

### **D.4.3 Flood Frequency Analysis Methods**

This section outlines general features of statistical methods used in a flood insurance study, including providing basic statistical tools that are frequently needed. It is recommended that extremal analysis of annual maxima be performed using the Generalized Extreme Value (GEV) Distribution with parameters estimated by the Method of Maximum Likelihood. The discussion in this section is illustrative only; guidelines for application of these tools in specific instances are provided in other sections of this appendix.

#### **D.4.3.1 The 1% Annual Chance Flood**

The primary goal of a coastal Flood Insurance Study (FIS) is to determine the flood levels throughout the study area that have a 1% chance of being exceeded in any given year. The level that is exceeded at this rate at a given point is called the 1% annual chance flood level at that point, and has a probability of 0.01 to be equaled or exceeded in any year; on the average, this level is exceeded once in 100 years and is commonly called the 100-year flood.

The 1% annual chance flood might result from a single flood process or from a combination of processes. For example, astronomic tide and storm waves combine to produce the total high water runup level. There is no one-to-one correspondence between the 1% annual chance flood elevation and any particular storm or other flood-producing mechanism. The level may be produced by any number of mechanisms, or by the same mechanism in different instances. For example, an incoming wave with a particular height and period may produce the 1% annual chance runup, as might a quite different wave with a different combination of height and period.

Furthermore, the flood hazard maps produced as part of an FIS do not necessarily display, even locally, the spatial variation of any realistic physical hydrologic event. For example, the 1% annual chance levels just outside and just inside an inlet will not generally show the same relation to one another as they would during the course of any real physical event because the inner waterway may respond most critically to storms of an entirely different character from those that affect the outer coast. Where a flood hazard arises from more than one source, the mapped level is not the direct result of any single process, but is a construct derived from the statistics of all sources. Note then that the 1% annual chance flood level is an abstract concept based as much on the statistics of floods as on the physics of floods.

Because the 1% annual chance flood level cannot be rigorously associated with any particular storm, it is a mistake to think of some observed event as having been the 1% annual chance event. A more intense storm located at a greater distance might produce the same flood level, or the same flood level might be produced by an entirely different mechanism, such as by a tsunami from a distant landslide or earthquake. Furthermore, if a particular storm were, in fact, the so-called 100-year event, it could not be so everywhere, but only in its effect at a particular point.

The 1% annual chance flood level is a consequence solely of the areawide flooding mechanisms recognized for a particular location. That is, there may be mechanisms that are not taken into account, but that could also produce water levels comparable to the 1% level or that could

contribute to the 1% level. For example, tsunamis occur in all oceans, so even the Atlantic Coast is vulnerable to tsunami attack at some frequency. The Great Lisbon earthquake of 1755 (with magnitude approaching 9) produced a large Atlantic tsunami that was felt in the New World; however, tsunamis are not recognized as areawide flood sources for the Atlantic Coast. Similarly, advances in science may from time to time reveal new flood mechanisms that had not previously been recognized; for example, only in recent years has the physics of El Niños been clarified and their contribution to coastal flood levels recognized.

#### **D.4.3.2 Event vs. Response Statistics**

The flood level experienced at any coastal site is the complicated result of a large number of interrelated and interdependent factors. For example, coastal flooding by wave runup depends upon both the local waves and the level of the underlying still water upon which they ride. That still water level (SWL), in turn, depends upon the varying astronomic tide and the possible contribution of a transient storm surge. The wave characteristics that control runup include amplitude, period, and direction, all of which depend upon the meteorological characteristics of the generating storm including its location and its time-varying wind and pressure fields. Furthermore, the resulting wave characteristics are affected by variations of water depth over their entire propagation path, and thus depend also on the varying local tide and surge. Still further, the beach profile, changing in response to wave-induced erosion, is variable, causing variation in the wave transformation and runup behavior. All of these interrelated factors may be significant in determining the coastal flood level with a 1% annual chance of occurrence.

Whatever methods are used, simplifying assumptions are inevitable, even in the most ambitious response-based study, which attempts to simulate the full range of important processes over time. Some of these assumptions may be obvious and would introduce little error. For example, a major tsunami could occur during a major storm, and it might alter the storm waves and runup behavior and dominate the total runup. However, the likelihood of this occurrence is so small that the error incurred by ignoring the combined occurrence would be negligible. On the other hand, the conclusion might not be so clear if the confounding event were to be storm surge rather than a tsunami because extreme waves and surge are expected to be correlated, with high waves being probable during a period of high surge.

These guidelines offer insight and methods to address the complexity of the coastal flood process in a reasonable way. However, the inevitable limitations of the guidance must be kept in mind. No fixed set of rules or cookbook procedures can be appropriate in all cases, and the Mapping Partner must be alert to special circumstances that violate the assumptions of the methodology.

##### **D.4.3.2.1 Event-Selection Method**

A great simplification is made if one can identify a single event (or a small number of events) that produces a flood thought to approximate the 1% flood. This might be possible if, for example, a single event parameter (such as deep-water wave height) is believed to dominate the final runup, so the 1% value of that particular item might suffice to determine the 1% flood. In its simplest form, one might identify a significant wave height thought to be exceeded with only 1% chance, and then to follow this single wave as it would be transformed in propagation and as it would run up the beach. This is the *event-selection method*. Used with caution, this method may

allow reasonable estimates to be made with minimal cost. It is akin to the concept of a design storm, or to constructs such as the standard project or probable maximum storms.

