Plotting positions via maximum-likelihood for a non-standard situation

D. A. Jones

Hydrology and Earth System Sciences, 1(2), *357–366* (1997) © EGS

Hydrology & Earth System Sciences

Plotting positions via maximum-likelihood for a non-standard situation

D.A. Jones

Institute of Hydrology, Wallingford, Oxfordshire, OX10 8BB, UK

Abstract

A new approach is developed for the specification of the plotting positions used in the frequency analysis of extreme flows, rainfalls or similar data. The approach is based on the concept of maximum likelihood estimation and it is applied here to provide plotting positions for a range of problems which concern non-standard versions of annual-maximum data. This range covers the inclusion of incomplete years of data and also the treatment of cases involving regional maxima, where the number of sites considered varies from year to year. These problems, together with a not-to-be-recommended approach to using historical information, can be treated as special cases of a non-standard situation in which observations arise from different statistical distributions which vary in a simple, known, way.

Introduction

The uses of plotting positions in the context of frequency analysis of annual maximum flow and rainfall data are well-known: see, for example, NERC (1975, Section 1.3) and Stedinger *et al.* (1993, Section 18.3). In the standard situation, where the assumption of statistically independent and identically distributed observations can be made, plotting positions can be used in four ways:

(a) informal, graphical estimation of a distribution function;
(b) formal fitting of a parametric distribution function;
(c) informal assessment of how well a fitted distribution matches the data;
(d) formal testing of the fit of a given parametric family of distributions.

The assumptions involved in the standard situation are plausible enough to cover a wide range of applications of plotting positions in hydrology. However, there are a number of potential applications for which the assumptions are no longer plausible, but for which it would be nice to be able to derive the equivalents of plotting positions which would then be potentially available for the above uses.

This paper develops a way of calculating plotting positions which are appropriate for situations in which the observations, while still independent, are no longer identically distributed. Such non-standard situations arise in a number a practical problems and two simple examples are outlined in the next paragraphs: discussion of these example is continued later. It will be obvious that, for a plotting-position procedure to be meaningful for non-identically distributed observations, the assumptions adopted must be such as to provide both a strong connection between the distributions of individual observations and a clear definition of the specific distribution to which the plotting positions are to relate and, of course, they must be realistic for the situation being modelled.

EXAMPLES OF NON-STANDARD SITUATIONS

Consider the analysis of annual maximum floods or rainfalls in a case where records for some years are incomplete. A standard approach is to discard the records for such years unless data are available for 'most' of the year. In the latter case, the incompleteness of the year is effectively ignored: the maximum value found in the available record for that year is treated exactly as if the record were complete. It may be possible to argue that the maximum value observed in the available portion of an incomplete year of record has distribution function F^s, where F is the distribution function for maxima in complete years and where s is the fraction of the year covered by the record. One could argue for this on two bases. Firstly, one could say that, since the distribution functions of maxima over 1, 2, 3, . . . years are F, F^2, F^3, . . ., the distribution for fractional-record years should be of the same form. Secondly, one could start from a 'peaks-over-threshold' approach to deriving the distribution of annual maxima, which yields the form F^s for fractional years if the assumption of non-seasonality is made. It also yields the same form, with a version of s based on relative rates of occurrence of events,

if seasonality is allowed in the event-occurrence process but not for the event-sizes.

Thus, in the case of annual maxima for possibly incomplete years, an appropriate model might be to say that a given observation arises from distribution function F^s, where s is a known value measuring completeness of the data-record for the given year and where $s = 1$ for a complete year. The problem is to assign to the observations a set of plotting positions appropriate to the underlying distribution function, F, which in this case is the distribution of annual maxima for complete years. A similar form of model is also appropriate for the problem of treating regional-maxima. Here an annual observation would be the largest daily or instantaneous value observed at any site in a region. Since the number of sites within a region varies from year to year, the observations are not identically distributed but, on the basis of empirical studies reported elsewhere (Dales and Reed, 1989), the distribution function of a regional-maximum can be related to the distribution function of annual maxima at a single site via a power-relationship. In this case, the power, s, would be a value depending on the number and density of the sites available, with $s = 1$ for a single-site region so that the underlying distribution being estimated by the procedures here would be the distribution of single-site annual maxima.

Plotting Positions

The chapter by Stedinger *et al.* (1993) provides a detailed review of the frequency analysis of extreme events for hydrological applications and, since this includes coverage of work on plotting positions, there seems little point in duplicating this effort here. For practical application in standard situations, no improvements have been found over the simple plotting position formulae recommended by Cunnane (1978).

For present purposes, the question of defining plotting positions will be taken to mean providing, for each $x(r)$ in a set of ranked observations $\{x(i); i = 1, \ldots, N\}$, a value p_r which estimates the value of the 'underlying' distribution function F at $x(r)$. That is,

$$p_r \approx F\{x(r)\}. \qquad (1)$$

This interpretation is one of the usual interpretations given to plotting positions. The alternative interpretation, not used in the derivations here, is that used to define 'unbiased' plotting positions: this requires that p_r should be the value of F evaluated at the mean or median value of $x(r)$ evaluated over the distribution of possible outcomes. This approach is difficult in the present, non-standard, case both because of the non-identically distributed observations and because it would be necessary to decide whether or not to work with distributions conditioned on the pattern found in the ordered list $\{x(i)\}$ for the different original distributions. In contrast, the approach via Eqn. (1) is straightforward.

