WATER RESOURCES RESEARCH, VOL. 29, NO. 2, PAGES 271–281, FEBRUARY 1993

Some Statistics Useful in Regional Frequency Analysis

J. R. M. HOSKING AND J. R. WALLIS

IBM Research Division, Yorktown Heights, New York

Regional frequency analysis uses data from a number of measuring sites. A "region" is a group of sites each of which is assumed to have data drawn from the same frequency distribution. The analysis involves the assignment of sites to regions, testing whether the proposed regions are indeed homogeneous, and choice of suitable distributions to fit to each region's data. This paper describes three statistics useful in regional frequency analysis: a discordancy measure, for identifying unusual sites in a region; a heterogeneity measure, for assessing whether a proposed region is homogeneous; and a goodness-of-fit measure, for assessing whether a candidate distribution provides an adequate fit to the data. Tests based on the statistics provide objective backing for the decisions involved in regional frequency analysis. The statistics are based on the L moments [*Hosking*, 1990] of the at-site data.

INTRODUCTION

Regional Frequency Analysis: An Index Flood Procedure

A common problem in many aspects of environmental engineering is that of estimating the return period of rare geophysical events such as extreme floods or precipitations for a site or a group of sites. Regional frequency analysis uses data from several sites to estimate the frequency distribution of the observed data at each site.

Suppose that data are available at N sites in a region, with sample size n_i at site i, and let $Q_i(F)$ be the quantile of nonexceedance probability F at site i. The key assumption of an index flood procedure is that the region is homogeneous, that is, that the frequency distributions of the N sites are identical apart from a site-specific scaling factor, the index flood. We may then write

$$Q_i(F) = \mu_i q(F), \qquad i = 1, \cdots, N. \tag{1}$$

Here μ_i is the index flood. We shall take it to be the mean of the at-site frequency distribution, though any location parameter of the frequency distribution may be used instead, for example, *Smith* [1989] uses the quantile $Q(0.9)$. The remaining factor in (1), $q(F)$, is the regional quantile of nonexceedance probability F. The regional quantiles $q(F)$, $0 < F < 1$, form the "regional growth curve," which defines a dimensionless regional frequency distribution common to all sites.

The mean is naturally estimated by $\hat{\mu}_i = \bar{Q}_i$, the sample mean at site i. Other location estimators such as the median or a trimmed mean could be used instead.

The dimensionless rescaled data $q_{ij} = Q_{ij}/\hat{\mu}_i, j = 1, \cdots, n_i, i = 1, \cdots, N$, are the basis for estimating $q(F)$. It is usually assumed that the form of $q(F)$ is known apart from p undetermined parameters $\theta_1, \cdots, \theta_p$. We consider index flood procedures in which the parameters are estimated separately at each site, the site i estimate of θ_k being denoted by $\hat{\theta}_k^{(i)}$. The at-site estimates are combined to give regional estimates:

$$\hat{\theta}_k^{(R)} = \sum_{i=1}^{N} n_i \hat{\theta}_k^{(i)} \Big/ \sum_{i=1}^{N} n_i. \tag{2}$$

This is a weighted average, with the site i estimate given weight proportional to n_i because for regular statistical models the variance of $\hat{\theta}_k^{(i)}$ is inversely proportional to n_i. Substituting these estimates into $q(F)$ gives the estimated regional quantile $\hat{q}(F)$. This method of obtaining regional estimates is essentially that of *Wallis* [1980], except that the weighting proportional to n_i is a later addition, suggested by *Wallis* [1982]. Somewhat different methods were used by *Dalrymple* [1960] and *Natural Environment Research Council* [1975].

The site i quantile estimates are obtained by combining the estimates of μ_i and $q(F)$:

$$\hat{Q}_i(F) = \hat{\mu}_i \hat{q}(F). \tag{3}$$

L Moments

For a random variable X with cumulative distribution function F the quantities

$$\beta_r = E\{X[F(X)]^r\} \tag{4}$$

are probability-weighted moments, defined by *Greenwood et al.* [1979] and used by them to estimate the parameters of probability distributions. *Hosking* [1986, 1990] defined L moments to be linear combinations of probability-weighted moments:

$$\lambda_{r+1} = \sum_{k=0}^{r} p_{r,k}^* \beta_k, \tag{5}$$

where

$$p_{r,k}^* = (-1)^{r-k} \binom{r}{k} \binom{r+k}{k}. \tag{6}$$

L moment ratios are the quantities

$$\tau_r = \lambda_r/\lambda_2, \qquad r = 3, 4, \cdots. \tag{7}$$

Paper number 92WR01980.
0043-1397/93/92WR-01980$05.00

L moments are similar to but more convenient than probability-weighted moments because they are more easily interpretable as measures of distributional shape. In particular, λ_1 is the mean of the distribution, a measure of location; λ_2 is a measure of scale; and τ_3 and τ_4 are measures of skewness and kurtosis, respectively. The L CV, $\tau = \lambda_2/\lambda_1$, is analogous to the usual coefficient of variation.

The foregoing quantities are defined for a probability distribution but, in practice, must often be estimated from a finite sample. Let $x_1 \le x_2 \le \cdots \le x_n$ be the ordered sample. Define

$$l_{r+1} = \sum_{k=0}^{r} p^*_{r,k} b_k, \tag{8}$$

where

$$b_r = n^{-1} \sum_{j=1}^{n} \frac{(j-1)(j-2)\cdots(j-r)}{(n-1)(n-2)\cdots(n-r)} x_j. \tag{9}$$

Then l_r is an unbiased estimator of λ_r. The estimators $t_r = l_r/l_2$ of τ_r and $t = l_2/l_1$ of τ are consistent but not unbiased. The quantities l_1, l_2, t_3, and t_4 are useful summary statistics of a sample of data. They can be used to judge which distributions are consistent with a given data sample [*Hosking*, 1990, section 3.5]. They can also be used to estimate parameters when fitting a distribution to a sample, by equating the sample and population L moments [*Hosking*, 1990, section 4.1].

L moments are applicable to regional frequency analysis: the parameters $\theta_1, \cdots, \theta_p$ are taken to be the L moments $\lambda_1, \tau, \tau_3, \cdots, \tau_p$ and are estimated by the corresponding sample L moments of the at-site q_{ij} statistics. Recent research [*Hosking et al.*, 1985; *Lettenmaier and Potter*, 1985; *Wallis and Wood*, 1985; *Lettenmaier et al.*, 1987; *Hosking and Wallis*, 1988; *Potter and Lettenmaier*, 1990] has shown that index flood procedures based on probability-weighted moments or L moments yield robust and accurate quantile estimates.

We use the following notation for regional L moments. Assume that the analysis concerns a group of N sites. Sample L moment ratios at site i are denoted by $t^{(i)}$, $t_3^{(i)}$, $t_4^{(i)}$, etc. Superscript (i) will be omitted if possible without confusion. Group average L moment ratios, with sites weighted proportionally to their record lengths, are

$$\bar{t} = \sum_{i=1}^{N} n_i t^{(i)} \Big/ \sum_{i=1}^{N} n_i, \qquad \bar{t}_r = \sum_{i=1}^{N} n_i t_r^{(i)} \Big/ \sum_{i=1}^{N} n_i,$$

$$r = 3, 4, \cdots.$$

Stages in Regional Frequency Analysis

Regional frequency analysis typically involves four stages, three of which involve subjective judgement. This paper describes statistics, based on L moments, that provide objective support for these judgements.

Screening of the data. As with any statistical analysis, the first stage is a close inspection of the data, so that gross errors and inconsistencies can be eliminated. For screening the data we describe a discordancy measure D_i. This identifies unusual sites, those whose at-site sample L moments are markedly different from those of the other sites in the data set. The discordancy measure provides an initial screening of the data and indicates sites where the data may merit close examination.

