

Reply

LASSE MAKKONEN

VTT Technical Research Centre of Finland, Espoo, Finland

(Manuscript received and in final form 15 April 2010)

ABSTRACT

This reply addresses the use of order statistics in extreme value analysis. The author has previously proposed in this journal that the distribution-dependent estimators of plotting position in extreme value analysis should be abandoned and replaced by the Weibull formula. It was also demonstrated that the use of the wrong plotting positions has resulted in underestimation of the probability of extreme-weather events. Cook's comments challenge these developments and defend the previously presented plotting methods. In this reply it is outlined that the Weibull formula provides the exact probability P_I of nonexceedance in order-ranked data. Hence, there is no sampling error related to P_I . This renders Cook's primary arguments invalid. The specific critical comments by Cook are also replied to and are shown to be unfounded.

1. Introduction

Estimating the probabilities of extreme-weather events is crucially important for avoiding weather-related disasters. Such estimates are the basis of external loads in codes and regulations for structural design, for example. An important method in this estimation is the extreme value analysis based on order-ranked observations of weather variables, typically annual extremes. By the extreme value analysis the probability distributions of the potentially hazardous weather phenomena can be estimated and the probability of exceeding a certain critical value can be determined. An important starting point in the extreme value analysis is associating probabilities with the observed extreme variable values.

It has been proposed that the commonly used distribution-dependent estimators of these associated probabilities, the so-called plotting positions, should be abandoned and replaced by Weibull's (1939) formula (Makkonen 2006, 2008a,b). It was shown by Makkonen (2006) that an underestimation of the probability of extreme events has resulted from the use of many other plotting positions historically, and that this has had an adverse impact on building codes and other means for optimal design against extreme-weather events.

Cook (2010) challenges these developments and claims to disprove the findings by Makkonen (2006). According to Cook (2010), "the unbiased estimators developed since 1939 are valid and provide a significant improvement in accuracy."

It is outlined in this reply that the Weibull formula provides the exact distribution-free probability P_I of nonexceedance by a random observation drawn from the population. There is no sampling error related to P_I . Hence, the use of "estimators" for plotting positions is both unnecessary and misleading. The critical comments by Cook (2010) are replied to and are shown to be unfounded.

2. Probability and relative sample frequency in order-ranked data

A fundamental confusion has existed for more than 50 years on the concepts of "probability" and "sample probability." Cook (2010) makes an attempt to clarify these concepts in section 3 of his paper with poor success. He uses notations P_I for the probability associated with what he refers to as an "ideal invariant process" and P for sample probability but yet denotes by P the probability defined as the limiting value of the sample relative frequency.

Since this confusion is the root of many errors encountered in extreme value analysis, the wording sample probability is not used in this paper. Such a concept is highly misleading since probability is a property of a *random process* and not of a *sample*. In this paper "sample relative frequency" and notation r are used for what Cook

Corresponding author address: Lasse Makkonen, VTT Technical Research Centre of Finland, Box 1000, 02044 VTT, Espoo, Finland.
E-mail: lasse.makkonen@vtt.fi

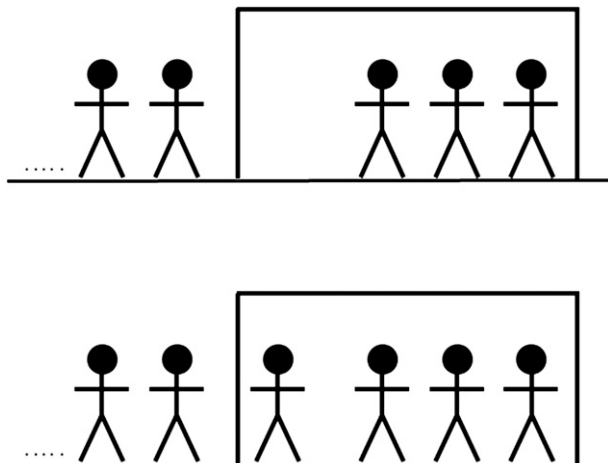


FIG. 1. Examples of nonexceedance probabilities in random sampling. (top) There are three people in a room ($n = 3$) ranked in order from the shortest ($m = 1$) to the tallest ($m = 3$). The probability that a person randomly drawn from this sample is not taller than the second tallest ($m = 2$) of them is $2/3 = m/n$. However, the probability that a person randomly drawn from the population of all persons is not taller than the second-tallest person in the room is $1/2 = m/(n + 1)$. This can be seen by imagining that (bottom) this randomly chosen person has entered the room. Upon reranking, this person may fall in any of the new ranks first, second, third, and fourth, in two of which (first and second) the person is not taller than the second-tallest person in the original sample.

and many others call “sample probability” or even just “probability.”

