Archive

BIRTH OF A PARENT: THE WAKEBY DISTRIBUTION FOR MODELING FLOOD FLOWS

John C. Houghton

Working Paper No. MIT-EL77-033WP

October 14, 1977

Acknowledgements

This research was done in the Environmental Systems Program in the Division of Applied Science at Harvard University while the author was a Ph.D. candidate. Many of the ideas originated with Professor Harold A. Thomas, Jr. The author wishes to thank the U.S. Geological Survey which sponsored the majority of this research.

Introduction

Flood frequency analysis is a tool used in forecasting the frequencies of future floods. In general, a past record is fit with a statistical distribution function which is then used to make inferences about future flows. Many distributions and various ways of fitting them are already in use or have been proposed. Slack et al. (1975), Benson (1968), and others have attempted to choose an appropriate model for flood records from among the alternate traditional distributions and fitting procedures. In Slack et al. (1975), Monte Carlo techniques were used to generate synthetic flows from various background distributions. These samples were in turn fit with various assumed distributions. In their notation, the parent distributions were labeled "F-distributions" and the assumed distributions labeled "G-distributions". The search for a parent distribution constitutes a different problem from estimating the design event. This paper deals with the world of F-distributions; a following paper by this author (Houghton 1977b) is concerned with the G-world.

This paper introduces a new five-parameter distribution, which we have named the Wakeby, as a substitute for traditional F-distributions. We define the Wakeby distribution and show how it overcomes certain deficiencies associated with traditional distributions. In Houghton (1977b), a variant of the Wakeby distribution is tested using a new fitting procedure. In both papers we follow convention in assuming independent and identically-distributed observations from each sample; serial correlation and non-stationarity are assumed to be insignificant.

Rationale For A New Distribution

The Wakeby distribution has five parameters, a significant increase from the two or three in standard distributions. There must be good reason for introducing a new distribution, particularly if it absorbs more degrees of freedom than those distributions currently in use. The instability of higher moments and their functions, such as the coefficient of skew, is well known. They often add more noise than signal to estimation procedures for conventional distributions. Although the Wakeby distribution has five parameters, neither the higher sampling moments nor even the sample variance are used to estimate those parameters. Hydrologists and engineers in past years have occasionally felt the need to go beyond three parameters, but it was recognized that the use of higher moments than the third would introduce too much error into the estimation process. The estimation procedure developed for the Wakeby distribution circumvents this problem.

In traditional estimation procedures, the smallest observations can have a substantial effect on the right-hand side of the distribution. But the left-hand side does not necessarily add information to an estimate of a quantile on the right-hand side. Indeed, since floods are not known to follow any particular distribution, it seems intuitively better to divorce the left-hand side from the right. It will be shown that the Wakeby does exactly that. There is also some reason to believe that none of the standard distributions have the properties on their left-hand sides that may, in fact, reflect nature. If, in reality, the lowest observations follow the left-hand tail of a low-skew lognormal distribution, and the highest observations follow the right-hand tail of

a high-skew lognormal distribution, no conventional three-parameter distributions would model it accurately. They lack enough kurtosis for any given skew. Fitting a three-parameter curve to a five-parameter nature would distort the whole fit, including the higher quantiles. The so-called "separation effect" presented by Matalas et al. (1975) can be explained by this argument.

There is also the practical test of what the Wakeby distribution is able to do when used in other contexts. If a search for generic categories of floods in different regions of the nation is successful for Wakeby parents but not for others, then there is more reason for its adoption. Similarly, it has been difficult to find a regional skew. If, for example, there is more success in finding a regional d (d is the shape parameter of the right-hand tail), then the Wakeby has significant advantage over conventional three-parameter distributions. These two concepts are evaluated in Houghton (1977a).

