Open Access

The Effect of Declustering in the r-Largest Maxima Model for the Estimation of H_S-Design Values

Takvor H. Soukissian* and Panagiota M. Arapi

Hellenic Centre for Marine Research, Institute of Oceanography, PO BOX 712, 19013 Anavyssos, Attica Greece

University of the Aegean, School of the Environment, Department of Marine Sciences, Marine Sciences Building, University Hill, 811 00 Mytilene, Greece

Abstract: A major problem often encountered in design sea-state prediction is the limited amount of available extreme-type wave data. The Annual Maxima model is consistent with the conditions of the mathematical background of Extreme Value Theory, yet its application raises statistical uncertainties in cases where the initial data population is limited. Due to this, alternative models of similar theoretical background have been developed to describe extreme values, including the "r-largest maxima method". A main problem in applying this model refers to the appropriate selection of a sample comprising the r independent maxima within each year of the available time series: since in nature environmental extremes tend to appear in clusters, the native time series under examination should be appropriately "de-clustered" to satisfy the independency assumption. Some established declustering procedures refer to: a) the selection of a "Standard Storm Length", b) the combination of a run length k and a relatively high threshold value u (Runs declustering), c) the estimation of wave energy reductions between consecutive sea-state systems (DeClustering Algorithm) and d) the selection of the three largest monthly maxima of each year of the initial significant wave height time series (triple annual maximum series). The aim of this paper is to assess the effect of the aforementioned declustering procedures on the numerical results obtained by the r-largest model.

Keywords: r-largest maxima, design value, extreme waves, significant wave height, declustering.

1. INTRODUCTION

The term "extreme events" has not been compactly defined in literature and its use depends on the purpose and the subject of each particular study. The scarcity of appearance, the magnitude and the socio-economic impacts of an extreme event can be some of the factors which should be considered when defining these events. One persuasive definition, which embodies many of these factors and is often used in statistics, is that extremes are the events corresponding to the tail of the probability distribution of the process under study. In modern literature, great concerns have been expressed for such events in many diverse areas of application and therefore the statistical analysis of their behaviour has been of primary interest.

The branch of probability theory which, through a firm theoretical foundation, provides the stochastic description of such events is known as "Extreme Value Theory" (EVT). The most important result of EVT states that, under appropriate conditions, the normalized maximum of a stochastic sequence of independent and identically distributed (iid) random variables converges in distribution to a random variable which follows the Generalized Extreme Value (GEV) distribution family, [1]:

$$H_{GEV}\left(x;\mu,\sigma,\xi\right)=\exp\left[-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right],\quad 1+\xi\left(\frac{x-\mu}{\sigma}\right)\geq 0 \tag{1.1}$$

where $-\infty<\mu<\infty$ is the location parameter, $\sigma>0$ is the scale parameter and $-\infty<\xi<\infty$ is the shape parameter. The above parameters are to be estimated from the data available and this is usually done using the method of maximum likelihood.

In most applications though, the available data is usually in the form of time series and do not necessarily satisfy the conditions of the classical extreme value theory. The most important hindrance, especially in environmental applications, is that serial dependence is almost always an inherent characteristic of the time series. It is highly related to the time interval between consecutive events: the smaller the time interval, the higher the dependence while events separated by larger time lags could be assumed to be independent [2]. In this connection, although EVT cannot be directly applied on the routinely collected physical data, under some statistical pre-processing it is possible to use its results on appropriately selected independent sub-series of the original data set, consisting of extreme-type data. For example, a most commonly used approach in this respect is to segment the initial data set into large blocks of the same size and extract the maximum value of each block. Attempting to be compatible with the theoretical assumptions, these blocks are

*Address correspondence to this author at the Hellenic Centre for Marine Research, Institute of Oceanography, PO BOX 712, 19013 Anavyssos, Attica Greece; Tel: +30-2291076420; Fax: +30-2291076323;
E-mail: tsouki@ath.hcmr.gr

1874-835X/11 2011 Bentham Open

usually chosen to correspond to a time period of one year (or season) and the corresponding extreme value model is known as the "Block (Annual) Maxima Model" (hereafter AMM).

Having attained an iid sample comprising the, say, annual or seasonal maxima, the EVT implies that these maxima follow the GEV distribution; hence estimates of extreme quantiles of the distribution of the block (annual) maximum can be obtained by inverting equation (1.1), [1]:

$$H_{GEV}^{-1}\left(1-p;\mu,\sigma,\xi\right)=x_P=\begin{cases}\mu-\dfrac{\sigma}{\xi}\left\{1-\left[-\ln\left(1-p\right)\right]^{-\xi}\right\}, & \xi\neq 0,\\ \mu-\sigma\ln\left[-\ln\left(1-p\right)\right], & \xi=0,\end{cases} \quad (1.2)$$

where $H\left(x_P\right)=1-p$.

In equation (1.2), the term x_P essentially denotes the level which, to a reasonable degree of accuracy, is expected to be exceeded in average once every $1/p$ years and thus, it is commonly referred to as the *return level* (or the *design value*) associated with the *return period* $1/p$.

This approach, however, despite its unchallengeable benefits, raises in practice significant statistical uncertainties, since the sample size of the required extreme-type data (i.e. the block maxima series) is usually very small. Additionally, a common objection to this approach is that by taking into account, for example, only the annual maximum, some of the excluded (for being the second, third, etc annual largest) values can be greater than the maximum of another year. To this end, alternative methods have been developed that allow the extraction of larger sets of extreme-type data, the most important of which is the Excesses-over-Threshold family of models and the r-largest maxima model; see [1], [3-5]. The theoretical origin of these methods is based on the framework of EVT and thus, though they utilize larger and enriched extreme-type data sets, the independency assumption is still a prerequisite. As a result, prior to extreme value analysis, the time series under examination should be appropriately "declustered" to satisfy the independency assumption, since in nature extremes tend to appear in clusters (i.e., exhibiting strong dependency in neighbouring extreme values).

