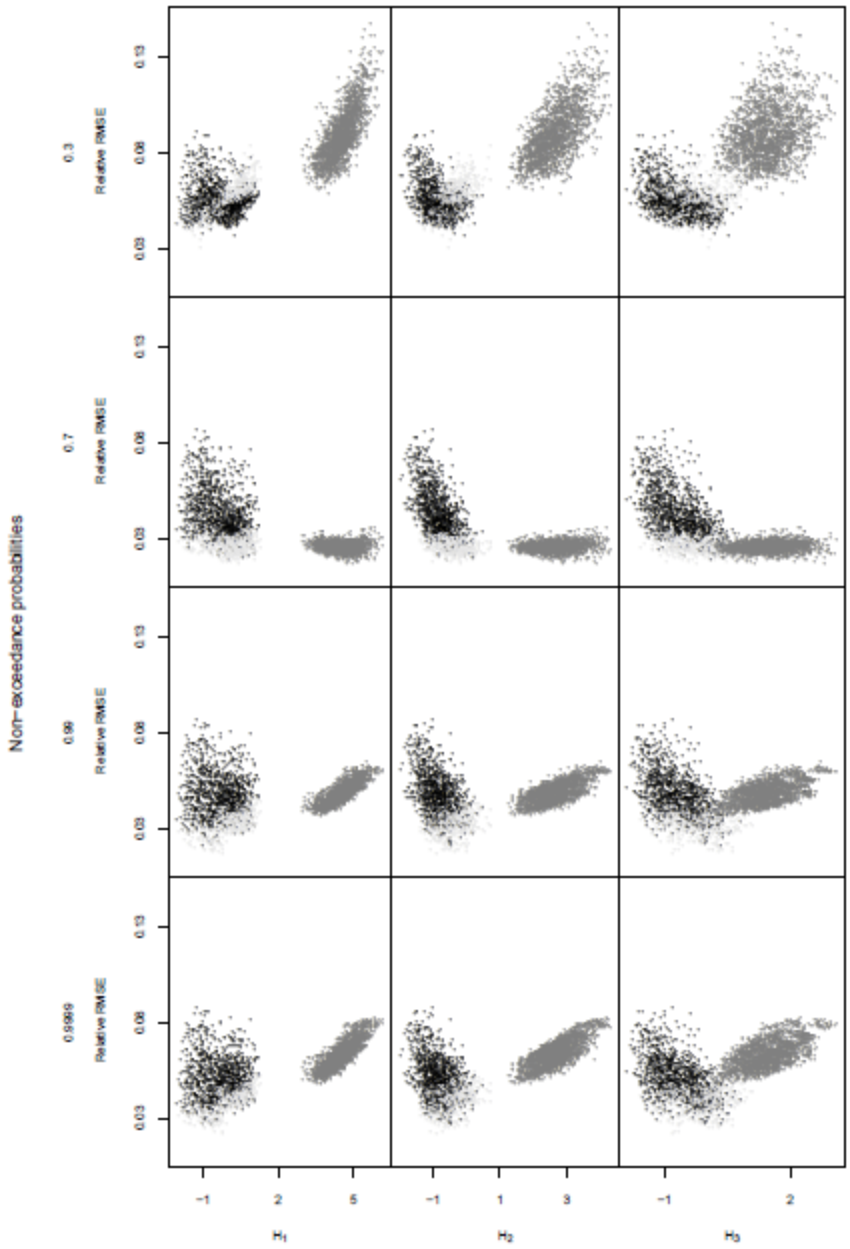

Figure 7 H_1, H_2, and H_3 plotted against relative RMSE. Regions with multipliers less than or equal to 0.6 are colored black; regions including site 46 are colored dark gray; regions with multipliers higher than 0.6 and not containing site 46 are colored light gray

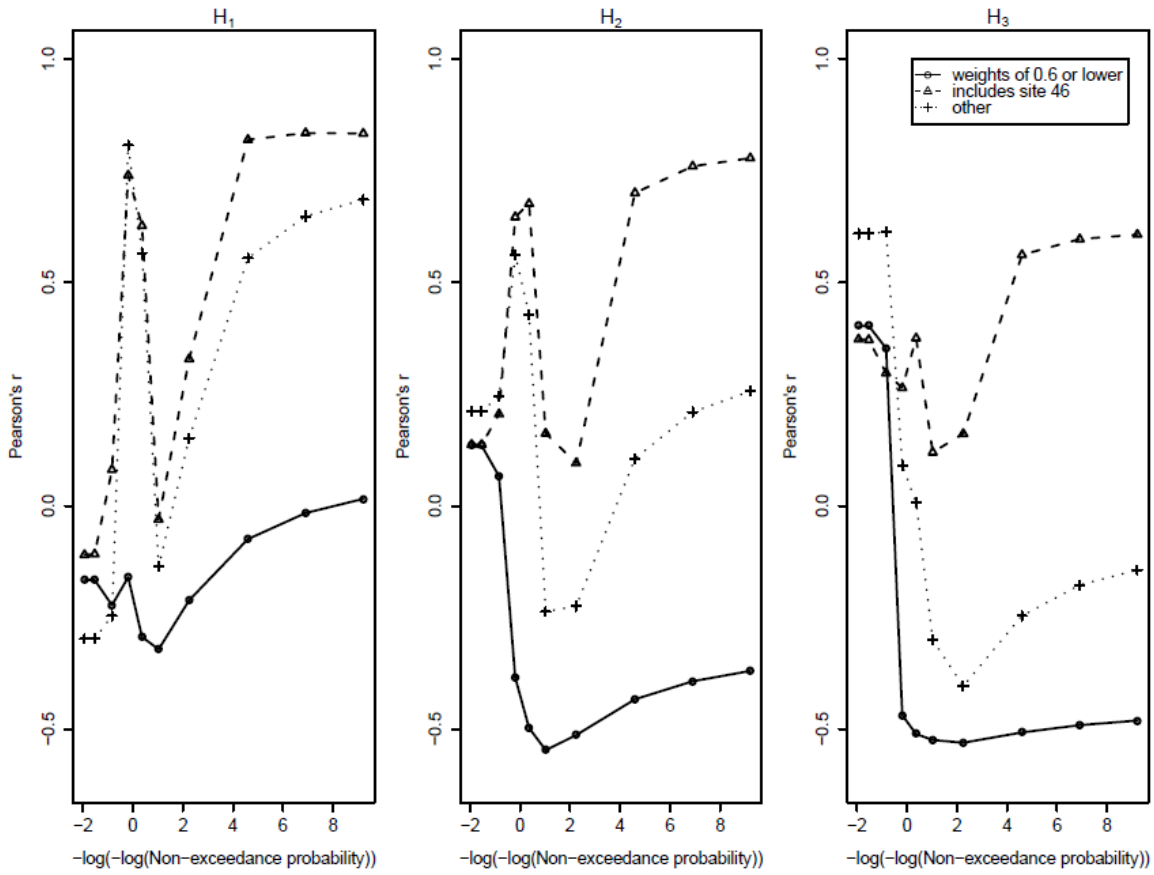

Figure 8 Pearson's r of the linear fit between relative RMSE and H_1, H_2, and H_3.

The distribution of regions in Figure 7 is separated into two groups based on whether site 46 is a member. Site 46 is the gauge with the lowest L-CV, L-skewness and L-kurtosis in the analyzed dataset. In L-moment ratio graphs such as Figure 2 it is an outlier, and regions containing it have higher heterogeneity than regions that do not. As a result, a high-heterogeneity cluster composed of regions containing site 46 appears on Figure 7 and is colored in dark gray. The clusters are most separated for H_1, while for H_3 they slightly overlap.

For most regions, the H-RMSE relationship measured using Pearson's r in Figure 8 becomes more positive as RMSE decreases and non-exceedance probability increases.

For a small population of low-heterogeneity regions not containing site 46, however, Pearson's r becomes more negative as RMSE decreases and non-exceedance probability increases. The group of regions not containing site 46 is therefore split into two clusters that exhibit different patterns of variation; these clusters can be defined by whether the region's fitted multiplier for RMSE simulation is 0.60 or below, or 0.65 or above (black or light grey in Figure7, respectively). The former cluster generally contains lower-heterogeneity regions than the latter, but there is significant overlap. Only regions not containing site 46 had heterogeneities low enough to necessitate low multipliers, creating simulated regions with small variance from the mean.

2.5: Discussion

In the simulation experiment the performance gap between H_2 and $H_1$is relatively narrow, indicating that L-skewness offers a useful amount of heterogeneity information in the presence of L-CV variation. H_2 is consistently a slightly less faithful proxy for error than H_1 across the simulated regions, but like H_1 it can also be plotted against percent RMSE added due to heterogeneity and threshold values can be described as equivalent to a range of percent RMSE added. Analogously to the process used for associating H_1's thresholds to ranges of percent RMSE added, Figure 3 can be visually inspected to estimate the range of added RMSE spanned by simulated data with H_2 near a threshold value. $H_2 = 0.25$ could in this way be described as equivalent to a 10-50% range of RMSE increase and $H_2 = 0.5$ as equivalent to a 30-90% range of added RMSE. Note that H_2 has a less linear relationship with RMSE than H_1, implying that the ranges of RMSE associated with H_2 should be wider than those associated with H_1. Further

supporting these proposed thresholds is the consistent 4:1 relationship between H_1 and H_2 across the simulated dataset (Figure 9).

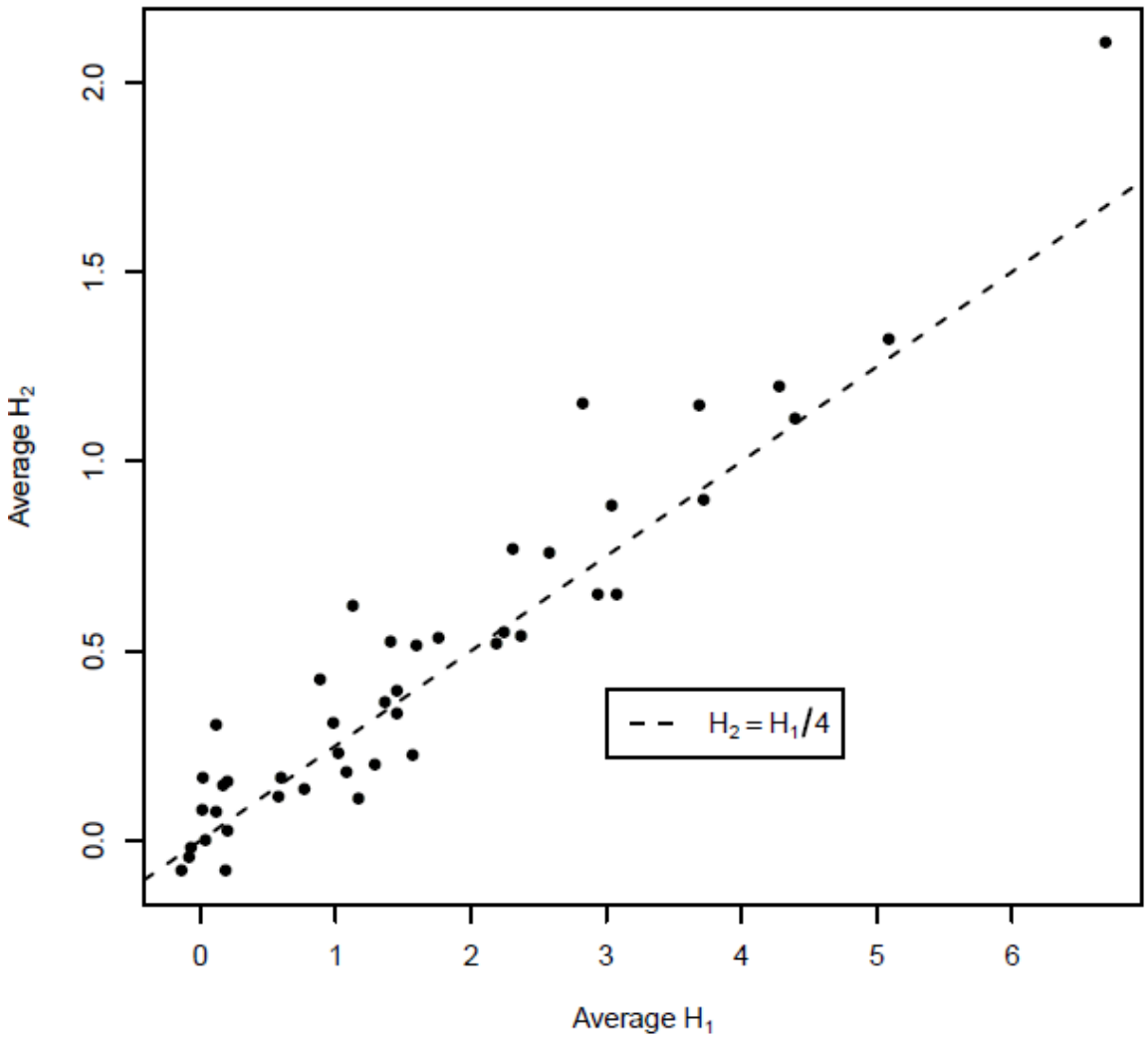

Figure 9 Average H_1 plotted against average H_2 across all simulations.

In the enumeration experiment, the population of regions fitted with low multipliers is characterized by negative H_1 values. This can indicate high between-site

cross-correlation in a region (Hosking and Wallis, 1997), a regional characteristic that could be related to the lesser linearity of the H-RMSE relationship for the population of regions fitted with low multipliers. The inter-site regional correlations calculated using Equations 10-13, however, were similar for regions generated using low and high multipliers. The association between low multipliers and a less positively linear H-RMSE relationship also may be related to the small between-site variation in L-moment ratios for simulated regions created using low multipliers in the Monte Carlo estimation procedure for quantile error.

Regions with higher multipliers lacking site 46 make up a low-RMSE, low-H cluster with characteristics blending those of the low-multiplier and site-46-containing clusters. As the order of the L-moment ratios used increases from H_1 to H_2 to H_3, this cluster transitions in Figure 8 from a site 46 profile toward a low-multiplier profile. All three populations of regions see the relationship between Pearson's r and non-exceedance frequency grow more linear as the H statistic correlated with RMSE increases from H_1 to H_2 to H_3. Pearson's r scores generally increase from H_1 to H_2 to H_3.

The H-RMSE relationship, in general, becomes more linear as RMSE decreases. Regions fitted with multipliers higher than 0.60 tend to exhibit a positive linear correlation between H and RMSE. In this respect, the high-heterogeneity population of regions including site 46 are especially strong. This suggests that at these L-moment ratio magnitudes, the relationship between homogeneity and quantile error of a region becomes less linear at low heterogeneities. The procedures used for estimating error and

heterogeneity may introduce variance to the relationship. Nevertheless, all three H statistics appear to be related to Monte Carlo RMSE estimates.

2.6: Summary and Conclusion

Hosking and Wallis (1997) find that H_1 has power as a proxy of error while H_2 and H_3 do not, but simulation and enumeration studies conducted here paint a more nuanced picture. H_1 remains the favored heterogeneity statistic across simulated and real-world datasets across a wide range of skewness, but H_2 is nearly as effective. The efficacy of H_2 has been obscured through the application of thresholds constructed with reference to the linear relationship between H_1 and percent RMSE added, which has a higher slope than the H_2-percent RMSE added relationship.

Both the enumeration experiment on real data, which compared heterogeneity statistics to estimated RMSE, and the simulation experiment, which compared heterogeneity statistics to percent RMSE added due to heterogeneity, found linear relationships between H and RMSE. The enumeration experiment reported total error without isolating the error due to heterogeneity. For the enumeration experiment, other error components not related to heterogeneity may have played a role in lowering Pearson's r scores and introducing multiple populations of regions with different linear relationships to estimated RMSE.

The enumeration experiment used quantile error estimation procedures adapted from Hosking and Wallis (1997), in which regional variance in L-CV was replaced with linearly varying values spanning a smaller range than the sample data. Linear variation in one L-moment ratio was replaced with shrinkage multipliers for each at-site value of all

L-moment ratios, creating a region more comparable to the original sample data. The adapted method can be described as an application of shrinkage theory.

H_2's consistently lower numerical values cause poor performance in significance tests conducted using thresholds based on H_1, specifically the H = 2 heterogeneity threshold, but the linear relationship between H_2 and percent RMSE added is nearly as striking as that for H_1 and offers justification for analogous H_2-specific thresholds. The thresholds derived for H_2 above are one possible alternative, but H_2's inferior linearity with respect to error makes H_2 thresholds less reliably indicative of a given range of error added than H_1 thresholds. For this reason, the performance of an H_2 threshold at separating heterogeneous from homogeneous regions is not likely to match the performance of H_1 thresholds. However, an H_2 threshold derived from the H_2-percent RMSE added relationship is much more likely to offer acceptable performance than an H_2 threshold derived from the H_1-percent RMSE added relationship.

This study supports the use of H_1 as the primary heterogeneity estimator and quantifies the degree of its superiority over H_2 and H_3. Although H_2 is less linear with respect to error than H_1, it also conveys heterogeneity information and with appropriate thresholds can offer a supplementary indication of regional heterogeneity. H_2 offers a great deal of linearity with respect to error due to heterogeneity and is therefore a useful supplementary measure to H_1. H_1 need not stand alone representing regional heterogeneity.

2.7: Acknowledgements

Data were provided by the State Climatology Office, Minnesota Department of Natural Resources - Division of Ecological and Water Resources. This work also used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. Financial support provided by the George Mason University Presidential Scholarship is gratefully acknowledged. The authors also wish to thank Jason Giovannettone for his assistance.

3: THE RELATIONSHIP BETWEEN MONTE CARLO ESTIMATORS OF HETEROGENEITY AND ERROR FOR DAILY TO MONTHLY TIME STEPS IN A SMALL MINNESOTA PRECIPITATION GAUGE NETWORK

Abstract

Precipitation quantile estimates are used in engineering, agriculture, and a variety of other disciplines. Index flood regional frequency methods pool normalized gauge data, assuming homogeneity among the constituent gauges of the region, and output unitless regional quantile estimates which are rescaled at each gauge. Because violation of the homogeneity hypothesis is a major component of quantile estimation error in regional frequency analysis, heterogeneity estimators should be "reasonable proxies" of the error of quantile estimation. In this study three Monte Carlo heterogeneity statistics tested in Hosking and Wallis (1997) are plotted against Monte Carlo estimates of quantile error for all five-or-more-gauge regionalizations in a twelve-gauge network in the Twin Cities region of Minnesota. Upper-tail quantiles with non-exceedance probabilities of 0.75 and above are examined at time-steps ranging from daily to monthly. A linear relationship between heterogeneity and error estimates is found and quantified using Pearson's r score. Two of Hosking and Wallis (1997)'s heterogeneity measures, incorporating the coefficient of variation in one case and additionally the skewness in the other, are found to be reasonable proxies for quantile error at the L-moment ratio values characterizing

these data. This result, in addition to confirming the utility of a commonly used coefficient of variation-based heterogeneity statistic, provides evidence for the utility of a heterogeneity measure that incorporates skewness information.

3.1: Introduction

Precipitation quantile estimation is important in a variety of fields, including civil engineering and agriculture. Estimation of upper quantiles is particularly important for evaluating the likelihood of major flood events, for which stream flow gauges are also used. Regardless of data source, accurate statistical estimation of uncommon extreme events requires large sample size for acceptable accuracy. Statistical methods which can improve the utility of low-sample-size data help analysts tackle these problems.

Linear moments (Greenwood et al., 1979) are often used as parameters for probability distributions by statistical hydrologists in place of the conventional moments due to the lower bias of their sample estimators, especially in cases of small sample size or high skew, and their lack of algebraic bounds such as those depending on sample size for conventional moment estimators (Vogel and Fennessey, 1993). They are more robust to outliers and offer more reliable inferences about the identity and parameterization of an underlying probability distribution from small sample size (Hosking, 1990).