In the standard situation, a plotting position formula of the general form

$$p_r = (r - a)/(N + 1 - 2a), \qquad (2)$$

is often used (Cunnane, 1978), where a is a constant determining which one of a variety of specific plotting position formulae is actually used. Values of a in the range 0.375 to 0.50 are usually recommended and individual countries or organisations typically have specific favoured values. For example, Cunnane (1978) suggests using the Gringorten plotting positions, which result from choosing $a = 0.44$, if the EV1 distribution is likely to be appropriate, and suggests $a = 0.40$ as a simple, general purpose choice. Note that, in a recent review from a statistical viewpoint (Hyndman and Fan, 1996), the value $a = \frac{1}{3}$ is suggested. The plotting positions derived in this paper for the non-standard case can not be expressed in a simple explicit formula similar to Eqn. (2), but instead require the iterative solution to a simple non-linear equation.

Even in the standard case, there are no straightforward theoretically-based procedures for deriving plotting positions. Theoretical analysis via the unbiased plotting position approach indicates that the 'best' plotting positions depend upon the form of the common or underlying distribution function F. Such results are thus of little use in deriving plotting positions when, as is usual, the form of the distribution function is unknown and, indeed, is the subject of the investigations in which they are to be used. Guo (1990a) discusses plotting positions specifically targeted at the General Extreme Value distribution, and concludes that the best results are obtained with the simple, non-distribution-specific, formula in Eqn. (2).

For the non-standard case dealt with here, an approach based on maximum-likelihood theory has been adopted. While this theory is straightforward in itself, it does not provide a direct route to an estimate of the required quantity, as expressed by Eqn. (1): rather, a somewhat arbitrary modification of the approach has to be made involving the weighting of estimates of F just below and just above each observation. It turns out that weights can be chosen in such a way as to yield identical results to Eqn. (2) when the approach is applied in the standard case. In the absence of alternative suggestions, it seems reasonable to adopt the same type of weights in deriving plotting positions for the non-standard situation on the basis that the overall procedure will then produce the accepted plotting positions when applied in the standard situation. The new procedure, while not yielding a simple expression directly comparable to Eqn. (2), retains the desirable feature that no assumption is needed about the form of the underlying distribution.

One interpretation of the procedure being suggested is that 'maximum-likelihood' is used to guide the combination of the information provided by the original, unordered, observations which arise from different-but-related distributions, while the use of weights is a minor

but important feature which allows the usual plotting positions to be recovered in the standard situation. It should be recalled that the 'good' properties of the standard plotting position formulae in Eqn. (2) are very often assessed in a context which emphasises the 'unbiased' approach to interpreting plotting positions. This is not necessarily a contradiction to the discussion above: any seemingly negative comments about requirements of unbiasedness for plotting positions relate only to their usefulness in deriving ways of calculating plotting positions for non-standard situations, not to assessing how they perform.

Model Formulation

ASSUMPTIONS

It is convenient to begin by outlining the statistical model being used. Some practical situations in which such a model may be appropriate have already been described, and some further discussion is given later. Note that this model is only one special case of a class of models for non-identically distributed observations. Let $\{X_i; i = 1, \ldots, N\}$ denote statistically independent random variables corresponding to a set of observations $\{x_i\}$, and let the observations be complemented by a set of known values or covariates $\{s_i\}$, so that the observed data set actually consists of the set of pairs:

$$D = \{ x_i, s_i ; i = 1, \ldots, N\}.$$

The random variables $\{X_i\}$ are assumed to be related to a common underlying distribution function F in the following way: the random variable $X = X_i$, with covariate $s = s_i$, has distribution function F^s (ie. the s'th power of F). In the applications here, the covariates used are proportional to an underlying notional sample size attributable to each observation, so that covariate s_i refers to a 'size' or 'size parameter' associated with observation x_i. The distribution function F is unknown, but the assumption is made that it has no discrete components. This means that, under the assumptions of the model, ties have zero probability. For practical purposes any ties in the x-values may be broken arbitrarily: however, the results for plotting positions would depend on how this is done, unless the size parameters are also identical. Suppose that the data-set of pairs, D, is rearranged so that the observations $\{x_i\}$ are in increasing order, giving the ordered data-set:

$$D_o = \{ x(i), s(i) ; i = 1, \ldots, N\}.$$

The aim of the procedure is to use the ordered data set D_o to calculate a set of values $\{p_i\}$ such that p_r estimates F, evaluated at $x(r)$. Thus the distribution being estimated corresponds directly to the value of the size parameter $s = 1$. A case in which all the size parameters are unity is the same as the standard situation of identically distributed observations. The discussion at the end of this section indicates a situation in which an alternative to the assumption that an observation has distribution function F^s arises in a natural way: still other forms of relationships between the distributions of observations are possible.

