Jerson Kelman [1]

STATISTICAL APPROACH TO FLOODS

ABSTRACT

The usual approach to the calculation of $x(T)$, the annual maximum daily streamflow associated with recurrence interval T, is to fit a probability distribution to a set of observations of annual maxima. The choice of the probability distribution is often based on asymptotic results. We investigate this model selection criterion through evaluation of the errors in estimating of $x(T)$ for a Markovian daily flow stochastic process.

The design of spillways or flood control storage requires the complete calculation of the T flood hydrograph, rather than just the peak value. Questions regarding the evolution of reservoir storage could be solved if a large number of daily streamflow sequences were available to be used in the evaluation of the frequency of failure of each tentative design. The utility of stochastic daily streamflow models is discussed, particularly the question of how to reduce the computer time necessary to generate a large number of synthetic daily sequences.

1. INTRODUCTION

It is usual to involve hydrologists in the design of hydraulic structures which are subjected to streamflows up to the critical event called the "design flood". When the failure of the structure can have catastrophic consequences, the design flood is often calculated through a hydrometeorological approach, which provides an upper bound to observed storms with the purpose of defining an event that "with all likelihood" will never happen. Descriptions of this methodology for applications in temperate regions are found in the literature (for example, WMO, 1973), but for tropical regions

[1] Electrical Energy Research Center—(CEPEL), Caixa Postal 2754–CEP 20001–Rio de Janeiro, Brazil

I. B. MacNeill and G. J. Umphrey (eds.), Stochastic Hydrology, 193–225.

there are only a limited number of references (Myers, 1981).

The design flood can also be calculated through the flood frequency analysis, which is the subject of this paper. Flood frequency analysis is a set of procedures that make use of statistics for assigning the exceedance probability to each flood event.

In some engineering problems one needs only to define the peak flow $x(T)$, as for example when designing a levee. Most of the work done in statistics deals with this kind of problem; namely, how to calculate the flow that will be exceeded in any year with probability p. For major hydraulic structures, T is sometimes chosen to be as large as 10,000 years. The usual approach to calculation of $x(T)$ is to fit a probability distribution $\hat{F}(\cdot)$ to a set of observations of m annual maxima $\{x_1, x_2, \ldots, x_m\}$ and obtain an estimate, $\hat{x}(T)$.

Several questions may be raised in connection with this approach:

(a) What is the population probability distribution from which $\{x_1, x_2, \ldots, x_m\}$ was sampled?

(b) What is the probability distribution associated with the smallest mean square error (or mean absolute error) for the estimator $\hat{X}(T)$?

(c) How large is this error?

(d) What is the probability of under-designing, such that $P\{\hat{X}(T) < x(T)\}$?

The answer to questions (a) and (b) may be different because the errors in the parameters of the population distribution may be high. There are several results available in the literature aimed at answering questions (c) and (d) when the population distribution is known; that is, when the estimation procedure is the only source of error (for example, Kottegoda, 1980). However, results are not easily obtained when the population distribution is unknown.

The first asymptotic distribution of extreme value theory is often used as an approximation for the unknown population distribution. One of the main results of this theory states that if the random variables Y_i are independent with a common distribution of exponential type, then the maximum defined as $X = \max\{Y_1, Y_2, \ldots, Y_n\}$, will have the following large sample probability distribution (Gumbel, 1958):

$$\lim_{n\to\infty} F(x) = \exp\left[-\exp\left(-\Psi(x-\mu)\right)\right]. \qquad (1)$$

This asymptotic distribution, sometimes referred to as the Gumbel distribution, is valid even when the random variables Y_i are weakly dependent, which is the case when the correlation between Y_i and Y_{i+k} goes to zero with increasing k (Cramer and Leadbetter, 1967). However, there are probability distributions for Y with either asymptotic distribution for X, or with

distributions associated with the second (also called Frèchet) or third (also called Weibull) asymptotic distribution.

Since most probability distributions used in hydrology are of exponential type, such as the normal, the log-normal and the gamma, it is understandable why the Gumbel distribution seems to be a suitable approximation to the unknown population distribution of X. The term "approximation" is introduced because equation (1) is used for finite n (up to 365) and also because the daily flows Y_i are not identically distributed. The adequacy of this approximation will be discussed in Section 2.

Another frequently used approach to the selection of an approximate probability distribution for X, not necessarily confined to the set of asymptotic distributions, is to examine a number of candidate distributions and pick the one that most closely fits the data. Obviously, the goodness-of-fit measure has to take into account the number of parameters of each distribution.

Comparative studies have been made with data from a great number of streamflow gauges with a view to obtaining a standardized distribution of the annual maximum. In the United States the Water Resources Council (USWRC, 1967) suggested the use of the log-Pearson III distribution and later furnished further guidelines regarding the estimation procedure (USWRC, 1977). This recommendation created a great deal of controversy. It has been noted by Wallis (1981) that the 500-year-flood divided by the size of the drainage area may vary over five orders of magnitude for streamflow gauges located in a small hydrologically homogeneous region.

In England (N.E.R.C., 1975) six different goodness-of-fit measures led to inconclusive results. The final recommendation of the British study was to use a specific probability distribution for each region of Great Britain. These distributions, the so-called "Regional Growth Curves", also have been subjected to well-founded criticism (Hosking *et al.*, 1985).

One may question if goodness-of-fit is a reasonable criterion for selecting an approximation to the annual maxima probability distribution. In fact, a good fit is valid only in the range of the annual maximum for which there are observations available, usually associated with small recurrence intervals. However, what matters is the unknown fit for large T values. Houghton (1977) and Moreira *et al.* (1983) have shown that the best "interpolating distribution" (the best fit) is not necessarily the best "extrapolating distribution" (the best estimator of $x(T)$, T large). In Section 3, it is shown how the minimization of the mean absolute error of $\hat{X}(T)$ may be used as an alternative criterion for selecting the approximate probability distribution of the annual maximum.

The only question raised thus far is that of how to estimate the peak flow, $X(T)$. However there are other engineering problems which require

the inflow volume for different durations; for example, the sizing of the flood storage in a man-made reservoir. In this regard, one is required to calculate a flood storage with a failure recurrence interval of T. The design of a spillway presents a similar problem. In this case, it is possible to attenuate the flood in the so-called "safety storage", which is situated above the flood control storage. Whenever there is some water in the safety storage, the operational rule is to empty it as quickly as possible. Therefore, the only limitation on the outflow rate is set by the hydraulic conditions of the spillway. These will not be constant. Furthermore, as this is an operation required for dam protection, no constraints regarding downstream flooding are taken into account while the safety storage is being voided. The problem is to calculate jointly the spillway capacity and the safety storage for an overtopping of the dam event with the recurrence interval T. If the dam is earthfilled, overtopping will likely mean a dam break with catastrophic downstream effect, and T is therefore assumed very large, say 1000 or 10,000 years. Obviously, the larger the spillway capacity the smaller will be the safety storage, andvice versa.

Questions regarding the evolution of reservoir storage could be solved easily if a large number of daily flow sequences were available to be used in the evaluation of the frequency of failure for each tentative design. Obviously, these frequencies would only be reasonably close to the respective probabilities of failure if the number of simulations were at least one order of magnitude larger than the recurrence interval being considered. For flood control calculations this means that the number of daily sequences should be of the order of 500 and for spillway design of the order of 100,000. But the stream records are seldom longer than $m = 100$ years. This paradox can be circumvented if a daily stochastic streamflow model is used to produce as many synthetic sequences as necessary.

Several features of flood volume modelling and daily streamflow modelling are discussed in Section 4, in particular the question of how to reduce the computer time necessary to generate a large number of synthetic daily sequences.

2. THE FIRST ASYMPTOTIC EXTREME VALUE PROBABILITY DISTRIBUTION (GUMBEL)

Let us assume that non-stationarity in the daily flow process can be neglected during a flood season that lasts for n days. In this case, it is easy to obtain some insight into how the Gumbel distribution approximates the true distribution of the annual maximum daily streamflow. Initially, let us accept the unrealistic assumption that the daily streamflows $Y_1, Y_2, \ldots, Y_n$

are independent random variables. In this case, the probability distribution of $X = \max\{Y_i\}$ is simply

$$F_x(x;n) = P(X \leq x) = P\left(\cap_{i=1}^{n} Y_i < x\right) = [F_y(x)]^n . \qquad (2)$$

Figure 1 shows the graphs of $F_x(x;n)$ for different n values for the case where the Y_i are normally distributed with $E(Y_i) =$ var $(Y_i) = 1$, $\forall i$. The horizontal axis is such that a plot of the Gumbel distribution would form a straight line; that is, the variable g is such that

$$g = -\ln(-\ln F_x(x;n)). \qquad (3)$$

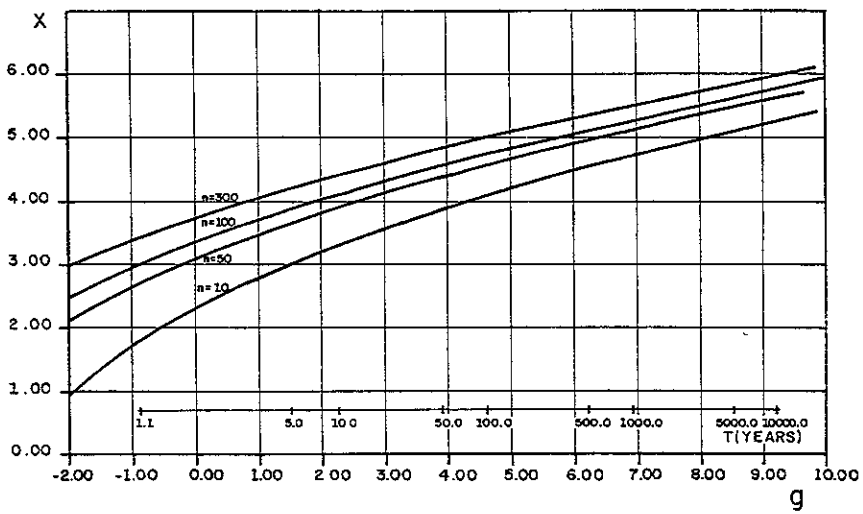

Figure 1. *Probability distribution of* $X = \max\{Y_i,\ i = 1, \ldots, n\}$. $E(Y_i) =$ var $(Y_i) = 1$, (Y_i, Y_j) *independent,* Y_i *normally distributed.*

The main facts to be observed from Figure 1 are:

a) The curves cannot be approximated by straight lines, meaning that the use of the Gumbel distribution would result in error. Of course, this has been known at least since Gumbel's comment (1958, pp. 219) about a graph similar to Figure 1 (See Figure 6.2.1 (3) in the above reference, which incidentally has a minor mistake): "For the normal distribution, however, the approach is very slow. The curves for $n = 100$, 200, 500 and 1000 taken from Tippet (1925) depart sensibly from a straight line, if we go outside the interval 0.05 to 0.95".

b) As typical steamflow records are generally no longer than 30 years, straight lines fitted to the empirical probability distributions of X, in the range $T = 1$ to $T = 30$, will tend to overestimate $x(T)$, for large T values.

