Coastal Engineering 56 (2009) 1013–1021

Contents lists available at ScienceDirect

Coastal Engineering

journal homepage: www.elsevier.com/locate/coastaleng

Automated threshold selection methods for extreme wave analysis

Paul Thompson [a,*], Yuzhi Cai [b], Dominic Reeve [a], Julian Stander [b]

[a] C-CoDE, School of Engineering, University of Plymouth, Devon, PL4 8AA, UK
[b] School of Mathematics and Statistics, University of Plymouth, Devon, PL4 8AA, UK

ARTICLE INFO

Article history:
Received 18 September 2008
Received in revised form 6 May 2009
Accepted 8 June 2009
Available online 6 August 2009

Keywords:
Bootstrap
Covariate dependent thresholds
Distribution with Generalized Pareto tail
Generalized Pareto Distribution
GPD
JOINSEA
Return level confidence intervals

ABSTRACT

The study of the extreme values of a variable such as wave height is very important in flood risk assessment and coastal design. Often values above a sufficiently large threshold can be modelled using the Generalized Pareto Distribution, the parameters of which are estimated using maximum likelihood. There are several popular empirical techniques for choosing a suitable threshold, but these require the subjective interpretation of plots by the user.

In this paper we present a pragmatic automated, simple and computationally inexpensive threshold selection method based on the distribution of the difference of parameter estimates when the threshold is changed, and apply it to a published rainfall and a new wave height data set. We assess the effect of the uncertainty associated with our threshold selection technique on return level estimation by using the bootstrap procedure. We illustrate the effectiveness of our methodology by a simulation study and compare it with the approach used in the JOINSEA software. In addition, we present an extension that allows the threshold selected to depend on the value of a covariate such as the cosine of wave direction.

1. Introduction

The successful design of a reliable and effective coastal defence structure can be associated primarily with knowledge of future extreme conditions which the defence must withstand. Typically, coastal defences are designed to provide sufficient protection against flooding or erosion to a desired return level associated with a particular return period, e.g. 100 years. The estimation of return levels and their uncertainty therefore has considerable engineering importance, especially in the area of coastal defence design. Statistical methodology for such estimation tasks is required as its input data about the extreme values of the conditions of interest.

There are two main methods for defining extremes. The first is based on dividing the time period over which the data are collected into blocks, with the most extreme value in each block being used for future analysis (e.g. daily or monthly maxima). The second method is based on exceedances over a specified threshold. In this paper we concentrate on the excesses over a threshold and provide an automated and computationally inexpensive threshold specification technique. Before presenting our technique, it is necessary to discuss how excesses over a suitable threshold can be modelled and analysed statistically.

Let y be a value taken by the variable of interest, for example wave height, and let u be a threshold. Provided u is sufficiently large, values of y greater than u can be modelled using the generalized Pareto Distribution (GPD); see Coles (2001), for example. The cumulative distribution function H of the GPD takes the form:

$$H(y) = 1 - \left[1 + \frac{\xi(y-u)}{\sigma_u}\right]^{-1/\xi}, \tag{1}$$

where $y>u$ and $1+\xi(y-u)/\sigma_u>0$. The parameters σ_u and ξ control the scale and shape of the distribution. Here we use the notation σ_u to emphasize that the scale parameter changes with the threshold u, although we will drop the subscript u when this emphasis is no longer needed; the shape parameter ξ does not change with u. The parameters σ_u and ξ need to be estimated from available data, and this can be done using maximum likelihood estimation, as discussed in detail in Coles (2001) and Smith (1985). Usually, the selection of an appropriate threshold u is performed on a visual basis and so can have a range of associated errors. These visual procedures require prior knowledge of the accurate interpretation of threshold choice plots, such as the Mean Residual Life plot, to achieve a satisfactory model fit; again see Coles (2001) for examples. We illustrate the difficulties associated with the interpretation of the Mean Residual Life plot in Section 2.

Threshold selection has received some additional attention in the literature, for example, Dupuis (1999) presents a guide to threshold selection based on robustness considerations, while Tancredi et al. (2006) adopt a Bayesian approach and discuss how to take account of threshold uncertainty. The methods presented in these papers are complicated to implement and can be computationally demanding;

* Corresponding author.
E-mail address: p1thompson@plymouth.ac.uk (P. Thompson).

0378-3839/$ – see front matter
doi:10.1016/j.coastaleng.2009.06.003

see Section 2.3 for further discussion of Tancredi et al. (2006) and Guillou and Hall (2001) for related methodology. The automated threshold selection method that we will present requires little external input other than the variable of interest, and is considerably simpler and easier to implement than the approaches proposed in these papers.

We have also extended our threshold selection method to allow threshold choice to depend on a covariate such as the cosine of wave direction, where our specific aim is to account for the directional effect when modelling wave height or wave period using GPDs. The practical advantage of our extended procedure is that it automatically identifies the wave directions associated with the highest waves and consequently can provide better estimation of wave height return levels.

The rest of this paper is organized as follows. In Section 2 we present our automated threshold selection technique and compare it with one of the currently available subjective approaches. We also describe a bootstrap procedure for assessing the effect of uncertainty on return level estimation. In Section 3 we describe a simulation study aimed at quantifying the effectiveness of our method. In Section 4 we compare our approach with the existing methodology used in the JOINSEA software (see Wallingford, 1998a,b). In Section 5 we extend our method to allow threshold choice to depend on a covariate. Finally, in Section 6 we present some concluding comments.

2. Automated threshold selection technique

2.1. Theoretical basis

When fitting the GPD to data, the scale and shape parameters σ_u and ξ can be estimated using maximum likelihood estimation. To achieve a good model fit, we need to choose a suitable value of the threshold u. Commonly used techniques involve visual assessment of threshold choice plots and rely upon prior experience of their interpretation; see Tawn and Coles (1994) and Davidson and Smith (1990). Such plots are found in Coles (2001) for GPDs fitted to rainfall data. We shall discuss one of these plots, the Mean Residual Life plot, in Section 2.2.2 below. Another of these techniques plots parameter estimates of GPDs fitted using a range of thresholds against the threshold, and is the basis for our automated threshold selection methodology. We now outline our automated method for threshold selection.

Let $u_1,\ldots, u_n$ be n equally spaced increasing candidate thresholds. Let $\hat{\sigma}_{u_j}$ and $\hat{\xi}_{u_j}$ be maximum likelihood estimators of the scale and shape parameter based on data above the threshold u_j, $j=1,\ldots, n$. Finally, let u be a suitable threshold, that is one for which values of $y>u$ can be modelled using the GPD. It follows from Coles (2001), page 83 that, provided $u\le u_{j-1}<u_j$,

$$\sigma_{u_{j-1}} = \sigma_u + \xi\left(u_{j-1} - u\right) \text{ and } \sigma_{u_j} = \sigma_u + \xi\left(u_j - u\right). \tag{2}$$

