Advances in Geosciences, 2, 25–30, 2005
SRef-ID: 1680-7359/adgeo/2005-2-25
European Geosciences Union

Advances in Geosciences

Wave-height hazard analysis in Eastern Coast of Spain – Bayesian approach using generalized Pareto distribution

J. J. Egozcue[1], **V. Pawlowsky-Glahn**[2], **and M. I. Ortego**[1]

[1]Dept. Matemàtica Aplicada III, U. Politècnica de Catalunya, Spain
[2]Dept. Informàtica i Matemàtica Aplicada, U. de Girona, Spain

Received: 24 October 2004 – Revised: 2 March 2005 – Accepted: 4 March 2005 – Published: 21 March 2005

Abstract. Standard practice of wave-height hazard analysis often pays little attention to the uncertainty of assessed return periods and occurrence probabilities. This fact favors the opinion that, when large events happen, the hazard assessment should change accordingly. However, uncertainty of the hazard estimates is normally able to hide the effect of those large events. This is illustrated using data from the Mediterranean coast of Spain, where the last years have been extremely disastrous. Thus, it is possible to compare the hazard assessment based on data previous to those years with the analysis including them. With our approach, no significant change is detected when the statistical uncertainty is taken into account. The hazard analysis is carried out with a standard model. Time-occurrence of events is assumed Poisson distributed. The wave-height of each event is modelled as a random variable which upper tail follows a Generalized Pareto Distribution (GPD). Moreover, wave-heights are assumed independent from event to event and also independent of their occurrence in time. A threshold for excesses is assessed empirically. The other three parameters (Poisson rate, shape and scale parameters of GPD) are jointly estimated using Bayes' theorem. Prior distribution accounts for physical features of ocean waves in the Mediterranean sea and experience with these phenomena. Posterior distribution of the parameters allows to obtain posterior distributions of other derived parameters like occurrence probabilities and return periods. Predictives are also available. Computations are carried out using the program BGPE v2.0.

1 Introduction

Hazard studies require estimation of model parameters, but normally estimation has to be carried out using few data. This causes an important statistical uncertainty on the parameter values. However, engineering practice often ignores this uncertainty and describes hazard by point estimators of parameters. A consequence of this kind of practice is that, when the data base is naturally augmented in time, the hazard parameters seem to change, and sometimes the change might be large. A case like this occurred at the Mediterranean coast of Spain. Several large storms occurred in the last 3 years with a considerable damage for coastal infrastructure, and therefore some practitioners – and even authorities – have claimed for updated wave-height hazard assessment to detect possible changes.

Our aim is to show that, when uncertainty of the estimates is taken into account, the detected changes on the hazard are not significant up to now. Hence, there is no justification – at least from our point of view – to attribute these last events to hypothetical climate changes.

To attain an estimate of statistical uncertainty in the estimation of parameters we need a model for occurrence of events and a procedure of estimation. We have selected the most simple model for event occurrence, the Poisson process, and magnitudes of events have been modelled by a Generalized Pareto Distribution (GPD). Thus, we have to estimate 3 parameters (Poisson rate, and scale and shape parameters for GPD) and a reference threshold. The used technique is essentially Bayesian.

Section 2 briefly describes the Poisson-GPD model. Section 3 presents the Bayesian estimation. Finally, in Sect. 4 we present results obtained from wave-height data of Palamós buoy.

2 The Poisson-Generalized Pareto Distribution model

The following model for events occurring in time and their magnitude is standard in hazard analysis. Details can be found in Embrechts et al. (1997), Davison and Smith (1990), and Grandell (1997). In a Poisson process, events are defined as time-points. In our case, we call such events storms, and they are defined as follows. A storm starts when the recorded wave-height in the reference device (a recording

Correspondence to: J. J. Egozcue
(juan.jose.egozcue@upc.es)

buoy) is greater than h_0=2 m and has been less than h_0 for at least 4 consecutive days. It finishes when the recorded wave-height is less than h_0 and remains at this level for at least 4 days. We define the magnitude of the storm as the maximum wave-height recorded during the storm and the occurrence time of the storm is taken as the time instant of such a record. Wave-height can be recorded in several ways. Here we adopt the spectral wave-height, H_{m0}. It is evaluated integrating the wave spectral density on positive frequencies and then taking the square root. It is similar to the so called significant wave-height, the mean of wave-heights conditioned to be higher than the 2/3-quantile of wave-height for individual waves.

