STATISTICAL APPROACH TO FLOODS*

Jerson Kelman

Electrical Energy Research Center - (CEPEL)

Caixa Postal 2754 - CEP 20001 - Rio de Janeiro , Brazil

ABSTRACT

The most 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. This model selection criteria is investigated through the evaluation of the errors on the estimation of x(T) for a Markovian daily flow stochastic process.

The design of spillways and/or flood control storage requires the calculation of the T flood hydrograph, rather than just the peak value. Questions regarding the evolution of reservoir storage could be easily solved if a large number of daily streamflow sequences were available to be used in the evaluation of the frequency of failure of each temptative 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.

* Presented at the Symposia on Statistics in Honour of Professor V.W. Joshi's 70th Birthday, University of Western Ontario , Canada, May 27-31, 1985.

1. INTRODUCTION

Hydrologists are usually called to colaborate in the design of hydraulic structures aimed to support streamflows up to a critical event, the so called "design flood". When the failure of the structure can have catastrophic consequences, the design flood is often calculated through a hydrometeorological approach, which maximizes the previously observed storms with the purpose of reaching an event that "with all likelihood" will never happen. It is interesting to note that this "upper limit" for storms is called the probable maximum precipitation, although no probability calculation is used in its derivation. Descriptions of this methodology for applications in temperate regions are easily found (for example WMO 1973) but for tropical regions there are few references, as for example the work of 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 exceedence probability to each flood event.

In some engineering problems one only need 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. Engineers like to call $T=p^{-1}$ as the "recurrence interval" in years, as this is the expected time interval for the ocurrence of the first flow larger than $x(T)$. For major hydraulic structures T is sometimes chosen to be as large as 10 000 years. The most usual approach for the calculation of $x(T)$ is to fit a probability distribution $\hat{F}(.)$ to a set of observations of m annual maxima $\{x_1, x_2, \ldots, x_m\}$ and get the 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, that is, $P(\hat{X}(T) < x(T))$?

The answer to questions (a) and (b) may be different because the errors on the parameter estimation of the population distribution may be so high that the use of a wrong distribution, possibly with few parameters, is perhaps the best choice. There are several results available in the literature aimed to answer 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 available when the population distribution is unknown.

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

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

This asymptotic distribution, also called Gumbel, is still valid even when the random variables Y_i are weakly dependent, as 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 no asymptotic distribution for X, or else, associated to the second (also called Frèchet) and third (also called Weibull) asymptotic distributions, rather than to the first.

Since most of the probability distributions used in Hydrology are of the exponential type, as for example the normal 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. Section 2 will discuss the question of how good is this approximation.

The other usual approach to the selection of an approximate probability distribution of X, not necessarily confined to the set of the asymptotic distributions, is to examine a number of candidate distributions and pick up 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.

Comparison studies have been made with data from a great number of streamflow gauges aiming to get a standardized distribution of annual maximum. In the United States the Water Resources Council (USWRC, 1967) suggested the use of the log-Pearson III distribution and later on furnished further guidelines regarding the estimation procedure (USWRC, 1977). This recommendation created a great deal of controversy. For example Wallis (1981) proved that the 500-year-flood divided by 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", have also been subject to a well founded criticism (Hosking et al, 1985).

One may wonder if goodness of fit is a reasonable criteria for selecting an approximation for 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 usually really matters is the unknown fit for large T values. Houghton (1977) and Moreira (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 criteria for selecting the approximate probability distribution of the annual maximum.

So far only the question of how to estimate the peak flow x(T) has been mentioned. However there are some engineering problems which require also the inflow volume for different durations. For example, the sizing of the flood storage in a man made reservoir. In this case the operation rule of the reservoir is such that if any water is in the flood storage, the operation target is to empty the storage as fast as possible, constrained by an upper limit on the outflow, beyond which downstream flooding would happen. There is always the possibility that the inflow volume for a particular duration is so large that the attenuation in the flood storage is not enough to avoid downstream losses. The problem is how to calculate a flood storage with a failure recurrence interval of T (typical value, T = 50 years), for a given upper limit on the outflow rate.

A similar problem is the design of a spillway. 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 operation rule is to empty it as quick as possible. Therefore the only limitation on the outflow rate is set by the hydraulic conditions of the spillway and it will not be constant. Furthermore, as this is an operation required for dam protection, no constraints regarding downstream flooding are taken into account meanwhile the safety storage is being voided. The problem is how 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 dam break with catastrophic downstream effects, and T is therefore assumed very large, say 1000 or 10000 years. Obviously the larger is the spillway capacity the smaller is the safety storage, and vice-versa.

Questions regarding the evolution of reservoir storage could be easily solved if a large number of daily flow sequences were available to be used in the evaluation of the frequency of failure of each temptative 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 focused 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 the non-stationary of the daily flow process can be neglected during the flood season, that lasts for n days. In this case it is easy to get some insight on how the Gumbel distribution approximates the true distribution of annual maximum of daily streamflow. Initially let's accept the very 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

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

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

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

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 (1958, pp. 219) comment about a graph similar to Figure 1 (Gumbel's book Figure 6.2.1(3), 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 streamflow records are 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 that the Y_i are log-normally distributed with $E(Y_i)=var(Y_i)=1$, $\forall i$. Again the curves can't be approximated by straight lines, but opposite 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 Figure 1 and 2 are different, meaning that the marginal distribution of daily flow Y_i is an extremely relevant aspect to be considered when estimating x(T).(Grigoriu, 1979).

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, pgs. 119,120,136,146), $F_x(x,n)$ will converge to the Gumbel distribution with growing n in very 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 called as "heavy-tailed". More precisely, it can be said that if the conditional mean excedance defined as $E(Y-y|Y>y)$ is a decreasing (increasing) function of y- at least for sufficiently large y- then the probability distribution function of Y is light (heavy) tailed (Bryson, 1974).

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

$$Y_i = \exp(W_i) \text{ and}$$

$$W_i = \alpha + \gamma (W_{i-1}) - \alpha) + \beta (1 - \gamma^2)^{0.5} N_i \qquad (4)$$

where

N_i is a standard normal and $E(N_i N_k) = \begin{cases} 0, & k \neq i \\ 1, & k = i \end{cases}$

We can't expect that this simple Markovian process will actually resemble daily streamflows, but it is useful to get some insight on 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 easily derived

$$E(Y_i) = \exp(\alpha + \beta^2/2) \qquad (5a)$$

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

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

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

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

$$\text{var}(Y_i|y_{i-1}) = [\exp(2(\beta^2(1-\gamma^2) + \alpha(1-\gamma))) + -\exp(\beta^2(1-\gamma^2) + 2\alpha(1-\gamma))]\, y_i^{2\gamma} \qquad (5f)$$

Assuming $\alpha = \ell n\, 2^{-0.5} = -0.35$, $\beta = (\ell n 2)^{0.5} = 0.83$ and $\gamma = [\ell n 2]^{-1}\, \ell n(1+0.95) = 0.96$, it is possible to show, by back substitution in the above equations, that $E(Y_i) = var(Y_i) = 1$, skew $(Y_i) = 4$ and 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 largest is the streamflow today, the less precise will be the flow forecast for tomorrow.