Identification of homogeneous regions. The next stage in regional frequency analysis is the assignment of the sites to regions. A "region," a set of sites whose frequency distributions are (after appropriate scaling) approximately the same, is the fundamental unit of regional frequency analysis. For identifying homogeneous regions we must be able to test whether a proposed region is acceptably close to homogeneous. This can be done by calculating summary statistics of the at-site data and comparing the between-site variability of these statistics with what would be expected of a homogeneous region. We describe a heterogeneity measure H that performs this test: the summary statistics that it uses are the sample L moments.

Choice of a regional frequency distribution. After a region has been identified the final stage in the specification of the statistical model is the choice of an appropriate frequency distribution from which to obtain $q(F)$ in (1). For this purpose we describe a goodness-of-fit measure Z. This indicates whether a candidate distribution gives a good fit to a region's data: specifically, whether the regional average L moments are consistent with those of the fitted distribution.

Estimation of the regional frequency distribution. The parameters of the regional frequency distribution are estimated by combining the at-site estimates to give the regional average (2). Algorithms for estimating a distribution's parameters from its L moments have been given by *Hosking* [1991]. No subjective judgement is involved.

Use of the D_i, H, and Z statistics in formal significance tests requires knowledge of the sampling distributions of the statistics. These distributions are not accurately known and depend in a complex way on the record lengths at the sites, the degree and nature of cross correlation between the sites' data, and the true regional frequency distribution. They can be derived theoretically under the assumption that sample L moments of at-site data are independent and normally distributed, and they can be obtained by Monte Carlo simulation for any specification of the regional frequency distribution and record lengths. On the basis of these considerations we suggest numerical values of the statistics that can be used in the decision-making process. However, these values should be regarded as rough guidelines rather than as strict decision criteria.

DISCORDANCY MEASURE

Aim

Given a group of sites, the aim is to identify those sites that are grossly discordant with the group as a whole. Discordancy is measured in terms of the L moments of the sites' data.

Heuristic Description

Regard the L moments (L CV, L skewness, and L kurtosis) of a site as a point in three-dimensional space. A group of sites will yield a cloud of such points. Flag as discordant any point that is far from the center of the cloud.

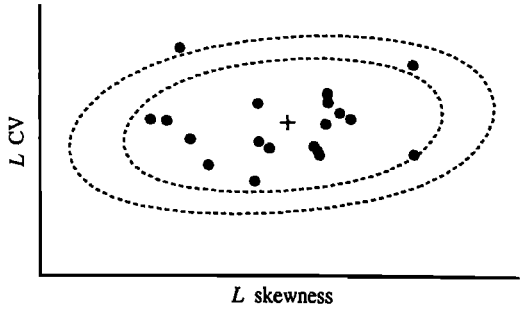

Fig. 1. Definition sketch for discordancy.

"Far" is interpreted in such a way as to allow for the correlation between the sampled L moments.

For example, consider Figure 1. For convenience we consider L CV and L skewness only. The center of the cloud of points, marked by a plus, is the point whose coordinates are the group average values of L CV and L skewness. We construct concentric ellipses with major and minor axes chosen to give the best fit to the data (as determined by the sample covariance matrix of the sites' L moments). "Discordant" points are those that lie outside the outermost ellipse.

Formal Definition

Let $\mathbf{u}_i = [t^{(i)}\ t_3^{(i)}\ t_4^{(i)}]^T$ be a vector containing the t, t_3, and t_4 values for site i. Let

$$\bar{\mathbf{u}} = N^{-1} \sum_{i=1}^{N} \mathbf{u}_i$$

be the (unweighted) group average. Define the sample covariance matrix

$$\mathbf{S} = (N-1)^{-1} \sum_{i=1}^{N} (\mathbf{u}_i - \bar{\mathbf{u}})(\mathbf{u}_i - \bar{\mathbf{u}})^T.$$

Define the discordancy measure for site i as

$$D_i = \tfrac{1}{3}(\mathbf{u}_i - \bar{\mathbf{u}})^T \mathbf{S}^{-1} (\mathbf{u}_i - \bar{\mathbf{u}}).$$

Large values of D_i indicate sites that are most discordant from the group as a whole and are most worthy of investigation for the presence of data errors.

Use

Two uses for the discordancy measure are envisaged. First, at the outset of the analysis it may be applied to a large group of sites, all those within some large geographical area. The idea is that sites with gross errors in their data will stand out from the other sites and be flagged as discordant. Sites flagged as discordant at this stage should therefore be closely scrutinized for errors in the recording or transcription of data or for sources of unreliability in the data (such as a recording gage having been moved or for man-induced changes of the site's frequency distribution over time).

Later in the analysis, when homogeneous regions have been at least tentatively identified, the discordancy measure can be calculated for each site in the proposed region. If any site is then discordant with the region as a whole, the possibility of moving that site to another region should be considered. It must be borne in mind, however, that a site's L moments may differ by chance alone from those of other physically similar site; for example, an extreme but localized meteorological event may have affected only a few sites in a region. If such an event is approximately equally likely to affect any of the sites in the future, then it is correct to treat the entire group of sites as a homogeneous region, even though some sites may appear to be discordant with the region as a whole.

Notes

D_i is a standard discordancy measure for multivariate observations. *Wilks* [1963] proposed an outlier measure that is equivalent to $\max_i D_i$. For a univariate observation, D_i reduces to $(u_i - \bar{u})^2/s^2$, the squared studentized residual; the maximum absolute studentized residual has been widely used as an outlier measure since *Thompson* [1935]. The average of D_i over all sites is 1.

It is not easy to choose a single value of D_i that can be used as a criterion for deciding whether a site is unusual. If it is assumed that the $\mathbf{u}_i$ are drawn from independent identical multivariate normal distributions, then for large regions the statistics $3D_i$ have, approximately, independent chi-square distributions with 3 degrees of freedom, and approximately 3% of the D_i values exceed 3. Thus we tentatively suggest $D_i \geq 3$ as a criterion for declaring a site to be unusual. However, it is advisable to examine the data for the sites with the largest D_i values, regardless of the magnitude of these values.

The use of an unweighted average in the definition of $\bar{\mathbf{u}}$ is preferred to the weighted average used in the homogeneity and goodness-of-fit measures described below. A weighted average allows for greater variability in short records and would permit a short-record site to be further from the group average before being flagged as discordant. Here, however, we are at an early stage of the analysis and we want to isolate the unusual sites and their potential data errors regardless of the record length.

Example

We use an illustrative set of annual maximum precipitation data obtained from the U.S. Historical Climatology Network [*Karl et al.*, 1990]. The data are for the "North Cascades" region, one of 23 climatic divisions of the continental United States used by *Plantico et al.* [1990]. Data are available for 19 sites. Record lengths and L moment ratios are given in Table 1 and illustrated in Figure 2.

The D_i values for the North Cascades data are also given in Table 1. There is no evidence of gross errors in the data. The largest D_i value is 2.63 for site 6, which has high L CV and low L skew. In our experience this is not a particularly large value. It might be worthwhile to shift this site to another region if there are physical grounds for doing so, but the entire group of sites is at this stage a plausible candidate for being a homogeneous region.