The number of storms in a given time t is assumed to be Poisson distributed, $N\sim$Poisson(λ), and, therefore,

$$\mathrm{P}[N = n|\lambda, t] = \frac{(\lambda t)^n \exp[-\lambda t]}{n!} \ , \ n = 0, 1, 2, \dots \ , \qquad (1)$$

which defines a homogeneous Poisson process. For climatic events, we relax the validity of Eq. (1) to values of t being an integer number of years to avoid seasonality effects. The Poisson rate, λ, is the expected number of events in one year; its inverse, τ=1/λ, is the return period of events in years. In the process of estimation we use a new parameter, $z=-\log_{10}\lambda=\log_{10}\tau$. The reason for it is a matter of scale: we are mainly interested in very low values of λ, or very large values of τ, and we normally discriminate values of τ by its order of magnitude. This facilitates visualization, computation and point estimation.

Each storm is evaluated by the maximum H_{m0} recorded during the storm. Again, the scale of H_{m0} seems to be relative: a wave-height of 15.0 m is similar to another one of 15.5 m; but 1.0 m is clearly double of 0.5 m; furthermore, a sea of 0m wave-height is an idealized, impossible sea. These facts are accounted for by taking logarithms (in base 10 for easy reading of labels). Accordingly, we define $X=\log_{10} H_{m0}$ as a measure of storm-magnitude. Each event has a random value of X, which is assumed independent from event to event and from the Poisson process. Also, we assume that the X's for different events have the same distribution function F_X. There are difficulties in identifying a model for F_X, but the Generalized Pareto Distribution (GPD) provides a flexible enough model, particularly when excesses of X over a high threshold are considered, due to asymptotic properties of such a distribution (Pickands, 1975).

The excess of X over a reference threshold u, defined as $Y=X-u$ given that $X>u$, is modelled by a GPD, whose expression is

$$F_Y(y|\xi, \beta) = \left(1 + \frac{\xi y}{\beta}\right)^{-\frac{1}{\xi}} \ , \ 0 < y < y_{\sup} \ , \qquad (2)$$

where ξ is a shape parameter and β is a scale parameter (Embrechts et al., 1997). The value of ξ defines the so-called domain of attraction (DA). The Weibull DA corresponds to GPD's such that $\xi<0$ and $y_{\sup}=-\beta/\xi$, i.e. the support of Y, and hence of X, is limited with a finite upper tail. The Fréchet DA is characterized by $\xi>0$ and $y_{\sup}=+\infty$, and the support of Y becomes infinite. GPD's in the Fréchet DA have heavy upper tails, i.e. very large excesses are likely, although central values (mean or median) may be low. Finally, if ξ=0 and $y_{\sup}=+\infty$, the GPD belongs to the Gumbel DA and Eq. (2) takes the limit form

$$F_Y(y|\xi = 0, \beta) = 1 - \exp\left[-\frac{y}{\beta}\right] \ , \ 0 < y < +\infty \ , \qquad (3)$$

which is an exponential distribution.

Once the parameters of the GPD have been estimated, properties of the marked Poisson processes allow us to compute all hazard parameters, like return periods or exceedance probabilities in a defined lifetime. For instance,

$$\tau(x) = \frac{\tau(u)}{1 - F_Y(x - u|\xi, \beta)} \ , \ u_0 \le u \le x \ , \qquad (4)$$

where $\tau(x)$=1/$\lambda(x)$, is the return period of events for which X is larger then x. Also, non-exceedance probabilities of the threshold $x\ge u$ in a lifetime L are

$$\mathrm{P}[N(x) = 0|\lambda(u), \xi, \beta] = \exp[-\lambda(x)L] \ , \qquad (5)$$

where $\lambda(x)=\lambda(u)[1-F_Y(x-u|\xi, \beta)]$. We should remark that Eqs. (4) and (5) provide values of hazard parameters assuming that the distributional parameters $\lambda(u)$, ξ and β are known. However, as they are estimates, statistical uncertainty affects their values and, accordingly, hazard parameters, like those involved in Eqs. (4) and (5), become also random. Bayesian estimation provides a way of dealing with this issue.