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

$$F_X(x;n) = P(Y_1 \leq x,\ Y_2 \leq x, \ldots, Y_n \leq x) = \underbrace{\int_{-\infty}^{\frac{\ell n x - \alpha}{\beta}} \cdots \int_{-\infty}^{\frac{\ell n x - \alpha}{\beta}}}_{n} \Phi_n(\underline{u})\, d\underline{u} \tag{6}$$

where Φ_n is the n-variate density function of the standard normal. This n-fold integral is very difficult to be obtained for large n values.

A first possible approximation for $F_X(x;n)$ is (Rosbjerg, 1979)

$$F_X(x;n) \simeq F_1(x;n) = P(Y_1 \leq x) \prod_{i=2}^{n} P(Y_i < x \mid Y_{i-1} < x) =$$

$$= [\Phi_1 (\tfrac{\ell n\, x - \alpha}{\beta})]^{2-n} [\Phi_2 (\tfrac{\ell n\, x - \alpha}{\beta}, \tfrac{\ell n\, x - \alpha}{\beta}; \gamma)]^{n-1},\ n \geq 2$$

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 for $F_X(x;n)$ is to assume 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 between upcrossing (K) is exponentially distributed with mean rate (Grigoriu, 1979):

$$\mu(x) = P(Y_{i+1} > x,\ Y_i \leq x)$$

$$= \Phi_1\left(\frac{\ell n\ x-\alpha}{\beta}\right) - \Phi_2\left(\frac{\ell n\ x-\alpha}{\beta}, \frac{\ell n\ x-\alpha}{\beta};\ \gamma\right) \qquad (8)$$

or simply

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

That is

$$F_K(k) \simeq 1 - \exp\left((\Phi_2 - \Phi_1)k\right) \qquad (9)$$

But

$$P(X \leq x) = F_X(x;n) = P(K>n) \simeq 1 - F_K(n) = \exp\left((\Phi_2 - \Phi_1)n\right)$$

In short, the second approximation to $F_X(x;n)$ is

$$F_2(x;n) = \exp\left((\Phi_2 - \Phi_1)n\right) \qquad (10)$$

A third approximation can be obtained through the Monte Carlo approach, using (4) to generate s sequences $\{Y_1, Y_2, \ldots, Y_n\}_j$ $j=1,s$. 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 , std.dev. (X)=2.23, skew (X)=2.74, Kurt (X)=18.72

These values are very different from the descriptors of the Gumbel distribution (skewness of 1.14 and Kurtoses of 5.4). Also for comparison it is displayed the curve for the independent process, which is exactly calculated by Eq. 2.

It can be noted in Figure 3 that the time persistence of daily streamflows does not play a role as relevant as the marginal distribution (see also Figure 1), although the time persistence can't 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 error on the estimation of X(T). For example, the error on the approximation of x(1000) in this particular case was of the order of 12%. This isn't too much when one thinks about all other sources of uncertainty usually found in the study of floods. But since we are talking about an avoidable error, the recommendation in this subject is to adopt the empirical distribution, $F_3(x;n)$, rather than the approximations $F_1(x;n)$ or $F_2(x;n)$.

Now we would like to know how good 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 some other fitting probability distribution, as well as to a "time series approach". Therefore we will be considering three alternative approaches for estimating x(T) and we want to find out which of them will lead in

the 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 estimate $\hat{\psi}$ and $\hat{\mu}$ (Eq. 1) are found through the iterative algorithm

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

$$\psi_o = 1.28/s_x \tag{11b}$$

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

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

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

where

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

b) The exponential distribution (EXD)- there are several competitive distributions to the first asymptotic, as for example the gamma, the log-pearson type III, the generalized extreme value, and others. The two parameters exponential was selected here for reasons that will become clear in the next section. It's probability distribution is defined as

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

It can be easily shown that skew (X) = 2 and Kurt(X) =9

The adopted estimation procedure is

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

and

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

c) The time series approach (TSA) - it uses the transformed daily streamflow record $\{\ell n\ y_i,\ i=1,\ n\}_j, j=1,\ m$ in order to estimate α, β and γ. The estimates are used in equation (7) to get $F_1(x(T);\ n)$ and ultimately $x(T)$. According 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 has been ruled out from the Monte Carlo experiment which description follows because it would be computationaly unfeasible.

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 as above defined.

Equation 4 was used to synthetize s = 1000 sets of m = 20 years of "streamflow data", each year with a "flood season" of n = 100 days. The three above described alternatives were employed to each set in order to estimate $x(T)$ for T= 100, 1000 and 10000 years. That is, $F_X(x)$ is respectivaly 0.99, 0.999 and 0.9999. The results are displayed in Table 1:

Table 1. Results of the Monte Carlo Experiment

Method	x(100) = 11.46			x(1000) = 18.99			x(10000) = 30.32		
	BIAS	STDV	RMSE	BIAS	STDV	RMSE	BIAS	STDV	RMSE
GUD	-3.11	1.75	3.57	-7.59	2.49	7.99	-15.88	3.27	16.21
EXD	0.18	2.47	2.48	-1.91	3.75	4.21	- 7.80	5.03	9.28
TSA	1.43	3.25	3.55	2.24	6.21	6.60	2.49	10.77	11.05

Columns

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

The estimator $\hat{X}(T)$ associated to the GUD method has the smallest variance but on the other hand it has such a large bias

that it would be highly non recomended in this particular case. For example $E(\hat{X}(T))$ is roughly half the true value for T = 1000 or 10000. Also for these two T values, confidence intervals around an estimate $\hat{x}(T)$ will not tend to contain the true value $x(T)$, if these confidence intervals are calculated by the usual procedure. That is, if X is distributed as Gumbel and if the method of maximum likelihood is employed, then $\hat{X}(T)$ is asymptotically distributed as normal with $E(\hat{X}(T))=x(T)$ and var $(\hat{X}(T))$ given by (Henriques, 1981),

$$\text{var}_A(\hat{X}(T))=\frac{\text{var}(X)}{m}(0.67+0.37(\ell n(-\ell n(1-T^{-1})))^2-0.33\ \ell n\ (-\ell n(1-T^{-1}))) \qquad (14)$$

For example, for T=1000, var $(X)=2.23^2$, m=20, Eq. 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. Assuming a particular estimate $\hat{x}(T)$ as 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 a wrong choice in this particular situation. This is a hint against the belief spread among hydrologists that the asymptotic theory for extremes is a sound approach to flood modeling.

The estimator $\hat{X}(T)$ associated to 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 rational for this hypothetical loss function is that an 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 10000, although its estimator $\hat{X}(T)$ is sistematically the one with the largest variance.

3. PROBABILITY DISTRIBUTION FOR ANNUAL MAXIMUM

The exponential distribution (Eq. 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 very 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 streamflow annual maximum is not new. Slack et al (1975) developed a Monte Carlo experiment in which many random samples of different sizes were produced by parent population distritubions $F(x)$ 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, either to make the role of $F(x)$ or of $G(x)$: 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 10000 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 $G(x)$ be either Wakeby, Gumbel or Log-normal. The Wakeby distribution is well suited for Monte Carlo studies because it can reproduce different shapes of the probability distributions usually employed in Hydrology and also because it is easy to generate synthetic samples. The random variable X distributed as Wakeby is defined as

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

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 Landewehr et al (1980) was that the Gumbel and Log-normal distributions resulted on 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 Landewehr et al (1980), adding the two parameter exponential distribution (Eq. 12) to the list of the G(x) distributions. He found out that for T larger than 200 years the exponential distribution with the parameters estimated through the method of moments resulted on the smallest cumulative (along the populations) mean squared error. His conclusion was 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 populations F(x). For this reason, Damazio et al (1983) used regional Wakeby distributions of annual maximum, estimated for brazilian basins by a procedure suggested by Wallis (1980), for playing the F(x) role. Again the exponential distribution (Eq. 12) turn 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 on flood events which preceded the gauged record is available. In some basins there are physical evidences about flood events ocurred 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 date of ocurrence of these events. As it is not obvious if the inclusion of this kind of information actually decreases the error on the estimation of 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 palaeological information is negligible.

