Nordic Hydrology, 25, 1994, 301-312

Statistical Estimates of PMP Values

Jónas Elíasson

University of Iceland, Reykjavik

The article discusses two statistical methods to estimate PMP values, the Hershfield and the NERC methods. Neither method offers any explanation why the PMP values can be calculated by the use of unbounded statistical distributions, but both methods include the use of envelope curves that are not independent of the region. Bounded data that fits an unbounded distribution must deviate from the distribution for high return periods and tend to a limiting value, and then there exists, a limiting reduced variate that can be used to find the PMP value. When the distribution is $EV1$, the limiting reduced variate can be defined by a mapping transformation, or by cutting off the distribution. It is shown that when Hershfield or NERC methods are used, the limiting reduced variate is included in the PMP values and can be separated from regional parameters. It is suggested that the limiting reduced variate, that depends solely on return period, may more easily be transferred between regions than the other parameters. This may be a great help in finding *PMP* values in regions where observations are not extensive enough to define limiting return periods with necessary certainty.

A case study with data from Iceland demonstrates, that using the limiting reduced variate, similarities emerge in the Icelandic data and the NERC *PMP* that justify the acceptance of the NERC method.

Introduction

Probable Maximum Precipitation (*PMP*) is defined as follows:

> *...the greatest depth of precipitation for a given duration meteorologically possible for a given size storm area at a particular location at a particular time of year, with no allowance made for long-term climatic trends.*
> (WMO 1986).

Knowledge of Probable Maximum Precipitation is increasingly important in engineering design. This has led to an important development in statistical methods to estimate *PMP*. The earliest work was in the U.S.A. Here Hershfield suggested that *PMP* values could be calculated by adding 15 standard deviations to the mean, but later he extended the method in Hershfield (1965), to include an envelope curve. In Britain a more systematic method was introduced by NERC in the FSR (1975), based on a thorough study on the tails of the distributions of annual maxima. Recently Norway followed (Förland and Kristoffersen 1988; 1989) by adapting the NERC method to Norwegian conditions.

Little information is available on *PMP* in Iceland, undoubtedly because the number of meteorological observations is scarce, especially in the interior of the country. In many stations observation years are few. In short, the Icelandic meteorological data is inadequate in order to utilise in full the statistical methods devised so far.

Station values have been calculated and published Elíasson (1991). There, the Hershfield and NERC methods were used without any attempt to find if they were either supported or contradicted by the statistical evidence that might possibly be extracted from the inadequate data. The problem is, that any attempt to calculate *PMP* values for Iceland must be based on parameters that may be transferred between regions on a limited scale at least. These parameters have to be identified and separated from regional parameters that contemporary science has proved to vary between regions. This problem does not arise in regions where adequate data is available.

The Limiting Reduced Variate y_{lim}

Consider any statistical distribution function describing the variable X

$$P(X<x) = F(\frac{x}{a} + b) \qquad (1)$$

The distribution function *F(x)* is a monotonous function and may be inverted to give

$$y = F^{-1}F(\frac{x}{a} + b) = \frac{x}{a} + b \qquad (2)$$

y is often called the *reduced variate* of F. In testing if some data $X_1, X_2, \ldots X_n$ fits our distribution we usually take F^{-1} of our *plotting frequencies* $P_1, P_2, \ldots P_n$

$$Y_n = F^{-1}(P_n) \qquad (3)$$

and fit the $X_1, X_2, \ldots X_n$ to them by using an estimator to find the value of the scaling constant a and the location constant b in Eq. (2), which becomes in a shorter version

$$y = \frac{x}{a} + b \tag{4}$$

The variables y and x must be bounded to the same side and unbounded to the same side, as Eq. (4) is linear. That y can take any value in $(-\infty,\infty)$ while x is bounded in *e.g.* $(-\infty, x_{PM})$ – where x_{PM} is some probable maximum value – should really be out of the question. But this is what is done when probable maximum precipitation is estimated by statistical methods.

But if we instead of a linear relationship use

$$y = \frac{x}{a} + b + \frac{k}{(x/a)+b} \tag{5}$$

where k is a negative constant, and y is the reduced variate of some unbounded distribution function, y will tend to infinity as $x/a + b$ approaches zero.