Dalrymple (1960) established the index flood method, in which gauge records are normalized by an index such as the mean and pooled. From these regional data, unitless regional frequency estimates are made which are multiplied by the at-site indices to produce frequency estimates for each gauge. This assumes that after normalization, quantile estimates for all gauges in a region can be described by a single frequency

distribution. The degree to which this homogeneity assumption is violated is estimated by heterogeneity statistics – generally the homogeneity assumption is considered to be valid if estimated regional heterogeneity is below some threshold.

Regional frequency analysis using linear moments, sometimes abbreviated RFA-LM e.g. Núñez et al. (2011) and Maeda et al. (2013), uses the index flood approach and therefore requires the use of a validated heterogeneity statistic. Hosking and Wallis (1993) proposed three heterogeneity statistics, denoted H_1, H_2, and H_3 or "the H statistics" in this study. The use of H_1 was preferred, based on simulation experiments demonstrating that H_1 was a "reasonable proxy" for error. These statistics, most notably H_1, have been used to defend the utility of candidate regionalizations in several related subfields of hydrology.

For rainfall gauge data, monthly (Núñez et al., 2011) and annual precipitation totals (Guttman, 1993; Lin and Chen, 2006; Parida and Moalafhi, 2008; Modarres and Sarhadi, 2011; Dikbas et al., 2012) are among the time steps and recurrence intervals to which the H statistics have been applied. They have also been used to defend regionalizations of gauges for annual maximum rainfall over assumed durations of 5-120 min (Adamowski et al., 1996), 5 min – 24 hrs (Alila, 1999; Smithers and Schulze, 2001), 1-24 hours (Norbiato et al., 2007; Um et al., 2010), 24 hours (Huff and Angel, 1992; Szolgay et al., 2009; Gabriele and Chiaravalloti, 2013), 1-72 hrs (Bradley, 1997), and 1-7 days (Kyselý et al., 2007; Kyselý and Picek, 2007; Yang et al., 2010; Ngongondo et al., 2011).

The regionalization of stream flow gauge data has been defended using the H statistics for cases including annual peak discharge or maximum flood (Zrinji and Burn, 1996; Burn and Goel, 2000; Kjeldsen et al., 2002; Jingyi and Hall, 2004; Abida and Ellouze, 2006; Atiem and Harmancioğlu, 2006; Rao and Srinivas, 2006; Srinivas et al., 2007; Noto and La Loggia, 2009; Gaume et al., 2010; Guse, 2010; Saf, 2010; Hussain, 2011; Kar et al., 2012; Rianna et al., 2012; Seckin et al., 2013), partial duration series for different numbers of peaks in a year (Pham et al., 2013), and annual 7-day minimum flow (Modarres, 2008; Dodangeh et al., 2013).

Santos et al. (2011) applied the H statistics to the Standardized Precipitation Index (SPI), a drought index calculated using annual maximum and peak-over-threshold precipitation data, while Feng et al. (2013) applied them to maximum daily precipitation in the three days after a drought. For gridded precipitation data the H statistics have been used in regional frequency analyses of annual totals (Satyanarayana and Srinivas, 2009) and for maximum annual precipitation at 1-5 day time steps (Marx and Kinter, 2007).

Other homogeneity tests include the Lu-Stedinger test (Lu and Stedinger, 1992), used in Gaál and Kyselý (2009), Alila's S test (Alila, 1994) used in Alila (1999), and the K-sample Anderson-Darling test (Scholz and Stephens, 1987), which is ascertained as superior to H_1 for highly skewed simulated data by Viglione et al. (2007). Chebana and Ouarda (2007) extend the Hosking and Wallis homogeneity test to the multivariate case, as applied in Sadri and Burn (2011), while Castellarin et al. (2008) propose a correction to the test accounting for cross-correlation between gauge records. Bhuyan et al. (2010)

compares L-moments to the LH-moments of Wang (1997) for annual maximum flood data.

No single investigation can evaluate the utility of the H statistics for all the data types to which they have been applied. However, daily precipitation gauge data can be aggregated to longer time steps, allowing for the characterization of the H statistics' performance across data with a wide range of statistical properties. The limited size and spatial extent of the gauge network used in this study minimize the effects of climatic variation due to distance and topography. The study area's small spatial and wide temporal scope allows the relative performance of the H statistics to be assessed across wide ranges of variance, skew, and kurtosis while controlling for many of the sources of unpredictability in real data.

The remainder of the paper is structured as follows. First, the study area and the data set are explicated in Section 2. Next the steps of the analysis are described, beginning with the aggregation of daily data into longer time steps in Section 3.1 and continuing to a discussion of L-moment theory in Section 3.2. In this section formulae for the linear moments are introduced (Section 3.2.1), followed by methods for estimating a candidate regionalization's heterogeneity (Section 3.2.2), choosing and fitting a probability distribution for a candidate regionalization (Section 3.2.3), and estimating error (Section 3.2.4). Section 3.3 describes the method by which the H statistics are validated through comparison to estimates of quantile error. Section 4 contains the results of the analysis, starting with a description of the variance in L-moment ratios and the number of data points equal to zero as time step is varied (Section 4.1). Next, the

probability distributions to which regionalizations of each time step were fitted are presented in Section 4.2, after which the correlations between heterogeneity and error estimates across time steps and non-exceedance frequencies are presented in Section 4.3. Finally, Section 5 offers conclusions.

3.2: Data

The data used in this study, comprising daily precipitation totals measured by volunteers across Minnesota, were collected and quality controlled for inclusion into Perica et al. (2013). This process winnowed the volunteer high-density gauge network to 341 gauges, 57 of which did not skip a full month between the start and end of the data record. Twelve such gauges were present in the Minneapolis-St. Paul or Twin Cities region (Figure 10, Table 4). These gauges provide thousands of measurements from four decades while spanning slightly more than a quarter of a degree in latitude and longitude. In the remainder of this paper, the relationship between Monte Carlo estimates of heterogeneity and quantile error of estimation is investigated for all 3,302 valid (five- or more-gauge) regionalizations within this dense gauge network.

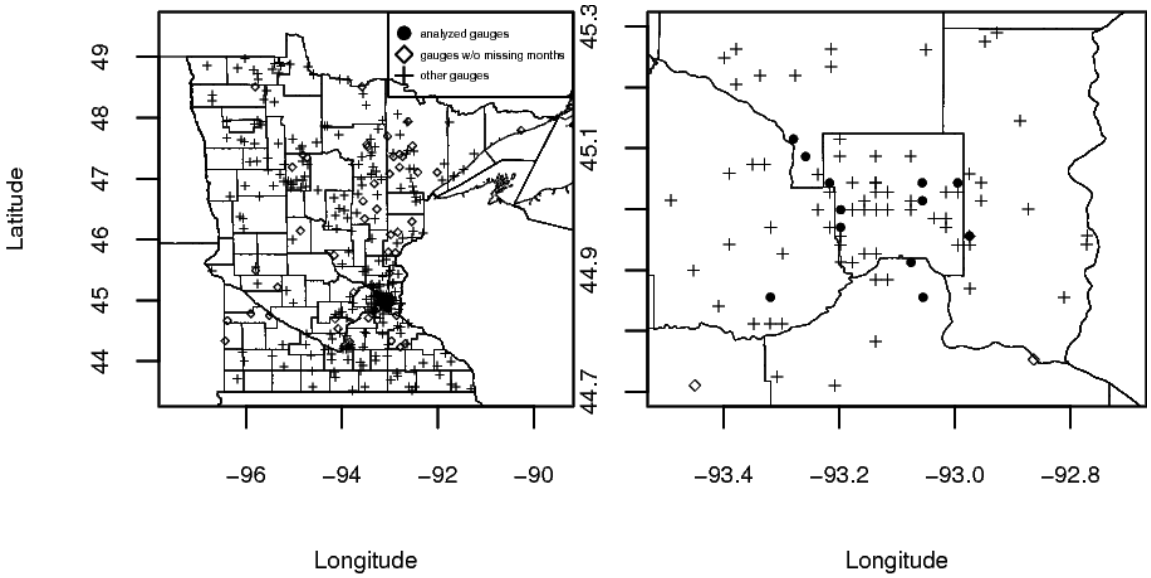

Figure 10 Spatial extent of selected gauges, gauges without missing months, and remainder of dataset.

Table 4 Precipitation gauge dataset

Gauge ID #	# days	Start date	End date	Latitude	Longitude
11	1714	Feb 1982	Feb 1996	44.86	-93.32
35	1191	Jan 1999	Jan 2009	45.01	-93.06
39	2419	Jan 1978	Aug 1998	45.04	-92.99
46	2765	Dec 1980	Nov 2008	45.12	-93.28
78	1518	Dec 1975	Apr 1990	44.97	-93.20
104	3751	Jan 1976	Nov 2008	45.09	-93.26
149	1488	Aug 1993	Nov 2008	44.86	-93.05
150	1787	Jul 1993	Nov 2008	44.91	-93.08
266	1859	May 1992	Oct 2008	45.00	-93.20
268	1438	Sep 1990	Nov 2004	45.04	-93.06
272	1134	Sep 1998	Nov 2008	45.04	-93.22
328	3572	Sep 1979	Dec 2007	44.96	-92.97

3.3: Analysis

3.3.1: Aggregation of higher time steps from daily precipitation totals

Daily precipitation totals were aggregated across time steps ranging from one to thirty-five days; intervals that included one or more days of missing record were eliminated. Data were collated for several starting points within each time step and all possible starting points for selected time steps. For example, all seven starting points for the seven day time step were analyzed. For short time steps a significant proportion of the

data consists of time steps with zero recorded precipitation; these "true zeroes" were preserved in the analysis.

3.3.2: Regional Frequency Analysis using Linear Moments (RFA-LM)

Hosking and Wallis (1997) outlined an index-flood regionalization approach to hydrological quantile estimation, denoted here as RFA-LM. The mean of at-gauge or at-site data is used to normalize the data prior to regional pooling. After a candidate regionalization's degree of deviation from the assumption of regional homogeneity is found to be below a certain threshold, a probability distribution is chosen and fitted to the L-moment ratios of the normalized regional data. A unitless regional 'growth curve' of quantile estimates across a range of non-exceedance probabilities is multiplied by each gauge's at-site mean to produce at-site quantile estimates. The statistical package "lmomRFA", which is written in the programming language R (R Core Team, 2012), is used to implement this method.

Regionalization and L-moment analysis according to the methods of Hosking and Wallis has been applied to Minnesota for totals across durations from one month to five years (Werick, 1994), for daily data (Huff and Angel, 1992), and for five-minute to sixty-day durations (Perica et al., 2013). Frequency estimates from the latter study with upper and lower bounds of a 90% confidence interval are available at http://hdsc.nws.noaa.gov/hdsc/pfds/.

3.3.3: Linear Moments

Hosking and Wallis (1997) define the L-moments λ_{r+1} in Equation 15 for k from 0 to r:

Equation 15

$$\lambda_{r+1} = \sum_{k=0}^{r} p^*_{r,k}\beta_k$$

Where $p^*_{r,k}$ is defined in Equation 16:

Equation 16

$$p^*_{r,k} = \frac{(-1)^{r-k}(r+k)!}{(k!)^2(r-k)!}$$

And β_k is the probability weighted moment, defined in Equation 17 for the quantile function x(u):

Equation 17

$$\beta_k = \int_0^1 x(u)u^r du$$

The sample estimate of β_r is b_r, estimated using Equation 18 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:

Equation 18

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

The L-moments of particular concern are usually the first four: λ_1 is the mean of a data record, λ_2 measures the dispersion, λ_3 measures skewness, and λ_4 measures kurtosis. From these L-moments, L-moment ratios are constructed; the L-CV or coefficient of L-variation is $\tau = \lambda_2/\lambda_1$, while L-skewness and L-kurtosis are calculated using r = 3 and 4 respectively in Equation 19:

Equation 19

$$\tau_r = \lambda_r / \lambda_2$$

The sample estimate of λ_r is l_r and the sample estimate of τ_r is t_r with the sample estimate of the L-CV denoted as t; these are calculated using equations 1-5, substituting b_r for β_r.

3.3.4: Heterogeneity statistics

Heterogeneity statistics quantify the degree to which a candidate regionalization deviates from the hypothesis of homogeneity – identical L-moment ratios at all of the candidate regionalization's constituent gauges. Three statistics defined by Hosking and Wallis (1997), denoted H_1, H_2, and H_3 are used in this study. Following the method in that monograph, pooled gauge data are fitted to the four-parameter Kappa distribution and simulated regions with the same number of sites and record lengths as the real data are generated. Equations for the four-parameter Kappa distribution can be found in the Appendix of Hosking and Wallis (1997).

A measure V of between-site L-moment ratio variation is calculated for the regional data and for each of many iterations of each simulated region. Three variants of V and therefore of H are identified. H_1 calculates V from the second L-moment ratio, while H_2 considers the second and third and H_3 the third and fourth. The (i) superscript indicates the at-site L-moment ratio value for gauge i; the R superscript indicates the regional average L-moment ratio. N is the number of sites in the region and n_i is the number of data points at site i. Equations 20-23 are used to calculate V and H for the

three variants. In Equation 23 μ_V is the mean of V across all simulated regions and σ_V is the standard deviation.

Equation 20

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

Equation 21

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

Equation 22

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

Equation 23

$$\mathrm{H} = \frac{V - \mu_V}{\sigma_V}$$

Hosking and Wallis (1997) calculated the ratio between root mean square errors (RMSEs) of simulated regions with heterogeneous linear moments and their equivalent homogeneous regions and found a positive linear relationship between error added due to heterogeneity and H_1. H_1 was therefore considered a "reasonable proxy" of the error associated with heterogeneity. H_2 and H_3 were not found to possess this quality. The authors recommended the use of H_1, with $H_1 > 2.0$ indicating heterogeneity and $H_1 < 1.0$ indicating homogeneity. Intermediate values of H_1 between 1.0 and 2.0 indicate possible heterogeneity.

3.3.5: Choice of a distribution

Three-parameter probability distributions, which are fitted to the mean and the second and third L-moment ratios, cannot model all L-skewness/L-kurtosis combinations. For a given L-skewness (the third parameter of these probability distributions) there can be only one L-kurtosis; L-kurtosis as a function of L-skewness, a relationship unique to each three-parameter probability distribution, can be plotted against the L-skewness and L-kurtosis of data to ascertain which distribution fits the data most accurately. In this way, linear moment ratios can be plotted on an L-skewness/L-kurtosis plot as a graphical aid in fitting distributions (Vogel and Fennessey, 1993).

Hosking and Wallis (1997)'s Z statistic is a Monte Carlo statistic based on the difference between a candidate regionalization's average L-kurtosis and the expected L-kurtosis of a three-parameter distribution fitted to the candidate regionalization's average mean, L-CV, and L-skewness. This quantifies the relationship used in graphical estimation and provides a reasonable measure of goodness of fit for three-parameter distributions.

Simulated data are generated from a given three-parameter distribution after fitting that distribution to the candidate regionalizations' average L-moments. The Z statistic is calculated from the bias B_4 and the standard deviation σ_4 of L-kurtosis, t_4. Z scores for a three-parameter distribution DIST, Z^{DIST}, are calculated using L-kurtosis for each gauge m within a region, $t^{(m)}_4$, and regional average L-kurtosis, t^R_4, across N_{sim} simulations, as well as the L-kurtosis of each fitted distribution τ^{DIST}_4, in equations 24-26:

Equation 24

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

Equation 25

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

Equation 26

$$Z_{DIST} = \frac{\tau_4^{DIST} - t_4^R + B_4}{\sigma_4}$$

The distribution with the lowest Z score provides the best fit to the data. The "somewhat arbitrary" convention of accepting the hypothesized distribution when Z^{DIST} has absolute value less than 1.64 is derived from the assumption of a standard Normal distribution, representing a 90% confidence level. However, the assumption of homogeneity underlies the assumption of Normality, undermining the justification for this criterion so that it is "not recommended as a formal test". Equations for τ^{DIST}_4 for five three-parameter distributions, the Generalized logistic, Generalized Pareto, Generalized extreme-value, Pearson type III, and Lognormal, are available in table A.3 of the Appendix of Hosking and Wallis (1997).

3.3.6: Estimation of quantile error

Hosking and Wallis (1997) establish a Monte Carlo simulation routine that estimates the root mean square error of the regional quantile estimate for a given quantile. The method requires not only the true L-moment ratios of the regionalized gauges but

also a simulated region, based on the true data but with less L-moment ratio variation. Observed data incorporates sampling error, which increases L-moment ratio variation; using the unadjusted observed data as the basis for Monte Carlo generation of regions would result in simulated regions with inaccurate, high sampling errors. The average observed L-moment ratio values are preserved in this method, but variation around these values is reduced to below the observed level. Theoretical justification for scaling variation toward the mean can be found in the literature discussing shrinkage estimators, beginning with Stein (1956), who showed that the multivariate sample mean is inadmissible. Lindley and Smith (1972) discuss shrinkage estimators in a Bayesian context.

The degree to which observed L-moment variation is dampened toward the regional mean in the simulated region is determined by testing a multiplier of 0.75 then adding or subtracting 0.10 depending on whether this simulated region generates lesser or greater heterogeneity, measured using H_1, than observed data from the constituent gauges of the candidate regionalization. Coefficient of L-variation, L-skewness, and L-kurtosis values at each site are all multiplied by the same multiplier. The multiplier is increased or decreased by 0.10 until the observed regional heterogeneity is passed, at which point the multiplier halfway between the two whose heterogeneities bracket the observed regional heterogeneity is also tested. Among these three multipliers, the multiplier generating the heterogeneity closest to true regional heterogeneity is selected.

Once a simulated region with a known fraction of the observed L-moment variation has been defined, quantile estimates at a range of non-exceedance frequencies F

are calculated for M randomized instantiations of this region, $\widehat{\mathrm{Q}}_i^{[m]}(F)$. Relative root mean square error is calculated for each frequency F across N gauges and M simulations with reference to the quantiles present in the observed data $\mathrm{Q}_i\ (F)$ using Equation 27:

Equation 27

$$\mathrm{R}^R(F) = N^{-1}\sum_{i=1}^{N}\sqrt{M^{-1}\sum_{m=1}^{M}\left(\frac{\hat{Q}_i^{[m]}(F) - Q_i(F)}{Q_i(F)}\right)^2}$$

3.3.7: Evaluation of heterogeneity-error relationship across time steps

In this study the "reasonable proxy" relationships between heterogeneity statistics and error are quantified by plotting, for all analyzed time steps and starting points, each of the H statistics against estimated relative RMSE for the 3,302 valid regionalizations of the twelve-gauge network. The multiplier described in Section 3.2.4 is used as a threshold to identify a population of regionalizations with very low multipliers characterized by lack of a linear relationship between heterogeneity and error. This anomalous population, whose heterogeneity-error relationship may be obfuscated by the near-homogeneity of the simulated regions used to estimate error, represents a small minority of all possible regionalizations.

Because the large majority of regionalizations with high multipliers are observed to form linear clusters when quantile error of estimation is plotted against each of the H statistics, the Pearson's r statistic is a reasonable summary of the degree to which heterogeneity statistics are proxies of error. Pearson's r is then calculated for this

population of regionalizations, indicating the degree to which a positive linear correlation exists between heterogeneity and error across the analyzed data set after accounting for limitations of the error estimation routine.

For two vectors x and y with n entries each whose mean values are $\bar{x}$ and $\bar{y}$, Pearson's r is calculated using Equation 28:

Equation 28

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

3.4: Results

3.4.1: Aggregation of daily data

The method of aggregation described in Section 3.1 is applied to twelve Twin Cities precipitation gauges with daily records. Because the starting point of "Day One" is arbitrary (whether determined by reference to a calendar or, as here, starting with the first day of recorded data in the oldest of the twelve analyzed gauges), all starting points have equal relevance. At-site L-moment ratios are calculated for all possible starting points and time steps from one to thirty days in length. In Figures 11 and 12 the solid line indicates the average across all twelve gauges and all possible starting points for the time step on the x axis; the dotted lines indicate the maximum and minimum value at any of the gauges and starting points. The fraction of time steps without measured precipitation, denoted "true zeroes" here, decreases quickly as time step increases (Figure 11).