Application 1: Annual Maxima with Incomplete Records

The arguments presented in the initial description of this problem suggest that the proposed method should be useful in taking account of incomplete years of record in frequency analyses of annual maxima. However, caution is needed because reliance is being placed on assumptions or arguments of the type outlined, which are not required for ordinary annual-maxima analyses. Particular caution would be needed in the presence of strong seasonality.

By considering this type of example application, it is possible to gain an idea of the effects on the plotting positions that might arise from certain extreme patterns of the sizes $\{s_i\}$. In the present application, the sizes are simply the fractions of a year of record available for each annual maxima. Suppose that out of the total of N years, N_c have complete data, while the remaining years have records for very small fractions of the year. Suppose that the pattern of observations is such that the incomplete-record years provide the highest observations overall: then one might expect the assigned plotting positions to be rather similar to those that would be derived for N complete years. In contrast, if the incomplete-record years provide the lowest observations overall, the plotting positions assigned to the N_c complete-record years should be close to those that would have been assigned if only the complete-record years had been available. While the approach to calculating plotting positions suggested here does not lead to explicit formulae, the approximations presented in the Appendix tend to verify these speculations.

Application 2: Regional Maxima

The second application suggested in the introduction was that concerned with regional maxima. In this case the basic data available consist of a set of values $\{z_{ij}\}$, where z_{ij} denotes the annual maximum rainfall at site j in year i. Here j denotes the site-number in the full network of sites, but not all sites are available for every year: let C_i denote the set of sites available for year i. It is assumed that the values $\{z_{ij}\}$ have been standardised to remove site-dependent scale effects. Suppose that a new data-set, $\{x_i\}$, is constructed, consisting of the yearly network-maxima defined by

$$x_i = \max\{z_{ij}; j \in C_i\}.$$

In practical cases, there is statistical dependence between the annual maxima for different sites within the same year. However, a model developed by Dales and Reed (1989) suggests that the effect of this dependence can be allowed for by calculating a size parameter, s_i, for each year which provides a measure of the effective number of independent sites in that year. The values s_i depend upon the number

of gauges, and the effective areal-extent of the gauges, in the set C_i and are calculated according to a simple formula given by Dales and Reed (1989, Table 8.5). This suggests that a model for the data-set of pairs $\{x_i, s_i\}$ is that the network-maximum in a given year, x, is the realisation of a random variable with distribution function F^s, where F is the distribution function of the annual maximum, z, for a single site. Thus F, the distribution of single-site maxima, is the 'typical' distribution in the terminology of Dales and Reed (1989).

As part of an analysis of the overall data-set, $\{z_{ij}\}$, with the intention of estimating return periods for single-site maxima, it may be reasonable to consider what information about this can be derived from the yearly network-maxima alone. The idea here is that an analysis of network-maxima will tend to concentrate more on higher values of rainfall than does the overall analysis. An example of such an analysis would be a graphical display of the network-maxima using plotting positions relating to the typical (single-site) distribution F.

Application 3: Historical Data

As is perhaps implicit in Application 1, the model might be applied to certain types of 'historical data', where part of the information in a data-set consists of one, or perhaps several, items which specify the maximum values recorded in certain given (long) time-periods. This type of problem will not be discussed at length here, since usually other information about the 'historical period' would be available which should be included in any analysis: this information might be of the form '. . . and all other yearly values were below . . .'. The treatment of this type of historical information, whilst outside the scope of the model used here, is discussed by Hirsch and Stedinger (1987), Guo (1990b) and Stedinger *et al.* (1993, Section 18.6.3). Use of the plotting positions derived in this paper is not recommended for 'historical data' unless one can be certain that all of the information available is represented within the underlying model.

DISCUSSION

All of the applications outlined here have related to annual-maxima. The corresponding model for annual-minima is that an observation with size parameter s has distribution function equal to $1 - (1 - F)^s$. Plotting positions for this case could be developed by a parallel argument to that employed here, or else via the usual steps that are used to derive results for minima from those for maxima. However, the applicability of this type of model for the incomplete-record case for annual minima is more doubtful than for maxima because one would expect seasonality to have a stronger effect. At present, there are no models for regional extremes of drought indices similar to those for high rainfalls developed by Dales and Reed (1989).

Plotting Positions for the Standard Situation

As already discussed, the standard case is first examined separately in order to develop an approach to deriving plotting positions which match those used in practice. The approach chosen is based on maximum-likelihood estimation and is developed by considering first the question of providing an estimate of $p = F(x)$, for a given fixed value of x: this x is any arbitrary value but, specifically, cannot be chosen on the basis of the sample data $\{x_i\}$. The initial analysis here is fairly standard, but is given in a little more detail than strictly necessary in order to provide a simple guide to the steps used later for the non-standard case.