The inevitable difficulty with the event-selection method is that multiple parameters are always important, and it may not be possible to assign a frequency to the result with any confidence because other unconsidered factors always introduce uncertainty. Smaller waves with longer periods, for example, might produce greater runup than the largest waves selected for study. A slight generalization of the event-selection method, often used in practice, is to consider a small number of parameters – say wave height, period, and direction – and attempt to establish a set of alternative, “100-year” combinations of these parameters. Alternatives might be, say, pairs of height and period from each of three directions, with each pair thought to represent the 1% annual chance threat from that direction, and with each direction thought to be associated with independent storm events. Each such combination would then be simulated as a selected “event”, with the largest flood determined at a particular site being chosen as the 100-year flood. The probable result of this procedure would be to seriously underestimate the true 1% annual chance level by an unknown amount. This can be seen easily in the hypothetical case that all three directional wave height and period pairs resulted in about the same flood level. Rather than providing reassurance that the computed level were a good approximation of the 100-year level, such a result would show the opposite – the computed flood would not be at the 100-year level, but would instead approximate the 33-year level, having been found to result once in 100 years from each of three independent sources, for a total of three times in 100 years. It is not possible to salvage this general scheme in any rigorous way – say by choosing three, 300-year height and period combinations, or any other finite set based on the relative magnitudes of their associated floods – because there always remain other combinations of the multiple parameters that will contribute to the total rate of occurrence of a given flood level at a given point, by an unknown amount.

#### **D.4.3.2.2 Response-based Approach**

With the advent of powerful and economical computers, a preferred approach that considers all (or most) of the contributing processes has become practical; this is the *response-based approach*. In the response-based approach, one attempts to simulate the full complexity of the physical processes controlling flooding, and to derive flood statistics from the results (the local response) of that complex simulation. For example, given a time history of offshore waves in terms of height, period, and direction, one might compute the runup response of the entire time series, using all of the data and not pre-judging which waves in the record might be most important. With knowledge of the astronomic tide, this entire process could be repeated with different assumptions regarding tidal amplitude and phase. Further, with knowledge of the erosion process, storm-by-storm erosion of the beach profile might also be considered, so its feedback effect on wave behavior could be taken into account.

At the end of this process, one would have a long-term simulated record of runup at the site, which could then be analyzed to determine the 1% level. Clearly, successful application of such a response-based approach requires a tremendous effort to characterize the individual component processes and their interrelationships, and a great deal of computational power to carry out the intensive calculations.

The response-based approach is preferred for all Pacific Coast FISs.

#### **D.4.3.2.3 Hybrid Method**

Circumstances may arise for which the Mapping Partner can adopt a *hybrid* method between the event-selection and response-based extremes; this hybrid method may substantially reduce the time required for repeated calculations. The Mapping Partner must use careful judgment in applying this method to accurately estimate the flood response (e.g., runup); detailed guidance and examples of the method can be found in PWA (2004).

The hybrid method uses the results of a response-based analysis to guide the selection of a limited number of forcing parameters (e.g., water level and wave parameter combinations) likely to approximate the 1% annual chance flood response (e.g., runup). A set of baseline response-based analyses are performed for transects that are representative of typical geometries found at the study site (e.g., beach transects with similar slopes; coastal structures with similar toe and crest elevations, structure slopes, and foreshore slopes). The results obtained for these representative transects are then used to guide selection of parameters for other similar transects within the near vicinity. The Mapping Partner may need to consider a range of forcing parameters to account for variations in the response caused by differences in transect geometry; a greater range of forcing parameters will need to be considered for greater differences between transect geometries.

The hybrid method simply postulates that if a set of wave properties can be found that reproduces the 1% annual chance flood established by a response-based analysis at a certain transect, then the same set of parameters should give a reasonable estimate at other transects that are both similar and nearby.

#### **D.4.3.3 General Statistical Methods**

##### **D.4.3.3.1 Overview**

This section summarizes the statistical methods that will be most commonly needed in the course of an FIS to establish the 1% annual chance flood elevation. Two general approaches can be taken depending upon the availability of observed flood data for the site. The first, preferred, approach is used when a reasonably long observational record is available, say 30 years or more of flood or other data. In this *extreme value analysis* approach, the data are used to establish a *probability distribution* that is assumed to describe the flooding process, and that can be evaluated using the data to determine the flood elevation at any frequency. This approach can be used for the analysis of wind and tide gage data, for example, or for a sufficiently long record of a computed parameter such as wave runup.

The second approach is used when an adequate observational record of flood levels does not exist. In this case, it may be possible to simulate the flood process using hydrodynamic models driven by meteorological or other processes for which adequate data exist. That is, the hydrodynamic model (perhaps describing waves, tsunamis, or surge) provides the link between the known statistics of the generating forces, and the desired statistics of flood levels. These *simulation* methods are relatively complex and will be used only when no acceptable, more economical alternative exists. Only a general description of these methods is provided here; full

documentation of the methods can be found in the user's manuals provided with the individual simulation models. The manner in which the 1% annual chance level is derived from a simulation will depend upon the manner in which the input forcing disturbance is defined. If the input is a long time series, then the 1% level might be obtained using an extreme value analysis of the simulated process. If the input is a set of empirical storm parameter distributions, then the 1% level might be obtained by a method such as joint probability or Monte Carlo, as discussed later in this section.

The present discussion begins with basic ideas of probability theory and introduces the concept of a continuous probability distribution. Distributions important in practice are summarized, including, especially, the extreme value family. Methods to fit a distribution to an observed data sample are discussed, with specific recommendations for FIS applications. A list of suggested additional information resources is included at the end of the section.

#### D.4.3.3.2 Elementary Probability Theory

Probability theory deals with the characterization of random events and, in particular, with the likelihood of occurrence of particular outcomes. The word "probability" has many meanings, and there are conceptual difficulties with all of them in practical applications such as flood studies. The common frequency notion is assumed here: the probability of an event is equal to the fraction of times it would occur during the repetition of a large number of identical *trials*. For example, if one considers an annual storm season to represent a trial, and if the event under consideration is occurrence of a flood's exceeding a given elevation, then the annual probability of that event is the fraction of years in which it occurs, in the limit of an infinite period of observation. Clearly, this notion is entirely conceptual, and cannot truly be the source of a probability estimate.

An alternate measure of the likelihood of an event is its expected *rate of occurrence*, which differs from its probability in an important way. Whereas probability is a pure number and must lie between zero and one, rate of occurrence is a measure with physical dimensions (reciprocal of time) that can take on any value, including values greater than one. In many cases, when one speaks of the probability of a particular flood level, one actually means its rate of occurrence; thinking in terms of physical rate can help clarify an analysis.