In other basins there is some historical information based on the memory of old people who eventually knows the highest river stage in his own life span and, with luck, also in his parentes life span. In these cases the most that the hydrologist can expect is to know that was the highest water level that ocurred before systematic measurements started, going back h years from now. This lenght 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 generaly smaller than 50). Cohn (1984) developed new techniques to incorporate this kind of historical information. He assumed F(x) to be log-Pearson III and adopted the log-normal distri - bution (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 find out what is the performance of the Exponential and of the Gumbel distributions when historical data is available for moderate h (up to 150). They defined a set of twelve popula- tion distributions F(x) of the Wakeby form, called W-1,W-2,...W-12. All the twelve have one single mode, a positive lower limit and no upper limit. Their skewness and kurtosis were selected in order to resemble typical values of the 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) and three other skewness levels corresponding to 1.5, 2.0 and 2.5 were also investigated. For each skewness level, three kurtosis values were considered, the lowest one in each case corresponding to the Log-normal distribution. Table 2 shows the main character- istics of each distribution. It should be noted that all of them have unit expected value and coeficient of variation arbi trarily chosen as 0.49.

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

The relative mean absolute error was calculated for each population F(X) and each estimation procedure.

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

Figure 5 shows the variation of MAE (10000) 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 on the record lenght m actually increases the error!

In Figure 5 it can also be noted that an increase on the length of time h has a very small effect on the error.

Table 3 shows the estimation procedure with the smallest MAE (10000) for each of the pairs (h,m) and each of the twelve population Wakeby distributions. Inside the parenthesis it is shown the corresponding MAE. It should be noted that the exponential distribution was the winner in all cases but the W-1.

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 abruptaly when it is not the winner. Therefore a reasonable criteria for selecting the most robust estimation procedure is to search for the one that has the highest minimum efficiency along the twelve populations. That is, the minimax criteria seems to be adequate 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. There are some cases that B=1 (method of moments) would be preferable and others were B=2 (method of maximum likelihood). As a rule of thumb the method of moments seems to be indicated whenever $h \leq 4m$ and otherwise the method of maximum likelihood.

The fact that the exponential distribution came out of this section 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 imediately antecedent to the year of the design were available. In other words , several estimates of x(10000) were done for different "windows" of 10 years sliding over the streamflow record.

The estimates of x(10000) ranged from 13000 m^3/s to 40000 m^3/s. As in 1983 the peak flow of 17000 m^3/s was actually observed, a catastroph could be occorred in several circumstances. Fortunately the spillway was designed through hydrometeorological methods and the capacity is 26000 m^3/s, very close to the estimate of x(10000) when the full 42 years of record are used.

Kelman and Damázio (1985) have studied the probability distribution of the recurrence intervals associated with estimates $\hat{x}(10000)$ from different record lenghts (m) sampled from 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 10000 years flood event. Since underdesigning of a flood structure is much more serious than overdesigning, the authors have suggested a "safety factor", as it is so usual in the engineering practice, to be used whenever the streamflow record is small. This safety factor was developed under the assumption that when the target is x(10000), the probability of hitting some value smaller than x(100) should be at most 0.01. The safety value J was empirically derived as

$$J = -0.107 + 5.48m^{-0.5} - 63.26m^{-2} + 169.63m^{-2.5}, \quad m \leq 23$$

$$J = 1, \quad m > 23 \qquad (18)$$

Since this study was done for the exponential distribution, the author's recommended equation for estimating the 10000 years flood event is:

$$\hat{x}(10000) = J\,(\bar{x} + 8.21\, s_x) \qquad (19)$$

4. DAILY STREAMFLOW MODELLING

Let us suppose that 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, it is meant 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 of the equation

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

where

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

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. This would be perhaps the case for annual streamflows. However the strong time persistence of daily streamflow make it necessary to search for alternative solutions of Eq. 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 to 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, very often "smoothing functions" are used to assure that the function $w^*(d)$ is indeed non-decreasing. The flood control storage is selected as

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

which is equivalent to

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

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

$$P(V > v_B) = P(W(1) > v_B + y^* \text{ or } W(2) > v_B + 2y^* \text{ or } \ldots)$$

$$\geq P(W(d_c) > v_B + d_c\, y^*) = p \tag{23}$$

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

Other possibility for calculating v^* is to apply Eq. (20b) to each flood season of the streamflow record, resulting on 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 very difficult to define the probability distribution of V, for positive V, unless m is excepcionaly large. As this is seldom the case, a stochastic model may be employed to produce as many synthetic flood seasons as necessary to estimate v^*, through the empirical probability distribution of 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 on dam overtopping (Kelman and Damazio, 1983).

There are several daily streamflow models described in the literature, as for example those suggested by Quimpo (1967) Treiber and Plate (1975), Kelman (1977, 1980), Weiss (1977) O'Connell and Jones (1975) and Yakowitz (1979).

However these models have seldom been reported as useful in flood studies. Few exceptions could be mentioned as for example Plate (1979), Taesombut and Yevjevich (1979), Bulu (1979), Kelman and Damazio (1983). Perhaps the lack of popularity of daily streamflow models is due to the 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 the skepticism and 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 very simple models, usually conceived on a semi-empirical basis, give best results. Perhaps this is so because simple models tend to be parcimonious on the number of model assumptions, even at the cost of not being parcimonious on 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 modelling, it is better to leave data "speak for itself", so to say, rather than imposing some tight pre-conceived 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 ocurrence of events which are not themselves remarkable, but that can be used as "building blocks" to synthetize 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 (whis is not the DIANA model). Is 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_i be the mean flow on day t and

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

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

$$\text{A} - \begin{cases} Z_i > 0 \rightarrow a = 1 \\ Z_i \leq 0 \rightarrow a = 2 \end{cases}$$

$$\text{B} - \; q_{j-i} \leq Z_{i-1} < q_j \rightarrow b = j$$

$$\text{C} - \; \tau_{m-1} \leq i < \tau_m \rightarrow c = m$$

The vector $\underset{\sim}{q} = (q_0, q_1, q_2, \ldots, q_j, \ldots, q_r)$ partitions the range of daily flows into r intervals whereas the vector $\underset{\sim}{\tau} = (\tau_0, \tau_1, \tau_2, \ldots, \tau_m, \ldots, \tau_s)$ partitions the flood season duration into s intervals. Therefore each value Z_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 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 $\underset{\sim}{q}$ which will divide th 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 $\underset{\sim}{\tau}$. Care must be taken to

avoid classes with scarcity of sample points. In fact the number of observations in each class should be large enough to allow the use of the associated empirical distribution.

The persistence of daily streamflow 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 t value, according to classification C.