The first step is to write down the likelihood function of the data, which represents the information contained in the data-set about the unknown parameter p. Because of the non-parametric nature of the assumptions being made, all of the information in the data about p is contained in the set of indicator random variables $\{\mathcal{J}_i\}$, which have observed values $\{j_i\}$, where

$$\begin{aligned} \mathcal{J}_i &= 1, \quad X_i \leq x, \\ &= 0, \quad \text{otherwise.} \end{aligned} \tag{3}$$

In the standard situation the random variables $\{\mathcal{J}_i\}$ are independent and identically distributed with

$$\text{Prob}\,\{\mathcal{J}_i = 1\} = p = F(x). \tag{4}$$

Hence the contribution to the likelihood function from the i'th observation is either a factor of p, if $j_i = 1$, or a factor of $(1 - p)$, if $j_i = 0$. Suppose now that exactly r of the observations satisfy $x_i \leq x$ or, equivalently, $j_i = 1$. Suppose also that one now chooses to work with the ordered data $x(i)$, with corresponding indicator variables, $j(i)$: then

$$\begin{aligned} j(i) &= 1, \quad i = 1, \ldots, r \\ &= 0, \quad i = r + 1, \ldots, N. \end{aligned}$$

Hence the log-likelihood, denoted by $L_r(p)$, is given by

$$\begin{aligned} L_r(p) &= \Sigma \log[p^{j(i)}\{1 - p\}^{1-j(i)}], \\ &= \Sigma^r \log p + \Sigma_{r+1} \log(1 - p), \end{aligned} \tag{5}$$

where the first summation, Σ^r, is over $i = 1, \ldots, r$ and the second, Σ_{r+1}, is over $i = r + 1, \ldots, N$, and where the value of the second summation is zero for $r = N$. Note that this notation for the summations is used again later. Eqn. (5) simplifies is an obvious way, but for later purposes it is convenient to rearrange it as follows.

$$\begin{aligned} L_r(p) &= r \log p + (N - r) \log(1 - p), \\ &= N \log p + (N - r)\{\log(1 - p) - \log p\}, \end{aligned} \tag{6}$$

It is further convenient to work with a modified form of the likelihood equation, defined by equating to zero the expression for p times the derivative of the log-likelihood with respect to p. This yields

$$N - (N - r)\{1/(1 - p)\} = 0. \tag{7}$$

These slightly unusual steps are chosen to give a simple way of overcoming certain numerical difficulties experienced with the solution for the non-standard situation discussed later, where an iterative solution technique is required.

The log-likelihood function is maximised at the value $p = r/N$, and the same estimated value for $F(x)$ is found for any value of x between the observations at rank r and $(r + 1)$. This means that the estimated value changes from $(r - 1)/N$ to r/N when moving from just below to just above $x(r)$. In a sense, no estimate is produced for a value of x strictly equal to $x(r)$, although formally the value r/N could be used. However, it seems reasonable to attempt to use some compromise between the estimates applicable on either side of $x(r)$. One possibility is to use a straightforward weighted average of the separate maximum-likelihood estimates. However, the averaging-of-estimates approach suffers from the difficulty that the results are not invariant to simple transformations: for example, different results would arise from averaging estimates of p compared with working with 'return-period' $1/(1 - p)$, averaging the estimated return-periods, and converting back to probabilities. Note that, for graphical purposes using an EV1 reduced-variate scale, the transformed variable of principal concern is the 'reduced variate' $y = -\log(-\log p)$.

An alternative approach is to argue as follows, and work with a weighted log-likelihood function. The log-likelihood function $L_{r-1}(p)$ summarises the information in the data about p for a value of x immediately below the r'th observation, while $L_r(p)$ summarises the information for a value immediately above. Thus one could argue for taking a (weighted) average of the two functions to form an objective function to maximise in order to define an estimate for p at a given observation. In the standard case, it turns out that the estimate obtained is exactly the same as that obtained by averaging the separate individual estimated probabilities (with the same weights), but this is no longer true in the non-standard case. If weights b and $(1 - b)$ are used, the objective function to be maximised is

$$\begin{aligned} L^*(p) &= b\, L_{r-1}(p) + (1 - b)\, L_r(p), \\ &= N \log p + \{b(N - r + 1) + (1 - b)(N - r)\} \\ &\quad \{\log(1 - p) - \log p\}, \\ &= N \log p + \{N - r + b\}\{\log(1 - p) - \log p\}, \end{aligned} \tag{8}$$

which is maximised at $p_r = (r - b)/N$. There is no reason why different weights should not be used in the estimation of p for different points $x(r)$: thus b may be a function of r and N. One way of choosing b is to ensure that the results produced for plotting positions agree with those obtained from the usual formula, Eqn. (2). Thus b is defined by

$$p_r = (r - b)/N = (r - a)/(N + 1 - 2a),$$

which yields

$$b = \{aN + (1 - 2a)r\}/\{N + 1 - 2a\}. \tag{9}$$

It can seen that b varies from approximately a to $(1 - a)$ as r varies from 1 to N.