To begin, a number of elementary probability rules are recalled. If an event occurs with probability  $P$  in some trial, then it fails to occur with probability  $Q = 1 - P$ . This is a consequence of the fact that the sum of the probabilities of all possible results must equal unity, by the definition of total probability:

$$\sum_i P(A_i) \equiv 1 \quad (\text{D.4.3-1})$$

in which the summation is over all possible outcomes of the trial.

If  $A$  and  $B$  are two events, the probability that either  $A$  or  $B$  occurs is given by:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \quad (\text{D.4.3-2})$$

If  $A$  and  $B$  are mutually exclusive, then the third term on the right-hand side is zero and the probability of obtaining either outcome is the sum of the two individual probabilities.

If the probability of  $A$  is contingent on the prior occurrence of  $B$ , then the *conditional probability* of  $A$  given the occurrence of  $B$  is defined to be:

$$P(A | B) \equiv \frac{P(AB)}{P(B)} \quad (\text{D.4.3-3})$$

in which  $P(AB)$  denotes the probability of both  $A$  and  $B$  occurring.

If  $A$  and  $B$  are *stochastically independent*,  $P(A/B)$  must equal  $P(A)$ , then the definition of conditional probability just stated gives the probability of occurrence of both  $A$  and  $B$  as:

$$P(AB) = P(A)P(B) \quad (\text{D.4.3-4})$$

This expression generalizes for the joint probability of any number of independent events, as:

$$P(ABC\dots) = P(A)P(B)P(C)\dots \quad (\text{D.4.3-5})$$

As a simple application of this rule, consider the chance of experiencing at least one 1% annual chance flood ( $P = 0.01$ ) in 100 years. This is 1 minus the chance of experiencing no such flood in 100 years. The chance of experiencing no such flood in 1 year is 0.99, and if it is granted that floods from different years are independent, then the chance of not experiencing such a flood in 100 years is  $0.99^{100}$  according Equation D.4.3-5 or 0.366. Consequently, the chance of experiencing at least one 100-year flood in 100 years is  $1 - 0.366 = 0.634$ , or only about 63%.

#### D.4.3.3.3 Distributions of Continuous Random Variables

A continuous random variable can take on any value from a continuous range, not just a discrete set of values. The instantaneous ocean surface elevation at a point is an example of a continuous random variable; so, too, is the annual maximum water level at a point. If such a variable is observed a number of times, a set of differing values *distributed* in some manner over a range is found; this fact suggests the idea of a *probability distribution*. The observed values are a *data sample*.

We define the *probability density function*, PDF, of  $x$  to be  $f(x)$ , such that the probability of observing the continuous random variable  $x$  to fall between  $x$  and  $x + dx$  is  $f(x) dx$ . Then, in accordance with the definition of total probability stated above:

$$\int_{-\infty}^{\infty} f(x) dx = 1 \quad (\text{D.4.3-6})$$

If we take the upper limit of integration to be the level  $L$ , then we have the definition of the *cumulative distribution function*, CDF, denoted by  $F(x)$ , which specifies the probability of obtaining a value of  $L$  or less:

$$F(x \leq L) \equiv \int_{-\infty}^L f(x) dx \quad (\text{D.4.3-7})$$

It is assumed that the observed set of values, the sample, is derived by random sampling from a *parent distribution*. That is, there exists some unknown function,  $f(x)$ , from which the observed sample is obtained by random selection. No two samples taken from the same distribution will be exactly the same. Furthermore, random variables of interest in engineering cannot assume values over an unbounded range as suggested by the integration limits in the expressions shown above. In particular, the lower bound for flood elevation at a point can be no less than ground level, wind speed cannot be less than zero, and so forth. Upper bounds also exist, but cannot be precisely specified; whatever occurs can be exceeded, if only slightly. Consequently, the usual approximation is that the upper bound of a distribution is taken to be infinity, while a lower bound might be specified.

If the nature of the parent distribution can be inferred from the properties of a sample, then the distribution provides the complete statistics of the variable. If, for example, one has 30 years of annual peak flood data, and if these data can be used to specify the underlying distribution, then one could easily obtain the 10-, 50-, 100-, and 500-year flood levels by computing  $x$  such that  $F$  is 0.90, 0.98, 0.99, and 0.998, respectively.

The entirety of the information contained in the PDF can be represented by its moments. The *mean*,  $\mu$ , specifies the *location* of the distribution, and is the first moment about the origin:

$$\mu = \int_{-\infty}^{\infty} x f(x) dx \quad (\text{D.4.3-8})$$

Two other common measures of the location of the distribution are the mode, which is the value of  $x$  for which  $f$  is maximum, and the median, which is the value of  $x$  for which  $F$  is 0.5.

The spread of the distribution is measured by its variance,  $\sigma^2$ , which is the second moment about the mean:

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \quad (\text{D.4.3-9})$$

The *standard deviation*,  $\sigma$ , is the square root of the variance.

The third and fourth moments are called the skew and the kurtosis, respectively; still higher moments fill in more details of the distribution shape, but are seldom encountered in practice. If the variable is measured about the mean and is normalized by the standard deviation, then the *coefficient of skewness*, measuring the asymmetry of the distribution about the mean, is:

$$\eta_3 = \int_{-\infty}^{\infty} \left(\frac{x - \mu}{\sigma}\right)^3 f(x) dx \quad (\text{D.4.3-10})$$

and the *coefficient of kurtosis*, measuring the peakedness of the distribution, is:

$$\eta_4 = \int_{-\infty}^{\infty} \left(\frac{x - \mu}{\sigma}\right)^4 f(x) dx \quad (\text{D.4.3-11})$$

These four parameters are properties of the unknown distribution, not of the data sample. However, the sample has its own set of corresponding parameters. For example, the *sample mean* is:

$$\bar{x} = \frac{1}{n} \sum_i x_i \quad (\text{D.4.3-12})$$

which is the average of the sample values. The *sample variance* is:

$$s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 \quad (\text{D.4.3-13})$$

while the sample skew and kurtosis are:

$$C_s = \frac{n}{(n-1)(n-2)s^3} \sum_i (x_i - \bar{x})^3 \quad (\text{D.4.3-14})$$

$$C_K = \frac{n(n+1)}{(n-1)(n-2)(n-3)s^4} \sum_i (x_i - \bar{x})^4 \quad (\text{D.4.3-15})$$

Note that in some literature the kurtosis is reduced by 3, so the kurtosis of the normal distribution becomes zero; it is then called the *excess kurtosis*.