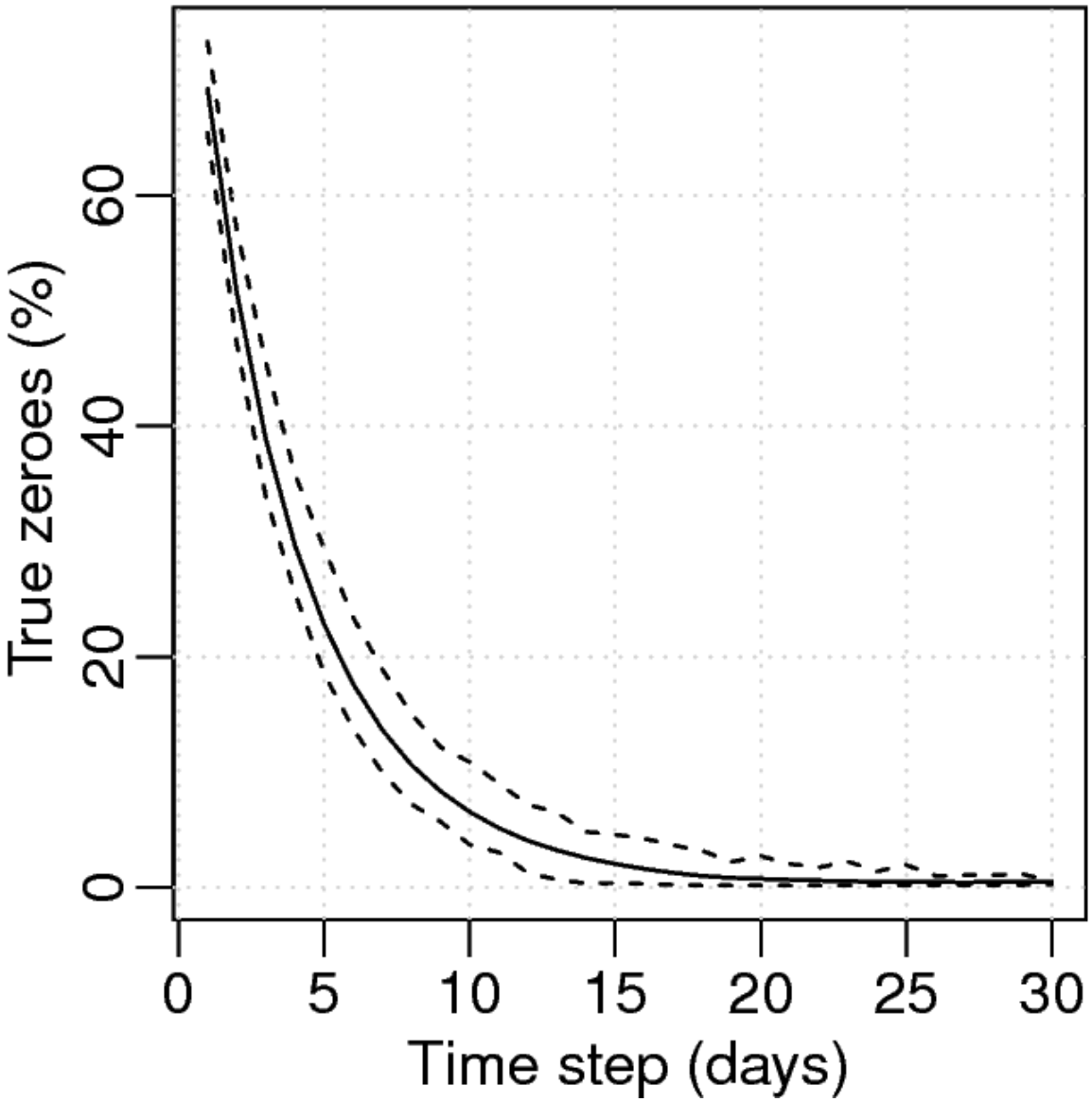

Figure 11 Percentage of intervals in which no precipitation was recorded. Solid line is mean of non-zero percentages across all sites and starting points, dashed lines are maximum and minimum percentage at any site or starting point

Generally, an inverse relationship is observed between length of time step and magnitude of linear moment ratios. Sample size decreases as length of time step increases; decreasing sample size may also drive an increase in variation around the average. The Twin Cities of Minnesota present a wide range of L-moment ratio values

across the spectrum from daily to monthly precipitation aggregates. The coefficient of L-variation ranges from 0.4 to 0.9, L-skewness from 0.8 to 0.2 and L-kurtosis from 0.6 to below 0.1 (Figure 12).

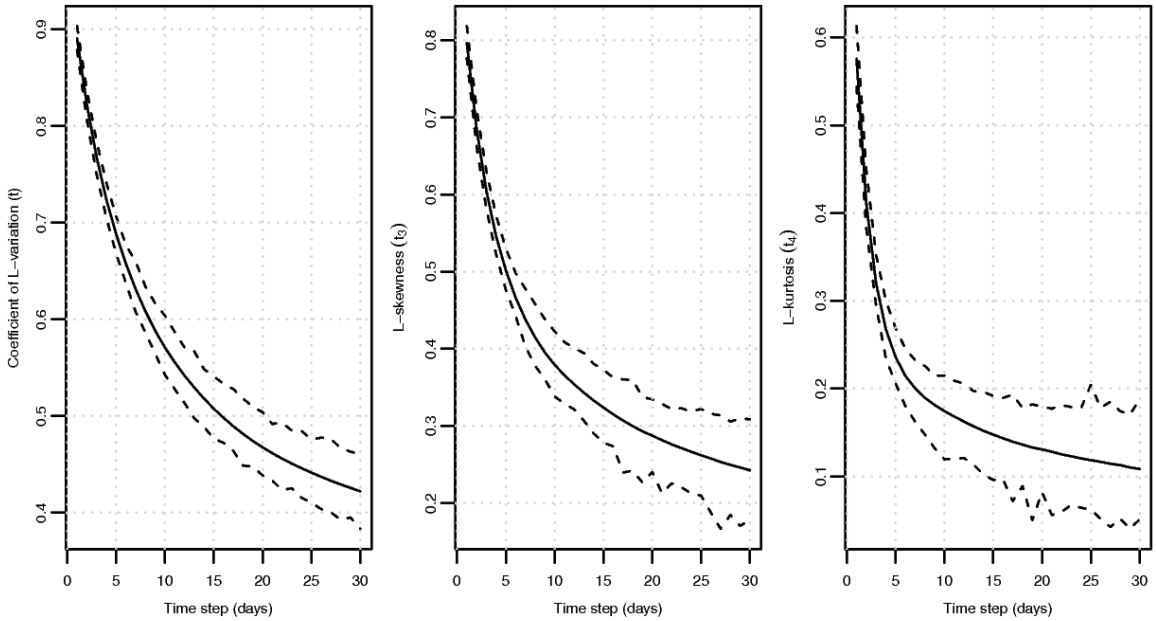

Figure 12 (a) Coefficient of L-variation (t). (b) L-skewness (t_3). (c) L-kurtosis (t_4). Solid line is mean across all sites and starting points, dashed lines are maximum and minimum value at any site or starting point

3.4.2: Fitting a distribution

Average L-moment ratios having been calculated for each 5-or-more-site regionalization of the twelve gauge records at each time step and starting point, an appropriate distribution can be fit to each regionalization. The Z score method of Section 3.2.3 is used for this purpose. Along with heterogeneity and error results, Z scores are presented in this study at the starting point of Day One for time steps one through nineteen, the even numbers from 20 to 30 inclusive, and the 35-day time step. Results are

presented at all starting points for time steps seven and fourteen. Other starting points and time steps were evaluated through the same methods and produced similar results. As Figure 13 illustrates, the L-kurtosis/L-skewness relationship at these gauges is similar to that produced by the Pearson type III distribution for less than fourteen days and to the Generalized Pareto for longer time steps.

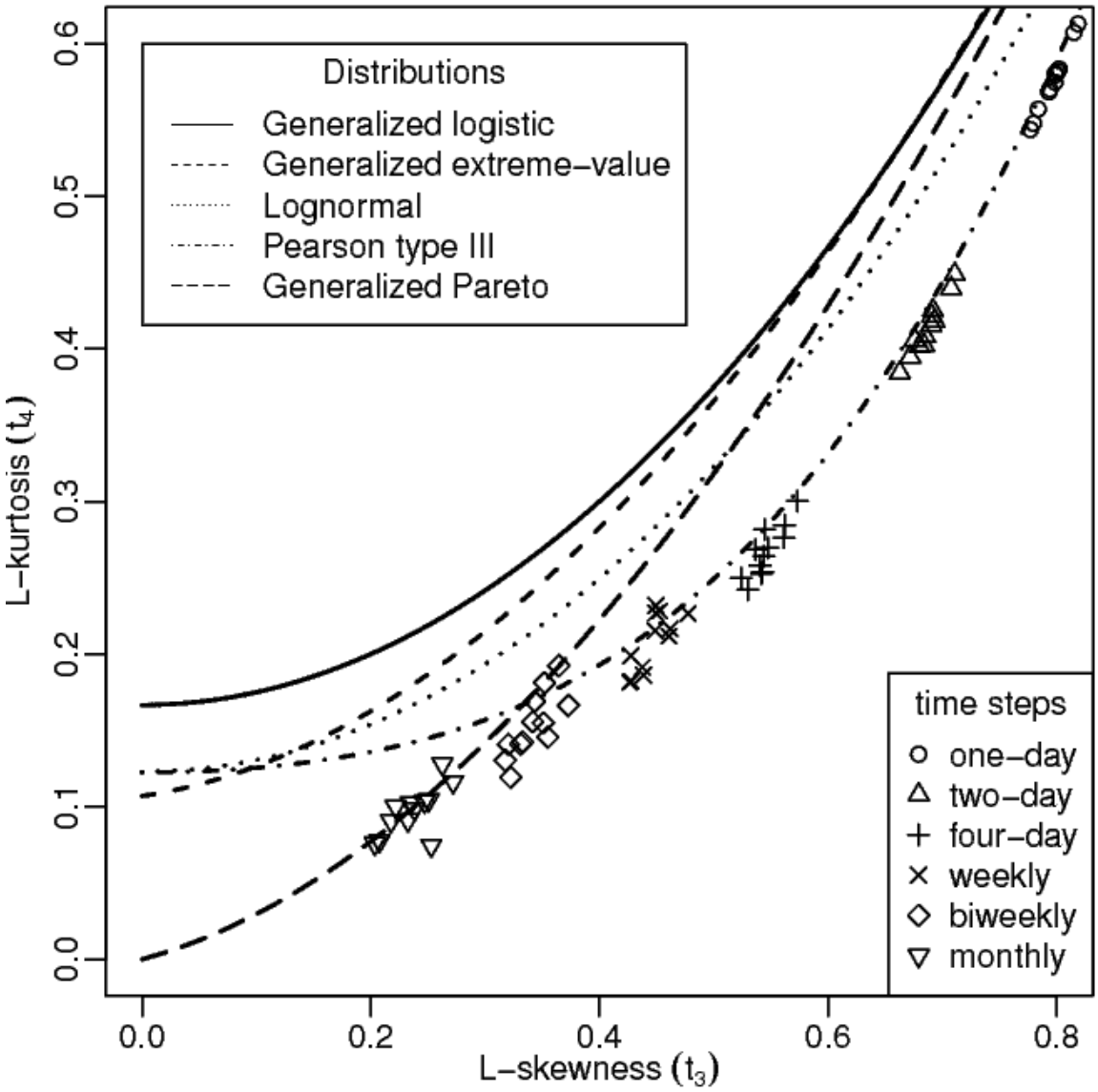

Figure 13 Graphical juxtaposition of observed data and five three-parameter probability distributions at a starting point of Day One. Points: L-skewness and L-kurtosis (t3 and t4) of analyzed gauges at selected time steps. Lines: L-kurtosis outputted by three-parameter distributions as a function of L-skewness

Shifting the starting point for a time step can result in markedly different patterns across an L-kurtosis/L-skewness graph. Figure 14 illustrates that for a 7-day or weekly time step, the gap between two clusters of sites for a starting point of Day One disappears when using a starting point of Day 4. Conversely, the 14-day or biweekly time step has a densely packed cluster of sites for a starting point of Day One; an outlier site appears when the starting point is shifted to Day 7. These patterns, although they are the result of random variation, can have the effect of changing the distribution that best fits the data.

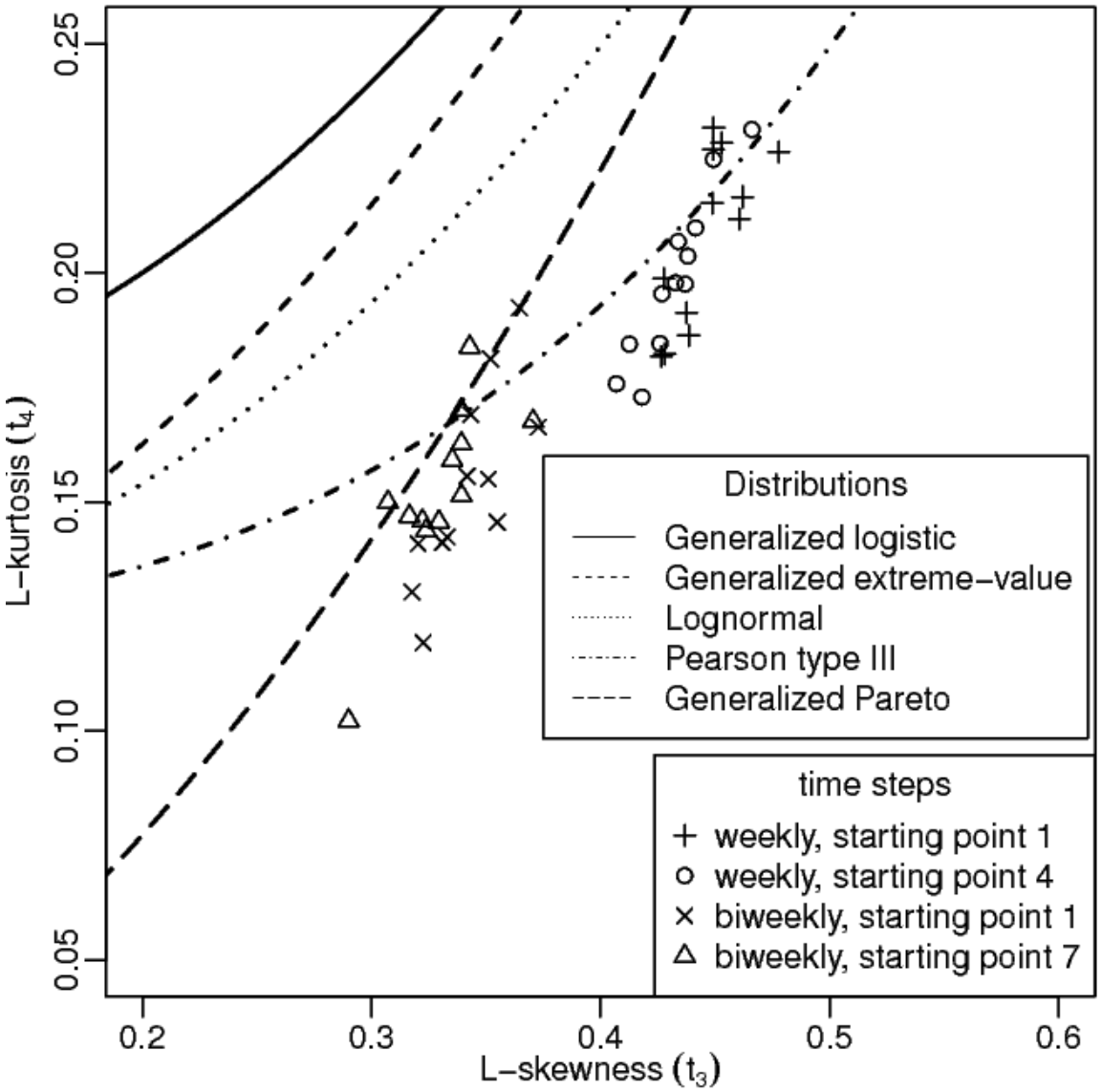

Figure 14 Graphical juxtaposition of observed data and five three-parameter probability distributions. Points: L-skewness and L-kurtosis (t_3 and t_4) of analyzed gauges at 7- and 14-day time steps for two starting points each. Lines: L-kurtosis outputted by three-parameter distributions as a function of L-skewness

For each regionalization, Z scores are calculated for each of the five three-parameter distributions listed above. Either the Generalized Pareto (GPA) or the Pearson Type III (PE3) distributions offer the best fit across all regionalizations in the data set, for all starting points and time steps. Figure 15 illustrates the percentage of regionalizations

best fitting these distributions across time steps for the starting point of Day One. $|Z^{GPA}|$ is consistently below 1.64 at long time steps for all or nearly all regionalizations, but $|Z^{PE3}|$ is often above 1.64 at short time steps. No other Z statistic's absolute value is less than 1.64 for any regionalization across seventy tested time step-starting point combinations.

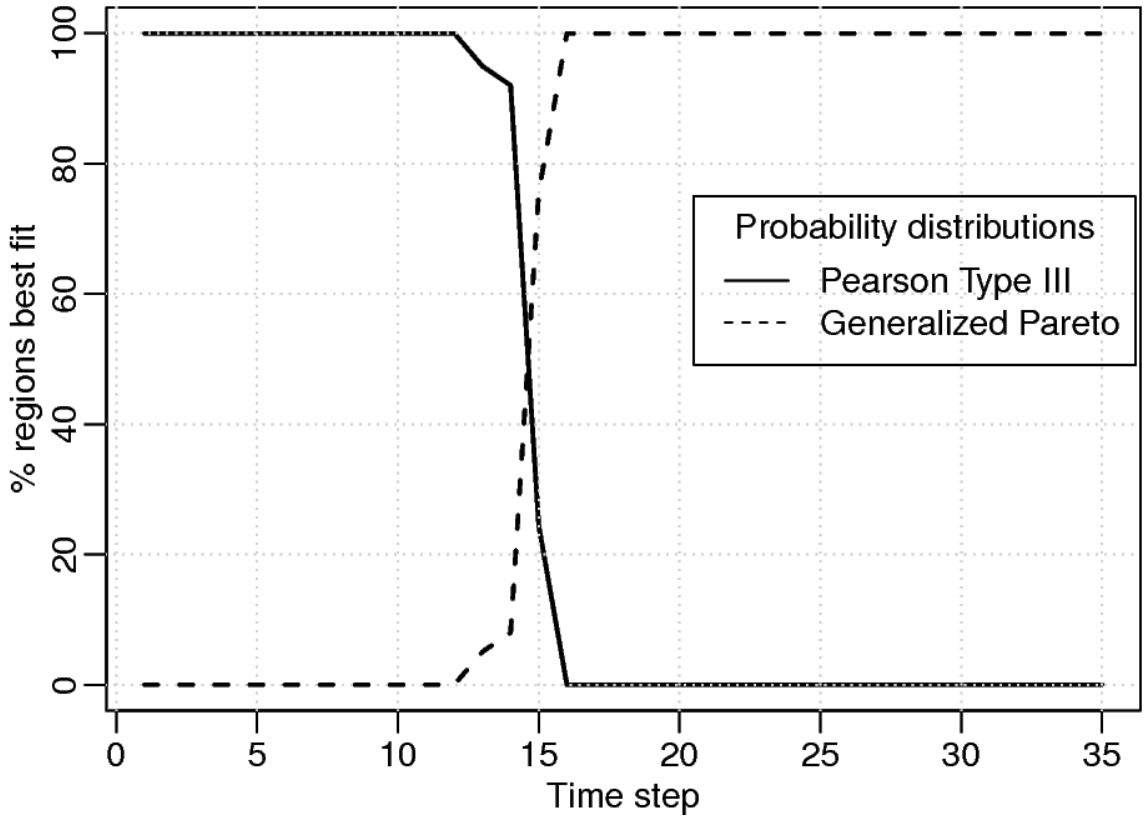

Figure 15 Percentage of regionalizations across a range of time steps (using starting point of Day One) with best fit for Pearson Type III and Generalized Pareto distributions

For time steps near fourteen days in length, the percentage of regionalizations best fit using the Pearson Type III or Generalized Pareto distributions varies as the starting

point changes. Figure 16 illustrates the percentage of regionalizations best fit by Pearson type III and Generalized Pareto for the fourteen-day time step across all starting points. Random variation in the L-moment ratios of each constituent gauge as starting point is varied results in each regionalization's average L-kurtosis moving closer to or farther from the predicted value for each distribution. Because the Z statistic is a Monte Carlo statistic, random sampling also contributes to variance in the Z scores for each distribution, which can affect the ranking of the distributions. At the biweekly time step, most starting points result in the Generalized Pareto distribution being the best fit for more than three quarters of regionalizations, while several other starting points result in the same majority fitting to the Pearson Type III distribution. While both distributions were used to generate data for all starting points at the fourteen-day time step, the results were very similar.