Under this concept, a main problem in applying the r-largest maxima model, concerns the selection of the sample comprising the r independent maxima within each year of the available time series; thus a declustering procedure is required. In environmental applications, a frequently used declustering method refers to the selection of a "Standard Storm Length" (abbreviated hereafter as SSL), which is essentially the minimum time distance required between successive events (maxima) in order for them to be considered as statistically independent. In the available literature, the selection of SSL is elaborated in a more or less arbitrary way, since there is a lack of physical explanation concerning the presumed statistical independency of the clusters' maxima. Another available declustering procedure is the Runs Declustering, which, though oriented for generating extreme-type data for the Peak-Over-Thresholds (POT) and the Generalized Pareto Distribution (GPD) models, it is a typical method which can be also applied for obtaining maxima of independent clusters from any given time series. This procedure comprises the combination of two declustering parameters, a threshold u and a run length k, assuming that "distant enough" excesses of a "high enough" threshold can be considered as independent. Apparently, special care should be taken in the determination of u and k, but, as in the SSL case, there is no formal procedure to do so and the selection is performed using common sense. In addition to the above, a new declustering algorithm is introduced tailored especially for discriminating independent clusters (sea state systems) in significant wave height (H_S) time-series [5]. The Declustering Algorithm (DeCA) involves the estimation of wave energy reductions (ER) from one sea-state system to the next sea-state system, pleading that if the wave energy content falls under a specified percentage, the corresponding sea state systems can be (safely) considered as independent clusters. Finally, in the present paper the method of the "triple annual maximum series" (hereafter referred as TAM) [6], is also treated as a very simple declustering procedure, since it involves the selection of the three largest monthly maxima of each available year of the initial H_S time series, assuming that they are statistically independent. Let us note here that declustering for stationary series is also an important problem, theoretically resolved by introducing the extremal index, denoted as θ. In a way, in this case there exists a firm theoretical basis for efficiently dealing with clusters; however, the statistical estimators for θ are yet not devoid of problems, in the sense that the selection of some of the involved parameters is largely arbitrary [7].

The present paper aims at the assessment of the effect of the aforementioned declustering procedures on the numerical results obtained by the r-largest model. The structure of the paper is as follows: after a short review of the r-largest method (Section 2), the declustering procedures are introduced and described in some detail in Section 3. The numerical results regarding the design values of significant wave height and the associated return periods, obtained by implementing the various declustering procedures and applying the r-largest model, are presented and discussed analytically in Section 4. Finally, in Section 5 some concluding remarks are provided.

2. THE r-LARGEST MAXIMA MODEL

The Annual Maxima model, though adequately aligned with the conditions of the mathematical background of EVT, raises significant statistical uncertainties in cases where the initial data population is limited, as only a small fraction of the available extreme-type data is used. In an attempt to overcome this difficulty, alternative statistical models of similar theoretical background have been developed, including the "r-largest maxima method". For a review on the alternative extreme value methods see for example [1]. The r-largest maxima model, firstly introduced by Weissman [8], is actually an alternative method for estimating the parameters of the GEV distribution taking into account not only the first, but also the second, third, etc largest values of each available year, expanding in this way the extreme-type data set used for the statistical inference. Applications of this method can be found in [9-11], while a theoretical description of the r-largest maxima model is given below, see also [1]:

Let $\{X_n\} = X_1, X_2, \ldots, X_n$ be a stochastic sequence of independent and identically distributed (iid) random variables following the same cumulative distribution function (cdf) and $M_n = \max\{X_n\}$ follow the GEV distribution for some sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ as $n \to \infty$. Let also $X_{k:n} \equiv M_n^{(k)} = k$-th largest of $\{X_1, X_2, \ldots, X_n\}$. Then, for fixed k

$$\Pr\left[\frac{M_n^{(k)} - b_n}{a_n} \le x\right] \to H_{(k)}(x) = \exp\{-\tau(x)\} \sum_{s=0}^{k-1} \frac{\{\tau(x)\}^s}{s!}, \quad (1.3)$$

where $\tau(x) = \left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}$ and $1+\xi(x-\mu)/\sigma > 0$.

As stated in equation (1.3), under appropriate normalization, as $n \to \infty$ the k-th order maximum of $\{X_n\}$ follows a distribution function which has the same parameters as the GEV distribution of the maximum M_n. The following result provides the joint density of the entire vector $\vec{M}_n^{(r)}$: under the same conditions as above, the limiting probability density function (pdf) of the last (appropriately normalized) r largest maxima (order statistics), $\left\{\frac{M_n^{(1)} - b_n}{a_n}, \frac{M_n^{(2)} - b_n}{a_n}, \ldots, \frac{M_n^{(r)} - b_n}{a_n}\right\}$, as $n \to \infty$, is given by the following relation:

$$f\left(x^{(1)}, x^{(2)}, \ldots, x^{(r)}\right) = \exp\left\{-\left[1+\xi\left(\frac{x^{(r)}-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \times \prod_{k=1}^{r} \frac{1}{\sigma}\left[1+\xi\left(\frac{x^{(k)}-\mu}{\sigma}\right)\right]^{-1/\xi-1}. \quad (1.4)$$

For $\xi \to 0$, eq. (1.4) reduces to:

$$f\left(x^{(1)}, x^{(2)}, \ldots, x^{(r)}\right) = \exp\left\{-\exp\left[-\left(\frac{x^{(r)}-\mu}{\sigma}\right)\right]\right\} \times \prod_{k=1}^{r} \frac{1}{\sigma} \exp\left[-\left(\frac{x^{(k)}-\mu}{\sigma}\right)\right]. \quad (1.5)$$

Apparently, for $r=1$, eq. (1.4) reduces to the GEV distribution and eq. (1.5) to the Gumbel cdf.

To estimate the model parameters in eq. (1.4) the maximum likelihood is used. In this case, the likelihood function is:

$$L(\mu,\sigma,\xi) = \prod_{i=1}^{m}\left(\exp\left\{-\left[1+\xi\left(\frac{x_i^{(r_i)}-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \prod_{k=1}^{r_i} \frac{1}{\sigma}\left[1+\xi\left(\frac{x_i^{(k)}-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1}\right), \quad (1.6)$$

for

$$\frac{1+\xi\left(x^{(k)}-\mu\right)}{\sigma} > 0, \quad k = 1,\ldots,r_i, \quad i = 1,\ldots,m,$$

and the maximization of the corresponding log-likelihood function is performed numerically. See for example [1] and references cited therein.