Figure 2 shows the graphs of $F_x(x; n)$ for different n values for the case where the Y_i are log-normally distributed with $E(Y_i) =$ var $(Y_i) = 1$, $\forall i$. Again the curves cannot be approximated by straight lines, but, in contrast to the case of Figure 1, the use of the Gumbel distribution will tend to underestimate $x(T)$ for large T values. Furthermore, it should be noted that the vertical scales used in Figures 1 and 2 are different, meaning that the marginal distribution of daily flow Y_i is relevant when estimating $x(T)$ (Grigoriu, 1979).

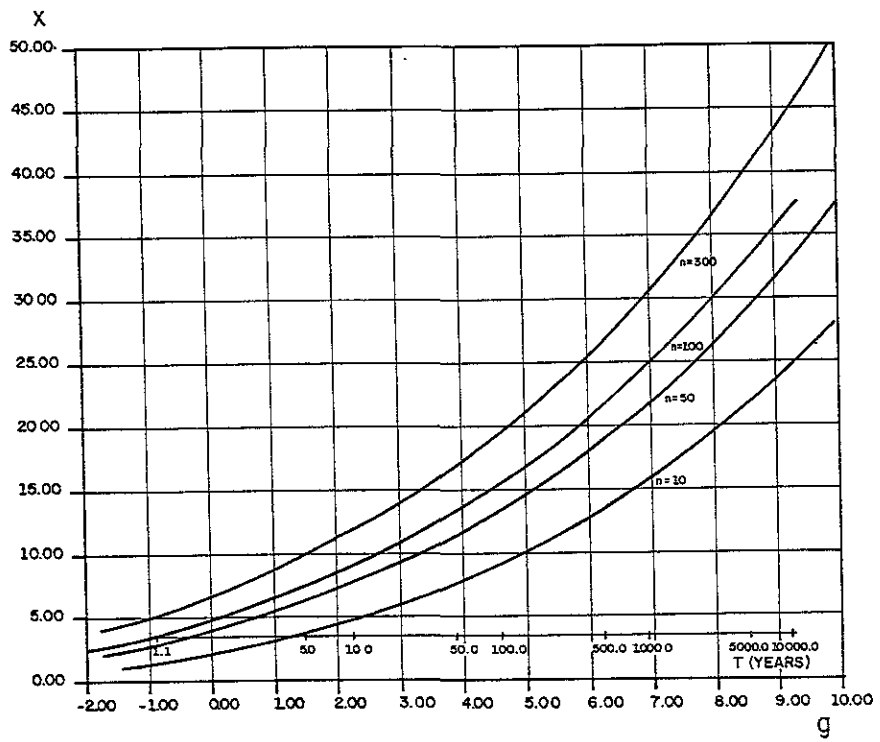

Figure 2. *Probability distribution of* $X = \max\{Y_i,\ i = 1, \ldots, n\}$. $E(Y_i) = \text{var}(Y_i) = 1$, (Y_i, Y_j) *independent,* Y_i *normally distributed.*

The differences between Figures 1 and 2 are due to the tail behaviour of the two distributions. Although the normal and the log-normal distributions are of the exponential type (Gumbel, 1958, pages 119, 120, 136, 146), $F_x(x, n)$ will converge to the Gumbel distribution with increasing n in different ways. In fact the normal distribution is "light-tailed", in the sense that its density function goes to zero, for increasing y, more rapidly than an exponential density function. The converse is true for the log-normal distribution, which is "heavy-tailed". More precisely, it can be said that if the conditional mean exceedance defined as $E(Y - y \mid Y > y)$ is a decreasing (increasing) function of y— at least for sufficiently large y— then the probability distribution of Y is light (heavy) tailed (Bryson, 1974).

Now assume that Y_i, the streamflow on day i, is such that

$$Y_i = \exp(W_i)$$

and

$$W_i = \alpha + \gamma(W_{i-1} - \alpha) + \beta(1 - \gamma^2)^{\frac{1}{2}} N_i, \tag{4}$$

where N_i is standard normal and

$$E(N_i N_k) = \begin{cases} 0, & k \neq i \\ 1, & k = i. \end{cases}$$

We cannot expect that this simple Markovian process will actually resemble daily streamflows, but it is useful in providing some insight into how the time persistence of the process affects the use of the Gumbel distribution as an approximation for extreme values.

Obviously the marginal distribution of Y_i is log-normal and the following properties can be derived easily:

$$E(Y_i) = \exp(\alpha + \beta^2/2), \tag{5a}$$

$$\text{var}(Y_i) = \exp(2\alpha + \beta^2)\exp(\beta^2 - 1), \tag{5b}$$

$$\text{skew}(Y_i) = (\beta/\alpha)^3 + 3(\beta/\alpha), \tag{5c}$$

$$\text{corr}(Y_i, Y_{i+k}) = \exp(\beta^2\gamma^k) - 1, \tag{5d}$$

$$E(Y_i \mid y_{i-1}) = \exp[\beta^2(1 - \gamma^2)/2 + \alpha(1 - \gamma)]\, y_{i-1}^{\gamma}, \tag{5e}$$

and

$$\text{var}(Y_i \mid y_{i-1}) = \left[\exp\Big(2\big(\beta^2(1 - \gamma^2) + \alpha(1 - \gamma)\big)\Big) - \exp\big(\beta^2(1 - \gamma^2) + 2\alpha(1 - \gamma)\big)\right] y_i^{2\gamma}. \tag{5f}$$

If one assumes that $\alpha = \ln 2^{-\frac{1}{2}} = -0.35, \beta = (\ln 2)^{\frac{1}{2}} = 0.83$ and $\gamma = [\ln 2]^{-1} \ln(1 + 0.95) = 0.96$, it is possible to show, by back substitution

in the above equations, that $E(Y_i) = \text{var}(Y_i) = 1$, $\text{skew}(Y_i) = 4$ and $\text{corr}(Y_i, Y_{i+1}) = 0.95$. These values are typical for daily streamflow time series of large rivers. The regression of Y_i given y_{i-1} is practically coincident with the straight line $0.95\, y_{i-1} + 0.05$ for values of y_{i-1} larger than 0.5 and the autocorrelation is practically coincident with 0.95^k, for values of k smaller than 10.

The stochastic process defined by (4) is heteroscedastic, which is a feature in agreement with the hydrological experience that the larger the streamflow is today, the less precise will be the flow forecast for tomorrow.

The probability distribution of $X = \max_i \{Y_i\}$ is

$$F_X(x;n) = P(Y_1 \le x,\ Y_2 \le x, \ldots, Y_n \le x) = \int_{-\infty}^{\frac{\ln x-\alpha}{\beta}} \cdots \int_{-\infty}^{\frac{\ln x-\alpha}{\beta}} \phi_n(\mathrm{u})\mathrm{du}, \tag{6}$$

where ϕ_n is the n-variate density function of the standard normal. This n-fold integral is difficult to evaluate for large values of n.

A first approximation to $F_X(x;n)$ is the following (Rosbjerg, 1979):

$$F_X(x;n) \simeq F_1(x;n) = P(Y_1 \le x) \prod_{i=2}^{n} P(Y_i < x \mid Y_{i-1} < x) = \left[\Phi_1\left(\frac{\ln x-\alpha}{\beta}\right)\right]^{2-n} \left[\Phi_2\left(\frac{\ln x-\alpha}{\beta}, \frac{\ln x-\alpha}{\beta}; \gamma\right)\right]^{n-1}, \quad n \ge 2.$$

In short, this approximation is

$$F_1(x;n) = \Phi_1^{2-n}\Phi_2^{n-1}, \tag{7}$$

where Φ_1 and Φ_2 are the standard normal probability distributions respectively for the univariate and bivariate (with correlation coefficient γ) cases.

A second possible approximation to $F_X(x;n)$ may be obtained by assuming that the upcrossings of the $\{Y_i\}$ process with regard to the threshold level x, for large x, is a Poisson process. As such, the waiting time, (K), between upcrossings is exponentially distributed with the following mean rate (Grigoriu, 1979):

$$\mu(x) = P(Y_{i+1} > x,\ Y_i \le x) = \Phi_1\left(\frac{\ln x-\alpha}{\beta}\right) - \Phi_2\left(\frac{\ln x-\alpha}{\beta}, \frac{\ln x-\alpha}{\beta}; \gamma\right) \tag{8}$$

or simply

$$\mu(x) = \Phi_1 - \Phi_2.$$

Hence

$$F_K(k) \approx 1 - \exp\big((\Phi_2 - \Phi_1)\, k\big). \tag{9}$$

However,

$$P(X \le x) = F_X(x;n) = P(K > n) \approx 1 - F_K(n) = \exp\big((\Phi_2 - \Phi_1)\, n\big).$$

Therefore, in our short notation the second approximation to $F_X(x;n)$ may be written:

$$F_2(x;n) = \exp\big((\Phi_2 - \Phi_1)\, n\big) \tag{10}$$

A third approximation can be obtained through the Monte Carlo approach by using (4) to generate s sequences $\{Y_1, Y_2, \ldots, Y_n\}_j, j = 1, \ldots, s$. Since each sequence is associated with one extreme value observation, a sample $(x_1, x_2, \ldots, x_s)$ can be produced. Therefore, it is possible to estimate $F_X(x;n)$ by $F_3(x;n)$, the empirical probability distribution of X. In fact, $F_3(x;n)$ converges to $F_X(x;n)$ with growing s.