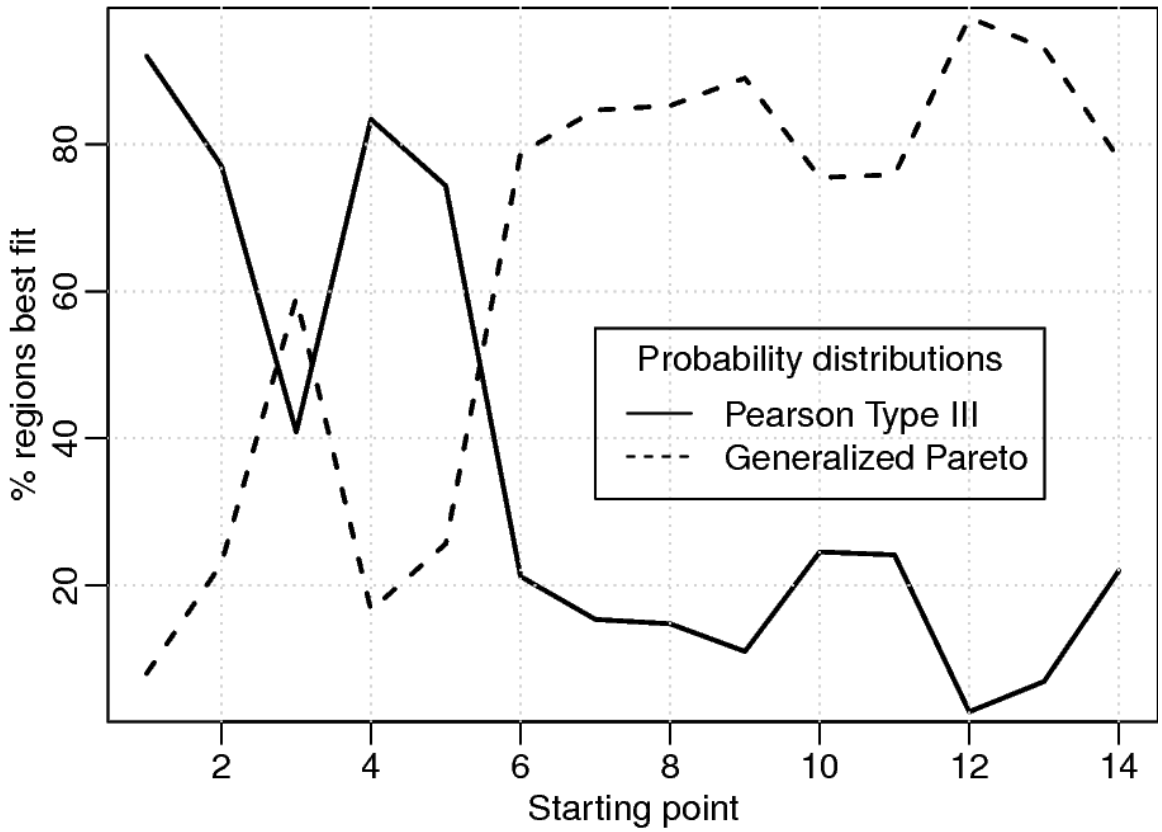

Figure 16 Percentage of regionalizations at 14-day time step across all starting points with best fit for Pearson Type III and Generalized Pareto from a set of five three-parameter distributions

Distributions are fit using equations in the Appendix of Hosking and Wallis (1997). While the estimated quantiles for low non-exceedance values can be negative when true zeroes are included in the analysis, a physically impossible result, this study focuses on the upper tails of the distribution. Quantile estimates at all analyzed sites, starting points, and time steps were positive for non-exceedance probabilities considered here (0.75 and greater).

3.4.3: Comparing heterogeneity and error estimates

All 3,302 valid regionalizations of the twelve-gauge dataset are evaluated for heterogeneity and error using Monte Carlo simulation at several time steps ranging from one to thirty-five days. After screening for low-multiplier regionalizations whose observed L-moment ratio variation is less than required for the accurate estimation of quantile error, correlations between error and the H statistics for the remaining regionalizations are measured using Pearson's r.

The multiplier step produces a minority of regionalizations at every time step for which a small fraction of the observed L-moment ratio variation was required to create a synthetic region matching the true H_1. Almost all of these regionalizations had negative H_1 values. Despite their low heterogeneity, these regionalizations' estimated errors are often high when fitted by the Pearson Type III or Generalized Pareto distributions. Because simulated quantiles are compared to true quantiles from the original data, it is unsurprising that estimated error is high when the simulated region's L-moment ratios are dissimilar to the original region. Regionalizations with a multiplier threshold of ≤ 0.50 are removed from the analysis due to their highly nonlinear clustering and the lack of a positive correlation between heterogeneity and error. Changing the threshold from 0.50 to 0.45 or 0.55 only slightly modifies the results. A threshold of 0.50 achieves acceptable performance at separating the main body of regionalizations, whose heterogeneity and error statistics are both high or low simultaneously, from the anomalous low-weighted regionalizations containing low heterogeneities and high errors.

High multipliers become less common as time step increases, but the most common multipliers for all time steps are above the threshold of 0.50 (Figure 17). Nevertheless, the number of regionalizations fitted with multipliers equal to or below 0.50 steadily increases at short time steps (Figure 18). For longer time steps the relationship is less well defined, but in general longer time steps result in a greater number of low-multiplier regionalizations. Most regionalizations at all time steps have high multipliers and pass this screening step. Figure 19 illustrates the number of low-multiplier regionalizations at all starting points for the weekly and biweekly time steps; some starting points have very few low-multiplier regionalizations while others have more.

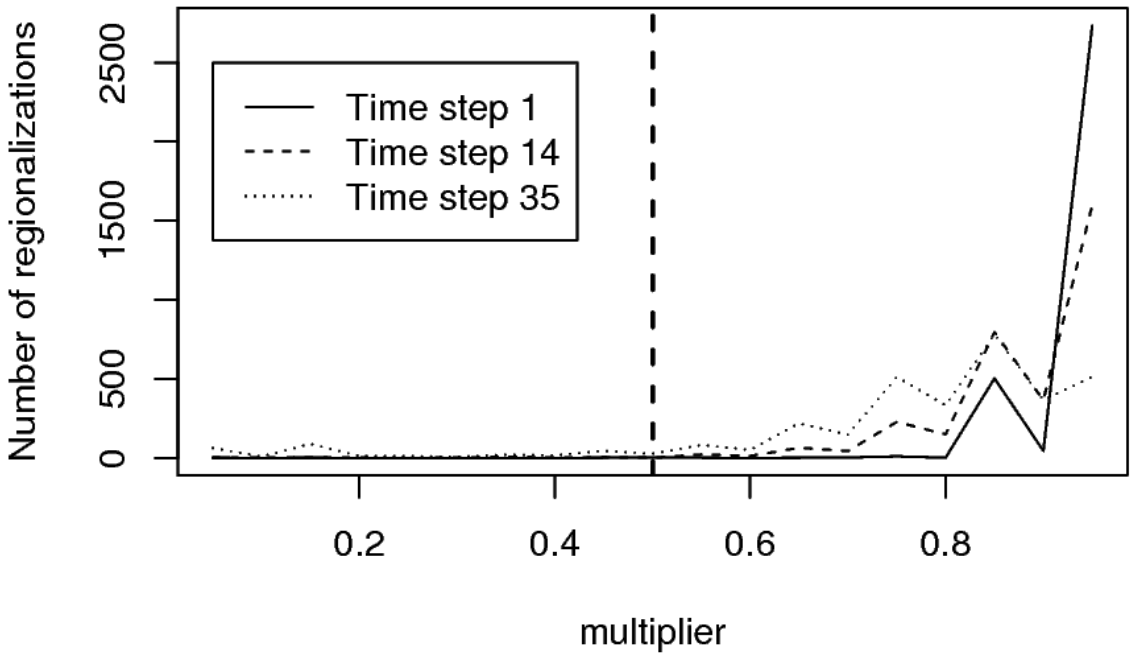

Figure 17 Number of regionalizations fitted to each multiplier at selected time steps for starting point of Day One. Low-multiplier threshold is illustrated as heavy dashed line.

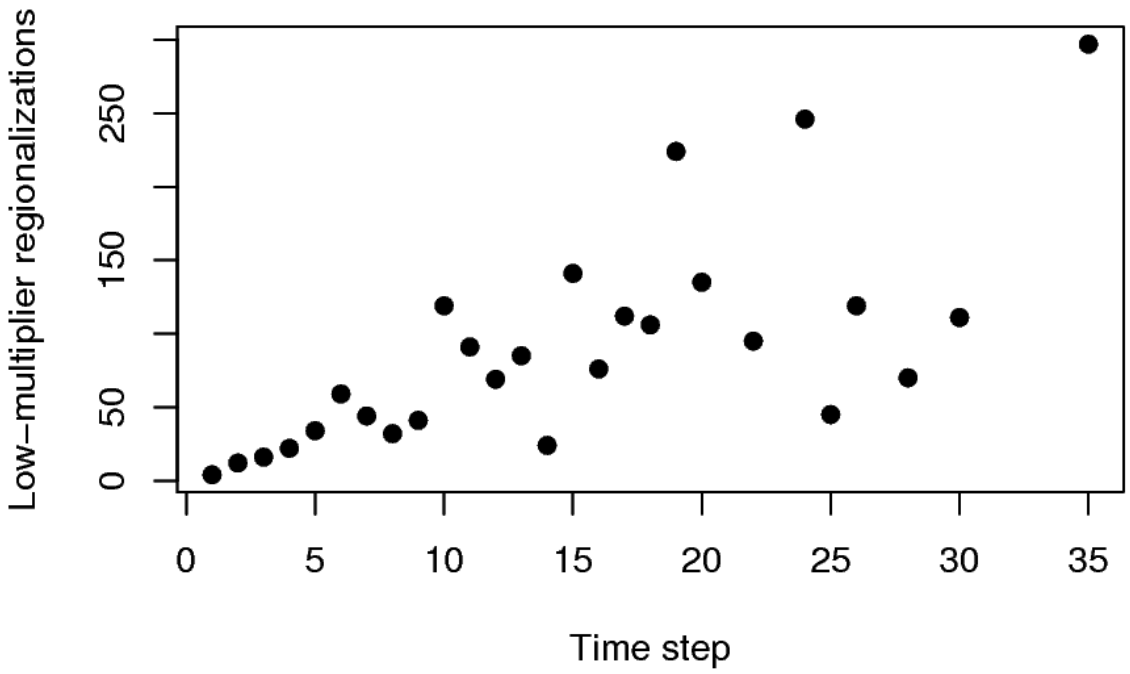

Figure 18 Number of regionalizations fitted with a multiplier less than or equal to 0.50 over a range of time steps for a starting point of Day One

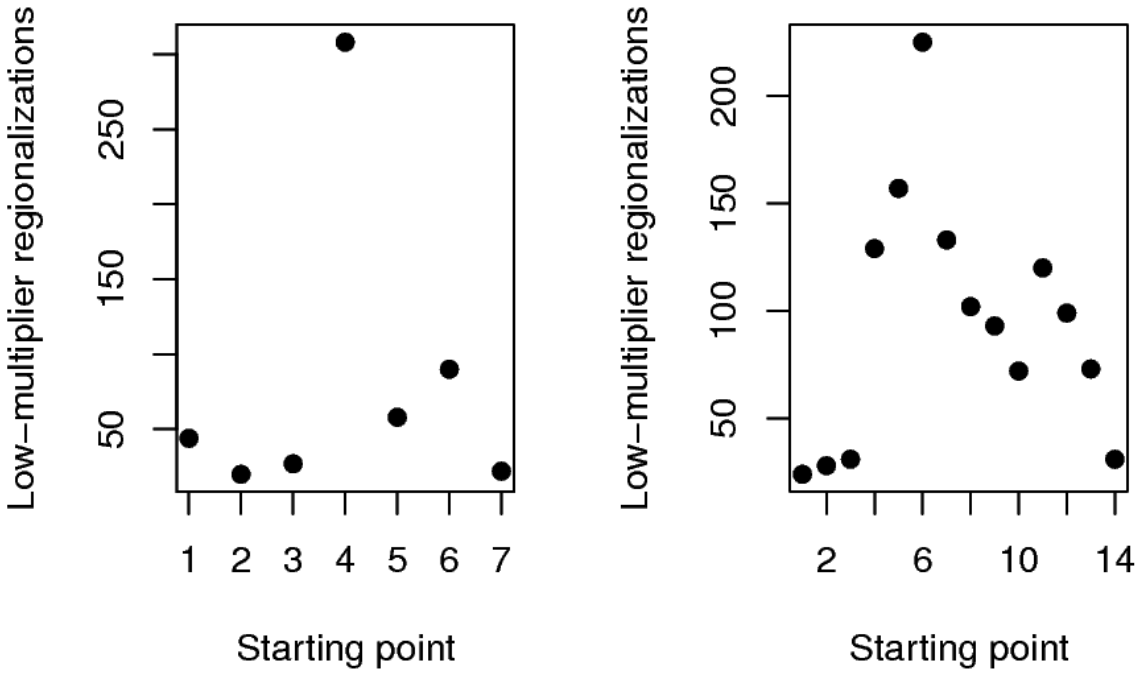

Figure 19 Number of regionalizations fitted with multiplier equal to or below 0.5 for all starting points at time steps of seven (left) and fourteen (right)

Pearson's r scores drawn from the relationship between heterogeneity and error are plotted for a given combination of starting point, time step, heterogeneity statistic, and non-exceedance frequency in Figure 20. While multiple linear relationships can be discerned in Figure 20 (the cluster with the highest heterogeneity contains all regionalizations incorporating site 328 which has outlier L-moment ratio values), only the diffuse cluster with multipliers less than or equal to 0.50 is eliminated before Pearson's r is calculated for the remaining regionalizations. Some starting points and time steps have multiple linear clusters visible in H-RMSE graphs, and some do not. These linear clusters are visible across many non-exceedance frequencies and for all three H statistics. The clusters can sometimes overlap, especially for H_3.

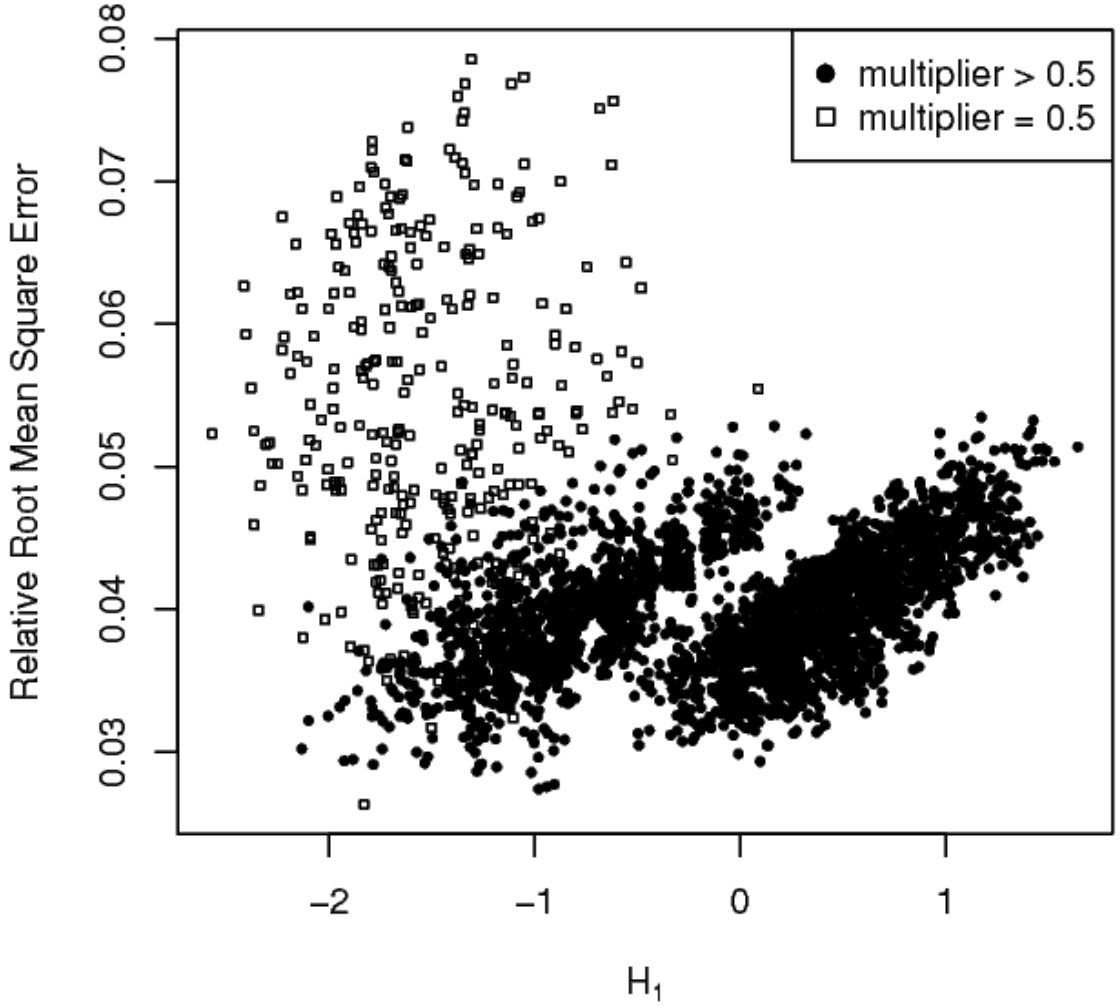

Figure 20 Relative Root Mean Square Error at the non-exceedance frequency of 0.90 plotted against H_1 for 3,302 regionalizations at a time step of 7 days and a starting point of Day 4. Regionalizations are differentiated by the magnitude of the fitted multiplier

For the heterogeneity statistics H_1 and H_2, correlation between heterogeneity and error for regionalizations with multipliers above 0.50 is slightly negative as measured by Pearson's r for non-exceedance frequencies below 0.90 (Figure 21). Pearson's r scores become slightly positive for all three heterogeneity statistics at a non-exceedance of 0.90. For a non-exceedance frequency of 0.95, the Pearson's r correlation is closer to 0.50,

while at non-exceedance frequencies of 0.99 and 0.999 correlations are consistently above 0.60, albeit with some drop-off at the longest time steps. These observations made from Figure 21 indicate the presence of a linear relationship between H_1 and H_2 and quantile error estimates for high non-exceedance frequencies across a wide range of time steps, and therefore, of L-moment ratio values.