As noted above the parameters of eq. (1.4) correspond directly to those of the GEV distribution of the maximum and therefore the r-largest maxima model is essentially an improved way to estimate them. At this point we must point out that the above results concern the largest values of a sequence, the terms of which are iid random variables. Since, in practical applications, the physical data rarely if ever are statistically independent, they cannot be used directly as recorded; therefore a declustering procedure is required. In addition to that, special care must be taken while selecting r to avoid issues of high variance (produced due to lack of large sample size) or bias (produced due to the violation of the asymptotic support of the model).

3. DECLUSTERING PROCEDURES

In nature, the extremes of most physical processes tend to occur consecutively (in clusters), a fact that violates the main assumption in EVT, i.e. that the variables under study are iid. To this end, the first step in the analysis of environmental characteristics, such as spectral characteristics of wind waves (e.g. significant wave height, spectral peak period, etc.), is to identify the independent clusters within the examined time series, or, as it is frequently expressed, to 'de-cluster' the time series. As a result, a number of declustering techniques have been developed, i.e., procedures that lead to identifying peak events in the data record which can be reasonably considered independent. In this context, since the r-largest maxima model shares the assumption that the involved variables are iid, the use of such procedures is required in this case as well.

However, declustering procedures involve a great deal of subjectivity, as Ferro and Segers [7], strongly support through their very explicit statement: "All declustering schemes proposed in the literature require an auxiliary parameter, the choice of which is largely arbitrary". It is, therefore, evident that special care should be taken in the application of any declustering scheme. In the next sections, the main techniques used for the declustering of significant wave height time series are presented, while their effect on the selection of the r largest maxima is assessed and discussed in Section 5.

3.1. The Standard Storm Length Model

Tawn suggested a simple procedure for the extraction of the r largest independent values, adopting the rational assumption that distant enough peak values can be considered as independent, see [11]. In particular, the time series is considered as a sequence of separate independent storms, each having a standard storm length (SSL), say τ. Under this assumption, the extraction of the maximum value within each storm for each year of the examined time series is expected to provide a series of independent extreme values. The r-largest of such values for each year could be essentially the required r largest independent annual events. For a given standard storm length τ, the algorithmic scheme proposed is the following:

1. Pick out the largest annual value from within the year of interest.
2. Discard all values lying at a distance (lag) less or equal to $\tau / 2$ from either side of the time instant where the value was chosen from.
3. Select the largest value of the remaining data.
4. Repeat steps 2 and 3 until the time series data is exhausted.
5. Repeat steps 1-4 for all the years of the available time series.

It is obvious that the selection of τ is of great importance: low values of τ could lead to inclusion of dependent extreme events (i.e., events occurring within the same storm-cluster), while high values of τ would generate few extreme type data leading thus to high variance in the obtained estimates. An important drawback of this method is that, in many cases, the storm length is chosen in an essentially arbitrary way, as, for example, from “typical” standard storm length values proposed in the relevant literature. Such suggestions can be found, for example in [12] where three choices of SSL, namely 24, 48 and 72h were explored; in [13] where the value of 72h was adopted; in [14, 15] where SSL was chosen equal to 30h, or in [16] where SSL was selected to vary from 5 to 7 days. Aiming at a less subjective choice of SSL value, Tawn [11] suggested estimating τ by examining the autocorrelation structure of the variable of interest. In particular, he suggested that a rationally selected τ could be the lag after which the process becomes very weakly correlated, see also [17]. In [18] this approach was utilized on time series of significant wave height from the North Sea, concluding, after a close examination of the autocorrelation function, that 480h would be an appropriate value for τ.

At this point, however, it should be noted that, for most environmental parameters, as for the significant wave height, the assumption of almost equally sized storms-clusters is not realistic, while the correlation structure could vary significantly between different annual data-sets, rendering the consideration of a constant and unique de-correlation time (SSL) for the entire population highly debatable. For some applications of this declustering model see [18-21].

3.2. Runs Declustering

Another declustering technique frequently applied on environmental data is Runs Declustering. It is based on the assumption that all values successively exceeding a selected threshold may be considered as a separate independent cluster, as long as they do not occur within a given time interval, [22]. This procedure was originally developed to generate extreme-type subseries particularly for the POT and the GPD models, but it can also be used to extract, from any given time series, the extreme-type sample required for any of the (extreme value) models which share the iid assumption.

The inductive scheme proposed in this case is the following:

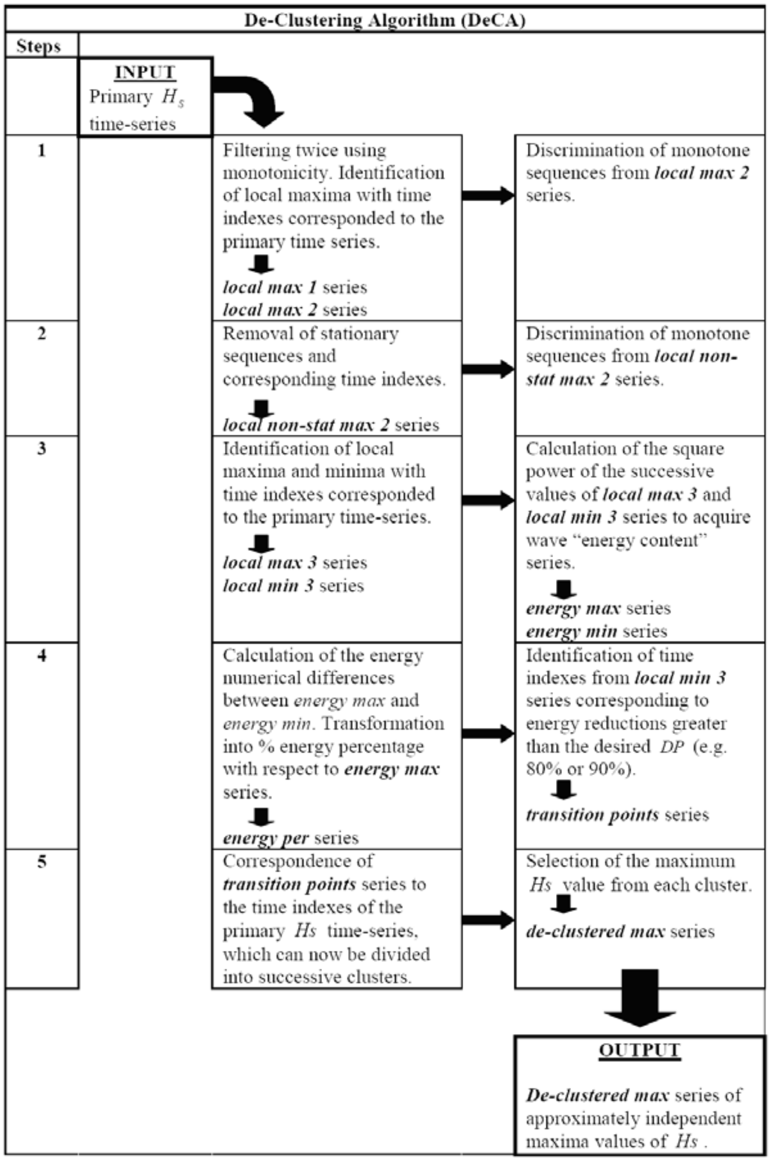
1. Specify a “high enough” threshold u such that the values of the process which exceed it to be considered as extreme-type events and specify an appropriate number of observations, say k (known as run length).
2. Specify clusters: The cluster begins when the threshold u is exceeded for the first time and ends once at least k consecutive observations fall below the threshold. Namely, two excesses of the threshold (are considered to) belong to the same storm (cluster) if they are separated by less than k consecutive non-exceeding values.
3. Extract the maximum value of the cluster and proceed to identify the next cluster (using steps 2 and 3).
4. Terminate the procedure when the time series data is exhausted.
5. Repeat steps 2-4 for all the available (annual) time series.