HETEROGENEITY MEASURE

Aim

The aim is to estimate the degree of heterogeneity in a group of sites and to assess whether they might reasonably

TABLE 1. Summary Statistics for the North Cascades Precipitation Data Set

Site	HCN Site Code	n	Mean	L CV	t_3	t_4	t_5	D_i
1	350304	98	19.69	0.1209	0.0488	0.1433	−0.0004	0.60
2	351433	59	62.58	0.0915	0.0105	0.1569	0.0020	1.02
3	351862	90	40.85	0.1124	0.0614	0.1541	−0.0058	0.38
4	351897	61	46.05	0.1032	0.0417	0.1429	−0.0022	0.23
5	352997	65	45.02	0.0967	−0.0134	0.1568	0.0173	0.93
6	353445	86	31.04	0.1328	−0.0176	0.1206	0.0235	2.63
7	353770	78	80.14	0.1008	0.0943	0.1967	0.0856	2.12
8	356907	72	41.31	0.1143	0.0555	0.1210	0.0487	0.45
9	357169	67	30.59	0.1107	0.0478	0.1371	0.0316	0.11
10	357331	99	32.93	0.1179	0.0492	0.0900	0.0225	1.61
11	357354	49	17.56	0.1308	0.0940	0.1273	0.0352	2.08
12	358466	61	69.52	0.1119	−0.0429	0.0927	−0.0061	1.52
13	450945	69	47.65	0.1018	0.0435	0.1446	−0.0056	0.31
14	451233	73	102.50	0.1025	0.0182	0.1047	−0.0221	1.30
15	453284	70	52.41	0.1054	−0.0224	0.1664	0.0035	1.58
16	454764	66	79.70	0.1174	0.0124	0.1317	−0.0176	0.29
17	454769	59	44.64	0.1115	−0.0346	0.1032	0.0083	1.04
18	457773	74	58.66	0.1003	0.0446	0.1450	−0.0379	0.43
19	458773	82	39.02	0.1046	0.0128	0.1583	0.0443	0.38

HCN, Historical Climatology Network.

be treated as a homogeneous region. Specifically, the heterogeneity measure compares the between-site variations in sample L moments for the group of sites with what would be expected for a homogeneous region.

Heuristic Description

In a homogeneous region all sites have the same population L moments. Their sample L moments will, however, be different, owing to sampling variability. Thus a natural question to ask is whether the between-site dispersion of the sample L moments for the group of sites under consideration is larger than would be expected of a homogeneous region. The situation is sketched in Figure 3. Let us consider how to measure the "between-site dispersion of sample L moments" and how to establish "what would be expected of a homogeneous region."

A visual assessment of the dispersion of the at-site L moments can be obtained by plotting them on a graph of L skewness versus L CV or L kurtosis. An alternative and simple measure of the dispersion of the sample L moments is the standard deviation of the at-site L CVs. It is reasonable to concentrate on L CV, since between-site variation in L CV has a much larger effect than variation in L skewness or L kurtosis on the variance of the estimates of all quantiles $Q_i(F)$, except those in the far tail of the distribution with $F \geq 0.998$ [*Hosking et al.*, 1985]. To allow for the greater variability of L moments in small samples, averages should be weighted proportionally to the sites' record lengths.

To establish what "would be expected" we use simulation. By repeated simulation of a homogeneous region with sites having record lengths the same as those of the observed data, we obtain the mean and standard deviation of the chosen dispersion measure. To compare the observed and simulated dispersions, an appropriate statistic is

$$\frac{\text{(observed dispersion)} - \text{(mean of simulations)}}{\text{(standard deviation of simulations)}}.$$

A large positive value of this statistic indicates that the observed L moments are more dispersed than is consistent with the hypothesis of homogeneity.

Finally, we must specify the region used in the simulations. If the observed sites do form a homogeneous region, this region's population L moments are likely to be close to the average of the sample L moments of the observed data. To avoid committing ourselves to a particular two- or three-parameter distribution, in the simulations we use a more general distribution such as the Wakeby [*Houghton*, 1978] or kappa [*Hosking*, 1988]. In this paper we have used a kappa distribution, a four-parameter distribution with quantile function

$$Q(F) = \zeta + \alpha\{1 - [(1 - F^h)/h]^k\}/k.$$

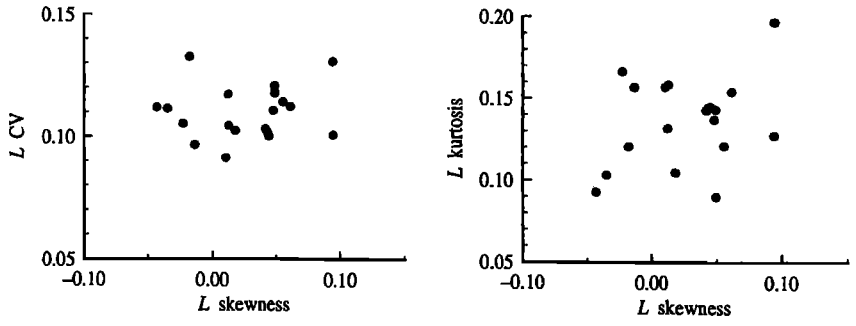

Fig. 2. L moment ratios of the North Cascades sites.

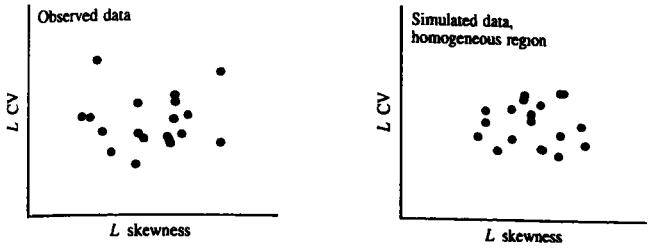

Fig. 3. Definition sketch for heterogeneity.

It includes as special cases the generalized logistic ($h = -1$), generalized extreme-value ($h \to 0$), and generalized Pareto ($h = +1$) distributions. Its L moments can be chosen to match the group average L CV, L skewness, and L kurtosis of the observed data, using an algorithm from *Hosking* [1991]. The feasible values of L skewness and L kurtosis for the kappa distribution are illustrated in Figure 4. The distribution has a finite upper bound if $k > 0$ (the region below the $k = 0$ line in Figure 4) and a hyperbolically decreasing upper tail if $k < 0$.

Formal Definition

Calculate the weighted standard deviation of the at-site sample L CVs:

$$V = \sum_{i=1}^{N} n_i (t^{(i)} - \bar{t})^2 \Big/ \sum_{i=1}^{N} n_i.$$

Fit a kappa distribution to the group average L moments 1, $\bar{t}$, $\bar{t}_3$, $\bar{t}_4$.

Simulate a large number N_{sim} of regions from this kappa world; the regions are homogeneous and have no cross correlation or serial correlation, and sites have the same record lengths as their real world counterparts. For each simulated region calculate V. From the simulations determine the mean and standard deviation of the N_{sim} values of V. Call these μ_V and σ_V.

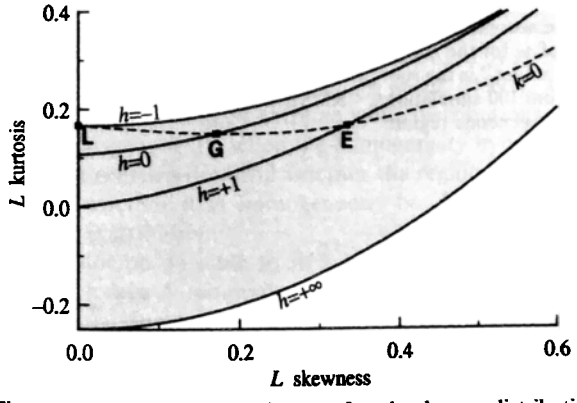

Fig. 4. L moment ratio diagram for the kappa distribution. Shaded area shows the L skewness and L kurtosis values attained by the kappa distribution. Special cases include the logistic (L), Gumbel (G), exponential (E), generalized logistic ($h = -1$), generalized extreme value ($h = 0$) and generalized Pareto ($h = +1$) distributions.

Calculate the heterogeneity measure

$$H = \frac{(V - \mu_V)}{\sigma_V}.$$

Declare the region to be heterogeneous if H is sufficiently large. We suggest that the region be regarded as "acceptably homogeneous" if $H < 1$, "possibly heterogeneous" if $1 \le H < 2$, and "definitely heterogeneous" if $H \ge 2$.