3 Bayesian estimation

Once a suitable threshold u has been selected, the estimation of $z=z(u)=-\log_{10}\lambda(u)$, ξ, and β, is required. According to the Bayesian paradigm, they are assumed to be random variables. Their joint probability densities, $f_{z\xi\beta}(z, \xi, \beta)$ and $f_{z\xi\beta}(z, \xi, \beta|D)$, account for their uncertainty before (*prior*) and after (*posterior*) the data sample, symbolized by D.

Prior density represents our knowledge about parameters previous to D. A further assumption is that $z(u)$ is independent from (ξ, β), i.e. $f_{z\xi\beta}(z, \xi, \beta)=f_z(z)\cdot f_{\xi\beta}(\xi, \beta)$. Bayes' theorem is then

$$f_{z\xi\beta}(z, \xi, \beta|D) = L(z, \xi, \beta|D) \cdot f_z(z) \cdot f_{\xi\beta}(\xi, \beta) \ , \qquad (6)$$

where the likelihood of the data can also be factorized as

$$L(z, \xi, \beta|D) = \mathrm{P}[N(u) = n|z, t_0] \cdot \prod_{j=1}^{n} f_Y(y_j|\xi, \beta) \ , \qquad (7)$$

where D has been made explicit as the number $N(u)=n$ of excesses $y_j=x_j-u$ over u.

The posterior density in Eq. (6) is itself the result of the Bayesian estimation, but it is also the basis to obtain the distribution of hazard parameters, like return periods (Eq. 4), occurrence probabilities (Eq. 5), or others. As an interesting example, assume that the estimated GPD is in the Weibull

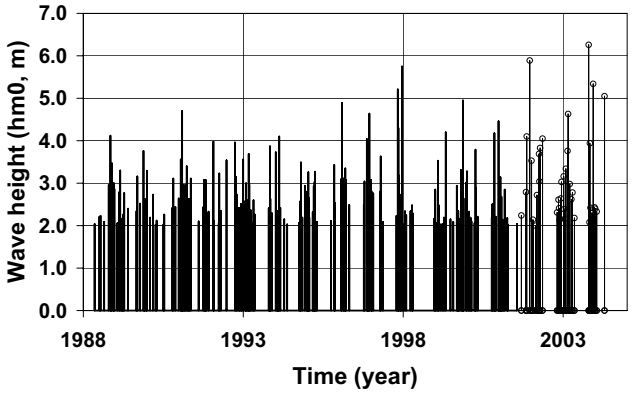

Fig. 1. Storms at Palamós buoy. Vertical bars represent H_{m0} in meters. Last years, assumed more hazardous, are marked with circles. First period, no markers.

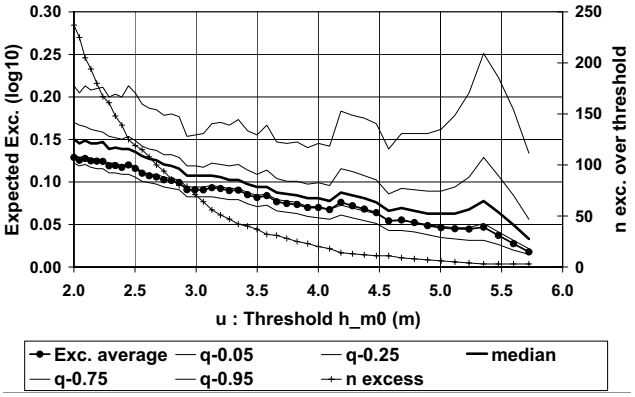

Fig. 2. Estimates of the mean excess over threshold. Circle marker, sample estimate; thick line, median of the posterior estimate; thin lines, 0.05, 0.25, 0.75, 0.95 quantiles of the posterior. Cross marker, number of excesses over threshold, scale on the secondary axis.