Hence,

$$\sigma_{u_j} - \sigma_{u_{j-1}} = \xi\left(u_j - u_{j-1}\right). \tag{3}$$

Furthermore, the standard maximum likelihood theory, as discussed in Coles (2001), tells us that $E[\hat{\sigma}_{u_j}]\approx\sigma_{u_j}$ and $E[\hat{\xi}_{u_j}]\approx\xi$, for any j such that $u_j>u$. Let

$$\tau_{u_j} = \hat{\sigma}_{u_j} - \hat{\xi}_{u_j} u_j, \quad j = 1, \ldots, n, \tag{4}$$

and consider the differences

$$\tau_{u_j} - \tau_{u_{j-1}}, \quad j = 2, \ldots, n; \tag{5}$$

it follows from the above results about the expected values of maximum likelihood estimators and from Eq. (3) that $E[\tau_{u_j}-\tau_{u_{j-1}}]\approx 0$. Moreover, we can appeal to the same theory to conclude that $\tau_{u_j}-\tau_{u_{j-1}}$ approximately follows a normal distribution. The variability of this difference does not itself measure the variability associated with our threshold selection procedure. This distributional result suggests the following procedure for finding a suitable threshold u:

(1) Identify suitable values of equally spaced candidate thresholds $u_1<u_2<\cdots<u_n$. We found that setting $n=100$ gives good results. We take u_1 to be the median and u_n to be the 98% quantile of the data, unless fewer than 100 values exceed this value, in which case u_n is set to the 100th data value in descending order. Our procedure performs well in such circumstances. Less reliable results were obtained from smaller data sets.

(2) If u is a suitable threshold, then all differences $\tau_{u_j}-\tau_{u_{j-1}}$ have an approximate normal distribution with mean 0 provided $u\le u_{j-1}<u_j$. If u is unsuitable, then these differences may not follow a normal distribution. This suggests that a suitably applied test for normality is an effective method to determine u.
The Pearson's Chi-square Test is used as a test of goodness of fit to establish whether or not the observed differences are consistent with a normal distribution with mean 0; see Greenwood and Nikulin (1996). Initially, we consider $u=u_1$ and perform the Pearson normality test based on all the differences $\tau_{u_2}-\tau_{u_1}, \tau_{u_3}-\tau_{u_2},\cdots, \tau_{u_n}-\tau_{u_{n-1}}$. If the null hypothesis of normality is not rejected, u is taken to be a suitable threshold. If the null hypothesis is rejected, then we consider $u=u_2$, remove $\tau_{u_2}-\tau_{u_1}$ from the set of differences considered, and repeat the above procedure. We have found from a simulation study that a size 0.2 Pearson normality test generally performs most consistently over a range of normality tests and sizes. Reducing the size of the test has the effect of lowering the chosen threshold.

(3) Step 2 is repeated until the Pearson's normality test indicates that the differences are consistent with a normal distribution with mean 0. If this does not happen, u_n is returned with a warning. Our experience is that this latter situation occurs rarely.

The above steps can be performed quickly, so yielding a procedure that is computationally inexpensive. We implemented our method in the freely available, open source statistical environment

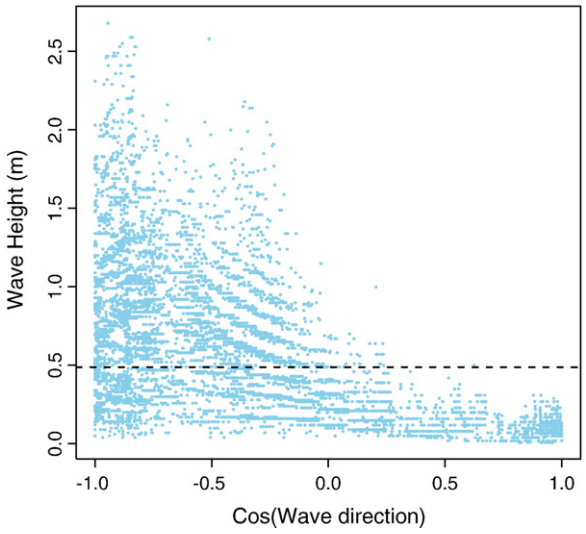

Fig. 1. Scatter plot of wave height against the cosine of wave direction for 10,000 values from the Selsey Bill data set. The horizontal line was produced by applying our automated threshold selection procedure to the wave height observation, taking no account of the cosine of wave direction.

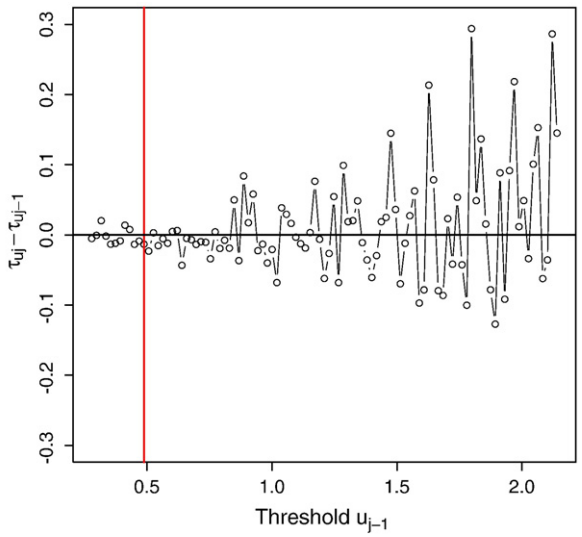

Fig. 2. Graph of the differences $\tau_{u_j} - \tau_{u_{j-1}}$ against threshold u_{j-1} for the wave height data. The vertical line indicates the automated threshold selection choice.

R Development Core Team (2009), which is becoming more widely used in engineering and related areas.

2.2. Practical examples

2.2.1. Coastal wave data

We now apply the method presented in Section 2.1 to a real data set. The data used in this example relate to conditions near the Selsey Bill area (Hawkes, personal communication). They were generated using the hindcast technique (see Reeve et al., 2004, for example) based on wind records. The data set consists of hourly hindcast measurements of the variables significant wave height, wave period and wave direction over a time span of 27 years. Wave hindcasting attempts to create the wind-wave conditions, and cannot account for the swell component. In this example we take a random sample of 10,000 observations from the data set. The resulting values are typical of data that are collected in similar studies and satisfy the independence assumption that underlie the maximum likelihood theory. A plot of wave height against the cosine of wave direction is shown in Fig. 1.

Our automated threshold selection technique was applied to these wave height observations and indicated 0.487 m as a suitable threshold. This threshold is also shown in Fig. 1. The values of the cosine of wave direction were not used in finding this threshold. Fig. 2 plots differences $\tau_{u_j} - \tau_{u_{j-1}}$ against threshold u_{j-1}, and as described in Section 2.1 is the basis of our threshold selection procedure. Fig. 3 shows diagnostic plots, as discussed by Coles (2001) and produced by the freely available ismev package of Coles and Stephenson (2006) run in R Development Core Team (2009). Such plots are now used routinely, and so have not been edited here; detailed explanation is provided in the caption. These diagnostic plots indicate that the fitted GPD model is satisfactory. Both the probability and quantile plots show that there is little difference between empirical and fitted values from the model, indicating a good fit. Similarly, there is a reasonable agreement between the data and the estimated return levels and associated 95% confidence envelope, and between the histogram of the data values above the chosen threshold and the fitted generalized Pareto density. This example shows that our proposed methodology can provide an automated, simple and computationally inexpensive

