Distribution functions of extreme sea waves and river discharges

Fonctions de distribution des vagues de mer et des décharges de rivières extrêmes

PIETER H.A.J.M. VAN GELDER, IAHR Member, Section of Hydraulic Engineering and Geosciences, Delft University of Technology, The Netherlands. Stevinweg 1, 2628 CN Delft, P.O. Box 5048, 2600 GA Delft, The Netherlands. Tel.: +31-15-2786544; fax: +31-15-2785124; e-mail: P.H.A.J.M.vanGelder@Tudelft.nl

CONG V. MAI, Faculty of Coastal Engineering, Water Resources University of Vietnam, 175 Tayson, Dong Da, Hanoi, Vietnam & Section of Hydraulic Engineering, Faculty of Civil Engineering and Geosciences, Delft University of Technology, The Netherlands. Stevinweg 1, 2628 CN Delft, P.O. Box 5048, 2600 GA Delft, The Netherlands. Tel.: +31 15 278 4735; fax: +31 15 278 5124; e-mail: C.MaiVan@Tudelft.nl

ABSTRACT

Some of the important elements to be considered by the designer in the field of water defenses and hydraulic structures include the determination of the maximum environmental loads such as maximum wave height, maximum water level and maximum river discharge at locations of interest. These maxima can be estimated by various statistical methods based on observations. The main point of interest is the behaviour of each method for predicting these extrema and its deviation from the true value. An overview of statistical methods is given to determine the extreme values of river and sea related variables. The method of Regional Frequency Analysis (RFA) is proposed to predict the occurrence probabilities of these extrema. The RFA intends to extend the dataset by pooling the data of neighboring locations and establishes the regional growth curve of a so-called “homogenous region”. An important aspect is the formation of the homogeneous regions because only for those regions data can be pooled. Statistical methods will also be presented to derive homogeneous clusters. Applications include the prediction of extreme river discharges in Northwest and Central Europe and extreme waves along the Dutch North Sea coasts.

RÉSUMÉ

Certains des éléments à considérer par l’ingénieur opérant dans le domaine des défenses fluviales et maritimes et des structures hydrauliques comprennent la détermination des chargements environnementaux maximaux tels que la hauteur des vagues, le niveau des eaux et la décharge d’eau maximaux aux points critiques. Ces maxima peuvent être estimés par diverses méthodes statistiques basées sur des observations. Le point d’intérêt principal est le comportement de chaque méthode pour prédire ces extrema et leur déviation de la valeur réelle. Les méthodes statistiques pour prédire les valeurs extrêmes des rivières et de la mer sont passées en revue. La méthode de l’analyse fréquentielle régionale (AFR) est proposée pour prédire les probabilités d’occurrence de ces extrema. L’AFR consiste à étendre le jeu de données en groupant les données de régions voisines et établir la courbe de croissance régionale d’une région dite homogène. Un aspect important est la formation de régions homogènes car seules ces régions peuvent être groupées. Des méthodes statistiques sont aussi présentées pour définir des groupes homogènes. Ces méthodes peuvent être appliquées aux décharges extrêmes des rivières au Nord Ouest et au centre de l’Europe ainsi qu’aux vagues extrêmes le long des côtes Holländaises de la mer du Nord.

Keywords: Extreme values, goodness of fit, homogeneity, L-moments, maximum wave height, probability distribution, robust statistics, wind setup

1 Introduction

In water defense and hydraulic engineering, important elements to be considered by the designer include the determination of the maximum environmental variables such as the significant and the maximum wave height, the maximum water level (sea, river, and lake), the maximum river discharge and its corresponding water level at a location of interest. The design of hydraulic engineering structures and insurance risk calculations, usually rely on knowledge of the occurrence frequency of these extreme events. The estimation of these frequencies is, however, difficult because extreme events are by definition rare and data records are often short. In other words: The uncertainties related to the distribution analysis of extreme values are high. The parameters of these distribution functions for estimation of extreme values can be estimated by various methods. The main point of interest is the behaviour of each method for predicting the p-quantile, i.e., the value which is exceeded by the random variable with probability p, where p ≤ 1 (van Gelder, 2000). The estimation of extreme quantile corresponding to a small probability of exceedance is commonly required in the risk analysis of hydraulic structures. Such extreme quantiles may represent

Revision received August 28, 2007! Open for discussion until December 31, 2008.

280
design values of environmental loads (wind, waves, snow, and earthquake), river discharges, and flood levels specified by design codes and regulations.

It is desirable that the quantile estimate is unbiased, that is, its expected value should be equal to the true value. It is also desirable that an unbiased estimate be efficient, i.e., its variance should be as small as possible. The problem of unbiased and efficient estimation of extreme quantiles from small samples is commonly encountered in the civil engineering practice. For example, annual flood discharge data may be available for the past 50 to 100 years and on that basis one may have to estimate a design flood level corresponding to a 1000 to 10,000 years return period. Due to this fact the statistical extrapolation to predict extreme values is often contaminated by sampling and model uncertainty. This has motivated the development of approaches to enlarge the sample size in the extreme value analysis.

This paper presents actual research available to estimate the extreme waves, surges and river distributions including techniques to enlarge the dataset by using available information. The method of Regional Frequency Analysis (RFA) is applied to predict the occurrence probabilities of extreme values of these variables. Two specific applications are addressed in this paper: (i) Determination of probability distributions of river data and (ii) estimation of extreme maximum values for sea data.