Finally, given the correct choice of parameters, the Wakeby distribution can generate synthetic flows in the pattern of a lognormal distribution or any of the other conventional distributions. But the reverse is not true. There are shapes of the distribution function of a Wakeby that cannot be mimicked by any of the three-parameter distributions. Thus, not only can the Wakeby provide patterns of flow not possible with these other distributions, but it can also serve as an organizational construct. Each of the traditional distributions is a subset of the parameter space of the Wakeby. It is possible to fit a single distribution, the Wakeby, with many combinations of parameter values that are easily compared, rather than several distribution functions, each with a different analytical form. The Wakeby distribution is a grand parent.

The Properties of the Wakeby Distribution

The Wakeby distribution is most easily defined as an inverse distribution function:

$$x = -a(1-F)^{b} + c(1-F)^{-d} + e \quad , \tag{1}$$

where F is the uniform $(0,1)$ variate. The equation is written so that a, b, c, and d are always positive, and e is sometimes positive. The first moment about zero (mean) is

$$\mu_1'(x) = e - \left[\frac{a}{b+1} - \frac{c}{1-d}\right]. \tag{2}$$

The second moment about the mean (variance) is

$$\mu_2(x) = \frac{c^2}{1-2d} - \frac{2ac}{1+b-d} + \frac{a^2}{1+2b} - \left[\frac{c}{1-d} - \frac{a}{1+b}\right]^2. \tag{3}$$

The parameter e is a location parameter, and further moments about zero for the variate $\upsilon = x-e$ are:

$$\mu_3'(\upsilon) = \frac{c^3}{1-3d} - \frac{3c^2a}{1+b-2d} + \frac{3ca^2}{1+2b-d} - \frac{a^3}{1+3b} \tag{4}$$

and

$$\mu_4'(\upsilon) = \frac{c^4}{1-4d} - \frac{4c^3a}{1+b-3d} + \frac{6c^2a^2}{1+2b-2d} - \frac{4ca^3}{1+3b-d} + \frac{a^4}{1+4b}. \tag{5}$$

For $d > 1$ the mean is infinite; for $d > .5$ the variance is infinite, etc.

The Wakeby is similar to a five-parameter member of the Tukey family of lambdas (Joiner and Rosenblatt 1971). Given values of a and b that are typical of flood records, the $-a(1-F)^b$ term generally has no effect on x if F is above .25 . Thus the Wakeby can be thought of in two parts. The right-hand tail $c(1-F)^{-d}+e$, and the left-hand tail $-a(1-F)^b$, which is in effect an adjustment to the graph of $c(1-F)^{-d}+e$.

Distributions of the order statistics are easier to calculate with the Wakeby than with some other distributions. For the portion of the distribution which is not affected by the term involving a and b , the distribution of the k^{th} observation is

$$f\left(x_{(k)}\right) = \left[\frac{\Gamma(n+1)}{\Gamma(k)\Gamma\ (n-k+1)\ cd}\right]\left[1-\frac{\left(x_{(k)}-e\right)^{-\frac{1}{d}}}{c}\right]^{k-1} \cdot \left[\left(\frac{x_{(k)}-e}{c}\right)^{-\frac{1}{d}}\right]^{n-k+1+d} \cdot \qquad (6)$$

The distribution of ranks on the lower tail is not analytical, but percentiles are easily calculated by applying the Wakeby distribution as a transformation on the percentiles of a Beta distribution.

This apparently new Wakeby has roots in older models. One of the first distributions used to model floods was the Fuller formula: $x = a + b(T-1)^c$, where $T = \frac{1}{1-F}$, a formulation which is nearly identical to the right-hand tail of the Wakeby.

Fitting Selected Flood Records

The U. S. Geological Survey provided for us a tape of selected streamgaging records throughout the continental United States. About 1,400 high-quality stations out of a total of more than 10,000 in operation were selected for the tape; those selected had the longest records with minimal regulation and diversion. These 1,400 stations have variable numbers of years of records and are often discontinuous. In this research, the procedure for choosing n-year records from these 1,400 stations is the same as that used in Matalas et al. (1975). We selected for the majority of our research forty-six gaging stations which had been operating continuously for sixty years or more, truncating to sixty years those with longer records.