Consider the probability of the maximum of an independent identically distributed random variable with the distribution function $F(x)$. Largely following Leadbetter *et al.* (1983), we can write

$$M_n = \mathrm{Max}(X_1, X_2, \ldots X_n)\,,\; P(M_n < x) = F^n(x) \tag{6}$$

As $F(x) < 1$, $F^n(x)$ vanishes as $n \to \infty$ for a fixed x. The problem of finding any limit of $P(M_n < x)$ is therefore to find a proper scaling of x that depends on n in such a way, that the distribution of the scaled value has a limit as n increases. In other words, for every n we must be able to find two scaling parameters, $a_n = a(n)$ and $b_n = b(n)$, such that

$$\lim_{n\to\infty}\left(P(M_n < \frac{x}{a_n} + b_n) = F^n(\frac{x}{a_n} + b_n)\right) = G(x) \tag{7}$$

Extreme value theory now tells us that if the limit function $G(x)$ exists, the resulting distribution function is max-stable, which means that for any z we can always find two constants such that

$$G^z(a(z)x + b(z)) = G(x) \tag{8}$$

Extreme value theory tells us furthermore, that the only distributions that are max-stable are the 3 types of *extreme value distributions,* $EV1$, $EV2$ and $EV3$, where $EV1$ is the well known Gumbel distribution and the $EV3$ is known as the Weibull distribution. All 3 can be shown to be special cases of a General Extreme Value distribution GEV.

If the transformation Eq. (5) is to be used, it is an advantage if it does not alter the max-stable properties of the $EV1$ distribution function. We now define the distribution function Eq. (9), which is the result of applying the transformation Eq.

(5) to $EV1$. For convenience we set a and b in Eq. (5) to 1 and zero respectively. This can be done without any loss of generality, using in Eqs. (9) and (10) a new x' to denote the scaled x in Eq. (5) leads to the same result

$$F(x) = \exp(-\exp(-y)) \quad \text{for} \quad x < 0 \ ; \quad F(x) = 1, \quad \text{for} \quad x \geq 0 \tag{9}$$

Using scaling dependent on n, we can investigate the limit Eq. (7) of this function and find

$$\lim_{n\to\infty}\left(P(M_n < \frac{x}{a_n} + b_n) = F^n(\frac{x}{a_n} + b_n)\right) = \exp(-\exp(-(\frac{x}{a_n} + b_n) - \frac{k}{(x/a_n) + b_n} + \log n))$$

$$\cong \exp(-\exp(-\frac{x}{a_n} - b_n - \frac{k}{b_n}(1 - \frac{x}{a_n b_n}) + \log n)) \to \exp(-\exp-(x))$$

with two possible solutions for the scaling parameters, the first is

$$b_n = \frac{1}{2}\log n + \sqrt{(\frac{1}{2}\log n)^2 - k} \to \log n$$

$$a_n = 1 - \frac{k}{b_n^2} \to 1$$

and the second is

$$b_n = \frac{1}{2}\log n + \sqrt{(\frac{1}{2}\log n)^2 - k} \to \frac{k}{\log n}$$

$$a_n = 1 - \frac{k}{b_n^2} \quad a_n = -\frac{(\log n)^2}{k} > 0 \tag{10}$$

The convergence is independent of the value of the constant k for the first set of parameters, but for large n, $x/a_n + b_n$ jumps to the positive side of the x-axis where $F(x) = 1$. The parameters of the other set depend on k, but $x/a_n + b_n$ stays on the negative side of the x-axis for large n where $F(x) < 1$. The distribution Eq. (9) is therefore in the domain of attraction of the max-stable $EV1$ for all $k < 0$. Taking k very close to zero will make the distribution function Eq. (9) follow $EV1$ very closely up to the point $x/a + b = 0$, but just before that point $F(x)$ makes a jump up to 1, but without becoming discontinuous in any point. But for $k = 0$, $F(x)$ will make a finite jump from the value e^{-1} to 1 in $x/a + b = 0$. Using the non-linear transformation Eq. (5) maps the interval $(-\infty,\infty)$ in a one to one manner on the interval $(-\infty,0)$. But we can make the distribution of x follow the $EV1$ distribution as closely as we wish up to zero.

