REPRESENTATIONS OF RAINFALL AND RUNOFF BY THE DESCENDING EXPONENTIAL

by
Donald A. Parsons
Hydraulic Engineer, Oxford, Mississippi

INTRODUCTION

Many applications to rainfall and runoff data have shown the general satisfactoriness of the descending Exponential Function to represent the distributions of magnitude of these things. Specifically included are point rainfall amounts in given time periods up to a week or more, depending upon the frequency of occurrence, point rainfall amounts in storms, peak flood flows, flood volumes, and runoff amounts in given time periods. Schafmayer and Grant (1) used the Exponential Function for Chicago rainfall.

Accepting this function as representative of the distribution in magnitude of single events, other functions can be derived therefrom that represent the largest event in groups of successive events from the exponentially distributed values. Also, still other functions can be derived that express the distributions of the means and sums of the magnitudes of groups of successive events from a population with an exponential distribution.

FAMILIES OF DISTRIBUTION FUNCTIONS

Two related families of distribution functions are found that closely represent rainfall and runoff magnitudes. The one function common to both families is the basic distribution function for single events. It may be represented by the cumulative distribution function.

$$q = 1 - p = 1 = e^{-cx} \qquad (1)$$

where x is the magnitude of an event, p is the probability that x will be exceeded in a single trial, q is the probability that x will not be exceeded in a single trial, c is a constant, and e = 2.718. A list of symbols with definitions accompanies the text.

This cumulative distribution function of single events represents the population of larger values of rainfall or runoff and neglects many events that may be classed as belonging to the group of small things.

A family of distribution functions

$$Q_m = q^m = (1-p)^m = (1-e^{-cX})^m \qquad (2)$$

gives the probability that the largest of m random values of x will not exceed X. Note the lower case x is a value of any one of the single events, whereas capital X is the value of the largest x in the group of m single events. There is a distribution function for capital X's for each value of m.

When the origin of coordinates is changed from X = 0 to the modal value of X by introducing into equation (2) the notation

$$\epsilon = X - \tilde{X} \qquad (3)$$

where ϵ (epsilon) is the difference between a value of X and the modal value of X, the function (2) becomes

$$Q_m = [1 - \frac{e^{-c\epsilon}}{m}]^m \qquad (4)$$

The limiting form of this distribution as m approaches an infinite number is the skewed function

$$Q_m = e^{-e^{-c\epsilon}} \tag{5}$$

which is the so-called extreme value function, derived by Fisher and Tippett (2), and actively sponsored in recent years by Gumbel (3) as a function for use in the description of the distribution of hydrologic events, especially annual peak flood flows. Actually, it is found to be just one of a family of functions which should yield to some other one in the family for values of m below about 20 to 30 for best representation of the data. Most runoff data have mean annual $\bar{m}$ values smaller than this.

The derivation of (5) from (4) is simply done as follows: As m becomes very large, i.e., as $m \longrightarrow \infty$, equation (4) becomes indeterminate. Differentiate to aid in evaluating the equation at the limit, letting Q_m, ϵ and m be variable:

$$\frac{dQ_m}{dm} = Q_m^{\frac{m-1}{m}} e^{-c\epsilon} \left[\frac{1}{m} + \frac{d\ (c\epsilon)}{dm} \right] \tag{6}$$

If $m \longrightarrow \infty$

$$\frac{dQ_m}{dm} = Q_m e^{-c\epsilon} \frac{d(c\epsilon)}{dm} \tag{7}$$

and

$$\frac{dQ_m}{Q_m} = e^{-c\epsilon}\, d(c\epsilon) \tag{8}$$

integrating $\ln Q_m = -e^{-c\epsilon}$ (constant of integration = 0) (9)

and

$$Q_m = e^{-e^{-c\epsilon}} \tag{10}$$

Formulas (2) and (4) are cumulative distribution functions for the largest items from groups of items from the Exponential Distribution, formula (1). They are the same except for the origin of measurement. Formula (2) is more meaningful because it involves actual magnitudes X. Henceforth, this family of distribution functions will be referred to as the "Distribution of the Largest."

Other statements of the distribution of X, the largest in groups of m drawn from an exponentially distributed population are:

$$cX = -\ \ln[1-Q_m^{1/m}] \tag{11}$$

or

$$\frac{X}{\bar{X}} = \frac{cX}{\overline{cX}} = \frac{-\ln[1-Q_m^{1/m}]}{\sum_{1}^{m} 1/k} = f_1\ (m,\ Q_m) \tag{12}$$

or

$$\frac{X}{\bar{X}} = -\ cv_X\ \frac{\ln[1-Q_m^{1/m}]}{(\sum_{1}^{m} 1/k^2)^{1/2}} = cv_X\ f_2\ (m,\ Q_m) \tag{13}$$

Equation (13) is in the form used by Brakensiek (20) and Zingg.

The other family of distribution functions is

$$Q_n = \frac{1}{\Gamma(n)} \int_o^{cn\bar{x}} (cn\bar{x})^{n-1} e^{-cn\bar{x}} d(cn\bar{x}) \qquad (14)$$

which is the Incomplete Gamma Function, giving the probability that the means or sums of n random values of x from the population represented by Eq. (1) will not exceed $\bar{x}$ or $n\bar{x}$, respectively. There is a distinct distribution function for each value of n. The limiting form of (14) as n becomes increasingly large is the symmetrical normal distribution. Formulas (2), (4), (5), and (14) are deduced from formula (1) by strict mathematical and probability principles (4). They are, therefore, intimately related and consistent with one another.

The development and use of the functions (1), (2), (4), and (14) is with the assumption that effects of the items in the group of small things are negligible. There are situations, especially in arid areas, when this is not so; and studies might well be made to find ways of representation of these smaller things. Also assumed to be negligible in effect is the time variation between events. Whereas the analysis in the case of equations (2), (4) and (14) assumes a fixed number of events, m or n, in each time period, this is not actually the case. Nevertheless, all of them have been found helpful in the representation of hydrologic data.

Table 1 summarizes the three distribution functions: I, for single events; II, for the largest in groups of m single events; III, for the means, $\bar{x}$, and sums, $\overline{nx}$, of groups of n single events.