Performance

The performance of H as a heterogeneity measure was assessed in a series of Monte Carlo simulation experiments. For each of a number of artificial regions, 100 replications were made of data from the region, and the accuracy of quantile estimates and values of the heterogeneity measure H were calculated. N_{sim}, the number of regions simulated in the computation of H, was 500. Simulation results are given in Table 2. Regions were specified by the number of sites in the region, the record lengths at each site, and the frequency distribution at each site. Frequency distributions were generalized extreme-value distributions at each site and were specified by their L moments τ and τ_3; the at-site mean was, without loss of generality, set to 1 at each site. Three types of region were used in the simulations: homogeneous; heterogeneous, with L CV and L skewness varying linearly from site 1 through site N; and "bimodal," with half the sites having one distribution and half another. These regions test the ability of H to detect heterogeneity both when the frequency distributions vary smoothly from site to site and when there is a sharp difference between the frequency distributions at two subsets of sites. The base region for the simulations has $N = 21$, $n_i = 30$ at each site and regional average values of 0.2 for both τ and τ_3. Variations on this region include changing N to 6 or 11, changing n_i to 60, and changing the regional average τ to 0.1 or 0.3 with appropriate changes in τ_3. Both homogeneous and heterogeneous variants of these regions were simulated.

Quantile estimates were obtained by regional analysis, fitting a generalized extreme-value distribution to the regional average L moments. The root-mean-square error (RMSE) of the quantile estimates $\hat{Q}_i(F)$ was calculated for each site and was divided by the true quantile to obtain a relative RMSE. The "RMSE of quantiles" columns in Table 2 give the average, over all sites in the region, of this relative RMSE. For purposes of estimating extreme quantiles, we consider the true measure of heterogeneity to be the amount by which the error in the quantile estimates is greater for the observed region than for a homogeneous region with the same values for N, the n_i, and the regional average L moments. This error cannot be calculated for observed data because the underlying frequency distribution is unknown, but it can be found for simulated data. It is the "RMSE relative to homogeneous region" entry in Table 2.

Figure 5 summarizes the relationship between the average H value for a simulated world and the RMSE of quantile estimates for that world relative to a homogeneous region. In general, the relationship is fairly well defined, showing that H is indeed a reasonable proxy for the likely error in quantile

TABLE 2. Simulation Results for the Heterogeneity Measure H

World Type	τ Average	τ Range	τ_3 Average	τ_3 Range	N	n	RMSE of Quantiles 0.01	0.10	0.99	0.999	RMSE Relative to Homogeneous Region 0.01	0.10	0.99	0.999	Average H
Homogeneous	0.20	0	0.20	0	6	30	0.104	0.072	0.103	0.159	1.00	1.00	1.00	1.00	0.11
	0.20	0	0.20	0	11	30	0.090	0.071	0.089	0.124	1.00	1.00	1.00	1.00	0.10
	0.20	0	0.20	0	21	30	0.080	0.070	0.081	0.104	1.00	1.00	1.00	1.00	0.06
	0.10	0	0.05	0	21	30	0.039	0.033	0.036	0.042	1.00	1.00	1.00	1.00	0.00
	0.30	0	0.30	0	21	30	0.143	0.115	0.132	0.168	1.00	1.00	1.00	1.00	0.02
Heterogeneous 30%	0.20	0.06	0.20	0.06	6	30	0.140	0.092	0.129	0.201	1.35	1.28	1.25	1.26	0.91
	0.20	0.06	0.20	0.06	11	30	0.125	0.088	0.114	0.167	1.39	1.24	1.28	1.35	1.08
	0.20	0.06	0.20	0.06	21	30	0.115	0.086	0.106	0.148	1.44	1.23	1.31	1.42	1.19
	0.10	0.03	0.05	0.06	21	30	0.052	0.040	0.053	0.068	1.33	1.21	1.47	1.62	1.42
	0.30	0.09	0.30	0.09	21	30	0.199	0.145	0.162	0.227	1.39	1.26	1.23	1.35	1.07
Hetergeneous 50%	0.20	0.10	0.20	0.10	6	30	0.186	0.118	0.164	0.256	1.79	1.64	1.59	1.61	2.09
	0.20	0.10	0.20	0.10	11	30	0.168	0.112	0.147	0.220	1.87	1.58	1.65	1.77	2.51
	0.20	0.10	0.20	0.10	21	30	0.157	0.108	0.138	0.200	1.96	1.54	1.70	1.92	2.96
	0.10	0.05	0.05	0.10	21	30	0.068	0.049	0.072	0.096	1.74	1.48	2.00	2.29	3.53
	0.30	0.15	0.30	0.15	21	30	0.271	0.184	0.204	0.303	1.90	1.60	1.55	1.80	2.49
Homogeneous	0.20	0	0.20	0	6	60	0.071	0.051	0.074	0.112	1.00	1.00	1.00	1.00	−0.02
	0.20	0	0.20	0	11	60	0.061	0.050	0.062	0.087	1.00	1.00	1.00	1.00	0.28
	0.20	0	0.20	0	21	60	0.057	0.050	0.057	0.071	1.00	1.00	1.00	1.00	0.07
Heterogeneous 30%	0.20	0.06	0.20	0.06	6	60	0.116	0.075	0.108	0.166	1.63	1.47	1.46	1.48	1.55
	0.20	0.06	0.20	0.06	11	60	0.104	0.071	0.095	0.140	1.70	1.42	1.53	1.61	2.16
	0.20	0.06	0.20	0.06	21	60	0.098	0.070	0.088	0.126	1.72	1.40	1.54	1.77	2.41
Heterogeneous 50%	0.20	0.10	0.20	0.10	6	60	0.166	0.103	0.147	0.228	2.34	2.02	1.99	2.04	3.53
	0.20	0.10	0.20	0.10	11	60	0.151	0.097	0.131	0.199	2.48	1.94	2.11	2.29	4.51
	0.20	0.10	0.20	0.10	21	60	0.143	0.094	0.124	0.184	2.51	1.88	2.18	2.59	5.45
Heterogeneous 30%	0.20	0.06	0.20	0.06	21	(varies[a])	0.125	0.096	0.110	0.149	1.45	1.26	1.28	1.38	1.37
	0.20	0.06	0.20	0.06	21	(varies[b])	0.119	0.088	0.113	0.158	1.38	1.17	1.31	1.46	0.80
	0.20	0.06	0.20	0.06	21	(varies[c])	0.120	0.092	0.111	0.150	1.41	1.24	1.32	1.44	1.63
	0.20	0.06	0.20	0.06	21	(varies[d])	0.119	0.091	0.110	0.151	1.35	1.18	1.25	1.37	0.79
Homogeneous	0.20	0	0.20	0	2	30	0.155	0.080	0.145	0.254	1.00	1.00	1.00	1.00	0.00
	0.20	0	0.20	0	4	30	0.115	0.074	0.118	0.192	1.00	1.00	1.00	1.00	−0.01
	0.20	0	0.20	0	10	30	0.092	0.070	0.090	0.129	1.00	1.00	1.00	1.00	0.11
	0.20	0	0.20	0	20	30	0.080	0.069	0.080	0.102	1.00	1.00	1.00	1.00	0.13
Bimodal 20%	0.20	0.04	0.20	0.04	2	30	0.183	0.100	0.165	0.284	1.18	1.25	1.14	1.12	0.62
	0.20	0.04	0.20	0.04	4	30	0.149	0.093	0.143	0.231	1.30	1.26	1.21	1.20	0.71
	0.20	0.04	0.20	0.04	10	30	0.131	0.090	0.120	0.178	1.42	1.29	1.33	1.38	1.00
	0.20	0.04	0.20	0.04	20	30	0.126	0.091	0.111	0.158	1.57	1.32	1.39	1.55	1.56
Bimodal 30%	0.20	0.06	0.20	0.06	2	30	0.213	0.119	0.188	0.319	1.37	1.49	1.30	1.26	1.27
	0.20	0.06	0.20	0.06	4	30	0.183	0.113	0.169	0.271	1.59	1.53	1.43	1.41	1.48
	0.20	0.06	0.20	0.06	10	30	0.169	0.110	0.150	0.226	1.84	1.57	1.67	1.75	2.03
	0.20	0.06	0.20	0.06	20	30	0.166	0.112	0.142	0.208	2.07	1.62	1.77	2.04	3.02
Bimodal 50%	0.20	0.10	0.20	0.10	2	30	0.289	0.167	0.248	0.414	1.86	2.09	1.71	1.63	2.83
	0.20	0.10	0.20	0.10	4	30	0.267	0.162	0.234	0.374	2.32	2.19	1.98	1.95	3.30
	0.20	0.10	0.20	0.10	10	30	0.258	0.160	0.220	0.336	2.80	2.29	2.44	2.60	4.59
	0.20	0.10	0.20	0.10	20	30	0.256	0.162	0.214	0.323	3.20	2.35	2.67	3.17	6.57