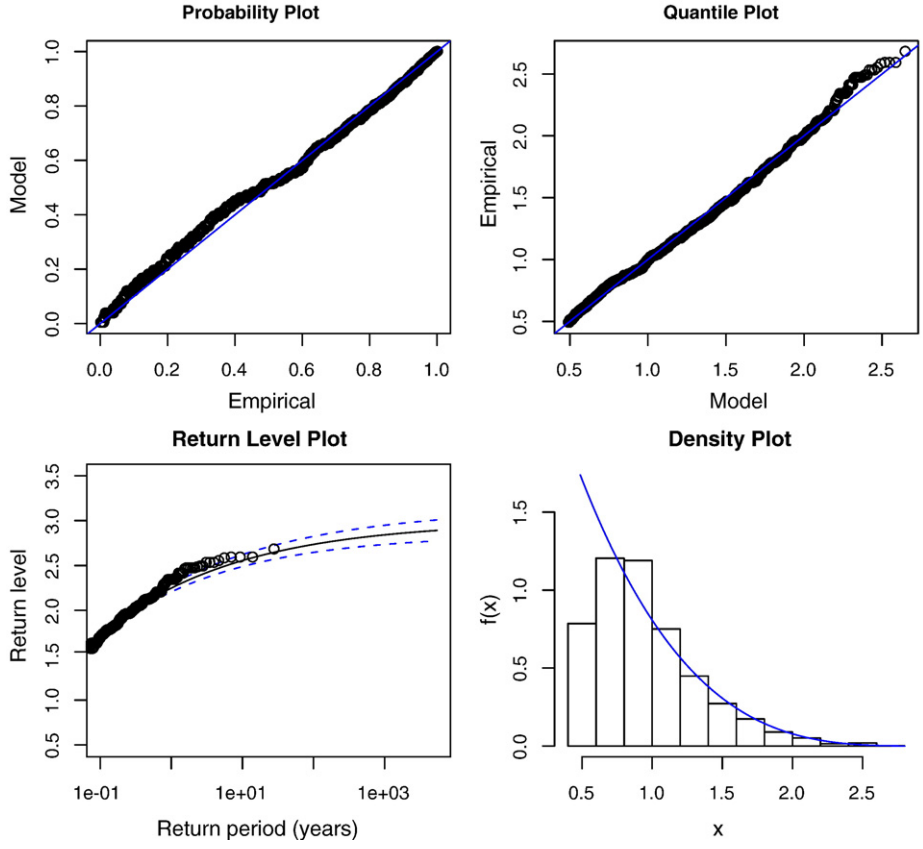

Fig. 3. Diagnostic plots for the GPD fit when the threshold is chosen using our automated threshold selection approach applied to the wave height data. This plot was generated using the ismev package (Coles and Stephenson, 2006). In the third plot Return level refers to wave height (m). In the fourth plot x refers to the wave height (m), and $f(x)$ to its probability density. See text for discussion of the individual plots.

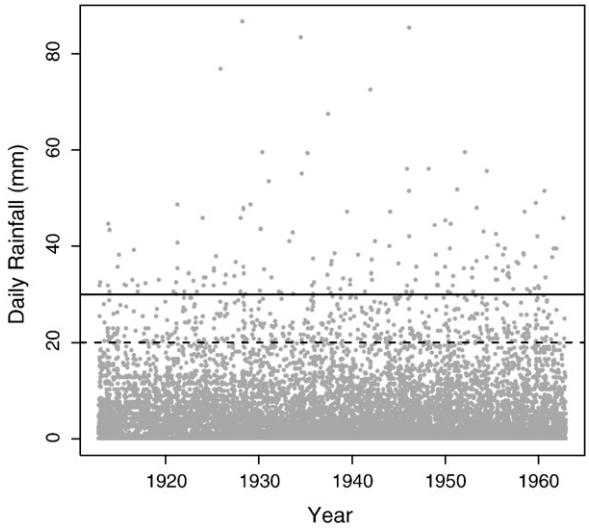

Fig. 4. Scatter plot of daily rainfall data against time. The dashed line shows our automated threshold choice, while the unbroken line is the threshold value recommended by Coles (2001).

threshold selection method that avoids the need for subjective interpretation of threshold choice plots with all their possible errors.

2.2.2. Daily rainfall data

We now compare the automated threshold selection method presented in Section 2.1 with a currently available subjective method by applying them to a data set considered by Coles (2001). The data comprise daily rainfall accumulations at a location in south west England recorded over the period 1914–1962. Coles (2001) presents this example to illustrate the currently available threshold selection techniques. Fig. 4 shows a plot of the data together with the threshold of 30 mm as recommended by Coles (2001) and our own automated choice of 20 mm. Fig. 5 shows the Mean Residual Life plot (see Coles, 2001 for details) upon which Coles' bases his choice. A threshold is usually identified as a value beyond which the plot is linear (up to sampling error). The behaviour of the plot is linear (up to sampling error) beyond 60 mm, but few data points lie above this value. Linearity also occurs between 30 and 60 mm, and so Coles (2001) recommends a value of 30 mm. A similar argument could also be used

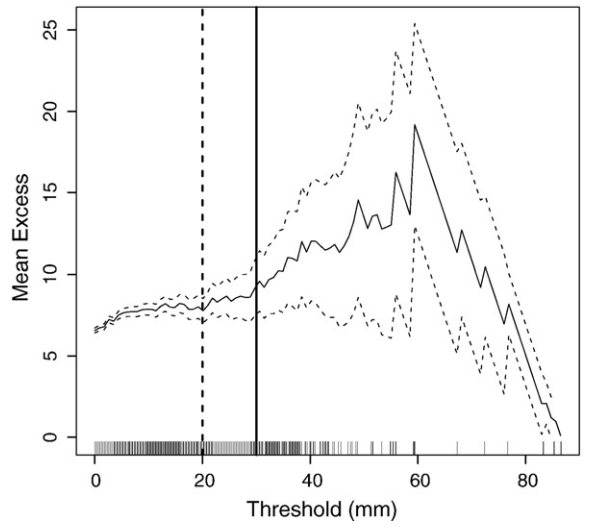

Fig. 5. Mean Residual Life plot for the daily rainfall data. The dashed line was produced by our automated threshold selection procedure. We have also added the threshold value recommended by Coles (2001) as the unbroken line. The individual values are indicated by a rug of dashes.

to justify our automated threshold choice of 20 mm. The subjective nature of and difficulties associated with the interpretation of the Mean Residual Life plot are well illustrated by this example. Fig. 6 shows comparisons of inferences (fitted densities, return levels and confidence intervals) from the fitted models based on each threshold. We can see that the fitted models are relatively similar indicating that our automated threshold selection technique compares well to the subjective procedure. Coles' (2001) threshold does yield fewer exceedances, which is the cause of the increased return level confidence interval widths in Fig. 6(B).

