CHALMERS | UNIVERSITY OF GOTHENBURG

PREPRINT 2009:29

Estimation of return values for significant wave height from satellite data

IGOR RYCHLIK
JESPER RYDÉN
CLIVE ANDERSON

Department of Mathematical Sciences
Division of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Göteborg Sweden 2009

Preprint 2009:29

Estimation of return values for significant wave height from satellite data

Igor Rychlik, Jesper Rydén, Clive Anderson

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg
SE-412 96 Göteborg, Sweden
Göteborg, June 2009

Preprint 2009:29
ISSN 1652-9715

Matematiska vetenskaper
Göteborg 2009

Estimation of return values for significant wave height from satellite data

Igor Rychlik[†], Jesper Rydén[‡] and Clive Anderson[*]

[†] Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg

[‡] Department of Mathematics, Uppsala University

[*] School of Mathematics and Statistics, University of Sheffield

June 15, 2009

Abstract

Estimation of extreme wave height across the oceans is important for marine safety and design, but is hampered by lack of data. Buoy and platform data are geographically limited, and though satellite observations offer global coverage, they suffer from temporal sparsity and intermittency, making application of standard methods of extreme value estimation problematical. A possible strategy in the face of such difficulty is to use extra model assumptions to compensate for lack of data. In this spirit we report initial exploration of an approach to estimation of extreme wave heights using crossings methods based on a log-Gaussian model. The suggested procedure can utilize either intermittent satellite data or regular time series data such as obtained from a buoy, and it is adapted to seasonal variation in the wave height climate. The paper outlines derivation of the method and illustrates its application to data from the Atlantic and Pacific oceans. A numerical comparison is made with the results of an annual maximum analysis for sites at which both satellite and buoy data are available. The paper concludes with a discussion of the applicability of the new approach, its relationship to other extreme value methods and desirable directions for further development.

Keywords: Return values, wave heights, crossing intensity, Gaussian processes, seasonality

1 Introduction

Significant wave height, H_s, at a particular location and time is a measure of the sea state, a local average of the prevailing wave heights or, equivalently, of variability of the sea surface elevation at that place and time. Estimates of the large values of H_s

likely to be encountered are important parameters for the design of vessels and marine structures. In naval architecture, for example, knowledge of extreme H_s is needed to assess the response of a ship travelling the oceans and continuously suffering stresses shortening its fatigue life. Estimates of extreme H_s across the oceans also play an important rôle in the work of the Naval Classification Societies in rating vessels for operation in different parts of the world. These uses point up the need for estimates of extreme H_s over the whole marine globe

The severity of the wave climate is typically summarized in terms of *return levels*, taken in this context to be high quantiles of the distribution of the annual maximum significant wave height. Suppose that $z(t)$ denotes H_s at time t for a particular location and that $M(z) = \max_{1 \leq t \leq 1} z(t)$ denotes its annual maximum. Then the T-year return level of H_s at the location, denoted by z_T, is defined by

$$\mathsf{P}(M(z) > z_T) = \frac{1}{T}. \tag{1}$$

Estimation of z_T clearly demands knowledge of the upper tail of the distribution of $M(z)$. Observations of H_s for the estimation may be available at some locations where there are buoys or platforms. Since the resulting data are usually in the form of regular time series of several observations per day, methods of estimation based on annual maxima or Peaks Over Threshold (POT) may be used. Such data are, however, limited both in number and geographical extent. Alternatives offering global coverage are observations of H_s from satellites, and reconstructions of H_s from numerical ocean-atmosphere models based on large-scale meteorological data. In this paper we focus on the first of these, the closer to direct observation of significant wave height. We report on initial exploration of a new method aiming to estimate location-specific H_s return levels over large ocean areas from satellite observations. The emphasis is on wide spatial coverage over optimal estimation at individual locations. We comment further on the use of model data in Section 4.

Though satellite observations of H_s offer global spatial coverage, they suffer from restricted temporal coverage; the satellite returns to the same location only at intervals of the order of days, and the returns are not equally spaced (see, for example, Anderson *et al* 2002). Thus the observed values, $z(t_i)$, at the location are made at times t_i that are sparse and irregular, posing difficulties for application of standard annual maxima and POT estimation methods. Evidently extra information or stronger assumptions are needed to compensate for the data limitations. This paper explores how far an assumption of Gaussianity of $\ln z(t)$ can carry the estimation, and to what extent and to what rarity of events it is applicable.

Our approach is based on the theory of level-crossings of stochastic processes. We take $z(t)$ to be modelled as a stochastic process and use the following relationship between crossings of a level u and the distribution of the maximum over a period of time:

$$\mathsf{P}(M(z) > u) = \mathsf{P}(z(0) > u) + \mathsf{P}(N(u;\, z) > 0,\, z(0) \leq u) \tag{2}$$

where $N(u;\, z)$ is the number of upcrossings of the level u during unit time (here one year). For reasonable models of z, as u increases, the probability $\mathsf{P}(z(0) > u)$ and the restriction $z(0) \leq u$ in the second probability in (2) become negligible, so that for large u,

$$\mathsf{P}(M(z) > u) \approx \mathsf{P}(N(u;\, z) > 0). \tag{3}$$

Furthermore, writing N for $N(u;\, z)$,

$$\mathsf{E}[N] - \frac{1}{2}\,\mathsf{E}[N(N-1)] \leq \mathsf{P}(N > 0) \leq \mathsf{E}[N].$$

When $\ln z$ is a Gaussian processes satisfying some smoothness assumptions, the term $\mathsf{E}[N(N-1)]$ tends faster to zero than $\mathsf{E}[N]$ (see, for example, Chapter 4 in Azais & Wschebor (2009)), and so

$$\mathsf{P}(M(z) > u) \approx \mathsf{E}[N]. \tag{4}$$

The expectation in (4) can be evaluated by means of Rice's formula (an approach referred to here as Rice's method); see, for example Marcus (1977). In the present paper, approximations for the expectation are proposed based on estimation of total variation of the process and the modelling of seasonality. An initial version of the methodology is sensitive to estimation of the total variation, and so an improved version based on upcrossings of an interval and referred to as the modified Rice method is developed and used to find conservative estimates of return levels.

The outline of the paper is as follows. In Section 2, the framework of crossing methods for estimation is outlined, introducing the Rice and modified Rice methods, and describing assumptions and estimation methodology. In Section 3, an application to satellite data is given, indicating the advantage of the methodology when data are sparse and intermittent. For comparison, the methodology is also applied to buoy data and results are compared to those based on annual maxima. Finally, Section 4 reviews the current limitations and applicability of the crossings-based approach.

2 Crossing methods for estimation of return values

The purpose in this section is to find reasonable estimates of $\mathsf{E}[N]$ in Equation (4).