2 Overview of study approach

It has long been recognized that many annual environment datasets are too short to allow for a reliable estimation of extreme events. Thus, the difficulties are related both to the identification of the appropriate statistical distribution for describing the data and to the estimation of the parameters of a selected distribution. Therefore, the distribution for one sample can be more accurately estimated by using information not just from that sample but also from the other related samples. In the environmental sciences, the data samples are typically measurements of the same kind of data made at different sites, and the process of using data from several sites to estimate the frequency distribution is known as regional frequency analysis. Regionalization provides a means to cope with this problem by assisting in the identification of the shape of potential parent distributions, leaving only a measure of scale to be estimated from the at-site data. This approach is used in this study to overcome the previously mentioned difficulties.

Regional flood frequency analysis involves two major steps: (1) Grouping of sites into homogeneous regions and (2) regional quantile estimates at the sites of interest. The performance of any regional estimation methods strongly depends on the grouping of sites into homogeneous regions. Geographically contiguous regions were used for a long time in hydrology, but have been criticized for being of arbitrary character, because the geographical proximity does not guarantee hydrological similarity. During the last ten years researchers have attempted to develop methods in which similarity between sites is defined in a multidimensional space of catchment-related or statistical characteristics.

A significant contribution to solve the delineation issue is the region-of-influence approach (e.g., Burn, 1990; Feaster and Tasker, 2002). This method dispenses completely with the classical notion of regions in that each site is allowed to have its own region. The site of interest is located at the center of gravity in a space of relevant flood and/or catchments characteristics, each weighted properly according to its relevance. The method also involves the choice of a distance threshold; only sites whose distance to the target site (in the weighted attribute space) does not exceed this threshold are included in the region-of-influence. An advantage of the region-of-influence method is that each site can be weighted according to its proximity to the site of interest in the estimation of a regional growth curve.

In this paper, the cluster analysis is used as a first attempt to group sites into homogeneous region. The delineation of a homogeneous region is closely related to the identification of the common regional distributions that apply within each region. A region can only be considered homogeneous if sufficient evidence can be established that the data at different sites of the region are drawn from the same parent distribution, except for the scale parameter. L-moment ratio diagrams and L-moment diagrams were used as popular tools for regional distribution identification, testing for outlier sites, identifying a regional distribution in numerous other studies (Chowdhury et al., 1989; Pilon and Adamowski, 1992; Vogel and Finnesey, 1993; Hosking and Wallis, 1993). Further, Hosking and Wallis (1997) developed several tests for regional studies. They gave guidelines for judging the degree of homogeneity of a group of sites, and for choosing and estimating a regional distribution. These L-moments related techniques are also applied in this study to test the sea wave and river discharge datasets.

3 Statistical background for RFA

3.1 L-moment statistics

L-moments are summary statistics for probability distributions and data samples. They are analogous to ordinary moments, because of providing measures of location, dispersion, skewness, kurtosis, and other aspects of the shape of probability distributions or data samples, based on the linear combination of the ordered data. The theoretical advantages of L-moments over ordinary moments and detail determination of L-moments for sample data and probability distributions can be found in Hosking and Wallis (1997).

3.2 Discordance measure

Suppose that there are N sites in the regions. Let \( u_i = [t_{i1}, t_{i2}, \ldots, t_{ik}]^T \) be a vector containing the \( t_{ij} \) values for site \( i \) with the superscript “T” denoting transposition of a vector. Let \( \bar{u} \) be the un-weighted group average. The matrix \( A \) of sums of squares and cross-products is defined by

\[
A = \sum_{i=1}^{N} (u_i - \bar{u})(u_i - \bar{u})^T. \tag{1}
\]
Then the discordance measure $D_i$ is

$$D_i = \frac{1}{n} \sum_{j=1}^{n} (x_{ij} - \bar{x}_j)^T A^{-1} (x_{ij} - \bar{x}_j)$$

A site is declared as discordance if its $D_i$ value exceeds the critical value. The critical value depends on the number of sites to be considered; for regions with more than 15 sites, the critical value is 3.0. The critical value decreases as the number of sites reduces (Hosking and Wallis, 1997).

### 3.3 K-means clustering

Suppose that there are already hypotheses concerning the number of clusters in the considered cases or variables. One may wish to form exactly $K$ clusters that are to be as distinct as possible. This is the type of research question that can be addressed by the K-means clustering algorithm. In general, this produces exactly $K$ different clusters of greatest possible distinction.

In a river map, a “hunch” may fall basically into eight different categories with regard to the physical aspects. It can be questioned whether this intuition can be quantified, i.e., does a K-means cluster analysis of the physical quantities indeed produce the eight clusters as expected. If so, the means on the different measures of physical quantities for each cluster would represent a quantitative way of expressing the hypothesis or intuition such as mountainous sites in cluster 1 or densely populated sites in cluster 2, etc.

Computationally, one may think of this method as an analysis of variance (ANOVA) “in reverse.” The method will start with $K$ random clusters, and then move objects between those clusters with the goal to (i) minimize the variability within the clusters and (ii) maximize the variability between the clusters. This is analogous to “ANOVA in reverse” in the sense that the significant test in ANOVA evaluates the between group variability against the within-group variability when computing the significant test for the hypothesis that the means in the groups are different from each other. In K-means clustering, objects are moved in and out of groups (clusters) to get the most significant ANOVA results.

Usually, as a result of the K-means clustering analysis, the means for each cluster on each dimension would be examined to assess how distinct the $K$ clusters are. Ideally, very different means for most, if not all dimensions, used in the analysis would be obtained. The magnitude of the $F$ values from the analysis of variance performed on each dimension is another indication of how well the respective dimension discriminates between clusters (Hartigan and Wong, 1979; Everitt, 1993).