Once the class mark vector $\tilde{q}$ and $\underset{\sim}{\tau}$ have been established, estimation of the transition probabilities $\pi_1, \phi_1, \pi_2, \phi_2, \ldots, \pi_s, \phi_s$, 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 value of b according to Y_{i-1} and of c according to t

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

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

VI) if u > π_c then a = 2 and go to (VIII)

VII) if u > ϕ_c then a = 1

VIII) sample 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, in the Grande River, Brazil, The input data to the model was a 32 years record of daily streamflows. The class marks chosen as $y_o=0$, $y_1=1000$, $y_2=2000$, $y_3=\infty$ (m3/s) and τ_o = Dec 1, τ_1 = Jan 1, τ_2 = Feb 1, τ_3 = Mar 1, τ_4 = Apr 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 to random variables "daily streamflow" and "annual maximum streamflow". It 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 say that the model itself cannot be rejected.

100,000 synthetic sequences were generated by a VAX 11/780 computer in 90 min 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 justo 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 on finding out which is the probability of an event A of this sample space, as well as to simulate the system's performance for several 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

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

where Y is a daily streamflow sequence.

Examining the algorithm of the proposed model, one can realize that it can be seen as a function that maps a 2n vector U, which componentes are independent standard uniformly distributed random variables U_i, i=1, 2n, into a n vector Y of dimension equal to n. Therefore Eq. 27 could be re-written as

$$h(U) = h'(Y) \begin{cases} = 1 \longleftrightarrow Y \in A \\ = 0 \longleftrightarrow Y \notin A \end{cases} \qquad (27b)$$

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

$$p = \int_y h'(y)\, f_y(y)\, dy = \int_u h(u)\, f_U(u)\, du \qquad (28)$$

where $f_y(\,)$ and $f_U(.)$ are respectively the multivariate density function 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 $y(j)=\{y_i, i=1,n\}_j, j=1,m$ are available, is given by

$$\hat{P} = \frac{1}{m} \sum_{j=1}^{m} h'(y(j)) \,, \qquad (28)$$

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

$$\text{var}(\hat{P}) = \frac{p\,(1-p)}{m} \qquad (29)$$

Examining again the algorithm of the proposed model, one can realize 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 more likelly to be close to 1. For example adopting for the marginal density the following expression:

$$f_{U_i^*}(u^*;\gamma) = (1-\gamma) + 2\gamma u_i^* \quad , \quad u_i^* \in (0,1) \quad , \quad \gamma \geq 0, \quad i=1,2,\ldots,2n \qquad (30)$$

Eq. 28 can be re-written as

$$p = \int_{u^*} \frac{h(u^*)\ f_U(u^*)}{f_{U^*}(u^*)}\ f_{U^*}(u^*;\gamma)\,du^* = E_{U^*}\left(\frac{h(u^*)\ f_U(u^*)}{f_{U^*}(u^*)}\right)$$

$$= E_{U^*}\left(\frac{h(u^*)}{f_{U^*}(u^*)}\right) \qquad 28\ (b)$$

Therefore a new estimator for p is given by

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

which is also unbiased. If $f_{U^*}(.)$ is properly chosen, the variance of $\tilde{P}$ may result smaller than the variance of $\hat{P}$. Mazumdar (1975) suggested that only a few independent variables U_i should be substituted bu 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 stay 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, $X = \max\{Y_i\}$, T = 100 years. According to Mazumdar (1975) the estimate of $\mathrm{var}(\hat{P})$ for $\gamma=\gamma_1$, when a set $\{y(j), j=1=m\}$ produced on the point $\gamma=\gamma_0$ is available, is proportional to

$$C(\gamma_0,\gamma_1) = \sum_j \frac{h'(y(j))}{f_{U^*}(u^*(j);\gamma_0)\, f_{U^*}(u^*(j);\gamma_1)} \qquad (32)$$

The optimal γ value can be found through an iterative search, that at each cycle uses Eq. 32 to find out the γ_1 that minimizes $\mathrm{var}(\hat{P})$. 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 4 cycles to $\gamma = 0.28$.

20 sequences of 500 flood sequences each were generated by the streamflow model with $\gamma = 0.28$. The empirical distribution probability of annual maxima was determined in each case, and the results are shown in Table 6.

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

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

m_{eq} defined as the number of synthetic sequences which are necessary to match $\mathrm{var}(\hat{P})$ (Eq. 29) with $\mathrm{var}(\tilde{P})$. As could be antecipated, $\tilde{P}$ is a better estimator than $\hat{P}$ for large recurrence intervals, and vice-versa.

5. CONCLUSIONS

a) Theory of extremes is not as useful for modelling flood streafmlows as has been often suggested. This is so because: a) one never knows to which of the asymptotic distributions, if to any, the distribution of $X=\max \{Y_i, i=1,n\}$ will approach as n goes to infinity: b) the transient behavior (n finite) may last for very large n values; c) the MSE of the estimator of x(T) associated to 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 the brazilian rivers.

c) Daily stochastic streamflow modelling is a suitable approach to the study of flood phenomena. The computer time reducing objective might be achieved by the importance sampling technique, although this topic must be further investigated and eventually will become obsolete due to the recent announcements of new computers with paralel processing capability.

ACKNOWLEDGEMENTS

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, ASCE, vol. 89, nºHY1, 1-23.

Bryson, M.C., 1974, "Heavy Tailed Distribution: Properties and Test, Tecnometrics", vol. 16, nº 1.

Bülu, A., 1979, "Flood Frequency Analysis Based on a Mathematical Model of Daily Flows", in Modelling Hydrologic Processes, Water Resources Publications.

Cohn, T.A., 1984, "The Incorporation of Historical Information in Flood Frequency Analysis", M.Sc. Thesis, Cornell University

Costa, J.P., Damazio, J.M., Pereira, M.V.F., Kelman, J., 1983, "Optimal Allocation of Flood Control Storage in a System of Reservoirs," Proceedings of the 7th National Seminar on Production and Transmission of Electric Energy", Brasilia, Brazil, in portuguese.

Cramer, H. and Leadbetter, M.R., 1967, "Stationary and Related Stochastic Processes", Wiley

Damazio, J.M., Moreira, J.C., Costa J.P., Kelman, J., 1983, "Selection of a Method for Estimating Streamflows With a Large Recurrence Interval", Proceedings of the 5th Brazilian Symposium of Hydrology and Water Resources, vol. 2, pp. 145, Blumenau, in portuguese.

Damazio, J.M., 1984, Comment on "Quantile Estimation With More or Less Floodlike Distributions by J.M. Landwehr, N. C. Matalas and J.R. Wallis, Water Resources Research, June.

Damazio, J.M. and Kelman, J., 1984, "Use of Historical Informa - tion for the Estimation of the Streamflow With a Recurrence Interval of 10000 Years", Technical Report, CEPEL 650/84, in portuguese.

Gomide, F.L.S., 1975, "Range and Deficit Analysis Using Markov Chains," Hydrology Paper nº 79, Colorado State University, USA.

Grigoriu, M., 1979, "On the Prediction of Extreme Flows, in "Inputs for Risk Analysis in Water Systems",Water Resources Publications.

Gumbel, E.J., 1958, "Statistics of Extremes, Columbia University Press.

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 Wallis, J.R., 1984, "Palaeoflood Hydrology and Flood Frequency Analysis", AGU Fall Meeting.