DA with probability 1, i.e. the storm-magnitude X is surely limited by $y_{\text{sup}}=-\beta/\xi$. Since ξ and β are considered random and jointly distributed as

$$f_{\xi\beta}(\xi,\beta|D) = f_{\xi\beta}(\xi,\beta)\cdot\prod_{j=1}^{m} f_Y(y_j|\xi,\beta)\ , \tag{8}$$

y_{sup} is also random and can be described by its probability density. A simulated sample of ξ and β generates a sample of y_{sup}. From this derived sample, central tendency parameters, like the median, provide point estimates, and the sample quantiles determine credible intervals for y_{sup}. These type of estimates of hazard parameters (return periods, exceedance probabilities, y_{sup}) will be used in Sect. 4.

Prior density for $z(u)$ has been assumed uniform for a very wide range of values, and $f_{\xi,\beta}$ has been assessed following the methods developed in Egozcue and Ramis (2001), which will be commented in the next section.

4 Case study: wave-height at Palamós buoy

Spanish port authority (Puertos del Estado, Spain) maintains buoy networks to record sea-waves and has provided the data used in this study. We have selected the buoy (Waveraider) placed at (41°49.8′ N, 3°11.2′ E) near Palamós (Catalonia, Spain). It records strong storms in the Eastern coast of the Iberian peninsula, preferably caused by strong North and East winds. Strong East winds are normally associated with heavy convective precipitation. After isolating storms as explained in Sect. 2, available data span 16 years, from May, 1988, to April, 2004. The maximum H_{m0} recorded was 6.26 m and it was recorded on 17 October 2003. Figure 1 shows these data. From September 2001 onwards, events are marked with circles. This period of 2 years and 7 months has been perceived as more hazardous than previous years, and we will compare the results for the first period of 13 years and 4 months (called short data set, S-data) with estimations obtained using the whole data set of 16 years (whole data set, W-data).

A first step in the estimation process is to determine a reference threshold for excesses. Storms have been defined with H_{m0} greater than 2 m, i.e. $X=\log_{10} H_{m0}>u_0=0.3010$. We look for a reference threshold $u\geq u_0$ such that excesses over it fit reasonably well to the GPD. A way to check this fit is to estimate the mean excess over a threshold. This mean excess, as a function of the threshold, is linear whenever the distribution of excesses is GPD with shape parameter $\xi<1$. The expression of the mean excess of X over u_1 GPD distributed is

$$\mathrm{E}[X-u|X>u] = \frac{\beta+\xi u}{1-\xi}\ ,\ u\geq u_1.$$

This fact is used to graphically check the fit to the GPD. The fit to a Weibull DA GPD is reflected in a negative linear trend of the mean excess ($\xi<0$). Positive linear trends correspond to Fréchet DA. We have estimated the mean excess over several thresholds. Figure 2 shows the result of two preliminary estimations of the mean excess for W-data: sample average of excesses (Embrechts et al., 1997) and a preliminar Bayesian approach; the posterior distribution of the mean excess is represented by the median (thick line, no markers) and some quantiles to give an idea of the uncertainty of the estimate (Egozcue and Tolosana-Delgado, 2002). As shown in Fig. 2, both estimators, the sample mean and the median of the posterior estimators, seem to be quite linear with increasing thresholds, thus suggesting a good fit for a reference threshold to be $u=u_0=0.3010$. A similar result is obtained for S-data.

Bayesian estimation allows us to define a prior density for the three parameters z, ξ, β. This is the Bayesian way of taking into account available information about the process that is previous to the data sample. As commented before, $f_z(z)$ is assumed uniform in a wide range of values and, therefore, its influence will be negligible. In turn, we put important information into $f_{\xi\beta}$, most of it delimiting the admissible domain of the parameters (Egozcue and Ramis, 2001). Figure 3 shows contours of the selected prior. The domain of the

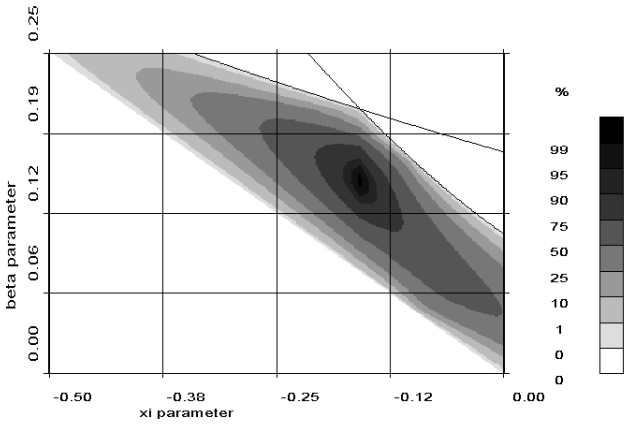

Fig. 3. Contours of the flat prior density in the (ξ, β) plane used in the estimation process. The boundary of the admissible domain is based on prior information.

