ON THE TWO STEP THRESHOLD SELECTION FOR OVER-THRESHOLD MODELLING

Pietro Bernardara (1,2), Franck Mazas (3), Jérôme Weiss (1,2), Marc Andreewsky(1), Xavier Kergadallan(4), Michel Benoît(1,2), Luc Hamm(3)

In the general framework of over-threshold modelling (OTM) for estimating extreme values of met-ocean variables, such as waves, surges or water levels, the threshold selection logically requires two steps: the physical declustering of time series of the variable in order to obtain samples of independent and identically distributed data then the application of the extreme value theory, which predicts the convergence of the upper part of the sample toward the Generalized Pareto Distribution. These two steps were often merged and confused in the past. A clear framework for distinguishing them is presented here. A review of the methods available in literature to carry out these two steps is given here together with the illustration of two simple and practical examples.

Keywords: peaks-over-threshold, extreme storm surges, extreme waves, declustering, extreme value theorem

INTRODUCTION

The accurate estimation of the probability of occurrences of extreme natural events is important for the protection of coastal areas. Traditionally, the estimation of these extreme events is done fitting a given probability distribution on a sample of historical observations or model output for a given phenomenon observed at a given site, usually as temporal series of the event-describing variable. The extreme value theory (Pickands 1975) offers a valid theoretical background. In particular, the extreme value central limit theorem states that the values exceeding a given threshold converge through a Generalized Pareto Distribution (GPD) if the original sample is composed by independent and identically distributed values (Pickands 1975). Following this theoretical result, the analysis of a sample of values exceeding a given threshold widely spread in extreme values analysis of natural events, together with the application of the GPD.

The observations of time series of environmental variables (e.g. temperature, wind, rainfall, discharge, sea level, sea surges), however, show strong temporal autocorrelation. Traditionally, in order to select independent events for the following statistical analysis, the concept of a physical threshold to overcome for defining an "extreme event" was widely used, and the Peaks-Over-Threshold (POT) sampling widely spread in the literature, see (Coles 2001; Smith 1984; Lang, Ouarda et al. 1999) among many others.

In the past (Lang, Ouarda et al. 1999), the threshold for the physical selection of independent extreme events and the threshold for the statistical sampling of extreme value asymptotically convergent toward GPD were confused and the same threshold was used both for sampling data and for meeting the hypothesis of extreme value theory.

Thus we propose to introduce a two-step threshold selection framework for over-threshold modeling (OTM), aiming to discriminate first a "physical threshold" for the selection of extreme and independent events (declustering) then a "statistical threshold" through an optimization procedure for the coherence with the hypothesis of the extreme value theory.

Though this framework is valid for all kinds of environmental variables and, more generally, for extreme analysis of time series of any variable, focus is made here on met-ocean variables (waves, surges, wind, sea levels…). Applications to an extreme surge analysis and to extreme wave analysis are thus presented for illustration.

TWO STEP THRESHOLD SELECTION FRAMEWORK DEFINITION

Physical declustering

Let call $Z_i(t)$ a time series (or, more generally, a spatial field) of realizations of a given natural variable Z (e.g. a significant wave height, a sea level, a surge, a wind speed…) at a given resolution Δt (i.e. time step).

[1] EDF R&D, Laboratoire National d'Hydraulique et Environnement (LNHE), 6 quai Watier, BP 49, 78401 Chatou cedex, France

[2] Université Paris Est, Saint-Venant Laboratory for Hydraulics, 6 quai Watier, BP 49, 78401 Chatou cedex, France