Hosking, J.R.M., Wallis, J.R., Wood, E.F., 1985, "An Appraisal of the U.K. Flood Studies Report - Nine Years on", Hydrological Sciences Journal (March).

Houghton, J.C., 1977, "Robust Estimation of the Frequency of Extreme Events in a Flood Frequency Context", Harward University, Cambridge, USA.

Hamanersley, J.M. and Handscomb, D.C., 1964, "Monte Carlo Method" Methuen and Co. Ltd.

Kelman, J., Damazio, J.M., Costa J.P., Pereira M.V.F., 1980, "Reservoir Operation for Flood Control", Brazilian Journal of Hydrology and Water Resources 2 (2), in portuguese.

Kelman, J., Damazio, J.M., Pereira, M.V.F., Costa, J.P., 1982, "Flood Control Restrictions for a Hydroelectric Plant" in "Decision Making for Hydrosystems Forecasting", Water Resources Publications.

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. and Damazio J.M., 1983, "Synthetic Hydrology and Spillway Design", XX Congress of the International Association for Hydraulic Research, Moscow.

Kelman, J. and Damazio J.M., 1984, "The 1982 Flood of the Iguaçu River at Salto Santiago", Brazilian Journal Engineering, Water Resources, vol. 2 - nº 2, in portuguese.

Kelman, J., Damazio, J.M., Costa, J.P., 1985a, "A Multivariate Synthetic Daily Streamflow Generator", Fourth International Hydrology Symposium, Fort Collins, Colorado, USA.

Kelman, J., Costa J.P., Damazio, J.M., Barbalho, V.M.S., 1985b, "Flood Control in a Multireservoir Systems" Fourth International Hydrology Symposium, Fort Collins, Colorado, USA.

Kelman, J., 1977, "Stochastic Modelling of Hydrologic Intermittent Daily Processes", Hydrology Paper 89, Colorado State University, USA.

Kelman, J., 1980, "A Stochastic Model for Daily Streamflow, Journal of Hydrology, 47.

Kottegoda, N.T., 1980, "Stochastic Water Resources Technology", The Macmillan Press.

Landewehr, J.M., Matalas, W.C. and Wallis, J.R., 1980, "Quantile Estimation with More or Less Floodlike Distributions", Water Resources Research, vol. 16, nº 3.

Mazumdar, M., 1975, "Importance Sampling in Reliability Estimation", Reliability and Fault Tree Analysis, SIAM, Philadelphia, USA, pp. 153-163.

Moreira, J.C., Damazio, J.M., Costa J.P., Kelman, J., 1983, "Estimation of Extreme Flows: Partial Series or Annual Maxima ?" Proceeding 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, in December 16, 1981.

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

O'Connell, P. and Jones, D.A., 1979, "Some Experience with the Development of Models for the Stochastic Simulation for Daily Flows", in "Inputs for Risk Analysis in Water Systems", Water Resources Publications.

Plate, E., 1979, "Extreme Value Models", in "In Inputs for Risk Analysis in Water Systems", Water Resources Publications.

Quimpo, R.G., 1967, "Stochastic Model of Daily Flow Sequences ", Hydrology Paper 18, Colorado State University, USA.

Rosbjerg, D., 1979, "Analysis of Extreme Events in Stationary Dependent Series", in "Inputs for Risk Analysis in Water Systems", in Water Resources Publications.

Rubinstein, R.Y., 1981, "Simulation and the Monte Carlo Method, Willey Series in Probability.

Slack, J.R., Wallis, J.R. and Matalas, N.C., 1975 "On the Value of Information to Flood Frequency Analysis", Water Resources Research, vol. 11, nº 5.

Treiber, B. and Plate, E.J., 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, Tchecoslovaquia.

U.S.W.R.C. (U.S. Water Resources Council), 1967, Uniform Technique for Determining Flood Flow Frequencies, Bulletin, nº 15.

U.S.W.R.C. (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, Vol. 13, nº 1.

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 Non-Parametric Markov Model for Daily River Flow", Water Resources Research, Vol. 15, nº 5.

Yevjevich, V. and Taesonbut, V., 1979, "Information on Flood Peaks in Daily Flow Series", in "Inputs for Risk Analysis in Water Systems", in Water Resources Publications.

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=10000
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

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	W-2	W-3	W-4	W-5	W-6	W-7	W-8	W-9	W-10	W-11	W-12
		A B C	A B C	A B C	A B C	A B C	A B C	A B C	A B C	A B C	A B C	A B C	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 (0,16)
	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 (0,25)
	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 2 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)

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

h	m	A B C 1 1 1	A B C 1 1 2	A B C 1 2 1	A B C 1 2 2	A B C 2 1 1	A B C 2 1 2	A B C 2 2 1	A B C 2 2 2
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

Table 5. Comparison Between Statistics of 31 Synthetic Sequences and 1 Historical Sequence, Each One of Them of 32 "Flood Seasons".

	DAILY STREAMFLOW				ANNUAL MAXIMUM STREAFMLOW			
	M E A N	STD.DEV.	S K E W	K U R T	M E A N	STD.DEV.	S K E W	K U R T
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