Type I is used to represent the distributions of the larger values of individual events, including; (1) point rainfall amounts in given time periods up to a week or more; (2) point rainfall amounts in storms; (3) peak flood flows; (4) flood volumes; (5) runoff amounts in given time periods up to a week or more.

Type II is used to represent the distributions of the largest items in groups of events of the five kinds listed for Type I. This includes: (1) annual or biennial maxima; (2) monthly maxima, provided the mean number of single events per month exceeds three or four, and; (3) the maxima for any number of events $m = \bar{m}Y$ in Y units of time, $\bar{m}$ being the average number of occurrences per unit of time. Thus the probability of any given value being greater or less than the maximum in Y years is expressed.

Type III is used to represent the distributions of sums of events such as: (1) annual or biennial rainfall amounts; (2) monthly rainfall amounts, provided the mean number per month is larger than 3 or 4; (3) annual runoff amounts.

The constant $c = 2.3/k$ is dependent in magnitude upon the unit of measure for the variable and upon all things affecting the value of x other than for the purely random variation within the limits described by the functions. The c for the type II functions is the same as for type I, the basic distribution, but the c in type III expressions differs in application because of the implicit inclusion of items from the group of small things in the totals.

The parameters, c and m or n, have physical significance related to the environment of the events under consideration in addition to their control

Table 1. Related Distribution Functions for Rainfall and Runoff
A

	Cumulative Distribution Function	Frequency Function
I	$q = 1-p = 1-e^{-cx}$	$y = \frac{dq}{dx} = ce^{-cx}$
II (1)	$Q_m = \left[1-e^{-cX}\right]^m$	$y = cm\left[1-e^{-cX}\right]^{m-1}e^{-cX}$
(2)	$Q_m = \left[1-\frac{e^{-c\epsilon}}{m}\right]^m$	$y = ce^{-c\epsilon}\left[1-\frac{e^{-c\epsilon}}{m}\right]^{m-1}$
(3)	$Q_m = \lim_{m\to\infty}\left[1-\frac{e^{-c\epsilon}}{m}\right]^m$	
	$= e^{-e^{-c\epsilon}}$ (Extreme Value Function)	$y = ce^{-\left[e^{-c\epsilon}+c\epsilon\right]}$
III	$Q_n = \frac{1}{\Gamma(n)}\int_0^{c\bar{n}\bar{x}} (c\bar{n}\bar{x})^{n-1}e^{-c\bar{n}\bar{x}}\,d(c\bar{n}\bar{x})$ (Incomplete Gamma Function)	$y = \frac{1}{\Gamma(n)}(c\bar{n}\bar{x})^{n-1}e^{-c\bar{n}\bar{x}}$

Table 1.
B

	Variable	Mode	Median	Arithmetic Mean*	Variance*
I	$cx = -\ln(1-q)$	$c\tilde{x} = 0$	$cx = \ln 2 = 0.693$	$c\bar{x} = 1$	$\chi_{cx} = 1$
	$= -\ln p$	$q = 0$	$q = 1/2$	$q = 1-\frac{1}{e} = 0.632$	
	$x = -k \log p$				
II (1)	$cX = -\ln\left[1-Q_m^{1/m}\right]$	$c\tilde{X} = \ln m$	$cX = -\ln\left[1-(\frac{1}{2})^{1/m}\right]$	$\bar{cX} = \sum_{k=1}^{m}\frac{1}{k}$	$\chi_{cX} = \sum_{k=1}^{m} 1/k^2$
	$X = -k \log\left[1-Q_m^{1/m}\right]$	$\tilde{x} = k \log m$			
		$Q_m = (\frac{m-1}{m})^m$	$Q_m = 1/2$		
(2)	$c\epsilon = c(X-\tilde{X})$	$\tilde{c\epsilon} = 0$	$c\epsilon = -\ln m\left[1-(\frac{1}{2})^{1/m}\right]$	$\bar{c\epsilon} = \sum_{k=1}^{m}\frac{1}{k} - \ln m$	$\chi_{c\epsilon} = \sum_{k=1}^{m}\frac{1}{k^2}$
	$= -\ln m\left[1-Q_m^{1/m}\right]$				
	$\epsilon = -k \log m\left[1-Q_m^{1/m}\right]$	$Q_m = (\frac{m-1}{m})^m$	$Q_m = 1/2$		
(3)	$c\epsilon = -\ln(-\ln Q_m)$	$\tilde{c\epsilon} = 0$	$c\epsilon = -\ln\ln 2 = 0.3665$	$\bar{c\epsilon} = 0.5772$	$\chi_{c\epsilon} = \frac{\pi^2}{6} = 1.645$
		$Q_m = \frac{1}{e} = 0.3679$	$Q_m = 1/2$	$Q_m = 0.5704$	
III	$c\bar{n}\bar{x}$	$n-1$	---	n	n
	$\bar{n}\bar{x}$	$\frac{n-1}{c}$	---	$\frac{n}{c}$	$\frac{n}{c^2}$
	$c\bar{x}$	$\frac{n-1}{n}$	---	1	$\frac{1}{n}$
	$\bar{x}$	$\frac{n-1}{cn}$	---	$1/c$	$\frac{1}{c^2 n}$

* the k in summations is different than the coefficient k.

of the shapes of the distribution curves. Skewness increases with decreasing m and n values. All of the parameters have geographical variations in magnitude, a knowledge of which would be helpful in deriving distribution functions in cases of scanty data and in removing some of the inherent variations in the data in studies of watershed factors affecting runoff, topographical features affecting rainfall, and other things. Only fragmentary studies of the geographical variations of the parameters have been made, but it is already known that $\bar{m}$ and $\bar{n}$ values, based upon the year, vary from over 50 to under 1, indicating a very large variation in skewness of the distribution functions. The mean depths of rainfall in storms and runoff in floods also vary considerably with location on the continent.

LIMITATIONS OF THE BASIC DISTRIBUTION

Confidence in the applicability of these functions for the representation of rainfall and runoff magnitudes has come and will continue to grow as favorable use experiences increase. For example, the satisfactory results of Barger and Thom (5) through use of the Gamma distribution for the representation of monthly rainfall amounts should increase confidence in the applicability of the basic Exponential Distribution from which the Gamma distribution has been derived. Other applications of the Gamma distribution to sums of events (rainstorms) are Friedman (6), Friedman (7) and Janes, and Thom (8).