[3] ARTELIA, Grenoble, France

[4] CETMEF, Brest, France

The *physical declustering* aims to identify a sample of independent and identically distributed (i.i.d) values X_i from the time series $Z_i(t)$. The physical declustering can be viewed as an identification procedure of “physical independent extreme observations” of the studied phenomenon. In this step, one should identify, through purely physical considerations, what will be called here *events*: storms, extreme surges or sea levels… The events can be characterized by a given duration that can be longer that the resolution Δt of the time series. In this case the sets of consequent values of the variable, belonging to same event, are called *clusters*. For each event, one should then define a random variable X describing the event, or a given characteristic of the event. One should note that X may represent instantaneous values (e.g. the maximum, or “peak”, of the event) or may be a mathematical transformation over the event duration (e.g., in hydrology, the daily average discharge of a river or the volume of the flood). A sample of N_T independent and identically distributed (i.i.d) values X_i is thus obtained. Its size is lower than the size of the time series $Z_i(t)$. The difference can be huge: typically, for three-hourly time series over several years, the sample size can decrease from 10,000 – 100,000 to 10 – 1,000. Note also that the X_i do not depend on the time any more. The definition of the actual variable to be used depends on the natural phenomenon involved, on the available data and on the aim of the study. The appropriate physical declustering technique also depends on the characteristics of the given natural variable and on the physic and dynamic characteristics of the observed process. For instance, the declustering of a wave height time series will require different techniques for cyclonic events or for mid-latitude storms. In general, the knowledge of the physics of the studied phenomenon drives the physical declustering choices.

Statistical optimization

Let define the random variable $Y = X - u_s$, given $X > u_s$. Y is the exceedance of X above u_s, which stands for *statistical threshold*. Thus a sample Y_i, of size N, can be defined from the sample X_i: $Y_i = X_i - u_s$, given $X_i > u_s$.

Within the theoretical framework of Over-Threshold-Modeling, and in particular according to the extreme value central limit theorem, the probability distribution of the Y_i must converge through the Generalized Pareto Distribution (GPD) provided that u_s is high enough.

The sub-sample of the Y_i contains the i.i.d. extreme values to be modeled by GPD and whose extrapolation will yield the estimated return levels (or extreme quantiles). The cumulative distribution function of the 2-parameter GPD is given by:

$$F_{Y;k,\sigma}(y) = 1 - \left(1 + k\frac{y}{\sigma}\right)^{-1/k} \qquad (1)$$

Where k is the shape parameter and σ is the scale parameter, with $y > 0$ for $k > 0$ and $y < -\sigma/k$ for $k < 0$.

The u_s threshold selection step is called here *statistical optimization.* The statistical optimization step is a purely statistical problem for which several methods have been proposed in the literature. It does not depend on the particular random variable (environmental or not) and it is general for every extreme value application.

Overview of the framework

A general overview on the two-step framework is depicted in Fig. 1 below.

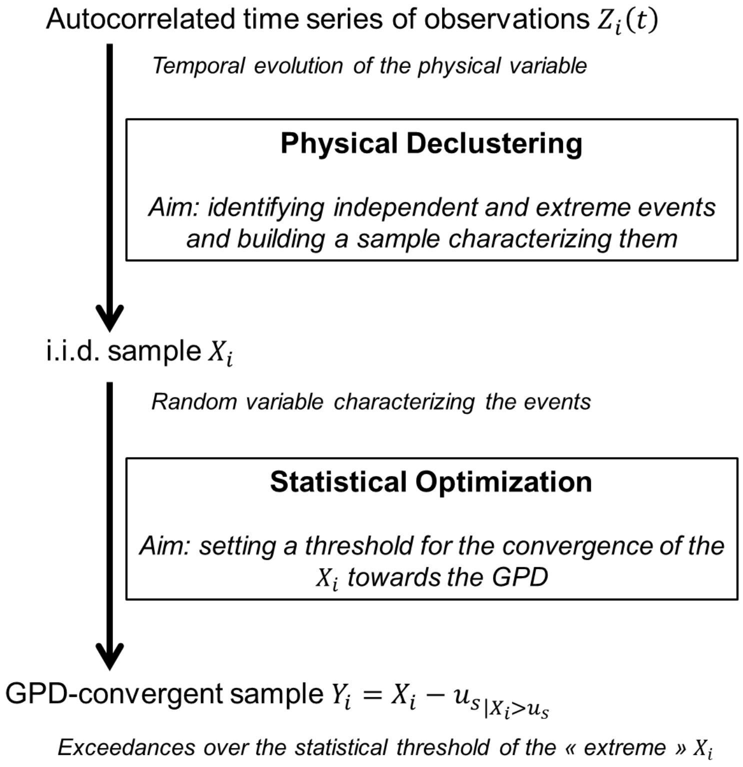
Autocorrelated time series of observations $Z_i(t)$
Temporal evolution of the physical variable
Physical Declustering
Aim: identifying independent and extreme events and building a sample characterizing them
i.i.d. sample X_i
Random variable characterizing the events
Statistical Optimization
Aim: setting a threshold for the convergence of the X_i towards the GPD
GPD-convergent sample $Y_i = X_i - u_{s|X_i>u_s}$
Exceedances over the statistical threshold of the « extreme » X_i