Initially, the forty-six records were fit with a three-parameter lognormal distribution using the method of moments. Most seemed by eye to fit adequately. However, some appeared to fit very poorly. This is substantiated analytically using the goodness-of-fit tests outlined below. One of these poorly-fitting records, #2, is used to illustrate the flexibility of the Wakeby distribution over the lognormal. The lognormal fit is displayed in Figure 1, and the Wakeby distribution applied to the same record is shown in Figure 2. The Wakeby fits much better. On the other hand, one might expect the reverse to be true also. That is, one might try to choose some of the forty-six floods that the lognormal fits well, but the Wakeby fits poorly. Actually, the Wakeby does a good job at duplicating the lognormal, but not conversely. At one point, we fit all forty-six floods with a four-parameter version of the Wakeby distribution. Nearly all forty-six appeared to fit adequately by eye, and all fit at least as well as the lognormal.

The Separation Effect

Matalas et al. (1975) presented a contradiction similar to that of the Hurst effect which they called the separation effect. They took each of the U.S.G.S. watershed regions in the United States and used all 30-year records from the master-file of 1,400 stations. For example, region #1 contained 178 such records. The coefficient of skew was calculated for each record. The standard deviation of the coefficients of skew was then plotted against the mean of the skews for that region. For 14 regions, the plot will contain fourteen such points, which are marked with an "X" in Figure 3.

An equivalent procedure can be applied to synthetic samples from a lognormal distribution. Sets of samples of 30 synthetic lognormal deviates are generated from a distribution with a particular skew. The coefficient of skewness is then calculated for each sample. Repeating this process several times for several background skews gives a frontier with averages and confidence interval as shown in Figure 3. This graph shows that for any given skew, the standard deviation is higher in nature than in traditional distributions. Matalas et al. (1975) included most of the commonly-used distributions, repeated the plots for 10-year and 20-year records, and found that none of the distributions could reproduce as high a standard deviation as that found in nature. This has been termed the "separation effect". Thus, nature has skews that are even more unstable than those generated by common distributions. Moreover the authors showed that this separation effect cannot be explained by small sample properties or by auto-correlation.

The Wakeby distribution was originally introduced to account for this effect. The three-parameter loglogistic, presented in Houghton (1977a),

shows more separation effect than the common distributions, but it could not mimic the separation effect noted in nature. What was needed was a distribution with a very thick right-hand tail and a left-hand tail thick enough to decrease average skews. This makes the middle part of the distribution function steeper than traditional skewed curves. The Wakeby distribution has this property. Original guesses at typical parameters of the distribution showed separation effects much larger than those found in nature. The set of parameters which make up a "Righteous Wakeby", as defined in Houghton (1977a), are in some sense typical ones for the data at hand. Figure 3 also shows the separation effect derived from that set of parameters. Ten sets of fifty replications each are plotted as a "1". They match extraordinarily well those found in nature. The separation effect can be duplicated by another means. Mixing lognormal parents of different skews produces a higher standard deviation for any average skew.

Goodness-Of-Fit Tests

Researchers in flow frequency analysis (see Matalas et al. 1973) recognize that it is difficult to apply conventional goodness-of-fit tests to flood records and discover meaningful results. Such tests do not seem to be powerful enough to distinguish among similar skewed distributions. If conventional tests could be used more effectively, it is likely that a common distribution would be agreed upon to model floods. Instead, there currently is a controversy over which distribution to use. However, the need for a more versatile distribution may be demonstrated by applying new goodness-of-fit tests that cast doubt on traditional distributions. We have chosen the lognormal distribution as the surro-

gate for traditional distributions. And where a fitting procedure needs to be identified, the method of moments is used. Our purpose is not to show that a majority of flood flow series cannot be modeled adequately with conventional distributions, but rather that a significant minority of records are fit poorly by the lognormal.