Figure 3 shows the graphs of the approximations for $n = 100$ days, which is a typical duration for the flood season. The graph of the second approximation was not plotted because it falls very close to $F_1(x;n)$. The third approximation, which is practically coincident with $F_X(x;n)$ for $T < 1000$, was obtained for $s = 10^5$ "flood seasons". The descriptors of the X variable are, according to the third approximation:

$$E(X) = 3.13,\ \text{std. dev.}\ (X) = 2.23,\ \text{skew}(X) = 2.74,\ \text{kurt}(X) = 18.72.$$

These values are different from the descriptors of the Gumbel distribution (skewness of 1.14 and kurtosis of 5.4). Also displayed for comparison is the curve for the independent process, which is calculated exactly using equation (2).

It can be noted in Figure 3 that the time persistence of daily streamflows does not play a role as significant as that of the marginal distribution (see also Figure 1), although the time persistence cannot be dismissed in this particular case. It should be noted that other Markovian processes with moderate auto-correlation coefficients may eventually be treated as independent, as far as extremes are concerned (Grigoriu, 1979).

The second comment on Figure 3 is that the Markovian approximation may lead to significant errors in the estimation of $x(T)$. For example, the error in the approximation of $x(1000)$ in this particular case was of the order of 12%. This is not large when one considers all other sources of uncertainty usually found in the study of floods. But since we are talking

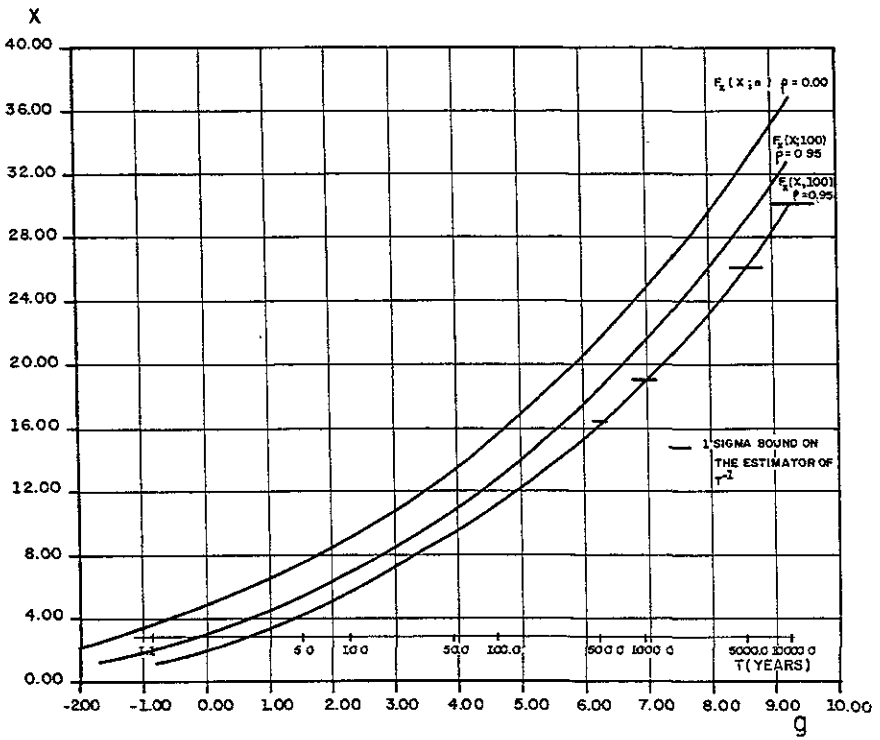

Figure 3. *Approximations to the probability distribution of* $X = \max\{Y_i, i = 1, \ldots, j\}$, $E(Y_i) = \text{var}(Y_i) = 1$, $\text{corr}\,(Y_i, Y_{i+1}) = 0.95$, Y_i *log-normally distributed.*

about an avoidable error, the recommendation on this subject is to adopt the empirical distribution, $F_3(x; n)$, rather than approximations $F_1(x; n)$ or $F_2(x; n)$.

Now we would like to know how valuable it is to fit the Gumbel distribution to a set of annual maxima streamflows, as far as the estimation of $x(T)$ is concerned. Furthermore, we would like to compare the accuracy of the resulting estimates with the accuracy associated with other fitting procedures for probability distributions, as well as with that associated with the "time series approach". Therefore, we will be considering three alternative approaches for estimating $x(T)$ and we want to discover which will lead, on average, to the smallest error. The three alternatives are:

a) Gumbel distribution (GUD). For a given set of annual maxima $(x_1, x_2, \ldots, x_m)$, the estimates $\hat{\psi}$ and $\hat{\mu}$ (equation 1) are found through the

iterative algorithm:

$$\psi_{j+1} = \psi_j - \frac{g(\psi_j)}{g'(\psi_j)}, \tag{11a}$$

$$\psi_0 = 1.28/s_x, \tag{11b}$$

$$g(\psi_j) = m\left[\frac{1}{\psi_j} - \bar{x} + \frac{\sum_i x_i \exp(-\psi_j x_i)}{\sum_i \exp(-\psi_j x_i)}\right], \tag{11c}$$

$$g'(\psi) = \frac{dg(\psi)}{d\psi}, \tag{11d}$$

and

$$\hat{\mu} = \frac{1}{\psi_j} \ln\left(\frac{m}{\sum_i \exp(-\psi_j x_i)}\right), \tag{11e}$$

where $\bar{x}$ and s_x are, respectively, the sample mean and the sample standard deviation.

b) The exponential distribution (EXD). There are several competitors to the first asymptotic distribution; for example, the gamma, the log-Pearson type III, the generalized extreme value, and others. The two-parameter exponential was selected here for reasons which will become clear in the next section. Its probability distribution is

$$F_X(x) = 1 - \exp\left[\frac{\delta - x}{\lambda}\right], \; x \geq 0, \; \lambda \geq 0. \tag{12}$$

It can be shown that $\text{skew}(X) = 2$ and $\text{kurt}(X) = 9$. The estimation procedure we adopt is:

$$\hat{\lambda} = \frac{m}{m-1}\left(\bar{x} - \min_i(x_i)\right)$$

and

$$\hat{\delta} = \min(x_i) - \frac{\lambda}{m}. \tag{13}$$

c) The time series approach (TSA). This uses the transformed daily streamflow record $\{\ln\ y_i, \; i = 1, \ldots, \; n\}_j$, $j = 1, \ldots, m$ to estimate α, β and γ. The estimates are used in equation (7) to get $F_1(x(T); n)$ and ultimately $x(T)$. In accordance with the observations related to Figure 3, it would be better to use $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ to get $F_3(x(T); n)$. However, this

Table 1. *Results of the Monte Carlo Experiment*

	Method	BIAS	STDV	RMSE
$x(100) = 11.46$	GUD	−3.11	1.75	3.57
	EXD	0.18	2.47	2.48
	TSA	1.43	3.25	3.55
$x(1000) = 18.99$	GUD	−7.59	2.49	7.99
	EXD	−1.91	3.75	4.21
	TSA	2.24	6.21	6.60
$x(10,000) = 30.32$	GUD	−15.88	3.27	16.21
	EXD	−7.80	5.03	9.28
	TSA	2.49	10.77	11.05

$$\text{BIAS} = \text{BIAS}(\hat{X}(T)) = E(\hat{X}(T) - x(T))$$
$$\text{STDV} = \text{STD.DEV.}(\hat{X}(T)) = \text{var}(\hat{X}(T))^{0.5} = (E(\hat{X}(T) - E(\hat{X}(T)))^2)^{0.5}$$
$$\text{RMSE} = (\text{MSE}(\hat{X}(T))^{0.5} = (E(\hat{X}(T) - x(T)^2)^{0.5}$$

has been ruled out from the Monte Carlo experiment, the description of which follows, because it would be computationally infeasible.

Let's assume that $x(T)$ must be estimated from a daily flow record of $m = 20$ years (a typical value) which was generated by the Markovian process with the parameters defined above.

Equation 4 was used to synthesize $s = 1000$ sets of $m = 20$ years of "streamflow data", each year with a "flood season" of $n = 100$ days. The three alternatives described above were applied to each set in order to estimate $x(T)$ for $T = 100$, 1000 and 10,000 years. That is, $F_x(x)$ is respectively 0.99, 0.999 and 0.9999. The results are displayed in Table 1.

The estimator $\hat{X}(T)$ associated with the GUD method has the smallest variance, but it has such a large bias that it would not be possible to recommend it in this particular case. For example, $E(\hat{X}(T))$ is roughly half the true value for $T = 1000$ or 10,000. Also, for these two T values, confidence intervals around an estimate $\hat{x}(T)$ will tend not to contain the true value $x(T)$, particularly if the confidence intervals are calculated by the usual procedure. That is, if X is Gumbel distributed and if the method of maximum likelihood is employed, then $\hat{X}(T)$ is asymptotically distributed as a normal variable with $E(\hat{X}(T)) = x(T)$ and $\text{var}(\hat{X}(T))$ given by (Henriques, 1981):

$$\text{var}_A(\hat{X}(T)) = \frac{\text{var}\,(X)}{m}(0.67 + 0.37\left(\ln\ (-\ln\,(1 - T^{-1}))\right)^2 - 0.33\ln\ (-\ln\,(1 - T^{-1}))). \quad (14)$$

For example, for $T = 1000$, $\text{var}(X) = (2.23)^2$, and $m = 20$, equation (14) yields $\text{var}_A(\hat{X}(T)) = (2.26)^2$, which is remarkably close to $\text{var}(\hat{X}(T)) = (2.49)^2$ of Table 1. If one assumes a particular estimate $\hat{x}(T)$ as being equal to $E(\hat{X}(T))$, and making the appropriate calculations, a 95% one sided confidence interval for the thousand-year flood would turn out to be equal to (11.40, 15.13), which is still far below the true value of 18.99. In conclusion, GUD would be an incorrect choice in this particular situation. This is a warning against the belief, widespread among hydrologists, that the asymptotic theory for extremes is a sound approach to flood modelling.

The estimator $\hat{X}(T)$ associated with the EXD method has the smallest mean squared error. It is the best choice, unless some loss function is used to penalize the negative bias more heavily than the positive bias. The rationale for this hypothetical loss function is that underdesign of a flood control structure has, in general, more serious consequences than an overdesign. If this is the case, the TSA would be the best choice for $T = 1000$ and 10,000, although its estimator $\hat{X}(T)$ is systematically the one with the largest variance.