The widespread acceptance of the extreme value function should tend to promote confidence in the family of type II functions of which the extreme value function is a member. On the other hand, there may be reason to have reservations about the complete acceptability of equation (1), the basic distribution of single events.

Equation (1) indicates that, as time passes and more and more events are observed, the magnitude of the largest observed event increases without limit. This violates our knowledge of the nature of things. It is unreasonable to believe, for example, that the depths of rainfall could exceed the diameter of the earth. Equation (1) gives values that are too large for those rarest of events whose probabilities of occurrence approach the infinitesimal. Yet, experience teaches that this fault of the basic distribution function is outside the range of its application. An attempt to show this has been made (4) by means of a study of the aerial distribution of extreme storm rainfall and a consideration of the observational potential during United States history.

FEATURES OF THE BASIC DISTRIBUTION

The cumulative distribution function for single events, $q = 1-e^{-cx}$ represents only the group of large things of the kind under consideration. The group includes all of the large items and some of the smaller ones. It is often found that about one-half of the items represented by the distribution function, the one-half that are small, are mixed up with the remaining small items of the kind. This makes it necessary to work with partial samples when analyzing the data. This, of course, is not to say that items of small magnitude are unimportant. They are unimportant only in the studies of large things.

Even though a function or functions could be found that would better represent all items of the kind, both large and small, the analysis of partial

samples would undoubtedly still be necessary. Both the number of recorded events and their measured magnitudes, in the range approaching zero value, are affected by a host of things that are extraneous to the events of larger magnitude. Many of them, an unknown proportion, are really then of a different kind and would need to be excluded from the analysis.

The frequency curve (sometimes called the probability density, or frequency function) is extremely skewed. The high point at $y = c$ is at $x = 0$. It is the dashed curve of Figure 1. The ordinate y/c and the abscissa cX in figure 1 are dimensionless; and are used to simplify the graphical representation of the distribution functions.

The arithmetic mean of values of cx, greater than $cx = a$, is

$$c\bar{x} = a + 1 \tag{15}$$

The arithmetic mean value of the population of x values is $1/c = k/2.3$, and the median or middle value is 0.693 times this, or 0.301 k. Given a sample from the population, estimates of the average value may be obtained by using the differences in successive values of x when arranged in order of magnitude, beginning with the largest. Thus successive estimates of the arithmetic mean value of the population are:

$$(x_1-x_2),\ 2(x_2-x_3),\ 3(x_3-x_4)\ \ldots\ldots\ (r-1)(x_{r-1}-x_r) \tag{16}$$

and the mean value of these estimates is

$$\frac{\sum_{r-1} x}{r-1} - x_r \tag{17}$$

or in words, an estimate of the arithmetic mean value of the population, using the r highest values, is the difference between the mean of values larger than the rth value and the value of the rth item in the array. Then the lower limit of values for tabulation and use in data analysis would be about seventy percent of this computed mean for the population. This estimate should increase in trustworthiness with increasing r, provided the values used are from the distribution of large things.

The arithmetic mean size of the largest item in relation to the sum of the magnitudes of all items in a sample of m is

$$\text{proportion of total amount} = (1/m)(1 + \ln m) \tag{18}$$

and, the average contribution of the proportion, p, of the larger events of the sample, to the sum of all m events, expressed as proportion of total amount is

$$\text{proportion of total amount} = p(1 - \ln p) \tag{19}$$

These things become of interest in applications to storm rainfall or flood water depths or volumes.

Examples of the average contribution of the one largest event to the sum of the magnitudes of all m events are as follows:

m	Contribution	m	Contribution
No. of Events	%	No. of Events	%
2	85	20	20
3	70	50	9.8
5	52	100	5.6
10	33	500	1.4

DATA ANALYSIS: GIVEN THE MAGNITUDES OF ALL LARGE EVENTS

The graphical analysis of successive recorded rainfall or runoff events is begun by ranking the data according to magnitude, beginning with the largest. The ranked data are plotted on semilogarithmic paper, with magnitude as ordinate along the rectangular scale and logarithm of rank as abscissa. The data are plotted at 0.3 less than the rank as previously (4) (9) determined to be proper. Thus the largest is plotted at log 0.7, the next to the largest at log 1.7, etc. The straight line drawn to represent the larger values of the array may be expressed.

$$x = x_1 - k \log r = k \log R - k \log r \quad (20)$$

$$= -k \log (r/R) = -k \log p \quad (21)$$

$$= -(1/c) \ln p \quad (22)$$

where x is the value exceeded by r events. R may be thought of as sample size. It is equal to the extrapolated value of rank at x = o, and p is the proportion of the R values larger than x. Also, the value of R is the number of events that occurred during the period of observation that were in the class of large things, as determined by application of the exponential function. The mean number of events per unit time, $\bar{m}$, is obtained by dividing R by the number of time units in the observation period.

An illustration of the procedure is shown in Figure 2 where the values for 15-minute intense rainfall in storms for 31 years of observation at Washington, D. C., as given by David Blumenstock (19), are used after making small corrections for likely instrument errors and the use of arbitrary clock intervals. The slope of the line is found to be k = 0.51. The number of storms of the high intensity class was R = 1180. Then the mean number of events per year was $\bar{m}$ = 38.

This sort of data treatment, using the magnitudes of all the larger events that occur, yields the multiplying factor, k, and the mean number of events per unit time, $\bar{m}$, for use in the distribution function of the largest to obtain numerical estimates of the largest expected event for given situations.

FEATURES OF THE DISTRIBUTION OF THE LARGEST

The family of distribution functions of the largest in samples from the exponential distribution has utility in two ways. It is useful in the analysis of published data that are the largest values in a given unit of time such as in each year. The analysis yields the value of $\bar{m}$ and also the value of k (or c) for the basic equation.

Having determined these parameters by some means, this distribution function then gives for any particular situation the probability that an event will not exceed any chosen magnitude within a chosen time period. Or conversely, it gives the magnitude for a chosen time period with a specified probability of not being exceeded. The complement, $(1-Q_m)$, of this probability is the probability of exceeding the magnitude one or more times. These things are fundamental to, and inherent in, the process of establishing design discharges for works of improvement involving stream flow.