#### D.4.3.3.4 Stationarity

Roughly speaking, a random process is said to be stationary if it is not changing over time, or if its statistical measures remain constant. Many statistical tests can be performed to help determine whether a record displays a significant trend that might indicate non-stationarity. A simple test that is very easily performed is the Spearman Rank Order Test. This is a non-parametric test operating on the ranks of the individual values sorted in both magnitude and time. The Spearman *R* statistic is defined as:

$$R = 1 - \frac{6 \sum_i (d_i)^2}{n(n^2 - 1)} \quad (\text{D.4.3-16})$$

in which *d* is the difference between the magnitude rank and the sequence rank of a given value. The statistical significance of *R* computed from Equation D.4.3-16 can be found in published tables of Spearman's *R* for *n* - 2 degrees of freedom.

#### D.4.3.3.5 Correlation Between Series

Two random variables may be statistically independent of one another, or some degree of interdependence may exist. Dependence means that knowing the value of one of the variables permits a degree of inference regarding the value of the other. Whether paired data  $(x,y)$ , such as simultaneous measurements of wave height and period, are interdependent or correlated is usually measured by their *linear correlation coefficient*:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} \quad (\text{D.4.3-17})$$

This correlation coefficient indicates the strength of the correlation. An  $r$  value of +1 or -1 indicates perfect correlation, so a cross-plot of  $y$  versus  $x$  would lie on a straight line with positive or negative slope, respectively. If the correlation coefficient is near zero, then such a plot would show random scatter with no apparent trend.

#### D.4.3.3.6 Convolution of Two Distributions

If a random variable,  $z$ , is the simple direct sum of the two random variables  $x$  and  $y$ , then the distribution of  $z$  is given by the convolution integral:

$$f_z(z) = \int_{-\infty}^{\infty} f_x(T) f_y(z-T) dT \quad (\text{D.4.3-18})$$

in which subscripts specify the appropriate distribution function. This equation can be used, for example, to determine the distribution of the sum of wind surge and tide under the assumptions that surge and tide are independent and they add linearly without any nonlinear hydrodynamic interaction.

#### D.4.3.3.7 Important Distributions

Many statistical distributions are used in engineering practice. Perhaps the most familiar is the *normal* or *Gaussian* distribution. We discuss only a small number of distributions, selected according to probable utility in an FIS. Although the normal distribution is the most familiar, the most fundamental is the uniform distribution.

##### D.4.3.3.7.1 Uniform Distribution

The uniform distribution is defined as constant over a range, and zero outside that range. If the range is from  $a$  to  $b$ , then the PDF is:

$$f(x) = \frac{1}{b-a}, \quad a \leq x < b, \quad 0 \text{ otherwise} \quad (\text{D.4.3-19})$$

which, within its range, is a constant independent of  $x$ ; this is also called a *top-hat* distribution.

The uniform distribution is especially important because it is used in drawing random samples from all other distributions. A random sample drawn from a given distribution can be obtained by first drawing a random sample from the uniform distribution defined over the range from 0 to 1. Set  $F(x)$  equal to this value, where  $F$  is the cumulative distribution to be sampled. The desired value of  $x$  is then obtained by inverting the expression for  $F$ .

Sampling from the uniform distribution is generally done with a random number generator returning values on the interval from 0 to 1. Most programming languages have such a function built in, as do many calculators. However, not all such standard routines are satisfactory. While adequate for drawing a small number of samples, many widely used standard routines fail statistical tests of uniformity. If an application requires a large number of samples, as might be the case when performing a large Monte Carlo simulation (see Subsection D.4.3.6.3), these simple standard routines may be inadequate. A good discussion of this matter, including lists of high-quality routines, can be found in the book *Numerical Recipes*, included in Subsection D.4.3.7, Additional Resources.

#### D.4.3.3.7.2 Normal or Gaussian Distribution

The normal or Gaussian distribution, sometimes called the bell-curve, has a special place among probability distributions. Consider a large number of large samples drawn from some unknown distribution. For each large sample, compute the sample mean. Then, the distribution of those means tends to follow the normal distribution, a consequence of the *central limit theorem*. Despite this, the normal distribution does not play a central role in hydrologic frequency analysis. The standard form of the normal distribution is:

$$f(x) = \frac{1}{\sigma(2\pi)^{1/2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$F(x) = \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)$$

(D.4.3-20)

#### D.4.3.3.7.3 Rayleigh Distribution

The Rayleigh distribution is important in the theory of random wind waves. Unlike many distributions, it has some basis in theory; Longuet-Higgins (1952) showed that with reasonable assumptions for a narrow banded wave spectrum, the distribution of wave height will be Rayleigh. The standard form of the distribution is:

$$f(x) = \frac{x}{b^2} e^{-\frac{x^2}{2b^2}}$$

$$F(x) = 1 - e^{-\frac{x^2}{2b^2}}$$

(D.4.3-21)

The range of  $x$  is positive, and the scale parameter  $b > 0$ . In water wave applications,  $2b^2$  equals the mean square wave height. The mean and variance of the distribution are given by:

$$\mu = b\sqrt{\frac{\pi}{2}} \tag{D.4.3-22}$$

$$\sigma^2 = b^2\left(2 - \frac{\pi}{2}\right)$$

The skew and kurtosis of the Rayleigh distribution are constants (approximately 0.63 and 3.25, respectively) but are of little interest in applications here.