"World type" is homogeneous, heterogeneous (L CV τ and L skewness τ_3 increase linearly from site 1 to site N), or bimodal (half the sites have high τ and τ_3, the other half have low τ and τ_3). All worlds have generalized extreme value frequency distributions. "Heterogeneous 30%" means that the (range of τ) ÷ (average τ) is 0.3. The second and third columns are the average and the range of τ for the region. The fourth and fifth columns are the average and the range of τ_3 for the region. N is the number of sites in the region, and n is the record length at each site. This is the same for all sites. "RMSE of quantiles" is the root mean square error of the estimated quantile, divided by the true value of the quantile; tabulated values are calculated from 100 simulations. "RMSE relative to homogeneous region" is RMSE of quantile divided by the same RMSE for the corresponding homogeneous region.

[a]Here $n = 50, 48, \cdots, 10$ at sites $i = 1, 2, \cdots, 21$.

[b]Here $n = 10, 12, \cdots, 50$.

[c]Here $n = 50, 46, \cdots, 14, 10, 14, \cdots 46, 50$.

[d]Here $n = 10, 14, \cdots, 46, 50, 46, \cdots, 14, 10$.

estimates. The $H = 1$ level is reached when the RMSE is 20–40% higher than for a homogeneous region; $H = 2$ is reached when the RMSE is 40–80% higher than for a homogeneous region. The main doubt concerning the H measure is excessive dependence on the number of sites in the region, particularly when the focus is on estimating quantiles that are not really extreme. In the "bimodal" regions, for example, the RMSE relative to a homogeneous region for the 0.1 quantile varies very little as the number N of sites in the region varies, but the average H value decreases steadily as N decreases. This means that H is better at indicating heterogeneity in large regions but has a tendency to give false indications of homogeneity for small regions. This effect is less marked at more extreme quantiles.

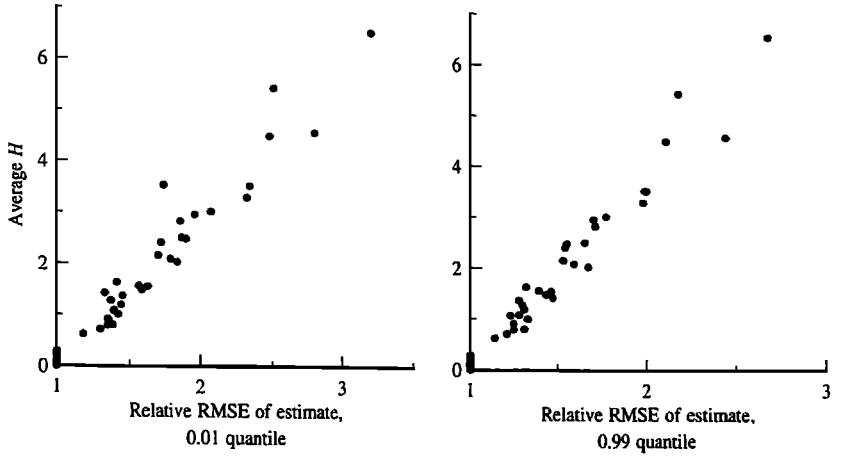

Fig. 5. Average H value and relative RMSE of quantile estimates for the simulated worlds of Table 2.

Use

To assess the heterogeneity of a proposed region, calculate H and compare it with the criteria given above. If the region is not acceptably homogeneous, some redefinition of the region should be considered. The region could be divided into two or more subregions, some sites could be removed from the region, or a completely different assignment of sites to regions could be tried.

Notes

The assessment of heterogeneity by comparison of L moments of observed data with those of data simulated from a homogeneous world has previously been made informally, using only one or two simulated worlds [*Wallis*, 1989; *Hosking*, 1990, Figure 7; *Pearson*, 1991; *Pilon et al.*, 1992; *Pilon and Adamowski*, 1992]. The use of H with a large number of simulations is a less subjective variant of this approach.

The value of N_{sim} should be chosen to achieve reliable estimates μ_V and σ_V. From simulations we judge that a value of $N_{\text{sim}} = 500$ should be adequate.

The use of a kappa distribution in the simulations is, as noted above, intended to avoid too early a commitment to a particular distribution as the parent of the observed data. This contrasts with homogeneity tests proposed by *Acreman and Sinclair* [1986] and *Chowdhury et al.* [1991], which involve fitting generalized extreme value distributions to the data. Using these tests, when the homogeneity hypothesis is rejected, it remains doubtful whether the region is heterogeneous or whether it is homogeneous but has some other frequency distribution.

It may not be possible to fit a kappa distribution to the regional average L moments. This occurs if $\bar{t}_4$ is too large relative to $\bar{t}_3$. In such cases we recommend that the generalized logistic distribution, a special case of the kappa distribution with the h shape parameter equal to -1, be used for the simulated world.

It is possible to construct heterogeneity measures in which V is replaced by other measures of between-site variability of sample L moments. We considered a measure based on L CV and L skewness,

$$V_2 = \sum_{i=1}^{N} n_i[(t^{(i)} - \bar{t})^2 + (t_3^{(i)} - \bar{t}_3)^2]^{1/2} \Big/ \sum_{i=1}^{N} n_i,$$

and a measure based on L skewness and L kurtosis,

$$V_3 = \sum_{i=1}^{N} n_i[(t_3^{(i)} - \bar{t}_3)^2 + (t_4^{(i)} - \bar{t}_4)^2]^{1/2} \Big/ \sum_{i=1}^{N} n_i.$$

V_2 and V_3 are the weighted average distance from the site to the group-weighted mean on graphs of t versus t_3 and of t_3 versus t_4, respectively. For both real world data and artificial simulated regions, H statistics based on V_2 and V_3 lack power to discriminate between homogeneous and heterogeneous regions: they rarely yield H values larger than 2 even for grossly heterogeneous regions. The H statistic based on V has much better discriminatory power. Similar results have been reported by *Lu* [1991]. The measure V is, of course, insensitive to heterogeneity that takes the form of sites having equal L CV but different L skewness, but this form of heterogeneity has relatively little effect on the accuracy of quantile estimates except very far into the extreme tails of the distribution, and it is in any case rare in practice, since sites with high L skewness tend to have high L CV too.

The H statistic is constructed like a significance test of the hypothesis that the region is homogeneous. However, we do not recommend that it be used in this way. Significance levels obtained from such a test would be accurate only under special assumptions: that the data are independent both serially and between sites and that the true regional distribution is kappa. We need to define a heterogeneity measure for regions that may not satisfy these assumptions, so we prefer not to use H as a significance test. It would be possible to generate simulated data which are correlated, but this would require much more computing time. A significance test is of doubtful utility anyway, since even a moderately heterogeneous region can provide quantile estimates of sufficient accuracy for practical purposes. Thus a test of exact homogeneity is of little interest.