### 3.4 Robust distances

One may think of the variables as defining a multidimensional space in which each observation can be plotted. Also, one can plot a point representing the means of all variables. This “mean point” in the multidimensional space is also called the centroid. The Mahalanobis distance is the distance of a case from the centroid in the multidimensional space. Thus, this measure provides an indication of whether or not an observation is an outlier. The classical Mahalanobis distance is defined as (Mahalanobis, 1934)

$$MD^2 = (x_i - T(X))^T C(X)^{-1} (x_i - T(X))$$

where $T(X)$ and $C(X)$ are the usual mean and covariance estimates

$$T(X) = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$C(X) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$$

and $x_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ for $i = 1, ..., n$ is the $i$th row of data matrix $X$; $n$ is the number of observations; and $p$ the dimension of space. Points whose $MD^2$ is large are flagged. It is well known, however, that the sample mean and covariance in a multivariate dataset are extremely sensitive to outliers. Rousseeuw and Leroy (1987) proposed to replace the classical mean and covariance in the expression of the Mahalanobis distance by their high breakdown point (HBP) robust analogues (such as the MVE (Minimum Volume Ellipsoid) or MCD (Minimum Covariance Determinant)).

The breakdown point (BP) is the smallest percentage of contaminated data that can cause the estimator to take on arbitrarily large aberrant values. The BP of the classical estimates based on the method of Maximum Likelihood, the method of Moments, or the method of least squares is zero.

The MCD estimator for location $T(X)$ is defined as the mean of the $k$ points of $X$, where $k$ is equal to $(n + p + 1)/2$, for which the determinant of the covariance matrix is minimal. Moreover, Vandev and Neykov (1993) showed that the breakdown point of the MCD is equal to $(n - k)/n$ if $k$ is within the bounds $(n + p + 1)/2 \leq k \leq n - p - 1$ and $n \geq 3(p + 1)$. Note that the number of observations should be at least three times larger than the dimensionality $+1$. If $k = (n + p + 1)/2$, then the BP is equal to $1/2$ asymptotically. For more information on the efficient algorithms for calculating the MCD and other robust covariance estimates with high breakdown points, see Rocke and Woodruff (1996).

To improve the efficiency of the estimates, Rousseeuw and van Zomeren (1990) performed one step improvements for the location and scatter as

$$T_1(X) = \frac{\sum x_i xi}{\sum w_i}$$

$$C(X) = \sum_{i=1}^{n} w_i (x_i - T_1(X))(x_i - T_1(X))^T / \sum w_i$$

where

$$w_i = \begin{cases} 1 & \text{if } (x_i - T_1(X))^T C(X)^{-1} (x_i - T_1(X)) \leq c \\ 0 & \text{otherwise} \end{cases}$$

with the cut-off value $c = X_{p,0.975}^2$. The observations with zero weights can be interpreted as outliers.
The main idea of the regional frequency analysis is to extend the estimation process in this regional frequency analysis. To aid the presentation, a formal definition is given as follows: Let \( Q_i(F) \), be the quantile function of the distribution at site \( i \).

The site dependence scale is naturally estimated as \( \hat{\mu}_i = \bar{Q}_i \), the sample mean of data at site \( i \). The basic dimensionless rescaled data \( q_j = Q_j/\hat{\mu}_j \), \( j = 1, \ldots, n_i \), \( i = 1, \ldots, N \), are used for estimating the regional growth curve \( q_i(F), 0 < F < 1 \).

How can homogeneous regions be derived on the basis of statistical techniques and physics-based considerations? Hosking and Wallis (1997) developed a unified robust approach to the RFA, based on the \( L \)-moments described by Hosking (1990) involving objective and subjective techniques for defining homogeneous regions, of assigning sites to regions, identifying and fitting regional probability distributions to data and testing hypotheses about distributions. By robustness Hosking and Wallis (1997) refer to statistics that work well even if the data are contaminated or the model assumptions are slightly violated. The advantages of their approach over the conventional method of moments and the maximum likelihood method are the smaller effect of outliers and more reliable inference from small samples, as the \( L \)-moments are a linear combination of data. The physical based consideration supports this data screening to see if datasets have similar physical characteristics for a homogeneous consideration and it later can be evident for excluding heterogeneity sites.

In general the four main steps of the RFA procedure are: (i) screening of data; (ii) identification of homogeneous regions; (iii) choice of a regional frequency distribution; and (iv) estimation of the regional frequency distribution (Hosking and Wallis, 1997). Hosking and Wallis (1997) recommend the standard discordance measure of Wilks (1963) to be used for identifying unusual sites in a region in terms of the sample \( L \)-moments ratios; a heterogeneity measure for assessing whether a proposed region is homogeneous; and a “goodness-of-fit measure” for assessing whether a candidate distribution provides an adequate fit to the data. The RFA is an iterative procedure. However, Hosking and Wallis (1997) emphasize physical reasoning rather than formal statistical significance in data processing. The Wilks’ discordance measure \( D_i \), based on the sample \( L \)-moments ratios (\( L-CV, L\text{-SKEW}, \text{ and } L\text{-KURT} \)) for each site as proposed by Hosking and Wallis (1997) should be used in the screening of data by discordant sites. It is well known that the standard measure of Wilks being equal to the Mahalonobis distance up to a fixed constant for the detection of multivariate outliers is not as robust as when based on the sample mean and the covariance matrix. Alternatives, based on robust estimates of multivariate location and scatter include the approach of Rousseeuw and van Zomeren (1990). The Wilks’ discordance measure \( D_i \), and several of its alternatives denoted by \( RD_i(\cdot) \) are applied in this study.