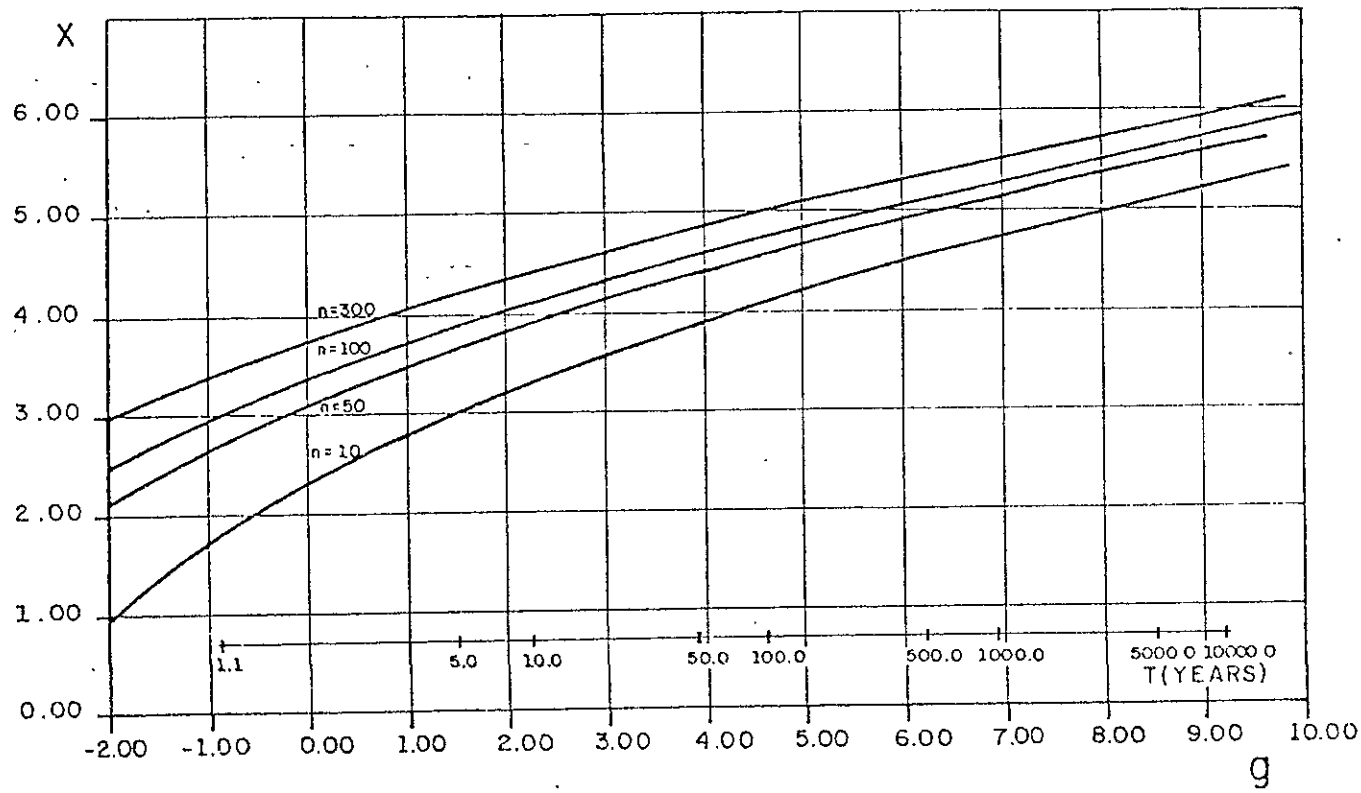


Figure 1. Probability Distribution of $X=\max\{Y_i,\ i=1,n\}$
$E(Y_i)=\mathrm{var}(Y_i)=1$, (Y_i,Y_j) independent, Y_i normally distributed