2.3. Using bootstrap percentile intervals to assess return level uncertainty

Uncertainty associated with inferences from the GPD model can depend on two sources: firstly, the uncertainty associated with estimating the scale and shape parameters from the available exceedances; secondly, the uncertainty associated with the selection of the threshold that defines these exceedances. Uncertainty in parameter estimation can be relatively small in comparison to the

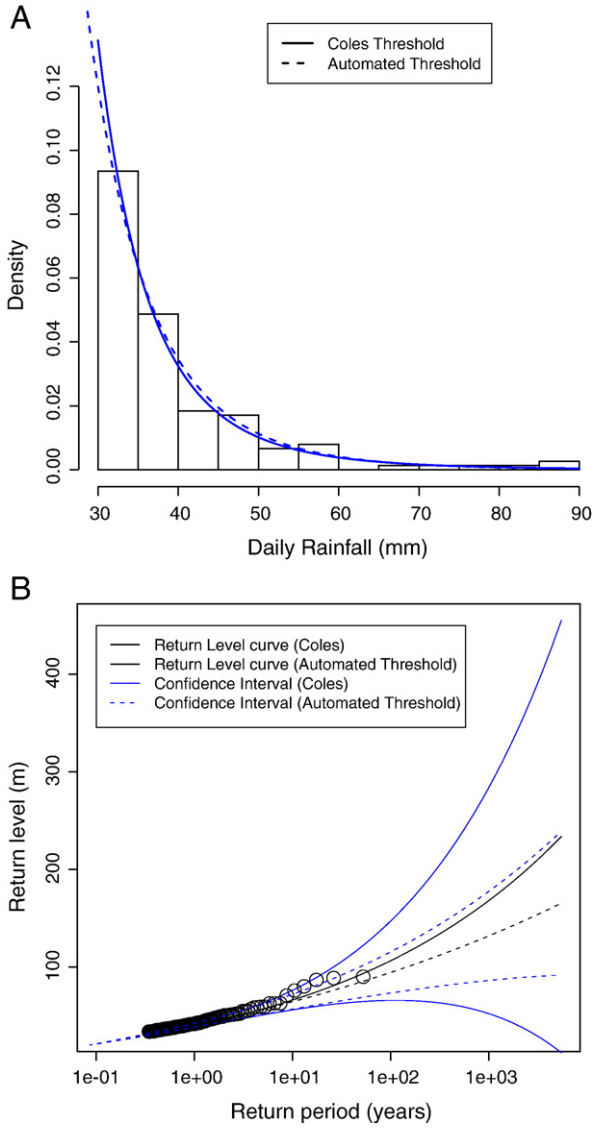

Fig. 6. (A): Histogram of the exceedances from threshold choice of 30 mm recommended by Coles (2001), together with the GPD fit (solid line). The GPD fit based on our threshold of 20 mm is also shown (dotted line). This GPD fit has been scaled so that the area under it above 30 mm is one. (B): Return level curves and confidence envelopes based on Coles' (2001) threshold (unbroken) and our threshold (dashed).

uncertainty in the choice of threshold. It is therefore important when discussing our technique to include the effect of the uncertainty associated with threshold choice in the inferential procedure.

As we saw in Section 1, return levels play a vital role in coastal engineering; see page 82 of Coles (2001) for a detailed discussion about the estimation of return levels and approximate confidence intervals from GPD fits. Standard software programmes, such as the ismev package estimate return levels and approximate confidence intervals, as shown in Fig. 3, but do not take into account uncertainty due to threshold selection.

In an important and innovative paper Tancredi et al. (2006) present a review of existing model based methodology to account for threshold uncertainty in GPD models, and then introduce their own technique. In contrast to conventional fixed threshold methods, Tancredi et al. (2006) work in the Bayesian framework and assume that the threshold is one of the parameters about which to make inference. To overcome the lack of a natural model below the threshold and to avoid over-restrictive parametric assumptions, they propose a flexible mixture of an unknown number of uniform distributions with an unknown range for below-threshold data. They consider it reasonable to expect different estimates of return levels and precision of estimates for different thresholds. This essentially leads to a Bayesian mixing of all reasonable threshold values and parameter estimates to determine an overall estimate of return levels and their uncertainty. Their approach is, however, highly computationally intensive, requiring the use of a reversible jump Markov chain Monte Carlo algorithm to cope with the unknown number of uniform distributions used for below-threshold modelling; see Green (1995). It also requires a number of prior assumptions to be made, although Tancredi et al. (2006) argue that return level estimation is more robust to these assumptions than to threshold choice in a fixed approach. Because of these drawbacks, we take a different approach to assess return level uncertainty based on the bootstrap procedure. Mooney and Duval (1993) and Efron and Tibshirani (1993) provide a basic summary of this procedure as follows:

(1) Set $b=1$.
(2) Draw a simple random sample of size m from the original data set $y_1,\ldots,y_m$ with replacement. We call this a bootstrap sample.
(3) For the bootstrap sample, calculate the quantity of interest, for example a specific return level, and call it $\hat{\theta}_b^{\star}$. We calculate the return level by first estimating the threshold using the methodology in Section 2.1. We then make use of this threshold when estimating the GPD model. Finally, we use the GPD parameter estimates to calculate the return level estimate.
(4) Increase b by 1 and repeat steps (2) and (3) a total of B times, where B is a large number. We present results for $B=1000$. Other values of B, ranging from 250 to 3000, yielded similar results.
(5) Construct a probability distribution by attaching a $1/B$ probability to each point, $\hat{\theta}_1^{\star}, \hat{\theta}_2^{\star},\ldots, \hat{\theta}_B^{\star}$.

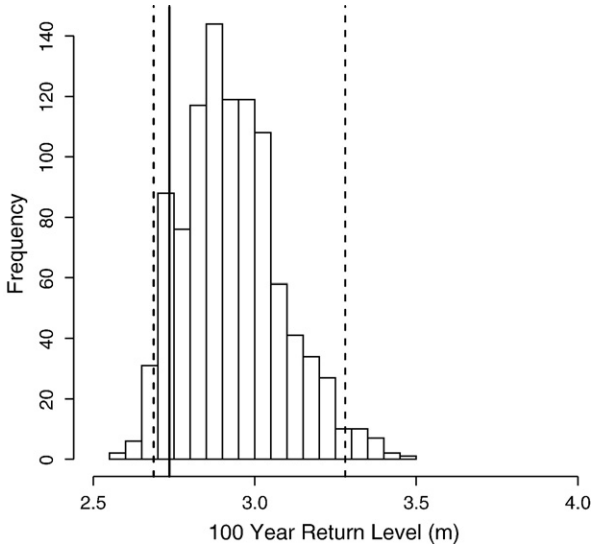

Fig. 7. Histogram of the bootstrapped 100 year return levels and associated 95% bootstrap percentile interval ($B=1000$ bootstrap iterations). The dashed lines are the percentile interval and the solid line is the return level based on the original data.

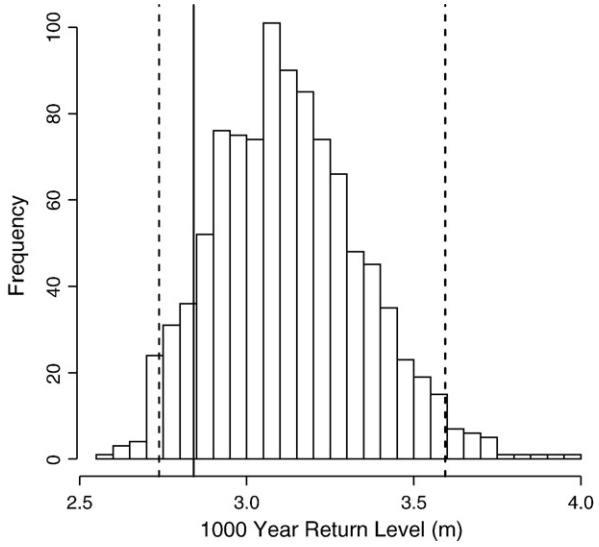

Fig. 8. Histogram of the bootstrapped 1000 year return levels and associated 95% bootstrap percentile intervals. The dashed lines are the percentile interval and the solid line is the return level based on the original data.