5 Case study: Rivers in north-western and central europe

5.1 Database

Datasets of daily river discharges of 213 locations in North Western and Central Europe over a time period in terms of location from 1 year to almost 200 years are available. The statistics of the annual averaged flows, the annual maximum flows, the annual minimum flows, and the mean annual precipitation, population density and elevation of the location were obtained from maps from BosAtlas [source: http://www.bosatlas.nl/hipsp/portal/bosatlas]. To examine a possible relationship between some of the variables, a classical correlation analysis was performed.
As can be seen from Table 1 no significant correlations were identified. The 213 stations are shown in Fig. 1. Figure 2 shows the daily river discharges of a river at a certain location including the yearly maximum, minimum and average discharges. The maximum values are of interest hereafter.

Following an initial data inspection, 168 sites were selected for further analysis. The remaining 45 sites were not used because of incomplete data values or too short observed periods (<15 years).

### 5.2 Formation of regions and estimation of regional distribution

The $K$-means clustering of site characteristics was performed following Hartigan and Wong (1979). The latitude, longitude, the elevation, the annual precipitation, the population density and the size of basin area were available from the database. The following transformations of the site characteristics were defined:

- **Size of basin area**: $Y = 3 \log(X)/\sigma(\log(X))$
- **Elevation**: $Y = \sqrt{X}/\sigma(\sqrt{X})$
- **Latitude and Longitude**: $Y = X/\sigma(X)$

As a result, eight clusters were derived from the $K$-means clustering algorithm. The next step included the testing for homogeneous regions based on Wilks discordance test and/or robust discordance distances. The at-site $L$-moment ratios were calculated for all sites at each cluster and the corresponding heterogeneity measure $H$ was determined. Hosking and Wallis (1997) consider a region to be acceptably homogeneous if $H < 1$, “possibly” homogeneous if $1 \leq H < 2$ and definitely heterogeneous if $H \geq 2$. If the region is not acceptably homogeneous, some redefinition of the region was considered by excluding some outliers, based on Wilks discordancy test and/or robust discordancy distances.

The test of all obtained regions based on the cluster analysis gave definitely heterogeneous regions by the $H$-measure. Therefore, the regions were reduced by excluding the discordant sites according to the robust distances, as the classical distance of Wilks in most of the cases was insignificant. This was expected based on the simulation results of van Gelder and Neykov (1999).

In this way, by several iterations of excluding discordant sites, the following six homogeneous regions resulted:

- Region 1: England and Northern Europe (10 sites)
- Region 4: South France (17 sites)
- Region 5: Thames River with 6 other sites (7 sites)
- Region 6: Western United Kingdom (11 sites)
- Region 8: Rhine and Danube rivers (10 sites)
- Region h33: Dutch and German sites (12 sites)

In total 67 sites or approximately 40% of all 168 sites were assigned to the above 6 homogeneous regions. Then the regional growth curves were estimated, followed by the estimations of the at-site quantile. For each region a goodness-of-fit statistic was calculated. For the regions with a “significant” correlated structure a modified algorithm of Hosking and Wallis (1997) for the RFA was used to improve the reliable estimates of the $H$-statistics and of the confidence bounds. Because of the high correlations of the annual maxima (every year added only one maximum value in the dataset) of the sites in the relatively small Dutch region, the modified algorithm appeared to be essential. The French region (20 sites) appeared to be extremely homogeneous with a low correlation structure. An application of the RFA...
for the river datasets in the Dutch regions results in Table 2. It presents the quantile estimates per location, as well as the RFA estimates of the regional growth curve and the goodness-of-fit measures for 4 distribution types.

It is emphasized that without the use of the robust discordancy measures it would have been impossible to derive the homogeneous regions so easily. If \( H > 1 \), then the robust distances of discordancy were also significant for some of the sites. This allows to conclude that the robust distances support the main decision making process, which is normally the most time-consuming part of the analysis. This gives an important contribution in performing a full RFA.

The largest possible homogeneous region that includes the Dutch rivers contains ten sites. These ten sites are made up by 3 Dutch sites, 1 Belgian site and 6 German sites. Adding low-lying Dutch rivers contains ten sites. These ten sites are made up by 3