Letting k tend to zero, makes the distribution Eq. (9) a $EV1$ distribution cut off at $x/a + b = 0$. Cutting off at a higher value of x does not make any difference for the limit Eq. (10) or the conclusion above, but it makes the distribution function become in the limit $k = 0$

$$F(x) = \exp(-\exp(-y)); \quad y < y_{\lim}; \quad F(y) = 1; \quad y \geq y_{\lim} \tag{11}$$

$$y = \frac{x}{a} + b \quad y_{\text{lim}} = \frac{x_{PM}}{a} + b$$

Using Eq. (11) instead of $EV1$ we are using a distribution function that contains the PM value. If we use the distribution Eq. (11) to predict probabilities of rare events, *e.g.* 1000-year flood, we can use it exactly as we use the $EV1$ distribution, but there is the important difference from $EV1$, that $F(x)$ in Eq. (11) stops in $x = x_{PM}$ and makes the finite jump there, from exp(-exp($-y_{\text{lim}}$)) to 1. The position of the jump is at $y = y_{\text{lim}}$ and depends on probability (return period) only. In the following we shall refer to y_{lim} as the limiting reduced variate. Eq. (11) may be called the *cut-off* $EV1$ d.f. just to have a name that is different from the name *truncated* d.f. that is sometimes used for d.f.'s where the probability in the cut off tail is distributed downwards. Eq. (11) will be used in the parameter separation process mentioned earlier.

In statistics, one would like to use the full transformation Eq. (5) instead of the cut-off Eq. (11). Then one must estimate k, a and b from the data. It may prove difficult to estimate k, a and b when the data follows $EV1$ closely, and does not deviate from it until the reduced variate is very high, as often is the case. Distributions with a small k, and cut off distributions will inevitably resemble each other, so it is possible to get around the problem of estimating k by using the cut off $EV1$ distribution Eq. (11). In a moment table in the appendix are listed three coefficients, $C1$, $C2$ and $C3$, that are to be used when a and b are estimated on basis of average value and standard deviation of the data. It may be seen that the coefficients have different values for different y_{lim} values, but they are so close to the $EV1$ value for high values of return period (and close to each other), so the estimate of a and b will not be largely effected by the value of y_{lim} if y_{lim} is only large enough. It is also possible to estimate a and b linear regression, this makes the estimate independent of $C1$, $C2$ and $C3$ in the appendix, *i.e.* independent of the actual value of y_{lim}.

Sometimes we may be able to obtain knowledge of y_{lim} in Eq. (11) from other sources. If we can estimate x_{PM}, y_{lim} will follow from Eq. (11), and *vice versa*, *e.g.* by using the linear regression estimate or a and b. In the following we shall see, how the Hershfield and NERC methods provide us with *a priori* regional information of y_{lim}

The Hershfield Method

In the Hershfield Method (Hershfield 1965; WMO 1986) the PMP value is written as

$$x_{PM} = x_M + K\ s \tag{12}$$

Where

x_{PM} – The *PMP* value
X_M – The average value of annual maxima
s – The standard deviation of annual maxima
K – Hershfields frequency factor. (A function of x_M)

Comparison of Eqs. (11) and (12) shows that K may be defined as

$$K = a\frac{y_{\lim}-y_M}{s} \tag{13}$$

where

y_M – reduced variate of x_M

This relation may then be used to find the $y_{\lim}$ that corresponds to a given K. Using the value 15 for K as originally proposed by Hershfield (1965), and a and s from the $EV1$ distribution (s/a = 1.28 and y_M = 0.5771) we get the value

$$y_{\lim} = 19.8 \tag{14}$$

This corresponds to a return period of over a billion years so it is a very high value, even for the East Coast of USA where humidity is high and the convective part of the precipitation is high too.

The NERC Method

The NERC method originates from Britain's FSR (1975), and is used in Britain. It has been adapted to Norwegian conditions for use there (Förland and Kristoffersen 1988; 1989).

The analysis in the FSR (1975), is based on the $M5$ parameter, defined as the annual precipitation maximum with return period of 5 years. To find this value, the reduced variate is defined according to $EV1$'s extreme value distribution. This makes the $M5$ value equal to the arithmetic average of the upper half of the raw data, which is a more stable estimate than just picking the 5-year value from the raw data.

The basic formula used in the FSR (Vol. 2 p. 16) is

$$\frac{x}{M5} = \exp(C(y-1.5)) \tag{15}$$

where y is $EV1$'s reduced variate (see Eq. (2)) and C a regional parameter depending on $M5$.

