WHAT IS THE PROBABILITY FUNCTION FOR LARGE TSUNAMI WAVES?

Harold G. Loomis
Honolulu, HI

ABSTRACT

Most coastal locations have few if any records of tsunami wave heights obtained over various time periods. Still one sees reference to the 100-year and 500-year tsunamis. In fact, in the USA, FEMA requires that at all coastal regions, those wave heights due to tsunamis and hurricanes be specified. The same is required for stream flooding at any location where stream flooding is possible. How are the 100 and 500-year tsunami wave and stream flooding heights predicted and how defensible are they? This paper discusses these questions.

Science of Tsunami Hazards, Vol. 24, No. 3, page 218 (2006)

PROBABILITY FUNCTION

The theory of probabilities for extreme events is a well developed subject and is routinely applied to: stream and river flooding, wind pressure, minimum and maximum rainfall, life expectancies, breaking of cables and fasteners and more. The theory and many applications are described in the book by Gumbel[1]. This is an advanced statistics book with lots of definitions and mathematics from which I am extracting a small part of the theory for application to tsunami wave heights. Following Gumbel, I use *f(x)* as the probability density function and *F(x)* as the probability distribution function which Gumbel calls simply the probability function, and I will use the same language. In other words,

$$F(x) = prob(X \le x)$$

Where X is the random variable, i.e. the result of an experiment or a measurement, and

$$F'(x) = f(x).$$

Let's assume that at a given location there actually is some probability function for wave heights of a series of tsunamis over time, and we want to determine what that probability function is and its parameters. At a given location it is assumed that each maximum wave height for a tsunami event is a realization of a random variable and that all of these random variables are drawn from the same probability function $F(x)$. It is this $F(x)$ that we want to determine. If the collection of n wave heights is arranged according to size, then the variable at each location in the sample has its own probability function, and it is not the same as the overall probability function from which the sample is drawn. This subject is called order statistics. We are interested in the probability function, $F_n(x)$, for the random variable X_n, the largest wave height in the sample. If someone were interested in minimum rainfall or minimum breaking strength, they would be interested in $F_1(x)$, the probability function for the smallest value of the sample. In order for the largest member of the sample to be $\le x$ it is necessary that every member of the sample be $\le x$, so

$$F_n(x) = F^n(x).$$

Note that the probability function for the largest value in the sample is different from the probability functions of the individual variables in the sample. Here is where extreme value statistics get interesting. It turns out that there are only 3 asymptotic forms for these extreme value probabilities, depending on whether the probability density functions

[1] The book by Gumbel referenced at the end has a bibliography of papers and examples on this subject pretty complete up until the year 1958.

for the individual random variable goes to zero like e^{-x}, like x^{-k}, or is bounded in some way. Note that this is not a limit as $n \to \infty$, (n is a fixed number!) but rather an asymptotic approximation as $x \to \infty$, which is appropriate because it is for large values of x that we want the probability function. If one knew exactly the original probability function one could evaluate $F^n(x)$ for any given x and n. However, even not knowing the original probability function, but reasoning in some way that it should fall off exponentially, or like a power of x, or that the range of x is limited in some way, we can still arrive at the asymptotic probability function for the largest wave height in the sample. In fact, the number n is not required to be known either as will be shown later

Taking first the case where the initial probability function is exactly exponential, we have

$$f(x) = \alpha e^{-\alpha x},$$
$$F(x) = 1 - e^{-\alpha x}.$$

In this case the asymptotic probability function is given by

$$F^n(x) = \left(1 - e^{-\alpha x}\right)^n.$$

The asymptotic value of this expression is given by

$$F^n(x) = \exp\left(-\exp\left(-\alpha\left(x - b\right)\right)\right) \qquad (1)$$

Such a double exponential function is surprising. One would never guess it from physical principles, but the above function is derived logically which will be demonstrated. That this is so is similar to the well known fact that

$$\lim(1 - x/n)^n = e^{-x} \quad \text{as} \quad n \to \infty.$$

What follows is not a proof (which can be found in Gumbel, in fact two of them) but rather a simple demonstration that the double exponential is reasonable. First a useful value u_n, the characteristic largest value, is defined as the largest value one would expect in a sample of size n, namely, the value for which $F(u_n) = 1 - 1/n$. If this is in the range where the probability function is approximately exponential, then

$$F(u_n) = 1 - e^{-\alpha u_n} = 1 - 1/n$$

from which

$$(1/n)e^{\alpha u_n} = 1$$

.