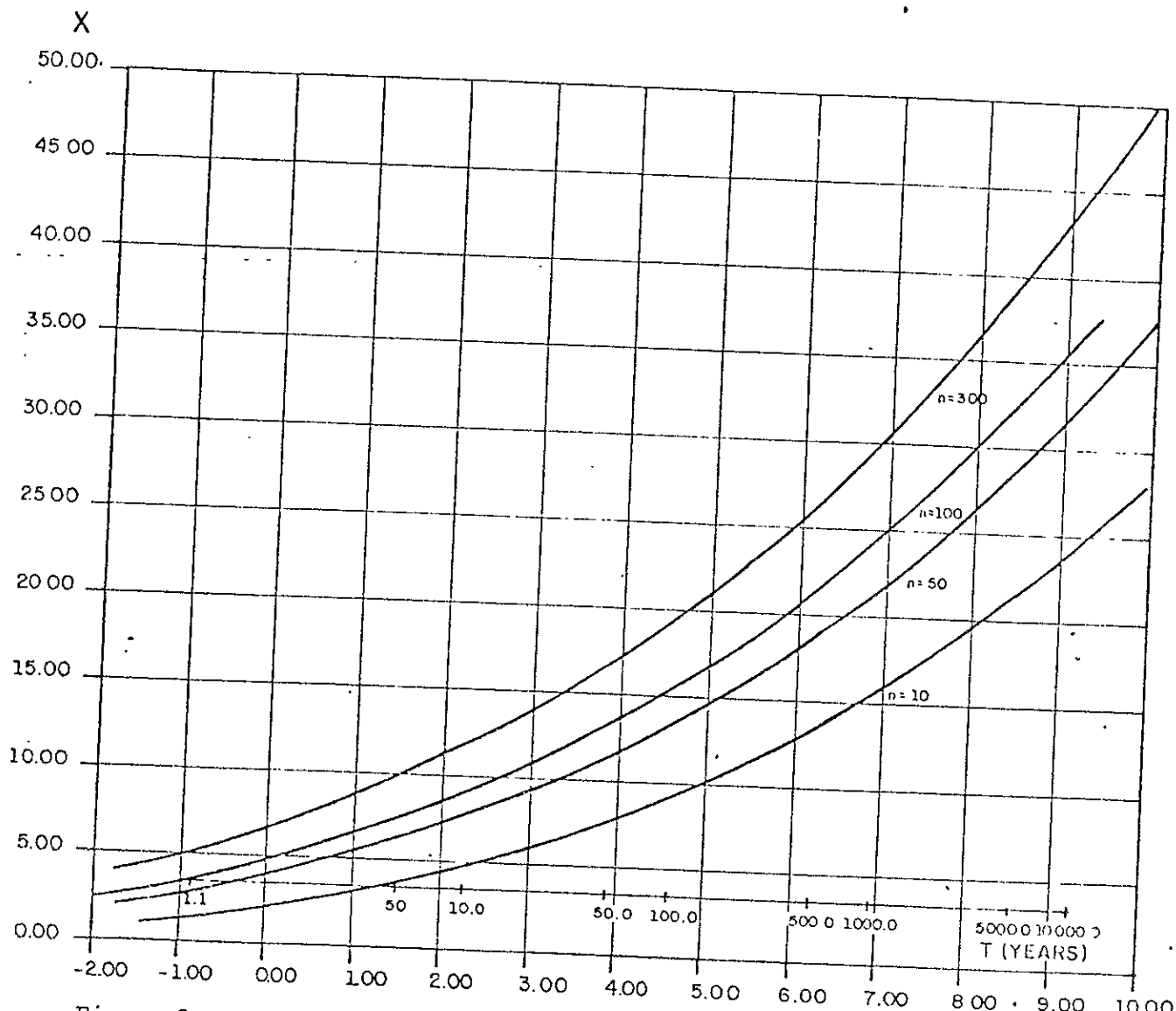


Figure 2. Probability Distribution of $X=\max_i\{Y_i, i=1,n\}$

$E(Y_i)=\text{var}\ (Y_i)=1$, (Y_i,Y_j) independent, Y_i normally distributed

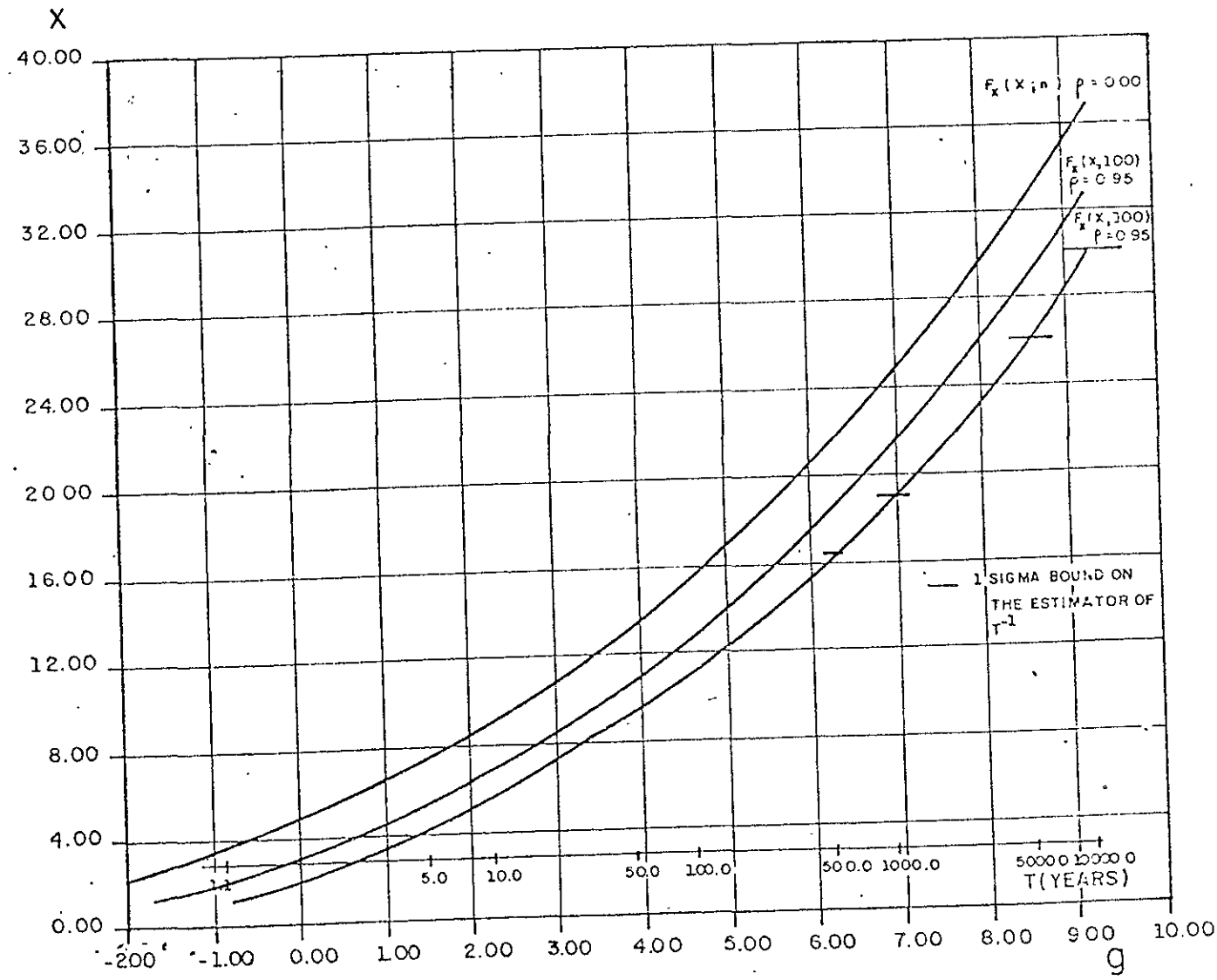


Figure 3. Approximations to the Probability Distribution of $X=\max_i \{Y_i, i=1,j\}$, $E(Y_i)=\text{var}(Y_i)=1$, corr $(Y_i, Y_{i+1}) = 0.95$, Y_i log-normally distributed

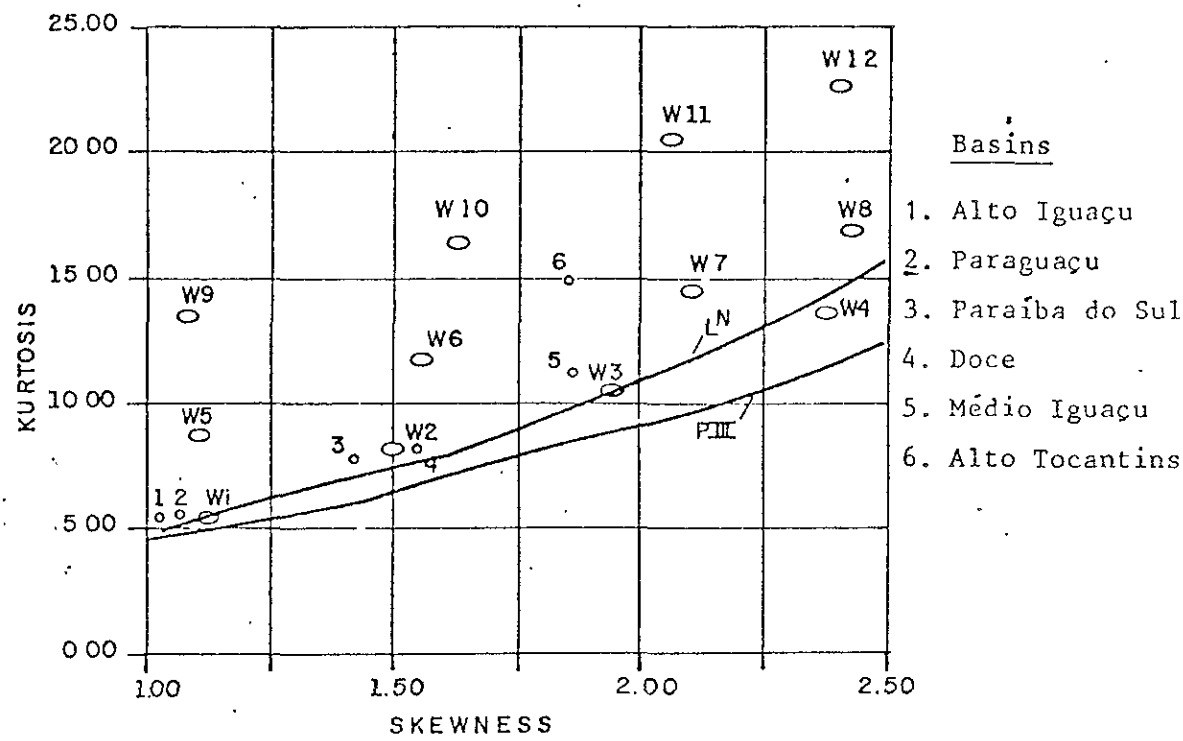
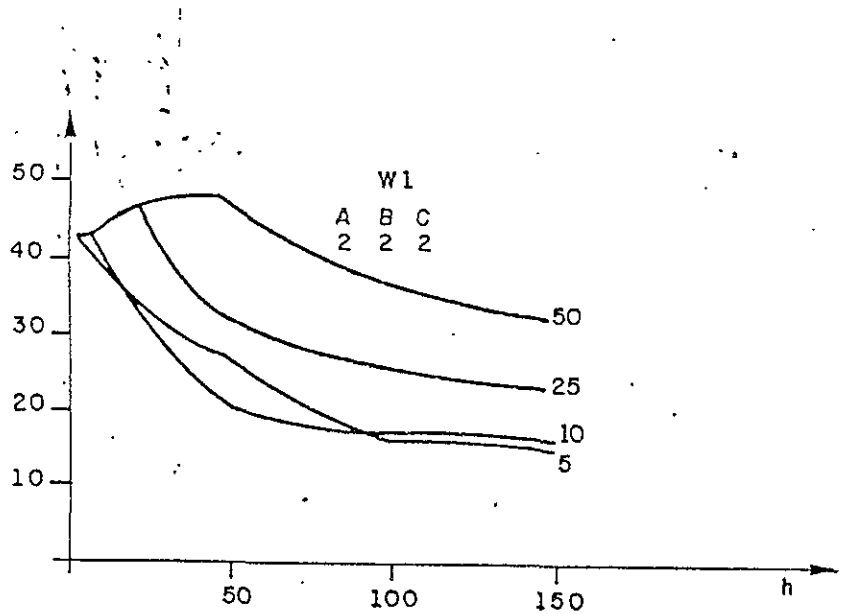
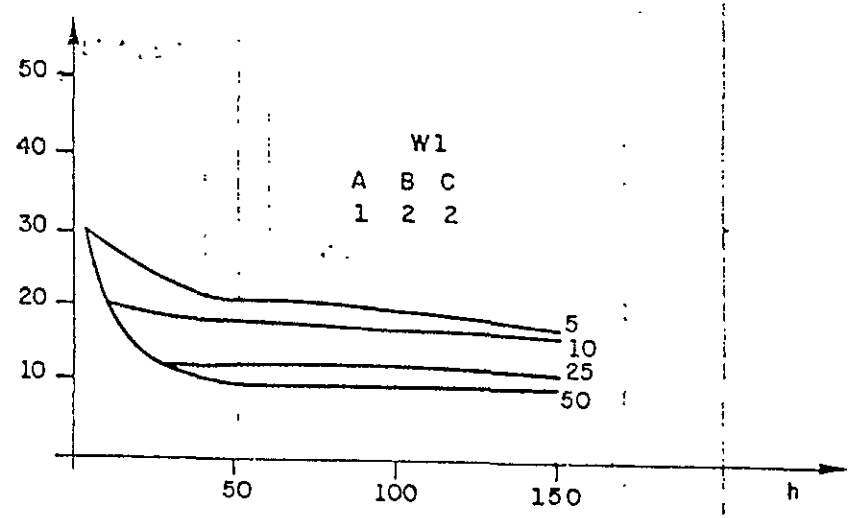
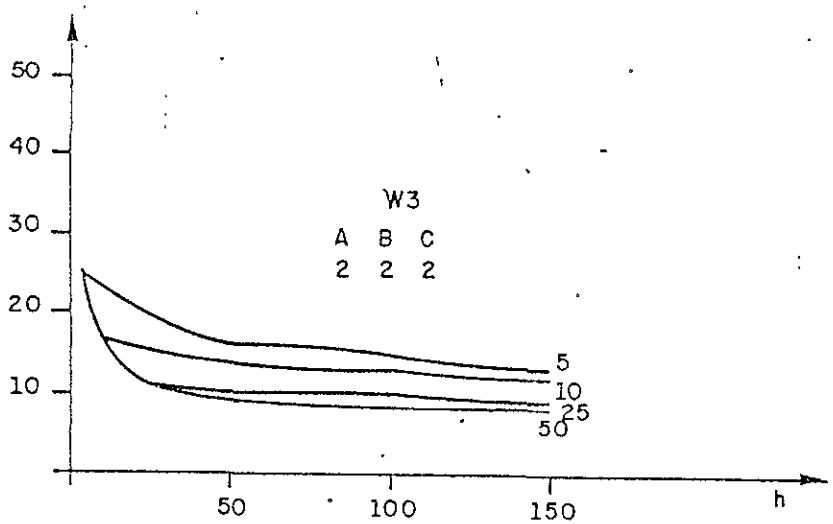