Figure 1. Overview of the two-step framework for OTM

Double threshold approach

The physical declustering step does not necessarily require the “over-threshold” concept. However, the techniques based on the exceedances of a threshold are quite intuitive and are widely spread.

In order to illustrate the proposed framework, and in coherence with the literature, we propose here to provide both physical declustering and statistical optimization trough a threshold approach, but distinguishing the two steps of the OTM as depicted in Fig. 1.

Thus, a first threshold u_p, called “physical threshold”, is used for physical declustering and the second threshold, u_s, comes from the statistical optimization procedure. This approach requires the hypothesis that $u_p \leq u_s$.

This approach is applied to two case studies, a wave height study and a regional frequency analysis of extreme skew surges (Bernardara, Andreewsky et al. 2011). That illustrates the fact that the methodological framework is valid in different fields of met-ocean hazard estimation, and more widely for any extreme analysis of time series.

CASE STUDY

Waves

We consider in this first illustrative example a time series of simulated three-hourly significant wave heights H_s off Marseilles, France (5.3104 W; 43.3460 N; water depth: 34 m). The duration of the data is $K = 13$ years, and the size of the time series is $n = 38005$ data. In order to ensure the homogeneity of the data (identically distributed), a decomposition of the sea states have been performed and only the swell components have been retained. The H_s time series is plotted in Fig. 2.

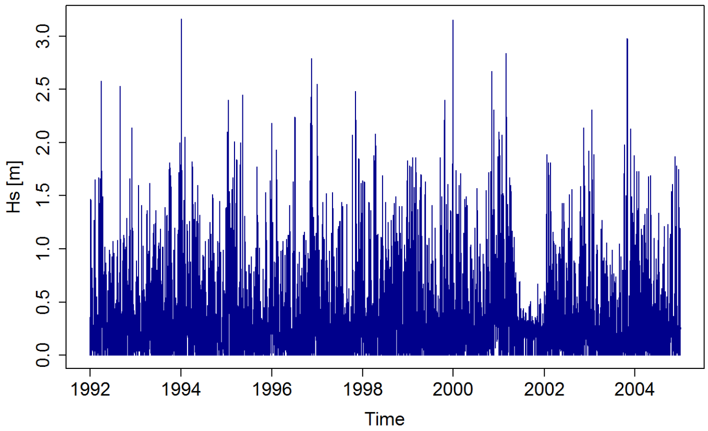

Figure 2. Swell significant wave height time series off Marseilles

The declustering procedure has been performed so as to obtain a sample of 10 storms per year in average, which is a physically sounding number of extreme events per year for the region. The physical threshold is thus set to $u_p = 1.4$ m, yielding $N_T = 130$ storm peaks X_i. A minimal duration of 24 hours between two storms has been set to ensure their independence. Furthermore, a minimal storm duration of 6 hours has been set (because very short events do not cause important damage to coastal structures) and fluctuations below the threshold within a same storm has been allowed for less than 12 hours. These parameters can be considered relevant for the physical threshold as chosen but it would not be the case for a threshold yielding one or two storms per year in average.