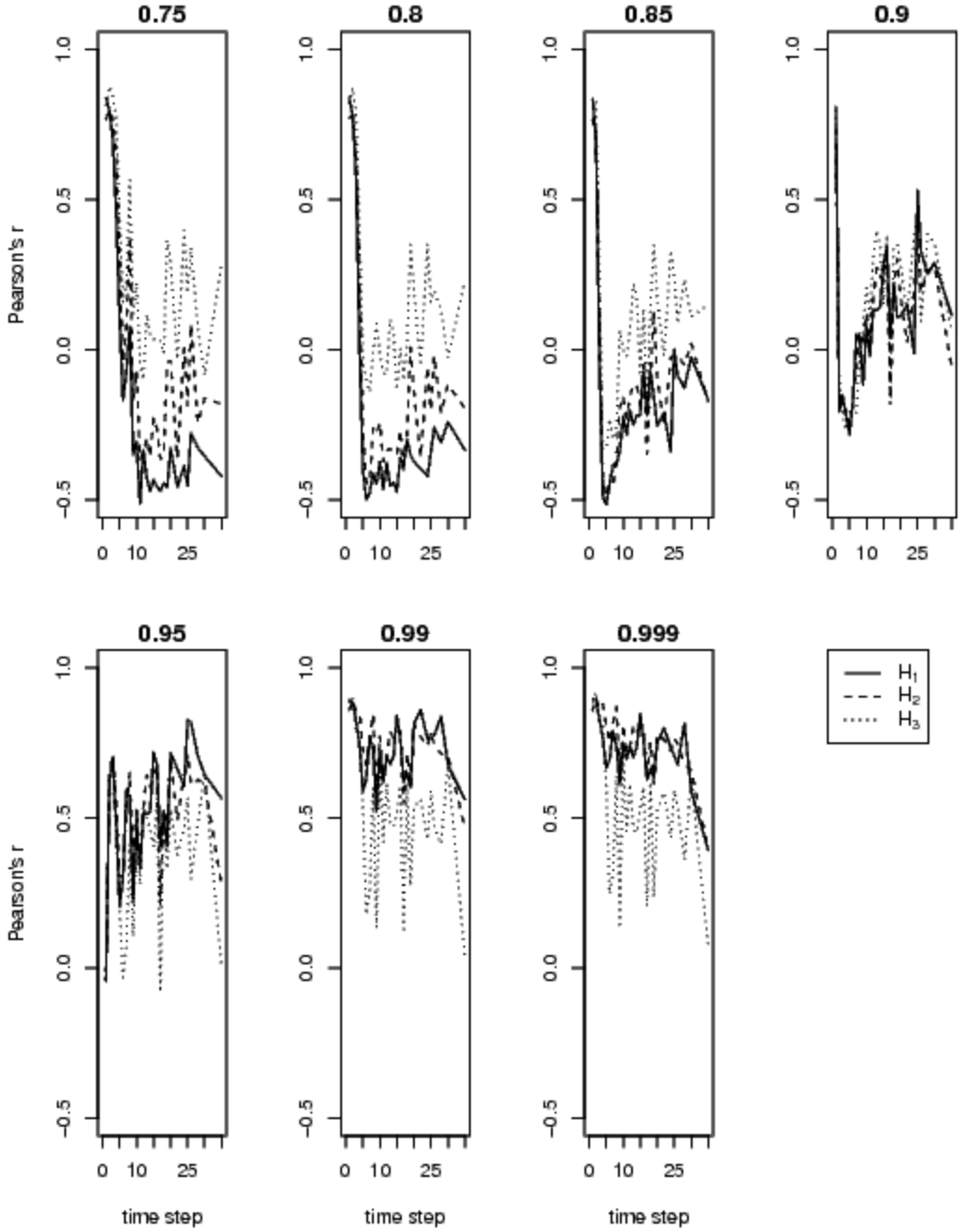

Figure 21 Pearson's r of H_1 (solid line), H_2 (dashed line), and H_3 (dotted line) for regionalizations with a multiplier of above 0.50 compared with estimated Root Mean Square Error at upper-tail non-exceedance frequencies across a range of time steps for a starting point of Day One

Figure 21 also illustrates the opposite tendency for the heterogeneity statistic H_3. This statistic exhibits near-zero correlation with RMSE for non-exceedance frequencies

below 0.95 but varies between 0.1 and 0.5 for higher non-exceedance frequencies. While H_3's Pearson's r scores are greater than those of H_1 and H_2 at lower frequencies, the correlations are still nearly zero and no strong linear relationship can be discerned using Pearson's r. However, H_3 lags behind where H_1 and H_2 exhibit their highest Pearson's r scores with reference to their correlation with RMSE.

3.4: Conclusions

Daily data from twelve Twin Cities precipitation gauges are aggregated using different time steps, and all valid regionalizations of these gauges are evaluated for heterogeneity and error. A positive linear correlation between two heterogeneity estimators and error estimates is observed for extreme quantiles (events with non-exceedance frequency of 0.95 or greater). These results offer evidence suggesting that H_1 and H_2, but not H_3, are reasonable proxies of RMSE for non-exceedance probabilities of 0.95 and above.

Regionalizations with low multipliers appear to have been affected by limitations inherent in the Monte Carlo error estimation mechanism used in this study. This method assumes that heterogeneity is produced by both sampling variance and the true scatter in L-moment ratio values within a region. Monte Carlo simulation involves the generation of samples, which adds sampling error. In order to prevent each iteration from being served a double dose of sampling error, the sampling error of the observed data must be ameliorated before fitting to the distribution. This is the role of the multiplier. When observed H_1 is very low – usually below zero – a multiplier of below 0.5 is necessary to reconcile the observed spread of L-moment ratios with observed H_1. The simulated

region thus produced is considerably different from the original data, so it is unsurprising that the error of quantile estimation increases. Issues leading to the fitting of low multipliers affected only a small minority of regionalizations even for the most heavily affected time steps and starting points.

The results of this analysis corroborate the simulation experiment in Hosking and Wallis (1997) indicating that one of the heterogeneity statistics used here is a "reasonable proxy" of the error of quantile estimation. In that monograph, results for H_2 and H_3 were not reported. The results of this analysis indicate a functional distinction between H_1 and H_2 on the one hand, which had poor correlations with RMSE at low non-exceedance frequencies but were useful proxies of estimated RMSE at high non-exceedance frequencies, and H_3 on the other, which was less effective at high non-exceedance frequencies, but only marginally more effective than H_1 and H_2 at low non-exceedance frequencies.

The relative efficacies of H_1 and H_2 are similar across this dataset and both present a linear relationship with error estimates despite the continued presence of error components not related to heterogeneity. The presence of multiple linear clusters for many combinations of H statistic, starting point and time step also had the effect of reducing Pearson's r scores between the H statistics and estimated relative RMSE. Evidence of a linear relationship between heterogeneity estimators and error estimates in real data strengthens the "reasonable proxy" argument for the heterogeneity estimators' accurate reflection of the heterogeneity error term in regional frequency analysis.

This methodology can be extended through expansion in scope. Several aspects of the study are amenable to further exposition. Other heterogeneity measures, including multivariate heterogeneity statistics, could be compared to error; indeed, other measures of error could be considered. The geographical scope of the analysis could be widened to include a broader area, or the relationship between heterogeneity and error could be investigated in a dataset from some other part of the world. If a larger gauge network is used and computational resources are insufficient for full enumeration, a random subset of regionalizations could be selected at which error and heterogeneity are estimated.

3.5: Acknowledgements

Data were graciously provided by the State Climatology Office, Minnesota Department of Natural Resources – Division of Ecological and Water Resources. Financial support provided by the George Mason University Presidential Scholarship is gratefully acknowledged.

4: DISCRIMINATORY POWER OF HETEROGENEITY STATISTICS WITH RESPECT TO ERROR OF PRECIPITATION QUANTILE ESTIMATION

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 "reasonable proxy" relationship with error caused by heterogeneity and through the ability of heterogeneity thresholds to detect heterogeneous regions. In this study, previous findings regarding three heterogeneity statistics proposed by Hosking and Wallis (1997), denoted H_1, H_2, and H_3, are revisited; the finding that H_1 is superior to H_2 and H_3 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 four times too high for application to H_2, and new thresholds are derived for H_2. Two nonparametric heterogeneity statistics are tested and found to achieve only the unsatisfactory performance level of H_3.

4.1: Introduction

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 effected by the pooling of multiple gauge records.

Heterogeneity statistics have been established in the literature based on their ability to detect heterogeneous regions (Viglione, 2007) and the degree to which the value of the heterogeneity statistic acts as a "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 "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.

In this study 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 here considered and heterogeneity thresholds are estimated for statistics with strong linear relationships to error.

This manuscript is subdivided into sections labeled Data, Analysis, Results, and Conclusions. The Data section describes the precipitation gauge dataset used in the enumeration study. The Analysis section introduces linear moments and the regional frequency analysis method, and then explicates heterogeneity and error statistics as well as methods used in the enumeration and simulation studies. The Results section describes the outcomes of these studies, and the Conclusions section reviews the results with reference to the literature and offers guidance to precipitation frequency analysts.

4.2: Data

Twelve daily precipitation gauges with more than a thousand 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 the 14th version of NOAA's Precipitation Frequency Atlas of the United States (Perica et al., 2013). Cartographic maps for all states currently covered by NOAA Atlas 14, 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 NOAA. Of these, 57 gauges which did not skip a full month between the start and end of the record were further considered and twelve clustered in the Minneapolis-St.Paul region were selected (Figure 22, Table 7). The longest record length of the twelve sites is 3751 days of non-zero precipitation and the shortest is 1134 days. The wet-day (non-zero) precipitation records of this group are used as the basis for the enumeration experiment.

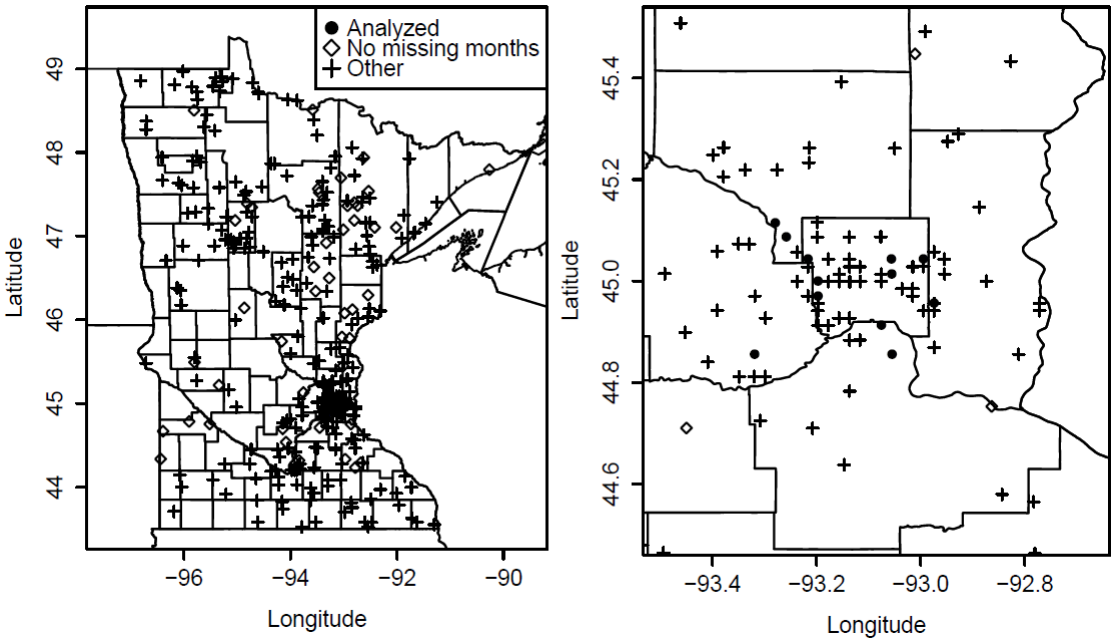

Figure 22 Map of Minnesota gauge network. Full state map on left, Twin Cities region on right.

Table 5 Characteristics of selected precipitation gauges

Gauge ID #	# days	Start month	End month	Latitude	Longitude
11	1714	Feb 1982	Feb 1996	44.86	-93.32
35	1191	Jan 1999	Jan 2009	45.01	-93.06
39	2419	Jan 1978	Aug 1998	45.04	-92.99
46	2765	Dec 1980	Nov 2008	45.12	-93.28
78	1518	Dec 1975	Apr 1990	44.97	-93.20
104	3751	Jan 1976	Nov 2008	45.09	-93.26
149	1488	Aug 1993	Nov 2008	44.86	-93.05
150	1787	Jul 1993	Nov 2008	44.91	-93.08
266	1859	May 1992	Oct 2008	45.00	-93.20
268	1438	Sep 1990	Nov 2004	45.04	-93.06
272	1134	Sep 1998	Nov 2008	45.04	-93.22
328	3572	Sep 1979	Dec 2007	44.96	-92.97

4.3: 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 outputted 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 here 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 this study. 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. The package applies FORTRAN code available at http://lib.stat.cmu.edu/general/lmoments. For the nonparametric bootstrap Anderson-Darling (AD) and Durbin-Knott (DK) test statistics the package "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.

4.3.1: 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 λ of a quantile function x(u) as the product of a polynomial $p^*_{r,k}$ and the probability-weighted moment β_k in Equations 29-31:

Equation 29

$$\lambda_{r+1} = \sum_{k=0}^{r} p^*_{r,k}\beta_k$$

Equation 30

$$p^*_{r,k} = \frac{(-1)^{r-k}(r+k)!}{(k!)^2(r-k)!}$$

Equation 31

$$\beta_k = \int_0^1 x(u)u^r du$$

The sample estimate of β_r is b_r, estimated using Equation 32 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:

Equation 32

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

λ_1 is calculated identically to the arithmetic mean, λ_2 measures variance, λ_3 measures skewness, and λ_4 measures kurtosis. L-moment ratios are often used to

represent the second through fourth moments. The L-CV or coefficient of L-variation is $\tau = \lambda_2/\lambda_1$, while L-skewness τ_3 and L-kurtosis τ_4 are calculated using Equation 33:

Equation 33

$$\tau_r = \lambda_r/\lambda_2$$

The sample estimate of λ_r is l_r the sample estimate of the L-CV is t, and the sample estimate of τ_r is t_r. These are calculated using Equations 34 and 35:

Equation 34

$$l_{r+1} = \sum_{k=0}^{r} p^*_{r,k} b_k$$

Equation 35

$$t_r = l_r/l_2$$

4.3.2: Homogeneity statistics

Because the L-moments analytical framework offers parameterizations for probability distributions, Monte Carlo statistics can be 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 more than regional heterogeneity increases it. However there must exist some threshold of heterogeneity at which error reduction due to increased sample size is cancelled out by increased 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.

4.3.3: 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 study 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. 5-minute, 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, 1997; Alila, 1999; Smithers and Schulze, 2001; Kyselý et al., 2007; Kyselý and Picek, 2007; Norbiato et al., 2007; Szolgay et al., 2009; Um et al., 2010; Yang et al., 2010; Ngongondo et al., 2011; Gabriele and Chiaravalloti, 2013). RFA-LM has also been used for quantile estimates of monthly (Núñez et al., 2011) and annual (Guttman, 1993; Lin and Chen, 2006; Parida and Moalafhi, 2008; Modarres and Sarhadi, 2011; Dikbas et al., 2012) precipitation totals. RFA-LM has also been widely applied to stream flow gauge data, particularly for quantile estimation of annual maximum floods (Zrinji and Burn, 1996; Burn and Goel, 2000; Kjeldsen et al., 2002; Jingyi and Hall, 2004; Abida and Ellouze, 2006; Atiem and Harmancioğlu, 2006; Rao and Srinivas, 2006; Srinivas et al., 2007; Noto and La Loggia, 2009; Gaume et al., 2010; Guse, 2010; Saf, 2010; Hussain, 2011; Kar et al., 2012; 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 following 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.

While three-parameter distributions are ultimately preferred for outputting regional quantile estimates, 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), one using only the L-CV (t), one using the L-CV and L-skewness (t_3), and one using the L-skewness and L-kurtosis (t_4). These statistics are denoted as V_1, V_2, and V_3 respectively in Equations 36-38. The R superscript indicates the regional average, and i is the index of the N sites in the region.

Equation 36

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

Equation 37

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

Equation 38

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

In Equation 39, μ_V is the mean and σ_V is the standard deviation of V across all simulated regions. H_1 is calculated using V_1, H_2 using V_2, and H_3 using V_3.

Equation 39

$$\mathrm{H} = \frac{V - \mu_V}{\sigma_V}$$

4.3.4: 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 (2007) expresses the Anderson-Darling (AD) and Durbin-Knott (DK) 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 for all i, the k-sample Anderson-Darling test statistic, here denoted as AD, can be calculated using Equation 40:

Equation 40

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

M_{ij} is the 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_{sim} values of AD which can be ranked low to high. A threshold equivalent to a 5% probability of heterogeneity is extrapolated from the 95^{th} 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. D_i is shown in Equation 41 below.

Equation 41

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

Because D_i is normal under the hypothesis of homogeneity, a statistic DK defined by Equation 42 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. Note that no simulations are required in the calculation of DK.

Equation 42

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

4.3.5: Root Mean Square Error (RMSE) 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.

4.3.6: 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. It follows that 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.

4.3.7: Goodness of fit for a probability distribution

Hosking and Wallis (1997)'s Z statistic is a Monte Carlo statistic measuring goodness of fit for a three-parameter distribution. Monte Carlo simulation using a given distribution parameterized by the mean, L-CV and L-skewness of the regional data

generates N_{sim} synthetic regions. The Z statistic measures how well the three-parameter distribution predicts the data's value of L-kurtosis, t_4. The bias B_4 and the standard deviation σ_4 of L-kurtosis, used in calculating Z, are calculated in equations 43 and 44. For a three-parameter distribution DIST with a fixed L-kurtosis after parameterization, τ^{DIST}_4, Z^{DIST} is calculated using the at-site L-kurtosis of site m, $t^{(m)}_4$, and regional average L-kurtosis t^R_4 in equation 45:

Equation 43

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

Equation 44

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

Equation 45

$$Z_{DIST} = \frac{\tau_4^{DIST} - t_4^R + B_4}{\sigma_4}$$

Hosking and Wallis (1997) list formulae providing τ^{DIST}_4 for five three-parameter distributions, the Generalized logistic, Generalized Pareto, Generalized extreme-value, Pearson type III, and Lognormal, in Table A.3.

4.3.8: Monte Carlo estimation of quantile error

Correlations between the constituent sites of the region are incorporated into the Monte Carlo process of simulated data generation by sampling from a multivariate Normal distribution with a covariance matrix that is based on the correlation between

constituent gauges at times when both recorded data. The Monte Carlo process by which simulated data are generated, described above for the Hosking-Wallis Monte Carlo heterogeneity statistics, can be modified to incorporate inter-site correlations. First, data fitting a multivariate Normal distribution are generated. This process requires a correlation matrix, which is calculated from the sample data as follows.

Average inter-site correlation $\bar{r}$ is calculated for the region using r_{ij}, the average correlation between each pair of gauges i and j in the dataset, in Equation 48. r_{ij} is calculated according to Hosking and Wallis (1997) using Equation 46:

Equation 46

$$r_{ij} = \frac{\sum_k (Q_{ik} - \bar{Q}_i)(Q_{jk} - \bar{Q}_j)}{\{\sum_k (Q_{ik} - \bar{Q}_i)^2 \sum_k \left(Q_{jk} - \bar{Q}_j\right)^2\}^{1/2}}$$

k is the index and n_{ij} is the number of time points for which both gauges i and j have data Q_{ik}, while $\bar{Q}_i$ is calculated in Equation 47.

Equation 47

$$\bar{Q}_i = n_{ij}^{-1} \sum_k Q_{ik}$$

Equation 48 calculates the average intersite correlation for the region, $\bar{r}$, from r_{ij}

Equation 48

$$\bar{r} = \{\frac{1}{2} N(N-1)\}^{-1} \sum_{1 \le i <} \sum_{j \le N} r_{ij}$$

Table 6 shows the correlation at the one-day time step between all twelve 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.