#### D.4.3.3.7.4 Extreme Value Distributions

Many distributions are in common use in engineering applications. For example, the log-Pearson Type III distribution is widely used in hydrology to describe the statistics of precipitation and stream flow. For many such distributions, there is no underlying justification for use other than flexibility in mimicking the shapes of empirical distributions. However, there is a particular family of distributions that are recognized as most appropriate for extreme value analyses, and that have some theoretical justification. These are the so-called *extreme value distributions*.

Among the well-known extreme value distributions are the *Gumbel* distribution and the *Weibull* distribution. Both of these are candidates for FIS applications, and have been widely used with success in similar applications. Significantly, these distributions are subsumed under a more general distribution, the GEV distribution, given by:

$$f(x) = \frac{1}{b} \left\{ 1 + c \left( \frac{x-a}{b} \right) \right\}^{\frac{1}{c}-1} e^{-(1+c(x-a)/b)^{-1/c}}$$

for  $-\infty < x \leq a - \frac{b}{c}$  with  $c < 0$  and  $a - \frac{b}{c} \leq x < \infty$  with  $c > 0$

$$f(x) = \frac{1}{b} e^{-e^{-(x-a)/b}} e^{-(x-a)/b} \quad \text{for } -\infty \leq x < \infty \quad \text{with } c = 0 \tag{D.4.3-23}$$

The cumulative distribution is given by the expressions:

$$F(x) = e^{-(1+c(x-a)/b)^{-1/c}}$$

for  $-\infty < x \leq a - \frac{b}{c}$  with  $c < 0$  and  $a - \frac{b}{c} \leq x < \infty$  with  $c > 0$

$$F(x) = e^{-e^{-(x-a)/b}} \quad -\infty \leq x < \infty \quad \text{with } c = 0 \tag{D.4.3-24}$$

In these expressions,  $a$ ,  $b$ , and  $c$  are the *location*, *scale*, and *shape* factors, respectively. This distribution includes the *Frechet* (Type 2) distribution for  $c > 0$  and the *Weibull* (Type 3)

distribution for  $c < 0$ . If the limit of the exponent of the exponential in the first forms of these distributions is taken as  $c$  goes to 0, then the simpler second forms are obtained, corresponding to the Gumbel (Type 1) distribution. Note that the Rayleigh distribution is a special case of the Weibull distribution, and so is also encompassed by the GEV distribution.

The special significance of the members of the extreme value family is that they describe the distributions of the extremes drawn from other distributions. That is, given a large number of samples drawn from an unknown distribution, the extremes of those samples tend to follow one of the three types of extreme value distributions, all incorporated in the GEV distribution. This is analogous to the important property of the normal distribution that the means of samples drawn from other distributions tend to follow the normal distribution. If a year of water levels is considered to be a sample, then the annual maximum, as the largest value in the sample, may tend to be distributed according to the statistics of extremes.

#### D.4.3.3.7.5 Pareto Distribution

If for some unknown distribution the sample extremes are distributed according to the GEV distribution, then the set of sample values exceeding some high threshold tends to follow the Pareto distribution. Consequently, the GEV and Pareto distributions are closely related in a dual manner. The Pareto distribution is given by:

$$F(y) = 1 - \left(1 + \frac{cy}{\tilde{b}}\right)^{-1/c} \quad \text{for } y = x - u$$

$$\text{with } \tilde{b} = b + (u - a) \quad \text{(D.4.3-25)}$$

where  $u$  is the selected threshold. In the limit as  $c$  goes to zero, this reduces to the simple expression:

$$F(y) = 1 - e^{-y/\tilde{b}} \quad \text{for } y > 0 \quad \text{(D.4.3-26)}$$

#### D.4.3.4 Data Sample and Estimation of Parameters

Knowing the distribution that describes the random process, one can directly evaluate its inverse to give an estimate of the variable at any recurrence rate; that is, at any value of  $1-F$ . If the sample consists of annual maxima (see the discussion in Subsection D.4.3.5), then the 1% annual chance value of the variable is that value for which  $F$  equals 0.99, and similarly for other recurrence intervals. To specify the distribution, two things are needed. First, an appropriate form of the distribution must be selected from among the large number of candidate forms found in wide use. Second, each such distribution contains a number of free parameters (generally from one to five, with most common distributions having two or three parameters) that must be determined.

It is recommended that the Mapping Partner adopt the GEV distribution for FIS applications for reasons outlined earlier: extremes drawn from other distributions (including the unknown parent distributions of flood processes) may be best represented by one member of the extreme value

distribution family or another. The remaining problem, then, is determination of the three free parameters of the GEV distribution,  $a$ ,  $b$ , and  $c$ .

Several methods of estimating the best values of these parameters have been widely used, including, most frequently, the methods of plotting positions, moments, and maximum likelihood. The methods discussed here are limited to point-site estimates. If statistically similar data are available from other sites, then it may be possible to improve the parameter estimate through the method of *regional frequency analysis*; see Hosking and Wallis (1997) for information on this method. Note that this sense of the word *regional* is unrelated to what is meant by *regional studies* discussed elsewhere in these guidelines.

#### D.4.3.4.1 Plotting Positions

Widely used in older hydrologic applications, the method of plotting positions is based on first creating a visualization of the sample distribution and then performing a curve-fit between the chosen distribution and the sample. However, the sample consists only of the process variable; there are no associated quantiles, and so it is not clear how a plot of the sample distribution is to be constructed. The simplest approach is to rank order the sample values from smallest to largest, and to assume that the value of  $F$  appropriate to a value is equal to its fractional position in this ranked list,  $R/N$  where  $R$  is the value's rank from 1 to  $N$ . Then, the smallest observation is assigned *plotting position*  $1/N$  and the largest is assigned  $N/N=1$ . This is clearly unsatisfactory at the upper end because instances larger than the largest observed in the sample can occur. A more satisfactory, and widely used, plotting position expression is  $R/(N+1)$ , which leaves some room above the largest observation for still larger elevations. A number of such plotting position formulas are encountered in practice, most involving the addition of constants to the numerator and denominator,  $(R+a)/(N+b)$ , in an effort to produce improved estimates at the tails of the distributions.