Plotting Positions for the Non-Standard Situation

A special case of the non-standard situation which can be readily dealt with arises when all of the size parameters are equal, but not unity: that is, $s_i = S$ for all i. In this case, the procedure for the standard case can be used to create estimates appropriate for the distribution function $\{F(x)\}^S$, and these can then be transformed to estimates appropriate for $F(x)$. For example, the weighted-likelihood procedure yields plotting positions of the form

$$p_r = \{(r - b)/N\}^{1/S}, \tag{10}$$

while, depending on the order in which the averaging and transformations are done, the averaging-of-estimates approach might yield a formula such as

$$p_r = b\, \{(r - 1)/N\}^{1/S} + (1 - b)\{r/N\}^{1/S}. \tag{11}$$

In the more general non-standard situation, the likelihood-based approach can readily be developed in a similar manner to that given for the standard case. In this situation, the indicator random variables $\{\mathcal{J}_i\}$ are independent but no longer identically distributed. In fact, the indicator variable $\mathcal{J}$ for a random variable with size s has

$$\text{Prob}\,\{\mathcal{J} = 1\} = p^s = \{F(x)\}^s, \tag{12}$$

where s varies from observation to observation. This means that, written in terms of the indicators and sizes, $j(i)$ and $s(i)$, when ordered in terms of the observations, the log-likelihood function appropriate to a fixed x for which exactly r observations are less than x, is given by

$$\begin{aligned} L_r(p) &= \Sigma \log[\{p^{s(i)}\}^{j(i)}\{1 - p^{s(i)}\}^{1-j(i)}], \\ &= \Sigma^r s(i) \log p + \Sigma_{r+1} \log(1 - p^{s(i)}), \end{aligned} \tag{13}$$

where the summations are as described following Eqn. (5). It is convenient to rearrange this to a form corresponding to eqn. 6: to do this, define N_s to be

$$N_s = \Sigma s(i), \tag{14}$$

where this summation ranges over all N observations. In the case of annual maxima with incomplete years, N_s measures the total overall record length, while for the regional maxima problem it measures the total effective number of independent site-years in the data-set. Then the rearranged expression for the log-likelihood is

$$L_r(p) = N_s \log p + \Sigma_{r+1} \{\log(1 - p^{s(i)}) - s(i) \log p\}. \tag{15}$$

The expression, corresponding to Eqn. (8), for the weighted log-likelihood to be used for estimating p at $x(r)$, is given by

$$\begin{aligned} L^*(p) &= b\, L_{r-1}(p) + (1 - b)\, L_r(p), \\ &= N_s \log p + b\, \{\log(1 - p^{s(r)}) - s(r) \log p\} \\ &\quad + \Sigma_{r+1} \{\log(1 - p^{s(i)}) - s(i) \log p\}, \end{aligned} \tag{16}$$

where the summation Σ_{r+1} again ranges over $i = r + 1$ to N. Then the modified likelihood equation for the weighted likelihood can be derived by following the same procedure as used in treating the standard case: ie. equating to zero the expression for p times the derivative of the log-likelihood with respect to p. This gives

$$N_s - b\ \{s(r)/(1 - p^{s(r)})\} - \Sigma_{r+1}\ \{s(i)/(1 - p^{s(i)})\} = 0. \quad (17)$$

This equation has a simple explicit solution in the case of the highest observation, $r = N$, which yields

$$p_N = \{1 - b\ s(N)/N_s\}^{1/s(N)}. \quad (18)$$

In other cases, Eqn. (17) can be solved for $p = p_r$ by standard numerical algorithms. It can be shown that the equation always has a single solution in the range 0 to 1. Hence, bisection root-finding procedures or the Newton-Raphson method can be used. The Appendix discusses a number of cases in which one can find either explicit formulae or approximations for the plotting positions.

If a way is required to express the uncertainty with which the plotting positions can be estimated, then it seems reasonable to apply the usual likelihood-based techniques to the weighted likelihood function even though it is not a likelihood function in the usual sense. However, if the uncertainty is required for incorporating into graphical displays, then it would actually be preferable to give this information for the probability, p, associated with any x, not just the observations. In this case the ordinary log-likelihood function, Eqn. (15), would be used. It is suggested that confidence limits be constructed so as to include values of p for which the log-likelihood function is sufficiently close to the value attained at the maximum-likelihood estimate, where the criterion for closeness is calculated as one-half times the appropriate percentage point of the χ^2 distribution with one degree of freedom.

Choice of the Weights b

As in the standard case, a choice for the weights $b = b_{r,N}$ has to be made. It was found earlier that the weights b required to reproduce the standard plotting position formulae would vary around the value ½, and one possibility for the non-standard case would be to adopt $b = ½$ as a standard choice. In the standard case this would yield the Hazen plotting positions. However, considerable effort has in the past been devoted to developing plotting position formulae (Cunnane, 1978) and it seems sensible to try to take this into account in any procedure developed for the non-standard case.