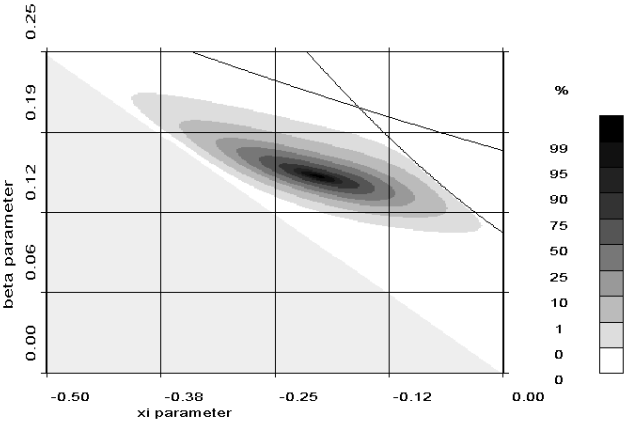

Fig. 4. Contours of (ξ, β)-posterior density for W-data. Probability is well centered in the assumed admissible domain.

ξ parameter is completely placed in the Weibull DA, as we assume the log-wave-height to be bounded. This assumption is physically based: wave-height is bounded by wave breaking, and wave breaking is connected to water-depth, which is clearly limited. Also, and more drastically, the fetch and the limited velocity of the wind cause boundedness of wave-height. We also assume events with H_{m0} as high as 8.0 m should be possible at this buoy. Consequently, GPD's (Weibull DA) for which this value is not attainable have been rejected. This determines the line in Fig. 3 which defines the lower boundary of the admissible domain. Furthermore, $H_{m0}>15$ m has been supposed almost impossible, and we assign a probability less than 10^{-5} to these events. This determines the curved boundary at the right of the figure. Similarly, we assume that the probability of a storm having H_{m0} greater than 5m is less than 0.1; this bounds the domain in the upper values of β. Finally, we assume that the density of excesses should decay more rapidly than a triangular probability density, $\xi>-0.5$.

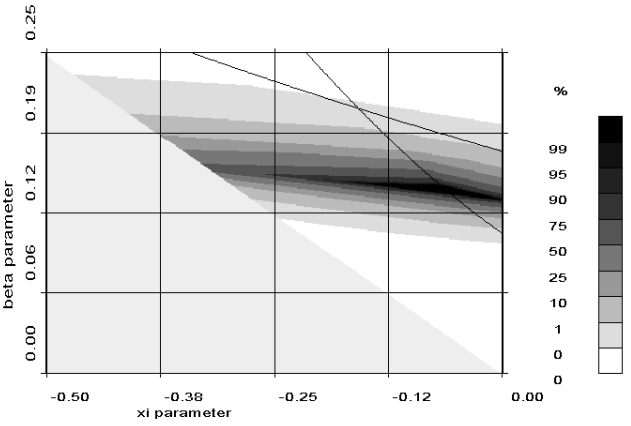

Fig. 5. Contours of significance of the Kolmogorov-Smirnov goodness of fit test for W-data.

The prior model for ξ, β is completed by fitting a flat density that is null in the above mentioned boundaries, as shown in Fig. 3. It is approximately centered at the GPD assigning probability 0.7 to storms with H_{m0} less than 3.0 m, a rough prior estimate of the most credible event.

The (ξ, β)-part of the contours of the posterior density are shown in Fig. 4 for W-data. When using S-data, a similar figure is obtained, although it is a little bit wider (less data) and the mode is placed at a slightly lower value of beta. The area containing representative probability of the posterior density is mainly determined by the amount of information, wider for less data. The fact that the mode is not very close to the border of the domain points out that the (log)-data correspond clearly to GPD's in the Weibull DA (finite tail), in agreement with our assumption.