Uncertainty in the quantity of interest – for example a specific return level – can be quantified by summarizing this probability distribution using a confidence interval. More precisely, we will use a bootstrap percentile interval. To obtain an $(1-\alpha)$-level interval we sort the B values $\hat{\theta}_1^{\star}, \hat{\theta}_2^{\star},\ldots, \hat{\theta}_B^{\star}$ in ascending order and select the $\left(\frac{\alpha}{2}B\right)^{\text{th}}$ and $\left(1-\frac{\alpha}{2}\right)B^{\text{th}}$ values as our confidence interval using the integer below and the integer above if these values are not themselves integers. We set $\alpha=0.05$, yielding 95% confidence intervals. We now present the result of applying the above bootstrap methodology to our data set. Fig. 7 shows a histogram of the bootstrapped 100 year return levels $\hat{\theta}_1^{\star},\ldots, \hat{\theta}_B^{\star}$ and the associated bootstrap percentile interval. Fig. 8 is an analogous plot for the 1000 year return level. These percentile intervals enable us to quantify the uncertainty in return level estimation in an accurate way, without ignoring threshold choice uncertainty and relying on the standard asymptotic theory outlined on page 82 of Coles (2001). Figs. 7 and 8 show that the bootstrap percentile interval widths are approximately 0.6 m for the 100 year return level and 0.8 m for the 1000 year return level, indicating that uncertainty about these estimates is not particularly large from an engineering point of view.

We remark that there are more refined methods for obtaining bootstrap confidence intervals. Venables and Ripley (2002) discuss 'normal', 'basic', BC_a and studentized confidence intervals, in addition to percentile confidence intervals in their Section 5.7; see Davison and Hinkley (1997) and Efron and Tibshirani (1993) for excellent and extensive further discussion. We chose to use percentile confidence intervals because they are simple to understand and implement.

3. Simulation study

In this section we investigate the performance of our automated threshold selection method by means of a simulation study. Fig. 9

shows a histogram of a data set comprising 10,000 simulated values of a random variable X with distribution function given by:

$$F(x) = \{(1-\beta)G_1(x) + \beta\}I[x > u] + G_2(x)I[x \le u], \quad x > 0, \tag{6}$$

where I is the usual indicator function and $\beta = P(X \le u)$. $G_1(x)$ is a GPD function with associated density function:

$$g_1(x) = \frac{1}{\sigma}\left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-(1/\xi+1)}, \quad x > u, \quad 1 + \frac{\xi(x-u)}{\sigma} > 0; \tag{7}$$

$G_2(x)$ is a truncated normal distribution function with associated density function

$$g_2(x) = \frac{\frac{1}{\alpha\sqrt{2\pi}}\exp\left(-\frac{(x-\gamma)^2}{2\alpha^2}\right)}{\int_0^\infty \frac{1}{\alpha\sqrt{2\pi}}\exp\left(-\frac{(x-\gamma)^2}{2\alpha^2}\right)dx}, \quad x > 0. \tag{8}$$

With this F, the distribution of the random variable X can be thought of as a mixture of a normal distribution truncated on $(0, u]$ and a GPD on (u, ∞) with weights β and $1-\beta$, with non-extreme values coming from the truncated normal and extreme values from the GPD. Given β and the parameters γ and α of g_2, we can find u from the condition:

$$\beta = \Pr(X \le u) = G_2(u) = \int_0^u g_2(x)dx \tag{9}$$

$$= \frac{\int_0^u \frac{1}{\alpha\sqrt{2\pi}}\exp\left(-\frac{(y-\gamma)^2}{2\alpha^2}\right)dy}{\int_0^\infty \frac{1}{\alpha\sqrt{2\pi}}\exp\left(-\frac{(y-\gamma)^2}{2\alpha^2}\right)dy}.$$

For the simulated data set shown in Fig. 9 we set $\beta = 0.9$, $\gamma = 2$ and $\alpha = 0.7$, and solved for u to obtain $u = 2.90$. We choose the parameter σ of the GPD so that there was no discontinuity at u in the probability density function of X. To do this we require

$$g_2(u) = (1-\beta)g_1(u) = \frac{1-\beta}{\sigma}. \tag{10}$$

With $u = 2.90$, this equation can easily be solved to yield $\sigma = 0.40$. We set the shape parameter ξ of the GPD to be 0.2. The resulting probability density function of X is shown in Fig. 9, together with the threshold $u = 2.90$ (dotted line).

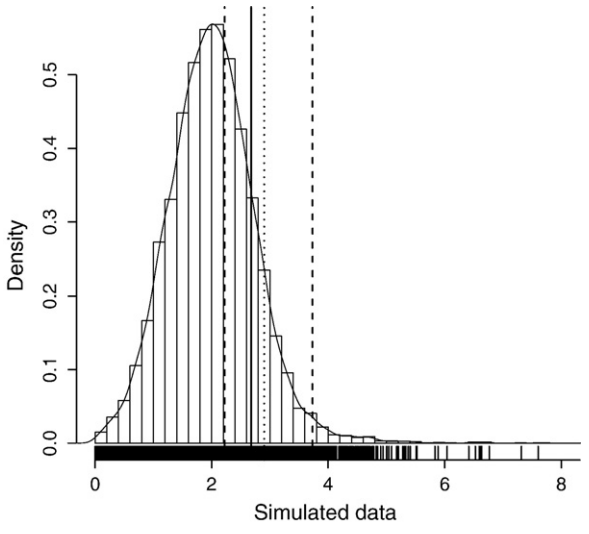

Fig. 9. Histogram of a data set of 10,000 simulated values of a random variable X with distribution function F. The associated probability density function is also shown. The individual values are indicated by a rug of dashes. Our automated threshold choice is indicated by a solid line, with the true threshold $u = 2.90$ being shown by a dotted line. The 95% bootstrap percentile intervals are also presented using dashed lines.

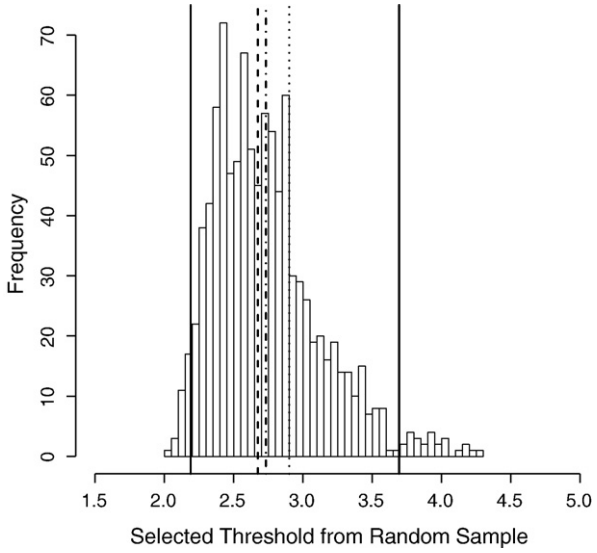

Fig. 10. Histogram of thresholds selected from 1000 random samples of size $N = 10,000$ from F. The mean and median of the automated threshold choices for the simulated data sets are shown by dot-dash and dashed lines respectively; while the true threshold $u = 2.90$ is shown by a dotted line. The 2.5% and 97.5% quantiles are shown as the outer solid lines.