Table 2: Regional frequency analysis of river datasets in Dutch region

<table>
<thead>
<tr>
<th>Site ID</th>
<th>Location</th>
<th>Scale</th>
<th>Shape</th>
<th>L-CV</th>
<th>L-Skew</th>
<th>N</th>
<th>Quantile estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.01 0.1 0.5 0.9 0.99 0.999</td>
</tr>
<tr>
<td>15</td>
<td>0.8626</td>
<td>0.2898</td>
<td>0.115</td>
<td>0.1826</td>
<td>0.0981</td>
<td>30</td>
<td>0.3788 0.6089 0.9666 1.4372 1.8979 2.244</td>
</tr>
<tr>
<td>16</td>
<td>0.8706</td>
<td>0.3041</td>
<td>0.1773</td>
<td>0.1833</td>
<td>0.0609</td>
<td>30</td>
<td>0.3373 0.5973 0.9785 1.4348 1.8268 2.0815</td>
</tr>
<tr>
<td>26</td>
<td>0.8837</td>
<td>0.3436</td>
<td>0.3054</td>
<td>0.1925</td>
<td>−0.012</td>
<td>30</td>
<td>0.2152 0.5574 1.0029 1.4429 1.7327 1.8724</td>
</tr>
<tr>
<td>124</td>
<td>0.8556</td>
<td>0.2999</td>
<td>0.1057</td>
<td>0.1902</td>
<td>0.1037</td>
<td>30</td>
<td>0.3586 0.5941 0.9633 1.4561 1.9478 2.3253</td>
</tr>
<tr>
<td>125</td>
<td>0.8543</td>
<td>0.2968</td>
<td>0.0947</td>
<td>0.1918</td>
<td>0.1105</td>
<td>30</td>
<td>0.3667 0.5968 0.9612 1.4558 1.961 2.3588</td>
</tr>
<tr>
<td>27</td>
<td>0.8904</td>
<td>0.3518</td>
<td>0.3493</td>
<td>0.193</td>
<td>−0.0359</td>
<td>30</td>
<td>0.1805 0.5498 1.0115 1.4387 1.6957 1.8075</td>
</tr>
<tr>
<td>28</td>
<td>0.8306</td>
<td>0.3142</td>
<td>0.04</td>
<td>0.2102</td>
<td>0.1445</td>
<td>30</td>
<td>0.3539 0.5642 0.945 1.5068 2.1508 2.7268</td>
</tr>
<tr>
<td>29</td>
<td>0.764</td>
<td>0.3351</td>
<td>−0.1146</td>
<td>0.2612</td>
<td>0.2458</td>
<td>30</td>
<td>0.2946 0.4975 0.8895 1.6243 2.7938 4.2933</td>
</tr>
<tr>
<td>30</td>
<td>0.9104</td>
<td>0.3455</td>
<td>0.4404</td>
<td>0.1828</td>
<td>−0.0838</td>
<td>30</td>
<td>0.1578 0.5622 1.0274 1.4037 1.5915 1.6575</td>
</tr>
<tr>
<td>31</td>
<td>0.8382</td>
<td>0.3207</td>
<td>0.0604</td>
<td>0.2109</td>
<td>0.1317</td>
<td>30</td>
<td>0.32 0.5587 0.9493 1.5078 2.121 2.644</td>
</tr>
<tr>
<td>Average for all sites</td>
<td>0.2945 0.5687 0.9695 1.4708 1.9719 2.4011</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regional L-moments ratio ( l_1, l_2, l_3, l_4 ) &amp; ( l_5 )</td>
<td>0.2692 0.5669 0.968 1.4685 1.9278 2.2479</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inter-site correlation</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distribution accepted and its goodness of fit measure</td>
<td>GLO GEV LN3 PE3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inter-site correlation</td>
<td>0.3500 0.6050 0.5950 0.6350</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6 Extreme value analysis of sea wave data

6.1 Review of wave statistical models

For the short-term scale of a few hours, Longuet-Higgins (1952) showed that the Rayleigh distribution is the most appropriate to describe the distribution of individual waves extracting from the sea water surface signal at a certain point and time. Without using a spectral description, various parameter estimation methods are available to determine the free parameter of the Rayleigh distribution. Green (1994) adapted the Rayleigh distribution to account for wave breaking in shallow waters. van Gelder and Vrijling (1999) investigated the performance of the parameter estimation methods for the Rayleigh distribution. These methods were compared with each other in terms of the relative bias and root mean squared error of the p-quantile \( (p < 1) \). Rodriguez et al. (1999) investigated the uncertainty of various sea state parameters, under which the significant wave height \( H \) resulted from the methods of spectral estimation. The use of the spectral estimator \( 4(m_0)^{1/2} \), in which \( m_0 \) is the 0th order spectral moment \( \int_0^{\infty} S(f)df \), appears to be a robust estimator, in the sense that it does not differ much with the choice of the estimation method of \( S(f) \). Three different spectral estimation methods, namely the Blackman–Tukey, the fast fourier transform and the maximum entropy resulted in similar estimates (Rodriguez et al., 1999).

Extreme wave height distributions on long term scales of years were investigated by numerous authors. Battjes (1970) noticed that symmetric distributions, such as the normal distribution, were not suitable to describe the long-term distribution for the significant wave heights. Skewed distributions, such as the Gumbel and Weibull distribution, fitted much better. To the same conclusions came Teng et al. (1993) and Teng and Palao (1996), who analyzed wave buoy data of the Pacific and Atlantic Oceans, Goda and Kobune (1990) with wave data from the Sea of Japan and the East China Sea, Rossouw (1988) with wave data from the Indian Ocean, van Vledder et al. (1993) with wave data from the Norwegian Coast, and Burchard and Liu (1994) with data from the Mediterranean Sea. Mathiesen et al. (1994) and Goda et al. (1993) developed a guideline for extreme wave analysis by using the 3-parameter Weibull distribution.

Guedes Soares and Ferreira (1995) proposed a model to describe the long-term extreme wave height distribution by...
averaging past distribution functions; if a priori distribution functions for the model parameters are given, over the possible values were conditioned. In this way they accounted for the long-term time-varying character of the significant wave height data.

A similar approach was adopted in this paper, by accounting for the variable space character of the significant wave height data. The proposed RFA will be applied to the distribution function of the significant wave heights on long-term scale. Instead of using long term observation records (some hundred years to some thousand years, which are obviously not available), RFA uses data from several sites in the so-called homogeneous regions, which have frequency distributions similar to the site of interest, in estimating event frequencies at that site.