Given a plot produced in this way, one might simply draw a smooth curve through the points, and usually extend it to the recurrence intervals of interest. This constitutes an entirely empirical approach and is sometimes made easier by constructing the plot using a transformed scale for the cumulative frequency. The simplest such transformation is to plot the logarithm of the cumulative frequency, which flattens the curve and makes extrapolation easier.

A second approach would be to choose a distribution type, and adjust its free parameters, so a plot of the distribution matches the plot of the sample. This is commonly done by least squares fitting. Fitting by eye is also possible if an appropriate *probability paper* is adopted, on which the transformed axis is not logarithmic, but is transformed in such a way that the corresponding distribution plots as a straight line; however, this cannot be done for all distributions.

These simple methods based on plotting positions, although widely used, are not recommended. Two fundamental problems with the methods are seldom addressed. First, it is inherent in the methods that each of  $N$  quantile bins of the distribution is occupied by one and only one sample point, an extremely unlikely outcome. Second, when a least squares fit is made for an analytical distribution form, the error being minimized is taken as the difference between the sample value and the distribution value, whereas the true error is not in the value but in its frequency position.

#### D.4.3.4.2 Method of Moments: Conventional Moments

An alternate method that does not rely upon visualization of the empirical distribution is the method of moments, of which there are several forms. This is an extremely simple method that generally performs well. The methodology is to equate the sample moments and the distribution moments, and to solve the resulting equations for the distribution parameters. That is, the sample moments are simple functions of the sample points, as defined earlier. Similarly, it may be possible to express the corresponding moments of an analytical distribution as functions of the several parameters of the distribution. If this can be done, then those parameters can be obtained by equating the expressions to the sample values.

#### D.4.3.4.3 Method of Moments: Probability-weighted Moments and Linear Moments

Ramified versions of the method of moments overcome certain difficulties inherent in conventional methods of moments. For example, simple moments may not exist for a given distribution form or may not exist for all values of the parameters. Higher sample moments cannot adopt the full range of possible values; for example, the sample kurtosis is constrained algebraically by the sample size.

Alternate moment-based approaches have been developed including probability-weighted moments and the newer method of linear moments, or L-moments. L-moments consist of simple linear combinations of the sample values that convey the same information as true moments: location, scale, shape, and so forth. However, being linear combinations rather than powers, they have certain desirable properties that make them preferable to normal moments. The theory of L-moments and their application to frequency analysis has been developed by Hosking; see, for example, Hosking and Wallis (1997).

#### D.4.3.4.4 Maximum Likelihood Method

A method based on an entirely different idea is the method of maximum likelihood. Consider an observation,  $x$ , obtained from the density distribution  $f(x)$ . The probability of obtaining a value close to  $x$ , say within the small range  $dx$  around  $x$ , is  $f(x) dx$ , which is proportional to  $f(x)$ . Then, the *posterior* probability of having obtained the entire sample of  $N$  points is assumed to be proportional to the product of the individual probabilities estimated in this way, in consequence of Equation D.4.3-5. This product is called the *likelihood* of the sample, given the assumed distribution:

$$L = \prod_1^N f(x_i) \quad (\text{D.4.3-27})$$

It is more common to work with the logarithm of this equation, which is the *log-likelihood*,  $LL$ , given by:

$$LL = \sum_1^N \log f(x_i) \quad (\text{D.4.3-28})$$

The simple idea of the maximum likelihood method is to determine the distribution parameters that maximize the likelihood of the given sample. Because the logarithm is a monotonic function, this is equivalent to maximizing the log-likelihood. Note that because  $f(x)$  is always less than one, all terms of the sum for  $LL$  are negative; consequently, larger log-likelihoods are associated with smaller numerical values.

Because maximum likelihood estimates generally show less bias than other methods, they are preferred. However, they usually require iterative calculations to locate the optimum parameters, and a maximum likelihood estimate may not exist for all distributions or for all values of the parameters. If the Mapping Partner considers alternate distributions or fitting methods, the likelihood of each fit can still be computed using the equations given above even if the fit was not determined using the maximum likelihood method. The distribution with the greatest likelihood of having produced the sample should be chosen.

#### **D.4.3.5 Extreme Value Analysis in an FIS**

For FIS extreme value analysis, the Mapping Partner may adopt the annual maxima of the data series (runup, SWL, and so forth) as the appropriate data sample, and then fit the GEV distribution to the data sample using the method of maximum likelihood. Also acceptable is the peak-over-threshold (POT) approach, fitting all observations that exceed an appropriately high threshold to the generalized Pareto distribution. The POT approach is generally more complex than the annual maxima approach, and need only be considered if the Mapping Partner believes that the annual series does not adequately characterize the process statistics. Further discussion of the POT approach can be found in references such as Coles (2001). The Mapping Partner can also consider distributions other than the GEV for use with the annual series. However, the final distribution selected to estimate the 1% annual chance flood level should be based on the total estimated likelihood of the sample. In the event that methods involve different numbers of points (e.g., POT vs. annual maxima), the comparison should be made on the basis of average likelihood per sample point because larger samples will always have lower likelihood function values.