A random sample $x_1, \ldots, x_N$ can be simulated from F as follows:

- Set $i = 1$. Simulate $y \sim N(\gamma = 2, \alpha^2 = 0.7^2)$;
- If $y < 0$ reject it;
- else if $0 < y < u$, set $x_i = y$ and increase i by 1;
- else if $y > u$ simulate $x \sim \text{GPD}(u = 2.90, \sigma = 0.4, \xi = 0.2)$, set $x_i = x$ and increase i by 1.
- Stop when $i = N + 1$.

We applied our automated threshold selection method to the simulated data set of size $N = 10,000$ shown in Fig. 9. The selected threshold took the value 2.678 m and can be seen to be close to the true value of $u = 2.90$. We next used the above simulation procedure to generate 1000 random samples of size $N = 10,000$ from F. We applied our threshold selection technique to each random sample; a histogram of these 1000 thresholds, together with 2.5% and 97.5% quantiles (2.189, 3.694), the true threshold $u = 2.90$ and mean $u_{\text{mean}} = 2.73$ and median $u_{\text{med}} = 2.67$ values of the distribution of estimated thresholds are shown in Fig. 10. The selected thresholds seem to be evenly and not very widely spread around the true threshold, suggesting that our method can recover a known threshold to a good degree of accuracy. Our method performed similarly well when applied to data sets simulated using different values of β, γ, α and ξ.

We now focus on the simulated data set shown in Fig. 9 and apply the bootstrap analysis discussed in Section 2.3, except that our bootstrap quantity of interest $\hat{\theta}_b^\star$ now becomes a selected threshold instead of a specific return level. Fig. 11 shows a histogram of the bootstrap threshold choices together with the 95% bootstrap percentile interval (2.225, 3.732), our automated threshold choice for the original simulated data set and the true threshold $u = 2.90$. The 2.5% and 97.5% quantiles found above have also been added. The 95% bootstrap interval is also shown in Figs. 9 and 11. We can see from these plots that the 95% bootstrap percentile interval is not very wide and contains the true and selected thresholds. The actual interval values of (2.225, 3.732) compare well with the 2.5% and 97.5% quantiles (2.189, 3.694) indicating that the bootstrap assesses well the uncertainty associated with our threshold choice procedure.

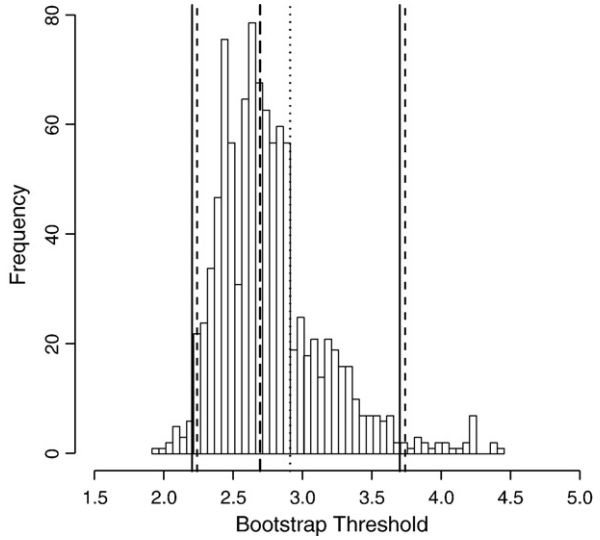

Fig. 11. Histogram of the bootstrap threshold choices. The automated threshold choice for the original simulated data set is shown as the large-dash line, while the true threshold $u = 2.90$ is the dotted line. The 95% bootstrap percentile interval is shown as the dashed lines, with the 2.5% and 97.5% quantiles from Fig. 10 being given using the outer solid lines.

In order to validate our bootstrap procedure further we performed an extensive study based on data sets simulated from distribution (6) to check the coverage of our bootstrap confidence intervals. Good results were obtained. We found that for the 1000 year return level, for example, the true coverage was 94%, very close to its 95% nominal level. The conclusion of all our simulation work is that our automated and computationally inexpensive procedure can recover a theoretical threshold from simulated data to a good degree of accuracy and that the bootstrap can be successfully used to assess associated uncertainties.

In the next section we give a further example of the application of our procedure by comparing it to an existing technique utilized in the JOINSEA software.

4. Comparison to the JOINSEA software

In this section we compare our new method with an existing technique used in the JOINSEA software (see Wallingford, 1998a,b). The JOINSEA approach for choosing an appropriate threshold assumes that extremes can be identified as exceedances over a 95% quantile. We now use the Selsey Bill data set introduced in Section 2.2 to compare our choice of threshold and fitted GPD with those obtained from the approach adopted in JOINSEA. Table 1 gives the results from the two approaches.

Fig. 12 shows again a scatter plot of wave height against the cosine of wave direction for the Selsey Bill data set, together with the two thresholds. The dashed line was obtained using our new threshold technique, while the solid line is the JOINSEA threshold. We see from Table 1 and Fig. 12 that the threshold values are very different, with the automated threshold being almost 1 m below the JOINSEA threshold. Fig. 13(A) and (B) shows comparisons of inferences (return levels, confidence intervals and fitted densities) from the fitted models based on each threshold. We can see that the resulting models are actually very similar indicating that our automated threshold selection technique is comparable to that of JOINSEA. The JOINSEA threshold yields fewer exceedances, which is the cause of the increased return level confidence interval widths in Fig. 13(A). The narrower confidence intervals yielded by our threshold selection technique, together with the fact that it is more model based, lead us to prefer our methodology over the JOINSEA approach. We also note that for data sets such as those simulated in Section 3 with $\beta > 0.95$ the JOINSEA approach is guaranteed to lead to non-extremes being included in future GPD analyses.

We applied our automated threshold selection technique to different data sets which varied in size and data collection location, and found it performed consistently well in terms of model goodness of fit. We felt that in the case of the Selsey Bill data our automated approach chose a relatively low threshold as a type of "average" threshold across the range of direction covariate values. This observation led us to extend our automated technique to allow the chosen threshold to vary with covariate value. We discuss our direction varying threshold methodology in detail in Section 5.

Table 1
The chosen threshold, number of exceedances, GPD parameter estimates and standard errors for our new automated threshold selection method and the approach adopted in the JOINSEA software.

	New technique	JOINSEA
Threshold value	0.487	1.480
Number of exceedances	5372	497
Maximum likelihood estimate, ξ	−0.230	−0.271
Maximum likelihood estimate, σ	0.576	0.405
Standard error, ξ	0.00952	0.04094
Standard error, σ	0.00940	0.02409

5. Extended automated threshold selection technique

We have seen that the Selsey Bill data set comprises information about wave direction as well as wave height. So far we have worked only with wave height. It is clear from Fig. 12 that the behaviour of wave height varies with wave direction. It therefore makes sense to include the directional effect in our automated threshold selection procedure, rather than to have a threshold that is constant over wave direction.

In extreme wave analysis directional effects are usually dealt with using one of two methods: either the data are split according to different directions with each separate data set being modelled independently, or the wave direction is included as a covariate as in Ewans and Jonathan (2006) and Jonathan and Ewans (2007), for example. In this section we propose a new approach to blocking the data.