The statistical optimization has been performed by studying the stability of the GPD shape parameter k and modified scale parameter $\sigma^* = \sigma - k u_s$ with respect to the statistical threshold u_s (Fig. 3), following the method described by (Mazas and Hamm 2011). On secondary axis, the value of λ (the mean number of Y_i per year) is given so as to visualize the evolution of the (normalized) sample size. A first “domain of stability” can be seen between roughly 1.5 and 1.8 m, then a second one between 1.87 and 2.2 m. Afterwards the sample size is too short and the parameter uncertainty is too great. The bias minimization requires to choose the highest domain of stability while the variance minimization needs as much data as possible: consequently the statistical threshold is set to $u_s = 1.87$ m, yielding $N = 43$ and $\lambda = 3.31$.

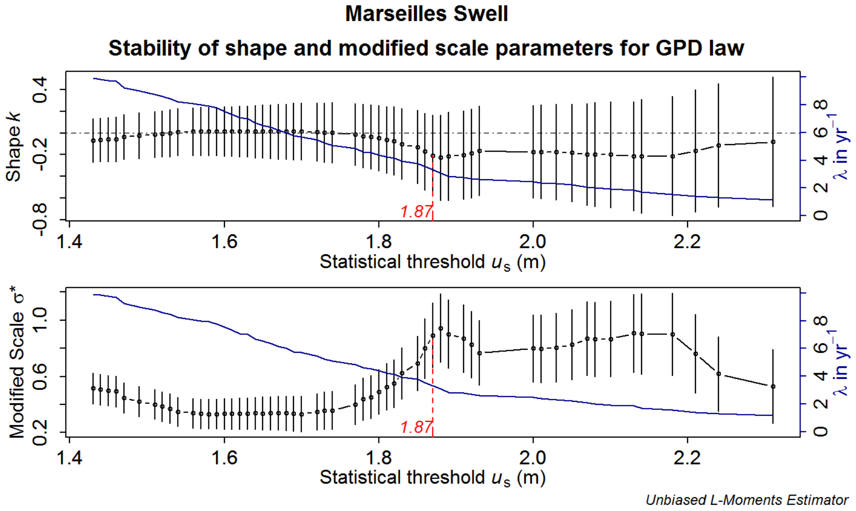

Figure 3. Stability of GPD shape and modified scale parameters for Marseilles’swell sample

The GPD parameters are estimated by the L-moments estimator. The resulting fit is illustrated in Fig. 4.