The criteria $H = 1$ and $H = 2$ are somewhat arbitrary but

are believed to be useful guidelines. If H were used as a significance test, then the criterion for rejection of the hypothesis of homogeneity at the significance level 10%, assuming normality for the distribution of V, would be $H = 1.28$. In comparison a criterion of $H = 1$ may seem very strict, but as noted above, we do not seek to use H in a significance test. From the simulation results in Table 2 an H value of 1 would typically arise from a region sufficiently heterogeneous that quantile estimates for it are 20–40% less accurate than for a homogeneous region with the same regional average L moments. For such a region it is still likely that regional estimation will yield much more accurate quantile estimates than at-site estimation, but it is possible that subdividing the region or removing a few sites from it may reduce its heterogeneity. We therefore regard this amount of heterogeneity as being on the borderline of whether a worthwhile increase in the accuracy of quantile estimates could be achieved by redefining the region, and thus we regard $H = 1$ as the limit at which seeking to redefine the region may be advantageous. Similarly, we regard $H = 2$ as a point at which redefining the region, if the available explanatory variables permit it, is very likely to be beneficial.

The validity of H as a heterogeneity measure is compromised if the selection of regions is based on sample L moments, for then the same data are being used both to choose regions and to test their homogeneity. One could, for example, define a region to consist of all sites with sample L CV within a certain small range. Such a region might yield a small value of H, but this would reflect only the pattern of noise, or sampling variability, in the data and have no physical significance. Valid use of H requires that assignment of sites to regions be based on external explanatory variables such as the physical characteristics or geographical location of the sites.

Example

For the North Cascades data the heterogeneity measure V is 0.0104. The group average L moments are

$$\bar{t} = 0.1103, \quad \bar{t}_3 = 0.0279, \quad \bar{t}_4 = 0.1366,$$

and the parameters of the fitted kappa distribution are

$$\xi = 0.9542, \quad \alpha = 0.1533, \quad k = 0.1236, \quad h = -0.2955.$$

Five hundred simulations were made of this kappa world. The V measures for the simulated worlds had an average of 0.0096 and a standard deviation of 0.0016. The calculated heterogeneity measure H is thus $(0.0104 - 0.0096)/0.0016 = 0.56$, and the region is acceptably homogeneous.

GOODNESS-OF-FIT MEASURE

Aim

Given a set of sites that constitute a homogeneous region, the aim is to test whether a given distribution fits the data acceptably closely. A related aim is to choose, from a number of candidate distributions, the one that gives the best fit to the data.

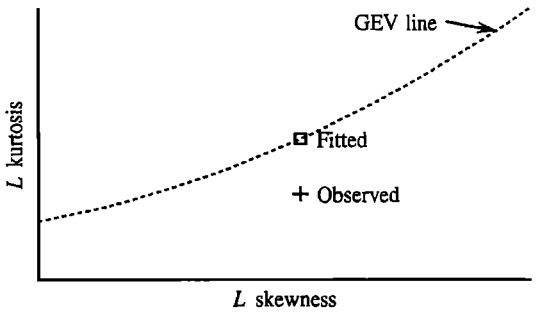

Fig. 6. Definition sketch for goodness of fit.

Heuristic Description

Assume that the region is acceptably close to homogeneous. Choice of a distribution for regions that are not homogeneous is discussed in the section on use below. The L moments of the sites in a homogeneous region are well summarized by the regional average; the scatter of the individual sites' L moments about the regional average represents no more than sampling variability. In most cases the distribution being tested will have location and scale parameters which can be chosen to match the regional average mean and L CV. The goodness of fit will therefore be judged by how well the L skewness and L kurtosis of the fitted distribution match the regional average L skewness and L kurtosis of the observed data. Fifth- or higher-order L moments could, in principle, be used too, but we have not found it necessary to do so.

To obtain a goodness of fit measure for a three-parameter distribution we argue as follows. Assume for convenience that the candidate distribution is generalized extreme-value (GEV) and that sample L skewness and L kurtosis are exactly unbiased. The GEV distribution fitted by the method of L moments has L skewness equal to the regional average L skewness. We therefore judge the quality of fit by the difference between the L kurtosis τ_4^{GEV} of the fitted GEV distribution and the regional average L kurtosis $\bar{t}_4$; see Figure 6. To assess the significance of this difference we compare it with the sampling variability of $\bar{t}_4$. Let σ_4 denote the standard deviation of $\bar{t}_4$, which we can obtain by repeated simulation of a homogeneous region with a GEV frequency distribution and sites having record lengths the same as those of the observed data. Then

$$Z^{\mathrm{GEV}} = (\bar{t}_4 - \tau_4^{\mathrm{GEV}})/\sigma_4$$

is a goodness of fit measure: small values of Z^{GEV} are consistent with the GEV being the true underlying frequency distribution for the region.

A difficulty with the procedure just described is that a separate set of simulations must be made for each candidate distribution to obtain the appropriate σ_4 value. In practice, it should be sufficient to assume that σ_4 is the same for all of the candidate three-parameter distributions; this is reasonable because all of the fitted distributions have the same L skewness and are therefore likely to resemble each other to a large extent. Given this assumption, it is also reasonable to assume that the best fitting kappa distribution has a σ_4 value close to those of the candidate distributions. Thus σ_4 can be obtained by repeated simulations of a kappa world; these

TABLE 3. Simulation Results for the Goodness-of-Fit Measure Z

					Percent Accepted				Percent Chosen			
τ	τ_3	N	n	World	GLO	GEV	LN3	PE3	GLO	GEV	LN3	PE3
0.10	0.05	21	30	GLO	74	3	7	6	91	0	9	0
0.10	0.05	21	30	GEV	2	90	82	85	1	61	23	15
0.10	0.05	21	30	LN3	7	82	89	89	5	36	43	15
0.10	0.05	21	30	PE3	6	85	89	90	4	39	41	15
0.20	0.20	21	30	GLO	80	25	16	2	89	10	1	0
0.20	0.20	21	30	GEV	34	95	89	53	15	53	20	11
0.20	0.20	21	30	LN3	15	90	93	71	4	39	31	26
0.20	0.20	21	30	PE3	1	52	71	90	0	6	23	70
0.30	0.30	21	30	GLO	86	54	18	0	85	14	1	0
0.30	0.30	21	30	GEV	74	95	69	8	36	47	15	0
0.30	0.30	21	30	LN3	13	85	95	36	4	35	50	11
0.30	0.30	21	30	PE3	0	5	41	93	0	0	14	85

Simulations are of homogeneous regions with specified values of L CV τ, L skewness τ_3, number of sites in region (N) and record length at each site (n). "World" is the true distribution used in the simulations. "Percent accepted" is the percentage of the simulations in which a candidate distribution gave an acceptable fit ($|Z| \leq 1.64$). "Percent chosen" is the percentage of the simulations in which a distribution was chosen as the best of the four candidates, in the sense of giving the smallest value of $|Z|$.

simulations can be the same ones used in the calculation of the heterogeneity measure described above.

We have so far taken the sample L moments t_3 and t_4 to be exactly unbiased. This is a very good approximation for t_3 but is not so good for t_4 when record lengths are short ($n_i \leq 20$) or the population L skewness is large ($\tau_3 \geq 0.4$). To overcome this problem, a bias correction for t_4 is used. Compare the fitted L kurtosis τ_4^{GEV} not with the regional average $\bar{t}_4$ itself but with the bias-corrected version $\bar{t}_4 - \beta_4$, where β_4 is the bias in the regional average L kurtosis for regions with the same number of sites and the same record lengths as the observed data. This bias can be obtained from the same simulations as those used to obtain σ_4.