Rice's formula in the non-stationary case (see, for example, Marcus (1977)) states that

$$\mathsf{E}[N] = \int_0^1 \mu_t(u; z)\,\mathrm{d}t. \tag{5}$$

Here $\mu_t(u;z)$ is the *upcrossing intensity* of the level u by z, or equivalently of $\ln u$ by $\ln z$, given by Rice's formula

$$\mu_t(u;z) = \mu_t(\ln u; \ln z) = \int_0^\infty s f_{\ln z(t), \dot{\ln} z(t)}(\ln u, s)\, \mathrm{d}s \tag{6}$$

when z is a continuously differentiable process satisfying some regularity assumptions (see Leadbetter et al. (1983, Chapter 7)) and $f_{\ln z(t), \dot{\ln} z(t)}$ denotes the density function of $(\ln z(t), \dot{\ln} z(t))$.

If $\ln z$ is a stationary Gaussian process with mean m and variance σ^2, the integral in Equation (6) can be evaluated explicitly, yielding

$$\mu_t(u;z) = \mu(u;z) = \frac{1}{2\sqrt{2\pi}} \frac{\gamma}{\sigma} \, \mathrm{e}^{-(\ln u - m)^2/2\sigma^2} \tag{7}$$

where $\gamma = \mathsf{E}[|\dot{\ln} z(t)|]$. The quantity γ can be estimated by the sample total variation

$$\gamma^* = \frac{1}{T_{\text{obs}}} \sum |\ln z(t_i) - \ln z(t_{i-1})|, \tag{8}$$

where T_{obs} is the length of the observation period in years and the t_i are the time instants of successive measurements.

In the following, the aim is to evaluate the expression in (6) in non-stationary cases. A model for the time-variation of the intensity in (6) is needed. Here a standard model incorporating seasonality is used, as presented in the following subsection.

2.1 Framework for seasonal modelling

Study of satellite measurements (Baxevani *et al* (2005)) suggests that for North Atlantic locations, $\ln z(t)$ can be reasoanably approximated by a normally distributed variable with seasonally varying mean and constant variance $\sigma^2 = \mathsf{V}[\ln z(t)]$. (Elsewhere, for example at Hawaiian buoy 51001, it has been observed that $\mathsf{V}[x(t)]$ depends on the season.) Accordingly, representing the mean by a simple cosine curve, we assume

$$\ln(z(t)) = m_0 + A\cos(2\pi t + \phi) + x(t) = m(t) + x(t), \tag{9}$$

where m_0 and A are constants, $m(t)$ denotes $\mathsf{E}[\ln z(t)]$ and $x(t) \sim \mathrm{N}(0, \sigma^2)$. The functions m and σ^2 are taken to depend on location. In the sequel we further assume that $m(t)$ varies much more slowly than $x(t)$ and that $x(t)$ is a stationary Gaussian process, at least approximately during the season of severe storms dominating the estimation of return values z_T.

2.2 Rice's method: computation of $\mathsf{E}[N]$

Assuming the model in (9), we find an approximation for $\mathsf{E}[N] = \mathsf{E}[N(u; z)]$ as follows. Obviously,

$$\frac{\mathrm{d}}{\mathrm{d}t} \ln z(t) = \dot{m}(t) + \dot{x}(t)$$

in distribution. For satellite data, often $|\dot{m}(t)| < 0.01\mathsf{E}[|\dot{x}(t)|]$ so that it is a reasonable approximation to take $\dot{m}(t)$ to be negligible relative to $\dot{x}(t)$. Thus

$$\mu_t(\ln u;\, \ln z) \approx \mu_t(\ln u - m(t);\, x) \approx \mu(\ln u - m(t);\, x), \tag{10}$$

where μ is the expression (7) for the stationary case. From (7) and (10) we arrive at the desired approximation

$$\mathsf{E}[N] \approx \frac{1}{2\sqrt{2\pi}} \int_0^1 \frac{\gamma}{\sigma}\, \mathrm{e}^{-[\ln(u)-m(t)]^2/2\sigma^2}\, \mathrm{d}t. \tag{11}$$

If σ^2 is small, then for high levels u a close approximation to the integral in (11) may be found (see Appendix) using Taylor's formula, giving the more explicit expression:

$$\mathsf{E}[N] \approx \frac{1}{4\pi} \frac{\gamma}{\sqrt{A(\ln(u) - m_0 - A)}}\, \mathrm{e}^{-(\ln(u)-m_0-A)^2/2\sigma^2}. \tag{12}$$

The required return level estimate z_T^* may be found by solving for u (replacing parameters by their estimates) in

$$\frac{1}{4\pi} \frac{\gamma}{\sqrt{A(\ln(u) - m_0 - A)}}\, \mathrm{e}^{-(\ln(u)-m_0-A)^2/2\sigma^2} = \frac{1}{T} \tag{13}$$

(or in a similar expression in which the left hand side is replaced by the integral which it approximates, evaluated numerically) where σ^2 is the variance of $x(t)$ and γ is the total variation in the stormy period. In practice, (11) and (12) lead to almost identical estimates of z_T. We call this method of estimation of return levels (by either (11) or (12)) *Rice's method.*

2.3 An Example

To examine the practicality of return level estimation based on (13) we discuss the case in which the residual process $x(t)$ is modelled as an Ornstein-Uhlenbeck process. This is an interesting process because it has very irregular sample paths, so will give a searching test of the procedure.

Example: Ornstein–Uhlenbeck process
The Ornstein–Uhlenbeck process is a zero-mean stationary Gaussian Markov process with covariance function

$$r(\tau) = \sigma^2\, \mathrm{e}^{-\alpha|\tau|},$$

where we take the time variable τ to be measured in seconds. For simplicity let $\alpha = 1$, $\sigma^2 = 1$. Moreover, assume that $m_0 = 0$ and $A = 0$ in (9). In actuality observations are made at discrete times, and so the model used is a discrete skeleton of the continuous time process. Thus the values $x_i = x(i\,\Delta t)$, where Δt is the discretization step, form an AR(1) time series. This is not a totally unrealistic model for measured signals from buoys, but the specific interest for the present discussion is that the model has very irregular sample paths so that $\mathsf{E}[\gamma^*]$ diverges to infinity as the discretization step approaches zero, and hence the currently proposed Rice method is likely to face difficulties.