Makkonen (2008a) pointed out that in random sampling the cumulative distribution function (CDF) of x_m must be interpreted as the empirical distribution function, which is defined based on the observed relative frequency r . The mean of $r_L(m, n)$ goes to the mean of the CDF value $P_I(m, n)$ when the number L of trials of r_m goes to infinity. Thus, Gumbel’s (1958) celebrated result, presented by different notations in Cook’s Eq. (9), can be interpreted as

$$P_I = m/(n + 1). \quad (1)$$

Here m is the order and n is the number of observations in a sample that is order ranked from the smallest ($m = 1$) to the largest ($m = n$). A particular benefit of using order statistics is that P_I is unaffected by the distribution.

The result of Eq. (1) can be described by elementary examples illustrated in Fig. 1. Suppose three people are in a room ($n = 3$) as in the top panel of Fig. 1. Let us rank them in order from the shortest ($m = 1$) to the tallest ($m = 3$). We first ask: What is the probability that a person randomly drawn from those *in the room* is not taller than the second-tallest person ($m = 2$) in the room? Two persons out of the three fall in this category, so that the answer is $2/3 = m/n$. Consider now a different question: What is the probability that a person randomly drawn

from *all people* is not taller than the second-tallest person in the room? Suppose that this randomly chosen person has entered the room, as shown in the bottom panel of Fig. 1. Now we have four people in the room, and, upon reranking, this new person may, with equal probability, fall in any of the new ranks first, second, third, and fourth. When the new person falls in the first or second rank then the new person is not taller than the second-tallest person in the original sample. Therefore, the nonexceedance probability is now $P_I = 2/4 = m/(n + 1)$.

These examples demonstrate that the nonexceedance probabilities in the two samplings are unique and depend on m and n only. They do not depend on the distribution of the variable. Furthermore, there is no error or bias related to them. The example also outlines that the argument given by Cook, that two probabilities $m/(n + 1)$ and m/n may emerge giving a nonunique plotting position, reflects the failure to distinguish between two different random processes: nonexceedance by a random observation drawn from a sample and by a random observation drawn from the population from which the sample originates.

3. The purpose of probability plotting

The purpose of probability plotting is to estimate the CDF of a variable. The CDF is defined as follows: For every real number x , the CDF of a real-valued random variable X is given by

$$x \rightarrow F_X(x) = P_I(X \leq x), \quad (2)$$

where the right-hand side represents the probability that the random variable X takes on a value less than or equal to x .

It should be clear, therefore, that when estimating the CDF, the purpose is to estimate the relationship between x and P_I . The probability P_I is, by the definition of CDF in Eq. (2), that of nonexceedance by a random observation in the *population*, often called “the next observation.” Therefore, when plotting on the probability axis for the purpose of estimating the CDF, Eq. (1) applies. Any deviation from Eq. (1) results in an error on the probability axis.

4. The role of sampling error

Cook (2010) refers to estimates of probability, indicating that sampling error plays a role in assessing the probability in extreme value analysis. The discussion above shows that no such role exists. When considering order-ranked data, the sampling error is related solely to the observed variable x , not to the probability of nonexceedance P_I . When dealing with a known CDF one

may, as a mathematical tool, fix a variable value X and treat the sample relative frequency r_L as a variable, as was done by Gumbel (1958). However, this should not be confused with the concept of probability P_I in order statistics of sampled data.

As will be further discussed in section 5, the probability P_I required to estimate a CDF cannot be derived from Cook's (2010) nonadditive transformations such as the "mean reduced variate," "sample return period," and "expectation of the return period." Cook's (2010) discussion on them and his related figures are, therefore, irrelevant to assessing a CDF based on order statistics. The confusion discussed above is also reflected in Cook's Eqs. (16)–(18) where P is used instead of the probability P_I that was used by Gumbel (1958) in making the transformation to the reduced variate $y = -\ln(-\ln P_I)$.

Cook (2010) states that in his Fig. 6 "the mean reduced variate \bar{y} is plotted against various estimates of probability on conventional Gumbel axes." The concept of "estimate of probability" is meaningless, however, since, in order statistics, the probability P_I is given by Eq. (1) and is not to be estimated.