Table 6 Correlation between twelve sites in dataset for the one day time step

Site ID	11	35	39	46	78	104	149	150	266	268	272	328
11	1	0	0.69	0.49	0.67	0.61	0.71	0.77	0.70	0.60	0	0.54
35	-	1	0	0.83	0	0.60	0.80	0.87	0.81	0.88	0.66	0.84
39	-	-	1	0.46	0.69	0.70	0.30	0.36	0.48	0.54	0	0.50
46	-	-	-	1	0.57	0.67	0.73	0.77	0.77	0.82	0.67	0.73
78	-	-	-	-	1	0.84	0	0	0	0	0	0.66
104	-	-	-	-	-	1	0.49	0.55	0.58	0.66	0.87	0.64
149	-	-	-	-	-	-	1	0.86	0.70	0.73	0.55	0.77
150	-	-	-	-	-	-	-	1	0.76	0.79	0.60	0.83
266	-	-	-	-	-	-	-	-	1	0.79	0.62	0.76
268	-	-	-	-	-	-	-	-	-	1	0.66	0.78
272	-	-	-	-	-	-	-	-	-	-	1	0.67
328	-	-	-	-	-	-	-	-	-	-	-	1

"Regsimq" uses $\bar{r}$ to create a correlation matrix of the form in Equation 49, with as many columns and rows as there are sites in the region. Note that the regional average correlation, not the specific inter-site correlation r_{ij}, is used in Equation 49.

Equation 49

$$R = \begin{bmatrix} 1 & \bar{r} & \bar{r} & \cdots & \bar{r} \\ \bar{r} & 1 & \bar{r} & \cdots & \bar{r} \\ \bar{r} & \bar{r} & 1 & \cdots & \bar{r} \\ \vdots & \vdots & & \ddots & \\ \bar{r} & \bar{r} & & & 1 \end{bmatrix}$$

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 'shape' given the distributional parameters while retaining the appropriate inter-site correlations. The function "regsimq" in "lmomRFA" is used to calculate quantile error estimates in R.

The shrinkage estimator procedure described above is applied to each site's L-moment ratios, which are moved toward or away from the mean by a multiplier that,

when the shrunken L-moments are used to parameterize Monte Carlo simulations, result in the closest H_1 to the original region. "Regsimq" generates multivariate Normal correlated data at each simulated gauge, transforming the data using the three-parameter distribution with the lowest absolute value of Z. The data are normalized and the simulated region's average L-moment ratios are found at each , which are fitted to the chosen distribution and used to output quantile estimates. Equation 50 estimates the root mean square error (RMSE) of the regional normalized quantile estimate $\widehat{\mathrm{Q}}_i^{(m)}$ as a predictor of at-site quantile estimate Q_i for a list of non-exceedance probabilities F.

Equation 50

$$\mathrm{R}^R(F) = N^{-1}\sum_{i=1}^{N}(M^{-1}\sum_{m=1}^{M}\{\frac{\widehat{Q}_i^{[m]}(F) - Q_i\ (F)}{Q_i(F)}\}^2)^{1/2}$$

4.3.9: 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 RMSE 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 RMSE 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. Here, 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. Pearson's r is defined with reference to two vectors x and y with n elements each and mean values of $\bar{x}$ and $\bar{y}$ in Equation 51:

Equation 51

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

4.3.10: 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 two days in length has two possible starting points, while a weekly time step has seven and a monthly thirty. Starting points are essentially arbitrary and are defined with reference to the first day of record in the oldest of the twelve 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.

Data with lower time steps have higher skewness, likely due to the highly seasonal nature of daily precipitation totals where many winter days will receive almost no precipitation while the summer months see large volumes of precipitation. As the time

step is increased, high-volume precipitation events that are shorter than the time step tend to become aggregated with low-precipitation periods and vice versa, reducing the degree to which total precipitation volume is concentrated in a small percentage of data points.

Analysts should be aware that for records with large sample size in areas with highly seasonal precipitation the most accurate quantile estimates for short time steps will be achieved by computing the linear moments of data from only the season in question. For the purposes of testing heterogeneity statistics across a wide range of L-moment ratio values, however, the characteristics of the full dataset without seasonal screening are favorable.

4.3.11: 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 'true' quantile estimate. The root mean square error (RMSE) 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 RMSE 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% RMSE added and H = 2 represents 40-80% RMSE added. H statistics based on V_2 and V_3 were found to rarely exceed these thresholds and

were not recommended as heterogeneity estimators. While the measure H that is proposed by the authors uses V_1, the authors provide calculations for H statistics based on V_2 and V_3 as well in publicly available FORTRAN and R code. In this study the three H statistics described by Hosking and Wallis (1997) are denoted H_1, H_2, and H_3, respectively.

Results for H_2 and H_3 were not presented in the original study. Here 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-RMSE 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 7 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 7 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 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 (a), (b), (c), and (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. RMSE estimates at the 0.01, 0.1, 0.99, and 0.999 non-exceedance frequencies are calculated, along with the five heterogeneity statistics, H_1, H_2, H_3, and AD are calculated at each of the 100 instantiations of each region using 500 simulations, while DK is not a Monte Carlo statistic and so was calculated once for each instantiation. The one hundred values of each heterogeneity statistic and RMSE estimate are averaged for each region in Table 7.

Table 7 Simulation study

Region type	L-CV		# of sites	Region type	L-CV		# of sites
	Average	Range			Average	Range	
Hom n = 30	0.2	0	6	Het 30% (a)	0.2	0.06	21
	0.2	0	11	Het 30% (b)	0.2	0.06	21
	0.2	0	21	Het 30% (c)	0.2	0.06	21
	0.3	0	21	Het 30% (d)	0.2	0.06	21
	0.15	0	21	Hom n = 30	0.2	0	2
Het 30% n = 30	0.2	0.06	6		0.2	0	4
	0.2	0.06	11		0.2	0	10
	0.2	0.06	21		0.2	0	20
	0.3	0.09	21	Bimodal 20% n = 30	0.2	0.04	2
	0.15	0.45	21		0.2	0.04	4
Het 50% n = 30	0.2	0.1	6		0.2	0.04	10
	0.2	0.1	11		0.2	0.04	20
	0.2	0.1	21	Bimodal 30% n = 30	0.2	0.06	2
	0.3	0.15	21		0.2	0.06	4
	0.15	0.075	21		0.2	0.06	10
Hom n = 60	0.2	0	6		0.2	0.06	20
	0.2	0	11	Bimodal 50% n = 30	0.2	0.1	2
	0.2	0	21		0.2	0.1	4
Het 30% n = 60	0.2	0.06	6		0.2	0.1	10
	0.2	0.06	11		0.2	0.1	20
	0.2	0.06	21				
Het 50% n = 60	0.2	0.1	6				
	0.2	0.1	11				
	0.2	0.1	21				

4.4: Results

4.4.1: Enumeration study results

The relationship between each of five heterogeneity statistics and estimated quantile RMSE is investigated at time steps ranging from daily to monthly. The distribution chosen to model the data is made on the basis of the Z scores for three-parameter distributions. Across hundreds of starting point – time step combinations, not

one region among all 3,302 possible enumerated regions in the dataset sees the Generalized logistic, Generalized extreme-value, or Generalized Normal distribution record the lowest Z score among the five distributions tested. All regions in simulations with time steps shorter than eleven days record the lowest absolute value of Z score for the Pearson type III distribution. Conversely, all tested time step – starting point combinations for starting points of greater than eighteen days in length saw the Generalized Pareto record the lowest absolute Z score for all enumerated regions.

For time steps of twelve to seventeen days in length some regions fit the Generalized Pareto better according to the Z test while others fit the Pearson type III better. The percentage of regions best fitting each varies in seemingly random fashion as starting point is varied, while time steps closer to twelve see more regions fit the Pearson type III than the Generalized Pareto, and vice versa for time steps closer to seventeen. For these intermediate time steps the distribution chosen has little effect on the observed 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 thirteen, while all regions for time steps of thirteen and below are modeled with the Pearson type III distribution.

Figure 23 offers a graphical analogue to the results of these Z tests; it also illustrates the similarity between the Pearson type III 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 fourteen-

day, time step near the intersection of the Pearson Type III and Generalized Pareto lines offers a graphical rationale for using Generalized Pareto at and above the fourteen-day time step and Pearson type III below.

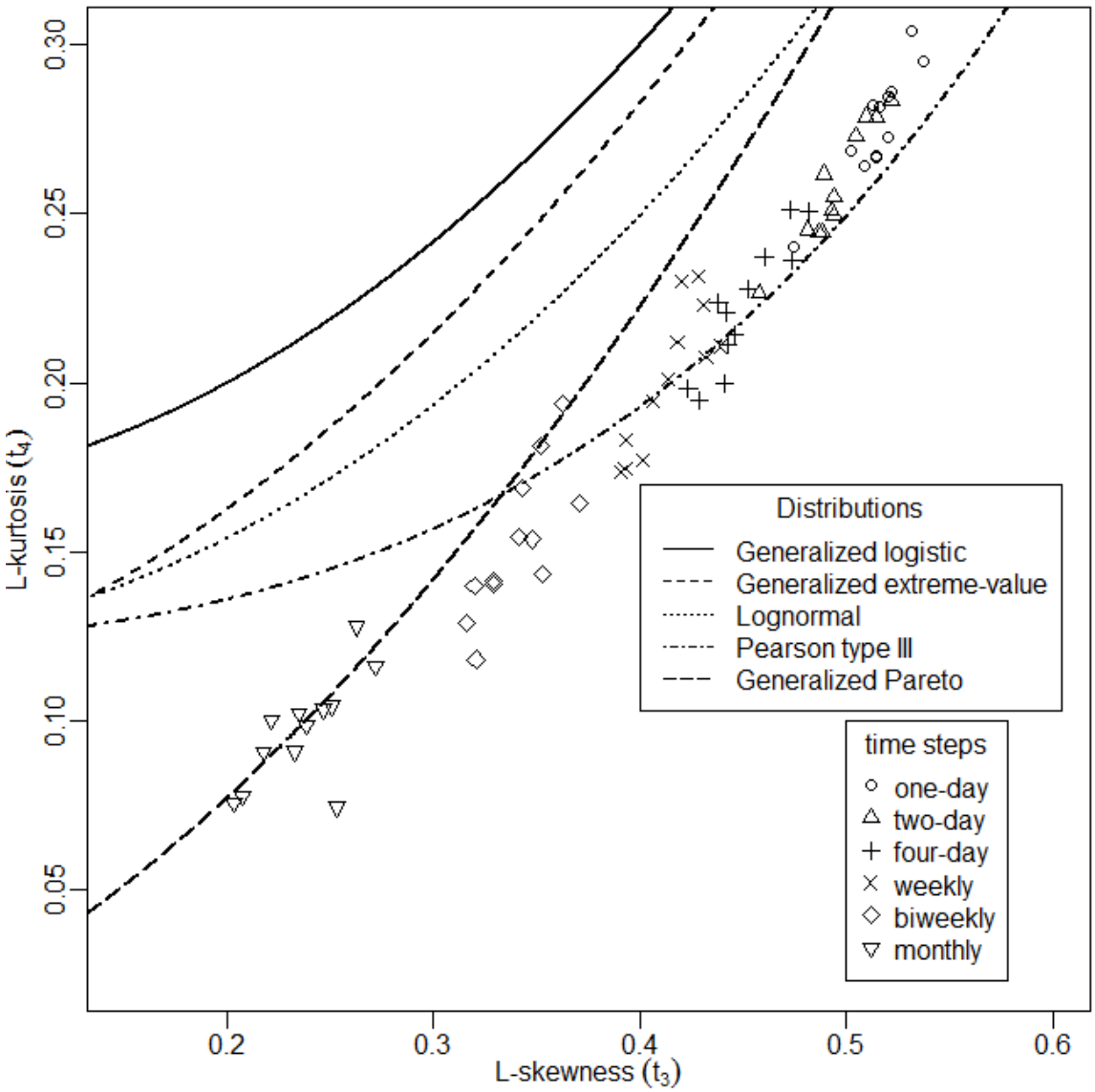

Figure 23 At-site L-skewness and L-kurtosis for different time steps at starting point of one plotted against curves representing L-moment ratios of data outputted by five three-parameter distributions.

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_1 and H_2 and to a lesser degree for H_3. AD and DK, instead of displaying a generally linear cluster with a positive slope, tend to display other patterns. Examples of H-RMSE plots at a non-exceedance probability of 0.999 are illustrated for the one- and fifteen-day time steps in Figures 24 and 25. In both cases two clusters can be discerned when H_1, H_2, or H_3 is plotted against RMSE, one at a low and one at a high heterogeneity, with the clusters having greater separation for H_1 than H_2, and for H_2 than H_3. This is more easily seen in the one-day than the fifteen-day time step, but in Figure 25 the distribution of regions when H_1 is plotted against RMSE is composed of two overlapping clusters, while for the H_3-RMSE plot the degree of overlapping has increased to the point where only one cluster is visible.

Despite the high-RMSE, low-heterogeneity points visible in the H_1, H_2 and H_3 plots in Figure 24, which are found to comprise the majority of regions with shrinkage multipliers below 0.5, a linear relationship with a positive slope between these heterogeneity estimators and RMSE exists for a large population of regions at the one day time step. The pattern seen in the H_1, H_2, and H_3 plots in Figure 25, where a more diffuse cluster exhibiting a positive slope comprises all regions, including low-multiplier ones, is broadly representative for time steps longer than three days. AD and DK do not exhibit linearity to the degree seen for H_1, H_2, and H_3.

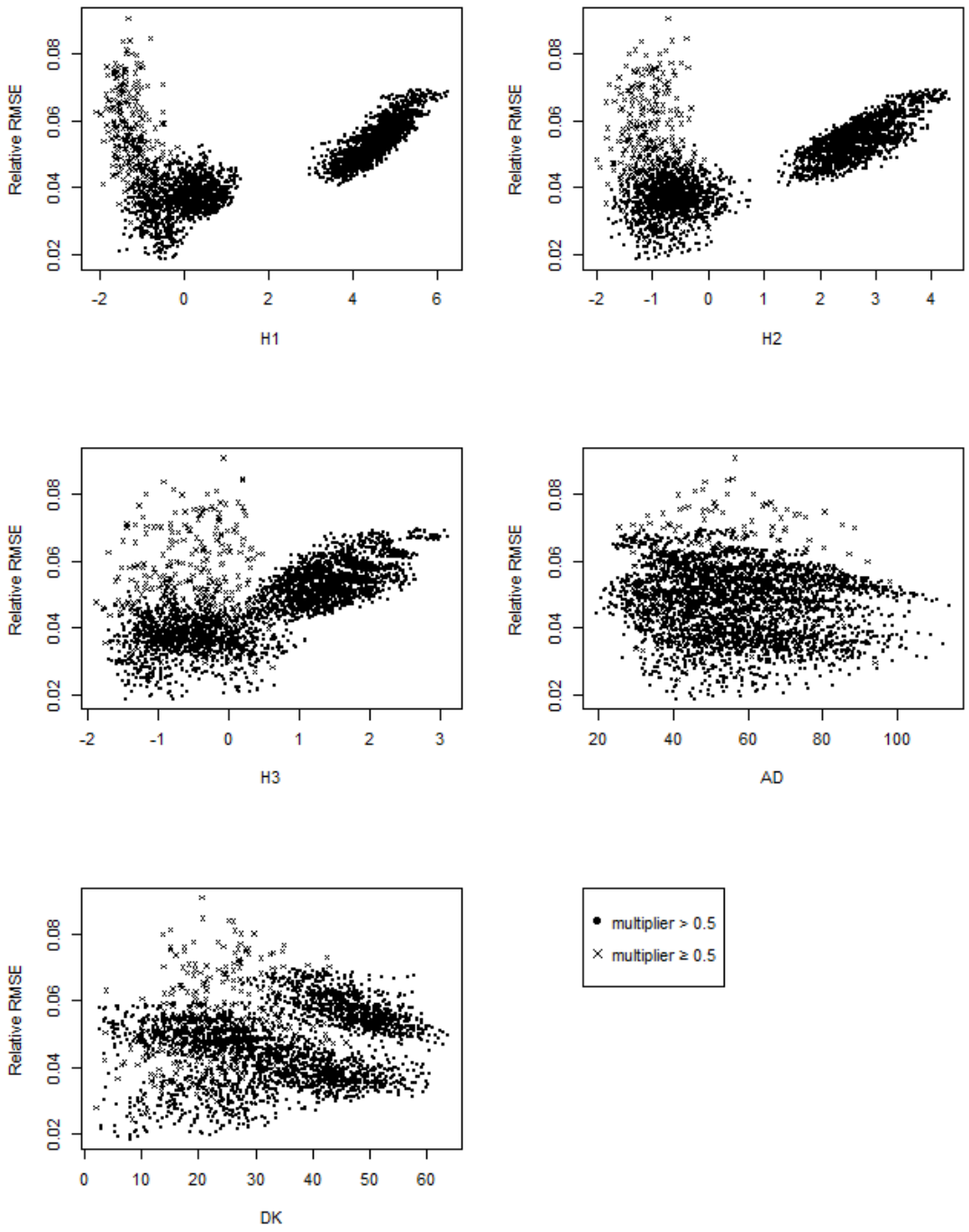

Figure 24 Estimated RMSE calculated using the Pearson type III distribution plotted against five heterogeneity statistics for the one day time step at 0.999 non-exceedance frequency

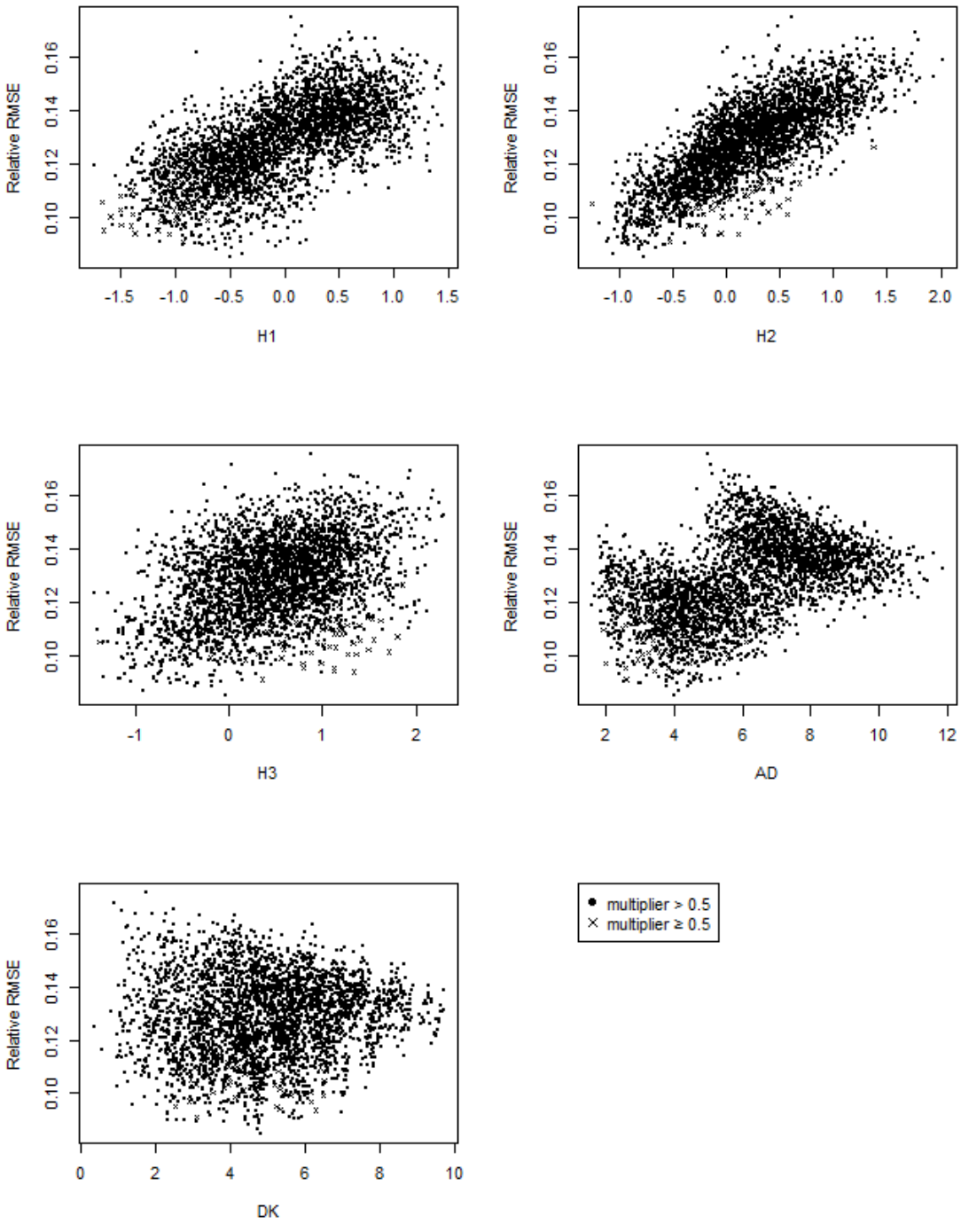
Relative RMSE
0.10
0.12
0.14
0.16
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
H1
Relative RMSE
0.10
0.12
0.14
0.16
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
H2
Relative RMSE
0.10
0.12
0.14
0.16
-1
0
1
2
H3
Relative RMSE
0.10
0.12
0.14
0.16
2
4
6
8
10
12
AD
multiplier > 0.5
multiplier ≥ 0.5
Relative RMSE
0.10
0.12
0.14
0.16
0
2
4
6
8
10
DK

Figure 25 Estimated RMSE calculated using the Generalized Pareto distribution plotted against five heterogeneity statistics for the fifteen day time step and starting point three at 0.999 non-exceedance frequency

Pearson's r between the five heterogeneity statistics and estimated RMSE at the 0.999 non-exceedance probability is plotted across time steps from daily to 35 days in length at starting point one in Figure 26. Similar patterns are observed for other starting points, while non-exceedance probabilities below 0.95 exhibit low Pearson's r for all heterogeneity statistics, possibly due to the predominance of non-heterogeneity-related components in the RMSE of quantile estimates. Here the patterns illustrated for the two cases above are summarized at all analyzed time steps into a single value indicating the degree of linearity between RMSE and each heterogeneity statistic. 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 RMSE. H_3 consistently exhibits a less linear relationship with RMSE than H_1 or H_2, while AD and DK's nonlinear relationship with RMSE visible in Figures 24 and 25 translates to consistently low Pearson's r scores.