The advantage of this method, as described in [23], is that it allows both the duration (persistence) of storms and the duration of intervals between them to vary according to the data, reflecting the inherent natural variability of these quantities. However, as in the SSL case, there are important issues regarding the selection of both the threshold u and the run length k, as improper choices can lead to either bias or to high variance. In absence of a more formal procedure, the selection of u and k is based on common sense judgement; thus, as suggested in [1], the method should be applied for several choices of u and k and the sensitivity of the results should be carefully examined. In cases where this declustering technique is used before applying the GPD method, the appropriate threshold values are selected by assessing and rating the fit of the GPD. In particular, in such cases, the appropriate threshold values are indicated by the stability of parameters’ estimates, in addition to the linearity in the mean residual life plot, making use of the fact that if GPD is the correct model for a (high enough) threshold value, say u_0, the mean excesses of u, for $u > u_0$, form a linear function of u. Under this concept, for example, in [24, 25], u was selected as an appropriate quantile of the data; this decision was supported by using the mean residual life plot. The run length value, on the other hand, is adopted following suggestions from the relevant literature. The rationale, under these selections, demands the run length to be large enough to cover the time period required for the phenomenon to be fully developed. In previous offshore studies [14, 15, 24, 26-30], these values were found to vary from 24 to 60h.

3.3. The Declustering Algorithm (DeCA)

Soukissian and Kalantzi [5] developed a declustering algorithm in order to locate sequences of approximately independent maxima from native H_S time series. The rationale

Table 1. Flow-chart of the Declustering Algorithm (DeCA), see [3]



De-Clustering Algorithm (DeCA)			
Steps			
	INPUT Primary H_S time-series		
1		Filtering twice using monotonicity. Identification of local maxima with time indexes corresponded to the primary time series. ***local max 1*** series ***local max 2*** series	Discrimination of monotone sequences from ***local max 2*** series.
2		Removal of stationary sequences and corresponding time indexes. ***local non-stat max 2*** series	Discrimination of monotone sequences from ***local non-stat max 2*** series.
3		Identification of local maxima and minima with time indexes corresponded to the primary time-series. ***local max 3*** series ***local min 3*** series	Calculation of the square power of the successive values of ***local max 3*** and ***local min 3*** series to acquire wave "energy content" series. ***energy max*** series ***energy min*** series
4		Calculation of the energy numerical differences between *energy max* and *energy min*. Transformation into % energy percentage with respect to ***energy max*** series. ***energy per*** series	Identification of time indexes from ***local min 3*** series corresponding to energy reductions greater than the desired *DP* (e.g. 80% or 90%). ***transition points*** series
5		Correspondence of ***transition points*** series to the time indexes of the primary *Hs* time-series, which can now be divided into successive clusters.	Selection of the maximum *Hs* value from each cluster. ***de-clustered max*** series
			OUTPUT ***De-clustered max*** series of approximately independent maxima values of *Hs* .

under this procedure is that significantly large wave energy reductions (hereafter ER) between local H_S maxima and the subsequent H_S minima indicate the transition to a different, approximately independent, "sea-state system". The extraction of the maximum from each such system results to the collection of the required independent extreme-type data. Apparently, as in the RL case, the persistence (duration) of the sea state systems (clusters) is not taken a priori as constant: being essentially considered as a random variable, the corresponding (random) sample is directly obtained from the data after a simple de-noising (filtering) procedure. In contrast to the other declustering approaches, in this case, the physics of the phenomenon is taken into account, by discriminating the independent clusters based on *wave energy reductions* between successive sea states.

A detailed presentation of the procedure is given in Table **1**, while the main steps are as follows:

1. Filter (de-noise) the time series twice using monotonicity.
2. Identify local maxima and minima, as well as their energy content ($\propto H_S^2 T_p$ or simply $\propto H_S^2$).
3. Estimate the relative difference between the energy content of each local maxima and the one of its subsequent local minima, i.e. the energy reduction (ER).
4. Once this ER falls below a predefined percentage, the cluster is terminated and the following value starts a different independent cluster.
5. Extract the maximum of each cluster.
6. Repeat steps 1-5 for all the available (annual) time series.

As a result, the parameter which can be considered as subjectively selected in this case is the ER percentage, while values over 80% have been proved, in practice, to be rational.

3.4. Triple Annual Maxima Series Approach

Sobey and Orloff [6] proposed a different and very simple approach on how to produce the data set, on which the r-largest maxima model would be applied. In particular, they used the series of monthly maxima, obtained from the initial H_S-time series and selected the three largest among them for each climate year. The resultant time series was named "triple annual maximum series", thus this method is referred to as the "triple annual maximum" (TAM) method (or model). Using arguments from the order statistics theory, the joint probability density function of the annual maximum, the annual second largest monthly wave and the annual third largest monthly wave were calculated, in order to produce the corresponding likelihood function. For each of the three maxima series they assumed the GEV or the log-normal distribution. Let us note that although this method resembles very much the r-largest maxima model, it is actually based on different principles. Under the same concept, in [31] the parameters of the GEV distribution were estimated using the four largest monthly maxima of each year.