Shapiro and Wilk Test

A very effective goodness-of fit test of normality for composite hypotheses has been introduced by Shapiro and Wilk (1965). The test is sensitive to both thick- and thin-tailed distributions as well as to asymmetrical distributions. The test itself requires no assumption about the mean or standard deviation, but to transform a three-parameter log-normal to a two-parameter normal, one must specify the location parameter before taking logs. We applied the test conservatively by searching over c-space, the location parameter, to maximize the significance level of the test. The log-space observations were then tested for normality by the Shapiro and Wilk method. We did not adjust the significance level for this degree of freedom, which would suggest that the number of rejections are in fact much higher than those presented in these results. The coefficients for $n = 60$ were not available, and so all gaging stations with $n = 50$ years of record were used. Results are shown in Table 1 using these 188 stations:

TABLE 1.

	Rejections	
SL	observed	expected
5%	14	9
2%	9	4
1%	7	2

In spite of the conservative application, more records were rejected than the expected number under the null hypothesis. It seems obvious that there are a portion of records which are not adequately portrayed by the lognormal distribution.

Kolmogorov-Smirnov Test

Traditional Kolmogorov-Smirnov testing involves simple hypotheses in which the parameters of the distribution are calculated without the aid of the sample itself. However, the Kolmogorov-Smirnov test has been adapted to composite tests for normal distributions by Lilliefors (1967). By searching over the unknown parameter c to maximize the significance level, as in the Shapiro and Wilk test, very few of the forty-six records could reject the null hypothesis of lognormality, even at the 20% level. However, using a value of c estimated using the method of moments (and disregarding the estimated a and b), virtually all of the records were not lognormal at the 1% level. This is another indication that the lognormal assumption, using the method of moments, is suspect.

Smirnov Distance Test

A third method for testing lognormality is one suggested by synthetic hydrology and the two-sample Smirnov distance test. All sixty observations were fit by a lognormal distribution using method of moments, and random samples of sixty observations were drawn from that parent distribution. The Smirnov two-sample test was then run to determine whether the original sixty and the synthetic sixty came from the same underlying distribution. The application of the test in this manner is probably also conservative; it does not reject some samples which should be rejected. However, the

results should be qualified at the same time; rejected samples might perhaps have been accepted if estimation techniques other than the method of moments were used. But conversely, applying alternate techniques could result in the rejection of samples which are presently accepted using the method of moments. All forty-six records of $n = 60$ were tested with replications of either 20 or 50. If the null hypothesis is true, the significance level of the results should be uniform (0,1), except that only discrete values are possible. We found that five of the forty-six were grossly non-lognormal. All five floods had 60% or more of the replications significant at the 5% level. Another six had significance levels that could probably be shown non-uniform by another goodness-of-fit test. This indicates that if one were using a fitted lognormal distribution to generate synthetic traces from any of these eleven records, the synthetic records would be fundamentally different from the original. It is pertinent to note that the U.S. flood records rejected by the Shapiro and Wilk test and this Kolmogorov-Smirnov application are nearly disjoint.

Fitting Procedure

Our fitting procedure uses the technique of probability plotting routines, as outlined in Houghton (1977a). It takes advantage of the separation properties of the left- and right-hand tails of the distribution. Phase one operates on the right-hand tail, phase two on the left-hand tail.

To start, choose some F_c which is a cutoff point. The curve corresponding to $F > F_c$ is analyzed in phase one, and that corresponding to $F < F_c$ in phase two. For phase one,

$$x_k = -a(1-F_k)^b + c(1-F_k)^{-d} + e \quad , \tag{7}$$

or alternatively,

$$\log(x_k - e + a(1-F_k)^b) = \log(c) - d\,\log(1-F_k) \quad , \tag{8}$$

for all x_k, such that $F_k > F_c$. Set $a = 0$ and $b = 1$, and assume an initial value for e. Then one can use linear regression to estimate c and d. A search is then made over e to minimize the sum of squares of the vertical distance from each observation point to the regression line. Plotting positions are postulated for each observation; we used the median plotting position rather than the mean in order to reduce positive bias. Phase one gives estimates of c, d, and e.