As the values of $m = \bar{m}Y$ are increased, the resulting frequency curves

become less skewed. However, there remains a considerable skewness at the largest value of m. The trend is shown in Figure 1. The limiting distribution of this series, as m increases toward an infinite value is located far to the right of Figure 1 extended, and is the so-called extreme value function.

A general picture of the relationships between the three central measures of the variable is given in Figure 3. The dashed line represents the distribution of single events. The curve with the hump represents the distribution of the highest in groups of 20 events. It is, for example, the distribution curve for maximum annual flood peaks, if there is an average of 20 floods per year of the class represented by the dashed curve. Or, it is the representation of 10 year flood peaks, if there is an average of only 2 floods per year.

The pictorial representations of size distributions of Figure 1 and 3 clearly convey the fact that the largest value in a group of m values has a very wide range in magnitude. Although it is customary and simpler to characterize a given variable by some central value, the arithmetic mean, the median, or the mode; a far more flexible and meaningful procedure is to think in terms of the probabilities that given values will or will not be exceeded. Where necessary or helpful to choose a central measure of the variable in this study the probable value or median has been preferred. Many representations of rainfall or runoff, especially for rainfall, are modal values. This is true when the so-called *"California method" of analysis is used. It is true in the case of Yarnell's (11) and the Miami Conservancy District's (12) representation of rainfall.

Use of the modal value of a distribution as a representation of a continuous variable is quite justified on the basis of ease of computation. There is, however, very little in the way of mathematical or probability concepts in its support. There is no "most probable" value of a continuous variable. The absurdity of universally adopting the modal value is illustrated by application to the basic distribution of single events. Here the modal value is zero.

Also avoided in this study is the use of the needless concepts of recurrence interval, exceedance interval and return period. Needlessness, variations in definitions, connotations of exactness that does not exist, and the power to confuse, all argue against their use.

DATA ANALYSIS: GIVEN THE LARGEST IN A PERIOD OF TIME

Two methods of data analysis are available, the method of moments and the graphical. The graphical method is especially preferred for the analysis of peak flood flow data. Table 2 is given as an aid to the graphical representation and analysis.

The (psi) ψ values of Table 2 are a function of m, the number of things from which the largest is selected, and Q_m, the probability that the value of the largest will not be exceeded by the largest in a random set of m.

---

* See Linsley (10) et al, p. 547.

Table 2. - Plotting Scales, Ψ, for Samples From the Population Having the Cumulative Distribution Function, $Q_m = [1 - e^{-cX}]^m$

Q_m	m=1	m=3/2	m=2	m=3	m=5	m=10	m=20	m=100	m=∞
.00	.70	.57	.47	.31	.11	-.17	-.47	-1.16	--
.01	.70	.59	.51	.42	.33	.26	.22	.19	.18
.05	.72	.63	.58	.51	.46	.41	.39	.37	.36
.10	.75	.67	.63	.59	.55	.51	.50	.48	.48
.20	.80	.75	.72	.70	.67	.65	.64	.64	.63
.30	.85	.83	.81	.80	.78	.77	.77	.76	.76
.40	.92	.91	.90	.89	.89	.88	.88	.88	.88
.50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.60	1.10	1.11	1.11	1.12	1.13	1.13	1.13	1.13	1.13
.70	1.22	1.24	1.25	1.27	1.28	1.28	1.28	1.29	1.29
.80	1.40	1.43	1.44	1.46	1.47	1.48	1.49	1.49	1.49
.90	1.70	1.74	1.76	1.78	1.79	1.80	1.81	1.82	1.82
.95	2.00	2.04	2.06	2.08	2.10	2.11	2.13	2.13	2.13
.99	2.70	2.74	2.76	2.79	2.81	2.83	2.83	2.84	2.84
mode:									
Ψ	.699	.744	.768	.791	.812	.826	.833	.840	.841
Q_m	.000	.193	.250	.296	.328	.349	.359	.366	.368
arithmetic mean:									
Ψ	1.133	1.123	1.118	1.110	1.104	1.098	1.094	1.093	1.092
Q_m	.632	.613	.604	.593	.584	.577	.574	.571	.570

They are computed * to give a straight line plotting of the data of slope k when the X values corresponding to the table values of Q_m are plotted along the ordinate at the corresponding Ψ values along the abscissa.

$$\Psi = 1 + \frac{X-X_{50}}{k} = 1 + \log \frac{1-(1/2)^{1/m}}{1-Q_m^{1/m}} \qquad (23)$$

and

$$X = k(\Psi-1) + X_{50} \qquad (24)$$

where (X_{50} = median value)

Analysis by the method of moments involves computation of the coefficient of variation, and entering Table 3 of reference (4) with $f(m) = (cv)^2$ to find the estimated m value for the population. The value of c is then found by dividing ν_1, the given arithmetic mean value of cX, by the arithmetic mean value of X. Space limitations preclude the inclusion of Table 3. Also, the graphical method of analysis is preferred.

The graphical analysis of samples from distributions of the largest value in groups of m events from the basic distribution begins with arrangement of the data in increasing order of magnitude. The data are then plotted on rectangular coordinate paper with X values (peak flows, flood volumes, etc.) along the ordinate and rank along the abscissa. The smallest value is plotted at r = 1/2, the next larger at r = 1½, and the largest at r = R-½ in accordance with previously determined procedure (4). R is the number of maxima observed and recorded. A curving line is fitted

* The computations involved in preparing Table 2 were facilitated by the use of National Bureau of Standards publications, numbers 14 (13) and 46 (14) of the applied mathematics series.

by eye through the plotted points, or successive points on the graph are connected by straight lines.

Then for each value of $Q_m = r/R$ of Table 2 an r value is mentally computed. The X value of the curve for each r value so obtained is then plotted at an abscissa value corresponding to its associated Ψ value in the table. The Ψ values for an arbitrarily selected m are used. A straight line is fitted by eye to the replotted values of X giving extra weight to the central values, and in the case of peak flows to those within-bank discharges (Ψ values less than about 1) for which the stage discharge relationship is usually more adequately established. Little weight is given to the extreme values.

The slope, k, of the line is determined. The modal value of X is obtained from the line at the modal value of Ψ for the selected m. Then the ratio of this modal value of X to k is equal to $\log \bar{m}$. If this computed value of $\bar{m}$ differs greatly from the arbitrarily chosen value, a replotting is desirable with a Ψ table m closer to the computed $\bar{m}$. However, it is found that the derived k's and $\bar{m}$'s are not very sensitive to moderate changes in the selected m.