Suppose that we wish to find the 100-year level for x; that is, the solution of the equation $\mathsf{P}(M(x) > u) = 0.01$. For such long periods the asymptotic result derived by Pickands (1969) is useful:

$$\lim_{S\to\infty} \mathsf{P}\big(\sqrt{2\ln(S)}\,(M_S(x) - k_S) \le u\big) = \exp(-\mathrm{e}^{-u}),$$

where M_S denotes the maximum over time S, and

$$k_S = \sqrt{2\ln(S)} + \frac{\ln(\ln(S)) - \ln(\pi)}{2\sqrt{2\ln(S)}}.$$

From this, with $S = 3.15 \times 10^7$, the 100-year value can be approximated by

$$x_{100}^{\mathrm{As}} = k_S + (-\ln(-\ln(1 - 1/100)))/\sqrt{2\ln(S)} = 6.8.$$

Suppose that we have sampled x every Δt seconds. Then

$$\mathsf{E}[\gamma^*] = \frac{2\sqrt{2(1 - r(\Delta t))}}{\Delta t\sqrt{2\pi}},$$

and hence, with

$$f_0 = \frac{\sqrt{2(1 - \exp(-\Delta t))}}{2\pi\,\Delta t},$$

the Rice method of Section 2.2 would be expected to give the estimate of the 100-year value as $x_{100}^{\mathrm{Ri}} = \sqrt{2\ln(100\,T\,f_0)}$, even without sampling variation. For $\Delta t = 1, 0.1, 0.01, 0.001$ we find that x_{100}^{Ri} is equal 6.35, 6.56, 6.74, 6.90, respectively. The asymptotic formula shows that these levels correspond, approximately, to return periods of 7.3, 24.0, 66.7, 179.5 years respectively, somewhat different from the nominal 100 years. Thus, unsatisfactorily, the estimates γ^* depend sensitively on observation frequency. □

2.4 Modified Rice method

The example above shows that the Rice method, being based on individual crossings of a level, is sensitive to local fluctuations of $\ln z$. An alternative is to base return level estimation on ‘storms’: episodes of generally raised sea-surface levels separated by calm periods of low levels. Compared to the Rice method the storms approach focusses on longer time periods and larger-than-local fluctuations during them, avoiding potential problems of unbounded total variation.

Mathematically speaking, we define a storm in terms of a pair of reference levels, $u_0 < u$ say; then a storm is an excursion above u_0 that also exceeds u. The number of storms during an observation interval of length T_{obs} is therefore the number of upcrossings of the interval $[u_0, u]$, as illustrated in Figure 1. We use the notation $N^{\text{S}}_{T_{\text{obs}}}(u_0, u; z)$ for this number and we adopt the convention that $N^{\text{S}}_1(u_0, u; z) = N^{\text{S}}(u; z)$. Suppose now that v is a level exceeding u. Neglecting the probability that we start to observe within a storm, we find by the same argument as for (3) the approximation:

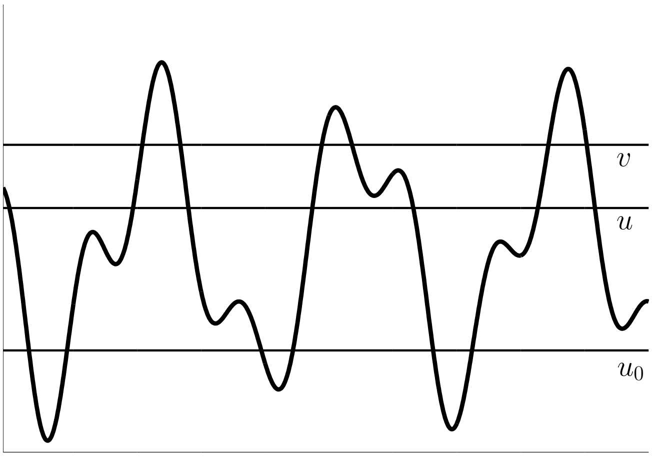

Figure 1: Illustration of crossings related to storms. Here, $N^{\text{S}}_{T_{\text{obs}}}(u_0, u; z) = 3$

$$\mathsf{P}(M(z) > v) \approx \mathsf{P}(N^{\text{S}}(v; z) > 0),$$

with equality as v approaches infinity. As before, the inequality

$$\mathsf{P}(N^{\text{S}}(v; z) > 0) \leq \mathsf{E}[N^{\text{S}}(v; z)]$$

is employed and a new estimate, called z_T^{Osc}, of the return value is defined as the solution to the equation

$$\mathsf{E}[N^{\text{S}}(z_T^{\text{Osc}}; z)] = \frac{1}{T}.$$

Since $\mathsf{P}(N^{\text{S}}(v; z) > 0) \leq \mathsf{E}[N^{\text{S}}(v; z)] \leq \mathsf{E}[N(v; z)]$, it follows that z_T^{Osc} is smaller than the estimate derived using Rice's method, but still conservative.

Example, continued

For the Ornstein-Uhlenbeck process x it is known (Rychlik 1996) that

$$\mathsf{E}[N^{\text{S}}(v; x)] = \frac{\alpha}{\sqrt{2\pi}} \left(\int_{u_0/\sigma}^{v/\sigma} \mathrm{e}^{\tau^2/2} \, \mathrm{d}\tau \right)^{-1}. \tag{14}$$

Taking $u_0 = 0$ and solving the equation $\mathsf{E}[N^{\mathrm{S}}(v)] = 0.01$ we obtain an upper bound for the 100-year return level x_{100}. In the present example the bound, x_{100}^{OU} say, is equal to 6.76, which is close to, but smaller than, the asymptotic estimate $x_{100}^{\mathrm{As}} = 6.8$. □

In general, except in some simple cases and Markov processes, no explicit formula for $\mathsf{E}[N^{\mathrm{S}}(v; z)]$ is known, so we seek an approximation. To do so we borrow an idea from Peaks-Over-Threshold methodology, making use of the intermediate threshold u with $v > u > u_0$. At high levels of v we expect exceedances of v to be rare isolated events, so that the proportion of storms (upcrossings of $[u_0, u]$) in which v also is exceeded, as measured say by $\mathsf{E}[N^{\mathrm{S}}(v; z)]/\mathsf{E}[N^{\mathrm{S}}(u; z)]$, can be expected to be approximately the same as the proportion of upcrossings of the level u that lead to upcrossings of v, $\mathsf{E}[N(v; z)]/\mathsf{E}[N(u; z)]$; and these proportions can be expected to become closer as v increases. This suggests the following as a reasonable approximation:

$$\mathsf{E}[N^{\mathrm{S}}(v; z)] \approx \mathsf{E}[N^{\mathrm{S}}(u;\, z)]\, \frac{\mathsf{E}[N(v; z)]}{\mathsf{E}[N(u; z)]}. \tag{15}$$

We now turn to return level estimation. The intensity of storms, $\mathsf{E}[N^{\mathrm{S}}(u;\, z)] = \lambda_u$ say, can be estimated by

$$\lambda_u^* = \frac{1}{T_{\mathrm{obs}}} N_{T_{\mathrm{obs}}}^{\mathrm{S}}(u; z).$$