Taking the logarithm of both sides of Eq. (15), using that $EV1$'s reduced variate of $M5$ is 1.5 gives

$$\log x = \log M5 + C(y-y_{M5}) \tag{16}$$

The analysis in the FSR (1975) bring about two main points. First it is established that C in Eq. (15) is a function of $M5$ only. This is important, as this means that the slope of the line Eq. (16) and the starting point on either axis, must be related. This point is discussed further in the case study in the next chapter.

The second point is, that the *growth factor* $x/M5$ depends on $M5$ in such a way that there exists an envelope curve that the growth factor does not exceed. This means that the envelope represents the highest probable values of x. The envelope curve is a function of $M5$, this is shown by Fig. 2.4 in Volume 2 of the FSR (1975) and discussed in the report. The value of the return period T, and the corresponding probability P, of the envelope curve may be determined as a function of $M5$, this shows that y_{lim} is a function of $M5$. As y in Eqs. (15) and (16) is $EV1$'s reduced variate, following approximate formula for this function may be evaluated

$$y_{\text{lim}} = 10.7 - 0.0071\, M5 \qquad (25 < M5 < 200) \tag{17}$$

Using Eq. (17) together with Eq. (15) and the regional values for C, the actual PMP values of Britain and Norway may be closely approximated within a few per cent error. In particular, PMP values for England and Wales in Table 4.2 in Vol. 2 of the FSR (1975) may be calculated by inserting the y_{lim} values into Eq. (15) together with the C deductible from Tables 2.7 and 2.9. The value of C may depend slightly on the region, it is different for Scotland and Northern Ireland, and Norway too. But Eq. (17) is always the same.

Eq. (17) is only valid in the quoted $M5$ range. Outside this range much less data are available and it has to be modified, but it still will be a funtion of $M5$.

A Case Study from Iceland

In Iceland, station values of the Probable Maximum Precipitation have been obtained using the Hershfield and NERC methods (Elíasson 1991). In order to obtain generalised estimates, the statistical distribution of 1-day annual maxima was investigated. Series of 1-day annual maxima of 11 stations, covering 60 years of observations, were selected. Each series was normalised by subtracting the average and dividing by the standard deviation. The 11 normalised series all have mean 0 standard deviation 1. Combining all 11 gives a series with 660 data points.

These data are shown in Fig. 1 with the normalised annual maxima plotted against $EV1$'s reduced variate. As can be seen the $EV1$ distribution fits the data very well. Other distributions fitted not so well, including log-$EV1$. Even though the data points are altogether 660, only one point shows a deviation from the line and it deviates downwards, opposite from what it should be were it indicating the existence of a maximum. The data is thus inadequate in order to predict an existence of a PMP value.

Then it was investigated if mean and standard deviations of the orginal 11 series were related. As the $EV1$ distribution fits the data so well, it was used in this

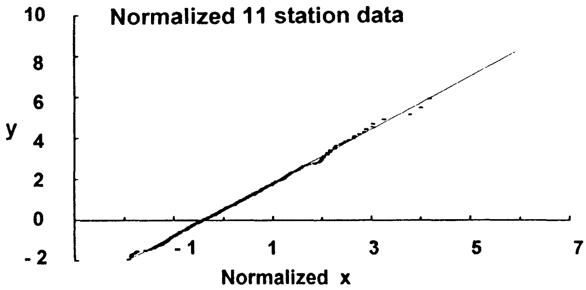

Fig. 1. $EV1$'s distribution fitted to 660 normalised 1-day annual maxima.

investigation, rewritten in the form Eq. (18) for convenience. Then it was investigated if C_i was related to $M5$ in a similar manner as the NERC method predicts it should be

$$x = M5 + C_i \, M5 \, (y - 1.5) \tag{18}$$

It turned out that stations with high $M5$ values tend to have high standard deviation too. Fig. 2 shows the 11 lines obtained when $EV1$'s distribution is fitted to data of the 11 stations. Visual inspection of Fig. 2 shows clearly that stations with high $M5$ also have a clear tendency for higher standard deviation.