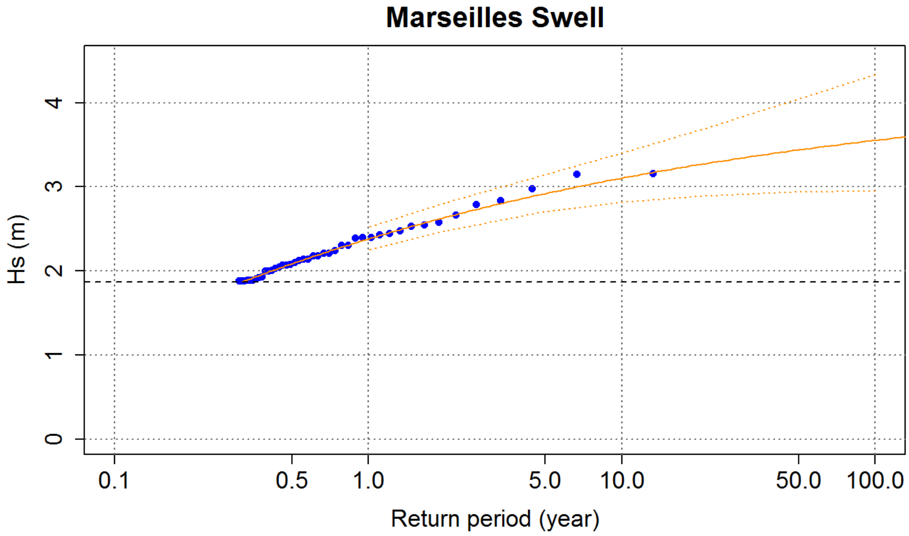

Figure 4. GPD fit for the swell off Marseille

Regional frequency analysis of skew surges

The regional frequency analysis (Cunnane 1988; Hosking and Wallis 1997) aims at gathering together the information from different sites to estimate a regional probability distribution for the occurrences of extreme values of a given phenomenon on the selected region. In particular, (Bernardara, Andreewsky et al. 2011) published recently a regional frequency analysis study on the skew surges on French coasts, showing that regional frequency analysis is a valuable solution because this phenomenon is homogenous enough in the region. (Bernardara, Andreewsky et al. 2011) uses a double threshold approach. In particular, a first physical threshold is used for the physical declustering of 18 skew surges time series observed at 18 sites along the French coasts. Here, u_p is estimated locally in order to obtain in average one skew surge per year. This value is lower than the average number of extreme surge in the region, but the regional approach, allowing collecting data from different sites, consents to reduce the information to the extreme part of the distribution. An illustration for the site of Arcachon is given in Fig. 5.

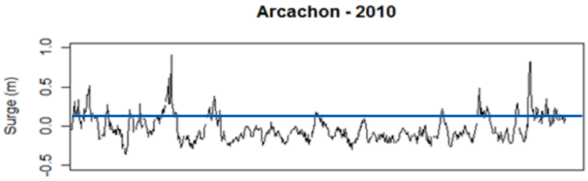

Figure 5. Illustration of physical declustering for a skew surges time series at Arcachon

Then a regional sample is built, containing the normalized skew surges from all sites. The statistical optimization for the choice of u_s is done on the regional sample, maximizing a χ^2 criterion for the fit of a GPD. See (Bernardara, Andreewsky et al. 2011) for details and results.

Note that, in this case, not only physical declustering and statistical optimization are treated as separated steps, but physical declustering is applied at local scale while statistical optimization is applied at regional scale.

CONCLUSIONS

This paper clarifies the general framework for an “over the threshold” exceedances modeling, distinguishing the physical declustering procedure from the statistical optimization.

We claim that the methodological separation of these two steps is useful and even necessary to avoid inconsistencies and to apply appropriately the existing threshold selection methods.

In particular, the double threshold approach is a first simple example of practical application of the framework, based on well-known threshold selection procedures.

The application to the two different met-ocean case studies, including a regional approach, demonstrates that the framework is relevant.

ACKNOWLEDGMENTS

The permission to publish the results of this ongoing research study was granted by the EDF (Electricité De France). The results in this paper should, of course, be considered as R&D exercises without any significance or embedded commitments upon the real behavior of the EDF power facilities or its regulatory control and licensing.

REFERENCES

Bernardara, P., M. Andreewsky, et al. 2011. Application of the Regional Frequency Analysis to the estimation of extreme storm surges. *J. Geophys. Res.*, 116(C02008).

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

Cunnane, C. 1988. Methods and merits of regional flood frequency analysis. *Journal of Hydrology*, 100, 269-290.

Hosking, J. R. M. and J. R. Wallis. 1997. Regional Frequency Analysis. An approach based on L-moments. Cambridge, Cambridge University Press.

Lang, M., T. B. M. J. Ouarda, et al. 1999. Towards operational guidelines for over-threshold modeling. *Journal of Hydrology*, 225, 103-117.

Mazas, F., and L. Hamm. 2011. A multi-distribution approach for determining extreme wave heights. *Coastal Engineering*, 58, 385-394.

Pickands, J. 1975. Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics*, 3(1), 119-131.

Smith, R. L. 1984. Threshold methods for sampling extremes. *Statistical extremes and applications*, T. d. O. J. Dordrecht, 621-638.