Formal Definition

Assemble a set of candidate three-parameter distributions. Reasonable possibilities include the generalized logistic (GLO), generalized extreme-value (GEV), generalized Pareto, lognormal and Pearson type III.

Fit each distribution to the group average L moments 1, $\bar{t}$, and $\bar{t}_3$. Denote by τ_4^{DIST} the L kurtosis of the fitted distribution, where DIST can be any of GLO, GEV, etc.

Fit a kappa distribution to the group average L moments 1, $\bar{t}$, $\bar{t}_3$, and $\bar{t}_4$.

Simulate a large number N_{sim} of regions from this kappa world; the regions are homogeneous and have no cross correlation or serial correlation, and sites have the same record lengths as their real world counterparts. (The fitting of a kappa distribution and simulation of kappa worlds can use the same computations as for the heterogeneity measure described above.) For the mth simulated region calculate the regional average L skewness $\bar{t}_3^{(m)}$ and L kurtosis $\bar{t}_4^{(m)}$.

Calculate the bias of $\bar{t}_4$,

$$\beta_4 = N_{\text{sim}}^{-1} \sum_{m=1}^{N_{\text{sim}}} (\bar{t}_4^{(m)} - \bar{t}_4),$$

the standard deviation of $\bar{t}_4$,

$$\sigma_4 = \left\{ (N_{\text{sim}} - 1)^{-1} \left[\sum_{m=1}^{N_{\text{sim}}} (\bar{t}_4^{(m)} - \bar{t}_4)^2 - N_{\text{sim}} \beta_4^2 \right] \right\}^{1/2},$$

and, for each distribution, the goodness-of-fit measure

$$Z^{\text{DIST}} = (\tau_4^{\text{DIST}} - \bar{t}_4 + \beta_4)/\sigma_4.$$

Declare the fit to be adequate if Z^{DIST} is sufficiently close to zero, a reasonable criterion being $|Z^{\text{DIST}}| \leq 1.64$.

Performance

The performance of Z as a heterogeneity measure was also assessed by means of Monte Carlo simulation experiments. Data were generated from artificial homogeneous regions with one of four frequency distributions: generalized logistic, generalized extreme-value, three-parameter lognormal (LN3), or Pearson type III (PE3). One thousand replications were made of data from each region. Each of these four distributions was also fitted to each region's data, and counts were kept of the number of times that each distribution was accepted as giving an adequate fit to the data, that is, $|Z| \leq 1.64$, and of the number of times that each distribution was chosen as giving the best fit among the four fitted distributions in the sense of giving the smallest value of $|Z|$. Z statistics were calculated with $N_{\text{sim}} = 500$.

Simulation results are given in Table 3. From the construction of Z the true distribution of the region should be accepted about 90% of the time. This is approximately true for all parent distributions except GLO, which is accepted less often. It is not clear why this should be so; it may reflect a tendency for σ_4 to underestimate the true variance of the regional average L kurtosis for GLO regions. The amount by which these numbers exceed the other entries in the "percent accepted" columns of Table 3 measures the ability of Z to distinguish between different distributions. This is achieved fairly well for the GLO distribution and, when τ_3 is relatively high, for the PE3, but in other cases the distributions are hard to distinguish. This reflects the similarity of

the quantiles of the GEV, LN3, and PE3 distributions when τ_3 is small. In particular, the LN3 and PE3 distributions both tend to the normal distribution when τ_3 tends to zero and are very similar when $\tau_3 = 0.05$; this explains the similarity of the corresponding rows of Table 3. The entries in the "percent chosen" columns show how well the Z statistic can be used to identify the correct distribution from among the four candidates. Again, this can be achieved particularly well for the GLO and for the PE3 with high τ_3.

Use

The procedure for a region that is acceptably homogeneous is as follows. Calculate Z for all candidate distributions. Flag as acceptable all distributions for which $|Z| \leq 1.64$. Calculate regional growth curves for all the "acceptable" distributions. If these growth curves are all approximately equal, for the scientific purposes of the application under consideration, then any of the "acceptable" distributions is adequate. To guard against the possibility that the region was misspecified, it is safest to choose from among the "acceptable" distributions the one that is most robust to such misspecification. If the growth curves are not approximately equal, there is a problem of scarcity of data: two models display differences that are statistically insignificant but operationally important. In this case, in which it has not been possible to confidently identify the best model, robustness becomes particularly important. Rather than choose a three-parameter distribution it may be better to use the five-parameter Wakeby distribution, which is particularly robust to misspecification of the underlying distribution function of a homogeneous region.

It may happen that none of the candidate distributions is accepted by the Z criterion. This sometimes occurs when the number of sites in the region or the at-site record lengths are large; in these circumstances σ_4 is small and Z can be large even if the regional average L skewness and L kurtosis are fairly close to those of one of the candidate distributions. If the regional average $(\bar{t}_3, \bar{t}_4)$ point falls between two distributions (or among three or more distributions) whose growth curves are approximately equal, for the scientific purposes of the application under consideration, then there is a problem of superabundance of data: two models display differences that are statistically significant but operationally unimportant. In this case it is reasonable to reclassify any of the operationally equivalent distributions as giving an "acceptable" fit to the data. Sometimes the regional average point does not lie between two operationally equivalent distributions; for example, it may lie above the generalized-logistic line. In these cases, no three-parameter distribution is acceptable, and a more general distribution such as the Wakeby or kappa should be used.

If the region is not acceptably homogeneous, there is no reason to suppose that a single distribution will give a good fit to every site's data. Nonetheless, fitting a single distribution can still yield much more accurate quantile estimates than fitting separate distributions to each site. The choice of distribution should be influenced by considerations of robustness; it is particularly important to use a distribution that is of robust to moderate heterogeneity in the at-site frequency distributions. The Wakeby distribution is a widely recommendable choice.

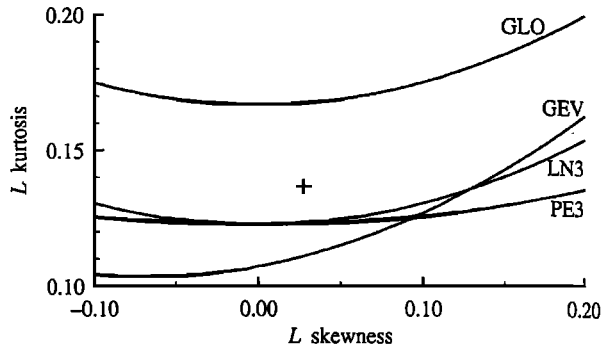

Fig. 7. Average L moments of the North Cascades data (plus), with L skewness versus L kurtosis relationships for some common distributions.

Notes

The criterion $|Z| \leq 1.64$ is somewhat arbitrary. The Z statistic has the form of a significance test of goodness of fit and has approximately a standard normal distribution if the at-site L kurtosis statistics have independent identical normal distributions. The criterion $|Z| \leq 1.64$ then corresponds to acceptance of the hypothesized distribution at a confidence level of 90%. However, the assumptions necessary for Z to be standard normal are unlikely to be exactly satisfied. Thus the criterion is a rough indicator of goodness of fit and is not recommended as a formal test.

The criterion $|Z| \leq 1.64$ is particularly unreliable if serial correlation or cross correlation is present in the data. Correlation tends to increase the variability of $\bar{t}_4$, and since there is no correlation in the simulated kappa world, the resulting estimate of σ_4 is too small and the Z values are too large. Thus a false indication of poor fit may be given. To overcome this problem it is possible to generate simulated data that are correlated, though this would require much more computing time.