The value of C_i can be calculated for each of the 11 lines. It is shown by the 11 points in Fig. 3. The empirical relation Eq. (19) is the broken line through the points

$$C_i = 0.1 + \frac{6}{M5} \tag{19}$$

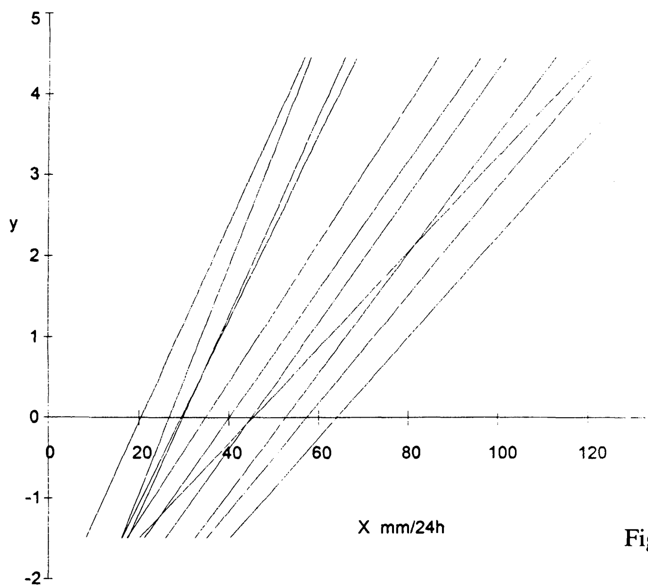

Fig. 2. $EV1$'s distribution fitted to 11-station data.

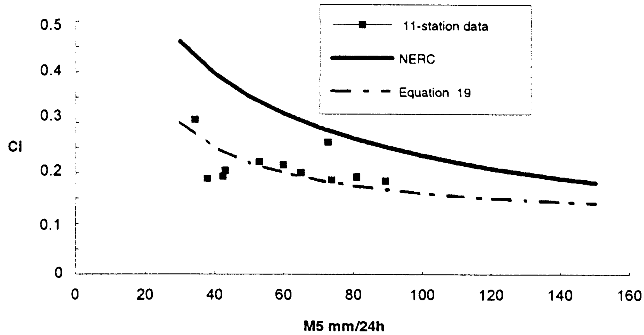

Fig. 3. 11-station data compared to NERC.

The solid line designated NERC in Fig. 3, is the C_i value calculated when the NERC *PMP* values are inserted in the left hand side of Eqs. (18) and (17) in the right hand side for the reduced variate y. This line shows how large C_i has to be, if Eqs. (17) and (18) should be used to calculate Icelandic *PMP*'s of the same value as NERC's. Comparison between the NERC curve and the similar Eq. (19) in Fig. 3 shows, that the NERC curve is indeed an envelope to the 11 data points.

Note that an independent estimate of C_i, using the 11 data points only, would result in a constant value, $C_i = 0.2$, because the data is inadequate to support the relation Eq. (19).

At this point the Hershfield and NERC methods can be compared using the result of the section on Hershfield's method. The first thing to note is the big difference in y_{lim}. The NERC value is about 10, at the most, but the Hershfield value about 20. The Hershfield *PMP* value will thus always be larger on the average than NERC, but the difference depends on the standard deviations as well as y_{lim}. We can use Eq. (12) to calculate $X_{PM,\ \text{Hershfield}}$, and Eq. (17) together with C from FSR (1975) and the NERC formula Eq. (16) to calculate $X_{PM,\ \text{NERC}}$. This makes it possible to calculate their ratio. Its average value for the Icelandic data is found to be 0.72 for $M5$ in the range 60 to 160mm/24h. Calculation of the average of the actual values of the 11 stations gives 0.76.

Conclusion

We can use statistical methods to calculate *PMP* values when an unbounded statistical distribution fits the data. We assume *a priori* that a limiting value exists, and then use a one to one transformation Eq. (5) that maps the unbounded interval $(-\infty,\infty)$ on the interval $(-\infty, PMP)$. Then the *PMP* value may be defined as the value of the distribution line $y = x/a + b$ that fits the data, evaluated for an empirical limiting reduced variate y_{lim}.

The definition above, makes it possible to derive the limiting values y_{lim} of the reduced variate for regions where there exist sufficient data to do so, and compare these values between regions and use them to compute *PMP* values where the available date are inadequate to derive local y_{lim} values.

In Iceland it has been possible to obtain valuable support to calculate *PMP* values this way. The *EV*1 distribution fits the data very well, and by comparing the y_{lim} values that may be derived from the NERC growth factor envelope curve used in England and Norway, it may be shown that the NERC curve envelopes the Icelandic data too. By using this y_{lim}, it is possible to arrive to the conclusion above, even though the Icelandic data is inadequate in order to define local y_{lim} values. The data available for the Icelandic region, can give the *M*5 values and a relation between the location parameter and the scaling parameter, but not an envelope curve for growth factors similar to the NERC curve. But the y_{lim} values obtained for Britain are used almost unchanged in Norway, and it is therefore concluded that they can be used for Iceland too.