Also, using (12) with $u_0 = \exp(m_0 + A)$, we find for the ratio of expectations on the right-hand side of (15)

$$\frac{\mathsf{E}[N(v; z)]}{\mathsf{E}[N(u; z)]} \approx \sqrt{\frac{\ln(u) - \ln(u_0)}{\ln(v) - \ln(u_0)}}\, \mathrm{e}^{-\left[(\ln(v)-\ln(u_0))^2 - (\ln(u)-\ln(u_0))^2\right]/2\sigma^2} \tag{16}$$

A convenient choice for u is $u = u_0 \exp(c\sigma)$, where c is a constant to be chosen. Then the T-year return level z_T satisfies

$$z_T = u_0 \exp(c_T \sigma) = \exp(m_0 + A + c_T \sigma), \tag{17}$$

where, by (16), c_T is the solution to

$$\lambda_u T = \sqrt{\frac{c_T \exp(c_T^2)}{c \exp(c^2)}}. \tag{18}$$

The return level z_T given by (17) and (18) is a function of the intensity of storms λ_u, and hence of the threshold $u = \exp(m_0 + A + c\sigma)$. We can estimate z_T by replacing unknown parameters in the determining equations by their own estimates. The parameters σ^2, A and m_0 can be estimated for any location from satellite data as described in Baxevani *et al* (2005). Once u is determined, the intensity λ_u can be estimated directly from observations as described above. Thus the remaining issue is the determination of the constant $c > 0$. In general this is a matter to be explored empirically. We discuss it in the light of findings from buoy data in Section 3.3.

Example, continued further
We apply the modified Rice method to the OU-process, calculating the intensity of storms by means of (14),

$$\lambda_u = \mathsf{E}[N^{\mathrm{S}}(u)] = \frac{1}{\sqrt{2\pi}} \left(\int_0^u \mathrm{e}^{\tau^2/2} \, \mathrm{d}\tau \right)^{-1} .$$

For $u = 2, 3, 4, 5, 6$ the estimates of x_{100} are found from (17) to be $6.54, 6.62, 6.67, 6.71$ and 6.74, respectively. The values should be compared with the bound $x_{100}^{\mathrm{OU}} = 6.76$ found earlier. We conclude that the method shows some promise even for this extremely irregular process. □

3 Implementation and Applications

In this section we use satellite and buoy data to explore the choice of c in the estimation of λ_u discussed in §2.4. We also illustrate application of the modified Rice method and compare its results to those of an analysis based on annual maxima.

3.1 Data

The satellite (altimeter) data used are measurements of H_s made by the TOPEX/Poseidon satellite at discrete locations along one-dimensional tracks over the oceans at different periods between October 1992 and January 1999. The data were obtained from the Southampton National Oceanography Centre (NOC). Drift in the TOPEX observations for 1997 to 1999 was corrected by the method of Challenor and Cotton (1999) and Caires and Sterl (2005).

The buoy data are from the National Data Buoy Center (NDBC) (`http://www.ndbc.noaa.gov`). Data from the six buoys listed in Table 1, all located in the northern hemisphere (see Figure 2), were used. Observations were hourly. As is often the case with buoy data, gaps in the time series occur; see Table 1. Moreover, the stormy seasons of the year often contained long time intervals without data (for these buoys: 8 in 21 years, 9 in 22, 9 in 20, 5 in 21, 6 in 20, 10 in 20 respectively).

Table 1. Buoys from NDBC

Buoy	Location	Observed period	Percentage missing obs.
44004	N 38.48, W 70.43	1 Jan 1983 – 31 Dec 2003	20
44005	N 43.19, W 69.16	1 Jan 1982 – 31 Dec 2004	20
44011	N 41.11, W 66.58	1 Jan 1985 – 31 Dec 2005	16
46001	N 54.31, W 146.53	1 Jan 1983 – 31 Dec 2003	12
46005	N 46.85, W 131.02	1 Jan 1983 – 31 Dec 2003	18
51001	N 23.43, W 162.21	1 Jan 1982 – 31 Dec 2002	22

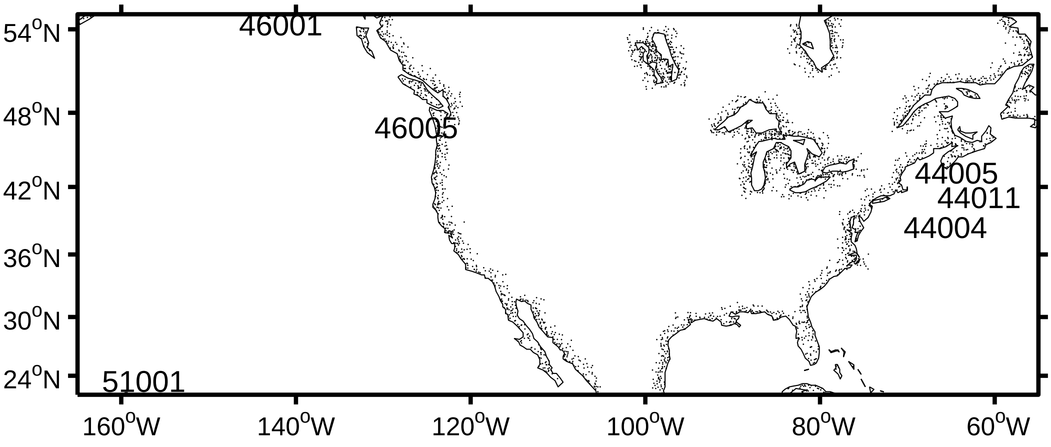

Figure 2: Locations of the buoys considered (Mercator projection).

3.2 Seasonal components and the residual process

An estimate $\hat{m}$ of the mean function $m(t) = m_0 + A\cos(2\pi\, t + \phi)$ was found for each of the six buoys, and the residuals $x(t) = \ln(z(t)) - \hat{m}(t)$ extracted. Normal QQ plots of these residuals taken approximately every fifth day (to avoid possible dependence effects) were examined. Figure 3 focusses on the upper 5% portions of these plots. The solid line in each plot corresponds to a Normal distribution with zero mean and a standard deviation equal to the sample standard deviation of the sampled residuals. The dotted lines are estimates of pointwise 2.5% and 97.5% percentiles of QQ plots from such a Normal distribution obtained by simulation of 1000 samples. It is seen that the observations from five of the buoys appear broadly consistent with the assumption of a stationary Normal distribution for $\ln H_s$ after seasonal mean-correction. In the case of Buoy 46005 there must be some doubt about such a model; residuals appear heavier-tailed than Normal in this case. On the other hand for Buoy 51001 the known seasonal change in variance appears not to result in departure of the largest residuals from a normal-induced model (a feature consistent with the upper tail of the distribution of residuals being dominated by a particular season). Though apparent consistency with assumption in five cases is far from confirmation of a stationary Normal model for residuals (not least because the hypothesis is empirically-based), nevertheless it motivates further examination of the crossings approach.