3. PROBABILITY DISTRIBUTION FOR ANNUAL MAXIMUM

The exponential distribution (equation (12)) was chosen as one of the alternatives for estimating $x(T)$ in the last section because extensive Monte Carlo studies have shown that this distribution is robust for fitting annual streamflow maxima (Damazio *et al.*, 1983; Damazio, 1984; Damazio and Kelman, 1984). In other words, using the exponential distribution to fit samples of annual maxima results in relatively good estimates of $x(T)$ across a range of possible parent distributions of X.

The search for a robust distribution for annual maximum streamflow is not new. Slack *et al.* (1975) developed a Monte Carlo experiment in which

random samples of different sizes were produced by parent population distributions $F(x)$ and then these samples were fitted by distributions $G(x)$, not necessarily of the same form as $F(x)$. In each case an estimate $\hat{x}(T)$ was found and the distance to the true value $x(T)$ measured. Four distributions were considered: the normal, the Gumbel, the three-parameter log-normal and the three-parameter Weibull. The authors considered sample sizes ranging from 10 to 90, population skewness ranging from 0 to 15 and recurrence intervals ranging from 10 to 10,000 years. They found that when $F(x)$ was a three-parameter distribution, the best $G(x)$ was not frequently of the same form of $F(x)$. Furthermore, they found that the choice of the best $G(x)$ in each case was more sensitive to the skewness of the corresponding $F(x)$ than to its general form.

Landwehr *et al.* (1980) selected six $F(x)$ distributions from the Wakeby family and allowed $G(x)$ to be either Wakeby, Gumbel or log-normal. The Wakeby distribution is well suited for Monte Carlo studies because it can reproduce the different shapes of probability distributions usually employed in hydrology and also because it lends itself to the easy generating synthetic samples. A random variable X distributed as Wakeby is defined as

$$X = m + a\left[1 - (1-U)^b\right] - c\left[1 - (1-U)^d\right], \tag{15}$$

where U is a random variable uniformly distributed in the interval (0, 1) and (m, a, b, c, d) are parameters. The major conclusion of Landwehr *et al.* (1980) was that the Gumbel and log-normal distributions resulted in a rather precise under estimation of extreme quantiles when playing the role of $G(x)$. However, this was not the case when $G(x)$ was adopted as the Wakeby distribution with parameters estimated through the probability weighted moments method.

Damazio (1984) repeated the study of Landwehr *et al.* (1980), adding the two-parameter exponential distribution (12) to the list of the $G(x)$ distributions. He found that for T larger than 200 years the exponential distribution with the parameters estimated through the method of moments resulted in the smallest cumulative (among the populations) mean squared error. He concluded that the exponential distribution should be considered by hydrologists as an alternative for modelling maximum annual series.

The conclusions from these Monte Carlo experiments depend naturally on the selection of the population distribution $F(x)$. For this reason, Damazio *et al.* (1983) used regional Wakeby distributions of annual maximum, estimating parameters for Brazilian basins by a procedure suggested by Wallis (1981). Again the exponential distribution (12) turned out to be the most robust among a large set of competitors such as: normal, two-parameter log-normal, three-parameter log-normal, two-parameter gamma, three-parameter gamma, generalized extreme values, Gumbel and Wakeby. The method of moments was adopted in all cases, with the exception of the Wakeby distribution, which was fitted through the probability weighted moments method. The second most robust distribution was the Gumbel.

The search for a robust estimator of $x(T)$ may be extended to the case when some information is available on flood events which preceded the gauged record. In some basins there is physical evidence of flood events that occurred thousands of years ago, such as landscape "scars" and mud layer deposits. Palaeoflood hydrology is a branch of the geophysical sciences that seeks the estimation of the magnitude and the date of occurrence of these events. Since it was not obvious that the inclusion of this kind of information actually decreased the error of estimation for $x(T)$, the subject was investigated by Hosking and Wallis (1984). They came to the conclusion that palaeohydrology information is most useful when estimating a three-parameter flood frequency distribution for a single site possessing only a short gauged record. When several independent and homogeneous gauged records from different sites are used in a regional flood frequency analysis, the value of paleohydrological information is negligible.

In other basins there may be some historical information based on the memory of old people who remembered the highest river stage in their own life span and, with luck, also in their parents' life span. In these cases the most that the hydrologist can expect to know is the highest water level that occurred before systematic measurements started. This length of time, h, is in general smaller than 150 years, which is not a short interval when compared to m, the number of years of a streamflow record (m is generally smaller than 50). Cohn (1984) developed new techniques for incorporating this kind of historical information. He assumed $F(x)$ to be log-Pearson III and adopted the log-normal distribution (a special case of the log-Pearson III) as $G(x)$. He found that the historical information was of tremendous value for reducing the mean squared error of the estimator of $x(10)$ and $x(100)$.

Damazio and Kelman (1984) developed Monte Carlo studies to investigate the performance of the exponential and Gumbel distributions when historical data is available for moderate h (up to 150). They defined a set of twelve population distributions $F(x)$ of the Wakeby form, (called W-1, W-2, ..., W-12). All twelve have a single mode, a positive lower limit and no upper limit. Their skewness and kurtosis were selected in order to resemble typical values for Brazilian rivers.

Figure 4 shows the chosen pairs of skewness and kurtosis, as well as some empirical data. The lowest skewness in the experiment was close to the Gumbel value (1.14); three other skewness levels corresponding to 1.5, 2.0 and 2.5 were also investigated. For each skewness level, three kurtosis values

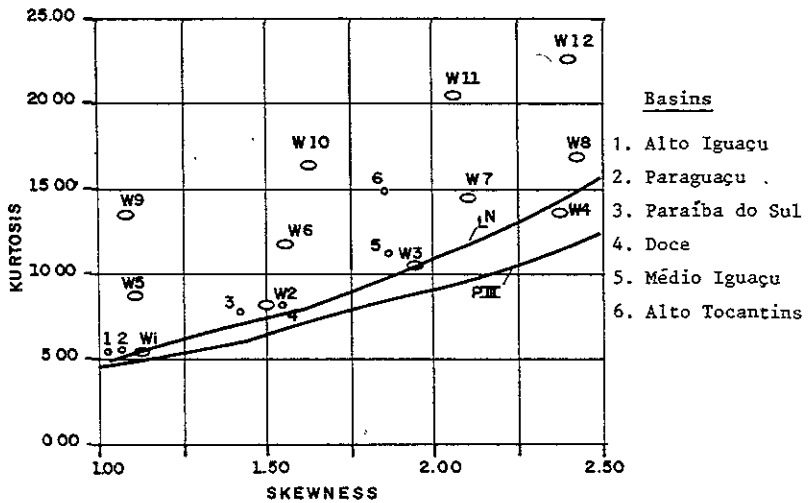

Figure 4. *Skewness and kurtosis of the Wakeby distributions.*

were considered, the lowest one in each case corresponding to the log-normal distribution. Table 2 shows the main characteristics of each distribution. It should be noted that all of them have unit expected value and coefficient of variation arbitrarily chosen as 0.49.

The Monte Carlo experiment was executed for $h = 50$, 100 and 150 years and $m = 5$, 10, 25 and 50 years. A large number of samples (k) were generated from the twelve Wakeby populations for each (h, m) pair. Each sample i $(i = 1, \ldots, k)$ was used to estimate $\hat{x}(T)$ by the eight alternative estimation procedures that are the combinations of the following three-way classification table:

$$A = \begin{cases} 1 & \text{– Gumbel Probability Distribution} \\ 2 & \text{– Exponential Probability Distribution} \end{cases}$$

$$B = \begin{cases} 1 & \text{– Method of Moments} \\ 2 & \text{– Method of Maximum Likelihood} \end{cases}$$

$$C = \begin{cases} 1 & \text{– Use Only Streamflow Record} \\ 2 & \text{– Use Streamflow Record + Historical Data} \end{cases}$$

The method of moments suggested by the USWRC (1977) was adopted for the case $(A = 1 \text{ or } 2,\ B = 1,\ C = 2)$. The method of maximum likelihood suggested by NERC (1975) for the case $(A = 1,\ B = 2,\ C = 2)$ and the method of maximum likelihood suggested by Damazio and Kelman (1984) for the case $(A = 2,\ B = 2,\ C = 2)$. Standard procedures were used in all cases with $C = 1$.

Table 2. *The Twelve Wakeby Distributions Used as Parent Distributions*

Wakeby	a	b	c	d	m	$E(X)$	Std.Var. (X)	Skew (X)	Kurt (X)	$x(T)$ $T = 1000$	$x(T)$ $T = 1000$
W-1	0.55	2.00	8.24	0.04	0.29	1.00	0.49	1.12	5.46	3.46	4.51
W-2	0.49	2.00	3.45	0.09	0.33	1.00	0.49	1.50	8.13	3.79	5.27
W-3	0.32	1.50	3.80	0.09	0.43	1.00	0.49	1.95	10.52	4.03	5.66
W-4	0.14	1.50	4.19	0.09	0.50	1.00	0.49	2.37	13.03	4.25	6.05
W-5	0.89	1.50	0.89	0.19	0.25	1.00	0.49	1.11	8.76	3.56	5.37
W-6	0.65	4.00	1.96	0.14	0.16	1.00	0.49	1.56	11.87	4.01	5.97
W-7	0.42	2.00	2.08	0.14	0.38	1.00	0.49	2.10	14.37	4.19	6.27
W-8	0.31	1.50	2.18	0.14	0.46	1.00	0.49	2.42	16.42	4.32	6.51
W-9	0.93	4.00	1.06	0.19	0.00	1.00	0.49	1.07	13.50	3.81	5.97
W-10	0.73	2.50	1.13	0.19	0.22	1.00	0.49	1.63	16.32	4.02	6.32
W-11	0.60	2.00	1.20	0.19	0.32	1.00	0.49	2.05	20.36	4.18	6.63
W-12	0.53	1.15	1.22	0.19	0.43	1.00	0.49	2.39	22.58	4.27	6.76

The relative mean absolute error was calculated for each population $F(X)$ and each estimation procedure; that is,

$$\text{MAE}\,(T) = \frac{1}{K}\sum_{i=1}^{K}\left|\frac{\hat{x}_i(T) - x(T)}{x(T)}\right|. \tag{16}$$

Figure 5 shows the variation of MAE (10,000) for the population W-1, which is "close" to the Gumbel and for the population W-3, which is "close" to the exponential. It is interesting to observe that when the "wrong" distribution is used to estimate $x(T)$, as when the Gumbel is used when the population is W-3 or when the exponential is used when the population is W-1, then an increase in the record length m actually increases the error!