In order to validate the fitting of GPD to the data, we have tested the goodness of fit using the Kolmogorov-Smirnov test. Figure 5 shows contours for the significance of that test for each (ξ, β) pair. The region with significance greater than 0.05 (good fit) covers the location of the posterior probability shown in Fig. 4, thus confirming a good fit for likely values of GPD-parameter values.

Once the posterior density of the three parameters in Eq. (6) has been obtained, we proceed to generate 1000 samples of (z, ξ, β). From these simulated samples several hazard parameters can be estimated. We present a comparison of exceedance probabilities in 50 years (Eq. 5), return periods (Eq. 4) and upper limit of the support of H_{m0} for both S and W-samples.

Figure 6 shows, for each threshold, the probabilities of exceedance in 50 years. A difference between the result for the W-sample and the S-sample can be observed: medians of the posterior exceedance probabilities apparently differ (curves with triangles, W-sample; curves with squares, S-sample). Exceedance probabilities seem to be higher for W-sample as suggested by Fig. 1. However, the median estimator for the S-sample follows approximately the 0.25 quantile of the W-sample. This means that, when taking into account the uncertainty of both estimates, we are unable to clearly distinguish

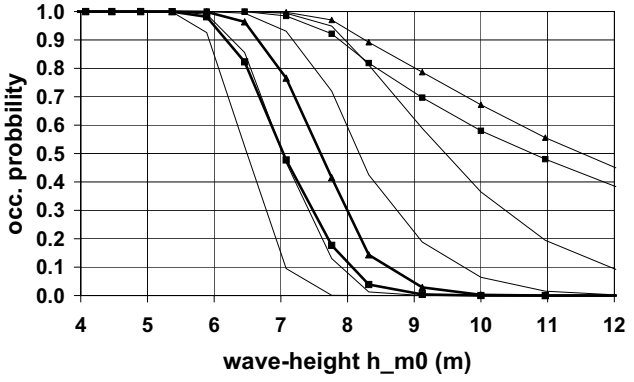

Fig. 6. Exceedance probabilities in 50 years for each threshold. Thick lines, median of the posterior estimate: triangles, W-data; squares, S-data. Thin lines, posterior quantiles 0.05, 0.25, 0.75 and 0.95 for W-data. Thin lines with markers: probability of the threshold to be attainable; triangles, W-data; squares, S-data.

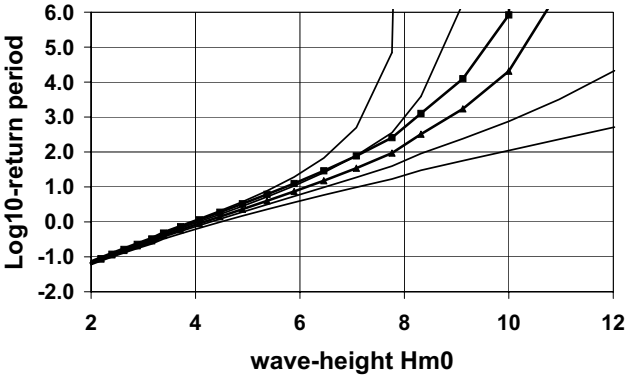

Fig. 7. Posterior estimate of $\log_{10}$-return periods. Median of the posterior estimate: triangles, W-data; squares, S-data. Lines without marker, posterior quantiles 0.05, 0.25, 0.75 and 0.95 for W-data.

between the results from the two samples. Figure 6 also shows estimated probabilities of each threshold being attainable for the two samples. They also differ from sample to sample, but differences are not substantial.

Figure 7 presents the estimated ($\log_{10}$) return periods for the W-sample (median with triangles, quantiles without markers). The increase of the statistical uncertainty with H_{m0} is clear from the separation of quantiles. These curves show a tendency to verticality. The reason for this is that the estimated GPD's are in the Weibull DA, i.e. they have an upper limit and values of H_{m0} behind them are not attainable. The estimates of the ($\log_{10}$) return periods (median; curve with squares) for the S-sample is also plotted in Fig. 7. As commented for exceedance probabilities, these differences, being appreciable, do not allow a rejection of stationarity of the extremal process.