In phase two, one assumes the values calculated in phase one for c, d, and e, and evaluates a and b by regression analysis applied to:

$$\log(-x_k + e + c(1-F_k)^{-d}) = \log(a) + b\,\log(1-F_k) \quad , \tag{9}$$

for all x_k such that $F_k < F_c$. Given new values of a and b, phase one is repeated, then phase two, etc. In practice, repetitions are usually unnecessary (i.e. the values of c, d, and e do not change with the updated values of a and b). In those cases where repetitions are needed, one repetition provides most of the change, and further repetitions tend to oscillate. Note that an F_c was assumed for the fitting procedure. In fact, the whole procedure is calculated for values of F_c for $0 < F_c < 1$. With $n = 60$, the cutoff point has been varied over

each 5^{th} sample between 5 and 55. The criterion by which we choose F_c and its associated parameters $\hat{a}$, $\hat{b}$, $\hat{c}$, $\hat{d}$, and $\hat{e}$, is a weighted sum of squares. This weighted sum of squares is a weighted sum of the ρ^2 values calculated in phase one and phase two. They are weighted by the proportion of observations in each phase. The F_c finally chosen is that one which maximizes the weighted ρ^2. For a high F_c, phase one would often result in some calculations involving the logs of negative numbers. In that case, the particular cutoff point and all others above it are excluded from further consideration. It seems likely that more elementary versions of the fitting procedure could be adopted with assumptions on F_c and repetitions of phase one and phase two. Other modifications are discussed in Houghton (1977a).

Conclusions

The Wakeby distribution has been shown to fit a set of U.S. flood records of high quality better than the lognormal distribution according to several goodness-of-fit tests. Furthermore, the Wakeby was able to "explain" the separation effect not evident in traditional distributions. A further use of the Wakeby distribution is presented in Houghton (1977a), in which a "handbook" set of Wakeby distributions are fit to various flood categories so that parameters are predetermined rather than estimated from the sample. In Houghton (1977b), the Wakeby is used to generate synthetic flows for Monte Carlo experiments. In that research study, the Wakeby distribution was employed both as a parent distribution and as a model in fitting the synthetic records.