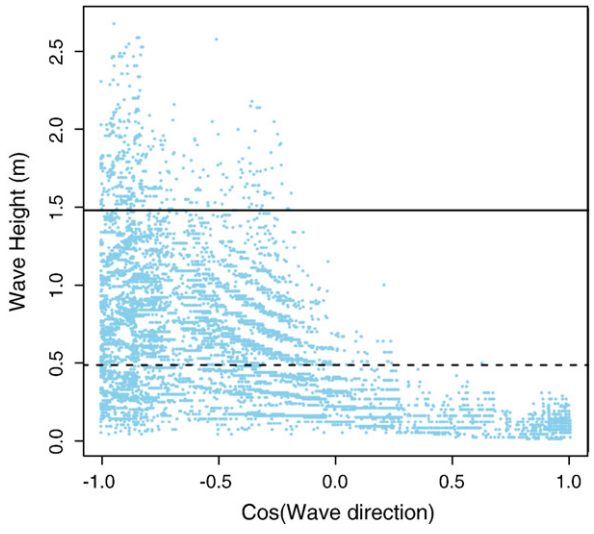

Fig. 12. Scatter plot of wave height against the cosine of wave direction for 10,000 values from the Selsey Bill data set. Our automated threshold choice is shown using the dashed line, while the solid line shows the threshold chosen by the JOINSEA software. Both threshold choices take no account of the cosine of wave direction.

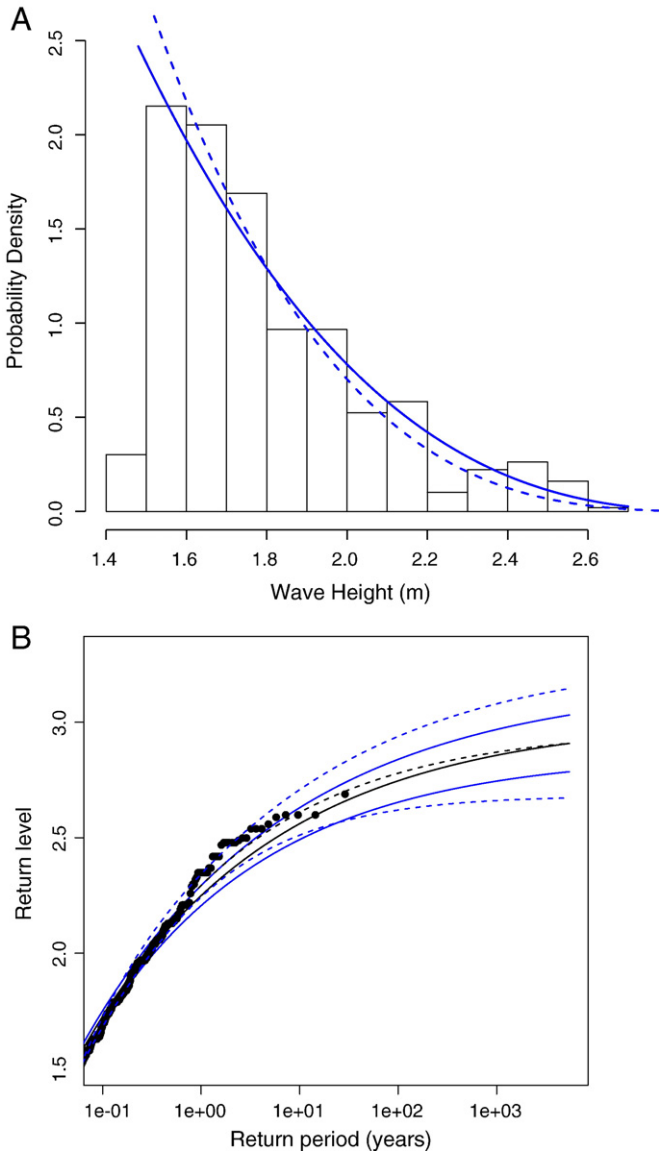

Fig. 13. (A): Histogram of the exceedances from the JOINSEA threshold choice, together with the GPD fit (dashed line). The GPD fit based on our threshold procedure is also shown (unbroken line). This GPD fit has been scaled so that the area under it above the JOINSEA threshold is one. (B): Return level curves and confidence envelopes from both automated (unbroken) and JOINSEA (dashed) threshold model fits.

Our approach is based on the automated threshold selection procedure that we have already presented and is as follows:

(1) First the data set is blocked according to the cosine of wave direction. The number of blocks is initially defined by the user; see Fig. 14(A) for example where the covariate axis is split into 40 equal width blocks. Each block is then altered iteratively to its optimum size as described in Eq. (2).

(2) The constant automated threshold selection procedure is applied to the data in each block. The block size can then be altered in order to achieve a satisfactory GPD fit in each block. If there is no sufficient number of observations within the block or if the block's optimal threshold choice does not define enough exceedances to achieve a good GPD fit, then the block is merged with the next consecutive block and the process is repeated. Through a simulation study we found that a sufficient number of observations would be the larger of the 5% of the total number of observations and 500, and a sufficient number of exceedances would be the maximum of 1% of the total number of observations and 50. The simulation study involved fitting a number of GPD models to different data sets and assessing the dependence of model fit quality on the number of observations and the number of exceedances. The merging of consecutive blocks is continued until the required minimum values for the number of observations and the number of exceedances for the merged block allows are reached. Our optimal blocks are shown in Fig. 14(B).

(3) Each block now has a constant optimal threshold associated with it. A separate GPD can be fitted to the wave height data within each block, and associated direction specific inferences about return levels can be made.

If the individual block thresholds shown in Fig. 14(B) are considered together, a piecewise constant threshold function is defined. A threshold that is continuous in the cosine of wave direction covariate can be obtained by applying a smoothing spline, for example. We did this using R Development Core Team's (2009) smooth.spline function; see Green and Silverman (1994), for example. The resulting smoothed direction varying threshold function is shown in Fig. 15.