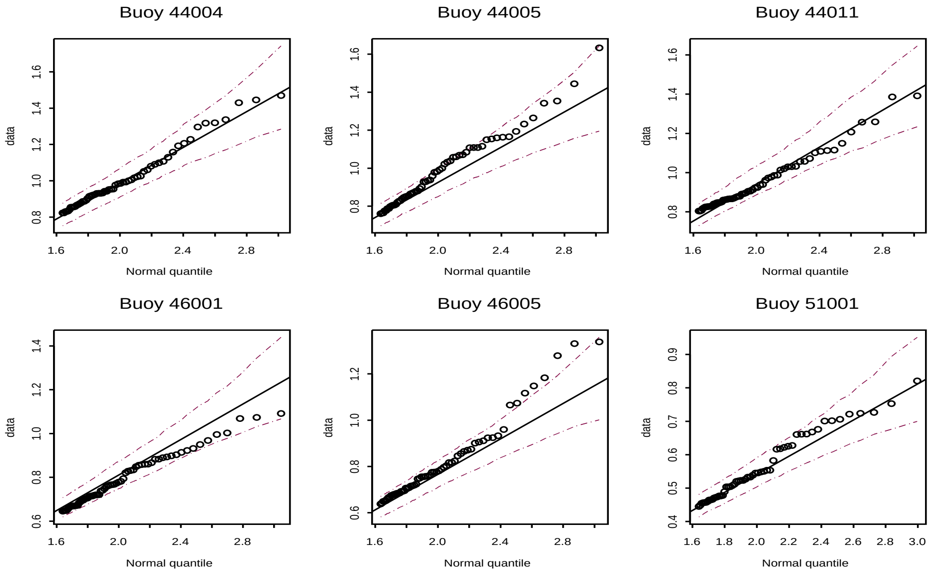

Figure 3: Upper 5% portions of Normal QQ plots of residual values. The solid line corresponds to a Normal distribution with zero mean and standard deviation matching that of the residuals. Dotted lines show estimated pointwise 95% limits based on 1000 samples simulated from this Normal distribution.

3.3 Choice of c and estimation of the intensity λ_u of storms

Once the seasonal model is fitted, estimation of λ_u in the modified Rice method of §2.4 hinges on choice of c. We therefore investigate, using the buoy data, how estimation of λ_u and, from it, return levels obtained by the modified Rice method are affected by different choices of c, taking $u = \exp(m_0 + A + c\,\sigma)$.

Temporarily we write λ_c for λ_u, and calculate the estimate λ_c^* for each buoy at $c = 2.0, 2.1, \ldots, 2.9$, then use the modified Rice method to calculate the resulting estimated 100-year return levels. Figure 4 shows the estimated z_{100} in relation to c. In the figure the dots correspond to estimates based on (18) with λ_u^* estimated individually, while the crosses show estimates obtained by replacing λ_u^* for each buoy and value of c by the average $\bar{\lambda}_c^*$ of the estimates for all six buoys We note that the two sets of estimates are quite similar; differences are mostly within one metre. These results give some grounds for hope that there is approximate regional stability in λ_u. Consequently, we propose to use (18) with the value $\bar{\lambda}_c$ deduced from Figure 4 to estimate 100-year values in regions where λ_u^* is not available. A balance has to be struck between model error, if c is too low, and lack of data if c is too high. In the rest of this paper we use $c = 2.2$ for the calculations using the modified Rice method.

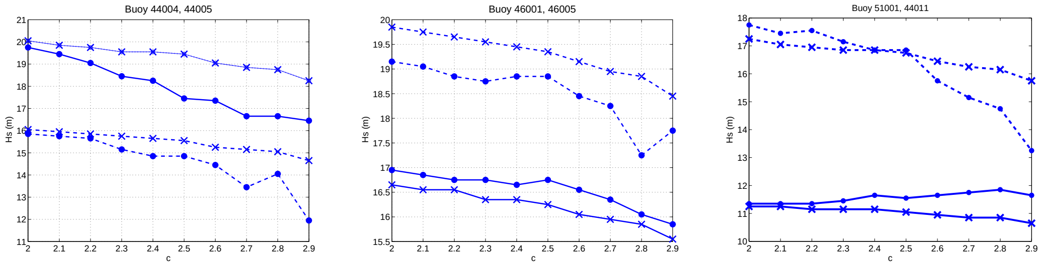

Figure 4: Estimates of 100-year significant wave height as estimated by modified Rice's method as a function of c. Dots: based on λ_u^*; X-marks: based on $\bar{\lambda}_u$. Left panel: Buoy 44004 (solid), Buoy 44005 (dashed). Centre panel: Buoy 46001 (solid), Buoy 46005 (dashed). Right panel: Buoy 51001 (solid), Buoy 44011 (dashed).

3.4 Global estimation of return levels from satellite observations

Baxevani, Caires and Rychlik (2009) give estimates of the seasonal parameters m_0, A, σ around the globe, based on satellite data. Together with the chosen λ_u^* of §3.3, these yield global conservative estimates of return levels of H_s by the modified Rice method. Figure 5 shows the estimated 100-year return levels. There are large uncertainties in estimated parameters for some locations; the effects are particularly visible for the Mediterranean between France and Sardinia where the estimated 100-year H_s is very large.

Satellite observations are available in particular at the locations of five of the buoys considered in §3.3 (the exception being buoy 46005). For comparison the estimated 100-year return levels (m) found by the modified Rice method for the parallel data sets are presented in Table 2.

Table 2. 100-year return values (in m) estimated by the modified Rice method ($c = 2.2$) based respectively on satellite data and buoy data

Buoy	44004	44005	44011	46001	51001
Satellite data	16.5	16.4	16.0	17.0	8.2
Buoy data	19.7	15.9	17.0	16.6	11.2