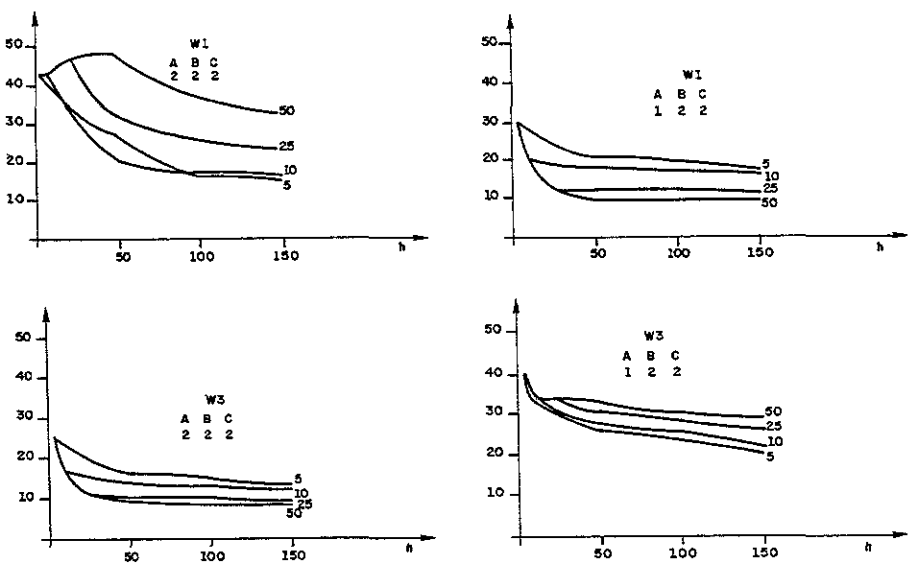

Figure 5. MAE *for W-1 and W-3*

Also, it can be noted from Figure 5 that an increase in the length of time h has a very small effect on the error.

Table 3 shows the estimation procedure with the smallest MAE (10,000) for each of the pairs (h, m) and for each of the twelve Wakeby populations. Inside the parentheses the corresponding MAE is shown. It should be noted that the exponential distribution was the winner in all cases except for the W-1 population.

Table 3. *Smallest MAE (10000)—Mean Absolute Error, $T = 10000$ and the Best Estimation Procedure*

A { 1. Gumbel Distribution; 2. Exponential Distribution
B { 1. Method of Moments; 2. Method of Maximum Likelihood
C { 1. Use Only Streamflow Record; 2. Use Streamflow Record + Historical Data

		W-1 A B C	W-2 A B C	W-3 A B C	W-4 A B C	W-5 A B C	W-6 A B C	W-7 A B C	W-8 A B C	W-9 A B C	W-10 A B C	W-11 A B C	W-12 A B C
150	50	1 2 2 (0.08)	2 1 2 (0.10)	2 2 2 (0.09)	2 2 1 (0.17)	2 1 2 (0.10)	2 2 2 (0.13)	2 2 1 (0.09)	2 2 1 (0.18)	2 1 2 (0.18)	2 2 2 (0.08)	2 2 1 (0.08)	2 2 1 (0.17)
	25	1 2 2 (0.09)	2 2 2 (0.11)	2 2 2 (0.11)	2 2 2 (0.17)	2 2 2 (0.11)	2 2 2 (0.11)	2 2 1 (0.13)	2 2 1 (0.20)	2 2 2 (0.12)	2 2 2 (0.12)	2 2 1 (0.10)	2 2 1 (0.18)
	10	1 2 2 (0.12)	2 2 2 (0.14)	2 2 2 (0.16)	2 2 2 (0.19)	2 2 2 (0.15)	2 2 2 (0.17)	2 2 1 (0.20)	2 2 2 (0.23)	2 2 2 (0.17)	2 2 1 (0.20)	2 2 1 (0.18)	2 2 1 (0.25)
	5	1 2 2 (0.13)	2 2 2 (0.16)	2 2 2 (0.17)	2 2 2 (0.19)	2 2 2 (0.18)	2 2 2 (0.21)	2 2 2 (0.24)	2 2 2 (0.24)	2 2 2 (0.22)	2 2 2 (0.25)	2 2 2 (0.25)	2 2 2 (0.28)
100	50	1 2 2 (0.08)	2 1 2 (0.11)	2 2 2 (0.09)	2 2 1 (0.17)	2 2 2 (0.10)	2 2 2 (0.16)	2 2 1 (0.09)	2 2 1 (0.17)	2 1 2 (0.18)	2 2 2 (0.09)	2 2 1 (0.08)	2 2 1 (0.17)
	25	1 2 2 (0.10)	2 2 2 (0.11)	2 2 2 (0.12)	2 2 2 (0.18)	2 2 2 (0.11)	2 2 2 (0.11)	2 2 1 (0.13)	2 2 1 (0.19)	2 2 2 (0.14)	2 2 2 (0.11)	2 2 1 (0.11)	2 2 1 (0.18)
	10	1 2 2 (0.13)	2 2 2 (0.14)	2 2 2 (0.17)	2 2 2 (0.21)	2 2 2 (0.15)	2 2 2 (0.17)	2 2 1 (0.20)	2 2 2 (0.24)	2 2 2 (0.18)	2 2 1 (0.18)	2 2 1 (0.19)	2 2 1 (0.26)
	5	1 2 2 (0.15)	2 2 2 (0.17)	2 2 2 (0.20)	2 2 2 (0.22)	2 2 2 (0.20)	2 2 2 (0.22)	2 2 2 (0.24)	2 2 2 (0.27)	2 2 2 (0.24)	2 2 2 (0.26)	2 2 2 (0.28)	2 2 2 (0.28)
50	50	1 2 1 (0.09)	2 1 1 (0.12)	2 2 1 (0.09)	2 2 1 (0.17)	2 1 1 (0.12)	2 1 1 (0.19)	2 2 1 (0.09)	2 2 1 (0.18)	2 1 1 (0.19)	2 2 1 (0.16)	2 2 1 (0.08)	2 2 1 (0.17)
	25	1 2 2 (0.10)	2 2 2 (0.13)	2 2 2 (0.12)	2 2 1 (0.19)	2 1 2 (0.13)	2 2 2 (0.13)	2 2 1 (0.13)	2 2 1 (0.19)	2 2 2 (0.19)	2 2 2 (0.10)	2 2 1 (0.12)	2 2 1 (0.18)
	10	1 2 2 (0.14)	2 2 2 (0.15)	2 2 2 (0.18)	2 2 2 (0.22)	2 2 2 (0.14)	2 2 2 (0.17)	2 2 1 (0.20)	2 2 1 (0.24)	2 2 2 (0.16)	2 2 2 (0.19)	2 2 1 (0.19)	2 2 1 (0.24)
	5	1 2 2 (0.16)	2 2 2 (0.18)	2 2 2 (0.21)	2 2 2 (0.24)	2 1 2 (0.20)	2 2 2 (0.23)	2 2 2 (0.26)	2 2 2 (0.28)	2 2 2 (0.25)	2 2 2 (0.26)	2 2 1 (0.28)	2 2 2 (0.31)
$h = n$	25	1 2 1 (0.11)	2 1 1 (0.16)	2 2 1 (0.12)	2 2 1 (0.18)	2 1 1 (0.13)	2 1 1 (0.21)	2 2 1 (0.13)	2 2 1 (0.20)	2 1 1 (0.21)	2 2 1 (0.15)	2 2 1 (0.10)	2 2 1 (0.18)
	10	1 2 1 (0.17)	2 1 1 (0.22)	2 2 1 (0.20)	2 2 1 (0.24)	2 1 1 (0.19)	2 2 1 (0.24)	2 2 1 (0.20)	2 2 1 (0.25)	2 1 1 (0.25)	2 2 1 (0.20)	2 2 1 (0.18)	2 2 1 (0.25)
	5	1 2 1 (0.25)	2 1 1 (0.30)	2 2 1 (0.10)	2 2 1 (0.33)	2 1 1 (0.26)	2 2 1 (0.29)	2 1 1 (0.29)	2 2 1 (0.33)	2 1 1 (0.31)	2 2 1 (0.27)	2 2 1 (0.27)	2 2 1 (0.32)

The efficiency of an estimation procedure for each Wakeby population can be defined as MAE* (T) / MAE (T), where MAE* (T) is the minimum error among all the estimation procedures and MAE (T) is the error for the particular estimation procedure under consideration. A robust estimation procedure is such that its efficiency does not drop abruptly when it is not the winner. Therefore, a reasonable criterion for selecting the most robust estimation procedure is to search for the one that has the highest minimum efficiency among the twelve populations. That is, the maximin criteria seems to be suitable in this particular situation. Table 4 shows the minimum efficiency for all pairs (h, m) and eight estimation procedures. According to the minimax criteria, it can be noted that $A = 2$ (exponential distribution) and $C = 2$ (streamflow record + historical data) are the best choices. In some cases $B = 1$ (method of moments) is preferable and in others $B = 2$ (method of maximum likelihood) is preferable. As a rule of thumb, the method of moments might be used whenever $h \leq 4m$; otherwise the method of maximum likelihood should be used.

The fact that the exponential distribution came out of this competition as the winner, which confirms and validates the conclusion of the previous section, does not mean that we have a reliable procedure for estimating $x(T)$, for T large. For example, Kelman and Damazio (1985) have studied what would be the design of the spillway for the Salto Santiago Dam in the Iguaçu River, if only 10 years of streamflow record immediately antecedent to the year of the design were available. In other words, several estimates of $x(10,000)$ were done for different "windows" of 10 years sliding over the streamflow record.

The estimates of $x(10,000)$ ranged from 13,000 m^3/s to 40,000 m^3/s. Since in 1983 the peak flow of 17,000 m^3/s was actually observed, a catastrophe could have occurred in several circumstances. Fortunately, the spillway was designed through hydrometeorological methods and the capacity is 26,000 m^3/s, very close to the estimate of $x(10,000)$ when the full 42 years of records are used.