5. Reply to the specific comments by Cook (2010)

In connection with his Eq. (20) Cook (2010) claims that there has been a fundamental error in Makkonen (2006) in that it was assumed that $R = \bar{T} = 1/\bar{Q}$. This is mischaracterization of the paper by Makkonen (2006) since none of the concepts R , \bar{T} , or \bar{Q} were used in that paper. Equations (9)–(12) in Makkonen (2006) deal with the concepts of R_I and Q_I when using Cook's notations.

A reply is given in the following to each of the specific numbered comments by Cook (2010):

- (i) The particular benefit of order statistics is that the probabilities related to the CDF are distribution free, as discussed in section 2 of this paper.
- (ii) The "bias" discussed by Cook arises from incorrectly plotting \bar{y} versus $-\ln(-\ln P_I)$ (see item vii below). Equation (1) gives the exact nonexceedance probability.
- (iii) The relationship between the rank of the observation and the corresponding return period R_I directly follows from the definition of R_I by P_I . This is Eq. (10) in Makkonen (2006).
- (iv) In the past, the methods of fitting a distribution to the data have been manipulated to compensate for the errors following from the use of a wrong plotting formula. This has resulted in widespread confusion (Makkonen 2008b). Such manipulations are unnecessary and lead to errors, since the correct probability $P_I = m/(n + 1)$ should be used in the first place.

(v) The purpose of the extreme value analysis is to determine the CDF, that is, the relationship between x and P_I [see Eq. (2)]. If, in an analysis, P_I is plotted wrongly, then the resulting CDF will be wrong. This fact has nothing to do with the distribution. Only after the plotting has been correctly done can one make conclusions about the distribution and find its parameters. One may, of course, be interested not only in the plot representing the CDF, but, for example, in the quantiles. These are determined based on the estimated CDF, which is, of course, not a distribution-free procedure.

(vi) Cook's claim that there was an error in Makkonen (2006) is not true because Cook's concepts, such as \bar{T} and \bar{Q} , were not used in that paper at all. The concept used by Makkonen (2006) is the return period $R_I = 1/(1 - P_I)$.

(vii) By "losing the linearity" Makkonen (2006) referred to the "bias" that is considered in Cook's Fig. 6. This figure represents the results of 10 000 trials. Each trial includes a sample of 9 values of random variable $y = x$ for which a CDF, F , of Fisher–Tippet type 1 (FT1) is assumed. The values in each trial have been order ranked to $y_{1,L}, \dots, y_{9,L}$ ($L = 1, \dots, 10\,000$) and the mean of each rank, that is, $\bar{y}_1 = E(y_{1,L}), \dots, \bar{y}_9 = E(y_{9,L})$, has been calculated. Last, the mean values \bar{y}_m have been plotted using different plotting positions. Since F is a non-linear transformation, $F[E(y_{m,L})] \neq E[F(y_{m,L})] = m/(n + 1)$. As a consequence, there is no reason to expect that plotting $m/(n + 1)$ against $F[E(y_{m,L})]$ would result in points on the straight line representing the fundamental distribution FT1. This is why Cook and many others regard plotting positions $m/(n + 1)$ as "biased."

When carrying out an extreme value analysis, however, we are *not* dealing with the case illustrated in Cook's Fig. 6. We have n order-ranked annual maxima x_1, \dots, x_n and we wish to know the non-exceedance probability P_I of each x_m next year. The answer to this question is $P_I = m/(n + 1)$. This means that, applying Gumbel's transformation, the plotting position is (x_m, y) , that is, $(x_m, -\ln\{-\ln[m/(n + 1)]\})$. In the extreme value analysis, any new annual maximum is added to the set of annual maxima, and *no mean of the type \bar{y}_m* is used as in Cook's Fig. 6.

Therefore, the demonstration in Cook's Fig. 6 and the related discussion on a "bias" is irrelevant to the correctly performed extreme value analysis. Because $F[E(y_{m,L})] \neq E[F(y_{m,L})]$, Cook's variable \bar{y} cannot be used to reproduce the correct probability P_I . This situation cannot be salvaged by manipulating P_I to "reproduce the linear form of the

datum distribution” as claimed by Cook. This is because, while reproducing the *right form* of the distribution, such manipulated values of P_I will provide *wrong probabilities* of nonexceedance, particularly at the tails of a distribution.