Estimation of upper limit of a distribution is always difficult because estimators normally behave poorly. However, Bayesian estimation allows to describe the statistical uncertainty, showing the reasons for this bad performance. Figure 8 shows the posterior densities of y_{sup} obtained for both samples. Again some differences can be observed; for instance, the mode for the S-sample is slightly smaller than that for the W-sample. However, both remain around 8.20–9.05 m, which is a small difference when compared with the dispersion of the estimated densities. A remarkable fact in this estimation is that the mode of y_{sup} has probability $\sim$0.85 of being attainable, as shown in Fig. 6, i.e. the probability that the true upper limit is less than 8.20–9.05 m is about 0.15.

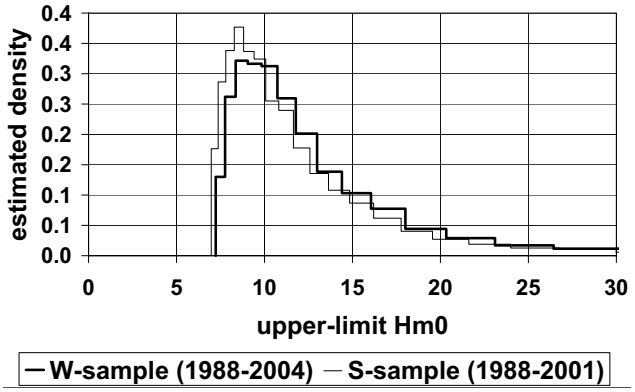

Fig. 8. Estimated posterior density of the upper limit of the support of H_{m0}. Thick line, W-sample; thin line, S-sample.

Table 1. Estimated return period in years (median and 0.05, 0.95 quantiles) for 7 m H_{m0} and for both S and W-samples.

threshold	sample	0.05-quantile	median	0.95-quantile
7.0	S	14.7	77.0	9720.0
7.0	W	9.8	34.4	498.0

All these estimates of hazard parameters describe the general characteristics of the wave-height hazard in Palamós, but one should take into account that the main characteristic is the statistical uncertainty of the estimated parameters due to the very limited span of the observation-time. In order to remark this fact, Table 1 gives the estimated return period in years (median and 0.05, 0.95 quantiles) for H_{m0}=7 m and for both samples.

5 Conclusions

The Bayesian analysis of the extremal series of storms at the Palamós buoy (16 years record) allowed to estimate both hazard parameters and their uncertainty. The series was modelled as a Poisson process, which events are storms, marked by the maximum attained wave-height (spectral height). The $\log_{10}$ wave-height was modelled by a Generalized Pareto Distribution. Both information from the sample and a prior were used. Estimated hazard parameters are affected by a large uncertainty.

The series from May, 1988 to April, 2004 (W-sample) was compared to a shorter one (S-sample) with endpoint at September, 2001. Observations in the last 2.6 years of W-sample were perceived as more hazardous than those of the S-sample. Differences confirming this impression have been observed. However, these differences are not enough to statistically state differences between the two series. This is mainly due to the above mentioned uncertainty affecting all estimates.

Acknowledgements. The authors thank O. Serrano (Puertos del Estado, Spain) for his encouragement and advice. Data were provided by Puertos del Estado, Spain. Research was supported under two agreements between Puertos del Estado, Spain and Universidad Politécnica de Cataluña, Spain. This research has received support from the Dirección General de Investigación of the Spanish Ministry for Science and Technology through the project BFM2003-05640/MATE.

Edited by: L. Ferraris
Reviewed by: anonymous referees

References

Davison, A. C. and Smith, R. L.: Models for exceedances over high thresholds, J. Roy. Statist. Soc. B, 52, 393–442, 1990.

Egozcue, J. and Ramis, C.: Bayesian Hazard Analysis of Heavy Precipitation in Eastern Spain, International Journal of Climatology, 21, 1263–1279, 2001.

Egozcue, J. and Tolosana-Delgado, R.: Program BGPE: Bayesian Generalized Pareto Estimation, edited by: Diaz-Barrero, J. L., ISBN 84-69999125, Barcelona, Spain, 2002.

Embrechts, P. C. K. and Mikosh, T.: Modeling Extremal Events, Springer Verlag, Berlin, Germany, 1997.

Grandell, J.: Mixed Poisson Processes, Chapman & Hall, London, UK, 1997.

Pickands, J.: Statistical inference using extreme order statistics, Annals of Statistics, 3, 119–131, 1975.