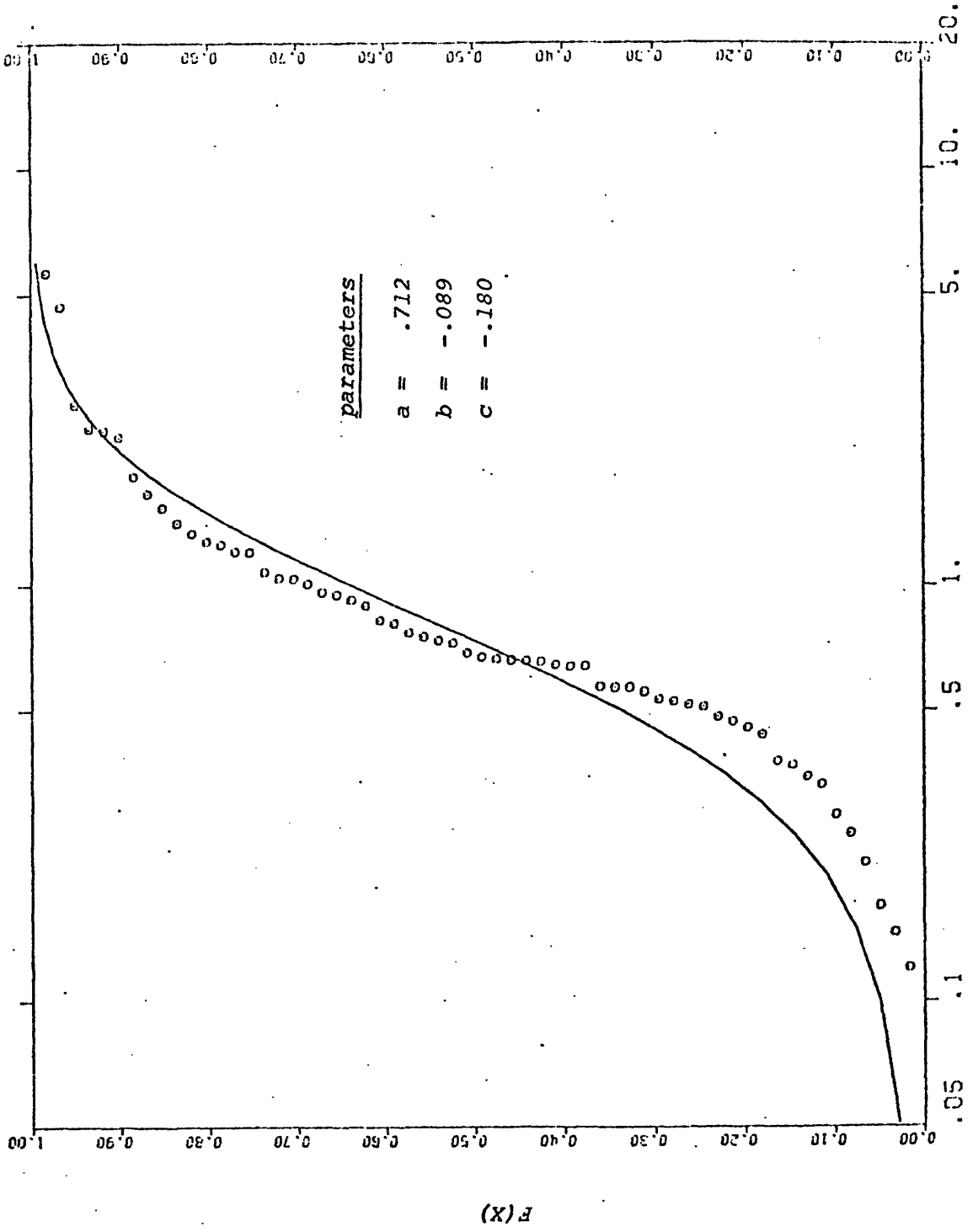


FIGURE 1. THREE-PARAMETER LOGNORMAL FIT TO FLOOD RECORD #2 (1180500)

