Data repository: Statistical methods

The following is written with reference to the Statistical Methods section, and explains more fully the terminology and methods used.

The survivor function, $S(x)$, gives the probability, P, of exceedance, such that

$$S(x) = P(X>x) = 1 - F(x) \qquad \text{(A)}$$

For a sample, $S(x)$ is calculated by rank ordering observations from $X_{i=0}$ to X_n, where n is the total number of observations, and $S(x_i) = (n-i)/n$. This may then be fitted to a continuous probability distribution survivor function (e.g., Weibull or log-logistic) and be used as a model for forecasting (e.g., Connor et al., 2003).

Probability plots provide a quick assessment of whether data may be described by a particular distribution. Plot axes are constructed by linearizing the cumulative distribution function, $F(x)$. If the data are described by the subject distribution they will plot a straight line, with slope a and intercept b. For the Weibull distribution, with repose times t, linearization results in the axes:

$$x = \ln(t) \text{ and } y = \ln(\ln(1/(1 - Q(t)))) \qquad \text{(B)}$$

where $Q(t)$ ($= F(x)$) is the unreliability, and is estimated by Benard's approximation to the median rank: $p = (i - 0.3)/(n + 0.4)$, where i is the observation rank. By definition b is the x-value at an unreliability of 0.632 (for $a = 1$ this is equal to the distribution mean).

Linearization of the log-logistic cumulative density function for probability plot construction, with slope a and intercept b, gives:

$$x = \ln(t) \text{ and } y = \ln[Q(t)/(1 - Q(t))] \qquad \text{(C)}$$

Such analysis can be undertaken using any plotting package, such as Excel.

With reference to Kolmogorov-Smirnov *P*-values (Massey, 1951): Since model parameters are estimated from the sample, our *P*-values are conservative; only a high value (perhaps $P>0.9$, for our n) suggests a model is suitable for describing the sample distribution. The Kolmogorov-Smirnov test does not assume a distribution, and assesses goodness of fit from the maximum deviation between the cumulative distribution function of the data and the model distribution.

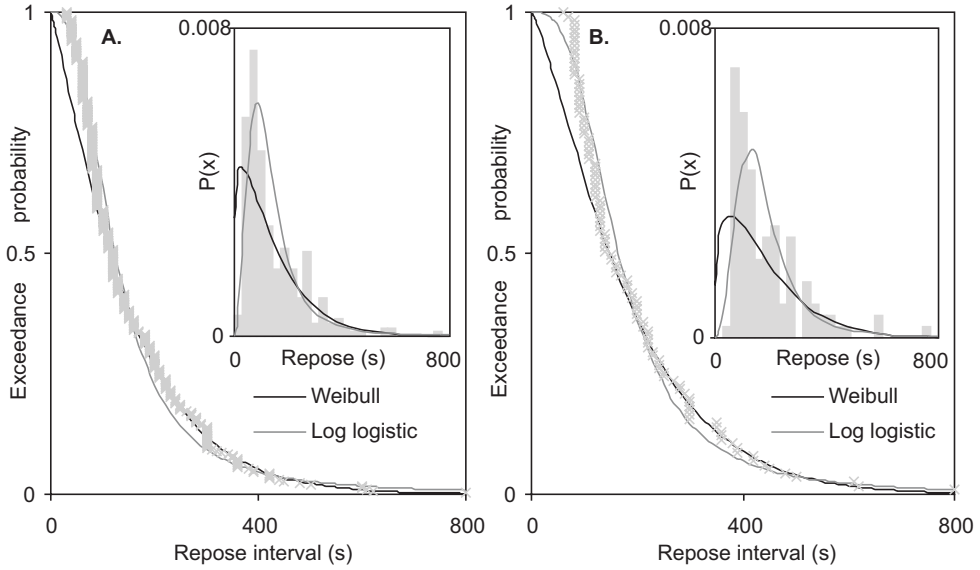
Data repository: Figures

Figure DR1. Survivor functions of best fitting Weibull and log-logistic models against explosion repose interval data from Anak Krakatau. A: 12–13/1/1960; B: 12/1/1960. Insets show histograms of data with the corresponding Weibull and log-logistic probability density functions.

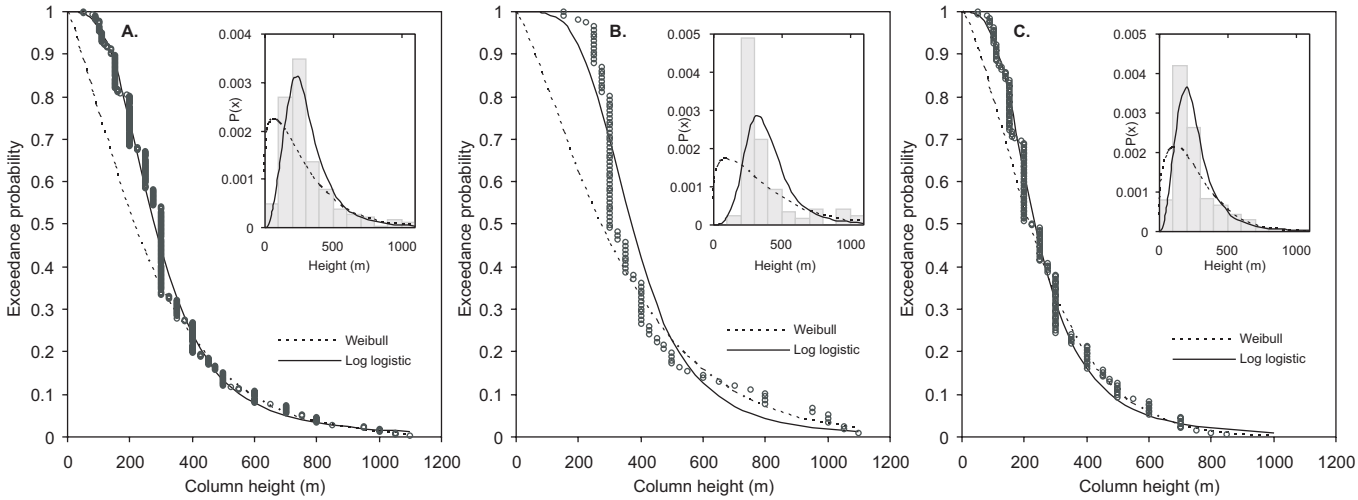
Figure DR2. Survivor functions of best fitting Weibull and log-logistic models against column height data from Anak Krakatau. A: 12–13/1/1960; B: 12/1/1960; C: 13/1/1960. Insets show histograms of data with the corresponding Weibull and log-logistic probability density functions.

Figure DR3. Cumulative distribution functions showing showing explosion data from the Kameni Islands against best-fitting Weibull and log-logistic curves. A: 12/02–25/3/1866, ash explosion frequency; B: 12/02–25/3/1866, steam explosion frequency; C: 12/02–25/3/1866, detonation frequency (Schmidt, 1872); D: 14/08–16/09/1925, explosion

frequency, during the initial paroxysmal stage of the 1925–2928 eruption; E: 23/09–14/10/1939, initial vigorous stage, explosion frequency at the Kténas dome; F: 15/10–24/11/1939, subsequent quieter stage, explosion frequency at the Kténas dome; G: 22/02–23/02/1926, explosion repose intervals during a late eruption period of reduced activity.


A.
B.
Exceedance probability
Repose interval (s)
P(x)
Repose (s)
Weibull
Log logistic
0
0.5
1
400
800
0.008

DR2007203 Sebastian Watt, Figure DR2 (figDR2.cdr)



Sebastian Watt, Figure DR3 (figDR3.cdr)