An illustration of the graphical method of analysis of samples from distributions of the largest is given in Figure 4 where 39 annual maxima peak discharges in the Embarrass River, Ste. Marie, Illinois, are plotted. The derived value of $\bar{m}$ by this method is 3.8. By the method of moments it is 3.1; and by analysis of single flood events in a manner similar to that of Figure 2 it is 4.4. The corresponding k values are 19,300, 20,100 and 20,100, respectively.

AN AID TO SOLUTION OF THE EQUATION FOR THE LARGEST EVENT

According to Table 1, the magnitude of the largest event associated with the probability, Q_m, of not occurring within Y units of time is

$$X = -k \log [1-Q_m^{1/\bar{m}Y}] \tag{25}$$

This can be expressed by

$$X = k(A + \log \bar{m}Y) \tag{26}$$

where $k=\log \bar{m}Y$ is the modal value of X. Table 3 gives values of A for selected values of $m = \bar{m}Y$ and Q_m. The general expression for A is

$$A = -\log m \ (1-Q_m^{1/m}) \tag{27}$$

For the particular case that m = 1,

$$A = -\log (1-Q_m) \tag{28}$$

and for the case that m becomes very large ($m \longrightarrow \infty$)

$$A = -0.434 \ln (-\ln Q_m) \tag{29}$$

$$= -\log (-2.3 \log Q_m)$$

Table 3.--Values of A in the equation $X = k(A+\log \bar{m}Y)$

Q_m / m	0.50	0.80	0.90	0.95	0.98	0.99
1	0.30	0.70	1.00	1.30	1.70	2.00
2	0.23	0.67	0.99	1.30	1.70	
5	0.19	0.66	0.98	1.29		
10	0.17	0.66				
20	0.17	0.65				
50	0.16					
∞	0.159	0.651	0.977	1.290	1.695	1.998

GENERALIZED REPRESENTATION OF INTENSE, SHORT-PERIOD RAINFALL

A special application of equations (25) and (26) is for short period intense rainfall amounts. Satisfactory representations have been obtained by letting $k = g \log (10H+1)$ where g varies in value geographically and H is the intense period of storm rainfall in hours, up to two hours. The value of g at Washington, D. C., for example, is 0.94 as determined in the analysis of 15-minute rainfall in Figure 2. At Oxford, Mississippi it is 0.925. At Elkins, W. Va. it is 0.64.

APPLICATIONS TO DESIGN PROBLEMS

Whereas much of the thinking in the past has been based upon modal values of the largest, and in this report the median value of the largest is stressed; there is need to progress to the concept of other probabilities of occurrence within limited time periods. To provide additional information about the distribution in time of the larger values of the variable, and for an aid in the transition in thinking from that of a fixed maximum within a given time period to that of probabilities of occurrence within limited time periods, equation (30) is given. If through use of equation (25) the value of the variable with probability Q_m of not occurring in Y years is equated to the probable value in Y_{50} years, the period of time, Y_{50}, is obtained in terms of Y and Q_m. This value of time approaches current concepts of return period or recurrence interval. Thus

$$Y_{50} = \frac{-0.3Y}{\log Q_m} \qquad (30)$$

In other words, the X value with a probability Q_m of not being exceeded in Y years is the probable maximum value in Y_{50} years.

As an example, in the design of a small hydraulic structure, the owner estimates a useful life of 25 years, and the probable damages to be suffered by exceeding the capacity warrants the assumption of a 10 percent chance of this happening one or more times in the 25-year period. Then Y = 25 years and $Q_m = 0.90$ and

$$Y_{50} = \frac{-0.3(25)}{\log 0.90} = 164 \text{ years} \qquad (31)$$

Then the capacity should be equal to the probable maximum flow for a 164-year period.

Although the individual owner may be willing to assume such a risk as that of this example, his consultant in design may not be willing to do so. The maximum size of the event to be experienced by the consultant or design agency is larger, and is dependent upon the structure-years of existence of

those structures of concern, rather than on the existence of one structure. Just as for the individual owner with the design conditions of equation (31), the consultant has only a 50-50 chance of complete containment of all the flood waters in 164 structure-years of service.

Take another illustration. Assume that a consulting firm or agency is to design 10 structures with an average expected life of Y = 50 years each; and it has somehow arrived at the conviction that design procedures should be such that there is only a 10 percent chance that the capacity of one or more of the N = 10 structures will be exceeded in the 500 structure-years of planned service. Then the probability that the capacity of any particular one of the structures will not be exceeded during its expected life of 50 years is nearly 99 percent, and the 50-50 design period, Y_{50}, for each structure is 3289 years. This of course assumes a geographical dispersion of the structures that assures essentially, independent meteorological events for each.

Table 4 gives the project design periods (the years to be considered in estimating the size of the hydrologic event) for selected risks and planned structure-years of service. In other words the project design period is the length of time for which the chances are 50 percent that the design value of the variate, rainfall or runoff, will not be exceeded and 50 percent that it will be exceeded one or more times. The values in Table 4 seem large. A comparison of these time periods and probabilities with the relevant language that is in vogue for project design criteria (such as design for a 50-year flood), along with the lack of extensive numbers of failures, suggests that there has been great implicit dependence upon the use of large factors of safety in design.

Table 4.--Project design periods for selected risks, years

Life of Project	Probability of being exceeded one or more times, percent							
Structure-years	0.1	0.5	1	2	5	10	25	50
1	693	138	69	34	13.5	6.6	2.4	1
5	3464	691	345	172	68	33	12	5
10	6928	1383	690	343	135	66	24	10
25	17320	3457	1724	858	338	164	60	25
50	34640	6914	3448	1716	676	329	120	50
100	69280	13828	6897	3431	1351	658	241	100
500	346400	69141	34484	17155	6757	3289	1205	500
1000	692801	138283	68968	34310	13513	6579	2409	1000
5000	3,464,003	691413	344838	171548	67567	32894	12047	5000

FEATURES OF THE GAMMA DISTRIBUTION; AND APPLICATIONS

Analysis of data for which the Gamma Distribution is applicable involves the computation of the arithmetic mean and the variance. Each item of a sample is a sum $S = n\bar{x}$ of several amounts; as the amount of rainfall in a year is the sum of the amounts in several storms.