- (viii) The issue of feasibility of mean values of variables can be illustrated by a simple example. Consider two consecutive periods of 30 yr of observations of a variable a certain value of which is exceeded by an annual maximum once in the first period and 5 times in the second period. The estimate of the return period is then 30 yr for the first period and 6 yr for the second period. Using the definition by Cook for \bar{R} , the “mean return period” is then $(30 + 6)/2 = 18$ yr. Cook implies that this is a meaningful result and can be used in addressing the probability. In this example, however, we have 60 yr of observations and 6 events of exceedance. Therefore, the return period is 10 yr. This points out that Cook’s claim of a principle that the best estimate is always the mean is not true when the variable in question is nonadditive. Worse still, there is no way how the correct return period R_I (10 yr) can be traced back from a mean of the nonadditive variable \bar{R} (18 yr).
- (ix) The distribution-dependent “estimators” should be abandoned because of a very fundamental reason. This paper and two other recent papers (Makkonen 2008a,b) aim at making that reason clear: “Estimator of probability” is an inappropriate concept. In order statistics, the nonexceedance probability of the m th value is not a variable. It is known from Eq. (1) and is not to be estimated. Of course, “estimators” of a probability, such as those presented in Cook’s Table 4, are worse than the exact value.

6. Discussion and conclusions

The extreme value analysis got on the wrong tracks, starting from its invention a century ago, when P_I was set to m/n . It was soon realized that this method cannot be used at the tails of the distribution. To deal with this problem, an “estimator” of P_I was proposed as $(m - 1/2)/n$ by Hazen (1914). Subsequent to this, many other “estimators” were proposed, ending up with numerical methods such as those advocated by Cook (2010).

This unfortunate state of affairs arose because it was not observed at the time that m/n is not the probability of exceedance by a random observation in a population, but by that in a sample (see example in Fig. 1). As outlined in this paper, the correct probability $P_I = m/(n + 1)$ should have been used in the first place, making all “estimators” unnecessary. This correct formula was proposed by Weibull

(1939) and later by Gumbel (1958), but they did not justify it appropriately. Gumbel (1958) derived the result $E(r_L) = m/(n + 1)$ and recommended it as the plotting position but did not identify it as P_I . His derivation was based on considering r_L as a variable with a fixed value of x . This way of thinking subsequently initiated several erroneous methods in which r_L is confused with P_I with the consequence that P_I is considered as a variable to be estimated.

As a consequence, all but those estimates of the probability of natural hazards that have been made by Eq. (1) are incorrect. More than 10 erroneous plotting methods have been presented over a century (Makkonen 2008b). The practice of using plotting methods in building codes and other regulations has varied, so that the errors have found their way into statistics of extremes in many, but not all, countries. Only a few significant text books have recommended the correct plotting formula to be used (Gumbel 1958; Madsen et al. 1986; Coles 2001).

The paper by Makkonen (2006) and more recent papers (Makkonen 2008a,b) were attempts to direct the extreme analysis to the correct path by showing why the plotting of x must be done by P_I and that $P_I(m, n) = m/(n + 1)$. Cook’s (2010) comments advocate the widely practiced erroneous methodology and represent the conventional misunderstandings of the use of order statistics in extreme value analysis. In this reply, all of his critical comments on Makkonen (2006) have been shown to be unfounded and the correct methodology of this important issue was outlined.

Acknowledgments. Thanks are given to Matti Pajari for many fruitful discussions and comments on the manuscript. This work was supported by EWENT and Fate-Defex projects.

REFERENCES

- Coles, S., 2001: *An Introduction of Statistical Modeling of Extreme Values*. Springer, 224 pp.
- Cook, N., 2010: Comments on “Plotting positions in extreme value analysis.” *J. Appl. Meteor. Climatol.*, **50**, 255–266.
- Gumbel, E. J., 1958: *Statistics of Extremes*. Columbia University Press, 375 pp.
- Hazen, A., 1914: Storage to be provided in impounding reservoirs for municipal water supply. *Trans. Amer. Soc. Civ. Eng. Pap.*, **1308**, 1547–1550.
- Madsen, H. O., S. Krenk, and N. C. Lind, 1986: *Methods of Structural Safety*. Prentice Hall, 403 pp.
- Makkonen, L., 2006: Plotting positions in extreme value analysis. *J. Appl. Meteor. Climatol.*, **45**, 334–340.
- , 2008a: Bringing closure to the plotting position controversy. *Commun. Stat. Theory Methods*, **35**, 460–467.
- , 2008b: Problems in the extreme value analysis. *Struct. Saf.*, **30**, 405–419.
- Weibull, W., 1939: A statistical theory of the strength of materials. *Ing. Vetensk. Akad. Handl.*, **151**, 1–45.