6.2 The sea wave datasets

From nine measurement stations (Figure 3 and Table 3) on the Dutch North Sea, data were available from the Dutch Ministry of Transport, Public Works and Water Management (RIKZ). The station name, the coordinates and the acronym are indicated. The following information is available at the nine sites: date (yyyy mm dd); time (hhmm: UT); and the significant wave height \( H_{00} \) (m). The data are peaks over thresholds. They are made identically independent distributed by using a filter of 48 hours. Each site contains between 20 to 70 peaks over the period 1979–1993. As stressed by Castillo and Sarabia (1992), and Ferreira and Guedes Soares (1998) peaks over thresholds provide a modern and soundly based input for extrapolation problems. A regionalized frequency analysis of POT data of the nine stations will be performed below.

6.3 RFA of wave heights

The objective is to identify regional probability distributions for the extreme wave heights. For each site \( L \)-moment ratios \( (L-CV, L-Skewness, L-Kurtosis) \), discordance measure and robust distances were calculated (Table 4). The weighted means (based on the record-lengths) of \( L-CV, L-Skewness, \) and \( L-Kurtosis \) are 0.0609, 0.2330, and 0.1064, respectively. The number of sites should be at least 3 times larger than the dimensionality +1 in order to use robust discordancy measures (van Gelder and Neykov, 1999). Because the dimensionality is 3 \( (L-CV, L-Skewness, \) and \( L-Kurtosis) \), the required number of sites should be at least 10. However, if this requirement for the moment is neglected then the station MPN should be considered as discordant according to the Mahalanobis distance \( (2D(i)) = 6.47; \) the 90% critical value is 6.25, see Table 4, second row.

The \( L \)-moment diagrams are shown in Fig. 4 including the unusual sites (outside the inner \( 1 \)-sigma or outer \( 2 \)-sigma ellipse) whose data need a closer examination; i.e., those sites whose \( L \)-moments are notably different from those of the other sites in the data set. Site MPN is the left-lower star in Fig. 4 (left). It is far outside of the \( 2 \)-sigma ellipse. Together with the rather high Mahalanobis distance of MPN (Table 4), it was decided to exclude site MPN from the RFA. Thus, the RFA was performed for the remaining 8 sites.

A RFA consists of the following three steps: (i) the Wilks discordancy measure for outliers; (ii) the heterogeneity measures; and (iii) the goodness-of-fit measures to find the “best” fit distribution. From step (i), it followed that MPN was a discordant site. Figure 5 shows the frequency exceedance curves for the nine locations (MPN with crosses). It was difficult to notice the discordant behaviour of site MPN graphically. In Fig. 6, after a data

---

Table 3 Locations of wave height measurements

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>K13 platform</td>
<td>10.176</td>
<td>583.334</td>
<td>53°13′04″</td>
<td>3°13′13″</td>
<td>K13</td>
</tr>
<tr>
<td>Schiermonnikoog noord</td>
<td>206.527</td>
<td>623.438</td>
<td>53°35′44″</td>
<td>6°10′00″</td>
<td>SON</td>
</tr>
<tr>
<td>Eierlande Gat</td>
<td>106.514</td>
<td>587.985</td>
<td>53°16′37″</td>
<td>4°39′42″</td>
<td>ELD</td>
</tr>
<tr>
<td>Huisduinen muntenorplaats</td>
<td>64.779</td>
<td>507.673</td>
<td>52°33′00″</td>
<td>4°03′30″</td>
<td>YM6</td>
</tr>
<tr>
<td>Noordwijk meetpost</td>
<td>80.443</td>
<td>476.683</td>
<td>52°16′26″</td>
<td>4°17′46″</td>
<td>MPN</td>
</tr>
<tr>
<td>Euro platform</td>
<td>9.963</td>
<td>447.601</td>
<td>51°59′55″</td>
<td>3°16′35″</td>
<td>EUR</td>
</tr>
<tr>
<td>Lichteiland Gouwe</td>
<td>36.779</td>
<td>438.793</td>
<td>51°55′33″</td>
<td>3°40′11″</td>
<td>LGE</td>
</tr>
<tr>
<td>Schouwenbank</td>
<td>11.244</td>
<td>419.519</td>
<td>51°44′48″</td>
<td>3°18′24″</td>
<td>SWB</td>
</tr>
<tr>
<td>Scheer west Wandelare</td>
<td>–7.797</td>
<td>380.645</td>
<td>51°23′32″</td>
<td>3°02′57″</td>
<td>SCW</td>
</tr>
</tbody>
</table>

[Source: RIKZ, http://www.gofflimaat.nl]
Table 4 Characteristics of 9 sites (N is the number of peaks, D(I) is the Wilks measure, and MD(I) is the Mahalanobis distance)