4. NUMERICAL RESULTS AND DISCUSSION

As mentioned in previous sections, extreme value analysis can be used to estimate the design values (return levels for determined return periods) for a properly processed data set of significant wave height. In the present work, independent extreme-type data sets were extracted from buoy observations, using the procedures presented in Section 4, and the return levels associated with return periods of 10, 20, … and 100 years were estimated using the r largest maxima model. In this section, the procedures used for the analysis, the corresponding results and their comparison are discussed in detail.

4.1. Data Pre-processing and Declustering

The data examined in the present study consist of hourly buoy observations of significant wave height and the corresponding average period retrieved from the National Oceanic and Atmospheric Administration (NOAA)/National Data Buoy Center (NDBC) website (http://www.ndbc.noaa.gov/). In particular, from the NDBC buoy data available, the time series were extracted from the station 44005 located in the gulf of Maine 78 NM east of Portsmouth, moored at a depth of 201.2m.

As in most applications of extreme value theory, the data was firstly formed into "climate years", i.e. from June to May. The available buoy observations cover the time period from 1978 to 2008, but there is a number of values missing. The "valid" climate years, that formed the final time series, were selected based on the data availability: a year was only taken into account when the percentage of missing values was below 20%. Application of this criterion resulted in 16 valid years which formed our primary H_S time series, on which the declustering procedures presented in Section 3 were applied.

In particular, the algorithmic scheme of SSL procedure, presented in Section 3.1, was applied considering four characteristic storm length values (τ), namely 72h, 168h, 336h and 480h. The run length procedure, on the other hand, is provided through a number of functions by the R toolkit which were utilized, in the present work, for a total of nine (3x3) u- k (threshold- interval) combinations. Specifically, the cluster intervals 24, 72 and 168 hours were adopted following the suggestions in the available relevant literature, while the threshold values 1.2m, 1.9m and 2.4m were selected corresponding to the 50, 75 and 85% quantiles of the entire data set respectively. DeCA ran for ER 80, 85, 90 and 95% (ER was calculated for both energy content estimations $H_S^2 T_P$ and H_S^2), while for the TAM approach, the first, second and third largest monthly maxima of each year were directly extracted.

The declustered annual data sets were then sorted and new series comprising the first, second, etc largest maxima for each available year were obtained to be further used in the r largest maxima analysis.

4.2. Application of the r -largest Maxima Model

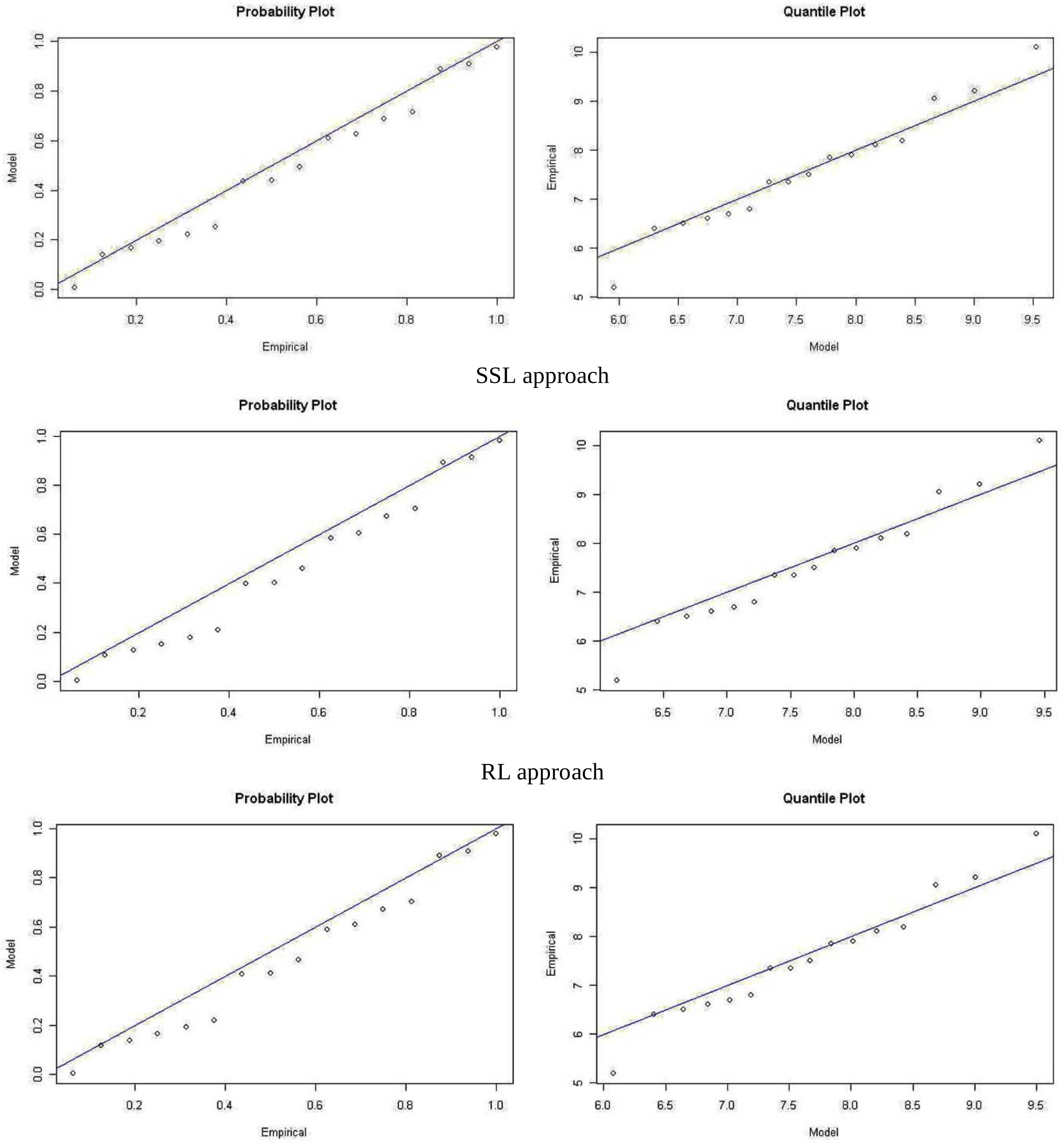
This data, in particular, were the input argument of the R toolkit through which the r-largest model was fitted, estimating the GEV parameters with the corresponding standard errors for different values of r ($r = 1, 2, ..., 10$)[1], using the maximum likelihood method. The diagnostics of these fits were then created in the form of a probability plot, a quantile plot, a return level plot and a histogram of data with fitted density. After an assessment of the obtained results, combined with theoretical arguments from the relevant literature, the value $r = 5$ was adopted as the most appropriate for the three declustering approaches, while, evidently, for the TAM series the analysis was applied for $r = 3$.