The definition of Z involves only the regional average L moments, not the individual at-site L moments. As noted above, the regional average gives a sufficient summary of the data when the region is homogeneous. When the region is heterogeneous, it is possible that a test that makes use of the at-site L moments might enable better discrimination between distributions. However, for heterogeneous regions we consider it more important that the chosen distribution be robust to heterogeneity than that it achieve the ultimate quality of fit. We therefore tend to prefer the Wakeby distribution for heterogeneous regions.

Example

For the North Cascades data the regional average L skewness and L kurtosis are $\bar{t}_3 = 0.0279$ and $\bar{t}_4 = 0.1366$. The position of the regional average relative to the τ_3-τ_4 relationships of four candidate three-parameter distributions is shown by the plus symbol in Figure 7. The Z values for four candidate distributions are generalized logistic, 3.59; generalized extreme value, −2.98; three-parameter lognormal, −1.51; and Pearson type III, −1.60. The lognormal and Pearson distributions give acceptably close fits to the regional average L moments. The growth curves for these two distributions are almost identical throughout the range of quantiles from 0.01 to 0.999, so either distribution would be

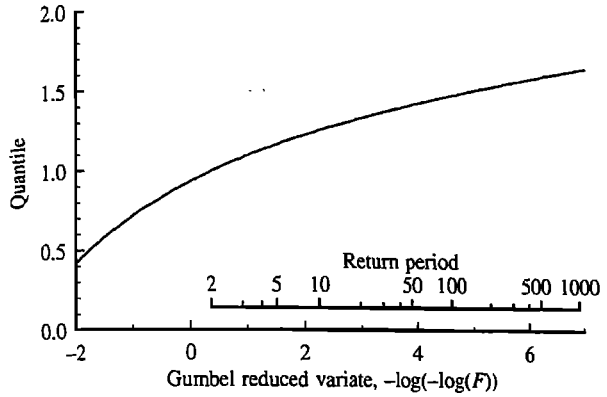

Fig. 8. Regional quantile function fitted to the North Cascades data.

an appropriate choice for this region. The growth curves are plotted, as though on extreme-value plotting paper, in Figure 8; the two curves are indistinguishable.

Analysis of the North Cascades data, that is, calculation of the D_i, H, and Z statistics, used 13 s of CPU time on an IBM 3090S computer.

SOFTWARE

A Fortran subroutine to calculate the statistics described in this paper is documented by *Hosking* [1991] and can be obtained by e-mail from the software repository StatLib. Send the request "send lmoments from general" to statlib@lib.stat.cmu.edu.

Acknowledgment. We are grateful to N. B. Guttman (National Climatic Data Center) for supplying data from the Historical Climatology Network.

REFERENCES

Acreman, M. C., and C. D. Sinclair, Classification of drainage basins according to their physical characteristics: An application for flood frequency analysis in Scotland, *J. Hydrol.*, *84*, 365–380, 1986.

Chowdhury, J. U., J. R. Stedinger, and L.-H. Lu, Goodness-of-fit tests for regional generalized extreme value flood distributions, *Water Resour. Res.*, *27*, 1765–1776, 1991.

Dalrymple, T., Flood frequency analyses, *U.S. Geol. Surv. Water Supply Pap.*, *1543-A*, 1960.

Greenwood, J. A., J. M. Landwehr, N. C. Matalas, and J. R. Wallis, Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form, *Water Resour. Res.*, *15*, 1049–1054, 1979.

Hosking, J. R. M., The theory of probability weighted moments, *Res. Rep. RC12210*, IBM Res., Yorktown Heights, N. Y., 1986.

Hosking, J. R. M., The 4-parameter kappa distribution, *Res. Rep. RC13412*, IBM Res., Yorktown Heights, N. Y., 1988.

Hosking, J. R. M., *L*-moments: Analysis and estimation of distributions using linear combinations of order statistics, *J. R. Stat. Soc., Ser. B*, *52*, 105–124, 1990.

Hosking, J. R. M., Fortran routines for use with the method of *L*-moments, version 2, *Res. Rep. RC17097*, IBM Res., Yorktown Heights, N. Y., 1991.

Hosking, J. R. M., and J. R. Wallis, The effect of intersite dependence on regional flood frequency analysis, *Water Resour. Res.*, *24*, 588–600, 1988.

Hosking, J. R. M., J. R. Wallis, and E. F. Wood, An appraisal of the regional flood frequency procedure in the UK flood studies report, *Hydrol. Sci. J.*, *30*, 85–109, 1985.

Houghton, J. C., Birth of a parent: The Wakeby distribution for modeling flood flows, *Water Resour. Res.*, *14*, 1105–1109, 1978.

Karl, T. R., C. N. Williams, F. T. Quinlan, and T. A. Boden, United States historical climatology network (HCN) serial temperature and precipitation data, *ORNL/CDIAC-30 NDP-019/R1*, Carbon Dioxide Inform. Anal. Cent., Oak Ridge Natl. Lab., Oak Ridge, Tenn., 1990.

Lettenmaier, D. P., and K. W. Potter, Testing flood frequency estimation methods using a regional flood generation model, *Water Resour. Res.*, *21*, 1903–1914, 1985.

Lettenmaier, D. P., J. R. Wallis, and E. F. Wood, Effect of regional heterogeneity on flood frequency estimation, *Water Resour. Res.*, *23*, 313–323, 1987.

Lu, L.-H., Statistical methods for regional flood frequency investigations, Ph.D. thesis, Cornell Univ., Ithaca, N. Y., 1991.

Natural Environment Research Council, *Flood Studies Report*, vol. 1, London, 1975.

Pearson, C. P., Regional flood frequency analysis for New Zealand data using *L*-moments, *Rep. WS 1417*, DSIR Hydrol. Cent., Christchurch, N. Z., 1991.

Pilon, P. J., and K. Adamowski, The value of regional information to flood frequency analysis using the method of *L*-moments, *Can. J. Civ. Eng.*, *19*, 137–147, 1992.

Pilon, P. J., K. Adamowski, and Y. Alila, Regional analysis of annual maxima precipitation using *L*-moments, *Atmos. Res. J.*, in press, 1992.

Plantico, M. S., T. R. Karl, G. Kukla, and J. Gavin, Is recent climate change across the United States related to rising levels of anthropogenic greenhouse gases?, *J. Geophys. Res.*, *95*, 16,617–16,637, 1990.

Potter, K. W., and D. P. Lettenmaier, A comparison of regional flood frequency estimation methods using a resampling method, *Water Resour. Res.*, *26*, 415–424, 1990.

Smith, J. A., Regional flood frequency analysis using extreme order statistics of the annual peak record, *Water Resour. Res.*, *25*, 311–317, 1989.

Thompson, W. R., On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation, *Biometrika*, *32*, 214–219, 1935.

Wallis, J. R., Risk and uncertainties in the evaluation of flood events for the design of hydraulic structures, in *Piene e Siccità*, edited by E. Guggino, G. Rossi, and E. Todini, pp. 3–36, Fondazione Politecnica del Mediterraneo, Catania, Italy, 1980.

Wallis, J. R., Hydrologic problems associated with oilshale development, in *Environmental Systems and Management*, edited by S. Rinaldi, pp. 85–102, North-Holland, Amsterdam, 1982.

Wallis, J. R., Regional frequency studies using *L*-moments, *Res. Rep. RC14597*, IBM Res., Yorktown Heights, N. Y., 1989.

Wallis, J. R., and E. F. Wood, Relative accuracy of log Pearson III procedures, *J. Hydraul. Eng.*, *111*, 1043–1056, 1985.

Wilks, S. S., Multivariate statistical outliers, *Sankhyā*, *25*, 407–426, 1963.

J. R. M. Hosking and J. R. Wallis, IBM Research Division, P. O. Box 218, Yorktown Heights, NY 10598.

(Received February 13, 1992;
revised August 5, 1992;
accepted August 17, 1992.)