Kelman and Damazio (1985) have studied the probability distribution of the recurrence intervals associated with estimates, $\hat{x}(10,000)$, from different record lengths (m) sampled from an exponential distribution. They found, for example, that when $m = 5$ there is a probability equal to 0.20 that the recurrence interval of the design flood will be smaller than 100 years, when one is actually trying to estimate the 10,000 years flood event. Since underdesigning of a flood structure is much more serious than overdesigning, the authors have suggested a "safety factor", for use whenever the streamflow record is small. This safety factor was developed under the assumption that when the target is $x(10,0000)$, the probability of hitting some value smaller than $x(100)$ should be at most no more than 0.01. The safety value α was

Table 4. *Minimum Efficiency of Each Estimation Procedure, MAE*(10000)/MAE(10000) among the 12 Wakeby Distributions. (* is the "winner")*

h	m	ABC 111	ABC 112	ABC 121	ABC 122	ABC 211	ABC 212	ABC 221	ABC 222
150.	50.	0.22	0.22	0.21	0.22	0.30	*0.31	0.17	0.25
150.	25.	0.26	0.27	0.26	0.28	0.34	0.37	0.19	*0.39
150.	10.	0.45	0.46	0.44	0.56	0.56	0.60	0.28	*0.75
150.	5.	0.44	0.50	0.41	0.78	0.46	0.61	0.30	*0.87
100.	50.	0.22	0.22	0.21	0.21	0.30	*0.31	0.16	0.22
100.	25.	0.29	0.36	0.28	0.30	0.38	*0.41	0.22	0.38
100.	10.	0.46	0.49	0.45	0.54	0.56	0.63	0.31	*0.76
100.	5.	0.49	0.58	0.47	0.86	0.50	0.69	0.34	*0.94
50.	50.	0.22	0.22	0.21	0.21	0.30	*0.30	0.19	0.19
50.	25.	0.29	0.30	0.31	0.31	0.40	*0.43	0.22	0.31
50.	10.	0.45	0.49	0.45	0.51	0.56	0.63	0.33	*0.70
50.	5.	0.55	0.64	0.52	0.76	0.55	0.82	0.36	*0.89
25.	25.	0.26	0.26	0.26	0.26	0.34	*0.34	0.23	0.23
10.	10.	0.45	0.45	0.44	0.44	0.56	*0.56	0.40	0.40
5.	5.	0.61	0.61	0.60	0.60	0.71	*0.71	0.58	0.58

derived empirically for the exponential distribution as follows:

$$\alpha = \beta\left[\frac{1+8.21\gamma}{\beta(1-\gamma)+9.21\gamma}\right], \tag{18a}$$

where

$$\beta = -0.107 + 5.48m^{-0.5} - 63.26m^{-2} + 169.63^{-2.5}, \quad m < 23 \tag{18b}$$

$$\beta = 1, \qquad m \geq 23, \tag{18c}$$

and α is the coefficient of variation.

The author's recommended equation for estimating the 10,000 years flood event for the spillway design of large dams is:

$$\hat{x}(10000) = \alpha(\overline{x} + 8.21 s_x). \tag{19}$$

4. DAILY STREAMFLOW MODELING

Let us suppose it is necessary to calculate the flood control storage v^* of a man-made reservoir located upstream from a city, in such a way that the probability of downstream flooding is equal to p. By downstream flooding, we mean that the daily outflow from the reservoir is greater than a critical value y^*. If V is the random variable "maximum flood volume to be attenuated in the reservoir during a flood season of n days", one is seeking the solution to the equation

$$P(V > v^*) = p, \tag{20a}$$

where

$$V = \max_{1 \le j \le k \le n} [0, ((Y_j + Y_{j+1} - \cdots + Y_k) - (k - j + 1)y^*)] \tag{20b}$$

and Y_i is the daily inflow to the reservoir on day i.

If the random variables Y_i and Y_j were independent $\forall i \neq j$, then the probability distribution of the maximum deficit derived by Gomide (1975) could be used. However, the strong temporal persistence of daily streamflows make it necessary to search for alternative solutions to equation (20).

Beard (1963) approached the flood control design problem by defining a set of random variables $(W(1), W(2), \ldots, W(d), \ldots, W(n))$ such that

$$W(d) = \max_i \left(\sum_{j=0}^{d-1} Y_{i+j},\ i = 1,\ 2, \ldots, n - d + 1 \right). \tag{21a}$$

There is a $(1 - p)$ inflow volume quantile $w^*(d)$, associated with each duration, which is defined as:

$$P(W(d) > w^*(d)) = p. \tag{21b}$$

The graph $(d, w^*(d))$ is usually a non-decreasing curve which is called the volume-duration relationship for probability of flood p. In practice the values $w^*(d)$ are calculated by fitting a probability distribution to each random variable $W(d)$. As the estimate of the quantile $w^*(d)$ may be eventually smaller than the estimate of $w^*(d + \Delta d)$, $\Delta d > 0$, due to sample variation, "smoothing functions" are often used to assure that the function $w^*(d)$ is indeed non-decreasing. The flood control storage is selected as

$$v_B = \max_d [w^*(d) - dy^*],\ d = 1,\ 2, \ldots, n, \tag{22}$$

which is equivalent to

$$v_B = w^*(d_c) - d_c y^*,$$

where d_c is called the critical duration. It should be noted that v_B is smaller than the true value v^* because

$$\begin{aligned} P(V > v_B) &= P(W(1) > v_B + y^* \text{ or } W(2) > v_B + 2y^* \text{ or } \ldots) \\ &\ge P(W(d_c) > v_B + d_c y^*) = p. \end{aligned} \tag{23}$$

In other words, this method results in a probability of downstream flooding greater than p.

Another possibility for calculating v^* is to apply (20b) to each flood season of the streamflow record, resulting in a random sample $(v_1, v_2, \ldots, v_m)$, where m is the number of years of record. A probability distribution for V is then fitted to the random sample and v^* is ultimately estimated. However, in several flood seasons the sampled V may be zero. In other words, there is a probability mass on zero, $P(V = 0) > 0$, and therefore the number of positive observations of V is smaller than the number of flood seasons m. Consequently it is difficult to define the probability distribution of V, for positive V, unless m is exceptionally large. As this is seldom the case, a stochastic model may be employed through the empirical probability distribution of V to produce as many synthetic flood seasons as necessary to estimate v^*.

If a stochastic model is available to produce thousands of daily streamflow sequences, it is possible not only to calculate the flood storage, but also to evaluate the safety of an existing or designed spillway. This can be done by simulating the reservoir evolution and counting the number of runs that result in dam overtopping (Kelman and Damazio, 1983).

There are several daily streamflow models described in the literature; for example, those suggested by Quimpo (1967), Treiber and Plate (1975), Kelman (1977, 1980), Weiss (1977), O'Connell and Jones (1979) and Yakowitz (1979). However, these models have seldom been reported as useful in flood studies. A few exceptions could be mentioned; for example, Plate (1979), Yevjevich and Taesombut (1979), Bulu (1979), and Kelman and Damazio (1983). Perhaps the lack of popularity of daily streamflow models is due to skepticism about the capability of these models to produce synthetic sequences with the same statistical properties as the single observed time series. This writer's experience is against this skepticism and is in favor of including these models in the hydrologist's tool kit. In fact, this writer and his colleagues at CEPEL have been applying successfully a multi-site daily streamflow model called DIANA (Kelman *et al.*, 1985a) to several flood studies in Brazil (Kelman *et al.*, 1980, 1982, 1983, 1984, 1985b; Costa *et al.*, 1983; Moreira *et al.*, 1983).

It has been our experience on large basins that simple models, usually conceived on a semi-empirical basis, give best results. Perhaps this is because simple models tend to be parsimonious in the number of model assumptions, even at the cost of not being parsimonious in the number of model parameters. When it comes to daily data, the information available is usually enough to support the option in favor of simple models, very often of a non-parametric type. In other words, in daily streamflow modeling, it is better to let the data "speak for itself", rather than imposing some tight preconceived stochastic process formulation. It should be noted, however, that we are referring to large basins which are not subjected to hurricanes. In such basins an exceptional flood may result from the joint occurrence of events which are not themselves remarkable, but that can be used as "building blocks" to synthesize hydrographs different from those observed in the past.

In order to illustrate these points, a model used by Kelman and Damazio (1983) for dam safety analysis will be briefly described (which is not the DIANA model). It might not represent the best balance of the *parameters* versus *assumptions* conflict. In fact, it is biased towards minimizing the role of the assumptions in favor of empirical evidence.

Let Y_t be the mean flow on day t and let

$$Z_i = Y_i - Y_{i-1}. \tag{24}$$

The Z_i are classified in a three way table according to the following criteria:

$$\begin{array}{lll} & Z_i > 0 & \rightarrow \quad a = 1 \\ A- & & \\ & Z_i \leq 0 & \rightarrow \quad a = 2 \\ B- & q_{j-1} \leq Z_{i-1} < q_j & \rightarrow \quad b = j \\ C- & \tau_{m-1} \leq i < \tau_m & \rightarrow \quad c = m \end{array}$$

The vector $\mathbf{q} = (q_0, q_1, q_2, \ldots, q_j, \ldots, q_r)$ partitions the range of daily flows into r intervals, whereas the vector $\boldsymbol{\tau} = (\tau_0, \tau_1, \tau_2, \ldots, \tau_m, \ldots, \tau_s)$ partitions the flood season duration into s intervals. Therefore, each value A_i may fall in one of the $2rs$ classes, according with the associated set (a, b, c). The class marks should be selected according to the peculiarities of the data. For example, one may guess that the falling (or rising) limb of the hydrographs behave differently for high and low flows and choose, by visual inspection, a component of $\mathbf{q}$ which will divide the two "states". Analogously one may observe that the floods in February "look different" from those of January and therefore choose the last day of January as one of the components of $\boldsymbol{\tau}$. Care must be taken to avoid classes with a scarcity of sample points; the number of observations in each class should be large enough to allow the use of the associated empirical distribution.

The persistence of daily streamflows is incorporated into the model through a seasonal two-state Markov chain representation:

$$\pi_c = P(Z_i \geq 0 \mid Z_{i-1} \geq 0) \tag{25}$$

and

$$\phi_c = P(Z_i < 0 \mid Z_{i-1} < 0), \tag{26}$$

where c depends on the i value, according to classification C.