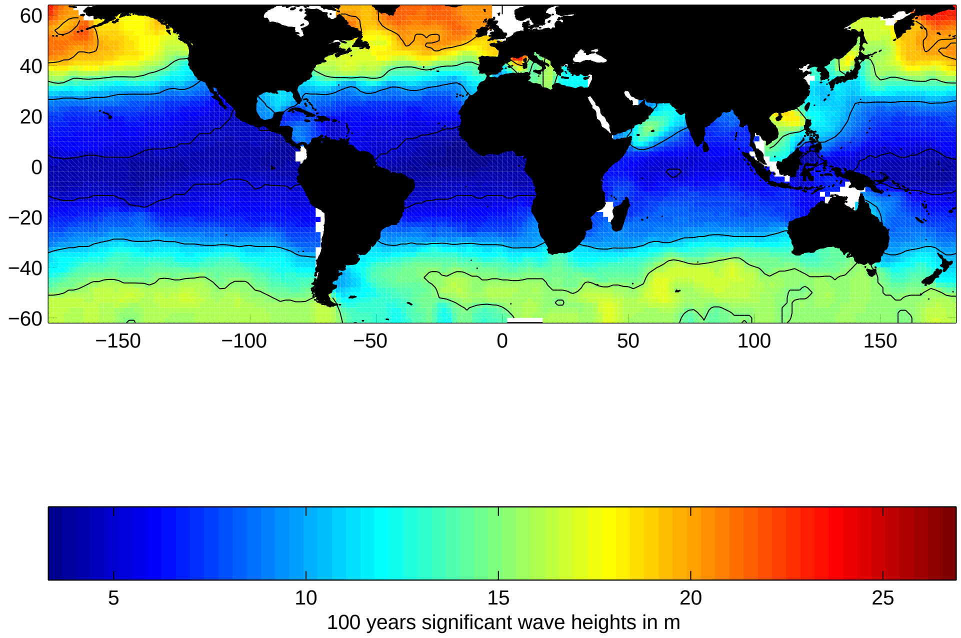

Figure 5: Estimated z_{100} by modified Rice's method. Contour lines are placed at 5, 10, 15 and 20 meters.

In Table 2 the modified Rice method was used with the storm intensity $\bar{\lambda}_c = 2.2$. The parameters were estimated using satellite data and buoy data, upper and lower rows, respectively. The estimates of parameters based on satellite data uses spatio-temporal measurements from 4 by 4 degree space regions and hence this model are based on smoother data, rendering less variability than the estimates based on buoys which are at fixed locations. However, the estimated 100-year significant wave heights are relatively close although conservative.

3.5 Comparison with method of annual maxima

Although the modified Rice method is proposed with satellite data in view, its use with buoy data allows it to be compared with the classical method based on the fitting of a Gumbel, or, more generally, a Generalized Extreme Value (GEV) distribution to annual maxima. Accordingly maximum likelihood return level estimates for 100 and 1,000 years found by the annual maxima approach based on Gumbel and GEV distributions are shown in Table 3 alongside corresponding estimates found by the modified Rice method. In the table, the estimated crossing intensity λ_c^* with $c = 2.2$ was used. For the annual maxima analyses we used seasonal years, beginning on 1 July. If, as here, substantial numbers of observations are missing during the stormy parts of the year, the observed annual maximum is likely to be less than the true maximum, and so we expect return levels to be under-estimated by an annual maximum method.

Table 3. Return values (in m) by different methods.

Buoy	Gumbel		GEV		Mod. Rice, $c = 2.2$	
	x_{100}	x_{1000}	x_{100}	x_{1000}	x_{100}	x_{1000}
44004	13.24	15.57	17.44	30.40	19.05	25.35
44005	11.08	13.36	10.30	11.61	15.65	20.85
44011	13.28	15.88	11.96	13.04	17.55	22.65
46001	15.61	18.65	13.44	14.30	16.75	20.55
46005	16.79	20.29	13.74	14.26	18.85	23.65
51001	13.65	16.76	12.11	13.51	11.35	13.85

We observe relatively large spreads in the estimates of the 100-year values using the annual maximum method, although the values are of the same magnitude.

4 Discussion

In this section we discuss limitations, applicability and possibilities for further development of the modified Rice method. A relationship with POT methods is described.

4.1 Precision and Seasonality

Standard errors: A limitation of the method as presented here is that it does not provide estimates of precision of the return level estimates. Precision estimates are, of course, essential for comparisons and for rational decision-making. A promising way to find them here appears to be by a bootstrap approach in conjunction with uncertainty estimates from the seasonal modelling. Further development in this direction is needed.

Seasonality: It would clearly be possible – and desirable – to model seasonality more flexibly than by a single sine-cosine function. Multi-frequency harmonic models and non-parametric regression models could be accommodated easily within the approach. The constancy of variance assumed in the seasonal model is similarly not essential; a varying σ^2 could be accommodated in the procedures at the expense of some extra numerical complexity. The treatment of seasonality in the approach, however, whether with simple or sophisticated models, tacitly assumes that seasonality in the extremes is the same as that in everyday observations, a strong assumption. For application, this is something to be checked, a substantial task for the whole globe.

4.2 The Gaussian assumption and relationship to the POT method

The Gaussian assumption enters through expression (12) which is used in (16) to find z_T from (17) and (18). Its rôle is to connect information about properties (of the physical processes generating H_s) at the moderate level u to information about properties at more extreme levels v. In essence the modified Rice method uses an empirical estimate of properties at an accessible level u and extrapolates to higher levels on the basis of an assumed Gaussian distribution linking u to higher levels. In this it is similar in spirit (though not in its linking distribution) to the POT method, as the following shows.

In the stationary case the POT method can be derived from an assumption that storms associated with exceedance of level u_0 occur through time according to a homogeneous Poisson process of rate ν_0, say, and have, independently, storm peaks S whose excesses over u_0 are distributed according to a Generalized Pareto distribution, so that

$$\mathsf{P}(S > s) = \{1 + \xi(s - u_0)/\tau\}^{-1/\xi}, \quad (s > u),$$

for suitable parameters $\tau > 0$ and ξ. Under this model the two-dimensional point process with points (t_i, s_i), where the t_i are the occurrence times of storms and s_i the corresponding peak sizes, is a Poisson process in the plane for which the number of points in regions of the form $[0, 1] \times (s, \infty)$ for $s > u_0$ is Poisson distributed with mean $\nu_0 \mathsf{P}(S > s)$. Suppose as before that $v > u > u_0$. Then the probability that the maximum H_s over a year exceeds v, being the probability that at least one storm peak exceeds v, is

$$\begin{aligned} \mathsf{P}(M > v) &= 1 - \mathrm{e}^{-\nu_0 \mathsf{P}(S>v)} \\ &\approx \nu_0 \mathsf{P}(S > v) \\ &= \nu_0 \mathsf{P}(S > u)\, \mathsf{P}(S > v | S > u) \end{aligned} \tag{19}$$

for v making the probability small. Storms associated with exceedances of level u are obtained by thinning those associated with level u_0 and so occur in a Poisson stream of rate $\nu_0 \mathsf{P}(S > u) = \nu_u$ say. Thus the probability (19) may be written as

$$\mathsf{P}(M > v) \approx \nu_u \frac{\mathsf{P}(S > v)}{\mathsf{P}(S > u)}. \tag{20}$$

To account for non-stationarity, the rate of occurrence of storms, ν_0, and the parameters ξ and τ of the generalized Pareto distribution are taken to vary with time. Similar arguments then show that (20) generalizes to

$$\mathsf{P}(M > v) \approx \int_0^1 \nu_u(t) \frac{\mathsf{P}(S_t > v)}{\mathsf{P}(S_t > u)} \,\mathrm{d}t, \tag{21}$$

where S_t denotes the generalized Pareto-distributed storm peak at time t. The POT method chooses a threshold u (possibly itself time-dependent) above which there are sufficient data and the generalized Pareto distribution fits well. It then estimates the rate function $\nu_u(t)$ empirically from the observed frequency of exceedances, estimates the generalized Pareto distribution from the magnitudes of the excesses, and obtains return level estimates from (21) or (19).