On the basis of Eqn. (9), the following choices appear reasonable:

$$b^{(1)} = \{aN + (1 - 2a)r\}/\{N + 1 - 2a\}; \quad (19)$$

$$b^{(2)} = \{aN_s + (1 - 2a)\ \Sigma^r\ s(i)\ \}/\{N_s + 1 - 2a\},$$
$$= \{(1 - a)N_s - (1 - 2a)\ \Sigma_{r+1}\ s(i)\ \}/\{N_s + 1 - 2a\}. \quad (20)$$

If one considers the case of the plotting position for the largest observation ($r = N$), for which an exact solution of the likelihood equation is known, and which is given by Eqn. (18), it is possible to conclude that Eqn. (20) provides a better choice for the weighting coefficient b than does Eqn. (19). In particular, it may be argued that the plotting position assigned to this observation should be very similar to that obtained for the largest observation in the standard case, but where the sample size is counted as being N_s. According to the non-standard model, the distribution function of the largest observation is $F(x)$ raised to the power N_s: this is the same as the distribution function of the largest of N_s observations having the underlying distribution function $F(x)$. It therefore follows that the formula for b should be chosen so that p_N, given by Eqn. (18), also satisfies

$$p_N \approx (N_s - a)\ /(N_s + 1 - 2a)$$
$$= 1 - (1 - a)\ /(N_s + 1 - 2a). \quad (21)$$

When Eqn. (20) is used to define b, the result for p_N is given by

$$p_N = \{1 - (1 - a)\ s(N)/(N_s + 1 - 2a)\ \}^{1/s(N)},$$

and it is clear that this matches the right hand side of Eqn. (21), exactly if $s(N) = 1$, and to a good approximation if N_s is moderately large compared to $s(N)$. It is of course possible to choose b for the highest observation in such a way that the required plotting position is exactly reproduced: this gives

$$b_{N,N} = [1 - \{(N_s - a)/(N_s + 1 - 2a)\}^{s(N)}]N_s/s(N). \quad (22)$$

While it might be possible to develop an appropriate extension of this to a more general one for $b_{r,N}$, there seems no convincing argument for adopting an expression of this complexity in preference to the simpler choice given by Eqn. (20).

Expression (20) for the weighting coefficient was constructed on the basis of a simple analogy with Eqn. (9) which gives the formula for b required in the standard case to reproduce the usual plotting positions. As before, the values of b vary in a fairly narrow range as r varies. It seems that the exact values used for b should not matter too much provided that they are reasonably close to ½. In some sense, the final plotting positions are mainly determined by the solution of the likelihood equation, rather than by b.

Given the simplicity of the proposed formula for b, it is natural to wonder whether some simple direct formula for the plotting positions can be found, thus avoiding the problem of finding the numerical solutions of the likelihood equations entirely. However, there seems to be no simple, potentially appropriate, formula for plotting positions which meets the two requirements:

(i) values should be between 0 and 1;
(ii) the formula should reproduce the value $(N_s - a)\ /(N_s + 1 - 2a)$ for $r = N$.

Given that the numerical solution of Eqn. (17) is easy, its use to derive plotting positions can be regarded as reasonably practicable. There is only a slight inconvenience attached to using such numerical procedures which might be overcome by an explicit formula, if one could be found.

An Example

The following example of the outcome of the procedure described here for deriving plotting positions has been created on the basis of randomly generated data, since then the 'true' plotting positions of the data-points are known, and can be compared with those calculated using the procedure. Results for two samples from the same model are shown in Table 1 so that some appreciation of the random sampling effects can be gained.

The situation being modelled is essentially that of the 'Regional Maxima' application discussed above. In this instance it is assumed that 20 years of data are available and that in each year the network-maximum value derives from 1, 3, 5, . . ., 39 effectively independent sites. The underlying distribution is considered to be either a uniform distribution or a standard EV1 distribution, the cases being considered in parallel. Thus in year i, the data value is either u_i, the maximum of s_i independent uniform variates, or x_i, the maximum of s_i independent EV1 variates. The usual transformation between uniform and EV1 variates is used to cover both cases simultaneously. Table 1 shows the results in terms of the variables $s(i)$, etc., ordered in terms of x or u. The results use Eqn. (20) to specify the weights b, with $a = 0.44$, to correspond to the Gringorten plotting positions, and they consist of the calculated plotting positions p_i and the corresponding EV1 'reduced variates' y_i. In this case, the "true" plotting position of the i'th ordered observation, $u(i)$, is $u(i)$ since this is the non-exceedence probability of this value under the underlying distribution. Similarly, the "true" reduced variate for the i'th ordered data point is $x(i)$. It can be seen that there is a reasonable match between $\{p_i\}$ and $\{u(i)\}$, and between $\{y_i\}$ and $\{x(i)\}$.

It is of a little interest to note the difference in the results from the two samples. Each sample resulted in a different ordering of the sizes $\{s(i)\}$, and thus different sets of plotting positions are suggested in the two cases. As would be expected, the largest network-maximum values tend to occur in years for which most sites are operating.

A limited set of somewhat more extensive simulation studies has been performed for examples similar to that given here. These have confirmed that the plotting positions given by the weighted likelihood method behave reasonably, in that scatter plots of the 'true' against the suggested plotting positions or reduced variates are centred about a one-to-one line, as would be hoped.