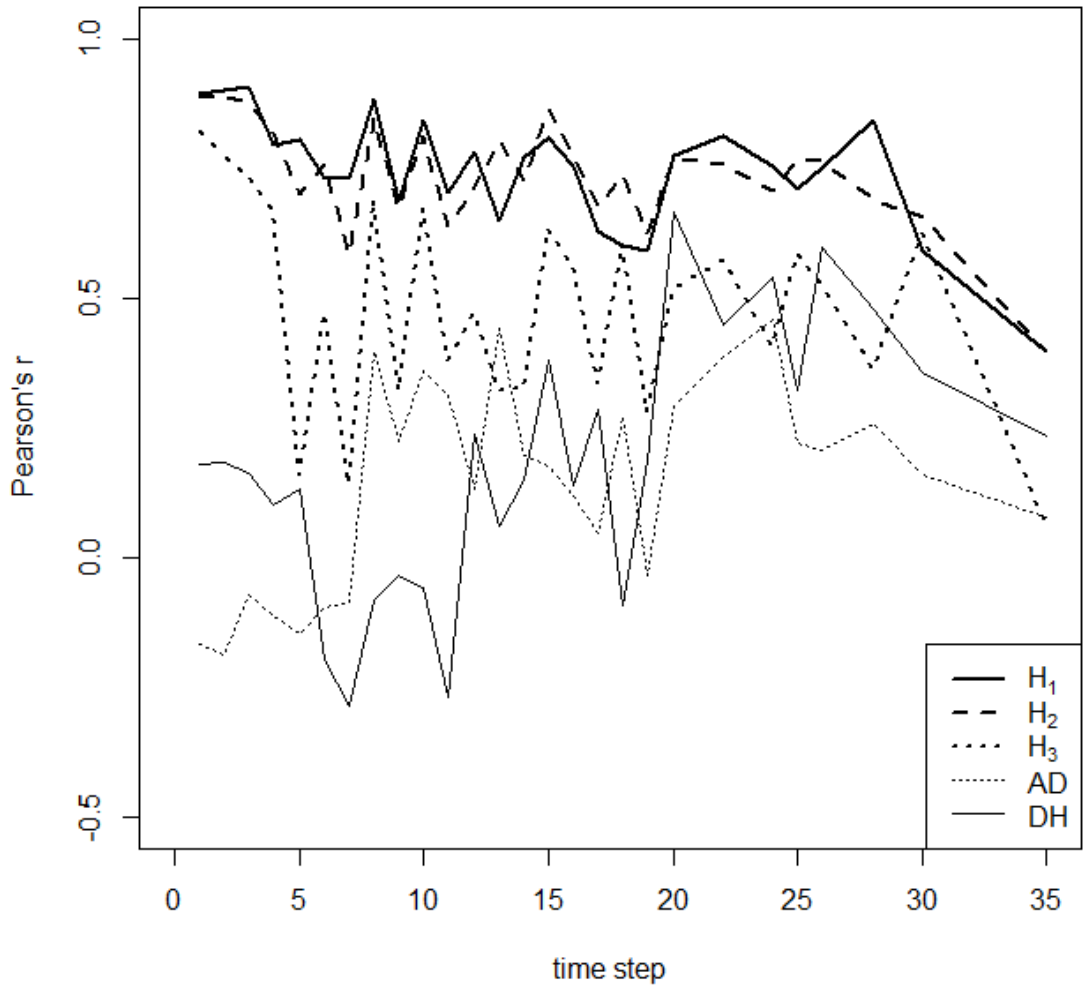

Figure 26 Pearson's r between estimated RMSE and five heterogeneity statistics at 0.999 non-exceedance frequency for 1-35 day time steps at starting point one

4.4.2: 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 RMSE added due to

heterogeneity. This relationship is depicted for the five heterogeneity statistics here considered in Figure 27 for the non-exceedance frequency 0.99. At non-exceedance frequencies of 0.01, 0.1, 0.99, and 0.999 the heterogeneity statistic-RMSE 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 RMSE added at all four non-exceedance frequencies, as seen in Figure 28.

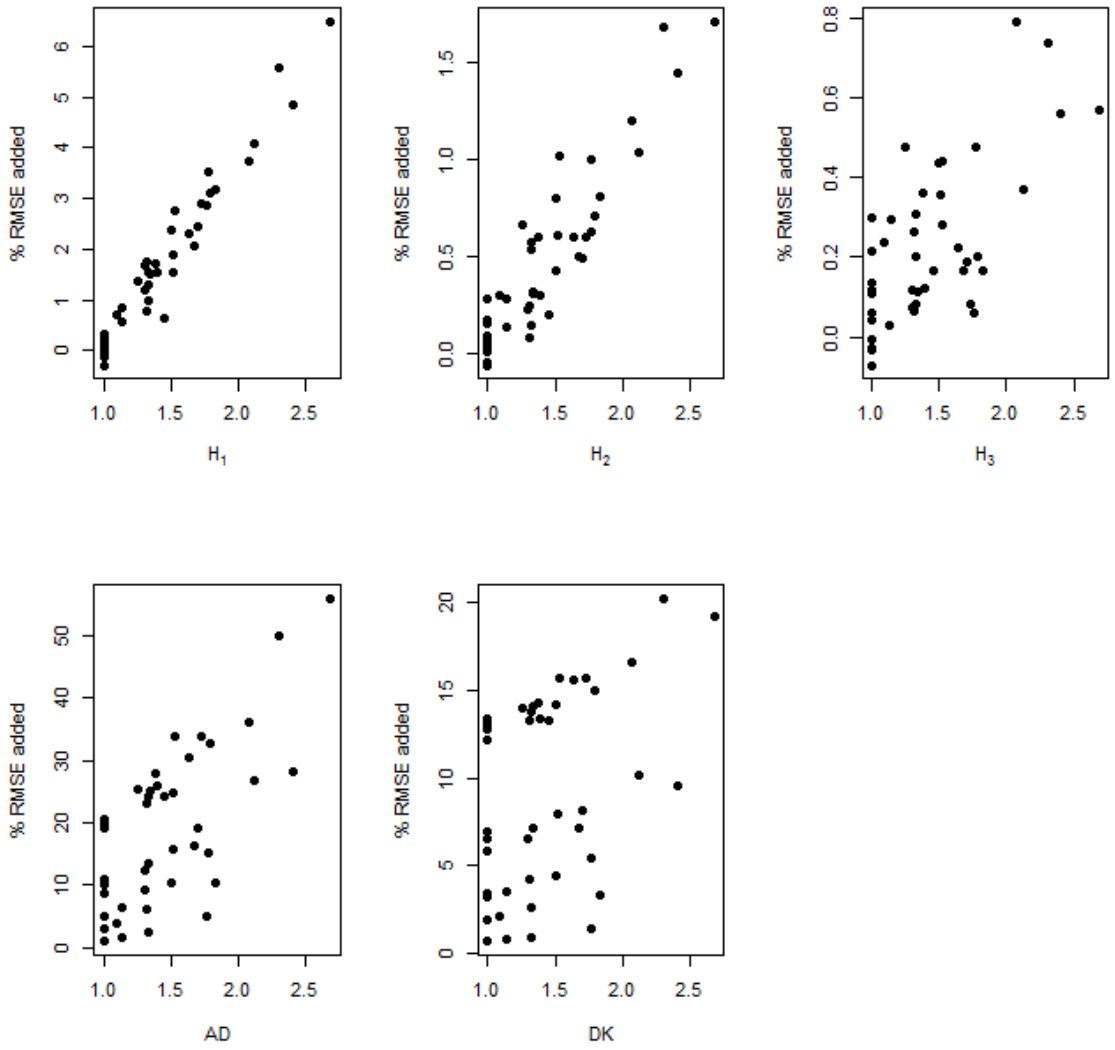

Figure 27 Percent RMSE added due to heterogeneity for simulated regions plotted against (a) H_1, (b) H_2 (c) H_3, (d) AD and (e) DK at non-exceedance probability of 0.99.

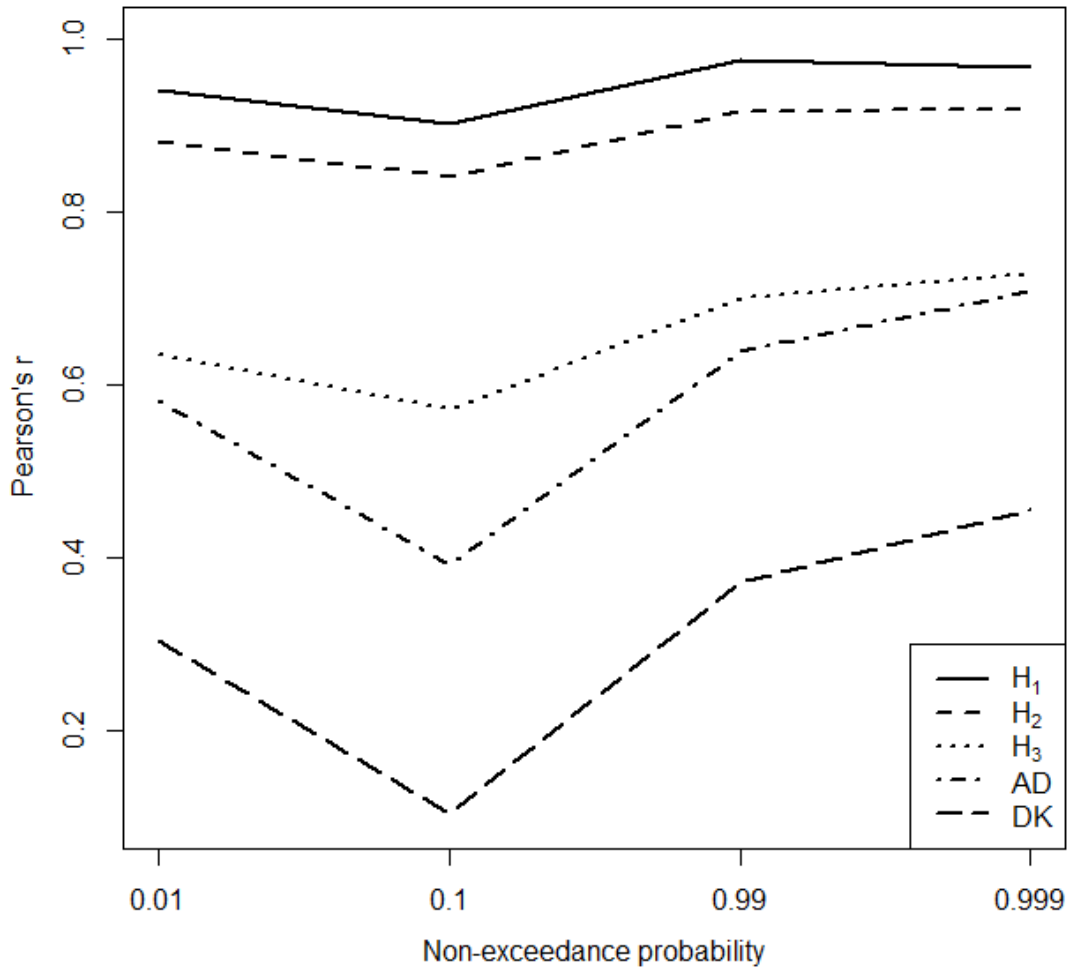

Figure 28 Pearson's r of linear fit between percent RMSE added due to heterogeneity and the heterogeneity statistics.

H_2 is found to consistently have approximately one-fourth the magnitude of H_1 across all simulations (Figure 29).

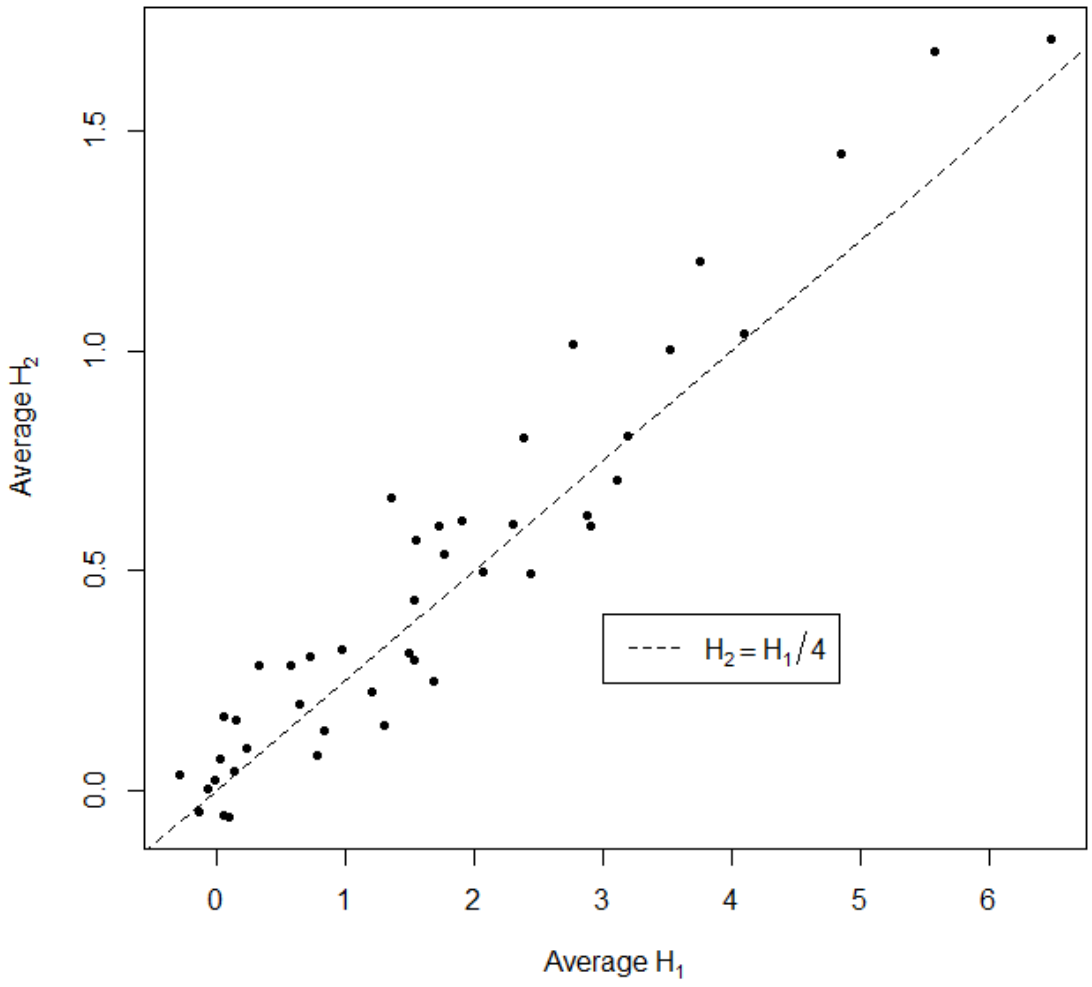

Figure 29 Average values of H_2 plotted against H_1 for all simulations

4.5: Conclusions

The Hosking-Wallis heterogeneity statistics H_1 and H_2 have superior performance to H_3 and the nonparametric statistics in both experiments. When heterogeneity-associated error is isolated in a simulation experiment a weak linear correlation can be

detected between error and the Anderson-Darling statistic's value, but for total estimated error in the enumeration experiment no linear relationship was noted. The relationship between error and the Durbin-Knott statistic was not well defined in either experiment.

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 these analyses 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-RMSE relationship severely understate the efficacy of H_2 at identifying heterogeneous regions.

The similarity between the H_1-RMSE and H_2-RMSE relationships, and the presence in the literature of arguments defending threshold values for H_1 based on the H_1-RMSE 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. It must be noted, however, that due to the slightly greater degree of scatter in an H_2-RMSE plot as compared to an H_1-RMSE plot the range of added RMSE that is considered equivalent to

a given H_2 value should be wider than the RMSE ranges described as equivalent to H_1 in Hosking and Wallis (1997). If $H_1 = 1$ is equivalent to 20-40% added RMSE, $H_2 = 0.25$ can be considered equivalent to 10-50% added RMSE; if $H_1 = 2$ is equivalent to 40-80% added RMSE, $H_2 = 0.50$ can be considered equivalent to 30-90% added RMSE. For both H_1 and H_2, equivalent RMSEs are approximations drawn from the empirical relationship between each H statistic and heterogeneity-associated RMSE in simulated experiments.

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.

H_1 and H_2 outperform the nonparametric statistics in the studies presented here, but some linearity could be detected for the Anderson-Darling statistic, indicating that even an L-moment based measure of error can exhibit linearity with reference to a nonparametric heterogeneity statistic. However, the relationship has a low enough Pearson's r that for any given Anderson-Darling value only a very wide range of equivalent RMSE can be described. For this reason Anderson-Darling threshold values analogous to those derived for H_1 and H_2 are not proposed.

5: CONCLUSION

Throughout history, major hydrological events such as floods and droughts have had devastating effects on human communities. Estimating the frequency of extreme hydrological events such as hundred-year floods is difficult without very large sample size, but reliable precipitation or streamflow gauges are scarce and rarely have more than a few decades of record, especially in developing regions of the world. Because of the important real-world ramifications of inaccurate estimates, such as floods due to dam failure or inadequate preparation for a famine caused by drought, a considerable amount of energy has been invested in reducing error.

Regionalization, the pooling of multiple gauges to increase sample size, is one proposed method. It requires the estimation of error caused by heterogeneity within the region so that the dissimilarity between gauges in the region can be shown to add less error than the increase in sample size removes. In recent decades, regionalization research has moved toward the use of linear moments and Monte Carlo simulation, known as regional frequency analysis using linear moments or RFA-LM. In the United States, regulatory agencies are debating changes to decades-old methods for estimating flood and storm magnitudes; RFA-LM methods represent a possible source of new ideas. This research, which indicates that two statistics are reliable indicators of error due to

heterogeneity, strengthens the case for RFA-LM as a reliable method for obtaining low-error estimates of extreme hydrological event magnitudes.