Once the class mark vector $\mathbf{q}$ and $\boldsymbol{\tau}$ have been established, estimation of the transition probabilities π_1, ϕ_1, π_2, $\phi_2, \ldots, \pi_s$, ϕ_s, and the grouping of the observed z_i values according to the corresponding (a, b, c) set, is a simple matter of data manipulation. Each synthetic daily flow sequence is produced according to the following algorithm:

I) $i = 0$; sample $q(0)$ from the last-day-of-dry-season flow empirical probability distribution; $a = 1$;

II) $i = i + 1$;

III) set the value of b according to Y_{i-1} and of c according to i;

IV) sample the u value from the uniform (0, 1) distribution;

V) if $a = 2$, go to (VII);

VI) if $u > \pi_c$ then $a = 2$ and go to (VIII);

VII) if $u > \phi_c$ then $a = 1$;

VIII) sample the z_i value from the empirical distribution of the (a, b, c) class;

IX) $y_i = y_{i-1} + z_i$;

X) if i is not the last day of the flood season go to (II).

The above algorithm was used by Kelman and Damazio (1983) to produce 100,000 synthetic daily streamflow sequences for the Furnas Dam, on the Grande River, Brazil. A 32 year record of daily streamflows provided input data for the model. The class marks chosen were: $y_0 = 0$, $y_1 = 1000$, $y_2 = 2000$, $y_3 = \infty$ (m^3/s); and τ_0 = Dec. 1, τ_1 = Jan. 1, τ_2 = Feb. 1, τ_3 = March 1, τ_4 = April 1 and τ_5 = May 1.

Figure 6 shows a comparison between the empirical probability distribution of annual maximum streamflow derived from the two sequences. The good matching, evident by eye inspection, can be confirmed by the chi-squared goodness-of-fit statistic of 1.01, using six grouping intervals.

Table 5 shows a comparison between the statistics associated with the random variables "daily streamflow" and "annual maximum streamflow". It

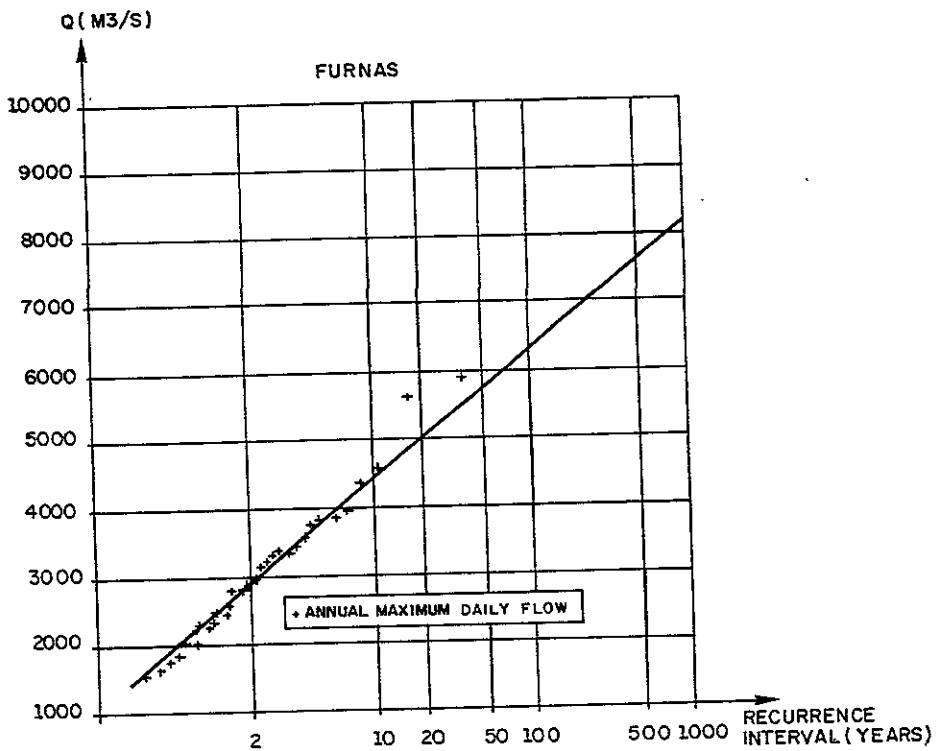

Figure 6. *Annual Maximum Distribution*

is found that the historical statistics are contained within the 95% confidence interval obtained from the synthetic realizations. In other words, one cannot reject the null hypothesis that the historical series was produced by the model. This is equivalent to saying that the model itself cannot be rejected.

The 100,000 synthetic sequences were generated by a VAX 11/780 computer in 90 minutes of CPU time and only 28 synthetic sequences were considered as "adverse hydrographs" for dam safety analysis. It seems to be a waste of computer time to generate 99972 sequences just to find out that they were not critical and consequently that they would not be necessary for simulation.

Let us assume that each streamflow sequence is a point of a sample space and we are interested in finding the probability of an event A in this sample space, as well as to simulate the system's performance for several

Table 5. *Comparison Between Statistics of 91 Synthetic Sequences and 1 Historical Sequence, Each One of Them of 32 "Flood Seasons".*

	DAILY STREAMFLOW				ANNUAL MAXIMUM STREAMFLOW			
	MEAN	STD. DEV.	SKEW	KURT	MEAN	STD. DEV.	SKEW	KURT
HIST	1210	720	1.65	6.86	3089	1031	0.88	3.66
SYNT	1288	873	1.76	7.82	3102	1081	1.03	4.82
MINIMUM	1119	669	0.99	3.98	2764	628	−0.02	2.28
AVERAGE	1288	862	1.57	6.46	3102	1048	0.76	3.66
MAXIMUM	1531	1152	2.44	11.50	3536	1424	1.69	7.16
$\hat{P}$(SYNT > HIST)	0.87	0.93	0.42	0.38	0.48	0.48	0.45	0.51

sample points that belong to A. In the previous paragraph the event A would be the set of the "adverse hydrographs". It would be convenient if the model could be biased in order to increase the likelihood that a sampled (synthetic) sequence belongs to A, without distorting the reliability on the evaluation of the probability of A. Kelman (1983) approached this question by using the importance sampling technique (Hammersley and Handscomb 1964; Rubinstein, 1981).

Let

$$h'(Y) = \begin{cases} 1, & Y \in A \\ 0, & Y \notin A \end{cases} \tag{27a}$$

where Y is a daily streamflow sequence.

The algorithm of the proposed model can be seen as a function that maps a $2n$ vector U, the components of which are independent standard uniformly distributed random variables U_i, $i = 1, \ldots, 2n$, into an n vector Y of dimension equal to n. Therefore, (27a) could be rewritten as

$$h(U) = h'(Y) = \begin{cases} 1, & Y \in A \\ 0, & Y \notin A. \end{cases} \tag{27b}$$

The probability of event A, $P(A) = p$, is given by

$$p = \int_y h'(y) f_y(y)\, dy = \int_u h(u) f_U(u)\, du, \tag{28a}$$

where $f_y(\cdot)$ and $f_U(\cdot)$ are respectively the multivariate density functions of Y and U. Obviously, $f_U(u)$ is 1 when u belongs to the domain of the random variable and 0 otherwise.

The usual estimator of p, when m sequences $y(j) = \{y_i,\ i = 1, \ldots, n\}_j$, $j = 1, \ldots, m$ are available, is given by

$$\hat{P} = \frac{1}{m} \sum_{j=1}^{m} h'\left(y(j)\right), \tag{29}$$

which is unbiased ($E(\hat{P}) = p$) and has variance given by

$$\text{var}\,(\hat{P}) = \frac{p(1-p)}{m}. \tag{30}$$

Examining again the algorithm of the proposed model, one realizes that if the u value of step IV is close to unity, the hydrograph will keep rising if it was already going up, or it will start rising if it was going down. Therefore, a way of increasing the number of "critical" synthetic sequences, keeping m constant, is to sample u values that are most likely to be close to 1. For example, adopting for the marginal density the following expression:

$$f_{u_i^*}(u^*; \gamma) = (1-\gamma) + 2\gamma u_i^*,\ u_i^* \in (0,\ 1),\ \gamma \geq 0,\ i = 1,\ 2, \ldots, 2n, \tag{31}$$

(28a) can be rewritten as

$$\begin{aligned} p &= \int_{u^*} \frac{h(u^*) f_U(u^*)}{f_{U^*}(u^*)} f_{U^*}(u^*; \gamma) du^* \\ &= E_{U^*}\left(\frac{h(u^*) f_U(u^*)}{f_{U^*}(u^*)}\right) \\ &= E_{U^*}\left(\frac{h(u^*)}{f_{U^*}(u^*)}\right). \end{aligned} \tag{28b}$$

Therefore, a new estimator for p is given by

$$\tilde{P} = \frac{1}{m} \sum_{j=1}^{m} \frac{h\left(U^*(j)\right)}{f_{U^*}\left(U^*(j)\right)} = \frac{1}{m} \sum_{j=1}^{m} \frac{h'\left(Y^*(j)\right)}{f_{U^*}\left(U^*(j)\right)}, \tag{31}$$

which is also unbiased. If $f_{U^*}(\cdot)$ is properly chosen, the variance of $\tilde{P}$ may be smaller than the variance of $\hat{P}$. Mazumdar (1975) suggested that only a few independent variables U_i should be substituted by independent U_i^* variables. With this in mind, a numerical example was performed assuming that $\gamma = 0$ (no "deformation") whenever $a = 2$ (hydrograph going down). In other words, γ was only allowed to be positive for $a = 1$, which means that the synthetic hydrographs will tend to have long rising limbs, as if some uncommon feature was imposed on the genesis of the flood, for example, a cold front that stays longer than usual over the basin being investigated.

The numerical example was done with the event A defined as $A = \{X > x_T\}$, where X is the annual maximum streamflow, $X = \max\{Y_i\}$, and $T = 100$ years. According to Mazumdar (1975), the estimate of $\text{var}(\tilde{P})$ for $\gamma = \gamma_1$, when a set $\{y(j),\ j = 1 = m\}$ produced at the point $\gamma = \gamma_0$ is available, is proportional to

$$C(\gamma_0, \gamma_1) = \sum_j \frac{h'\left(y(j)\right)}{f_{U^*}\left(u^*(j); \gamma_0\right) f_{U^*}\left(u^*(j); \gamma_1\right)}. \tag{33}$$

The optimal γ value can be found through an iterative search that at each cycle uses (33) to find out the γ_1 that minimizes $\text{var}(\tilde{P})$. This best γ_1 value is in turn used as the new γ_0 value in the next cycle. In the numerical example being considered the process converged in four cycles to $\gamma = 0.28$.