Table 1. Results for plotting positions for some simulated samples of data. Values u and x represent data-values, while p and y are the assigned plotting positions on uniform and EV1 scales, respectively

Sample 1 rank i	$s(i)$	$u(i)$	p_i	$x(i)$	y_i
1	7	0.695	0.640	1.012	0.807
2	1	0.738	0.756	1.191	1.272
3	3	0.788	0.806	1.435	1.535
4	9	0.833	0.851	1.699	1.826
5	17	0.853	0.887	1.843	2.121
6	25	0.878	0.913	2.035	2.396
7	33	0.921	0.932	2.502	2.649
8	15	0.936	0.943	2.709	2.836
9	5	0.948	0.950	2.926	2.963
10	11	0.954	0.956	3.049	3.090
11	21	0.973	0.962	3.580	3.245
12	27	0.979	0.968	3.868	3.425
13	29	0.980	0.974	3.923	3.619
14	39	0.987	0.979	4.314	3.835
15	35	0.987	0.983	4.329	4.071
16	31	0.989	0.987	4.537	4.325
17	37	0.992	0.990	4.792	4.624
18	13	0.994	0.993	5.154	4.992
19	23	0.996	0.996	5.518	5.509
20	19	0.999	0.999	7.069	6.558

Sample 2 rank i	$s(i)$	$u(i)$	p_i	$x(i)$	y_i
1	1	0.562	0.399	0.551	0.084
2	5	0.655	0.707	0.859	1.060
3	9	0.874	0.812	2.006	1.571
4	3	0.876	0.853	2.023	1.836
5	23	0.937	0.889	2.733	2.142
6	15	0.939	0.912	2.762	2.389
7	19	0.942	0.927	2.824	2.580
8	7	0.944	0.937	2.850	2.733
9	25	0.962	0.946	3.261	2.898
10	33	0.968	0.956	3.412	3.098
11	13	0.970	0.963	3.484	3.272
12	17	0.974	0.968	3.621	3.427
13	39	0.976	0.974	3.716	3.621
14	31	0.976	0.979	3.727	3.839
15	11	0.991	0.983	4.732	4.048
16	21	0.991	0.986	4.736	4.276
17	35	0.994	0.990	5.037	4.570
18	27	0.994	0.993	5.125	4.949
19	29	0.994	0.996	5.164	5.484
20	37	0.999	0.999	6.945	6.545

Acknowledgement

The work reported here arose in connection with research for the UK Flood Estimation Handbook as part of project FD0414 for the Ministry of Agriculture, Fisheries and Food.

Appendix A: Exact and Approximate Solutions

Explicit formulae for solutions, or approximate solutions, to the likelihood equation for the plotting position can be found in a few special cases. The plotting position, p_r, is the solution to Eqn. (17). Besides those cases outlined here, there are a few other special cases in which explicit solutions can be obtained, but they are omitted for brevity. In particular, if the set of sizes $\{s_i\}$ consists of values all of which are either a basic size or twice the basic size, then solution of Eqn. (17) becomes equivalent to the solution of a quadratic equation, while if the sizes are all either a basic size or three-times the basic size, then a cubic equation is involved. Other cases can be found which reduce to the solution of a quartic, for which an explicit solution could in principle be given. For present purposes it is convenient to treat the weight b as if it were constant, rather than varying with r as it would if the suggested weighting according to Eqn. (20) were adopted.

Case (i)

Suppose that the pattern of observations is such that, beyond the K'th ordered value of $x(i)$, all have equal values of $s(i)$:

$$s(i) = s(N), \quad i = K + 1, \ldots, N. \tag{A.1}$$

Then, as an extension of formula (18),

$$p_r = [1 - \{b + N - r\}\, s(N)/N_s]^{1/s(N)}, \quad r = K + 1, \ldots, N. \tag{A.2}$$

One would expect this case to arise frequently in problems for which most of the size parameters take the same value: for example in Application 1 dealing with incomplete records for annual maxima, in which case the largest observations would often arise from complete years, specifically those with size parameters equal to unity. Note that the plotting positions for the highest values do not depend on the number of the highest ranking values which share the size $s(N)$.

Case (ii)

As an extension of case (i), suppose that in addition, the lowest K observations have equal values of $s(i)$: that is

$$s(i) = s(1), \quad i = 1, \ldots, K. \tag{A.3}$$

Then Eqn. (A.2) becomes

$$p_r = [\{(r - K - b)s(N)\}/\{(N - K)s(N) + K\, s(1)\}]^{1/s(N)}, \quad r = K + 1, \ldots, N. \tag{A.4}$$

It can be seen from this that, if $s(1)$ is small compared with $s(N)$, the plotting positions assigned to the largest observations will be essentially the same as would have been produced if the smallest observations had not been included in the data-set.