In the course of this research, simulation and enumeration studies were conducted across a wide range of moment values for the purpose of evaluating the utility of various heterogeneity statistics. Both methodologies found that two statistics, H_1 and H_2, offered reasonable linearity with respect to both total and heterogeneity-associated RMSE. While H_1 has the slightly more linear relationship to error as measured using Pearson's r in both cases, it is based on only one L-moment ratio. While the assumption that "for most kinds of data, sites with high L-skewness tend to have high L-CV too" was sufficient for Hosking and Wallis (1997) to recommend the use of H_1, the sample estimate of each statistic is subject to random sampling error. Given that the purpose of the regionalization framework is to counteract the deleterious effects of small sample size, including sampling error, analysts may wish to evaluate whether their data suffer from this limitation as they ascertain which heterogeneity statistic or statistics to use as a diagnostic tool in their regionalization studies.

Precipitation analysts can use this information to justify the use of H_1, H_2, or both in conjunction as estimators of heterogeneity. Results had not previously been published indicating the degree to which H_2, H_3, AD, and DK are "reasonable proxies" to error; in addition, H_2's heterogeneity thresholds had previously been defined using H_1's relationship with error, which has a higher slope than H_2's relationship with error. By presenting the results of validation experiments and offering lower thresholds for H_2, this research aids precipitation analysts in quantifying the degree to which quantile estimates

from their regionalizations are affected by heterogeneity. Analysts using L-moment regionalization methods on stream flow and other physical variables may also find these results to be of interest, particularly the results of the simulation study.

5.1: Linearity with respect to error of heterogeneity statistics

H_1, which has long been a widely used heterogeneity statistic, is confirmed as a "reasonable proxy" of heterogeneity-associated error. Thresholds set for H_1 and justified with respect to error are confirmed to be in accordance with the linear relationship established by simulation experiments. However, H_2 is nearly as linear and across simulation experiments has a fairly consistent 1:4 relationship with H_1 in terms of magnitude. Nonparametric statistics have a much less linear relationship with error across these simulations, with the k-sample Anderson-Durbin test statistic (AD) slightly outperforming the Durbin-Knott statistic (DK). H_3 has a similar linearity to AD.

Enumeration experiments support these conclusions, with H_1 and H_2 again appearing to be in a class of their own in terms of Pearson's r for the relationship between each heterogeneity estimate and RMSE. AD and H_3 have similar r values at long time steps. Linear relationships are overall less strong, which may reflect the presence of many non-heterogeneity related components of error.

5.2: Selection of heterogeneity statistics

H_1 is the highest-performing statistic across the gamut of studies presented here; it is also the only statistic considered which uses only one L-moment ratio. Hosking and Wallis (1997) defended this by noting that real data tend to see equivalent variance in

different L-moment ratios, so that heterogeneity will manifest itself in both L-skewness and L-CV, not in only one or the other. This assumption, however, can be violated through the action of sampling error at low sample size, the very case for which regional frequency analysis is of greatest utility. If a low-sample-size dataset happens to exhibit greater heterogeneity in L-skewness than L-CV, H_1 would underestimate the heterogeneity while H_2 would be ideally equipped to incorporate information from both L-moment ratios.

The nonparametric statistics here considered did not exhibit high performance as "reasonable proxies" of error in the simulation or enumeration studies. Because the error metric used here is calculated through Monte Carlo procedures using L-moment ratios, it is perhaps unsurprising that Monte Carlo heterogeneity statistics based on L-moment ratios would offer a better proxy for error than statistics that were derived from a different theoretical basis. Nevertheless, a "reasonable proxy" relationship between quantile error associated with heterogeneity and these statistics has not been established.

Most datasets are likely to exhibit heterogeneity in all L-moment ratios, and H_1's ability to act as a "reasonable proxy" to error in these cases is unequaled by the other heterogeneity statistics whose performance was quantified in these analyses. Nevertheless, H_2 offers nearly equivalent performance and is robust too. Although the above argument could be logically extended to L-kurtosis, H_3 exhibits poor linearity with respect to error and is not recommended for use.

5.2.3: Heterogeneity thresholds for recommended statistics

Much of the basis for rejecting H_2 as inferior to H_1 has been its inability to recognize heterogeneous regions as measured by the number of such regions exceeding thresholds that were set with reference to the H_1-percent RMSE added due to heterogeneity relationship. When the simulation study used to derive these thresholds is used for H_2, however, a roughly 4:1 relationship between H_1 and H_2 values is noted. Both statistics are linear with respect to percent RMSE added, but the H_2-RMSE slope is one-fourth the H_1-RMSE slope. Thresholds set according to the H_1`-RMSE relationship, therefore, are roughly four times too high when applied to H_2, which is likely related to past findings that H_2 lacked the power to identify heterogeneous regions. When the methodology of Hosking and Wallis (1997) is used to find heterogeneity thresholds, appropriate thresholds for H_2 should have around one-fourth the magnitude of thresholds derived for H_1.

5.3: Future directions of research

This research represents a small subset of the approaches that could be used to validate heterogeneity statistics. Estimating error through nonparametric bootstrap methods may yield a better estimation of the performance of nonparametric heterogeneity statistics and could help isolate the degree to which L-moment-based heterogeneity statistics' linearity with respect to L-moment based estimates of error is due to the commonality of their design. Data from other parts of the world influenced by different climatic processes could be analyzed using the enumeration method, or a random subset of all possible enumerations could be evaluated for precipitation gauge networks with a

larger number of gauges. Other heterogeneity statistics, including alternative L-moment-based methods, could be analyzed for a "reasonable proxy" relationship with error. The relationship between H_1 and H_2 can be analyzed in a literature review, where the ratio between reported values for the two statistics can be compared to that obtained in the simulated experiments above.

REFERENCES

Abida, H., Ellouze, M., 2006. Hydrological delineation of homogeneous regions in Tunisia. Water Resources Management 20, 961-977, doi:10.1007/s11269-006-9017-3.

Atiem, I.A., Harmancioğlu, N.B., 2006. Assessment of regional floods using L-moments approach: The case of the River Nile. Water Resources Management 20, 723-747, doi:10.1007/s11269-005-9004-0.

Adamowski, K., Alila, Y., Pilon, P.J., 1996. Regional rainfall distribution for Canada. Atmospheric Research 42, 75-88.

Alila, Y., 1994. A regional approach for estimating design storms in Canada, Ph.D. thesis, Civil Engineering Department, University of Ottawa, Ottawa, Ontario, Canada.

Alila, Y., 1999. A hierarchical approach for the regionalization of precipitation annual maxima in Canada. Journal of Geophysical Research 104, 31,645-31,655.

Bhuyan, A., Borah, M., Kumar, R., 2010. Regional flood frequency analysis of north-bank of the River Brahmaputra by using LH-moments. Water Resources Management 24, 1779-1790, doi:10.1007/s11269-009-9524-0.

Bradley, A.A., 1998. Regional frequency analysis methods for evaluating changes in hydrologic extremes. Water Resources Research 34(4), 741-750, doi:10.1029:98WR00096.

Burn, D.H., Goel, H.K., 2000. The formation of groups for regional flood frequency analysis. Hydrological Sciences Journal, 45(1), 97-112, doi:10.1080/02626660009492308.

Castellarin, A., Burn, D.H., Brath, A., 2008. Homogeneity testing: How homogeneous do heterogeneous cross-correlated regions seem? Journal of Hydrology 360, 67-76, doi:10.1016/j.jhydrol.2008.07.014.

Chebana, F., Ouarda, T.B.M.J., 2007. Multivariate L-moment homogeneity test. Water Resources Research 43, W08406, doi:10.1029/2006WR005639.

Dalrymple, T., 1960. Flood frequency analyses: Manual of hydrology: Part 3. Flood-Flow techniques. United States Geological Survey: Water-Supply Paper 1543-A, Washington, DC.

Dikbas, F., Firat, M., Cem Koc, A., Gungor, M., 2012. Classification of precipitation series using fuzzy cluster method. International Journal of Climatology 32, 1596-1603, doi:10.1002/joc.2350.

Dodangeh, E., Soltani, S., Sarhadi, A., Shiau, J.T., 2013. Application of L-moments and Bayesian inference for low flow regionalization in Sefidroud basin, Iran. Hydrological Processes 28, 1663-1676, doi:10.1002/hyp.9711.

Feng, J., Yan, D., Li, C., Gao, Y., Liu, J., Regional frequency analysis of extreme precipitation after drought event in the Heihe River Basin, Northwest China. Journal of Hydrologic Engineering, doi:10.1061/(ASCE)HE.1943-5584.0000903.

Gaál, L., Kyselý, J., 2009. Comparison of region-of-influence methods for estimating high quantiles of precipitation in a dense dataset in the Czech Republic. Hydrology and Earth System Sciences 13, 2203-2219, doi:10.5194/hess-13-2203-2009.

Gabriele, S., Chiaravalloti, F., 2013. Using the meteorological information for the regional rainfall frequency analysis: An application to Sicily. Water Resources Management 27, 1721-1735, doi:10.1007/s11269-012-0235-6.

Gaume, E., Gaál, L, Viglione, A., Szolgay, J., Kohnová S., Blöschl, G., 2010. Bayesian MCMC approach to regional flood frequency analyses involving extraordinary flood events at ungauged sites. Jornal of Hydrology 394, 101-117, doi:10.1016/j.jhydrol.2010.01.008.

Greenwood, J.A., Landwehr, J.M., Matalas, N.C., Wallis, J.R., 1979. Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. Water Resources Research 15, 1049-1054, doi:10.1029/WR015i005p01049.

Guse, B.F., 2010. Improving flood frequency analysis by integration of empirical and probabilistic regional envelope curves. Ph.D. dissertation, Faculty of Mathematics and Natural Sciences, University of Potsdam, Potsdam, Germany.

Guttman, N.B., Hosking, J.R.M., Wallis, J.R., 1993. Regional precipitation quantile values for the continental United States computed from L-moments. Journal of Climate 6, 2326-2340.

Hosking, J.R.M., 1990. L-moments: Analysis and estimation of distributions using linear combinations of order statistics. Journal of the Royal Statistical Society, Series B 52, 105-124.

Hosking, J.R.M., 1993. Some statistics useful in regional frequency analysis. Water Resources Research 29(2), 271-281, doi:10.1029/92WR01980.

Hosking, J.R.M., Wallis, J.R., 1997. Regional frequency analysis: An approach based on L-moments. Cambridge University Press, New York. ISBN: 0521019400.

Huff, F.A., Angel, J.R., 1992. Rainfall frequency atlas of the Midwest. Illinois State Water Survey, Champaign. Bulletin 71.

Hussain, Z., Application of the regional flood frequency analysis to the upper and lower basins of the Indus River, Pakistan. Water Resources Management 25, 2797-2822. doi:10.1007/s11269-011-9839-5.

Jingyi, Z., Hall, M.J., 2004. Regional flood frequency analysis for the Gan-Ming River basin in China. Journal of Hydrology 296, 98-117, doi:10.1016/j.jhydrol.2004.03.018.

Kar, A.K., Goel, N.K., Lohani, A.K., Roy, G.P. Application of Clustering Techniques Using Prioritized Variables in Regional Flood Frequency Analysis—Case Study of Mahanadi Basin. Journal of Hydrologic Engineering 17, 213-223, doi:10.1061/(ASCE)HE.1943-5584.0000417.

Kjeldsen, T.R., Smithers, J.C., Schulze, R.E., 2002. Regional flood frequency analysis in the KwaZulu-Natal province, South Africa, using the index-flood method. Journal of Hydrology 255, 194-211, doi:10.1016/S0022-1694(01)00520-0.

Kyselý, J., Picek, J.,2007. Regional growth curve and improved design value estimates of extreme precipitation events in the Czech Republic. Climate Research 33, 243-255, doi:10.3354/cr033243.

Kyselý, J., Picek, J., Huth, R., 2007. Formation of homogeneous regions for regional frequency analysis of extreme precipitation events in the Czech Republic. Studia Geophysica Et Geodaetica 51, 327-344.

Lin, G.F., Chen, L.H., 2006. Identification of homogeneous regions for regional frequency analysis using the self-organizing map. Journal of Hydrology 324 1-9, doi:10.1016/j.jhydrol.2005.09.009.

Lindley, D.V., Smith, A.F.M., 1972. Bayes estimates for the linear model. Journal of the Royal Statistical Society. Series B 34 (1) 1-41.

Lu, L.H., Stedinger, J.R., 1992. Sampling variance of normalized GEV/PWM quantile estimators and a regional homogeneity test. Journal of Hydrology 138, 223-245, doi:10.1016/0022-1694(92)90166-S.

Maeda, E.E., Torres, J.A., Carmona-Moreno, C., 2013. Characterisation of global precipitation frequency through the L-moments approach. Area 45.1, 98-108, doi:10.1111/j.1475-4762.2012.01127.x.

Marx, L., Kinter III, J.L., 2007. Estimating the representation of extreme precipitation events in atmospheric general circulation models using L-moments. Report by Center for Ocean-Land Atmosphere Studies (COLA), Calverton, Maryland, USA.

Modarres, R., 2008. Regional frequency distribution type of low flow in north of Iran by L-moments. Water Resources Management 22, 823-841, doi:10.1007/s11269-007-9194-8.

Modarres, R., Sarhadi, A., 2011. Statistically-based regionalization of rainfall climates of Iran. Global and Planetary Change 75, 67-75, doi:10.1016/j.gloplacha.2010.10.009.

Ngongondo, C.S., Xu, C.Y., Tallaksen, L.M., Alemaw, B., Chirwa, T., 2011. Regional frequency analysis of rainfall extremes in Southern Malawi using the index rainfall and L-moments approaches. Stochastic Environmental Research and Risk Assessment 25, 939-955, doi:10.1007/s00477-011-0480-x.

Norbiato, D., Borga, M., Sangati, M., Zanon, F., 2007. Regional frequency analysis of extreme precipitation in the eastern Italian Alps and the August 29, 2003 flash flood. Journal of Hydrology 345, 149-166, doi:10.1016/j.jhydrol.2007.07.009.

Noto, L.V., Loggia, G.L., 2009. Use of L-moments approach for regional flood frequency analysis in Sicily, Italy. Water Resources Management 23, 2207-2229, doi:10.1007/s11269-008-9378-x.

Nuñez, J.H., Verbist, K., Wallis, J.R., Schaefer, M.G., Morales, L, Cornelis, W.M., 2011. Regional frequency analysis for mapping drought events in north-central Chile. Journal of Hydrology 405 352-366, doi:10.1016/j.hydrol.2011.05.035.

Parida, B.P., Moalafhi, D.B., 2008. Regional rainfall frequency analysis for Botswana using L-Moments and radial basis function network. Physics and Chemistry of the Earth 33 614-620, doi:10.1016/j.pce.2008.06.011.

Perica, S., coauthors, 2013. NOAA Atlas 14: Precipitation-frequency atlas of the United States. Volume 8 Version 2.0: Midwestern States. National Oceanic and Atmospheric Administration/National Weather Service.

Pham, H.X., Shamseldin, A.Y., Melville, B.W., 2013. Statistical properties of partial duration series and its implication on regional frequency analysis. Journal of Hydrologic Engineering, doi:10.1061/(ASCE)HE.1943-5584.0000916.

R Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0.

Rao, A.R., Srinivas, V.V., 2006. Regionalization of watersheds by fuzzy cluster analysis. Journal of Hydrology 318, 57-79, doi:10.1016/j.jhydrol.2005.06.004.

Rianna, M., Ridolfi, E., Lorino, L., Alfonso, L., Montesarchio, V., Di Baldassarre, G., Russo, F., Napolitano, F., 2012. Definition of homogeneous regions through entropy theory, paper presented at 3rd STAHY International Workshop on Statistical Methods for Hydrology and Water Resources Management, International Association of Hydrological Sciences International Commission on Statistical Hydrology, Tunis, Tunisia.

Satyanarayana, P., Srinivas, V.V., 2009. Regional frequency analysis of annual precipitation in data-sparse regions using large-scale atmospheric variables, paper presented at Symposium HS.2 at the Joint IAHS and IAH Convention, International Association of Hydrological Sciences and International Association of Hydrogeologists, Hyderabad, India.

Sadri, S., Burn, D.H., 2011. A Fuzzy C-Means approach for regionalization using a bivariate homogeneity and discordancy approach. Journal of Hydrology 401, 231-239, doi:10.1016/j.jhydrol.2011.02.027.

Saf, B., 2010. Assessment of the effects of discordant sites on regional flood frequency analysis. Journal of Hydrology 380, 362-375, doi:10.1016/j.jhydrol.2009.11.011.

Scholz, F.W., Stephens, M.A., 1987. K-sample Anderson-Darling tests. Journal of the American Statistical Association 82 (399), 918-924, doi:10.1080/01621459.1987.10478517.

Seckin, N., Cobaner, M., Yurtal, R., Haktanir, T., 2013. Comparison of Artificial Neural Network methods with L-moments for estimating flood flow at ungauged sites: The case of East Mediterranean River Basin, Turkey. Water Resources Management 27, 2103-2124. doi:10.1007/s11269-013-0278-3.

Smithers, J.C., Schulze, R.E., 2001. A methodology for the estimation of short duration design storms in South Africa using a regional approach based on L-moments. Journal of Hydrology 241, 42-52, doi: 10.1016/S0022-1694(00)00374-7.

Santos, J.F., Portela, M.M., Pulido-Calvo, I., 2011. Regional frequency analysis of droughts in Portugal. Water Resources Management 25, 3537-3558, doi:10.1007/s11269-011-9869-z.

Srinivas, V.V., Tripathi, S., Rao, A.R., Govindaraju, R.S., 2008. Reigonal flod frequency analysis by combining self-organizing feature map and fuzzy clustering. Journal of Hydrology 348, 148-166, doi:10.1016/j.jhydrol.2007.09.046.

Stein, C., 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proc. Third Berkeley Symp. on Math. Statist. and Prob. 1, 197-206.

Szolgay, J., Parajka, J., Kohnová, S., Hlavĉová K., 2009. Comparison of mapping approaches of design annual maximum daily precipitation. Atmospheric Research 92, 289-307, doi:10.1016/j.atmosres.2009.01.009.

Um, M.J., Yun, H., Cho, W., Heo, J.H., 2010. Analysis of orographic precipitation on Jeju-island using regional frequency analysis and regression. Water Resources Management 24, 1461-1487, doi:10.1007/s11269-009-9509-z.

Viglione, A., Laio, F., Claps, P., 2007. A comparison of homogeneity tests for regional frequency analysis. Water Resources Research 43, 1241-1249, doi:10.1029/2006WR005095.

Vogel, R.M., Fennessey, N.M., 1993. L moment diagrams should replace product moment diagrams. Water Resources Research 29, 1745-1752, doi:10.1029/93WR00341.

Wang, Q.J., 1997. LH moments for statistical analysis of extreme events. Water Resources Research 33, 2841-2848, doi:10.1029/97WR02134.

Werick, W.J., Willeke, G.E., Guttman, N.B., Hosking, J.R.M., Wallis, J.R., 1994. National drought atlas developed. Eos Trans. AGU 75, 89-90.

Yang, T., Shao, Q., Hao, Z.C., Chen, X., Zhang, Z., Xu, C.Y., Sun, L., 2010. Regional frequency analysis and spatio-temporal pattern characterization of rainfall extremes in the Pearl River Basin, China. Journal of Hydrology 380, 386-405, doi:10.1016/j.jhydrol.2009.11.013.

Zrinji, Z., Burn, D.H., 1994. Flood frequency analysis for ungauged sites using a region of influence approach. Journal of Hydrology 153, 1-21, doi:10.1016/0022-1694(94)90184-8.

BIOGRAPHY

Michael J. Wright graduated from Thomas Jefferson High School for Science and Technology, Alexandria, Virginia, in 2004. He received his Bachelor of Arts with a double major in History and Neuroscience from the University of Virginia in 2008. He received his Master of Science in Civil and Infrastructure Engineering from George Mason University in 2011. He has been employed at the Institute for Water Resources in the Army Corps of Engineers since 2010.