The arithmetic mean, $\bar{S} = \overline{n\bar{x}}$ of the N items of the sample is equal to n/c; and the variance is equal to $\chi = \frac{N}{N-1} s^2 = n/c^2$, where s is the sample

standard deviation. Then $c = \bar{S}/\chi$. Also $n = (\bar{S})^2/\chi = c\bar{S}$; and the mean of all the x's is 1/c. It should be remembered that in the usual application, to rainfall and runoff magnitudes, items from two distributions have been thrown together. The n and mean $\bar{x}$ therefore lose some of their rationality and should be considered by definition to be the number and mean size of events as determined by the application of the Gamma function. Knowing the values of n and c fixes the distribution function. Pearson's (15) tables of the Incomplete Gamma Function may then be used for values of n from 1 to 51 to find the probability Q_n associated with a given S value, or vice-versa. Table 5 shows the notational equivalents for this and Pearson's system of representation.

Table 5.--Notational equivalents for the Incomplete Gamma Function

Pearson's Notation	Equivalent Symbols of this paper
p	n-1
v	$cn\bar{x} = n(S/\bar{S})$
I(u,p)	Q_n
$u = v/(p+1)^{\frac{1}{2}}$	* $cn\bar{x}/n^{\frac{1}{2}} = c(n)^{\frac{1}{2}}\bar{x} = (n)^{\frac{1}{2}}(S/\bar{S})$

* Pearson's u is the ratio of $cn\bar{x}$ to the standard deviation of $cn\bar{x}$.

Another method of analysis of data from a population distributed in accordance with the Incomplete Gamma Function is to use Fisher's (16) maximum likelihood estimator. Thom's (17) estimate of n, based upon this method is

$$n = \frac{1 + (1+4A/3)^{\frac{1}{2}}}{4A} \qquad (32)$$

where $A = 2.3\ (\log \bar{S} - \overline{\log S})$

The Incomplete Gamma Function is especially suited to determinations of the probability of occurrence of excesses and deficiencies in annual or monthly rainfall and in annual runoff.

The n values for the number of floods per year are found to be roughly related and about equal to the corresponding $\bar{m}$ values for the number of floods per year as determined by analysis of peak flow values.

Approximate determinations of n values for flood occurrences can be made if drainage area, mean annual runoff and mean size of storm rainfall are known.

DATA ANALYSIS IN THE CASE OF SMALL VALUES OF m AND n

Implicit in the derivations (4) of the Distribution of the Largest and the Gamma distribution from the basic distribution of single events is the adoption of constant values of m and n, respectively, in successive groups of events. This is not truly the case in nature; and is an imperfection in the procedures. A limited study of the possible consequences in data analysis of ignoring the effect of variations in m and n from one group to another is in order.

Making the assumption that the time intervals between occurrences of

events are random, subject only to the restriction that in Y units of time m events occur, such that $m = \bar{m}Y$, the time distribution of occurrences is the Binomial distribution. Then the probability of getting exactly z events in a certain unit of time (or the proportion of the Y units of time in which z events occur) is

$$P = mCz\ p^z\ q^{m-z} \quad (z = 0,1, \ldots\ldots\ldots m) \tag{33}$$

which is the zth term of the expansion of $(p + q)^m$; where $p = 1/Y$ = the probability that a random event will fall within a certain time interval, and $q = (Y-1)/Y$ = the probability that a random event will not fall within a certain time interval.

It can readily be shown that when Y is large, i.e., for a long record of observation (strictly for an infinite time), the Binomial distribution approaches the Poisson distribution for which the proportion of the Y units of time in which z events occur is

$$P = \frac{e^{-\bar{m}}\ \bar{m}^z}{z!} \quad (z = 0, 1, \ldots\ldots\ldots m) \tag{34}$$

Because the basic data tend to become voluminous as magnitudes approach zero values, and because small values are contaminated by items from the group of smaller things, partial samples must be used in analysis of basic events. This prevents an experimental check upon the applicability of the Binomial and Poisson distributions; but there is little doubt about their representativeness. If so, a consideration of these distributions of events in time are highly important in data analysis for many cases. Table 6 shows the probabilities of having z events per unit time when the average number per unit is $\bar{m}$, in accordance with the Poisson distribution.

Table 6.--Probabilities of z events per unit of time according to the Poisson Distribution*

Av. No. per unit of time	Values of z					
$\bar{m}$	0	1	2	3	4	5
1/2	0.61	0.30	0.08	0.01		
1	0.37	0.37	0.18	0.06	0.02	
3/2	0.22	0.33	0.25	0.13	0.05	0.01
2	0.14	0.27	0.27	0.18	0.09	0.04
3	0.05	0.15	0.22	0.22	0.17	0.10
4	0.02	0.07	0.15	0.20	0.20	0.16

* From Burington and May (18) Table VII p. 259

The practice of disregarding, in the reporting and analysis of data, all events less than the largest in a given time period is wasteful of data. When this is done, it is necessary according to these concepts to use the Distribution of the Largest rather than the distribution for single events in the analysis. Furthermore, if the average number of events per unit of time is less than about 3 or 4, it is quite possible, as indicated in Table 6 that some units of time have passed without an occurrence of an event from the class of large things. In this case, and in the case when all of the few events from the class of large things are very small, the item of greatest magnitude in a time interval comes from the group of things that are small and are of no immediate concern.

There is then a mixing of data from two populations, and the small items in the array of values will be larger than they should be. This precludes the use of numerical methods of analysis, and makes questionable the results of graphical analysis in which some allowance for the mixing of data may be made. A still more wasteful procedure, but one that may be desirable in the case of a long record and low $\bar{m}$ value, is to double the time interval by using the larger item in successive pairs of items in chronological order. In the case of sums of events, where the Gamma function would be applicable, sums for greater time periods may be used where n values are very small. For example, the analysis of rainfall for two-year periods rather than for one may be desirable in semi-arid regions.