<table>
<thead>
<tr>
<th>Site</th>
<th>N</th>
<th>Acronym</th>
<th>L-CV</th>
<th>L-skewness</th>
<th>L-kurtosis</th>
<th>D(I)</th>
<th>MD(I)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>SCW</td>
<td>0.072</td>
<td>0.1191</td>
<td>−0.0316</td>
<td>1.96</td>
<td>5.21</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>MPN</td>
<td>0.0427</td>
<td>0.051</td>
<td>0.0725</td>
<td>2.43</td>
<td>6.47*</td>
</tr>
<tr>
<td>3</td>
<td>47</td>
<td>SWB</td>
<td>0.0485</td>
<td>0.2892</td>
<td>0.1423</td>
<td>1.35</td>
<td>3.61</td>
</tr>
<tr>
<td>4</td>
<td>33</td>
<td>LEG</td>
<td>0.0538</td>
<td>0.1924</td>
<td>0.0467</td>
<td>0.9</td>
<td>2.39</td>
</tr>
<tr>
<td>5</td>
<td>52</td>
<td>ELD</td>
<td>0.063</td>
<td>0.2246</td>
<td>0.1082</td>
<td>0.11</td>
<td>0.29</td>
</tr>
<tr>
<td>6</td>
<td>67</td>
<td>EUR</td>
<td>0.0606</td>
<td>0.2393</td>
<td>0.1477</td>
<td>0.45</td>
<td>1.19</td>
</tr>
<tr>
<td>7</td>
<td>72</td>
<td>K13</td>
<td>0.0613</td>
<td>0.2756</td>
<td>0.1139</td>
<td>0.17</td>
<td>0.44</td>
</tr>
<tr>
<td>8</td>
<td>58</td>
<td>SON</td>
<td>0.0692</td>
<td>0.2635</td>
<td>0.1509</td>
<td>1.23</td>
<td>3.29</td>
</tr>
<tr>
<td>9</td>
<td>61</td>
<td>YM6</td>
<td>0.0673</td>
<td>0.2876</td>
<td>0.078</td>
<td>0.42</td>
<td>1.11</td>
</tr>
</tbody>
</table>

From step (ii), the output was generated as shown in Table 5. These results indicate that the remaining 8 sites of the dataset are acceptably homogeneous according to the $H(1)$ statistics.

From step (iii), five three-parameter distributions Generalized Logistic (GL), Generalized Extreme Value (GEV), Generalized

The $H(2)$ and $H(3)$ statistics lack power to discriminate between homogeneous and heterogeneous regions.

In step (iii), five three-parameter distributions Generalized Logistic (GL), Generalized Extreme Value (GEV), Generalized
Pareto (GP), Lognormal (LN3), and Pearson type III (PEIII) were fitted to the regional data. Table 6 shows that the GP is the optimum distribution function, according to the Hosking and Wallis (1997) goodness-of-fit measure.

The regional frequency algorithm results in Table 7 and Fig. 7. From the regional quantile estimates, at-site quantile estimates can be determined by multiplying the regional quantile estimates with the at-site averaged values. The platform wave data is in cm. The RFA-based estimations can now be compared with the at-site estimations. From Fig. 7, the differences between the regional and at-site estimates can be in the order of 100 cm for the $10^{-4}$ quantile. Indeed the GP gives a satisfactory regional fit. It has a curvature downwards, indicating that there should be some maximum credible normalized extreme wave height.

### Table 6: Goodness of fit measures (based on 500 simulations)

<table>
<thead>
<tr>
<th>Distribution</th>
<th>L-moments</th>
<th>Z value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generalized Logistic (GL)</td>
<td>0.212</td>
<td>6.21</td>
</tr>
<tr>
<td>Generalized Extreme Value (GEV)</td>
<td>0.178</td>
<td>4.23</td>
</tr>
<tr>
<td>Lognormal (LN3)</td>
<td>0.165</td>
<td>3.48</td>
</tr>
<tr>
<td>Pearson type III (PEIII)</td>
<td>0.141</td>
<td>2.1</td>
</tr>
<tr>
<td>Generalized Pareto (GP)</td>
<td>0.097 $^{*}$</td>
<td>$-0.52$</td>
</tr>
</tbody>
</table>

### Table 7: Regional versus at-site parameter estimations of the GP based on L-moments

<table>
<thead>
<tr>
<th>Site no.</th>
<th>Acron</th>
<th>Location</th>
<th>Scale</th>
<th>Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regional GP</td>
<td>PCW</td>
<td>0.863</td>
<td>0.174</td>
<td>0.244</td>
</tr>
<tr>
<td>at-site 1 GP</td>
<td>SCW</td>
<td>318.7178</td>
<td>114.1414</td>
<td>0.5744</td>
</tr>
<tr>
<td>at-site 2 GP</td>
<td>SWB</td>
<td>405.0237</td>
<td>50.6725</td>
<td>0.1027</td>
</tr>
<tr>
<td>at-site 3 GP</td>
<td>LEG</td>
<td>451.8168</td>
<td>88.7547</td>
<td>0.3547</td>
</tr>
<tr>
<td>at-site 4 GP</td>
<td>ELD</td>
<td>483.6249</td>
<td>102.0713</td>
<td>0.2663</td>
</tr>
<tr>
<td>at-site 5 GP</td>
<td>EUR</td>
<td>414.5443</td>
<td>79.4354</td>
<td>0.2276</td>
</tr>
<tr>
<td>at-site 6 GP</td>
<td>K13</td>
<td>479.1845</td>
<td>81.9333</td>
<td>0.1358</td>
</tr>
<tr>
<td>at-site 7 GP</td>
<td>SON</td>
<td>461.3438</td>
<td>94.8344</td>
<td>0.1658</td>
</tr>
<tr>
<td>at-site 8 GP</td>
<td>YM6</td>
<td>448.9252</td>
<td>82.0869</td>
<td>0.1064</td>
</tr>
</tbody>
</table>

### 7 Discussion

A generalized Pareto distribution appears to be the optimum regional fit for the extreme wave heights at the Dutch North Sea coast. A convex curvature is present in the normalized regional growth curve leading to a regional upper limit of the significant wave height. This is physically relevant since the maximum wind-driven waves are limited by several energy dissipated processes, i.e., bottom friction induced wave reduction, white capping, wave breaking and wave-wave interactions.