Substituting this in the equation for $F^n(x)$, one has

$$F^n(x) = \left(1 - e^{-\alpha(x-u_n)} / n\right)^n$$

which gives the asymptotic expression (1) for $F^n(x)$.

A similar kind of argument works for the other two assumptions about the nature of the original $F(x)$ leading to moving the original probability function up to the exponential level. However, for tsunamis it seems like this first asymptotic expression is most reasonable, so we present only the first one.

So how does one find the 100-year and 500-year wave heights? You make an observational probability function out of the existing data, i.e. $X_1, X_2, ..., X_n$ are arranged in order of size and X_m is assigned the cumulative probability of *m/(n+1)* so that the plotting points are $\left(m/(n+1), X_m\right)$. Gumbel has several sections on the choice of what to use for plotting points and the one chosen seems to be the best decision. Then the chosen form of the true probability function is fitted to this observational probability function by adjusting α and *b*. α is like a scale factor and *b* is like the mean. These actually can be estimated from the data according to various statistical formulas, but since we are plotting the data anyway, it is easier to get them from the plot. Normally a change of variable is made so that the plot (if you are using the right probability function) can be fitted with a straight line. With a change of variable we have

$$-\ln(-\ln(F)) = \alpha(x - b)$$

and α is the slope of the line and b is its intercept. Once the line is plotted, one picks off the value of x where $F = .99$ and that is the height of the wave which will be exceeded with probability .01. If an event has probability p of occurring during each time unit, then $\widetilde{T}$, the average number of time units between such events will be 1/p. This is not a given but is the result of calculating the average return time. For this reason the probability paper usually has the probabilities scaled along the bottom axis and $\widetilde{T}$ scaled along the top. Similarly, for the 500-year wave, one picks off the wave height corresponding to $F = 1 - .002 = .998$. One can be suspicious of the 500-year wave prediction because one expects geological changes over that period of time, i.e. the sea level could rise significantly or a period of intense volcanic activity could occur. What the 100-year and 500-year predictions really mean is that given conditions as they are now, the first has a probability of .01/year and the second has a probability of .002/year.

Since the plotted data is scattered about a straight line (hopefully) it is obvious that there is uncertainty in drawing the line and thus in the predictions. Gumbel discusses this and in given instances shows how to calculate these uncertainties. Furthermore, in theory it is

possible that the largest value might lie above the line and might be larger that the wave height corresponding to $F = .01$. In other words, there may be a wave observed in a period of time shorter than 100years that actually exceeds the predicted 100-year wave.

There is a very useful added advantage if you have chosen the asymptotic probability function correctly then you can take the 50-year wave and scale it up to the 100-year wave by a simple arithmetic formula. That is you can scale from any time interval to any other time interval with this formula.

This asymptotic expression for waves with probability p can be used (approximately) to compare maximum wave heights for different time intervals. Suppose that $T_1 = 1/p_1$ and $T_2 = 1/p_2$, then

$$-\ln(-\ln(1-p_1)) = \alpha(x_1 - b)$$

and

$$-\ln(-\ln(1-p_2)) = \alpha(x_2 - b).$$

Making use of the series

$$\ln(1+x) = x + x^2/2 + x^3/3 + \ldots\ldots,$$

$$-\ln(1-p) = p - p^2/2 + p^3/3 - \ldots.$$

Since p is small, we can take only the first term of the series, so that the original equations can be written as

$$-\ln(p_1) = \ln(T_1) = \alpha(x_1 - b),$$

$$-\ln(p_2) = \ln(T_2) = \alpha(x_2 - b).$$

Subtracting the first from the second we have

$$\ln(T_2) - \ln(T_1) = \ln(T_2/T_1) = \alpha(x_2 - x_1),$$

so that

$$x_2 = x_1 + (1/\alpha)\ln(T_2/T_1).$$

This is very useful. If $T_2 = 2T_1$, then $x_2 = x_1 + (1/\alpha)\ln(2)$, or $x_2 = x_1 + (.693/\alpha)$.