Subsequently, the diagnostic plots for $r = 5$ were closely examined and the declustering parameter combinations that produced the "best fit" for each declustering procedure were determined. More precisely, the best fits were obtained as follows: in the SSL approach for ssl=480h, in the Run Length approach for u=1.9m, k=24h, and in DeCA for ER=90%.

The Probability-Probability and the Quantile-Quantile Plot corresponding to the above fits are presented in Fig. (**1**). At this point, we should note that the results of the DeCA approach presented in this Section refer to the ones obtained by calculating the energy content as $H_s{}^2 T_p$. This case was chosen for producing a slightly better fit than the one which ignores the spectral peak period.

The GEV parameter estimates followed by the corresponding standard errors for each approach are presented in Table **2**. In the cases of SSL, Run Length and DeCA, the

[1] Except for the Triple Annual Maxima case where r is taken a priori equal to 3.



SSL approach

RL approach

DeCA approach

Fig. (1). The Probability (left) and Quantile (right) plots produced for the selected (best fits) of the SSL, RL and DeCA approaches.

presented results were produced using the declustering parameters selected above by applying the best-fit criterion.

Since the focus of this paper is on the effect of the different declustering approaches on the obtained results, in Table **2** we should firstly concentrate on the standard errors. In this context, we notice that the largest standard errors occur for the TAM approach, the lowest for the Run Length approach, while the standard errors of the SSL, Run Length and DeCA approaches are of comparable magnitude. Let us also note that, in contrast to the other declustering procedures, the TAM approach is robust, in the sense that it finally results in a unique combination of GEV parameters.

The sensitivity of the SSL, Run Length and DeCA approaches with regard to the selection of the required subjective parameters, i.e. SSL, u-k and ER respectively, was also explored. In this respect, the GEV estimates (μ, σ and ξ), which were produced for various values of the aforementioned parameters, were plotted for each declustering procedure and are presented in Fig. (**2**).

Table 2. The Maximum Likelihood Estimates of the GEV Parameters and the Corresponding Standard Errors for each Declustering Method

	μ	std error	σ	std error	ξ	std error
SSL	7.1455	0.2194	1.0504	0.1159	-0.1570	0.0952
RL	7.2575	0.2049	0.9875	0.1030	-0.1688	0.0852
DeCA	7.2312	0.2111	1.0138	0.1082	-0.1672	0.0905
TAM	7.0757	0.2536	1.1325	0.1264	-0.1954	0.1222
AMM	7.0961	0.3159	1.1349	0.2208	-0.2166	0.1738

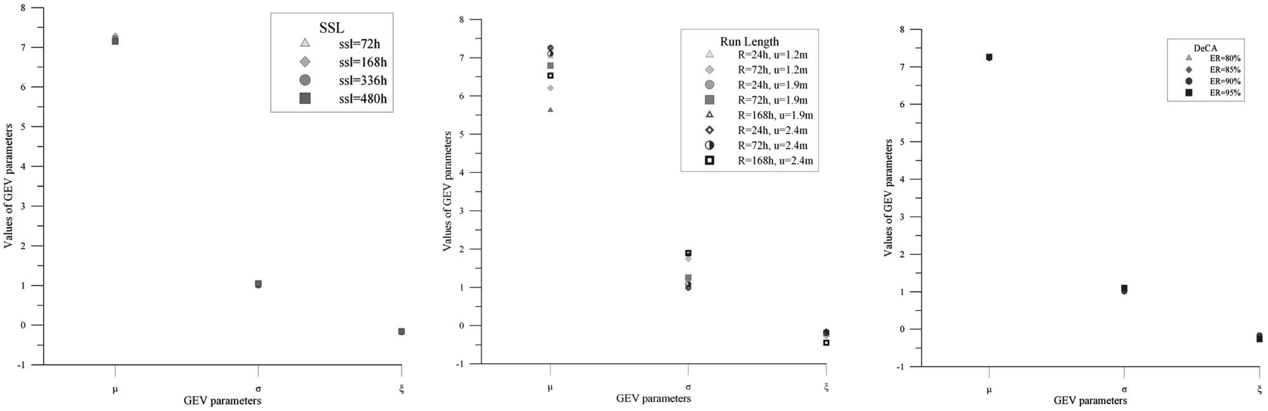


Fig. (2). GEV estimates produced using SSL (left), RL (centre) and DeCA (right) for the various values of declustering parameters.