References

Elíasson, J. (1991) Probable Maximum Precipitation in Iceland – Station Values, *Nordic Hydrology, Vol. 23(1),* pp. 49-56.

Förland, E., and Kristoffersen, D. (1988) Paaregnelig maksimal nedbör beregnet med ulike metoder. Fagrapport nr. 9/88 KLIMA. Det Norske Meteorologiske Institutt, Oslo (in Norwegian).

Förland, E., and Kristoffersen, D. (1989) Estimation of Extreme Precipitation in Norway, *Nordic Hydrology, Vol. 20(4/5),* pp. 257-276.

Hershfield, D. M. (1965) Method for Estimating Maximum Probable Precipitation, *Journal of American Waterworks Association, Vol. 57,* pp. 965-972.

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983) *Extremes and Related Properties of Random Sequences and Processes,* Springer Verlag.

Natural Environment Research Council (NERC) (1975) Flood Studies Report (FSR), Vol. II, London.

Sevruk, B., and Geiger, H. (1985) Distribution Types Currently in Use for Frequency Analysis of Extremes of Precipitation and Floods by Hydrological and Other Services, World Meteorological Organisation, Geneva.

WMO (World Meteorological Organisation) (1986) Manual for Estimation of Probable Maximum Precipitation; Second edition, WMO – No. 332, Operational Hydrology Report No. 1, Geneva.

First received: 14 September, 1993
Revised version received: 10 December, 1993
Accepted: 24 May, 1994

Appendix – The Cut-off Gumbel Distribution

The distribution function is defined as

$$F(y) = \exp(-\exp(-y)) \; ; \quad y < y_{lim} \; ; \qquad F(y) = 1 \; ; \quad y \geq y_{lim} \tag{A1}$$

The probability density function will be

$$f(y) = \exp(-\exp(-y))\exp(-y) \; ; \; y < y_{lim} \; ;$$
$$f(y) = (1-\exp(-\exp(-y_{lim}))) \; \delta(y-y_{lim}) \; ; \; y \geq y_{lim} \tag{A2}$$

Moments of $f(x)$; $x = a(y\text{-}b)$

$$Mn = \int_{-\infty}^{\infty} f(x)x^n dx =$$

$$\int_{-\infty}^{y_{lim}} e^{-e^{-y}} e^{-y}(a(y-b))^n dy + \int_{-\infty}^{y_{lim}} \delta(y-y_{lim})(1-e^{-e^{-y_{lim}}})(a(y-b))^n dy =$$

$$e^{-e^{-y}}(a(y-b))^n \Big|_{-\infty}^{y_{lim}} + (1-e^{-e^{-y_{lim}}})(a(y_{lim}-b))^n - na \int_{-\infty}^{y_{lim}} e^{-e^{-y}}(a(y-b))^{n-1} dy \tag{A3}$$

with

$$W_m(y_{lim}) = \int_{-\infty}^{y_{lim}} e^{-e^{z}} z^m dz \tag{A4}$$

Mn becomes

$$Mn = (a(y_{lim}-b))^n + n \sum_{i=0}^{i=n-1} \{ \binom{n-1}{i} (-1)^n W_{n-1-i} b^i a^n \} \tag{A5}$$

The two first moments become

$$M1 = \mu = a(C1-b) \quad M2 = \sigma^2 + \mu^2 = a^2(C2-bC3+b^2) \tag{A6}$$

The coefficients are

$$C1 = y_{lim} - W_0 \qquad C2 = y_{lim}^{\;2} + 2W_1 \qquad C3 = 2\,C1$$

y_{lim}	T, yrs	$C1$	$C2$	$C3$
1	3	0.241	0.601	0.481
2	8	0.446	1.188	0.893
3	21	0.528	1.584	1.056
4	55	0.559	1.796	1.118
5	149	0.570	1.897	1.141
6	404	0.575	1.943	1.149
7	1,097	0.576	1.964	1.153
8	2,981	0.577	1.972	1.154
10	22,027	0.577	1.977	1.154
12	162,755	0.577	1.978	1.154
$EV1,\infty$	∞	0.577	1.978	1.154

Address:
Háskóli Island,
Hjardarhagi 6, VI-II,
107 Reykjavik,
Iceland.