The modified Rice method on the other hand has (from (15))

$$\mathsf{P}(M > v) \leq \mathsf{E}[N^S(u; z)] \, \frac{\mathsf{E}[N(v; z)]}{\mathsf{E}[N(u; z)]}$$

which gives by (11)

$$\mathsf{P}(M > v) \leq \mathsf{E}[N^S(u; z)] \, \frac{\int \mathsf{P}(\ln z(t) > \ln v) \, \mathrm{d}t}{\int \mathsf{P}(\ln z(t) > \ln u) \, \mathrm{d}t}, \tag{22}$$

which if σ^2 is small is, by (12), approximately

$$\begin{aligned} \mathsf{P}(M > v) \quad &\leq \quad \mathsf{E}[N^S(u; z)] \, \frac{\mathrm{e}^{-(\ln(v) - m_0 - A)^2/2\sigma^2} / \sqrt{\ln(v) - m_0 - A}}{\mathrm{e}^{-(\ln(u) - m_0 - A)^2/2\sigma^2} / \sqrt{\ln(u) - m_0 - A}} & (23) \\ &\approx \quad \mathsf{E}[N^S(u; z)] \, \frac{\mathsf{P}(\ln z(t_0) > \ln v)}{\mathsf{P}(\ln z(t_0) > \ln u)} & (24) \end{aligned}$$

where t_0 is the time at which $m(t)$ attains its maximum. (The approximation (24) follows from (23) by the fact that Mills' ratio $(1 - \Phi(u))/\phi(u) \sim 1/u$ as $u \to \infty$, where Φ and ϕ denote the standard Normal distribution and density functions.)

We compare the POT and modified Rice approaches by comparing (21) with (22) or its approximation (24). We note:

(i) It is tacitly assumed in the modified Rice method that the occurrence rate λ_u of storms associated with exceedance of level u remains constant throughout the year. In this case the estimate λ_u^* in the modified Rice approach is the same as the estimate of ν_u in the POT approach. (If, however, storm occurrence is seasonal, then λ_u^* estimates the annual average of $\nu_u(t)$.)

(ii) The two approaches differ in their extrapolation from u to v, POT using a conditional distribution (which is still generalized Pareto), and modified Rice using a Gaussian distribution. Use of the generalized Pareto rests on a semi-parametric justification, the fact that this form is guaranteed at high enough levels whenever the tail of the distribution of S decays to zero in a regular manner (de Haan 1970). Use of the Normal distribution on the other hand is based on purely empirical evidence such as Figure 3. If the Normal assumption holds, then the Generalized Pareto distribution would be expected to give a reasonable approximation to it in the upper tail; on the other hand if the Normal assumption does not hold, then the Generalized Pareto would provide a flexible model for tail behaviour anyway. However, data at high levels would be needed to fit it (though see below).

(iii) The ability of the modified Rice method to yield estimates even when few observations of storm peaks are available (as is likely to be the case for satellite data) rests on the implicit assumption that the Gaussian distribution of $\ln H_s$ is universal at all levels, central and extreme, so that parameters estimated from non-extreme data are taken to be relevant to extremes too. In this respect the modified Rice method contrasts strongly with the POT method, which – as described above – attempts to use only extreme data for inference about extremes. An extension of the POT method, however, inspired by the crossings approach, would assume a Generalized Pareto distribution for $\ln H_s$ at high levels with seasonally varying location and local scale parameters that are the same as those at everyday levels (but without the everyday distribution itself necessarily being Generalized Pareto). Comparison of such an approach with the modified Rice method would be of interest. A point process framework would be natural for it.

(iv) The generalized Pareto distribution can represent both heavy and light tails. Even when a comparatively light tail is appropriate at very high levels, it is possible that at more moderate levels a heavy-tailed generalized Pareto distribution can give a good approximation (this is related to the penultimate phenomenon observed by Fisher & Tippett 1928). Thus extrapolation methods based on the Normal distribution may be anti-conservative in comparison to those based on the generalized Pareto. Though such anti-conservatism would act counter to the conservatism at the heart of the Rice method, it may suggest caution in the degree of extrapolation attempted.

(v) The modified Rice and POT approaches differ technically in the way in which seasonality is taken into account; the POT method averages $\nu_u(t)\mathsf{P}(S > v|S > u)$ over the seasons, but the modified Rice method (22) averages numerator and denominator separately in the extrapolation ratio, and does not, as noted in (i), explicitly allow for seasonality in λ_u.

4.3 Missing data and alternative approaches

Deeper investigation of the treatment of signals with large fractions of missing values is beyond the scope of this article. As pointed out by Rydén (2008), seasons with unusually low values may actually tend to increase the point estimates of return values.

Let us finally remark that for extreme winds, non-threshold methods have been proposed which take into account several values per year and occurrences of clustered values (Cook, 1982). A method for utilizing satellite data in the estimation of extreme H_s based on POT assumptions and asymptotic properties of threshold exceedances is outlined in Anderson *et al* (2001).

5 Conclusions

This paper reports initial exploration of a method by which return level estimates of H_s may be obtained from satellite data, compensating for data deficiency by stronger model assumptions. In the light of the discussion on Section 4 the following further investigations seems desirable:

(i) More extensive empirical checks on the robustness of the proposal of §3.3 for use of regional estimates $\bar{\lambda}_c$ when local estimates λ_u^* are not available.

(ii) As discussed in §4.1, methods to attach standard errors to the modified Rice estimates.

(iii) Further development of diagnostic methods to check whether Gaussian model assumptions are justified and investigation of the effect of departure from the assumptions.

(iv) Exploration of the degree of conservatism in estimates in relation to return period that results from use of the Rice inequality.

(v) Comparison of the modified Rice method and the extended POT method suggested in §4.2(iii).

(vi) Exploration of methods to combine estimates based on satellite data with those based on standard analyses of buoy, platform or numerical model data. More generally, exploration of estimation methods that utilize data from multiple sources. (We note that data from numerical models may not be completely consistent with directly-observed data from the same location because of model limitations, particularly at extreme levels. Estimation methods will need to allow for such systematic effects.)