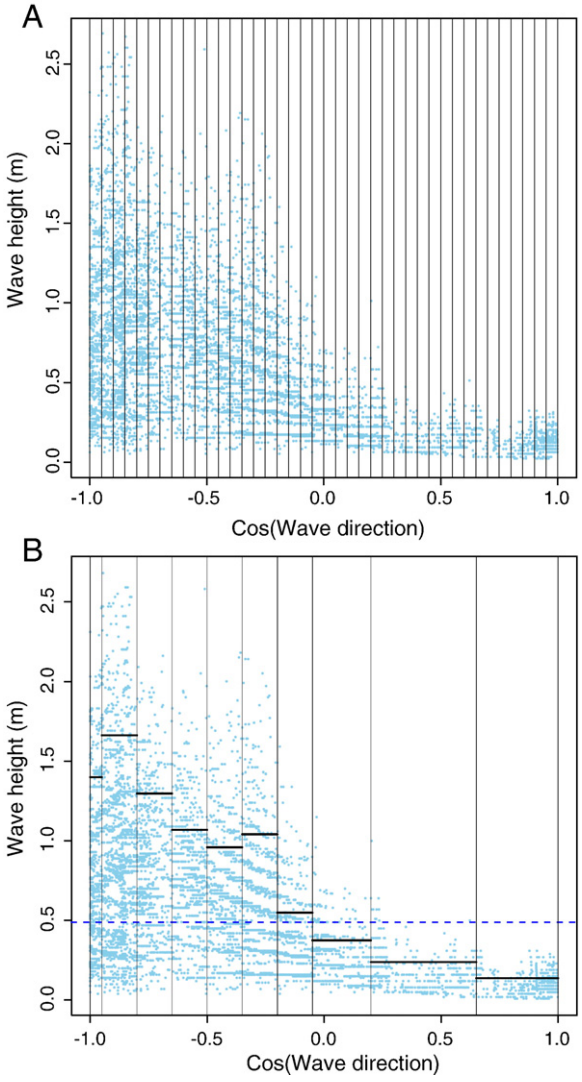

Fig. 14. (A): Scatter plot of wave height against the cosine of wave direction. The data have been split into 40 sections equally spaced along the covariate axis. (B): Scatter plot of wave height against the cosine of wave direction. The data have now been split into optimal blocks along the covariate axis. Individual automated thresholds have been chosen for each block and are shown by the solid horizontal lines. The dotted line shows the threshold chosen without reference to cosine of wave direction.

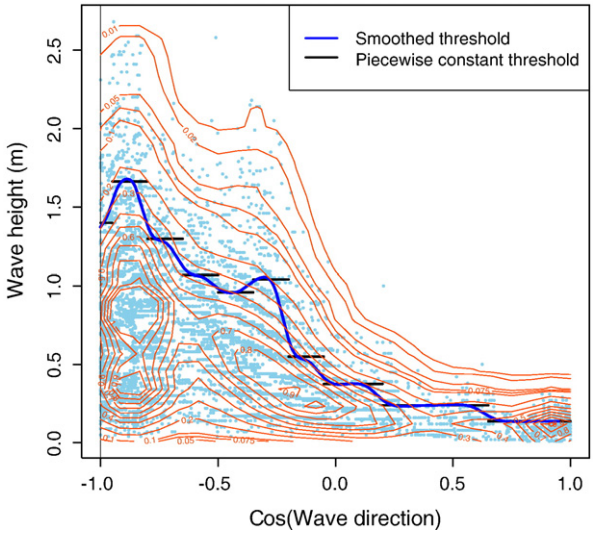

Fig. 15. The bivariate wave data with piecewise constant and smoothed covariate varying thresholds. Bivariate probability density estimate contours are overlaid on the scatter plot.

In order to justify further the choice of these direction varying thresholds we also show in Fig. 15 probability density contours for a bivariate kernel density estimate (calculated using the kde2d function of the MASS library; see Venables and Ripley, 2002) based on wave height and the cosine of wave direction. We see that the chosen thresholds align well with the tail of this probability density function across the range of cosine wave direction, supporting our direction varying threshold choice procedure. We conclude by remarking that, as mentioned, the more appropriate thresholds that this extended automated threshold selection technique provides can yield more accurate direction specific return level estimates. These in turn can lead to improved coastal defence designs that account for directional variations in extreme wave heights.

6. Concluding comments

In this paper, we have presented a new, automated, simple and computationally inexpensive method for selecting the threshold for the GPD in extreme value modelling. Our pragmatic method uses a series of normality tests to find an appropriate threshold choice for a given data set. We have contrasted our methodology with one of the currently available subjective approaches. We have shown the practical applicability of our method using an example from coastal engineering. We have demonstrated that our automated technique can recover a known threshold from a simulated data set to a good degree of accuracy. We have assessed the effect of the uncertainty associated with threshold selection on return level estimation using the bootstrap procedure. We have also provided comparisons of our new approach with the existing JOINSEA technique, pointing out improvements of our method over the existing one. In practice, our method can be seen as an additional tool that complements existing threshold selection methods.

We have extended our methodology to incorporate a direction covariate dependant threshold. This extension uses our automated threshold selection technique to segregate the data into optimal blocks based on goodness of fit and sample size requirements.

Our methodology can lead to more accurate return level estimates, with their uncertainty properly qualified, which can inform and enhance the coastal design process.

Acknowledgements

The authors acknowledge the support of a doctoral scholarship from the University of Plymouth and funding from the EPSRC projects RF-PeBLE (grant No. EP/C005392/1), LEACOAST 2 (grant No: EP/C013085/1) and BVANG (grant No: EP/C002172/1). We would also like to thank Dr. Peter Hawkes from HR Wallingford for the data set and supplementary information provided. We warmly thank the referees and the Editor-in-Chief for their thorough reading of the manuscript and suggestions that have led to considerable improvements in this paper.

References

Coles, S., 2001. An Introduction to Statistical Modelling of Extreme Values. Springer, London.
Coles, S. and Stephenson, A. (Original S functions) (R port and R documentation files), 2006. ismev: An Introduction to Statistical Modeling of Extreme Values. http://www.maths.lancs.ac.uk/stephena/, R package version 1.2.
Davidson, A.C., Smith, R.L., 1990. Models for exceedances over high thresholds. J. R. Stat. Soc., Ser. C, Appl. Stat. 52 (3), 393–442.
Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and their Application. Cambridge University Press, Cambridge.
Dupuis, D.J., 1999. Exceedances over high thresholds: a guide to threshold selection. Extremes 1 (3), 251–261.
Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, London.
Ewans, K., Jonathan, P., 2006. Estimating extreme wave design criteria incorporating directionality. 9th International Workshop on Wave Hindcasting and Forecasting, Victoria, Canada.
Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.
Green, P.J., Silverman, B.W., 1994. Nonparametric Regression and Generalized Linear Models. Chapman and Hall, London.
Greenwood, P.E., Nikulin, M.S., 1996. A Guide to Chi-squared Testing. Wiley (Probability & Statistics Series), New York.
Guillou, A., Hall, P., 2001. A diagnostic for selecting the threshold in extreme value analysis. J. R. Stat. Soc., B 63 (2), 293–305.
Jonathan, P., Ewans, K., 2007. The effect of directionality on extreme wave design criteria. Ocean Eng. 34, 1977–1994.
Mooney, C.Z., Duval, R.D., 1993. Bootstrapping: A Nonparametric Approach to Statistical Inference. Sage University Paper, London.
R Development Core Team, 2009. R: A language and Environment for Statistical Computing. 3-900051-07-0. http://www.R-project.org, Vienna, Austria.
Reeve, D., Chadwick, A., Fleming, C., 2004. Coastal Engineering: Processes, Theory and Design Practice. SPON, London.
Smith, R., 1985. Maximum likelihood estimation in a class of nonregular cases. Biometrika 72 (1), 67–90.
Tancredi, A., Anderson, C., O'Hagan, A., 2006. Accounting for threshold uncertainty in extreme value estimation. Extremes 9, 86–106.
Tawn, J.A., Coles, S.G., 1994. Statistical methods for multivariate extremes: an application to structural design. J. R. Stat. Soc., Ser. C, Appl. Stat. 43 (1), 1–48.
Wallingford, H.R., 1998a. The Joint Probability of Waves and Water Levels: JOINSEA. Report: TR71, HR Wallingford.
Wallingford, H.R., 1998b. The Joint Probability of Waves and Water Levels: JOINSEA. Report: SR 537, HR Wallingford.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, 4th Ed. Springer-Verlag, New York.