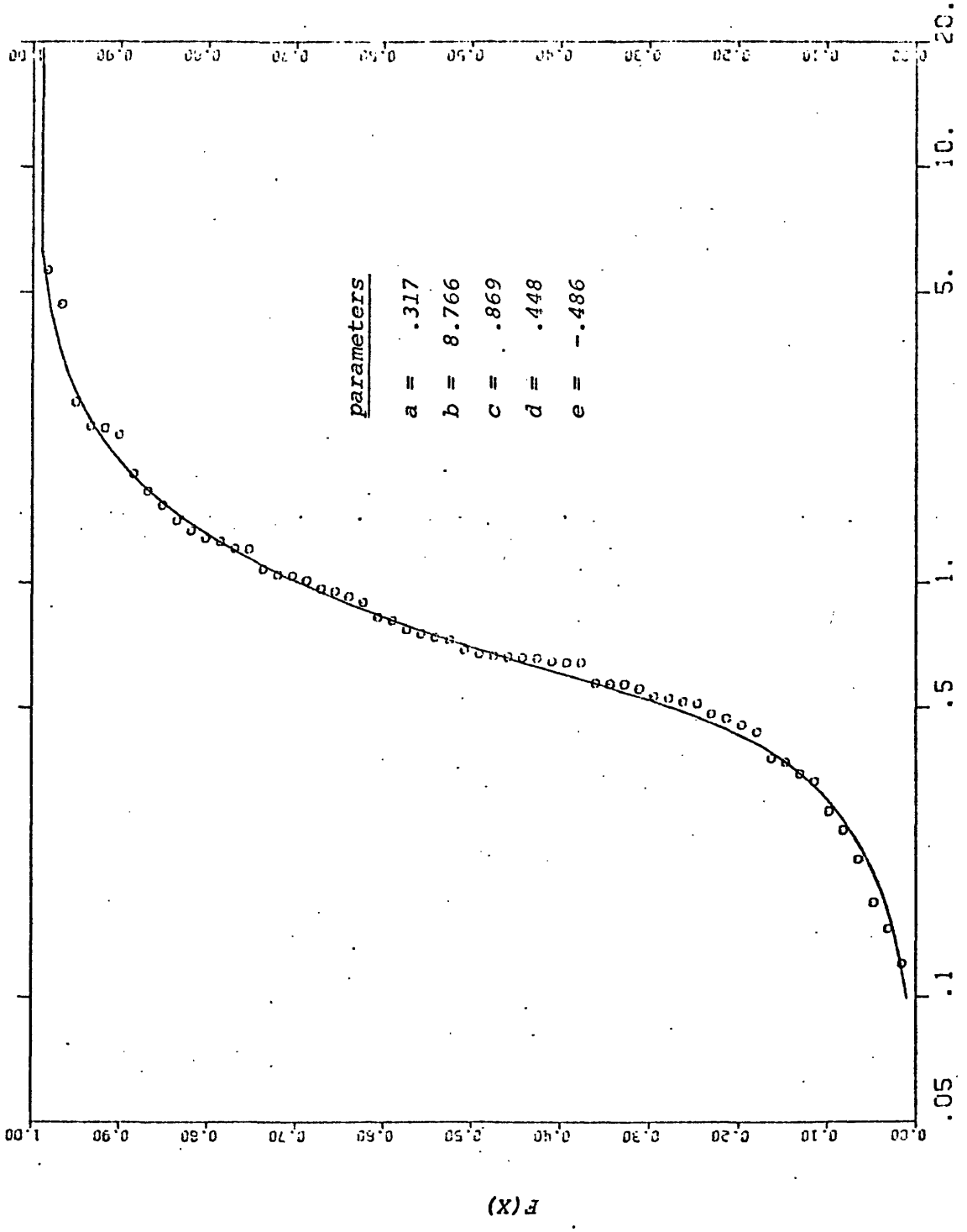
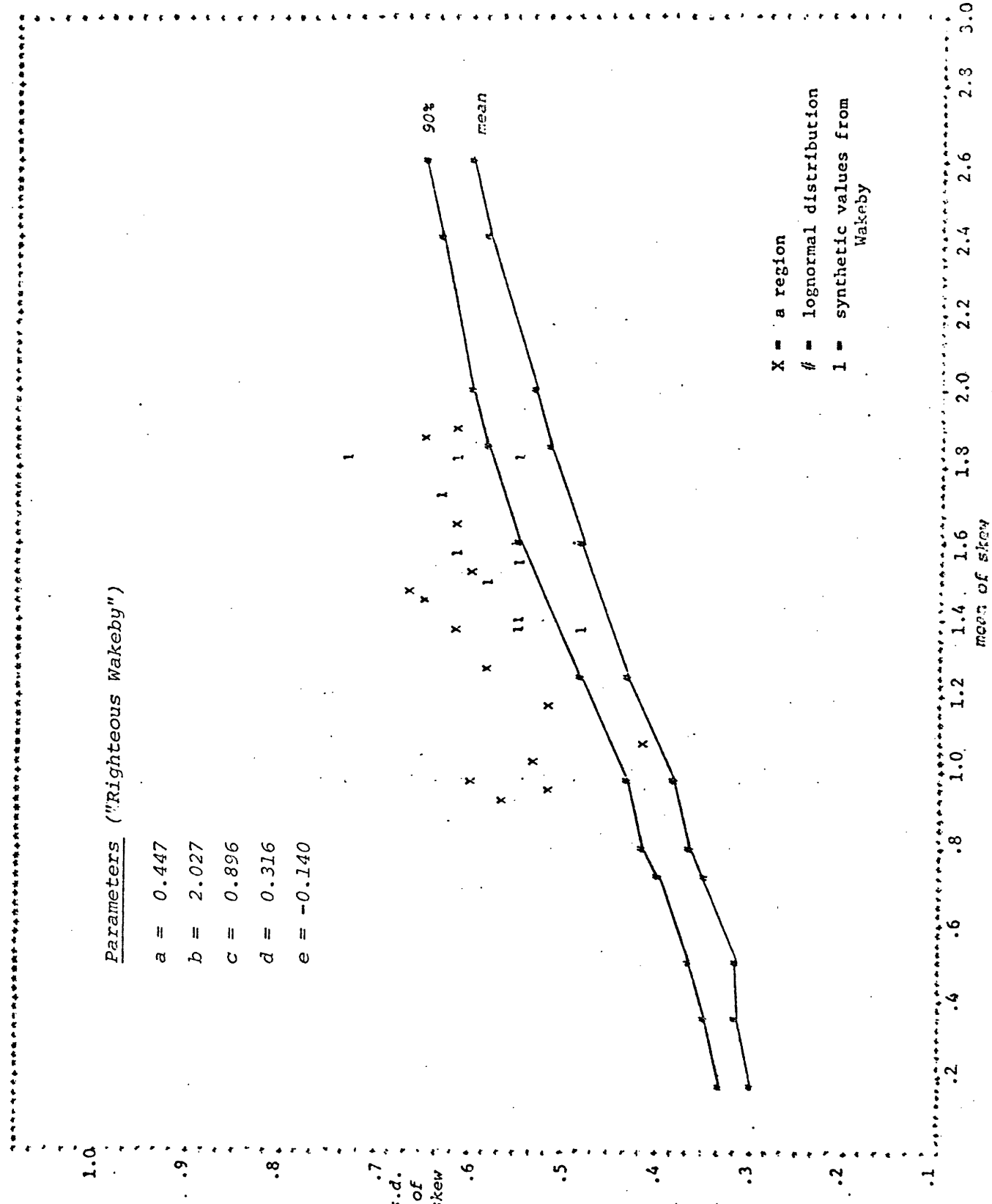