References

Anderson, C.W., Carter, D.J.T., and Cotton, P.D. (2001). Wave climate variability and impact on offshore design extremes. Shell International Report. Available from `http://clive-anderson.staff.shef.ac.uk/waves.pdf`

Azais, J-M, and Wschebor, M. (2009). *Level sets and extrema of random processes and fields.* Wiley.

Baxevani, A., Rychlik, I. and Wilson, R. (2005). A new method for modelling the space variability of significant wave height. *Extremes* **8**, 267-294.

Baxevani, A., Caires, S. and Rychlik, I. (2009). Spatio-temporal statistical modelling of significant wave height. *Environmetrics*, in press.

Bleistein, N. and R.A. Handelsman (1986). Asymptotic expansions of integrals. *Dover Publications.*

Caires, S. and Sterl, A. (2005). 100-year return value estimates for ocean wind speed and significant wave height from the ERA-40 data. *Journal of Climate* **18**, 1032-1048.

Challenor, P. and Cotton, P. D. (1999). Trends in TOPEX significant wave height measurement. Available as a pdf document at `http://www.soc.soton.ac.uk/JRD/TOPtren/TOPtren.pdf`

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values.* Springer-Verlag.

Cook, N.J. (1982). Towards better estimates of extreme winds. *Journal of Wind Engineering and Industrial Aerodynamics* **9**, 295-323.

Cramér, H. and Leadbetter, M.R. (1967). *Stationary and Related Stochastic Processes: Sample Function Properties and Their Applications.* Wiley. (Republication by Dover 2004.)

de Haan, L. (1970). *On Regular Variation and its Application to the Weak Convergence of Sample Extremes.* Math. Centre Tract 32, Amsterdam.

Fisher, R. A. & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc. Camb. Philos. Soc.*, **24**, 180–190.

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

Marcus, M. B. (1977). Level crossings of a stochastic process with absolutely continuous sample paths. *Annals of Probability* **5**, 52-71.

Pickands, J. (1969). Asymptotic properties of the maximum in a stationary Gaussian process. *Trans. Amer. Math. Soc.*, **145** 75-86.

Rychlik, I. (1996). Extremes, rainflow cycles and damage functionals in continuous random processes, *Stochastic Process. Appl.*, **63**, 97-116.

Rychlik, I. (2000). On some reliability applications of Rice's formula for the intensity of level crossings. *Extremes* **3**, 331-348.

Rychlik, I. and Rydén, J. (2006). *Probability and Risk Analysis. An Introduction for Engineers.* Springer-Verlag.

Rydén, J. (2008). Estimation of return values: Consequences of missing observations. *International Journal of Mathematical Education in Science and Technology* **39**, 357-363.

Appendix

We here outline the derivation of approximation for $\mathsf{E}[N(u;\, z)]$ in Eq. (12).

Consider again Eq. (11), rewritten as

$$\mathsf{E}[N(u;\, z)] = c_1 \int_0^1 \mathrm{e}^{-(\ln(v)-m(t))^2/2\sigma^2}\,\mathrm{d}t.$$

where $c_1 = (1/2\sqrt{2\pi})(\gamma/\sigma)$. If $m(t) = m_0 + A\cos(2\pi t + \phi)$ and σ^2 is small, then, for high v an approximation can be found using Taylor's formula, called also Laplace's method (Bleistein and Handelsman 1986).

A change of variables to eliminate ϕ alter limits of integration and using Taylor's formula on $m(t)$ $(m(0) = m_0 + A,\ m'(0) = 0,\ m''(0) = -A(2\pi)^2)$ along with the series expansion of the exponential function, we find

$$\begin{aligned}
\mathsf{E}[N(u;\,z)] &= c_1\int_{-1/2}^{1/2}\exp\left\{-\frac{1}{2\sigma^2}\left[\ln(v)-m(t)\right]^2\right\}\,\mathrm{d}t\\
&\approx c_1\int_{-1/2}^{1/2}\exp\left\{-\frac{1}{2\sigma^2}\left[\ln(v)-m_0-\frac{m''(0)}{2}t^2\right]^2\right\}\,\mathrm{d}t\\
&= c_1\int_{-1/2}^{1/2}\exp\left\{-\frac{1}{2\sigma^2}\left[(\ln(v)-m_0-A)^2\right.\right.\\
&\qquad\left.\left.-(\ln(v)-m_0-A)m''(0)t^2+\frac{t^4}{4}[m''(0)]^2\right]\right\}\,\mathrm{d}t\\
&= c_1\,c_2\int_{-1/2}^{1/2}\exp\left\{-\frac{1}{2\sigma^2}(A+m_0-\ln(v))m''(0)t^2\right\}\\
&\qquad\times\exp\left\{-\frac{1}{2\sigma^2}[m''(0)]^2\frac{t^4}{4}\right\}\,\mathrm{d}t\\
&= c_1\,c_2\int_{-1/2}^{1/2}\exp\left\{-\frac{1}{2\sigma^2}A(2\pi)^2(\ln(v)-m_0-A)t^2\right\}\\
&\qquad\times\left(1-\frac{2\pi^4A^2}{\sigma^2}t^4+o(t^4)\right)\,\mathrm{d}t
\end{aligned}$$

where

$$c_2=\exp\left\{-\frac{1}{2\sigma^2}(\ln(v)-m_0-A)^2\right\}.$$

Now, consider high levels v and small σ^2. We have that $v>u$ where $u=m_0+A+c\,\sigma$. Let $\tilde\sigma^2=\sigma^2/(4A\pi^2(\ln(v)-m_0-A))$, which for typical numerical values in this application $\sigma^2=0.2$, $A=0.3$ and $(\ln(v)-m_0-A)=4\sigma$ we have $\tilde\sigma=0.1$. Hence the limits in the integration correspond to five $\tilde\sigma$. Furthermore, it can be shown that the integral $\int t^4\exp(-0.5t^2/\tilde\sigma^2)\mathrm{d}t\approx 7.5\tilde\sigma^5$ and hence is negligible.

Hence, by a change of variables and identification of the standard normal distribution we arrive at the approximation

$$\begin{aligned}
\mathsf{E}[N(u;\,z)] &\approx c_1c_2\sqrt{2\pi}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}A(2\pi)^2(\ln(v)-m_0-A)t^2\right\}\mathrm{d}t\\
&= c_1c_2\sqrt{2\pi}\frac{\sigma}{2\pi}\frac{1}{\sqrt{A(\ln(v)-m_0-A)}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\tau^2/2}\,\mathrm{d}\tau\\
&= \frac{1}{4\pi}\frac{\mathsf{E}[V]}{\sqrt{A(\ln(v)-m_0-A)}}\exp\left\{-\frac{1}{2\sigma^2}(\ln(v)-m_0-A)^2\right\}.
\end{aligned}$$