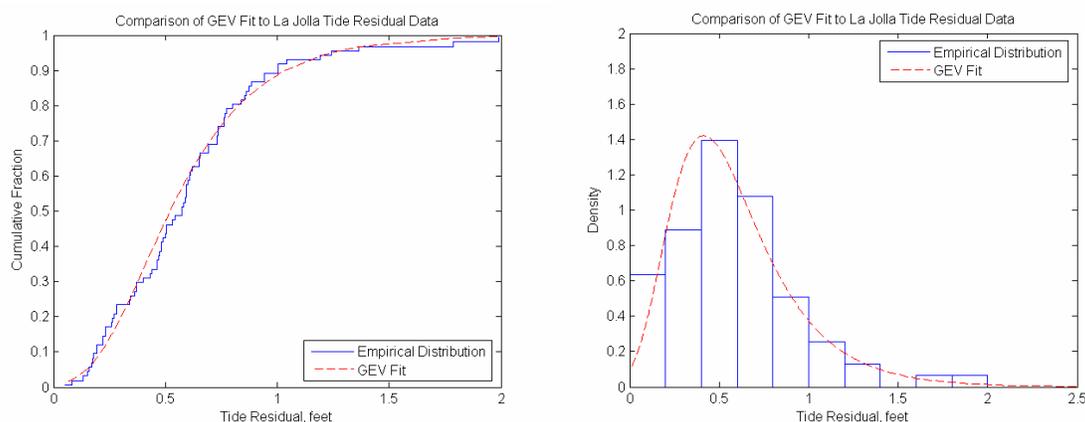
As an example of this process, consider extraction of a surge estimate from tide data. As discussed in Section D.4.4, the tide record includes both the astronomic component and a number of other components such as storm surge. For this example, all available hourly tide observations for the tide gage at La Jolla, California, were obtained from the National Oceanic and Atmospheric Administration (NOAA) tide data website. These observations cover the years from 1924 to the present. To work with full-year data sets, the period from 1924 to 2003 was chosen for analysis.

The corresponding hourly tide predictions were also obtained. These predictions represent only the astronomic component of the observations based on summation of the 37 local tidal constituents, so departure of the observations from the predictions represents the anomaly or residual. A simple utility program was written to determine the difference between corresponding high waters (observed minus predicted) and to extract the maximum such difference found in each year. Only levels at corresponding peaks should be considered in the analysis because small-phase displacements between the predicted and observed data will cause spurious apparent amplitude differences.

The resulting data array consisted of 80 annual maxima. Inspection of this file showed that the values were generally consistent except for the 1924 entry, which had a peak anomaly of over 12 feet. Inspection of the file of observed data showed that a large portion of the file was incorrect, with reported observations consistently above 15 feet for long periods. Although the NOAA file structure includes flags intended to indicate data outside the expected range, these points were not flagged. Nevertheless they were clearly incorrect, and so were eliminated from consideration. The abridged file for 1924 was judged to be too short to be reliable, and so the entire year was eliminated from further consideration.

Data inspection is critical for any such frequency analysis. Data are often corrupted in subtle ways, and missing values are common. Years with missing data may be acceptable if the fraction of missing data is not excessive, say not greater than one quarter of the record, and if there is no reason to believe that the missing data are missing precisely because of the occurrence of an extreme event, which is not an uncommon situation. Gages may fail during extreme conditions and the remaining data may not be representative and so should be discarded, truncating the total period.

The remaining 79 data points in the La Jolla sample were used to fit the parameters of a GEV distribution using the maximum likelihood method. The results of the fit are shown in Figure D.4.3-1 for the cumulative and the density distributions. Also shown are the empirical sample CDF, displayed according to a plotting position formula, and the sample density histogram. Neither of these empirical curves was used in the analysis; they are shown only to provide a qualitative idea of the goodness-of-fit.



**Figure D.4.3-1. Cumulative and Density Distributions for the La Jolla Tide Residual**

The GEV estimate of the 1% annual chance residual for this example was 1.74 feet with a log-likelihood of -19.7. The estimate includes the contributions from all non-astronomic processes, including wind and barometric surge, El Niño superelevation, and wave setup to the degree that it might be incorporated in the record at the gage location. Owing to the open ocean location of the gage, rainfall runoff is not a contributor in this case. *Note that this example is shown for descriptive purposes only, and is not to be interpreted as a definitive estimate of the tide residual statistics for this location for use in any application.* In particular, the predictions were

obtained from the NOAA website and so were made using the currently-adopted values of the tidal constituents. While this may be acceptable for an open ocean site such as La Jolla where dredging, silting, construction, and such are not likely to have caused the local tide behavior to change significantly over time, this may not be the case for other sites; the residual data should be estimated using the best estimates of the past astronomic components. *Nevertheless, this example illustrates the recommended general procedure for performing an extremal analysis using annual maximum observations, the GEV distribution, and the method of maximum likelihood.*

#### **D.4.3.6 Simulation Methods**

In some cases, flood levels must be determined by numerical modeling of the physical processes, simulating a number of storms or a long period of record, and then deriving flood statistics from that simulation. Flood statistics have been derived using simulation methods in FIS using four methods. Three of these methods involve storm parameterization and random selection: the Joint Probability Method (JPM), the Empirical Simulation Technique (EST), and the Monte Carlo method. These methods are described briefly below. In addition, a direct simulation method may be used in some cases. This method requires the availability of a long, continuous record describing the forcing functions needed by the model (such as wind speed and direction in the case of surge simulation using the one-dimensional [1-D] BATHYS model). The model is used to simulate the entire record, and flood statistics are derived in the manner described previously.

##### **D.4.3.6.1 JPM**

JPM has been applied to flood studies in two distinct forms. First, as discussed in a supporting case study document (PWA, 2004), joint probability has been used in the context of an event selection approach to flood analysis. In this form, JPM refers to the joint probability of the parameters that define a particular event, for example, the joint probability of wave height and water level. In this approach, one seeks to select a small number of such events thought to produce flooding approximating the 1% annual chance level. This method usually requires a great deal of engineering judgment, and should only be used with permission of the Federal Emergency Management Agency (FEMA) study representative.

FEMA has adopted a second sort of JPM approach for hurricane surge modeling on the Atlantic and Gulf coasts, which is generally acceptable for any site or process for which the forcing function can be parameterized by a small number of variables (such as storm size, intensity, and kinematics). If this can be done, one estimates cumulative probability distribution functions for each of the several parameters using storm data obtained from a sample region surrounding the study site. Each of these distributions is approximated by a small number of discrete values, and all combinations of these discrete parameter values representing all possible storms are simulated with the chosen model. The rate of occurrence of each storm simulated in this way is the total rate of storm occurrence at the site, estimated from the record, multiplied by each of the discrete parameter probabilities. If the parameters are not independent, then a suitable computational adjustment must be made to account for this dependence.