Differences between the at-site quantile estimates and the regional quantile estimates can be as high as 100 cm for the extreme extrapolations. It is better to rely on the regional quantile estimates for decision making, as is recommended also by Hosking and Wallis (1997). For further investigation, the data base should be extended with measurements from other stations in the neighbourhood; due to the internet availability, it is relative easy to obtain wave height measurement at various locations in the world. Adding more sites to the existing data base of 9 stations on the Dutch North Sea will result in more accurate predictions of the quantile estimates, maintaining the homogeneity condition. In that case more advanced methods based on robust distances can be used in the discordancy analysis of the sites.

Inter-site correlations are certainly present in this case study. Hosking and Wallis (1988), however, mentioned that their RFA-procedure is quite robust against inter-site correlations for low heterogeneity measures. A heterogeneity measure of 1.13 was obtained herein (Table 5), indicating a low value of heterogeneity, and therefore inter-site correlations will have no major disturbance on the estimates.

### 8 Conclusions

This study showed that the RFA is essential to resolve the difficulties of conventional methods in the identification of the appropriate statistical distribution for describing environmental datasets and estimating the parameters of a selected distribution. The RFA resolved this problem by “trading space for time” because of extending the dataset length by using as much information from neighbouring locations as possible. The first requirement is that all sites in the RFA must come from a homogeneous region. Testing and forming homogeneous regions involves K-means clusters, discordance measures and robust distance measure. Other stages of the RFA are performed on the basis of L-moment statistics and L-moment algorithms.

The application of RFA to analyse maximum river discharges in North West and Central Europe showed that in the regions considered, eight initial clusters were derived, based on K-means clusters and a large number of sites. However, only six clusters
were qualified for being homogeneous regions with a total of 67 sites (40% of 168 considered sites), based on regional testing criteria such as the heterogeneous measures, the discordance distance and the robust distance.

The PE III distribution corresponds to the optimum fitting to the river discharge data. Regional growth curve and averaged regional growth curve are nearly identical and lie in between all at-site’s growth curves. Only the at site growth curve of site 29 had an exceptionally rapid upward trend. Thus, this site can be considered as an outlier and should be excluded in RFA.

The regional frequency analysis was used subsequently to examine the extreme sea waves along the Dutch North Sea coast. It is evident that the waves on different locations are drawn from the same parent distribution, except for station MPN. The distance between the locations is rather small compared with the dimensions of the North sea, and the physical phenomena causing waves were wind speed and wind direction, in combination with the bathymetry of the North Sea. Thus, homogeneity has to be present in the data. The site MPN was excluded for RFA based on the robust discordance distance test. This was already clear from the physical consideration that the site MPN is located in relatively shallower water where incoming waves often break before reaching the location, and the statistical measures have confirmed this expectation. It is recommended to include more sites along the North Sea coast of other countries such as Belgium, England, Germany, France, Sweden, Norway, and Denmark. Apart from the collection of foreign sites, the process of generation of new sites can also be continued by splitting the data. The inference based on robust distances and H-measures complement and agreed each other for these particular data sets. It is expected that in general the robust distances can be better used than the discordance measure $D_i$ of Hosking and Wallis (1997). This should be proved by means of an extended Monte Carlo investigation.

The RFA for the sea wave data demonstrated that the General Pareto distribution is the optimum for being the regional distribution, fulfilling the physical properties of an upper limit to the significant wave height.

The RFA results in so-called “added years” to the observation series of a single measurement station by pooling the data of neighboring locations. The number of added years increases with the number of neighboring locations in the homogeneous region, as well as with the distances between locations. Quantifying the number of added years is subject of further research and closely related to the quantification of confidence bounds around the quantile levels. The smaller confidence bounds, the larger number of added years which can be obtained by pooling the data.

**Acknowledgements**

This work was supported by the European Community’s Sixth Framework Programme through the grant to the budget of the integrated Project FLOODsite, Contract GOCE-CT-2004-505420 and the Delft Cluster Project C71D28 ‘Safety against Flooding’, sub-workpackage WP-A3-3. The following authors are kindly acknowledged for their collaboration of previous research work on distribution analysis: N. M. Neykov, Pl. Neytchev, J. K. Vrijling, H. Chbab, and J. de Ronde.

**Notation**

$b_i$ = Sample probability weighted moments, order $r$
$C(X)$ = Covariance estimates
$D_i$ = Discordance measure
$F$ = Cumulative probability function
$H_w$ = Individual wave height at deep sea
$H_s$ = Significant wave height of an individual wave record
$i_l$ = Sample $L$-moments order $i$
$MD^2_i$ = Classical Mahalanobis distance
$N$ = Number of observation stations
$n$ = Length of an at-site single dataset
$NDBC$ = National Data Buoy Center
$q_i(.())$ = Regional growth curve
$Q_i(F)$ = Quantile function at site $i$ in a region
$Q_i$ = Observed datasets
$T(X)$ = Mean estimates
$t_r$ = Sample $L$-moment ratio, order $r$
$w_i$ = Weight factor
$X_1, X_2, \ldots, X_n$ = Data sample

**Greek symbols**

$\mu_i$ = Site-dependent scale factor
$\beta_r$ = Probability weighted moments, order $r$ of a frequency distribution function
$\lambda_i$ = $L$-moments order $i$ of a frequency distribution function
$t_r$ = $r$th Order $L$-moment ratio of a frequency distribution function
$\Phi(.)$ = Cumulative distribution function of the standard Normal distribution

**References**