FIGURE 2. WAKEBY DISTRIBUTION FIT TO FLOOD RECORD #2 (1180500)

FIGURE 3. SKEW EFFECT FROM "RIGHTEOUS WAKEBY"



Bibliography

Benson, Manuel A., "Uniform flood-frequency estimating methods for Federal agencies", Water Resources Research, vol. 4, no. 5, pp. 891-908, October 1968.

Houghton, John C., "Robust Estimation of the Frequency of Extreme Events in A Flood Frequency Context", Unpublished Ph.D. Dissertation, Division of Engineering and Applied Physics, Harvard University, 1977a.

Houghton, John C., "The Incomplete Means Estimation Procedure Applied To Flood Frequency Analysis", 1977b.

Joiner, Brian L., and J. R. Rosenblatt, "Some properties of the range in samples from Tukey's symmetric lambda distributions", J. Amer. Statist. Assoc., vol. 66, pp. 394-399, June 1971.

Lilliefors, Hubert W., "On the Kolmogorov-Smirnov test for normality with mean and variance unknown", J. Amer. Statist. Assoc., vol. 62, pp. 399-402, June 1967.

Matalas, N. C., and J. R. Wallis, "Eureka! It Fits a Pearson Type 3 Distribution", Water Resources Research, vol. 9, no. 2, pp. 281-289, April 1973.

Matalas, N. C., J. R. Slack, and J. R. Wallis, "Regional Skew In Search of a Parent", Water Resources Research, vol. 11, no. 6, pp. 815-826, December 1975.

Shapiro, S. S., and Wilk, M. B., "An analysis of variance test for normality (complete samples)", Biometrika, vol. 52, nos. 3 & 4, pp. 591-611, 1965.

Slack, J. R., J. R. Wallis, and N. C. Matalas, "On the value of information to flood frequency analysis", Water Resources Research, vol. 11, no. 5, pp. 629-647, 1975.