Case (iii)

As a different extension of case (i), suppose that the size $s(N)$ shared by the highest observations is close to zero. Then power series expansions lead to the approximation

$$p_r \approx \exp\{-(b + N - r)/N_s\} \approx \{1 - (b + N - r)/N_s\}, \quad r = K + 1, \ldots, N. \tag{A.5}$$

This indicates that the highest observation is given a plotting position which is a distance b/N_s below 1, and that the next highest observations have plotting positions at steps of $1/N_s$. Consider now the K'th observation: an approximation for the plotting position for this observation can be developed from the likelihood equation in two steps. First the contributions of the $(K+1)$'st to N'th observations are approximated using

$$(1 - p^{s(N)})/s(N) \approx -\log p, \quad \text{for small } s(N), \tag{A.6}$$

and then by

$$-\log p = [-\log p^{s(K)}]/s(K) = [-\log\{1-(1 - p^{s(K)})\}]/s(K) \approx (1 - p^{s(K)})/s(K), \tag{A.7}$$

where the assumption that p is close to 1 is made. It then follows that

$$p_{K,N} \approx [1 - \{b + N - K\}s(K)/N_s\,]^{1/s(K)}, \tag{A.8}$$

which shows that the plotting position assigned to this observation is similar to that which would have been obtained if the highest observations had all had size parameters equal to $s(K)$, apart from a possibly small effect on the total of the sizes, N_s.

The result (A.8) may also be compared with the result that would be obtained for the K'th observation if the higher observations, with small sizes $s(i)$, had been ignored. If the approximation is made that N_s is effectively unchanged, then

$$p_{K,K} \approx [1 - b\, s(K)/N_s\,]^{1/s(K)}. \tag{A.9}$$

It can therefore be seen that including high observations with small values of the size parameter $s(i)$ has a substantial effect on the plotting position of the next highest observation, compared with the effect of omitting them. In fact, if the further assumption is made that the highest remaining observations would have equal size parameters, then a useful estimate of the effect of omitting the highest observations can be obtained. Suppose that initial estimates are obtained by omitting the $L = N - K$ highest observations with equal but small values of $s(i)$. Then the approximations here suggest that the plotting position that would be assigned to the highest remaining observation, if all observations were included, would be close to that

assigned to the $(K - L)$'th highest observation out of the reduced set of K values.

Case (iv)

Suppose that the pattern of observations is such that the smallest K values of $x(i)$ all have equal values of $s(i)$, and that in total K^* values share this value of $s(i)$. Suppose that this common value is very much smaller than all of the other values of $s(i)$. Then an approximate formula for the lowest plotting positions is given by

$$p_r \approx \{(1 - b)/(K^* + 1 - r)\}^{1/s(1)}, \quad r = 1, \ldots, K. \quad \text{(A.10)}$$

This formula is derived by retaining only the dominant (lowest) powers of p, assuming p is small, in an approximation for the likelihood equation.

Case (v)

In a similar way to case (iv), an approximation for the plotting position when p is close to one can be derived by substituting the leading terms of a power-series expansion in $(1 - p)$ into the likelihood equation. This yields

$$p_r \approx \{1 - (b + N - r)/N_s\}, \quad \text{(A.11)}$$

provided that p_r is close to 1. A slightly more formal approach, which involves assuming that p_r has a power-series expansion in terms of $\varepsilon = 1/N_s$, gives

$$p_r \approx \{1 - \varepsilon\,(b + N - r)(1 + \tfrac{1}{2}\varepsilon t_r)\}, \quad \text{(A.12)}$$

where

$$t_r = b\{s(r) - 1\} + \Sigma_{r+1}\,\{s(i) - 1\}. \quad \text{(A.13)}$$

Here the assumption made in the power-series expansion is that N_s becomes large while $(N - r)$ and the contributions $\{s(r), \ldots, s(N)\}$ are fixed. Note that, in contrast to Eqn. (A.11), the higher order approximation used for Eqn. (A.12) shows an effect on the plotting position resulting from the ordering of the sizes $\{s(i)\}$, via Eqn. (A.13). A similar power-series expansion for $\log p_r$ can be derived.

References

Cunnane, C. (1978). Unbiased plotting positions—a review. *J. Hydrol.*, **37** (3/4), 205–222.

Dales, M.Y. and Reed, D.W. (1989). Regional flood and storm hazard assessment. Report No. 102, Institute of Hydrology, Wallingford, UK.

Guo, S.L. (1990a). A discussion on unbiased plotting positions for the General Extreme Value distribution. *J. Hydrol.*, **121**, 33–44.

Guo, S.L. (1990b). Unbiased plotting position formulae for historical floods. *J. Hydrol.*, **121**, 45-61.

Hirsch, R.M. and Stedinger, J.R. (1987). Plotting positions for historical floods and their precision. *Wat. Resour. Res.*, **23** (4), 715–727.

Hyndman, R.J. and Fan Y. (1996). Sample quantiles in statistical packages. *Amer. Statist.*, **50** (4), 361–365.

NERC (1975). *Flood Studies Report, Volume 1 (Hydrological Studies)*. Natural Environment Research Council, London.

Stedinger, J.R., Vogel, R.M. and Foufoula-Georgiou, E. (1993). Frequency analysis of extreme events. Chapter 18 of '*Handbook of Hydrology*' (Ed. Maidment, D.R.), 18.1–18.66, McGraw-Hill, New York.