My first reason for using the extreme value statistics (IUGG, Vancouver 1987)[2] was that it is widely used in many similar situations. Also, it seemed basically right because the asymptotic probability function was the right choice, given exponential fall off of the individual probabilities, no matter what the original probability function was. In continuing to reflect on the matter, some questions arise. First of all, what is the sample of wave heights from which the maximum is chosen? It could be the collection of wave heights in the near vicinity of the reported wave height which would surely be the largest.

I should point out that in the application of extreme value statistics to tsunamis, there is not enough data to really determine what probability function to use. The usual test is that the observed cumulative probability function will lie approximately on a straight line when plotted on the correct probability paper which is in effect, the choice of the correct probability function. However, since we have at most 5 values at any location (and many of those values are questionable) in Hawaii this is not a good test. Therefore the choice of the probability function will be mainly an exercise in logical reasoning.

How about augmenting the data with artificial values from imagined tsunamis? There are no probabilities connected with the imagined tsunamis so that doesn't expand the data for probability calculations. How about extending wave measurements of a given tsunami to places where no measurements were made by creating a numerical model of a given tsunami that agrees well at places where the tsunami was measured? This system, which was used for the FEMA maps has some validity. However, there still are too few data points to decide whether or not a straight line describes them well enough.

If I were to guess what the underlying probability function were for wave heights at any location, I would guess normal or Gaussian[3]. This is based on the Central Limit Theorem which says that a sum of random variables approaches the Gaussian whatever the probabilities of those random variables. In this case think of the many variables such as source size, location, mechanism, and all of the additional factors affecting runup size at any given shore location. Think of these as random variables. It seems that there are enough variables here to assume Gaussian for the total effect. Gumbel has a section in which he establishes that Gaussian qualifies as being essentially exponential so that the first extreme value probability function applies. (The conditions to qualify are actually broader than just falling off exponentially!)

Even if it is so that the probability function for wave heights at each point on the shoreline is Gaussian, the value reported and recorded should be treated as the 1st asymptotic probability function. The reason for this is that the wave height actually reported will be the largest of the wave heights from the immediate vicinity of that location.

---

[2] IUGG Tsunami Symposium, Vancouver, B.C., August 18-19, 1987

[3] We would be focusing on the larger tsunamis being fit to the upper end of the Gaussian probability function since it is probabilities of large tsunamis that we are looking for.

Given that the double exponential probability function for wave height is correct, there is another problem with the prediction of the tsunami wave height with return time 100 years. The following simple calculation will demonstrate the problem. Suppose one has estimated that the wave height h_1 is the height exceeded with probability .01, (or $F = .99$). In other words, *.01* is the probability that if a tsunami occurs, its size will exceed h_1. Suppose that on the average there are 5 significant tsunamis in 100 years. Then the probability that all 5 are less that h_1 is $(.99)^5 = .95$. A larger value h_2 must be found so that $F^5(h_2) = .99$, or $F = \sqrt[5]{.99} = .998$. This would, in fact, be the 500-year wave with probability .002 per tsunami. At the rate of 5 tsunamis/100 years, the probability of a tsunami exceeding h_2 would be $5x.002 = .01$, or on the average, once in 100 years.

The above suggests a scheme appropriate when tsunamis occur rather infrequently, say k per 100 years (based on experience.) Assume that the underlying probability function is the 1st asymptotic probability function. It is necessary to create the $-\ln(-\ln y).vs.x$ graph paper with your computer. The observed probability histogram points for the data from a given location are plotted. At this point you can pick off the values of α and *b* and solve for *x* for any value of *y* using the double exponential probability formula. Or graphically you can pick off the value of *x* for which $y = (.99)^{1/k}$ which gives the 100-year wave at that location.

How well will these methods predict the 100-year and 500-year waves? Unfortunately, or fortunately, we'll never know!

REFERENCES

Gumbel, E.J., Statistics of Extreme Values, Columbia University Press, 1957

Gumbel, E.J., Statistical Theory of Extreme Values and Some Practical Applications, National Bureau of Standards, Applied Math Series, No. 33