Table 6. *Results of the Importance Sampling Experiment*
$(m_{eq} = p(1-p)/\ \mathrm{var}(\tilde{P}))$ $\quad m = 500$

p	0.100	0.050	0.020	0.010	0.002	0.001
T (years)	10	20	50	100	500	1000
$q(T)(\mathrm{m}^3/\mathrm{s})$	4449	5054	5803	6393	7642	8206
$\widehat{CV}(\tilde{P})$	0.21	0.28	0.28	0.37	0.58	0.94
$CV(\hat{P})$	0.13	0.19	0.31	0.44	1.00	1.41
m_{eq}/(years)	204	242	625	723	1483	1131

Twenty sequences of 500 flood sequences each were generated by the streamflow model with $\gamma = 0.28$. The empirical probability distribution of annual maxima was determined in each case, and the results are shown in Table 6. Note that m_{eq} is defined as the number of synthetic sequences which are necessary to match $\mathrm{var}(\hat{P})$ (30) with $\mathrm{var}(\tilde{P})$; as could be anticipated, $\tilde{P}$ is a better estimator than $\hat{P}$ for large recurrence intervals, and vice versa.

5. CONCLUSIONS

(a) The theory of extremes is not as useful for modeling flood streamflows as has often been suggested. This is so because: (i) one never knows to which of the asymptotic distributions, if any, the distribution of $X = \max\{Y_i,\ i = 1, \ldots, n\}$ will approach as n goes to infinity; (ii) the transient behavior (n finite) may last for very large n values; and (iii) the MSE of the estimator of $x(T)$ associated with the first asymptotic distribution may be unacceptably large.

(b) The two-parameter exponential is the most robust distribution for estimating large return period flows for flood-like data typical of Brazilian rivers.

(c) Daily stochastic streamflow modeling is a suitable approach to the study of flood phenomena. The objective of reducing computer time might be achieved by the importance sampling technique, although this topic must be further investigated and may eventually become obsolete with the advent of computers with parallel processing capability.

ACKNOWLEDGMENTS

This research was suggested by ELETROBRAS. The help received from my colleagues at CEPEL, Jorge M. Damazio, Nelson Dias and Joari Costa, is gratefully acknowledged.

REFERENCES

Beard, L. R. (1963), "Flood control operation of reservoirs." *Journal of the Hydraulics Division, Proceedings of the American Society of Civil Engineers* **89**, 1–23.

Bryson, M. C. (1974), "Heavy tailed distributions: properties and tests." *Technometrics* **16**, 61–68.

Bulu, A. (1979), "Flood frequency analysis based on a mathematical model of daily flows." In *Modeling Hydrologic Processes*. Fort Collins, Colorado: Water Resources Publications.

Cohn, T. A. (1984), "The incorporation of historical information in flood frequency analysis." M.Sc. Thesis, Cornell University.

Costa, J. P., J. M. Damazio, M. V. F. Pereira, and J. Kelman (1983), "Optimal allocation of flood control storage in a system of reservoirs." In *Proceedings of the 7th National Seminar on Production and Transmission of Electric Energy*, Brasilia, Brazil, in Portuguese.

Cramer, H., and M. R. Leadbetter (1967), *Stationary and Related Stochastic Processes*. New York: Wiley and Sons.

Damazio, J. M. (1984), 'Comment on "Quantile estimation with more or less flood-like distributions" by J. M. Landwehr, N. C. Matalas and J. R. Wallis'. *Water Resources Research* **20**, 746–750.

Damazio, J. M., and J. Kelman (1984), "Use of historical information for the estimation of the streamflow with a recurrence interval of 10000 years". Technical Report, CEPEL 650/84, in Portuguese.

Damazio, J. M., J. C. Moreira, J. P. Costa, and J. Kelman (1983), "Selection of a method for estimating streamflows with a large recurrence interval." *Proceedings of the 5th Brazilian Symposium of Hydrology and Water Resources* **2**, 145, Blumenau, in Portuguese.

Gomide, F. L. S. (1975), "Range and deficit analysis using Markov chains." Hydrology Paper no. 79, Colorado State University.

Grigoriu, M. (1979), "On the prediction of extreme flows." In *Inputs for Risk Analysis in Water Systems*, ed. E. A. McBean, K. W. Hipel, and T. E. Unny, pp. 27–46. Fort Collins, Colorado: Water Resources Publications.

Gumbel, E. J. (1958), *Statistics of Extremes*. New York: Columbia University Press.

Hammersley, J. M., and D. C. Handscomb (1964), *Monte Carlo Methods*. London: Methuen.

Henriques, A. G. (1981), "Analysis of the frequency distribution of the annual maximum." National Laboratory of Civil Engineering (LNEC), Lisbon, Portugal, in Portuguese.

Hosking, J. R. M., and J. R. Wallis (1984), "Palaeoflood hydrology and flood frequency analysis." AGU Fall Meeting.

Hosking, J. R. M., J. R. Wallis, and E. F. Wood (1985), "An appraisal of the regional flood frequency procedure in the U.K." *Flood Studies Report, Hydrological Sciences Journal* **30**, 85–109.

Houghton, J. C. (1977), *Robust Estimation of the Frequency of Extreme Events in a Flood Frequency Context.* Cambridge, MA: Harvard University Press.

Kelman, J. (1977), "Stochastic modeling of hydrologic intermittent daily processes." *Hydrology Paper no. 89, Colorado State University, Fort Collins.*

Kelman, J. (1980), "A stochastic model for daily streamflow." *Journal of Hydrology* **47**, 235–249.

Kelman, J. (1983), "Floods and hydroplants." Thesis submitted in the competition for the full professorship in the hydraulics department of the Federal University of Rio de Janeiro.

Kelman, J., J. P. Costa, J. M. Damazio, and V. M. S. Barbalho (1985b), "Flood control in a multireservoir systems." Fourth International Hydrology Symposium, Fort Collins, Colorado.

Kelman, J., and J. M. Damazio (1983), "Synthetic hydrology and spillway design." XX Congress of the International Association for Hydraulic Research, Moscow.

Kelman, J., and J. M. Damazio (1984), "The 1982 flood of the Iguaçu River at Salto Santiago." *Brazilian Journal of Engineering, Water Resources,* Vol. 2–no. 2, in Portuguese.

Kelman, J, J. M. Damazio, and J. P. Costa (1985a), "A multivariate synthetic daily streamflow generator." Fourth International Hydrology Symposium, Fort Collins, Colorado.

Kelman, J., J. M. Damazio, J. P. Costa, and M. V. F. Pereira (1980), "Reservoir operation for flood control." *Brazilian Journal of Hydrology and Water Resources* **2**, in Portuguese.

Kelman, J., J. M. Damazio, M. V. F. Pereira, and J. P. Costa (1982), "Flood control restrictions for a hydroelectric plant." In *Decision Making for Hydrosystems Forecasting,* Water Resources Publications.

Kottegoda, N. T. (1980), *Stochastic Water Resources Technology.* New York: Macmillan.

Landwehr, J. M., N. C. Matalas, and J. R. Wallis (1980), "Quantile estimation with *more or less flood-like distributions." Water Resources Research* **16**, 547–555.

Mazumdar, M. (1975), "Importance sampling in reliability estimation." Reliability and Faulty Tree Analysis, *SIAM,* Philadelphia, pp. 153–163.

Moreira, J. C., J. M. Damazio, J. P. Costa, and J. Kelman (1983), "Estimation of extreme flows: partial series or annual maxima?" *Proceedings of the 5th Brazilian Symposium of Hydrology and Water Resources,* vol. 5, pp. 135, Blumenau, Brazil, in Portuguese.

Myers, V. A. (1981), "Estimation of probable maximum precipitation in tropical regions." Conference presented at ELETRONORTE, Brazilia, Brazil, on December 16, 1981.

N.E.R.C. (Natural Environment Research Center) (1975), Flood Studies Report, United Kingdom.

O'Connell, P., and D. A. Jones (1979), "Some experience with the development of models for the stochastic simulation of daily flows." In *Inputs for Risk Analysis in Water Systems,* ed. E. A. McBean, K. W. Hipel and T. E. Unny, pp. 287–312. Fort Collins, Colorado: Water Resources Publications.

Plate, E. (1979), "Extreme values models". In *Inputs for Risk Analysis in Water Systems,* ed. E. A. McBean, K. W. Hipel and T. E. Unny, pp. 3–26. Fort Collins, Colorado: Water Resources Publications.

Quimpo, R. G. (1967), "Stochastic model of daily flow sequences." Hydrology Paper No. 18, Colorado State University.

Rosbjerg, D. (1979), "Analysis of extreme events in stationary dependent series." In *Inputs for Risk Analysis in Water Systems,* ed. E. A. McBean, K. W. Hipel and T. E. Unny, pp. 69–75. Fort Collins, Colorado: Water Resources Publications.

Rubinstein, R. Y. (1981), *Simulation and the Monte Carlo Method.* New York: Wiley and Sons.

Slack, J. R., Wallis, J. R., and N. C. Matalas (1975), "On the value of information to flood frequency analysis." *Water Resources Research* **11**, 629–647.

Treiber, B., and E. J. Plate (1975), "A stochastic model for the simulation of daily flows." Symposium and Workshop on the Application of Mathematical Models in Hydrology and Water Resources, Bratislava, Czechoslovakia.

USWRC (U.S. Water Resources Council) (1967), *Uniform Technique for Determining Flood Flow Frequency.* Bulletin no. 15.

USWRC (U.S. Water Resources Council) (1977), *Guidelines for Determining Flood Flow Frequency.* Bulletin no. 17A.

Wallis, J. R. (1981) "Hydrologic problems associated with oilshale development." IFIP Conference, Italy.

Weiss, G. (1977), "Shot noise models for the generation of synthetic streamflow data." *Water Resources Research* **13**, 101–108.

World Meteorological Organization (WMO) (1983), "Manual for estimation of probable maximum precipitation." Operational Hydrology Report no. 1, WMO, no. 332, Genova, 190 pp.

Yakowitz, S. J. (1979) "A nonparametric Markov model for daily river flow." *Water Resources Research* **15**, 1035–1043.

Yevjevich, V., and V. Taesombut (1979), "Information on flood peaks in daily flow series." In *Inputs for Risk Analysis in Water Systems,* ed. E. A. McBean, K. W. Hipel, and T. E. Unny, pp. 171–192. Fort Collins, Colorado: Water Resources Publications.