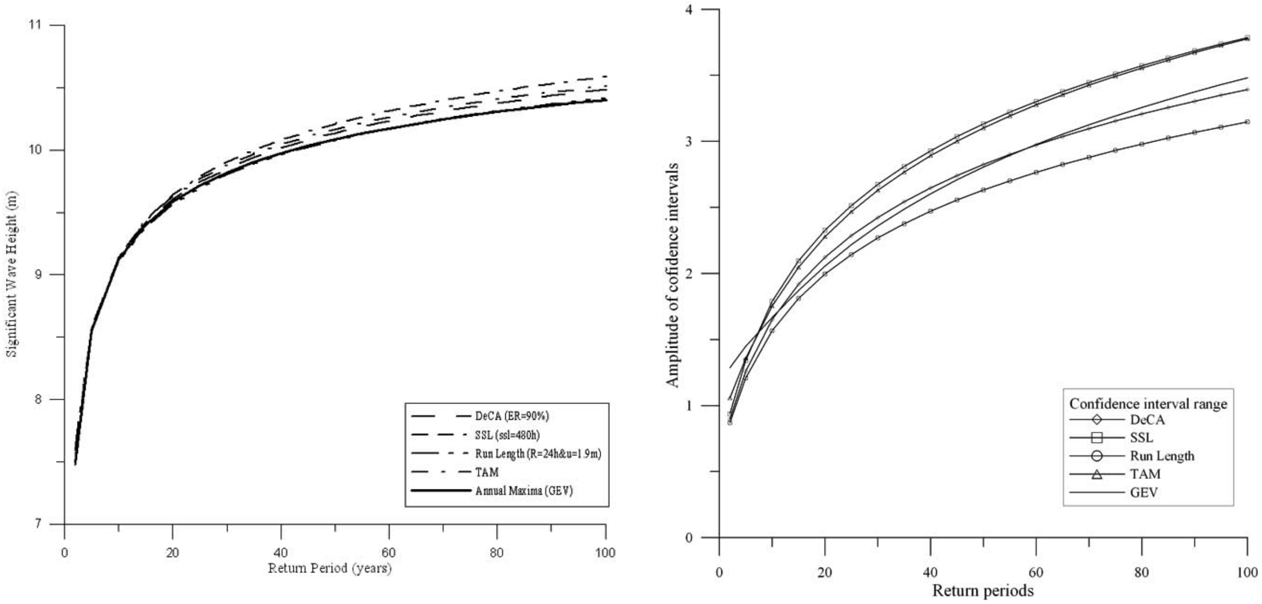


Fig. (3). The return levels estimated by the various approaches (left) and the widths of the corresponding confidence intervals (right) against return periods of 2 to 100 years.

As can be seen, the results obtained from the SSL and the DeCA approaches are very robust, as the GEV parameters exhibit very slight variations for the different values of SSL and ER respectively. More precisely, the resulted estimates for the SSL procedure exhibit a maximum deviation of 0.1338 in the values of μ, 0.0708 in the values of σ and 0.0201 in the values of ξ, while for DeCA the relevant values are 0.0425, 0.1084 and 0.0988 respectively. On the contrary, there is high variation among the results of the RL approach (1.6335 in μ, 0.9097 in σ and 0.3002 in ξ), revealing that this approach is highly sensitive to the selection of u and k.

Finally, the estimated GEV parameters obtained from all declustering procedures were used to calculate design values of significant wave height. The return levels obtained from the TAM, the AMM and the selected best fits of SSL, RL and DeCA approaches were plotted against return periods of up to 100 years (Fig. **3**, left). The widths of the corresponding confidence intervals, i.e. the differences between the upper and the lower confidence limit, obtained from all five approaches, are presented in Fig. **3** (right).

As can be seen in Fig. **3** (left), the calculated return levels do not exhibit significant differences among the various approaches, but in relation to the corresponding results of the AMM, they can be considered as slightly overestimating the obtained design values. In particular, for the return periods over 20 years, the TAM approach predicts the largest values of significant wave height, while the RL, of which the results are almost identical to those of the AMM, predicts the lowest. The return levels produced by the SSL and the DeCa approaches were in between the results of the TAM and RL. We should point out that the maximum relative deviation of the 100-year H_S design value is produced by the AMM and the TAM approaches and is approximately 2% (10.4 to 10.6m). On the other hand, as is depicted in Fig. (**3**) (right), the width of the confidence intervals varies significantly among the various approaches. In particular, the wider confidence intervals occur for the SSL and the TAM methods, while the narrower for the RL procedure. The confidence intervals obtained for the DeCA and the AM approaches are of moderate width.

5. CONCLUDING REMARKS

In this paper, the effect of the declustering procedures on the numerical results obtained by the r-largest maxima model was investigated. This model is more vulnerable to deviations from the iid assumption than the classical extreme value approaches (such as e.g. the annual maxima method), see [32] and therefore, the discrimination of independent cluster maxima in the initial time series is crucial. However, as mentioned in previous sections, almost all relevant declustering procedures involve the selection of a, more or less, arbitrary parameter. In the present study, the Standard Storm Length approach, the Run Length, the Triple Annual Maxima Series and the Declustering Algorithm were implemented, using a number of different values for the corresponding (subjectively selected) declustering parameters and the results were analytically examined.

In particular, a 16 year long time series of significant wave height was declustered by implementing the aforementioned procedures and the r-largest maxima model was applied on the obtained extreme type data in order to estimate the n-year design H_S sea-states, for $n = 10, 20, \ldots, 100$ years. The log-likelihood function was numerically maximized and the GEV distribution parameters with the corresponding standard errors were estimated. In this connection, the Run Length approach gave the lowest values of standard errors, while DeCA performed equally well, providing slightly larger values. Furthermore, the SSL and DeCA approaches exhibit a less dependent behaviour concerning the variation of the declustering parameters involved in the calculations with respect to the variability of the GEV distribution parameters. More precisely, both declustering procedures provided more stable GEV distribution parameters for different values of the declustering parameters.

On the other hand, regarding the design sea-state prediction, all declustering procedures provided very similar return periods - H_S -design value curves, as the maximum relative deviation was approximately 2%. Specifically, for return periods up to 20 years the obtained curves are almost indistinguishable. For return periods greater than 20 years, the SSL approach provides the largest design values (for SSL=480h), followed by the triple annual maxima method. The curve obtained by the RL procedure coincides with the curve obtained by the AMM, while DeCA provided slightly overestimated H_S -design values, in respect to the AMM.

Finally, more striking differences in the effect of the declustering procedures were detected in the confidence interval widths of the estimated design values. More precisely, the confidence interval width obtained by DeCA was found directly comparable to the one of the AMM for all values of the return period. The SSL and the TAM methods provided the widest confidence intervals, while the RL procedure gave the narrowest.

REFERENCES

[1] S. Coles, *An introduction to statistical modeling of extreme values*, USA, Springer, 2001.

[2] M. Ozger, and Z. Sen, “Return period and risk calculations for ocean wave energy applications”, *Ocean Eng.*, Vol. 35(17-18), pp. 1700-1706, 2008.

[3] T.H. Soukissian, G. Kalantzi, and I. Karagali, “Declustering of H_s-time series for applying the Peaks-Over-Threshold Method”, In: *Proc. 16th Int. Offshore Polar Eng. Conf.*, Vol. III, 2006, pp. 18-25.