The peak flood elevations for each storm are saved for subsequent determination of the flood statistics. This is done by establishing a histogram for each point at which data have been saved,

using a small bin size of, say, about 0.1 foot. The rate contribution of each storm, determined as described above, is summed into the appropriate elevation bin at each site. When this is done for all storms, the result is that the histograms approximate the density function of flood elevation at that site. The cumulative distribution is obtained by summing across the histogram from top down; the 1% elevation is found at the point where this sum equals 0.01. Full details of this procedure are provided in the user's manual accompanying the FEMA storm surge model (FEMA, 1987).

#### **D.4.3.6.2 EST**

The U.S. Army Corps of Engineers has developed a newer technique, EST, that FEMA approved for FIS; a full discussion can be found in Scheffner et al. (1999). The technique is based on bootstrap resampling-with-replacement, random-neighbor walk, and subsequent smoothing techniques in which a random sampling of a finite length historical-event database is used to generate a larger long-period database. The only assumption is that future events will be statistically similar in magnitude and frequency to past events.

The EST begins with an analysis of historical storms that have affected the study area. The selected events are then parameterized to define relevant input parameters that are used to define the dynamics of the storms (the components of the so-called input vectors) and factors that may contribute to the total response of the storm such as tidal amplitude and phase. Associated with the storms are the response vectors that define the storm-generated effects. Input vectors are sets of selected parameters that define the total storm; response vectors are sets of values that summarize the effects. Basic response vectors are determined numerically by simulating the historical storms using the selected hydrodynamic model.

These sets of input and response vectors are used subsequently as the basis for the long-term surge history estimations. These are made by repeatedly sampling the space spanned by the input vectors in a random fashion and estimating the corresponding response vectors. By repeating this step many thousands of times, an extremely long period of simulated record is obtained. The final step of the procedure is to extract statistics from the simulated long record by performing an extremal analysis on the simulated record as though it were a physical record.

#### **D.4.3.6.3 Monte Carlo Method**

As discussed above for the JPM approach, the Monte Carlo method is based on probability distributions established for the parameters needed to characterize a storm. Unlike JPM, however, these probability distributions are not discretized. Instead, storms are constructed by randomly choosing a value for each parameter by generating a random value uniformly distributed between 0 and 1, and then entering the cumulative distribution at this value and selecting the corresponding parameter value. Each storm selected by this Monte Carlo procedure is simulated with the hydrodynamic model, and shoreline elevations are recorded. Simulating a large number of storms in this way is equivalent to simulating a long period of history, with the frequency connection established through the rate of storm occurrence estimated from a local storm sample. The Monte Carlo method has been used extensively in concert with a 1D surge model by the State of Florida to determine coastal flood levels; see the 1D surge discussion in Section D.4.4 for additional information. An example of a Monte Carlo approach to the statistics of

wave runup in sheltered waters is shown in a supporting case study document (PWA, 2004). In that example, distributions were first developed for the forcing functions (winds, waves, water levels), and from them a long simulated time series was derived by Monte Carlo random sampling. Another example of a Monte Carlo analysis is shown in the context of event-based erosion in Section D.4.6; in that example, a Monte Carlo approach was used to relate bluff failure to bluff toe erosion.

#### **D.4.3.7 Additional Resources**

The foregoing discussion has been necessarily brief; however, the Mapping Partner may consult a large amount of literature on probability, statistics, and statistical hydrology. Most elementary hydrology textbooks provide a good introduction. For additional guidance, the following works might also be consulted:

##### **Probability Theory:**

*An Introduction to Probability Theory and Its Applications, Third Edition*, William Feller, 1968 (two volumes). This is a classic reference for probability theory, with a large number of examples drawn from science and engineering.

*The Art of Probability for Scientists and Engineers*, Richard Hamming, 1991. Less comprehensive than Feller, but provides clear insight into the conceptual basis of probability theory.

##### **Statistical Distributions:**

*Statistics of Extremes*, E.J. Gumbel, 1958. A cornerstone reference for the theory of extreme value distributions.

*Extreme Value Distributions, Theory and Applications*, Samuel Kotz and Saralees Nadarajah, 2000. A more modern and exhaustive exposition.

*Statistical Distributions, Second Edition*, Merran Evans, Nicholas Hastings, and Brian Peacock, 1993. A useful compendium of distributions, but lacking discussion of applications; a formulary.

*An Introduction to Statistical Modeling of Extreme Values*, Stuart Coles, 2001. A practical exposition of the art of modeling extremes, including numerous examples. Provides a good discussion of POT methods that can be consulted to supplement the annual maxima method.

##### **Statistical Hydrology:**

*Applied Hydrology*, Ven Te Chow, David Maidment, and Larry Hays, 1988. One of several standard texts with excellent chapters on hydrologic statistics and frequency analysis.

*Probability and Statistics in Hydrology*, Vujica Yevjevich, 1972. A specialized text with a lot of pertinent information for hydrologic applications.

##### **General:**

*Numerical Recipes, Second Edition*, William Press, Saul Teukolsky, William Vetterling, and Brian Flannery, 1992. A valuable and wide ranging survey of numerical methods and the ideas behind them. Excellent discussions of random numbers, the statistical description of data, and

modeling of data, among much else. Includes well-crafted program subroutines; the book is available in separate editions presenting routines in FORTRAN and C/C++.

**Software:**

Several open-source and commercial software packages provide tools to assist in the sorts of analyses discussed in this section. In particular, the S, S-PLUS, and R programming languages (commercial and open-source versions of a high-level statistical programming language) include comprehensive statistical tools. The R language package is available for free from the web site <http://www.r-project.org/>; several books discussing the use of R and S are available. Other well-known software packages include *Mathematica*, *Matlab*, *SPSS*, and *SYSTAT*.