Figure 4. Skewness and Kurtosis of the Wakeby Distributions







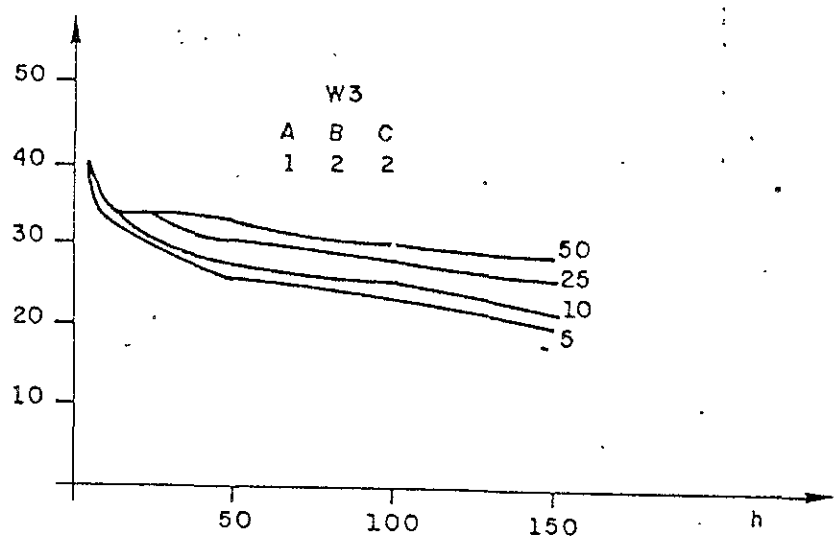


Figure 5. MAE for W-1 and W-3

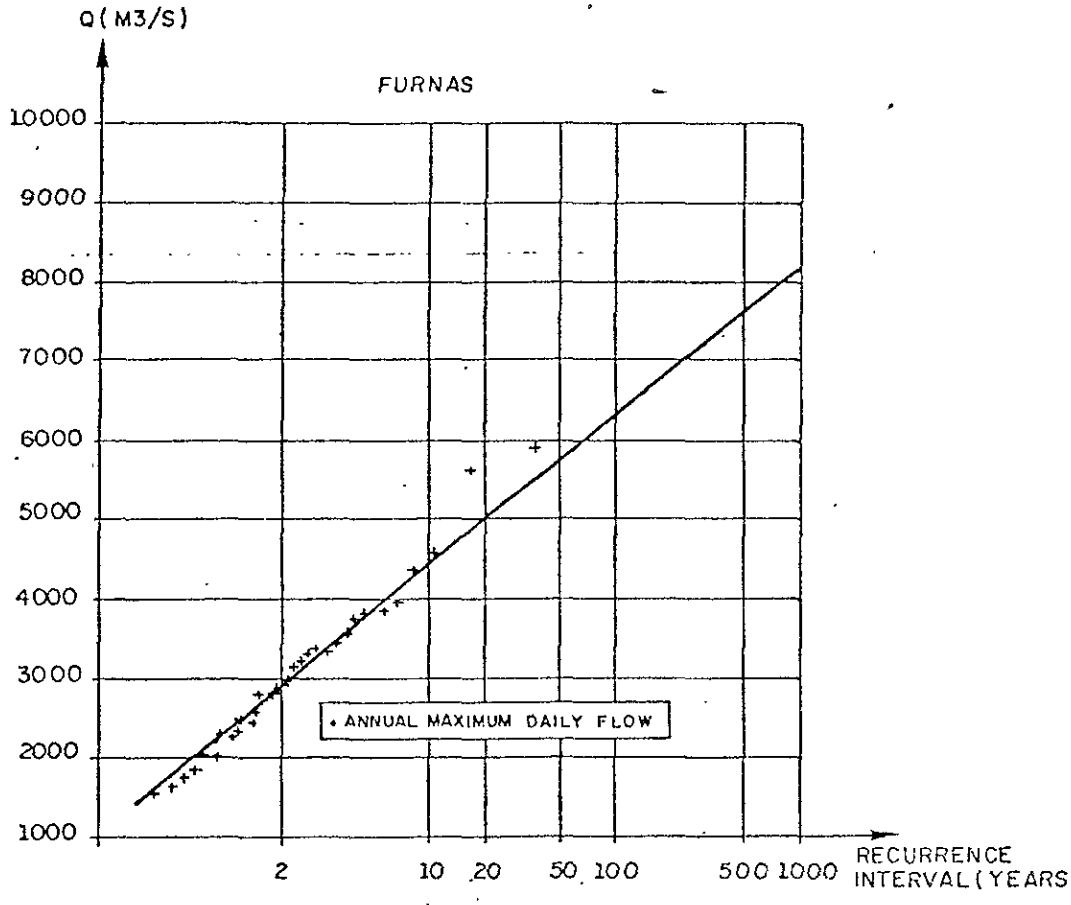


Figure 6. Annual Maximum Distribution