[4] T.H. Soukissian, and G. Kalantzi, “Extreme value analysis methods used for wave prediction”, In: *Proc. 16th Int. Offshore Polar Eng. Conf.*, Vol. III, 2006, pp. 10-17.

[5] T.H. Soukissian and G. Kalantzi, “A new method for applying the r-largest maxima model for design sea-state prediction”, *Int. J. Offshore Polar Eng.*, Vol. 19(3), pp. 176-182, 2009.

[6] R. J. Sobey, and L. S. Orloff, “Triple annual maximum series in wave climate analyses”, *Coastal Eng.*, Vol. 26(3-4), pp. 135-151, 1995.

[7] C. A. T. Ferro, and J. Segers, “Inference for clusters of extreme values”, *J. R. Stat. Soc. B*, Vol. 65, pp. 545-556, 2003.

[8] I. Weissman, “Estimation of parameters and large quantiles based on the k largest observations”, *J. Am. Stat. Assoc.*, Vol. 73, pp. 812-815, 1978.

[9] M. R. Leadbetter, G. Lindgren, and H. Rootzen, *Extremes and related properties of random sequences and processes*, Springer-Verlag, New York, 1983.

[10] R.L. Smith, “Extreme value theory based on the r largest annual events”, *J. Hydrol.*, Vol. 86, pp. 27-43, 1986.

[11] J. A. Tawn, “An extreme-value theory model for dependent observations”, *J. Hydrol.*, Vol. 101(1-4), pp. 227-250, 1988.

[12] S. Zachary, G. Feld, G. Ward, and J. Wolfram, “Multivariate extrapolation in the offshore environment”, *Appl. Ocean Res.*, Vol. 20(5), pp. 273-295, 1998.

[13] J. H. Alves, and I. R. Young, “On estimating extreme wave heights using combined Geosat, Topex/Poseidon and ERS-1 altimeter data”, *Appl. Ocean Res.*, Vol. 25(4), pp. 167-186, 2003.

[14] I. D. Morton, J. Bowers, and G. Mould, “Estimating return period wave heights and wind speeds using a seasonal point process model”, *Coastal Eng.*, Vol. 31(1-4), pp. 305-326, 1997.

[15] I. D. Morton, and J. Bowers, "Extreme value analysis in a multivariate offshore environment", *Appl. Ocean Res.*, Vol. 18(6), pp. 303-317, 1996.

[16] M. Mathiesen, Y. Goda, P. Hawkes, E. Mansard, M. J. Martin, E. Peltier, E. F. Thompson, and G. van Vledder, “Recommended practice for extreme wave analysis”, *J. Hydraulic Res.*, Vol. 32(6), pp. 803-814, 1994.

[17] G. Van Vledder, Y. Goda, P. Hawkes, E. Mansard, M. J. Martin, M. Mathiesen, E. Peltier, and E. Thompson, “Case studies of ex-

treme wave analysis: A comparative analysis." In: *Proc. 2nd Int. Symp. Ocean Wave Meas. Anal.*, New Orleans, Louisiana, 1993, pp. 978-992.

[18] C. G. Soares, and M. G. Scotto, "Application of the r largest-order statistics for long-term predictions of significant wave height", *Coastal Eng.*, Vol. 51(5-6), pp. 387-394, 2004.

[19] C. G. Soares, and S. Caires, "Changes in spectral shape due to the effect of finite water depth", In: *Proc. 14th Int. Conf. Offshore Mech. Arctic Eng., ASME,* 1995; pp. 547-556.

[20] C. G. Soares, and M. Scotto, "Modelling uncertainty in long-term predictions of significant wave height", *Ocean Eng.*, Vol. 28(3), pp. 329-342, 2001.

[21] A. Butler, J. E. Heffernan, J. E. Tawn, R. A. Flather, and K. J. Horsburgh, "Extreme value analysis of decadal variations in storm surge elevations", *J. Marine Syst.,* Vol. 67(1-2), pp. 189-200, 2007.

[22] R. L. Smith, "Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone", *Stat. Sci.*, Vol. 4(4), pp. 367-377, 1989.

[23] D. Walshaw, "Getting the most from your extreme wind data: a step by step guide", *J. Res. Natl. Inst. Stand. Technol.*, Vol. 99, pp. 399-411, 1994.

[24] L. Fawcett, and D. Walshaw, "Markov chain models for extreme wind speeds", *Environmetrics* Vol. 17, pp. 795-809, 2006.

[25] J. L. Wadsworth, J. A. Tawn, and P. Jonathan, "Accounting for choice of measurement scale in extreme value modeling", *Ann. Appl. Stat.*, In Press, 2010.

[26] S. Caires, and A. Sterl, "100-Year return value estimates for ocean wind speed and significant wave height from the ERA-40 data", *J. Clim.*, Vol. 18, pp. 1032-1048, 2005.

[27] M.A. Hemer, J.A. Church, and J.R. Hunter, "Waves and climate change on the Australian coast", *J. Coast. Res., SI,* Vol. 50, pp. 432-437, 2007.

[28] S. Caires, and M. R. A. van Gent, "Extreme wave loads", In: *Proc. 27th Int. Conf. Offshore Mech. Arctic Eng. ASME,* Estoril, Portugal, paper No. 57947, 2008.

[29] R. Smith, and I. Weissman, "Estimating the Extremal Index", *J. Roy. Stat. Soc. Ser. B (Methodological)*, Vol. 56(3), pp. 515-528, 1994.

[30] L. Fawcett, and D. Walshaw, "Improved estimation for temporally clustered extremes", *Environmetrics* Vol. 18, pp. 173-188, 2007.

[31] V. C. Panchang, C. Jeong, and D. Li, "Wave climatology in coastal maine for aquaculture and other applications", *Estuaries and Coasts,* Vol. 31(2), pp. 289-299, 2008.

[32] R.L. Smith, "Statistics of extremes, with applications in environment, insurance and finance", In: *Extreme Values in Finance, Telecommunications and the Environment,* B. Finkenstadt, and H. Rootzen, Eds., Chapman and Hall/CRC Press*:* London, 2003, pp. 1-78.

Received: November 20, 2010 Revised: November 20, 2010 Accepted: November 25, 2010