CONCLUSIONS

Generally satisfactory representations of the distributions of magnitude of discrete rainfall and runoff events are provided by the descending exponential function. A newly named family of distribution functions, called the "Distribution of the Largest," has been found to adequately portray the distributions of largest items to occur in given time periods. The skewness of these functions varies from the extreme of the exponential function to the moderate value of the extreme value function. The family of distributions known as the Incomplete Gamma Function is used to represent sums of rainfall or runoff events, for example, annual values of rainfall and runoff. The three functions, the descending exponential, the distribution of the largest, and the incomplete gamma function are consistent, the latter two being derivable from the first.

The parameters of the distribution functions that relate to magnitude and skewness have geographical variations that, if known, would help in selection of satisfactory prediction equations for cases of scanty data. The skewness is related to the mean number of events per unit time.

A more widespread adoption of probability concepts, and distribution functions like those presented, would be a more realistic approach to the determination of hydraulic structure capacities.

SYMBOLS AND DEFINITIONS

a = an arbitrarily set value of cx.

A = an approximating term in equation 26.

A = $2.3\ (\log_{10} \overline{S} - \overline{\log_{10} S})$ after Thom, equation 32.

c = a constant dependent in magnitude upon the unit of measure of the variable and upon all factors affecting the value of x other than purely random variation within the limits prescribed by the function.

c.v. = coefficient of variation = standard deviation divided by arithmetic mean.

${}_mC_z$ = the combination of m things taken z at a time.

e = the constant 2.71828 . . . . . .

$f(m)$ = function of m = the square of the coefficient of variation of X in Table 3 of reference (4).

$f(m,Q_m)$ = function of m and Q_m.

g = $k/\log(10H+1)$, a component of k that varies geographically in the representation of short period intense rainfall amounts.

H = the period of intense storm rainfall in hours.

$I(u,p)$ = Q_n. $I(u,p)$ is Pearson's notation for the probability that a random value from a population with a Gamma distribution characterized by p will not exceed u.

k = the sequence number of an item in a summation.

$k = 2.3/c$ = a constant dependent upon the unit of measure of x, and upon the quantitative values of factors affecting the x values other than chance.

ln = natural logarithm.

log = common or Briggs' logarithm.

m = the number in a group of random events with values x_1, x_2,.....x_m from which the largest value of x, designated X is chosen.

$\bar{m}$ = the average number of events per unit of time.

n = the number in a group of random events whose values are added together such that $n\bar{x} = x_1 + x_2 + \ldots\ldots + x_n$.

N = the number of items in a sample.

p = (1) the probability that the value x will be exceeded in a single trial; (2) the proportion r/R of the items of a sample that exceed x; (3) Karl Pearson's p = n-1 (Table 5).

P = probability of success; equations 33 and 34, the proportion of the Y units of time in which z events occur.

q = the probability that the value x will not be exceeded in a single trial.

Q = the proportion of a population less than a given value.

Q_m = q^m = the probability that all m randomly selected events will have values less than X; or in other words, the probability that the largest of m events will be less than X.

Q_n = the probability that the sum of n random items will not exceed $n\bar{x}$.

r = the item or sequence number of ranked members of a sample.

R = the total number of items in a sample.

s = the sample standard deviation.

S = $n\bar{x} = x_1 + x_2 + \ldots\ldots x_n$ = the sum of n random values of x.

t = time.

u = $v/(p+1)^{\frac{1}{2}}$ (Pearson) = $n^{\frac{1}{2}}\ (S/\bar{S})$.

v = $cn\bar{x} = n(S/\bar{S})$. v is Pearson's notation.

y = the rate of change of q, Q_m or Q_n with their respective variables; the ordinate to a frequency curve.

Y = units of time, usually years.

Y_{50} = the period over which a given magnitude of event would have a 50-50 chance of occurrence.

x = the magnitude of a single event chosen at random.

X = the value of the largest x in a random sample of m x's.

$\overline{x}$ = the arithmetic mean value of n x's; thus $n\bar{x} = x_1 + x_2 + \ldots. + x_n$.

$\tilde{X}$ = the modal value of X; the tilde over any other variable indicates the modal value of the variable.

X_{50} = the median value of X.

χ (chi) = the variance is the mean value of the squares of the deviations from the arithmetic mean and is equal to the square of the standard deviation σ (sigma). It is sometimes symbolized by (mu sub 2) μ_2 and called the second moment of the variable about its arithmetic mean.

ϵ epsilon = $X - \tilde{X}$ = the difference between a random value and the modal value of the largest in groups of m.

$\Gamma_{(n)}$ = the complete Gamma function of n.

ν_1 = nu sub 1 = the arithmetic mean.

ψ psi = $1 + \frac{X-X_{50}}{k}$ = a scale of values computed to give a straight line relationship of the ranked X values of a sample from a Distribution of the Largest in samples from an exponentially distributed population.

Σ sigma = summation.

$\longrightarrow \infty$ = approaches infinity.

REFERENCES

(1) Schafmayer, A. J., and Grant, B. E., "Rainfall Intensities and Frequencies", Proc. Am. Soc. Civil Engineers, vol. 63, no. 2, pp. 225-249, February (1937).

(2) Fisher, R. A., and Tippett, L. H. C., "Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample", Proc. Cambridge Phil. Sec., vol. 24, Pt. 2 (1928).

(3) Gumbel, Emil J., "Statistical Theory of Extreme Values and Some Practical Applications", Nat'l. Bur. of Standards, Applied Math. Series, no. 33 (1954).

(4) Parsons, Donald A., "Consistent Representations of Rainfall and Runoff Magnitudes", Research Report No. 334, Watershed Technology Research Branch, Soil and Water Conservation Research Division, Agricultural Research Service, US Department of Agriculture, October (1960). (Unpublished. Limited Supply. Single copies available on request.)

(5) Barger, G. L., and Thom, H. C. S., "Evaluation of Drought Hazard", Agronomy Journal, vol. 41, no. 11, pp. 519-526, November (1949).

(6) Friedman, Don G., "The Prediction of Long Continuing Drought in South and Southwest Texas", The Travelers Insurance Co., Occasional Papers in Meteorology, No. 1, September (1957).

(7) Friedman, Don G., and Janes, Byron E., "Estimation of Rainfall Probabilities", Storrs Agricultural Experiment Station, University of Connecticut, Bulletin 332, December (1957).

(8) Thom, H. C. S., "Three Chapters on Climatological Analysis", Manuscript of the U.S. Weather Bureau, August (1960).

(9) Borgman, L. E., "The Frequency Distribution of Near Extremes", Am. Geophysical Union, Journal of Research, vol. 66, no. 10, pp. 3295-3307, October (1961).

(10) Linsley, R. K., Jr., Kohler, M. A., and Paulhus, J. L., "Applied Hydrology", McGraw-Hill Book Co., Inc. (1949).

(11) Yarnell, David L., "Rainfall Intensity Frequency Data", US Department of Agriculture Miscellaneous Publication No. 204 (1935).

(12) Miami Conservancy District Technical Reports (Part v), "Storm Rainfall of Eastern United States" (1917).

(13) National Bureau of Standards, "Tables of the Exponential Function", No. 14 of the Applied Mathematics Series (1951).

(14) National Bureau of Standards, "Tables of the Descending Exponential", No. 46 of the Applied Mathematics Series (1955).

(15) Pearson, Karl, "Tables of the Incomplete Gamma Function", Cambridge University Press (1922).

(16) Fisher, R. A., "On the Mathematical Foundations of Theoretical Statistics", Philosophical Transactions Royal Society, Series A, vol. 222, pp. 309-368 (1941).

(17) Thom, H. C. S., "A Note on the Gamma Distribution", Monthly Weather Review, vol. 86, no. 4, pp. 117-122, April (1958).

(18) Burington, R. S., and May, D. C., "Handbook of Probability and Statistics", Handbook Publishers, Inc., Sandusky, Ohio (1953).

(19) Blumenstock, David I., "Rainfall Characteristics as Related to Soil Erosion", US Department of Agriculture Technical Bulletin 698 (1939).

(20) Brakensiek, D. L., and Zingg, A. W., "Application of the Extreme Value Statistical Distribution to Annual Precipitation and Crop Yields", US Department of Agriculture, Agricultural Research Service, ARS 41-13, February (1957).

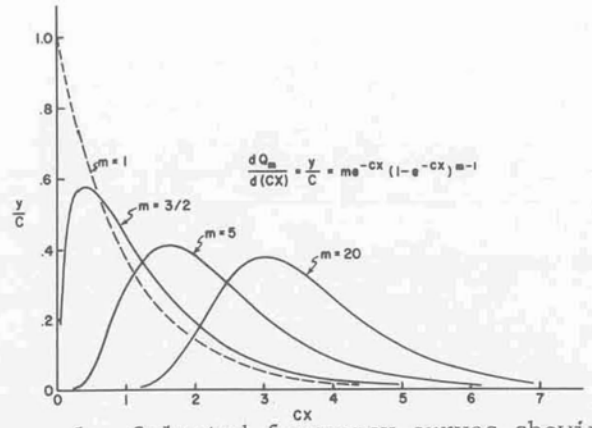

Figure 1.--Selected frequency curves showing the Distribution of the Largest in samples from the Exponential Distribution

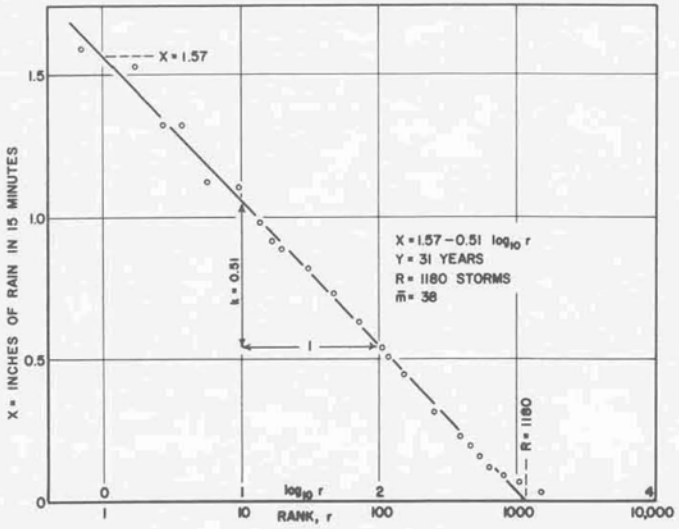

Figure 2.-- 15-minute rainfall at Washington, D. C.

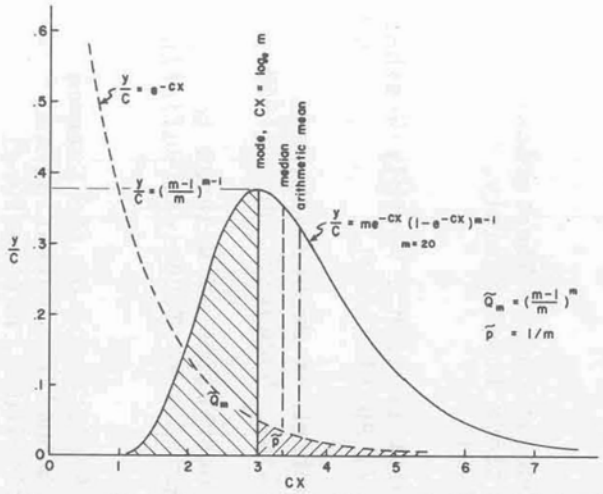

Figure 3.--Characteristics of the Distribution of the Largest in samples of m from the exponentially distributed population

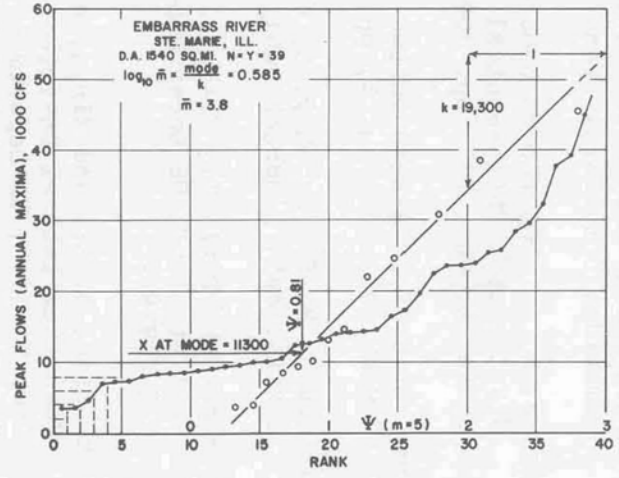

Figure 4.--Analysis of annual peak flows of the Embarrass River at Ste. Marie, Illinois